uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Hoonbin Hong
AUTEUR DE LA THÈSE / AUTHOR OF THESIS

Ph.D. (Mechanical Engineering)
GRADE / DEGREE

Department of Mechanical Engineering
FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Rotating Machinery Monitoring:
Feature Extraction, Signal Separation, and Fault Severity Evaluation

TITRE DE LA THÈSE / TITLE OF THESIS

Ming Liang
DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Bertrand Jodoin

Rong Liu

Michel Labrosse

Prakash Patnaik

Gary W. Slater
Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

Rotating Machinery Monitoring:

Feature Extraction, Signal Separation, and Fault Severity Evaluation

Hoonbin Hong

A thesis submitted to the Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Mechanical Engineering
Ottawa-Carleton Institute for Mechanical and Aerospace Engineering
University of Ottawa
Ottawa, Canada

December 2007

Your file Votre référence
ISBN: 978-0-494-49358-8
Our file Notre référence
ISBN: 978-0-494-49358-8

Abstract

Machine health monitoring systems (HMS) have been implemented for detection and diagnosis of different machinery faults to improve safety and reliability and cut down cost. However, the accuracy and timeliness of many such systems are still inadequate due to various noises and the interference of multiple signals. Therefore, many such systems tend to be either sluggish in fault detection or over-vigilant leading to false alarms. In addition, the literature has shown a lack of dependable and consistent fault severity measures for different defects. The existing fault severity measures often require comparisons with reference data. However, the reference data are not always available particularly for new machines or new operations. This obviously undermines their usefulness. This often causes either over-vigilance or slow actions, hence further hindering the diagnosis process. To provide accurate and timely detection and diagnosis decisions, this study focuses on three important subjects of fault diagnosis, i.e., fault detection, fault isolation, and fault severity assessment.

To improve the accuracy of fault detection, this study proposes three approaches: 1) A joint continuous wavelet transform (CWT) and autocorrelation method for extracting periodic fault features from a noisy signal mixture; 2) A kurtosis-based wavelet thresholding algorithm for de-noising and for detecting impulsive fault features; and 3) A generalized differentiation method for in-line oil-debris signature identification. The joint CWT and autocorrelation method can effectively remove fake (or spurious) impulses and hence improve the accuracy in detecting true fault signatures. Unlike many existing de-noising methods, the kurtosis-based wavelet thresholding algorithm can reduce both Gaussian noise and spurious impulses. Early detection of incipient fault is therefore possible due to the substantially purified signal. Different from the fault signatures in rotating machine elements, the oil-debris signature does not appear periodically but has a unique shape. Based on the generalized differentiation method, a signature extractor and an inflexion point locator are devised to extract in-line oil-debris signatures. The inherent low-pass filter property of two operators enables the algorithm to extract the oil-debris signature from the noisy signal without applying extra filters.

To identify multiple faults from a single data set from a rotating machine, a single-

channel signal separation method is proposed. Since the number of sources is not *a priori* information, the algorithm is suited for applications to mechanical fault diagnosis in which the number of fault sources is unknown.

To develop a reliable and consistent fault severity measure, the Lempel-Ziv complexity has been exploited and applied to the output from the CWT of the signal. A normalized Lempel-Ziv complexity gives a non-dimensional value between zero and one [0, 1], close to zero for a pure sinusoidal signal and one for a white Gaussian noise. The main objective is to develop a non-dimensional fault severity measure that is sensitive to fault severity and can be used in different applications without major modifications. Another objective is that the fault severity measure should be developed such that it can be used alone without relying on another set of reference data.

The proposed techniques for de-noising, fault detection and isolation, and fault severity have been evaluated in our lab or using data from the literature. The results have shown that the proposed methods can effectively purify signals, detect faults from vibration and oil-debris signals, separate multiple faults, and properly evaluate bearing fault severity.

Acknowledgments

I would like to express my appreciation to those who encouraged me to accomplish the study from Master to Ph.D. for the past five years. Special thanks to the supervisor Dr. Ming Liang for his support and guidance. I am also grateful to the Ph.D. thesis committee members, Dr. Prakash C. Patnaik (National Research Counsil Canada), Dr. Rong Liu (Carleton University), Dr. Bertrand Jodoin (University of Ottawa), and Dr. Michel Labrosse (University of Ottawa), for their patience in reviewing the thesis and providing valuable comments to improve the work.

I want to thank to Dr. Dongil Chang and Dr.(-ing) Choonghee Jo for close freindship and to my colleagues, Mr. Iman Soltani and Mr. Raymond Wang, in the lab of the University of Ottawa for the in-depth discussion on the academic problems.

Finally, I express all my gratitude to my family in Korea for their continuous encouragement.

Contents

List of Figures

List of Tables

Nomenclature

$[\ ]_{R-L}$	Riemann-Liouville fractional integral
α_b^j	damping factor of exponential decay of j-th vibration mode of bearing
$\bar{x}$	mean value of $x(t)$
β	contact angle of the balls on the races in a ball bearing
Δt_i	time difference between two consecutive time indices, i and $i+1$
Δt_m	mean value of time difference Δt_i
ϵ	load distribution factor in Stribeck equation
γ	Morlet wavelet parameter (=1)
$\Gamma(\ )$	Gamma function
$\hat{s}_i$	de-noised or estimated signal of s_i at time index i, a discrete $\hat{s}(t)$
$\hat{w}(\omega)$	Fourier transform of $w(t)$
$\hat{w}_{a,b}(\omega)$	Fourier transform of $w_{a,b}(t)$
$\hat{H}^n x(t)$	oil-particle signature with compensation for the magnitude variation
$\hat{s}(t)$	de-noised or estimated signal of $s(t)$

ω_0	Morlet wavelet parameter (≥ 5)
ϕ_f^i	phase of i-th harmonic of bearing model for impulses
ϕ_b^j	phase angle of exponential decay of j-th vibration mode of bearing
ϕ_l^k	phase angle of k-th harmonic of vibration model for non-uniform load
ρ_{WW}	autocorrelation coefficients of wavelet coefficients
σ_W	standard deviation of wavelet coefficients
$\sigma_{\Delta t}$	standard deviation of Δt_i
τ	time lag for calculating autocorrelation function
θ	circumferential angle of bearing around shaft
$\tilde{\rho}_{WW}$	Hilbert transform of ρ_{WW}
$\tilde{I}^n x(t)$	normalized inflexion points of oil-particle signature
$\tilde{s}_i$	data sampled from a fake impulses
$*$	complex conjugate
A	contact area between components
a	scale factor
A_b^j	amplitude of exponential decay of j-th vibration mode of bearing
A_d	area of a circular defect
A_f	amplitude of vibration model for impulses
A_l^k	amplitude of k-th harmonic of vibration model for non-uniform load
a_{best}	the best scale selected from energy and kurtosis
a_{max}	maximum scale level used in continuous wavelet transform

a_{min}	minimum scale level used in continuous wavelet transform
b	translation factor in time
BD	ball diameter of ball bearing
$c_N(N)$	Lempel-Ziv Complexity of sequence of length N
c_{nN}	normalized Lempel-Ziv complexity
$c_{UL,N}$	upper limit of Lempel-Ziv complexity of sequence of length N
$C_{WW}(\tau)$	covariance of wavelet coefficients
$D_L^1 x(t)$	the first-oder derivative defined as $\dfrac{x(t)-x(t+h)}{h} \quad (h \to 0)$
$D_R^1 x(t)$	the first-oder derivative defined as $\dfrac{x(t)-x(t-h)}{h} \quad (h \to 0)$
D_L^n	n-th order derivative of D_L^1
D_R^n	n-th order derivative of D_R^1
$E[\cdot]$	expected value
E_a	energy of wavelet coefficients at scale a
E_f	accumulated false energy due to spurious impulses
$EnvW(a_{best}, b)$	envelope signal of wavelet coefficients at the best scale a_{best}
F	force by which the components (ball and inner race) are pressed against each other
f	frequency in Hz
f^c	center frequency of the selected mother wavelet function
f_a^{ps}	pseudo frequency of scale a

f_b^j	resonance frequency of exponential decay of j-th vibration mode of bearing
F_s	sampling frequency
f_s	shaft rotating frequency
$FW_a(\cdot)$	Fourier transform of $W(a, b)$
$H^n x(t)$	oil-particle signature extracted from the input $x(t)$
$I^n x(t)$	inflexion points of oil-particle signature extracted from the input $x(t)$
Inp	input or wavelet coefficients to be de-noised
j	imaginary unit of complex value
K_a	kurtosis of wavelet coefficients at scale a
$Kurt(\cdot)$	kurtosis value
$L\{\ \}$	Laplace transform
$LZC\{\ \}$	function calculating Lempel-Ziv complexity
M	maximum value of the time lag τ
n	non-integer order used in the generalized differentiation
$n'(t)$	wavelet transform of the noise $n(t)$
$n(t)$	white Gaussian noise
N_1	number of harmonics induced by impulse series
N_2	modal order of bearing vibration
N_3	number of harmonics
n_e	number of rolling elements

Out_h	output of hard-thresholding rule
Out_s	output of soft-thresholding rule
$P_{\max,d}$	maximum contact pressure between ball and race when a defect exists
$P_{\max}$	maximum contact pressure between ball and race
P_{false}	false energy density calculated from fake impulses
PD	pitch diameter of ball bearing
PSD_i	power spectral density of signal i
r_d	radius of a circular defect
$r_{i,a}$	contribution ratio of scale a in signal i
$s(t)$	impulsive fault features generated by mechanical fault
s_i	original impulse, a discrete version of $s(t)$
$S_{i,a}(b)$	constructed wavelet coefficients of signal i in scale a at time b
T	period ($= 1/f$) in sec
t	time
t_i	discrete time moment at index i
t_{1a}	kurtosis-based threshold value
t_{2a}	kurtosis-based threshold value
t_{3a}	kurtosis-based threshold value
T_{thr}	threshold
$u(t)$	unit step function
$W(a,b)$	wavelet coefficient in scale a and at time b

$w(t)$	mother wavelet or basis wavelet function
w_H	weight of high-frequency carrier signal for fault severity
w_L	weight of low-frequency envelope signal for fault severity
$w_{a,b}(t)$	daughter wavelet of mother wavelet $w(t)$
$x(t)$	measured noisy signal mixture
$x_b(t)$	bearing-induced vibration including decay
$x_f(t)$	impulse series produced by the fault
$x_l(t)$	modulation effect caused by the nonuniform load distribution
$z_\rho(\tau)$	analytical signal or envelope

List of Acronyms

BSS	Blind Signal Separation
CMS	Condition Monitoring System
CWRU	Case Western Reserve University
CWT	Continuous Wavelet Transform
DWT	Discrete Wavelet Transform
EHMS	Engine Health Monitoring System
EMD	Empirical Mode Decomposition
FAA	Federal Aviation Administration
FFT	Fast Fourier Transform
FMEA	Failure Mode and Effect Analysis
GMF	Gear Meshing Frequency
HMM	Hidden Markov Model
HMS	Health Monitoring System
HUMS	Health and Usage Monitoring System

ICA	Independent Component Analysis
IID	Independent and Identically Distributed
IMF	Intrinsic Mode Function
MAP	Maximum *a Posteriori*
MLE	Maximum Likelihood Estimation
MSE	Mean Squared Error
NTSB	National Transportation Safety Board
ODM	Oil Debris Monitoring
PCA	Principal Component Analysis
PDF	Probability Density Function
PSD	Power Spectral Density
SCD	Spectral Correlation Density
SNR	Signal-to-Noise Ratio
SPWVD	Smoothed Pseudo Wigner Ville Distribution
SSR	Signal-to-Signal Ratio
STA	Synchronous Time Average
VFR	Visual Flight Rules

Chapter 1

Introduction

Early fault detection of machine elements is important to prevent catastrophic damage of a machine or a mechanical system. A condition monitoring system (CMS) or health monitoring system (HMS) has been developed to deal with this issue in this study. Many of today's machines are very complex, and the failures of the machines or their components are often unpredictable. It would therefore be risky to use the overly simplified maintenance strategies such as scheduled maintenance. For example, the condition of an airplane is directly affected by, among many other factors, the frequency of its flights, the weather conditions, the payloads, and the quality of the components. The statistics-based maintenance schedule may not best fit each individual plane. Hence, an automated and continuous condition monitoring system is essential to provide real-time information and improve the safety and reliability of a machine or vehicle. In this study, three important issues in fault diagnosis are addressed: fault feature extraction, signal separation, and fault severity. The following information is useful to help us understand the issues and define the research objectives.

Martin (1994) reviewed the three types of maintenance schemes in a standpoint of development. They are unplanned maintenance, planned or scheduled maintenance, and condition based maintenance.

The unplanned maintenance is the simplest strategy since maintenance takes place only when a breakdown occurs. It is obviously inefficient since the machine could be out of service at the unexpected times.

The planned or scheduled maintenance is adopted to partly resolve the problem of unplanned maintenance. In this strategy, maintenance takes place regularly based on the predetermined time interval. For example, aircraft maintenance plan is a typical practice of planned schedule maintenance. All aircraft components have their own planned maintenance schedule such as 500 hours, 1,000 hours, etc. It allows planning of maintenance resources and planning of downtime. However, it may lead to over-maintenance since many machine elements may be in excellent working conditions when the maintenance takes place.

The condition based maintenance is the most advanced policy. It can reduce unnecessary machine downtime and maintenance resources. It can also avoid the over-maintenance problem present in the planned maintenance policy. Nevertheless, it is impossible to on-line monitor the conditions of all machine elements, particularly for complex machinery. The advantages and disadvantages of three strategies are summarized in Fig. 1.1.

A health monitoring system is widely used in various fields such as machining, robotics and aviation. Two typical examples are the engine health monitoring system (EHMS) for a fixed-wing aircraft and the health and usage monitoring system (HUMS) for a rotary-wing aircraft or a helicopter. The EHMS is extensively applied from commercial airliners/military jet fighters to spacecraft (Volponi and Wood, 2005; Myers et al., 1985; Wilson et al., 2003) to avoid fatal accidents that may be caused by engine failure as much as possible. The National Transportation Safety Board (NTSB) has reported 1,396 civil helicopter accidents in the US from 1990 to 1996. Two thirds of these fell into low-priced helicopter mostly operated by Visual Flight Rules (VFR) without HUMS (Iseler and Maio, 2002). This indicates that the HUMS can reduce the fatal accidents. There were 192 turbine helicopter accidents globally in 1999. Of those, 49 were caused by engine failures and power loss, and 28 were directly due to mechanical failures, mostly in the drive train or gearboxes of the propulsion system (Learmount, 2000). That is, 40% of total accidents were caused by failures directly related to the mechanical systems (including engine failure) of the helicopters. According to the NTSB's accident data

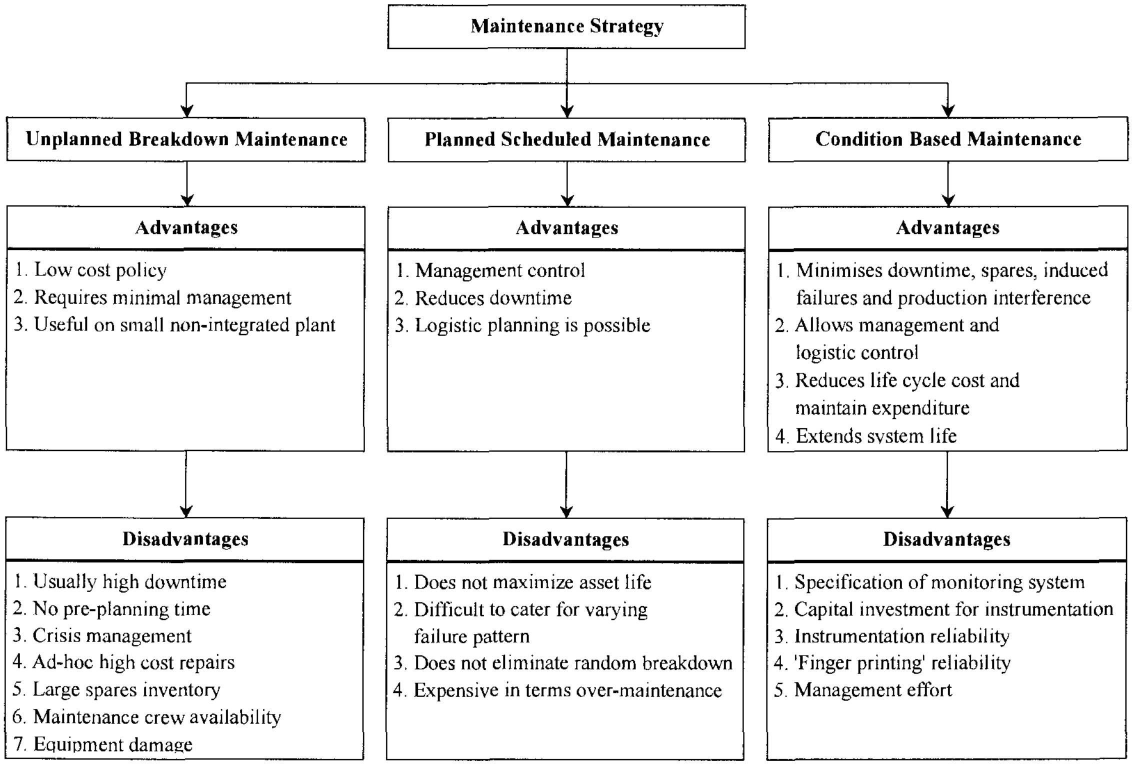


Figure 1.1. Effectiveness of different types of maintenance (from Davies (1990))

covering 7,571 US-registered transport aircraft from 1980 to 2001, the biggest cause is hardware-induced accidents (52%), followed by propulsion system related failures (36%). The Federal Aviation Administration (FAA) data covering 40,964 incidents of the US airplanes from 1998 to 2003 shows that 67% of incidents were caused by the combined system and component failure, malfunction, and power loss (Belcastro and Allen, 2007).

An HMS such as the EHMS and HUMS are therefore important to reduce the accidents caused by hardware-related defects because warning can be signaled before a small defect develops to the critical fault. The EHMS began to get attention extensively with the advent of commercial high bypass turbofan engines in the 1970s (Volponi and Wood, 2005) and the first HUMS for helicopter was certified in November 1991 for operation in the North Sea (Pipe, 2003). The main purpose of an HMS is to indicate the deterioration of a component so that an appropriate maintenance activity can take place to deal with

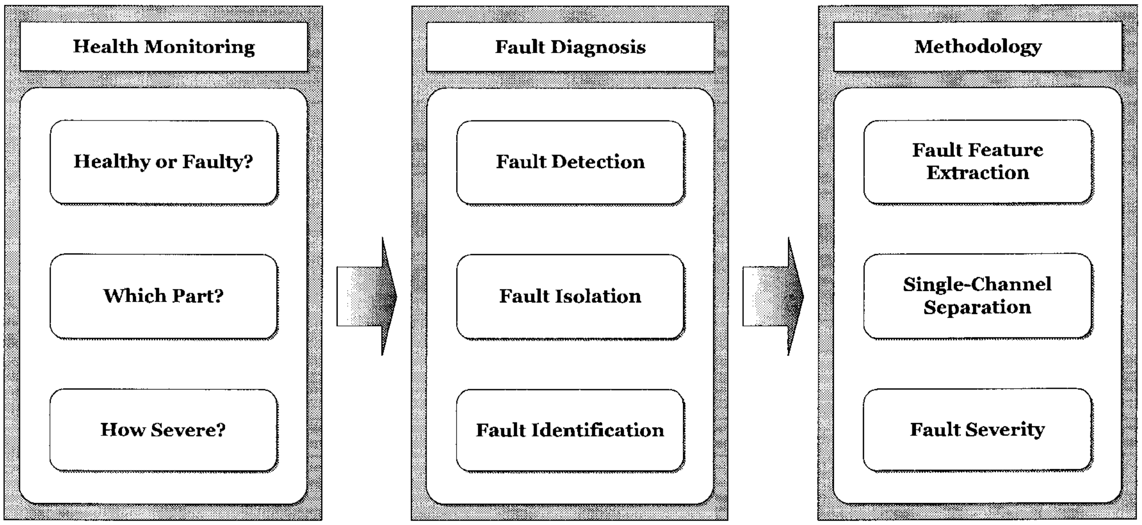


Figure 1.2. Three principal issues of health monitoring and the scope of the thesis

the problem. Since then, many studies have been conducted to improve the reliability of HMS, accomplishing a success rate of 70% in detecting defects (Pipe, 2003).

1.1 Thesis outline

A good machine health monitoring system should be able to answer the following three principal questions which are detailed in Fig. 1.2:

1. Is the machine healthy or faulty?
2. If faulty, which part of the machine is defective? and
3. How severe is the fault?

These questions can be answered by addressing three main issues: fault detection, fault isolation, and fault identification (Jardine et al., 2006). To address these three issues many methods have been proposed over the years. Nevertheless, further research is needed to improve the accuracy and reliability of these methods, which is the main purpose of this study. More specifically, this thesis focuses on the following topics to better address the above three issues:

1. Fault feature extraction from
    - Periodic signal
    - Non-periodic signal (for both vibration and oil-debris signals)
2. Single-channel signal separation
3. Fault severity evaluation.

1. **Fault feature extraction**

   Although HMS has been adopted in many industries over the years, there are two main problems present in many such systems: over-reaction (false alarm) and slow/no reaction (no alarm). False alarms inevitably cause unnecessary interruption of normal operations while slow or no reaction to an existing fault could lead to machine failures. The two problems are mainly caused by unreliable feature extraction algorithms for detecting relevant failure signatures (Tumer and Bajwa, 1999). Slow or no reaction is also caused by features blurred or tainted by noise. In this study, different de-noising and feature extraction methods will be developed for two of the most frequently encountered types of signals in machine fault detection, i.e., periodic signals and non-periodic signals (vibration and oil-debris signals).

2. **Single-channel signal separation**

   Another reason of no-alarm or slow reaction is the lack of sufficient data measurement (Tumer and Bajwa, 1999). In reality, the data available for fault diagnosis are often limited because it is impractical to monitor every element of a complex machine. In addition, a measured signal is easily masked by other sources, which makes it difficult for an HMS to detect faults hence timely warning cannot be provided. In view of this, a single-channel signal separation technique is proposed to mitigate this difficulty and to improve the vigilance of an HMS. Single-channel signal separation is also an effective fault detection and isolation method for reducing no-alarm problem by retrieving a weak signature of fault from a signal mixture of multiple periodic features measured by single sensor.

3. **Fault severity evaluation**

   Identifying fault severity is directly related to the repair or replacement decision. Fault severity is normally evaluated by comparing measured fault feature data with reference data (or healthy condition). This requires keeping reference data for a wide range of machine operating conditions, which is obviously inefficient and impractical. In this study, a new method that does not require comparison with reference data will be introduced. The attempt is being made to devise a non-dimensional fault severity measure ranging between [0, 1] and hence it could be used for any applications without much modification.

Bearings are critical elements because they not only support the rotating structures but also carry most energy generated during machine operation. Bearing failure may result in expensive damage in most machines, from a simple tape recorder to complex petroleum and refinery facility (Peterborough, 1965; Latino, 1996). Latino (1996) quantified the financial loss of production caused by bearing failure in the petroleum and refinery facility. According to the FMEA (Failure Mode and Effect Analysis) result, the bearing failure accounts for 48% of total failure reasons, which costs about 46% of the total loss of production. Gear faults are another major source of machine failures. The significance of gear failure is well studied especially for rotorcraft. The main reason of rotorcraft accidents is gear failure in a power transmission (Zakrajsek, 1989; Zakrajsek et al., 1992). Based on the investigation of serious rotorcraft accidents, 32% of fatigue failures were due to engine and transmission components (Zakrajsek et al., 1992).

Since bearings and gears are the two essential components in any rotating machinery, the proposed algorithms are evaluated using experimental data obtained from faulty gears and bearings. The remaining part of the thesis is organized as follows. Chapter 2 briefly reviews the previous work pertinent to the three topics mentioned earlier, i.e., fault feature extraction, signal separation, and fault severity. Feature extraction methods are developed in Chapter 3 through Chapter 5 for periodic and non-periodic signals including both vibration and oil-debris signals. Chapter 6 details a new single-channel signal separation method followed by a new approach for fault severity evaluation presented in Chapter 7. Finally, Chapter 8 concludes the thesis.

Chapter 2

Literature Review

This chapter presents a review of previous work pertinent to the three topics mentioned in the previous chapter, i.e., fault feature extraction, signal separation, and fault severity.

2.1 Review of fault feature extraction methods

The methods for fault feature extraction can be classified as indirect methods (vibration analysis) and direct methods (oil analysis) depending on the type of fault signal to be diagnosed. They are reviewed as follows.

2.1.1 Feature extraction in vibration analysis

Most fault signals collected by a vibration sensor from rotating machinery contain sparsely distributed non-smooth impulses. Such signals are often contaminated by noise from various sources with a white Gaussian noise behavior. The noise makes it difficult to detect fault signature from the measured signal. De-noising can improve signal-to-noise ratio (SNR) by reducing noise components from the signal mixture and hence plays an important role in mechanical fault feature extraction.

The de-noising tasks in the context of mechanical fault detection include removing both the Gaussian-like noise and the spurious impulses, i.e., the impulses caused by noise of non-Gaussian nature or other interferences. There are several de-noising methods such as low-pass, high-pass and band-pass filtering, and wavelet thresholding. However, none of them can effectively remove both Gaussian-like noise and spurious impulses.

Even for the removal of Gaussian-like noises, the existing methods are often inadequate. For example, the filtering methods are useful only for removing noises residing in limited and known frequency ranges. However, as the white Gaussian noise spreads over the entire frequency spectrum, the simple filtering methods cannot be applied to remove such noises.

As an alternative, wavelet thresholding, due to its sparsity, locality, and multiresolution nature, has emerged as a simple yet effective de-noising tool. Much related work has been carried out for applications in image processing. In wavelet de-noising, two of the major issues are the selection of thresholds and the specification of thresholding rules. Donoho proposed a soft-thresholding rule with a universal threshold, called *VisuShrink* (Donoho, 1995). This method can provide an optimal estimate of the signal to be identified in the minimax sense — minimize the maximum error over an N-sample signal. Though it provides a smooth visual result, the universal threshold tends to yield overly smoothed results (Chang et al., 2000a) and may not be practical to be used for on-line machine fault detection due to its dependence on a large number of samples.

The *SURE* threshold derived from Stein's unbiased risk estimator (SURE) (Stein, 1981) is an alternative to the universal threshold. An associated de-noising method, *SureShrink*, can be considered as a hybrid threshold based on both universal threshold and SURE threshold. The idea is to select the better of the two and hence to improve de-noising results. The SURE threshold is not dependent on N (number of data samples) explicitly and is determined in a sub-band adaptive manner (Jansen, 2001).

Chang et al. (2000a,b) proposed *BayesShrink* with a threshold determined by minimizing the *Bayesian risk* when soft-thresholding rule is applied. They proved that their method can improve the de-noising performance for 2D-image de-noising problems.

The most commonly used thresholding rules are soft- and hard-thresholding rules. Soft-thresholding in some applications yields near-optimal results (Donoho and Johnstone, 1994) and leads to visually smoother effect than the hard-thresholding rule does. This is because the latter is discontinuous and causes abrupt changes around the threshold.

Yoon and Vaidyanathan (2004) proposed a customized thresholding rule that can be switched between soft- and modified hard-thresholding rules by selecting different thresholding parameters that are dependent on the input signal. They have shown that the customized thresholding rule with properly selected parameters gives better de-noising results in terms of mean squared error (MSE) compared with the original soft- and hard-thresholding rules.

Though the wavelet de-noising approach has shown to be a powerful tool for noise reduction in image processing, the research on wavelet de-noising in the context of mechanical fault detection is still very limited. A few early studies have been reported in Lin and Qu (2000), Lin (2001), and Lin et al. (2004). As Lin et al. (2004) point out, the wavelet de-noising methods reviewed above are all based on orthogonal wavelet transform and assume that the true signal is smooth. With these methods, the transient components with rapid changes such as impulses are treated as noise. Clearly, such methods are not suitable for mechanical fault signals containing predominantly non-smooth impulses.

For the above reason, Lin et al. (2004) explored the use of non-orthogonal wavelet transform and applied a maximum likelihood estimation (MLE) based thresholding rule introduced by Hyvärinen (1999) to de-noise mechanical fault signals. However, since Hyvärinen's MLE thresholding rule is based on the probability distribution function (PDF) of the impulses (the signal features to be identified), *a priori* information on the PDF of the signal of interest is required. Such information is not always available, particularly for unknown faults. Therefore, it would be a challenging task to apply such a method to de-noise unknown signals.

It should also be noted that the basis wavelet function has a direct impact on de-noising results. Most aforementioned methods used orthogonal wavelet basis functions and discrete wavelet transform (DWT). The orthogonal basis functions usually do not resemble the impulse signals and are unlikely to be effective for purifying mechanical fault signals. Furthermore, the DWT renders rather coarse grids in a large portion of the time-scale plane and lead to unacceptably low resolutions in the frequency range of interest. In view of these deficiencies, Lin and Qu (2000) used the non-orthogonal Morlet wavelet as a basis function and applied continuous wavelet transform (CWT) to purify mechanical fault signals, which was demonstrated to be effective.

2.1.2 Feature extraction in oil analysis

Several different approaches are available in oil analysis including magnetic chip detection, spectroscopy (X-ray and infrared), laser optical analysis, and in-line oil analysis. Chip detection uses magnet to capture and isolate the metallic debris from the oil flow path. However, a magnetic chip detector is not effective at early detection of bearing damage because of its poor capture efficiency and is not able to detect non-ferromagnetic metal debris (Miller and Kitaljevich, 2000). Moreover, it is prone to false alarm due to the build up of ferromagnetic debris on chip detector.

X-ray or IR (infrared) spectroscopy measures material's characteristic spectrum since each element emits energy at its characteristic wavelength when it is energized to a certain energy level (Humphrey et al., 1998; Whitlock et al., 1998; Akochi-Koble et al., 1998; Garry and Doll, 1998; Powell, 1998; Toms et al., 1998a; Toms and MacIsaac, 2005; Toms et al., 1998b). A well-trained personnel is necessary to interpret the spectrum of X-ray or IR spectroscopy. Hence, it is unsuitable for daily oil analysis.

Laser optical analysis is a high resolution and high speed image processing technique (Reintjes et al., 1998; Tucker et al., 1998). A camera takes a digital image at high speed of 500 frames per second and an image processing technique is applied to extract the oil debris. This method can detect a particle less than 100 μm. However, it is impractical for on-line application because of the computational requirements for high-speed image processing.

In-line oil-debris analysis has been widely researched and applied for the mechanical fault diagnosis of gearboxes (Howe and Muir, 1998; Dempsey, 2001; Miller and Kitaljevich, 2000; Dempsey, 2000). Howe and Muir (1998) introduced an in-line oil-debris sensor to replace the magnetic chip detector for the engine gearbox of a military helicopter. Along with the sensor, a signal processing unit was developed for continuous digital signal processing such as phase adjustment, filtering, and signature detection.

The effectiveness of the in-line oil debris sensor has been demonstrated by comparing the accumulated mass of particles detected by the sensor with the total mass of oil debris extracted by a filter. An oil-debris sensor measures three parameters: the number of particles, the approximate size, and the accumulated mass. The relationship of each parameter with the bearing condition is comprehensively studied and it is shown that the accumulated mass is a proper measure of bearing damage (Miller and Kitaljevich, 2000).

Dempsey (2001) applied fuzzy logic to verify that the accumulated mass of oil debris is a good indicator for gear damage and to identify a method to determine threshold that discriminates condition states, for example, healthy or damaged state.

Efforts have also been made to combine oil-debris analysis with vibration analysis for more reliable fault diagnosis. Since a typical vibration sensor is vulnerable to the transmission noise, considering both vibration and oil-debris signals could improve the accuracy of fault diagnosis. Dempsey (2000) investigated the relationship between the oil debris and vibration and attempted to incorporate the two methods for the gear damage detection.

To obtain correct measurement of the three parameters (the number of particles, the approximate size, and the accumulated mass), it is important to extract oil-debris signature from the measured noisy signal. A typical technique of extracting such a signature is a threshold-based algorithm, in which a signal with magnitude greater than the pre-determined threshold value is considered to be the oil-debris signature. It should be noted that extracting an oil-debris signature using a threshold-based method requires the following (Howe and Muir, 1998; Miller and Kitaljevich, 2000):

1. Threshold values

   In general, deciding threshold value is not straightforward; the particle detectability depends on the selected threshold value. Therefore, inappropriate threshold value may undermine the usefulness of this approach.

2. Digital filtering

   The sensor measures signal in a wide range of frequency depending on the sampling frequency. The informative feature of oil debris appears in a narrow frequency range around the flow speed. The filters can remove unnecessary signals from frequency ranges (determined by flow speed) that are not of interest and can help detect oil-debris signature.

2.1.3 Summary of fault feature extraction methods

As reviewed above, much work has been done to improve fault feature extraction method. Main contributions may be summarized as follows:

- Threshold based de-noising has been improved by introducing various threshold values and thresholding rules:
  - The Sureshrink method removes the dependency of universal threshold on N (Stein, 1981; Jansen, 2001)
  - The universal threshold with soft-thresholding rule approaches near-optimal result (Donoho and Johnstone, 1994; Donoho, 1995)
  - Some thresholding rules alleviate the difficulty in determining threshold by considering MLE (Hyvärinen, 1999)
  - BayesShrink (soft-thresholding rule) with threshold determined by minimizing the Bayes risk improves de-noising for 2D image (Chang et al., 2000a,b)
  - A customized thresholding rule, a combination of soft- and hard-thresholding rules, proves to be more effective than typical soft- and hard-thresholding rules (Yoon and Vaidyanathan, 2004)
- Threshold based wavelet de-noising has been applied to mechanical fault detection (Lin and Qu, 2000; Lin, 2001; Lin et al., 2004)
- For oil analysis, the following methods were developed.
  - X-ray fluorescence (Humphrey et al., 1998; Whitlock et al., 1998)
  - IR spectroscopy (Akochi-Koble et al., 1998; Garry and Doll, 1998; Toms et al., 1998a; Powell, 1998; Toms et al., 1998b; Toms and MacIsaac, 2005)

- ◦ Laser optical analysis (Reintjes et al., 1998; Tucker et al., 1998)
- ◦ Oil-debris signal analysis (Howe and Muir, 1998; Dempsey, 2001; Miller and Kitaljevich, 2000; Dempsey, 2000)

2.2 Review of signal separation methods

2.2.1 Blind signal separation

A rotating machinery system may consist of many components such as gears, bearings, and shafts. The data collected from these components during operation contains valuable information about the machine or component conditions. However, the measured data is often a mixture of periodic signals from several rotating components or sources. Therefore, the information for maintenance decisions is not readily available from the data and has to be extracted using appropriate signal processing techniques. Blind signal separation (BSS) is an effective tool for identifying sources of the signals blended in the collected data. The BSS approach has been successfully applied in many different areas including telecommunications, medical diagnosis, speech and image processing. Recently, it has also shown to be a very appealing technique for rotating machinery condition monitoring.

The BSS problems can be classified into three categories: *complete, under-complete* and *over-complete*. If the number of sensors is equal to that of signal sources, the problem is a complete BSS problem. Otherwise it is either under-complete if there are more sensors than signal sources, or over-complete when the number of sensors is less than that of sources. In the context of machine condition monitoring, it may not be realistic to provide a sensor for each signal source for at least the following reasons:

1. Enclosed environment – many machine components are placed in an enclosed environment, e.g. the gears and bearings in a gearbox, where sensors cannot be easily installed.

2. Interference with operations – Even if the sensors can be installed, they may ob-

struct normal operations due to the over crowded vicinity of the machine components.

3. Excessive sensor cost and care – The use of large number of sensors is not only costly but also knowledge demanding. It involves excessive operating effort and signal processing burden as well. This inevitably complicates the monitoring system and discourages industrial acceptance.

4. Impact on structural integrity – The mounting of certain sensors such as accelerometers may weaken the strength and impair stability of some machine components, which defeats the very purpose of machine monitoring.

5. Unknown number of signal sources – A single machine component may experience different types of faults at the same time or at different times. Hence, the number of signal sources changes with time. For example, a gear can be subjected to wear, fracture or both. Some faults may never occur before routine replacement of the component, making it unnecessary to install a dedicated sensor. However, the occurrence of certain faults is not known a priori and not always predictable. Hence, it is difficult to determine the number of sensors to be used even if it is economically justified and technically viable to have a dedicated sensor for each fault signal.

For the above reasons, the number of sensors is mostly much less than the number of faults or signal sources in an industrial setting. This gives rise to the use of over-complete BSS for signal separation and fault detection.

BSS has been applied to mechanical signal separation by several researchers. Gelle et al. (2000) separated two source signals from two signal mixtures measured by accelerometers and microphones. Li et al. (2001) applied sequential BSS method to analyze the noise source of diesel engine from the measured signal using single microphone. In Cao and Liu (1996), the authors analyzed the separability of a set of mixed sources by linear filters and further provided guidelines for the design of linear filters. Cichocki et al. (1997) proposed a neural network based sequential method for BSS problems with unknown number of sources. Li and Wang (2002) examined several ill-conditioned cases including the over-complete BSS problem. They further developed a cost function based sequential approach for separation of unknown number of signals.

Recently, a BSS special issue has been published to address the unique issues in application of BSS to mechanical signals, "which are typically characterized by an excessive complexity as composed for instance to communication signals" (Antoni and Braun, 2005). Several papers in the special issue, including Antoni (2005), Peled et al. (2005), Bouguerriou et al. (2005), Boustany and Antoni (2005), Bonnardot et al. (2005), and Puigt and Deville (2005), are of particularly interest to this study. Antoni (2005) proposed to separate vibration signals into contributions or components from periodic, random stationary and random nonstationary sources, thereby making the BSS problem much more tractable. Peled et al. (2005) reported a blind deconvolution approach to the separation of impulsive signals, a common mechanical fault signature, by maximizing kurtosis through an iterative algorithm. It has been shown to be effective particularly when sensors cannot be mounted close to the source locations, such as bearings in a mechanical system. With their approach, the adverse effect of structural resonance can be removed and hence better fault detection results can be expected. Cyclostationarity is adopted in Bouguerriou et al. (2005), Boustany and Antoni (2005), and Bonnardot et al. (2005) as a powerful signal separation tool to extract mechanical fault signatures. For example, bearing fault signal is considered the second order cyclostationary. These cyclostationarity-based methods performed very well even for noisy signals. Puigt and Deville (2005) further suggested a time-frequency (using short-time Fourier transform) ratio-based BSS method to identify the columns of the mixing matrix for separation of attenuated and time-delayed sources. Servière and Fabry proposed a modified principal component analysis (PCA) approach to separating periodic source signals from the noisy signal mixtures (Servière and Fabry, 2004, 2005). They have shown that the modified PCA outperformed the traditional PCA method in the context of spatially correlated noise and modulated sources.

2.2.2 Single channel signal separation

In machine condition monitoring, a single sensor such as an accelerometer can provide rich information about conditions of several machine components. For example, the measured data from an accelerometer mounted on a bearing housing is a mixture of signals reflecting conditions of the bearing inner race, outer race, rolling elements, and shaft. In addition, a single machine component may represent multiple signal sources, reflected by multiple faults of different characteristics. As mentioned earlier, this is further complicated by the fact that it is often unknown if a fault will ever occur and when it will occur. In this case, it is impossible and unnecessary to provide an individual accelerometer for each of these components. The separation of these signals from the single sensor signal source is called *single-channel blind signal separation* problem, an extreme case of the over-complete problem.

As Roweis (2001) points out, the signal sources may not be recovered from the single-channel observation using the unmixing algorithms even though it has shown to be effective for multi-channel signal mixtures. The reason is that the unmixing algorithms recover sources by re-weighting multiple observation sequences and thus they cannot be applied if only one observation signal is available. This view is also endorsed by other researchers. For example, Jang and Lee (2003) state that the conventional BSS methods require observations of multiple channels, at least two, and hence cannot be readily employed for single-channel problems. For this reason, Roweis (2001) suggests decomposing or re-filtering the signal mixture onto several sub-frequency bands and then applying a hidden Markov model (HMM) approach to generate masking signal for each of these sub-band signals. Though this method is very successful for speech signal separation, it cannot be used to separate mechanical signals. The reason is that the HMM theory is based on the stationary assumption, i.e., the state transition probabilities are independent of the actual time at which the transition occurs (Ertunc et al., 2001). This means that the state transition probabilities follow an exponential distribution (memoryless). This assumption is perfectly justified in speech signals but invalid for mechanical signals because the state transition probability of mechanical faults is generally affected by history or accumulated effects. For example, the effect of fatigue is not memoryless. Jang and Lee (2003) recently developed an independent component analysis (ICA) approach to the single-channel speech separation problem. This method has shown to be effective in separating two source signals. For a mixture with more sources, they suggested the use of the sequential extraction method. They also note that the separability depends on the selection and training of basis functions.

2.2.3 Summary of signal separation methods

BSS has been widely researched in various applications such as communications and mechanical fault detection. In summary, the previous work in BSS has accomplished the following:

- Analyzed signal separability and provided guidelines for the design of linear filters (Cao and Liu, 1996),
- Applied BSS to separate mechanical fault signals (Gelle et al., 2000; Li et al., 2001; Servière and Fabry, 2004; Antoni, 2005; Bouguerriou et al., 2005; Bonnardot et al., 2005; Boustany and Antoni, 2005; Peled et al., 2005; Puigt and Deville, 2005; Servière and Fabry, 2005),
- Separated vibration signals into periodic, random stationary and random nonstationary sources (Antoni, 2005),
- Adopted cyclostationarity to separate bearing fault signal, a second-order cyclostationary signal (Bouguerriou et al., 2005; Bonnardot et al., 2005; Boustany and Antoni, 2005),
- Modified PCA to separate periodic source signals from noisy signal mixtures (Servière and Fabry, 2004, 2005),
- Applied BSS in time-frequency domain by using short-time Fourier transform (Puigt and Deville, 2005),
- Applied sequential extraction to unknown number of signals (Cichocki et al., 1997; Li and Wang, 2002), and

- Applied BSS to single-channel signal separation (Roweis, 2001; Jang and Lee, 2003).

2.3 Fault severity evaluation

2.3.1 Review

A mechanical fault signal contains information not only about the machine health condition (healthy or faulty) but also the severity of fault. Depending on the fault severity, a decision whether the faulty machine part should be repaired or replaced can be made. Accurate information of fault severity can effectively reduce machine downtime and maintenance cost as well. The most commonly used measure of fault severity is the energy value in frequency or time-frequency domain. Covacece and Introni (2002) illustrated fault developments of ball bearing of helicopter gearbox using auto- and cross-power spectrum. Two indices, auto- and cross-power spectrum, increase their magnitudes as the fault develops. In the work by Loutridis (2004a), energy was successfully applied to the fault feature (intrinsic mode function, IMF) obtained from the empirical mode decomposition (EMD) method. Dalpiaz et al. (2000) compared the cepstrum and spectral correlation density (SCD) methods to reveal the sensitivity of two indices to fault development. According to the work by Dalpiaz et al. (2000), the cepstrum is insensitive to the crack evolution of a gear. On the other hand, SCD shows an increasing trend as the gear crack grows.

Baydar and Ball (2001) applied a time-frequency analysis method, smoothed pseudo-Wigner-Ville distribution (SPWVD), to evaluate gear failures. Two symptoms were reported for faulty state. First, the energy concentration appears around angular position corresponding to the fault. Second, frequency interference increases between the fundamental and the higher harmonics. As a fault develops, these two phenomena appear more clearly. Baydar and Ball (2003) applied another time-frequency analysis method, continuous wavelet transform (CWT), to different gear fault conditions. They adopted both amplitude and phase plots of CWT to analyze the faulty state. The finding is similar to that of their previous work using SPWVD, i.e., the same two symptoms appear as fault develops. Boulahbal et al. (1999) applied CWT to the synchronous time average (STA) signal of gear crack and chip to obtain the amplitude and phase plots. The result indicated that energy concentration happened at the gear meshing frequency (GMF) in the amplitude plot and phase changes drastically around the fault location. In the work by Loutridis (2004b), a local energy density was calculated in the time-scale domain to evaluate the relationship between energy level and gear fault development. The energy is obtained from the wavelet coefficients at a specific time instant corresponding to the largest energy value over scales. A polar wavelet amplitude map was applied to the STA signal of gear faults to display the fault severity and the circumferential location around the rotating axis (Meltzer and Dien, 2004). In their cases, energy concentration appeared and became stronger as fault develops.

Rubini and Meneghetti (2001) compared the envelope spectrum method and averaged wavelet amplitude spectrum approach in evaluating ball bearing's inner- and outer-race faults. Both methods were able to reveal the difference in spectral magnitude as the fault develops. According to their work, the inner race and the outer race showed opposite trend in spectral magnitude as fault size increases.

Yan and Gao (2004) used the Lempel-Ziv complexity as a numerical measure to examine bearing conditions. This non-dimensional complexity level is close to zero when a signal is sinusoidal function and approaches 1.0 when the signal's nonlinearity increases due to the growth of fault. They demonstrated the possibility of using the complexity as a severity measure and further indicated that the fault size as a severity measure is proportional to the complexity of the signal. Lempel-Ziv complexity provided possibility of fault severity as a non-dimensional measure and fault severity could be assessed without relying on comparison. However, several important issues are yet to be addressed.

1. A distinguishable measure is necessary for noise. A signal produced by mechanical fault has non-dimensional complexity close to 1.0 as fault develops. There is an ambiguity in complexity values between noise and faulty signal because pure noise also has non-dimensional complexity of 1.0.

2. A single measurement needs to provide satisfactory fault severity result. A complexity-based method requires many data measurements to get a reliable result on the fault severity. This could be an obstacle for the complexity-based severity measure to be applied in practice.

3. A theoretical explanation between complexity and fault severity is necessary.

2.3.2 Summary of previous fault severity research

The following summarizes the previous work related to fault severity measurement for gear and bearing faults.

- Evaluated fault severity of gear or bearing fault by comparing magnitude or energy of
  - Auto- and cross-power spectrum in frequency domain (Covacece and Introni, 2002),
  - SPWVD in time-frequency domain (Baydar and Ball, 2001),
  - Amplitude and phase plots of CWT (Boulahbal et al., 1999; Baydar and Ball, 2003),
  - Polar wavelet amplitude map of STA signal (Meltzer and Dien, 2004),
  - Wavelet coefficients of faulty gear signal at the local time instant (Loutridis, 2004b), and
  - IMF of faulty gear signal obtained from the EMD in time domain (Loutridis, 2004a).
- Investigated effectiveness of two indices by comparison of
  - Cepstrum and SCD for faulty gear signal (Dalpiaz et al., 2000) and
  - Envelope spectrum and averaged wavelet amplitude spectrum (Rubini and Meneghetti, 2001).

- Introduced a non-dimensional numerical index to evaluate bearing failures (Yan and Gao, 2004).

2.4 Motivation and objectives

2.4.1 Motivation

As reviewed in the previous sections, various approaches have been proposed to address the three aspects of fault diagnosis: fault feature extraction, signal separation, and fault severity. However, there are still many unsolved issues in on-line mechanical fault diagnosis as summarized in the following.

- Fault feature extraction
  - Spurious impulses, caused by noise or other interferences, existing in the de-noising result are often an obstacle to extracting accurate impulsive fault features due to their similarity in shape.
  - The commonly used MSE measure is not an appropriate evaluation criterion because it does not consider spurious impulses.
  - Most threshold based de-noising methods were developed for other purposes such as image processing. Hence, they are not suitable for extracting impulsive features of mechanical faults.
  - It is difficult to determine a proper threshold and select a suitable thresholding rule for threshold based feature extraction (both vibration and oil signals).
  - A threshold based algorithm for extracting oil-debris signatures may not be able to extract weak signature tainted by noise, making it ineffective for early fault detection.
- Signal separation
  - Most BSS algorithms are computation-demanding because they were based on iterative processes, which is not suitable for on-line application.

  - The number of sources is often unknown in mechanical fault diagnosis. Even though a sequential extraction method is available, it may accumulate error in the sources extracted at a later stage of calculation (Ye et al., 2004).
  - Single-channel signal separation research in mechanical fault diagnosis is very limited.

- Fault severity evaluation
  - Most methods are based on the comparison of two different health states, for example, healthy and faulty states. This requires keeping reference data measured from known machine state in a wide range of operating conditions. This is obviously inefficient for practical applications.
  - As an alternative, a complexity based method provides a non-dimensional fault index without comparison. However, it has the following drawbacks:
    - Multiple data measurements are required for a reliable index of fault severity. This imposes a burden for on-line applications.
    - An additional algorithm is necessary to distinguish the complexity by fault from the complexity by noise since both a severe fault signal and a noise signal have a non-dimensional complexity index close to 1.0.
    - A relationship between fault severity and complexity needs to be explained.

2.4.2 Objectives

Much work has been carried out for most areas of fault diagnosis. However, it is inadequate to use existing methods to handle impulsive fault signals in real-time applications without prior information on the faults. In this thesis, new methods are developed for mechanical fault diagnosis in the following three aspects:

1. Fault feature extraction from both vibration and oil-debris signals,

2. Single-channel signal separation, and

3. Fault severity evaluation.

To develop effective algorithms for on-line applications, the following goals will be considered for all the three above aspects:

1. No iterating processes will be used in any algorithms to avoid lengthy computation,
2. Single set of measured data from a single sensor will be used for simplicity, and
3. Reliable performance should be maintained under a noisy environment.

A. Objectives for fault feature extraction

To reduce false alarms and slow/no reaction in fault feature extraction, the objectives are to develop:

1. an algorithm to extract features by CWT and autocorrelation focusing on the
   - reduction of spurious impulses and
   - selection of the best scale related to fault feature from CWT.
2. a methodology to extract features by threshold based de-noising which should
   - be suitable for extracting sparsely distributed impulsive features,
   - define proper threshold values and thresholding rules, and
   - define an evaluation criterion.
3. a method to extract a non-stationary oil-debris signature which
   - mitigates the difficulty present in determining threshold and thresholding rule,
   - requires no additional filtering process, and
   - preserves the magnitude and phase information of oil-debris signatures.

B. Objectives for single-channel signal separation

To improve the vigilance of a machine health monitoring system, a single-channel signal separation method, a joint wavelet decomposition and Fourier transform approach, will be developed to separate multiple periodic fault signals (e.g., gear or bearing signals) from a single-channel mechanical fault signal mixture. The goals are again effectiveness and simplicity. More specifically, these methods should:

- be iteration-free,
- not rely on *a priori* information such as the number of sources, and
- only use the data from a single sensor.

C. Objectives for fault severity evaluation

The objective is to improve the Lempel-Ziv complexity based fault severity evaluation method to:

- describe the relationship between complexity and an analytical model of bearing faults,
- develop a non-dimensional quantitative index,
- assess fault severity based on a single data set from a single sensor and hence eliminate the comparison process as well as the requirement for reference data set, and
- distinguish the complexity by fault feature from the complexity by noise.

Part I

Fault Feature Extraction

Most mechanical fault signals measured by sensors, e.g., vibration signals collected by an accelerometer and an oil condition signals provided by an in-line oil-debris monitoring (ODM) sensor, are mixtures of multiple signal sources. For example, the vibration signal is normally acquired from the machine casing. Such a data mixture contains signals from different sources including machine faults, casing vibration, and interaction between machine components. The fault signature is clouded by other signal components, leading to a very low signal-to-noise ratio (SNR) (Dalpiaz and Rivola, 1997). Accurate and timely detection of fault signatures is often hindered by such a low SNR. For this reason, this part of the thesis will focus on fault feature extraction and de-noising. More specifically, a joint CWT and autocorrelation approach is proposed for extracting fault features from periodic signals (Chapter 3), a kurtosis based hybrid thresholding method is developed for de-noising mechanical signals (stationary and non-stationary) (Chapter 4), and a generalized differentiation method is adopted to design an in-line oil-debris monitoring algorithm (Chapter 5). The details of the proposed methods are presented in each respective chapter.

Chapter 3

Identification of Mechanical Fault Signatures using CWT and Autocorrelation[1]

It is important to extract fault features from signals that are polluted by noise or unwanted spurious signal components. Though the literature has reported various de-noising methods, they mostly focus on the reduction of Gaussian-like noise or noises reside in certain known frequency ranges. The removal of spurious impulses has not been adequately addressed. However, Gaussian-like noise and spurious impulses often co-exist in the collected mechanical signals. This obviously undermines the usefulness of the de-noised results provided by a method which is incapable of handling the spurious impulses. The spurious impulses are often non-periodic in nature, which adds additional complexity to the problem. In this chapter, a new feature extraction method is developed to purify the signal by reducing both Gaussian-like noise and spurious impulses. To this end, the autocorrelation coefficient function of the continuous wavelet-transformed (CWT) signal at the best scale level is used to estimate time instants when the impulses appear regularly. The best scale level is used to capture the trace of fault signature produced by mechanical fault of rotating machinery and is selected based on the energy and kurtosis of all the scales. Since a wavelet transform represents the similarity between the signal

[1]See Appendixes B.1, B.2, B.3 and B.8 for MATLAB scripts.

and the wavelet function, a function with an impulse-like shape, in this case, the higher energy (or higher similarity) and the higher kurtosis value (stronger impulsiveness) can be used to select the best scale. Using both energy and kurtosis will overcome the weakness of using one of them alone. For example, kurtosis is a good indicator of impulsiveness but it does not provide any clue for the relevance of a selected wavelet function. On the other hand, using energy alone, as detailed later, would be in favor of the highest scale (associated with the lowest frequency) where the fault feature signals may not always reside. A comb filter is then designed using normal distribution function and is applied to the signal in the time-scale domain to extract periodic fault features. The simulation and experimental results reported in later sections indicate that the proposed method can significantly reduce number of miss-identified or spurious impulses caused by noise. Therefore, it can improve the feature extraction performance by reducing uncertainty due to spurious impulses.

It should also be noted that many previous wavelet de-noising methods focused on image processing where the signals are assumed to be smooth, i.e., non-impulsive. They are not suitable for de-noising mechanical fault signals featuring sparsely distributed impulses. In fact, the spurious impulses, i.e., the impulses caused by noise of non-Gaussian nature or any other interference often appear in the de-noising results obtained by the wavelet methods. The main reason is that the threshold based methods de-noise signal only by manipulating signal magnitudes in wavelet domain. Typical filters such as low-pass, high-pass, and band-pass filters are not suitable to remove spurious impulses either because they are designed for certain frequency bands whereas the non-periodic spurious impulses may not settle in any such frequency bands.

The existence of spurious impulses is a major obstacle that makes identifying mechanical fault feature extremely challenging not only because of their non-periodic nature but also due to their similarity to the authentic fault impulses in shape. To remove spurious impulses effectively, information on the time locations (or instants) of impulses is essential. However, this information is very often unavailable because the measured data is normally contaminated by noise.

This chapter introduces a method that can identify the time moments of periodic impulsive fault signatures from the measured noisy signal mixture. Once this information is available, a comb filter can be applied to extract fault features in time-scale domain. Since the comb filter is applied to the estimated time moments, the spurious impulses can be removed effectively from the extracted fault feature. The proposed method does not use threshold or thresholding rule. Therefore, it can remove any uncertainties that may exist in selecting threshold or thresholding rule.

3.1 The proposed method

A signal measured from a defective rotating machine consists of a series of impulses and noise. An impact between the fault and mating mechanical part generates an impulsive signal. For example, the sudden contact between a chipped gear tooth and a mating gear tooth will create an impulsive signal. An impulsive fault feature is normally contaminated by noise produced by several sources such as environmental noise, other rotating mechanical parts of the same machine, and instrument noise. A noisy signal mixture can be described as:

$$x(t) = s(t) + n(t) \tag{3.1}$$

where $x(t)$ is the measured signal mixture, $s(t)$ is the impulsive fault feature and $n(t)$ is the noise assuming to be white Gaussian. The assumption of white Gaussian noise is based on the observations that the noise components are, as mentioned above, contributed by several different sources. It is well known that if a noise signal $n(t)$ consists of multiple signals from different sources, the signal mixture $n(t)$ will approximately follow Gaussian distribution according to the central limit theorem (Papoulis and Pillai, 1998).

The overall procedure of the proposed method is as follows. It is usually difficult to detect the weak impulses from the measured noisy signal mixture. Therefore, the first step is to extract the trace of such impulses from the noisy signal mixture, even though it may be too weak to observe. This is done using continuous wavelet transform (CWT). However, as different scale levels of CWT exhibit different signal natures of the associated frequency contents, it is necessary to select a scale that best reflects the impulsive nature of the mechanical fault. For the selected scale level, the autocorrelation coefficient function is calculated to reveal the periodic nature more clearly. The periodic peaks of autocorrelation coefficient function would identify the instants when the impulses appear regularly. Then the comb filter is applied to extract the periodic signature. To apply the comb filter in time domain, an acceptable time interval (or bandwith if the filter is applied in frequency domain) has to be determined. Fig. 3.1 explains the overall procedure that is summarized as follows:

Step 1: Perform CWT of the noisy signal mixture.

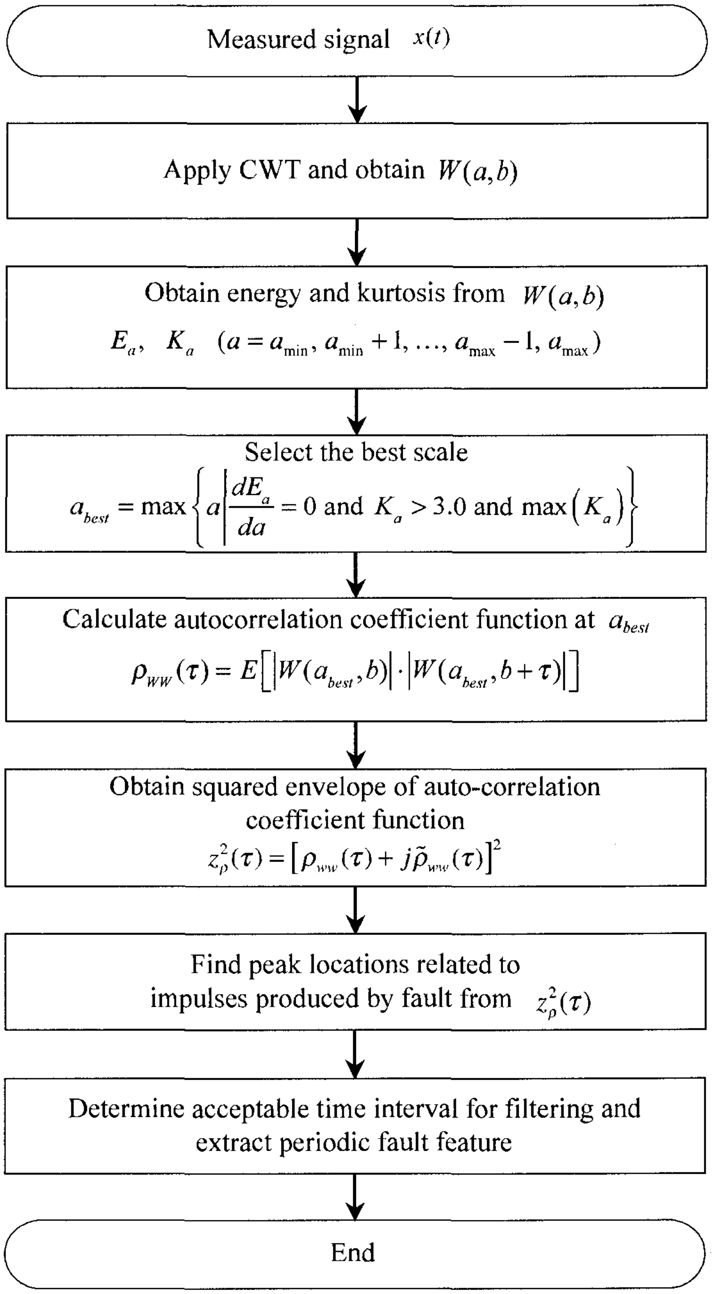


Figure 3.1. Block diagram of detection of impulse location

Step 2: Select the best scale based on the energy and the kurtosis distributions of the each wavelet scale.

Step 3: Obtain autocorrelation coefficient function of the wavelet coefficients at the best scale.

Step 4: Find peaks of the squared envelope of autocorrelation coefficient function corresponding to the hidden impulses.

Step 5: Determine the acceptable time interval around the peaks obtained in step 4 and apply a comb filter to the wavelet coefficients at the best scale to extract periodic fault feature.

3.1.1 Continuous wavelet transform (CWT)

A measured signal $x(t)$ is first transformed into time-scale domain by continuous wavelet transform (CWT).

$$W(a,b) = \int x(t)\, w_{a,b}^{*}(t)\, dt \tag{3.2}$$

where "$*$" indicates the complex conjugate, $W(a,b)$ is the wavelet coefficient, and $w_{a,b}(t)$ is the daughter wavelet defined as

$$w_{a,b}(t) = \frac{1}{|a|^{1/2}} w\left(\frac{t-b}{a}\right) \tag{3.3}$$

where a is the scale factor, b is the translation factor in time, and $w(\cdot)$ is the mother wavelet function or wavelet basis function. The factor $|a|^{-1/2}$ is used to ensure energy preservation (Goupillaud et al., 1984). In this work, CWT is selected because it guarantees higher resolution in time-frequency plane compared to the discrete wavelet transform (DWT) (Lin and Qu, 2000). The real Morlet wavelet is selected as a basis wavelet function for CWT because it is similar to the mechanical fault signature in shape (Lin and Qu, 2000; Lin et al., 2004), which is given by

$$w(t) = \exp\left(-\frac{\gamma^2 t^2}{2}\right)\cos(\omega_0 t) \tag{3.4}$$

where $\gamma = 1$ and $\omega_0 = 5$ are selected for admissibility conditions (Goupillaud et al., 1984) in this study.

The measured signal, $x(t)$, contains a series of impulses produced by mechanical fault with frequency f and period $T = 1/f$. Hence, it can be modeled as a sinusoidal function with frequency f as follows (Braun and Datner, 1979; Braun, 1980; McFadden and Smith, 1984; Wang and Kootsookos, 1998):

$$\begin{aligned} x(t) &= \sum A \cdot \delta(t - nT) + n(t) && (n = 1, 2, \ldots) \\ &\approx \cos(2\pi f t) + n(t) && (A = 1, \text{zero initial phase}) \\ &\approx \cos(\omega_1 t) + n(t) && (\omega_1 = 2\pi f) \end{aligned} \tag{3.5}$$

where the initial phase angle and the higher harmonics are neglected for simplicity. By substituting Eqs. (3.4) and (3.5) into Eq. (3.2), the wavelet coefficients of the measured signal, $x(t)$, can be obtained as follows (See Appendix A.1 for derivation.):

$$\begin{aligned} W(a,b) &= \int x(t) w_{a,b}^{*}(t) dt \\ &\approx \frac{1}{\sqrt{a}} \int \cos(\omega_1 t) \cdot \exp\left(-\frac{1}{2}\left(\frac{t-b}{a}\right)^2\right) \cos\left(\frac{\omega_0 (t-b)}{2a}\right) dt + n'(t) \\ &\approx \frac{1}{\sqrt{a}} \left\{ \cos(\omega_1 b) \cdot \frac{1}{4}\sqrt{2\pi a^2} \cdot \exp\left(-\left(\omega_1 - \frac{\omega_0}{2a}\right)^2 \Big/ \left(\frac{2}{a^2}\right)\right) \right. \\ &+ \cos(\omega_1 b) \cdot \frac{1}{4}\sqrt{2\pi a^2} \cdot \exp\left(-\left(\omega_1 + \frac{\omega_0}{2a}\right)^2 \Big/ \left(\frac{2}{a^2}\right)\right) \\ &+ \frac{1}{4}\sqrt{2\pi a^2} \sin(\omega_1 b) \cdot \exp\left(-\frac{\omega_0^2}{8}\right) \\ &- \frac{1}{4}\sqrt{2\pi a^2} \sin(\omega_1 b) \cdot \exp\left(-\left(\omega_1 - \frac{\omega_0}{2a}\right)^2 \Big/ \left(\frac{2}{a^2}\right)\right) \\ &\left. + \frac{1}{4}\sqrt{2\pi a^2} \sin(\omega_1 b) \cdot \exp\left(-\left(\omega_1 + \frac{\omega_0}{2a}\right)^2 \Big/ \left(\frac{2}{a^2}\right)\right) \right\} \\ &+ n'(t) \end{aligned} \tag{3.6}$$

where $n'(t)$ is the wavelet transform of the noise $n(t)$. The periodic nature of the mechanical fault appears more clearly in some of the scales of CWT and the best scale is selected in the following section.

3.1.2 Selection of the best scale level

Since each scale level shows different frequency contents of the measured signal, $x(t)$, only the scale level which best reflects the signal nature of the repetitive impulses should be selected. In this study, the energy and kurtosis criteria are jointly used for scale selection. The reasons are: a) the energy level is a measure of the shape similarity of the measured signal and the daughter wavelet function, b) unlike in the application for basis or node selection in wavelet packet transform (Nikolaou and Antoniadis, 2002), the energy criterion cannot be used alone in this study because higher energy level is also associated with higher scale level, and c) kurtosis is a good indicator of impulsiveness. Hence, the joint use of energy and kurtosis criteria would take into account of both the signal-wavelet similarity and the fault signature impulsiveness. The joint use of the two would also compensate for the incapability of the energy criterion in distinguishing whether the high energy level is due to high signal-wavelet similarity or high scale level. The energy of the wavelet coefficients in each scale level can be calculated by

$$E_a = \sum_b \{W(a,b)\}^2 \qquad a = a_{\min}, a_{\min}+1, \ldots, a_{\max}-1, a_{\max} \tag{3.7}$$

where E_a represents the energy of the wavelet coefficients in scale a. Several peaks may appear in the plot of E_a. Each scale associated with a peak energy value is considered as a candidate for the best scale. However, as indicated earlier, energy alone cannot guarantee the best scale since the magnitude of the wavelet coefficients increases as scale level increases for energy conservation (Goupillaud et al., 1984). The energy level becomes higher as scale level increases, which may mislead the selection process of the best scale. This is further explained below using energy distribution of daughter wavelets.

The Fourier transform of the Morlet daughter wavelets can be expressed as (Abbate et al., 2002):

$$\begin{aligned} w(t) &= e^{-t^2/2}e^{-j\omega_0 t} \\ \hat{w}(t) &= e^{-(\omega-\omega_0)^2/2} \\ \hat{w}_{a,b}(\omega) &= \sqrt{a}e^{-(\omega/a-\omega_0)^2/2}e^{j\omega b} \end{aligned} \tag{3.8}$$

where $\hat{w}(\omega)$ and $\hat{w}_{a,b}(\omega)$ are the Fourier transforms of Morlet wavelet, $w(t)$, and its daughter wavelets, $w_{a,b}(t)$, respectively. Fig. 3.2 illustrates the normalized (by the maximum energy value) energy distribution of daughter wavelet functions ($b = 1$ is used for illustration purpose). As shown in the figure, the energy level drastically increases with scale. Since the CWT is a series of band-pass filters with increasing amplification rate, wavelet coefficients will have larger amplitude as scale increases, which leads to higher energy levels.

To offset the above drawback of the energy criterion, the kurtosis distribution of wavelet coefficients is examined. The measured faulty signal contains impulses produced by mechanical fault and hence the best scale level should reveal the nature of these impulses. The energy level does not always reflect this nature since it is calculated based on the amplitude of the wavelet coefficients instead of the impulsiveness of a signal. A

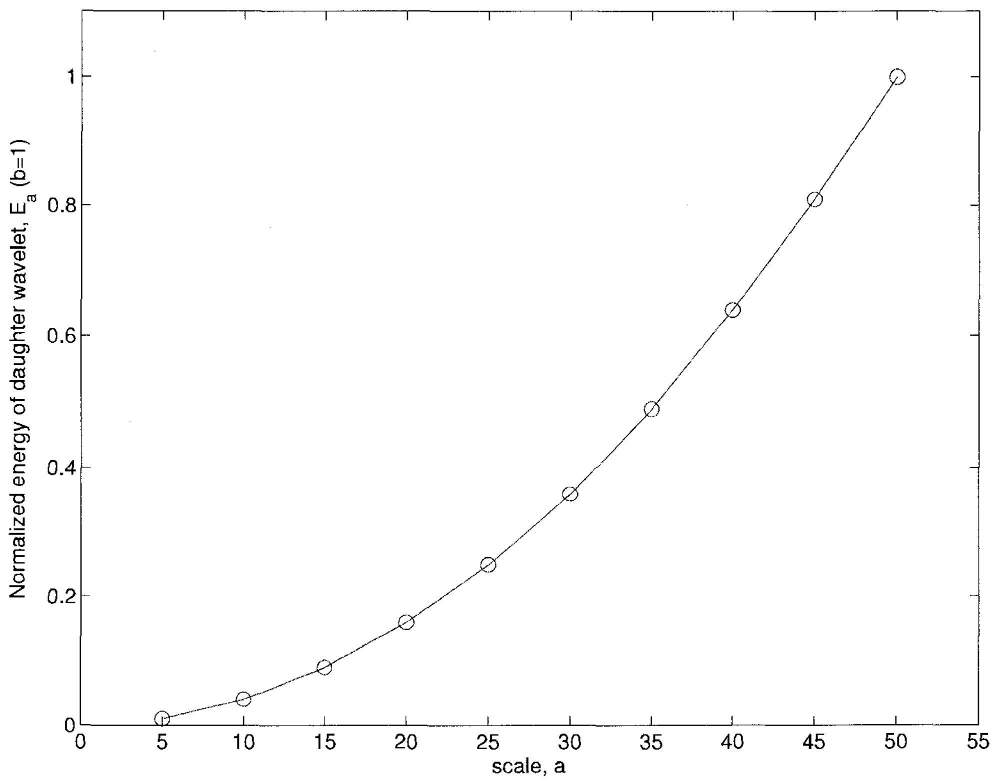


Figure 3.2. Energy distribution of daughter wavelet functions

high impulsiveness is a typical characteristic of a mechanical fault signal. Therefore, using energy value alone is not adequate to secure the best scale level that manifests the impulsive character of the mechanical fault signature. On the other hand, the kurtosis value is sensitive to the outliers or impulses generated by mechanical fault because it is proportional to the 4-th order higher moment defined as (Hyvärinen et al., 2001):

$$\text{kurtosis} = \frac{1}{N}\frac{\sum_{i=1}^{N}(x(i)-\bar{x})^4}{\sigma^4} \tag{3.9}$$

where $\bar{x}$ and σ are respectively the mean value and the standard deviation of the measured signal x, and N is the number of data samples of the measured signal. In this thesis work, N is set equal to the sampling rate. A signal with a Gaussian distribution yields a kurtosis value of 3.0 whereas the kurtosis of a signal with a sub-Gaussian distribution is less than 3.0. An impulsive signal with super-Gaussian distribution has a kurtosis value greater than 3.0, a typical feature of a mechanical fault. Therefore, considering both energy and kurtosis would provide more reliable scale selection process as follows:

$$a_{best} = \left\{ a \middle| \frac{dE_a}{da} = 0 \text{ and } \max(K_a) \text{ and } K_a > 3.0 \right\} \qquad a \in [a_{\min}, \ldots, a_{\max}] \tag{3.10}$$

where K_a is kurtosis at scale a and a_{best} is the best scale.

A mechanical fault signal is normally contaminated by noise and other interfering signals as simulated in Fig. 3.3. The periodic impulses are completely veiled by noise. To extract the trace of periodic impulses, kurtosis and energy values are calculated from the wavelet coefficients over scales as shown in Fig. 3.4. From these distributions, the best scale level can be selected by Eq. (3.10).

3.1.3 Autocorrelation coefficient function at the best scale

Based on Eqs. (3.6) and (3.10), the wavelet coefficients at the best scale a_{best} can be approximately expressed as

$$W(a_{best}, b) \approx C_1\cos(\omega_1 b) - C_2\sin(\omega_1 b) \tag{3.11}$$

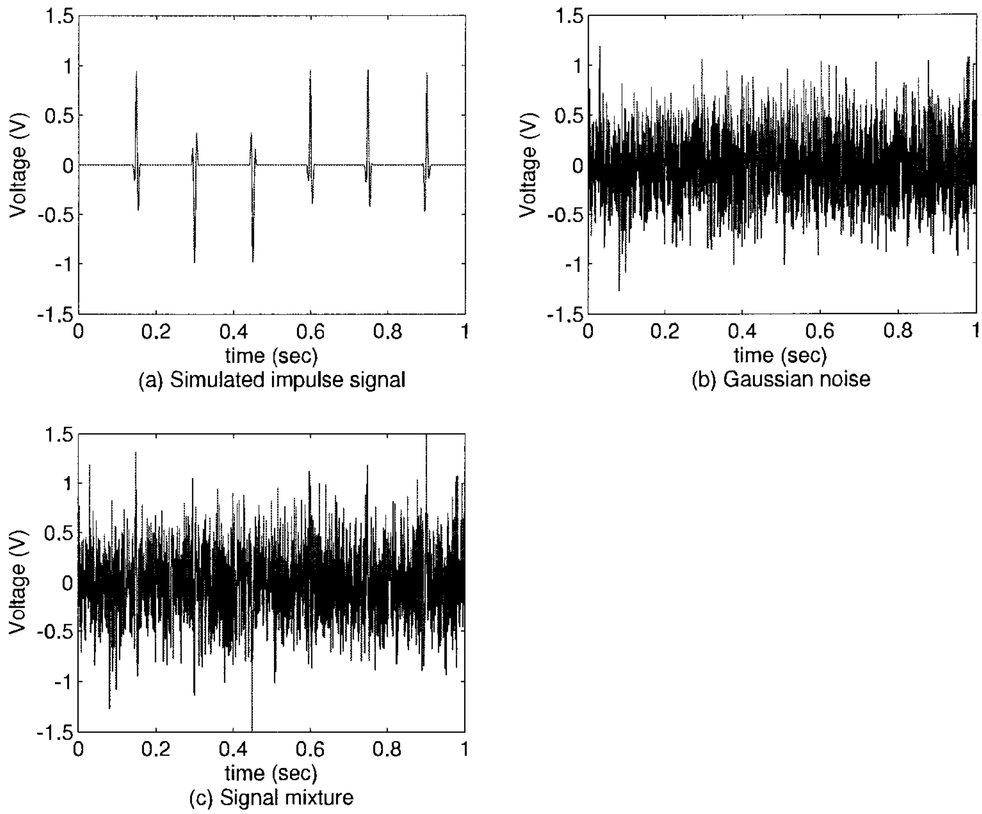


Figure 3.3. Simulated signal mixture

where

$$C_1 = \frac{1}{\sqrt{a}}\left\{\frac{1}{4}\sqrt{2\pi a^2}\cdot\exp\left(-\left(\omega_1-\frac{\omega_0}{2a}\right)^2\Big/\left(\frac{2}{a^2}\right)\right)\right.$$
$$\left.+\frac{1}{4}\sqrt{2\pi a^2}\cdot\exp\left(-\left(\omega_1+\frac{\omega_0}{2a}\right)^2\Big/\left(\frac{2}{a^2}\right)\right)\right\}$$

and

$$C_2 = \frac{1}{\sqrt{a}}\left\{-\frac{1}{4}\sqrt{2\pi a^2}\cdot\exp\left(-\frac{\omega_0^2}{8}\right)\right.$$
$$+\frac{1}{4}\sqrt{2\pi a^2}\cdot\exp\left(-\left(\omega_1-\frac{\omega_0}{2a}\right)^2\Big/\left(\frac{2}{a^2}\right)\right)$$
$$\left.-\frac{1}{4}\sqrt{2\pi a^2}\cdot\exp\left(-\left(\omega_1+\frac{\omega_0}{2a}\right)^2\Big/\left(\frac{2}{a^2}\right)\right)\right\}.$$

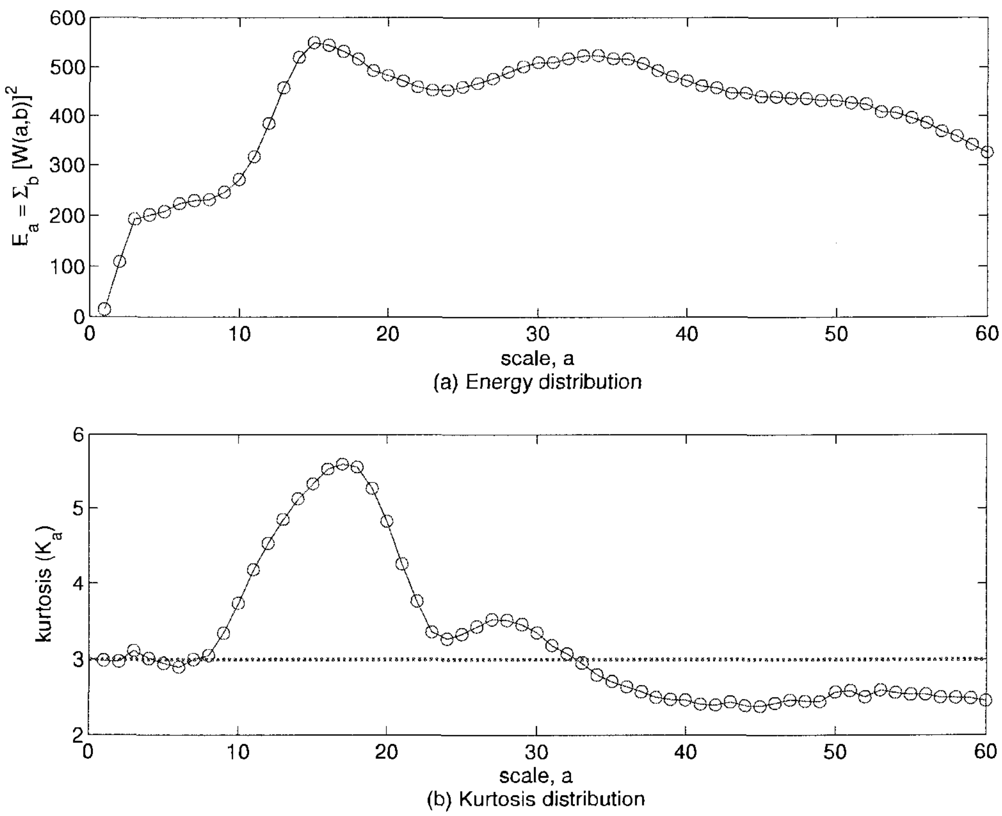


Figure 3.4. Energy and kurtosis distributions of wavelet coefficients over scales

It should be noted that, although the wavelet coefficients at the best scale contain frequency information of impulses of interest, they also include unwanted signal components such as noise that could result in spurious impulses. These unwanted components often cloud the periodicity of fault signature, making the signature undetectable. As suggested by Proakis and Manolakis (1996), autocorrelation can be used to detect the presence of a periodic signal corrupted by noise. Following this approach, the autocorrelation coefficients are calculated for the wavelet-transformed signal at the best scale level based on the following modified autocorrelation coefficient function based on Bendat and Piersol

(1991):

$$\begin{aligned}\rho_{WW}(\tau) &= \frac{C_{WW}(\tau)}{\sigma_W^2} \\ &= \frac{E\left[\left|W(a_{best},\ b)\right| \cdot \left|W(a_{best},\ b+\tau)\right|\right] - \left\{E\left[W(a_{best},\ b)\right]\right\}^2}{\sigma_W^2}\end{aligned} \tag{3.12}$$

where $E[\cdot]$ represents the expected value, τ is the time lag and $C_{WW}(\tau)$, $\rho_{WW}(\tau)$ and σ_W are respectively the covariance, the autocorrelation coefficient and the standard deviation of wavelet-transformed signal at the best scale a_{best}. Assuming the signal has a zero mean ($\bar{x} = E[x(t)] = 0$), the autocorrelation coefficient function can be written as (See Appendix A.2 for derivation.):

$$\begin{aligned}\rho_{WW}(\tau) &\approx \frac{1}{\sigma_W^2} E\left[\left|W(a_{best},\ b)\right| \cdot \left|W(a_{best},\ b+\tau)\right|\right] \\ &\approx \frac{1}{\sigma_W^2} \lim_{M\to\infty} \frac{1}{2M} \left| C_3 \cos(\omega_1 \tau) - C_4 \sin(\omega_1 \tau) \right| \\ &\approx \frac{1}{\sigma_W^2} \lim_{M\to\infty} \frac{1}{2M} \left| C_3 \cos\left(\frac{2\pi\tau}{T_1}\right) - C_4 \sin\left(\frac{2\pi\tau}{T_1}\right) \right|\end{aligned} \tag{3.13}$$

where

$$\begin{aligned}C_3 &= M\left(C_1^2 + C_2^2\right) + \left(C_1^2 - C_2^2\right) \frac{\sin(2\omega_1 M)}{2\omega_1} \\ C_4 &= 2C_1C_2 \frac{\sin(2\omega_1 M)}{2\omega_1}\end{aligned}$$

and M is the maximum value of the time lag τ for calculating autocorrelation coefficients. Eq. (3.13) indicates that the autocorrelation coefficient function of the wavelet-transformed signal has a clear periodicity related to the fault. The wavelet-transformed signal contains positive and negative values (Fig. 3.5b) since the measured signal (or acceleration) does as well. The rectified version of the autocorrelation coefficient function is used to facilitate the identification of the peak locations and hence the impulses because the peak locations are corresponding to the impulse locations.

3.1.4 Finding peaks of periodic fault features

An autocorrelation coefficient function normally displays many ripples (Fig. 3.5c) that are mostly due to the higher carrier frequency of the daughter wavelet and non-periodic

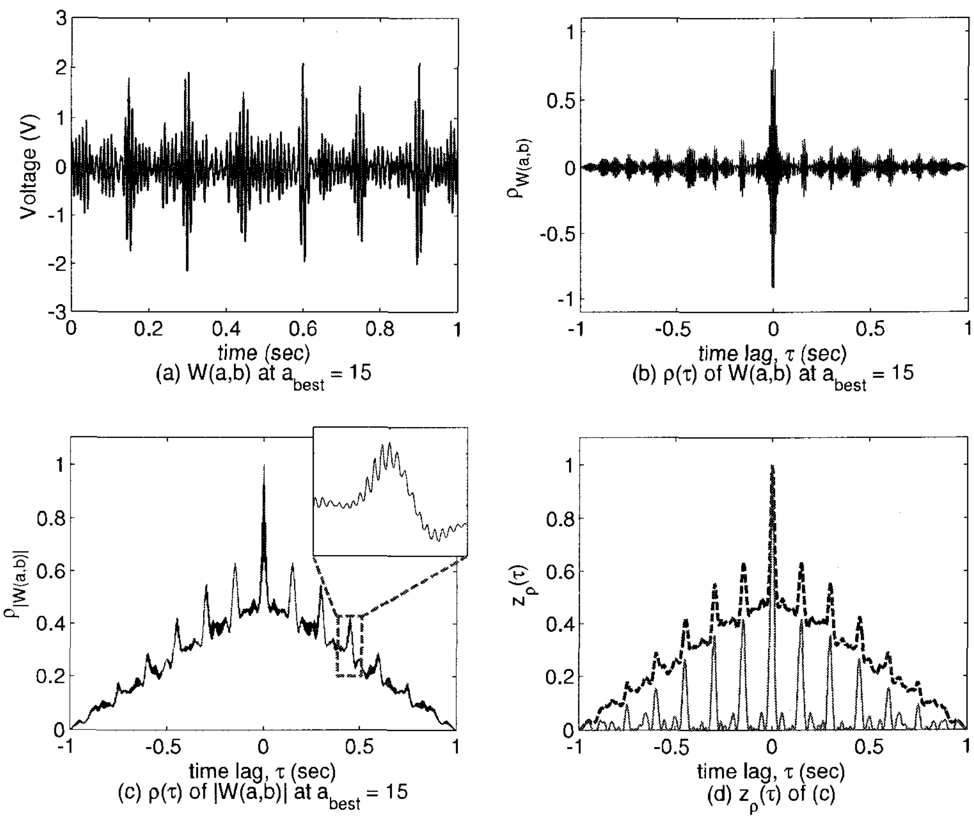


Figure 3.5. (a) Wavelet coefficients at the best scale a_{best}, (b) autocorrelation coefficient function of $W(a_{best}, b)$, (c) envelope of autocorrelation coefficient function of $|W(a_{best}, b)|$, and (d) rectified autocorrelation coefficient function (normalized by the maximum value)

signals such as noise. These ripples may cause difficulty to locate impulses. The envelope signal of the autocorrelation coefficient function can help estimate the peaks of impulses by reducing ambiguity, which can be done by removing the ripples. The envelope can be easily acquired by using the Hilbert transform (Bendat and Piersol, 1991).

$$z_\rho(\tau) = \rho_{WW}(\tau) + j\tilde{\rho}_{WW}(\tau) \tag{3.14}$$

where $z_\rho(\tau)$ is the analytical signal (or envelope), $\tilde{\rho}_{WW}(\tau)$ is the Hilbert transform of $\rho_{WW}(\tau)$ and j is an imaginary unit of the complex value. The Hilbert transform of real-valued function $\rho_{WW}(\tau)$ is defined by

$$\tilde{\rho}_{WW}(\tau) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\rho_{WW}(u)}{\tau - u} du. \tag{3.15}$$

The envelope signal obtained in Eq. (3.14) shows multiple peaks with constant period between peaks as shown in Eq. (3.13). These peaks indicate the time moments at which the impulses exist (Bendat and Piersol, 1991).

Since the absolute values of the wavelet-transformed signal are used in Eq. (3.12), there is an upward baseline on the left part of the envelope plot and a downward baseline on the right (Fig. 3.5c and d). As time lag increases, the peaks become less distinguishable because of the inclined baselines. To facilitate the selection of the impulse-associated peaks, the following actions are taken:

1. Push the baselines down to zero (Fig. 3.5d), and
2. Square the envelope of autocorrelation coefficient function to amplify the peaks (the solid line in Fig. 3.6).

Note that if the envelope $z_\rho(\tau)$ were calculated using the original autocorrelation coefficient function (Bendat and Piersol, 1991), i.e.,

$$\begin{aligned}\rho_{WW}(\tau) &= \frac{C_{WW}(\tau)}{\sigma_W^2} \\ &= \frac{E\big[W(a_{best},\ b)\cdot W(a_{best},\ b+\tau)\big] - \big\{E\left[W(a_{best},\ b)\right]\big\}^2}{\sigma_W^2}\end{aligned} \tag{3.16}$$

the autocorrelation coefficients and the squared envelope would be those shown in Fig. 3.5b and in Fig. 3.6 (dashed line), respectively. Obviously, the peaks of this plot (dashed line for $W(a_{best}, b)$ in Fig. 3.6) cannot be as easily identified as the one in the solid line (for $|W(a_{best}, b)|$) in Fig. 3.6.

3.1.5 Determining the acceptable time interval for the comb filter

To apply the comb filter, an acceptable time interval has to be determined. The time intervals between the peaks identified based on the autocorrelation function are not

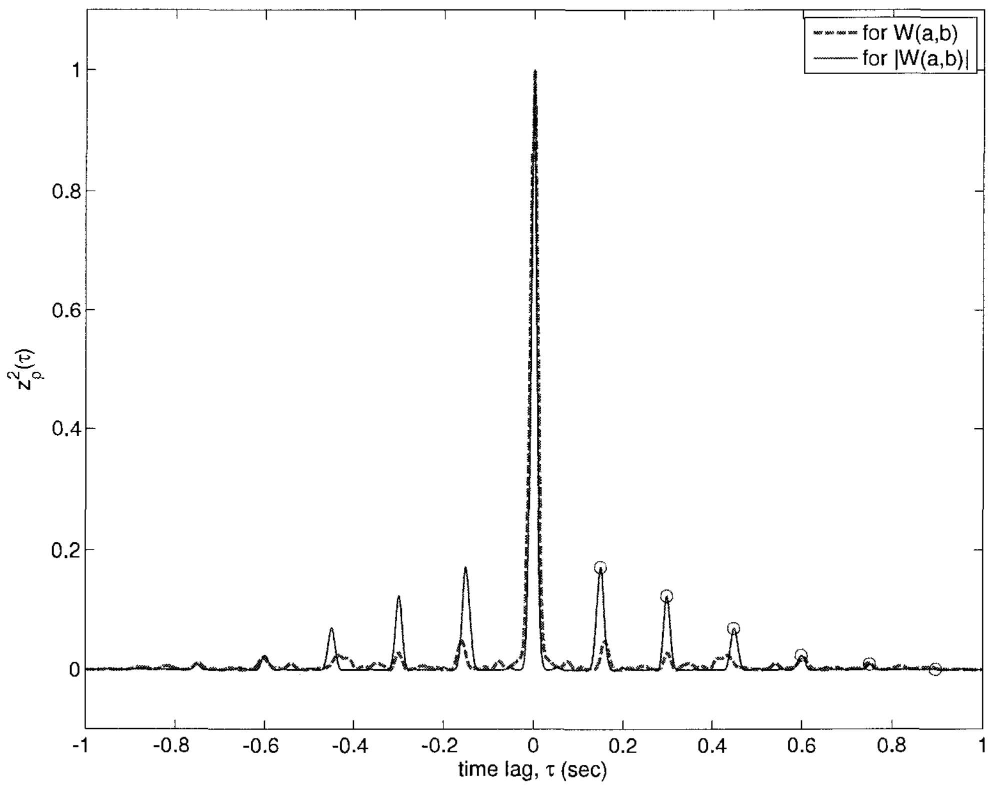


Figure 3.6. Comparison of two squared envelopes of autocorrelation coefficient functions obtained from $W(a_{best}, b)$ (dashed line) and $|W(a_{best}, b)|$ (solid line) at scale $a_{best} = 15$

identical. This error is caused by the decreased data samples as the time lag increases during the calculation of autocorrelation coefficients. To quantify this error, the time interval between two consecutive peaks is calculated:

$$\Delta t_i = t_{i+1} - t_i \qquad i = 1,\, 2,\, \ldots,\, J-1 \tag{3.17}$$

where t_i is the time instant determined from the peaks of autocorrelation coefficient function and J is the number of peaks obtained. The time interval can be considered as a random variable following normal probability density function (PDF) (Papoulis and Pillai, 1998).

$$f(\Delta t) = \frac{1}{\sigma_{\Delta t}\sqrt{2\pi}} \exp\left(-\frac{(\Delta t - \Delta t_m)^2}{2\sigma_{\Delta t}^2}\right) \tag{3.18}$$

where Δt_m and $\sigma_{\Delta t}$ are the mean and the standard deviation of time intervals. The possibility of the peak to be located within the range $[-3\sigma_{\Delta t}, 3\sigma_{\Delta t}]$ centered at t_i is 99.7%. In other words, the range of $[-3\sigma_{\Delta t}, 3\sigma_{\Delta t}]$ determines the acceptable time interval that contains impulses generated by mechanical fault (Abbate et al., 2002; Papoulis and Pillai, 1998). Hence, the normal distribution function shown in Eq. (3.18) can be used as a filter unit to extract fault features residing in the range of $[-3\sigma_{\Delta t}, 3\sigma_{\Delta t}]$ around the mean value, Δt_m. A comb filter can be composed by allocating this filter unit at the instants of the selected peaks (t_i's) in the time domain. Then, a comb filter specified by a normal distribution function, Eq. (3.18), is now ready to be applied to the wavelet-transformed signal and to extract the hidden fault signature. However, to preserve the magnitude of the original signal, a new function is used as a filter unit because $f(\Delta t_m) = 1$.

$$f(\Delta t) = \exp\left(-\frac{(\Delta t - \Delta t_m)^2}{2\sigma_{\Delta t}^2}\right) \tag{3.19}$$

This means no amplification or shrinkage of the original signal.

To apply a comb filter in time domain, it is necessary to consider the following situations. Since signal $x(t)$ and signal $x(t+t_0) \quad (0 < t_0 < t_1)$ have the same autocorrelation coefficient function, the time instant of the first impulse (t_0) should be known. If the impulses are located at t_i's ($i = 1, 2, \ldots, J$), the comb filter can be obtained by simply locating filter unit at t_i's. However, the impulses may be located anywhere between two consecutive t_i's because of translation, which requires estimating the deviation in time between the impulse and t_i. This time deviation is obtained by examining the time instant (t_0) of the maximum amplitude between time periods $[0,\ t_1]$. If $0 < t_0 < t_1$, the first impulse is located at time instant t_0. Hence, the comb filter is composed by locating filter units at time instants $(t_0, t_0 + t_1, t_0 + t_2, \ldots, t_0 + t_J)$. Otherwise, the filter units are located at time instants $(t_1, t_2, t_3, \ldots, t_J)$. To facilitate the selection of the maximum amplitude, the absolute value of the wavelet-transformed signal, $|W(a_{best}, b)|$, is used between time instants $[0,\ t_1]$.

3.1.6 Evaluation criterion

Traditionally, the de-noising results are often evaluated using the mean squared error (MSE). As detailed later in Chapter 4, the MSE tends to over-weigh the importance of the discrepancy between the original impulses and the impulses at the same time locations in the de-noised results. The confusion caused by the spurious impulses standing between the fault-induced impulses is largely overlooked. However, our main purpose here is to find the time moments when the fault-induced impulses appear. Preserving the magnitudes or sizes of the fault-induced impulses is not important in fault detection and should not be of our main concern. A good algorithm not only should be able to find the instants when the true fault-induced impulses occur but should be capable of clearing away any other impulses as well. Therefore, an ideal evaluation criterion should reflect such de-noising quality. In other words, the criterion should focus on the quantity and magnitudes of spurious impulses, both causing confusion in fault detection. The accumulated false energy of the spurious impulses can be used to quantify the combined effect of the number and magnitudes of spurious impulses and is adopted as performance criterion. It is defined as:

$$\begin{aligned} E_f &= \sum_{i=1}^{N} \tilde{s}_i^2 \\ \tilde{s}_i &= \begin{cases} \hat{s}_i & , \text{ if } s_i = 0 \\ 0 & , \text{ elsewhere} \end{cases} \end{aligned} \tag{3.20}$$

where s_i and $\hat{s}_i$ are the original impulse and the de-noised signal at time index i, respectively, N is the number of data samples, and E_f is the accumulated false energy contributed by spurious impulses.

3.2 Numerical illustration

A signal mixture (Fig. 3.3c) containing simulated impulses (sampled at 2,000Hz) 0.15 sec apart (Fig. 3.3a) and strong noise (Fig. 3.3b) is used to illustrate the procedure of the proposed method. As shown in Fig. 3.3c, it is difficult to detect the impulses directly from the signal mixture due to the strong noise. The proposed method is then applied according to the following steps:

Perform CWT and select the best scale

In this illustration, the Morlet wavelet is used as a basis wavelet function and the maximum scale level of 60 is applied since any scale above 60 is not of our interest. The time-scale domain information of the simulated signal mixture is then obtained by CWT. For each scale level, energy and kurtosis values are calculated and plotted in Fig. 3.4. As shown in Fig. 3.4a, there are two energy peaks around scales of 15 and 33. The two scale values associated with the two energy peaks are selected as the candidates of the best scale. The fact that the kurtosis value at scale 15 is substantially higher than that near scale 33 indicates that scale 15 is a better choice. In addition, as the kurtosis value at scale 33 is below 3, indicating a sub-Gaussian distribution, scale 33 can be excluded even though it has a high energy level. Therefore, scale 15 is selected as the best scale a_{best}.

Obtain autocorrelation coefficient function

An autocorrelation coefficient function is obtained based on the best scale (scale 15) selected in the previous paragraph. The wavelet-transformed signal and the envelope of the autocorrelation coefficient function are depicted in Fig. 3.5. Even though the selected best scale could reveal some weak trace of impulses (Fig. 3.5a), noise components make it difficult to distinguish the impulses associated with a mechanical fault. The envelope of autocorrelation coefficient function shows peaks corresponding to simulated impulses. To amplify the peaks caused by fault impulses, the autocorrelation coefficient function is squared and the result is shown in Fig. 3.6 after pushing the baseline to zero.

Find peaks of impulses and determine acceptable time interval

The time instants where the peaks appear are selected from the squared envelope of autocorrelation coefficient function. The selected peaks are marked with circles in Fig. 3.6. To illustrate the difficulty of determining an acceptable time interval, the fast Fourier

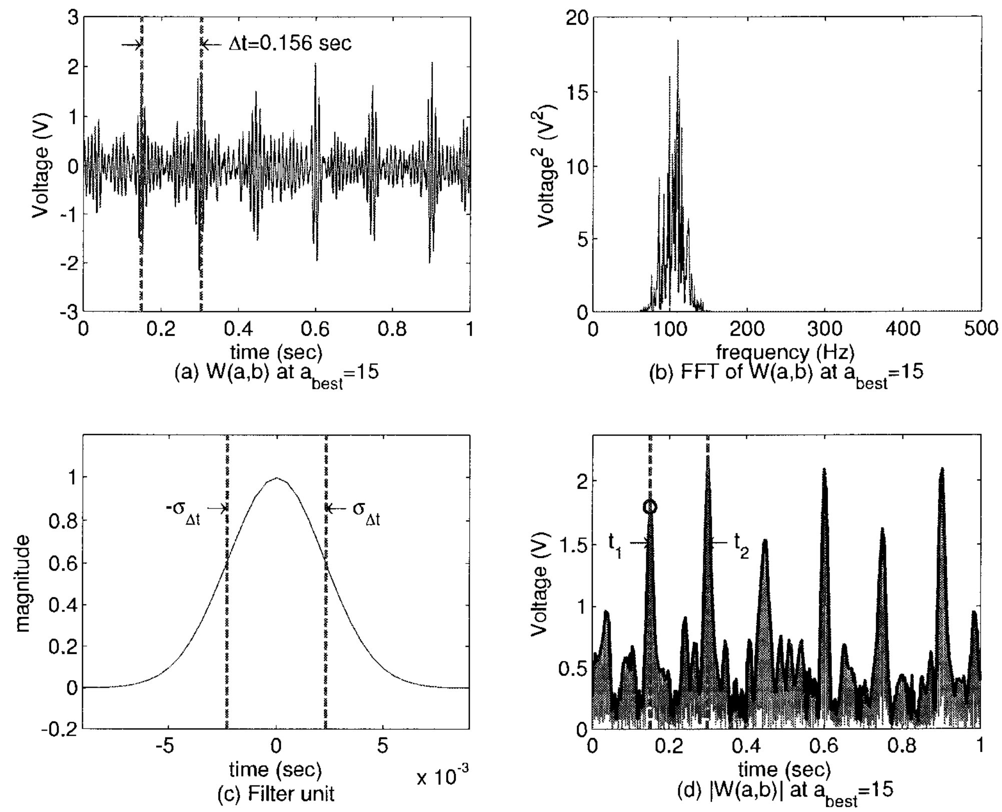


Figure 3.7. (a-b) Illustration of the difficulty in using regular bandpass filter, (c) the proposed filter unit, and (d) translation of impulses

transform (FFT) of the wavelet-transformed signal at the best scale a_{best} is shown in Fig. 3.7b. There are many peak frequencies around the carrier frequency about 100Hz, which makes it difficult to determine proper cut-off frequencies. The daughter wavelet of the Morlet wavelet at scale $a_{best} = 15$ has a pseudo frequency about 108Hz by the following relation (Mathworks, 2007).

$$f_a^{ps} = \frac{f^c}{a_{best} \cdot 1/F_s} = \frac{0.8125}{15 \cdot 1/2000} = 108.3\text{Hz}$$

where f_a^{ps}, f^c, and F_s are the pseudo frequency of scale a_{best}, the center frequency of Morlet wavelet function, and the sampling frequency of input signal, respectively. The frequency of impulses is about 6Hz ($\Delta t = 0.156$sec) as shown Fig. 3.7a. The frequency of 6Hz is not clearly shown in Fig. 3.7b. The wavelet-transformed signal at the best scale a_{best} shows frequency components mostly around the pseudo frequency, a center frequency of a daughter wavelet determined by the scale a_{best} or 108Hz in this example. Hence, the amplitude of the fault feature frequency is small. Its frequency located in thin tail, the far left side of the pseudo frequency, and hence cannot be displayed clearly in the Fourier transform plot. The comb filter unit determined from Eq. (3.19) is shown in Fig. 3.7c. This comb filter unit is applied at the time instants identified based on the autocorrelation coefficient function. The filter unit is located at time instants t_i's because time instants (t_i, in dashed line) selected from the autocorrelation coefficients coincide with the impulses of the fault feature ($t_0 = 0$, the first impulse marked with circle is located at t_1) as shown in Fig. 3.7d.

Apply filter to extract hidden impulses

To extract fault feature, a comb filter shown in Fig. 3.8a is applied to the wavelet transformed signal at scale a_{best} (Fig. 3.8b). The filtered signal and the simulated impulses are shown in Figs. 3.8 c and d, respectively. In terms of spurious impulses, the de-noised signal in Fig. 3.8c shows very good results. There are no significant spurious impulses between simulated impulses. The filtered signal shown in Fig. 3.8c yields zero false energy, i.e., $E_f = 0.0V^2$. For comparison, the same signal mixture shown in Fig. 3.3c is also de-noised using soft- and hard-thresholding rules with the universal threshold (Donoho, 1995). The original simulated impulses, the de-noising results obtained using the proposed method, the soft- and hard-thresholding rules are presented in Fig. 3.9 a, c, d and e, respectively. As shown in Fig. 3.9 d and e, the results of the two thresholding methods contain too many spurious impulses. In the soft-thresholding result, the original impulses are heavily truncated. Though the hard-thresholding rule does not truncate the original impulse as much as the soft-thresholding rule does, many spurious impulses are boosted to the level of the truncated impulses. In either case, it is very difficult to locate the original impulses. This can also be reflected by the high false energy values of the thresholding results that are $E_f = 0.236V^2$ and $E_f = 9.504V^2$ for the soft- and hard-

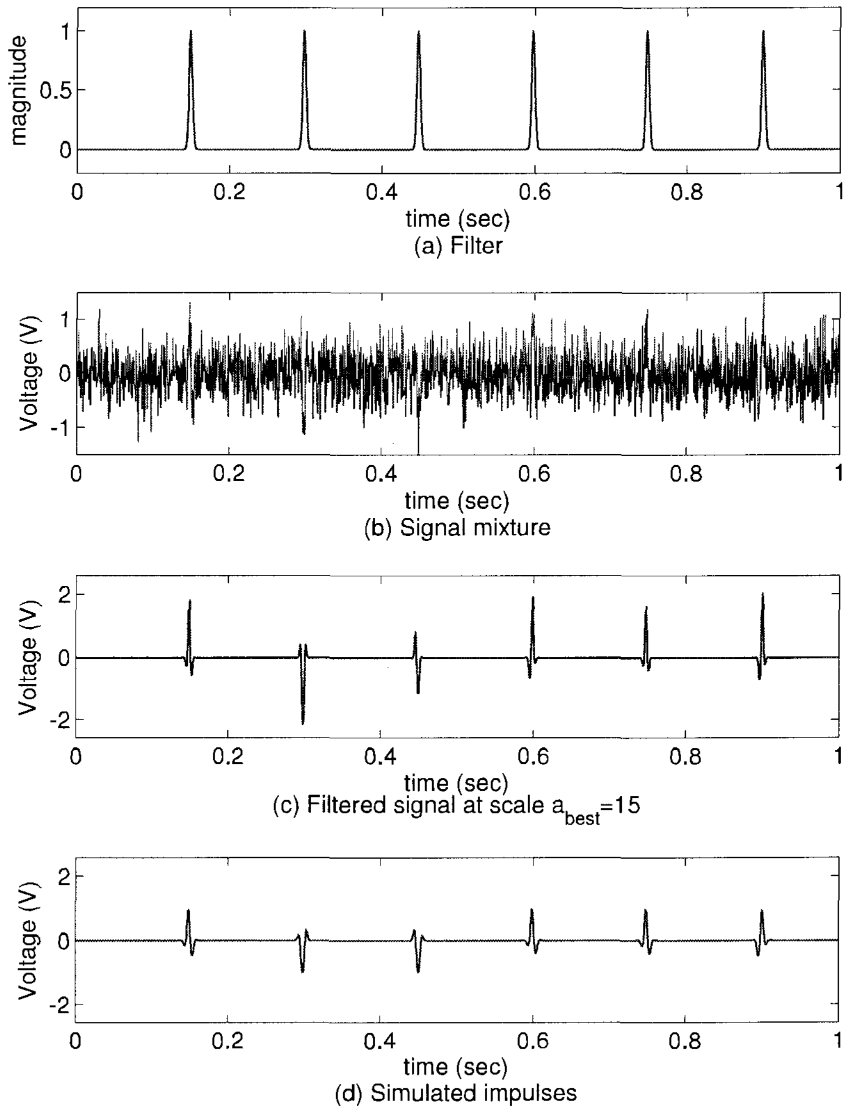


Figure 3.8. (a) Comb filter, (b) simulated signal mixture, (c) filtered signal, and (d) simulated impulses

thresholding rules, respectively. As shown in this comparison, the proposed method can remove spurious impulses and improve fault identification results substantially. Fig. 3.9 also illustrates that the MSE values are less relevant for our purpose. In fact, the MSE

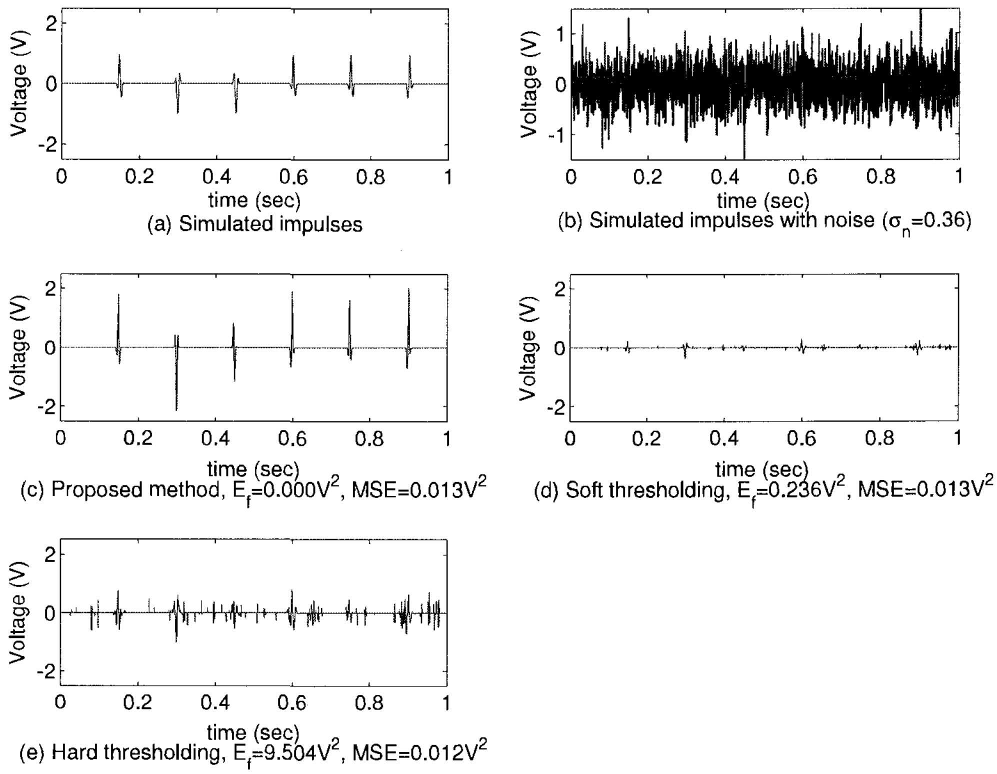


Figure 3.9. Comparison of the proposed method with soft- and hard-thresholding

values obtained from the hard- and soft-thresholding rules (0.012 and 0.013) are smaller than or equal to the one yielded by the proposed method (0.013), but both rules fail to extract any useful features.

3.3 Application to gear fault signal

The proposed method is applied to de-noising gear fault signal and facilitating the detection of the fault. The experimental setup of a gearbox simulator is shown in Fig. 3.10. The motor shaft is connected to the gearbox input shaft through a driving belt. The belt sheaves have diameter ratio of 1 to 2.56. Hence, the speed ratio of motor shaft to gearbox input shaft is 2.56 to 1. The motor shaft speed is 24Hz and the gearbox input shaft speed is about 9.4Hz. The faulty gear signal is acquired by an accelerometer (Montronix VS100-100 with 100mV/g sensitivity) mounted on the gearbox case at a sampling frequency of 12,000Hz. The acquired acceleration data is sent to a DAQ board (NI PCI-6040E with 12-bit resolution) for signal processing using MATLAB. Unnecessary high frequency components are removed by a digital Chebyshev I type filter with the following characteristics: cut-off frequency=240Hz and stop frequency=300Hz. Two cases, i.e., a healthy gear and a gear with one missing tooth, are considered. The gear with a missing tooth generates impulses of 9.4Hz or one impulse every 0.106 sec.

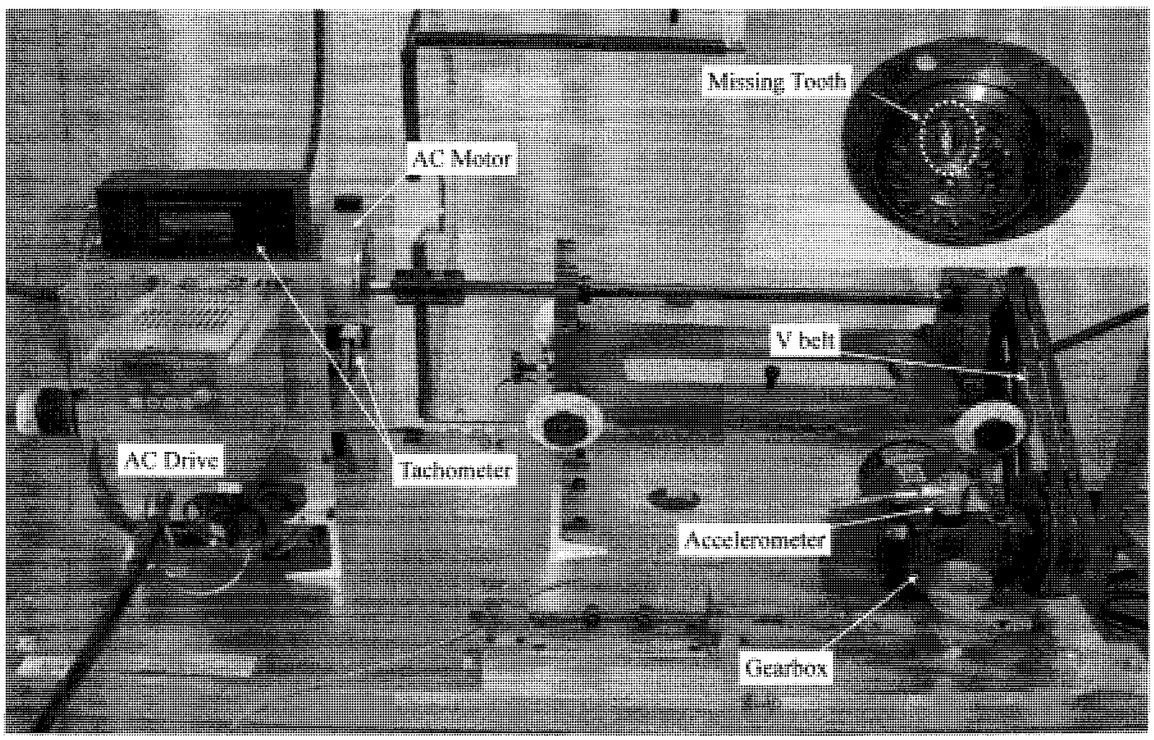


Figure 3.10. Experimental setup

3.3.1 Healthy gear

Fig. 3.11a shows the measured signal from the healthy gear. To select the best scale of CWT, the energy and the kurtosis are calculated and plotted in Fig. 3.11 b and c, respectively. As shown in Fig. 3.11b, there is an energy peak around scale 59. However, the kurtosis value around that scale is less than 3.0 (Fig. 3.11d). A signal with such a low kurtosis value can be considered as a sub-Gaussian noise and should not contain periodic impulses. According to Fig. 3.11b, there are no additional peaks observed in other scale ranges. Furthermore, Fig. 3.11d shows that the entire scale range from 35 to 60 belongs to sub-Gaussian state and thus this range can be excluded. For the rest of the scales

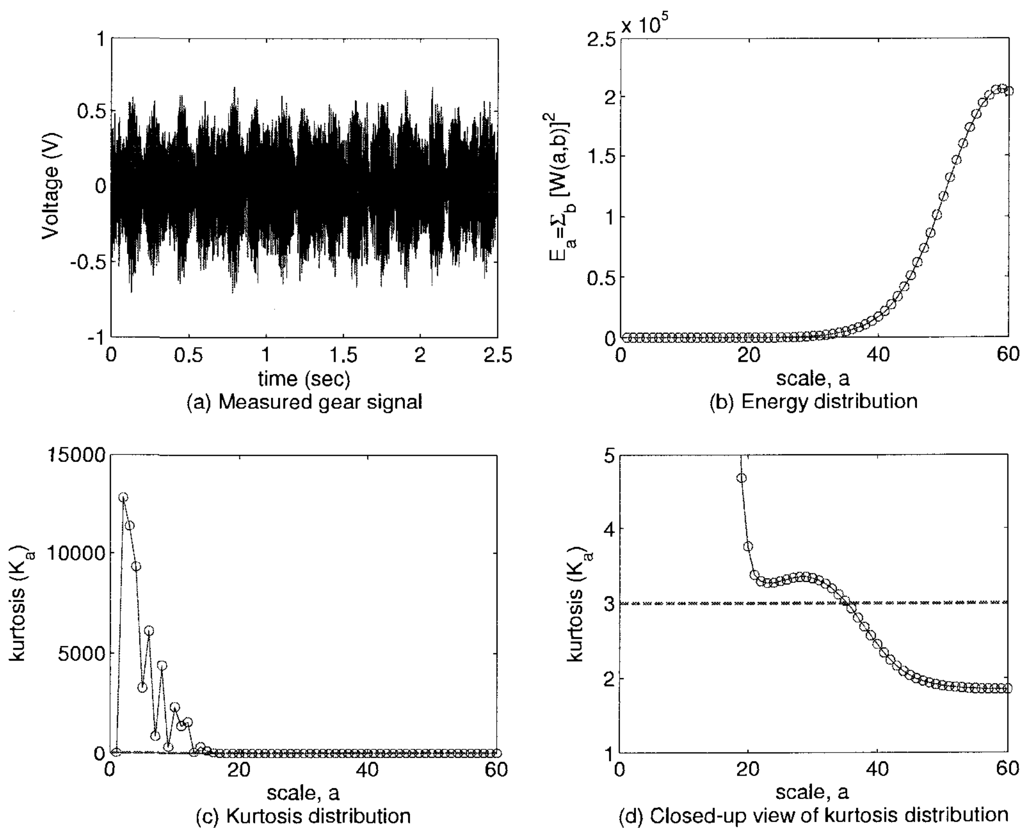


Figure 3.11. Healthy gearbox signal

from 0 to 35, the energy is practically zero and no peaks exist which indicates a complete mismatch between the daughter wavelets and the fault signal. The high kurtosis value alone may imply the existence of impulses and does not indicate the presence of any impulses related to the concerned fault. Hence, it can be concluded that the signal shown in Fig. 3.11a is in a healthy state.

3.3.2 Faulty gear with one missing tooth

The measured signal of the faulty gear is shown in Fig. 3.12a. Though the impulsive nature can be observed, it is not easy to identify the impulses related to the fault.

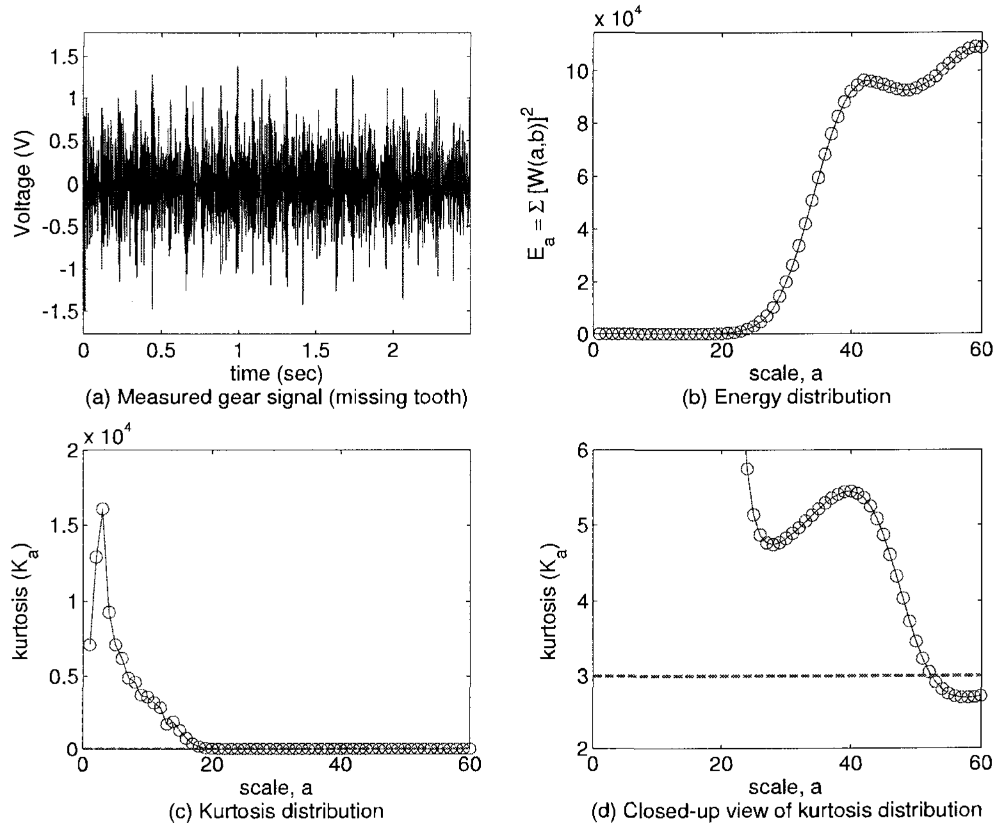


Figure 3.12. (a) Measured gear fault signal (missing tooth), (b) energy distribution of CWT, and (c-d) kurtosis values

Based on the energy and kurtosis distributions of the wavelet transformed signals shown in Fig. 3.12 b to d, scale $a_{best} = 42$ is selected as the best scale level. The wavelet transformed signal and its envelope of autocorrelation coefficient function are shown in Fig. 3.13. The periodic nature of impulses can be seen in Fig. 3.13a. However, there are too much noise components to separate the impulses corresponding to the mechanical fault. To select the time moments when the impulses reside, the autocorrelation coefficient function is obtained and shown in Fig. 3.13b. The selected time moments of impulses are shown in Fig. 3.13c with squared autocorrelation coefficient function. The

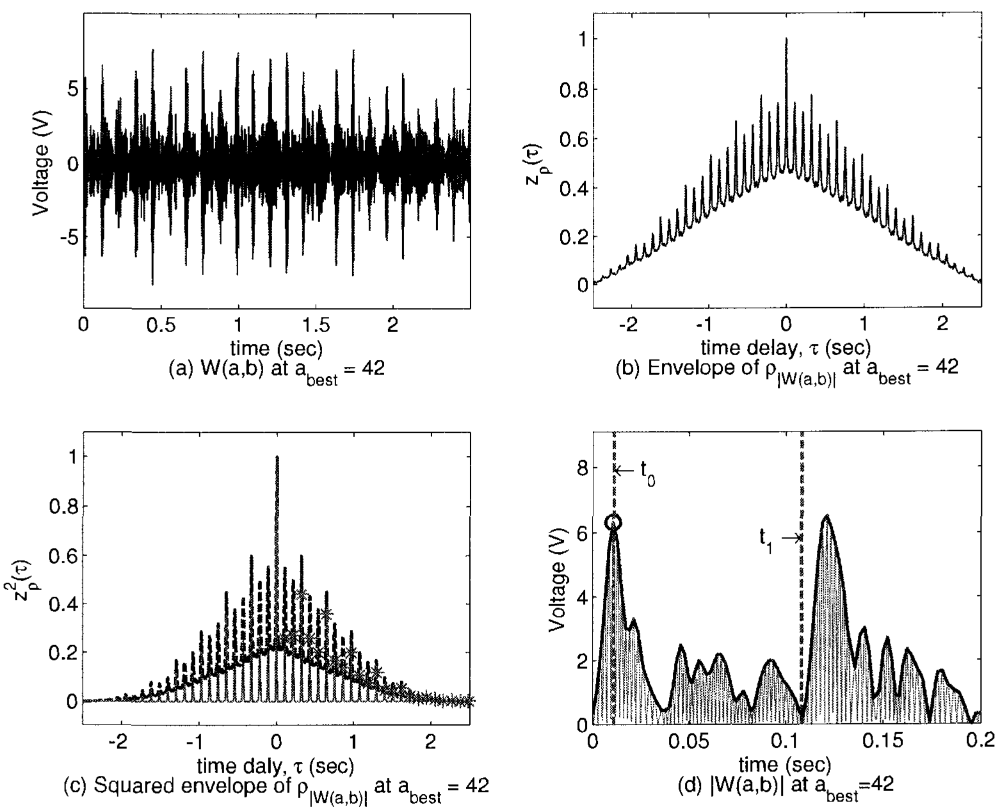


Figure 3.13. (a) Wavelet transformed signal, (b) envelope of autocorrelation function, (c) squared auto-correlation coefficient function with peaks (dashed line: before removing baselines, solid line: after removing baselines), and (d) translation of impulses at the best scale

selected locations are marked with asterisk "$*$". The translation of impulses is indicated by a circular marker and dashed line in Fig. 3.13d. Because $t_0 \neq 0$, the comb filter is composed by locating filter units at time instants $(t_0, t_0 + t_1, t_0 + t_2, \ldots, t_0 + t_J)$. After applying the filter, the extracted fault feature is shown in Fig. 3.14. The periodic impulses are clearly seen in this figure without any significant spurious impulses.

3.4 Discussion

Extracting fault feature is the basis for the reliable fault diagnosis. Most fault features are blended with noise signal generated by various sources, which makes it difficult to

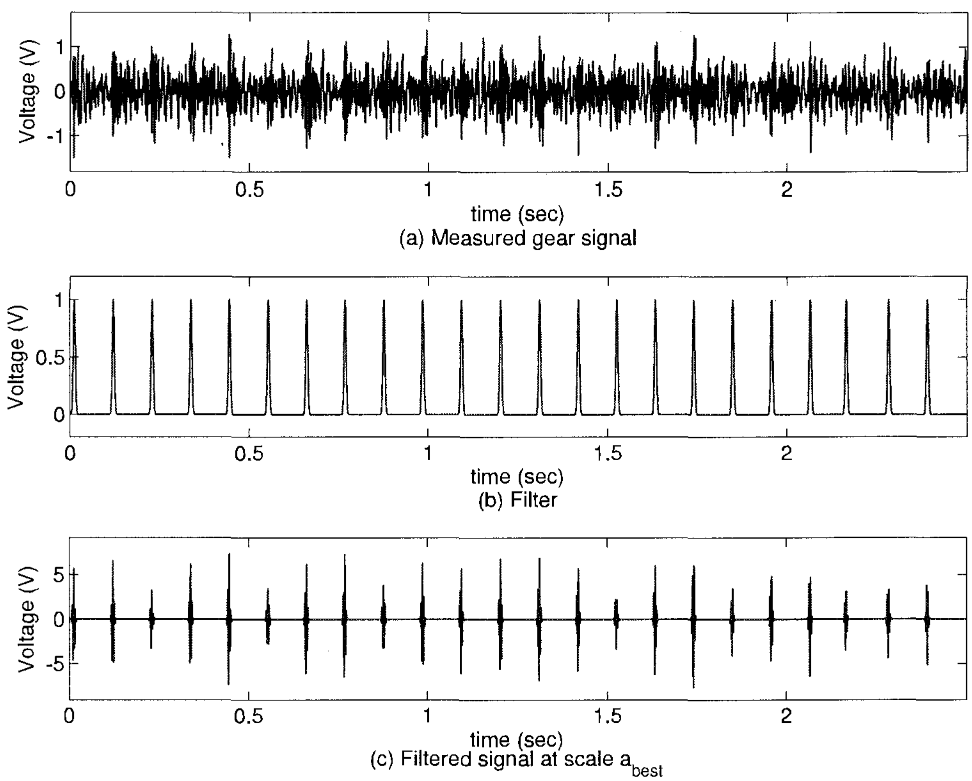


Figure 3.14. Extracted fault feature of missing tooth

identify fault features from measured signal. In this chapter, the fault feature extraction method is introduced for periodic signals. As demonstrated using both simulated and experimental data, the proposed method can reduce the spurious signals, the major obstacle in identifying the impulsive fault features. As a result, the impulsive fault can be effectively identified. The feature extraction method that can also be applied to non-stationary signals is discussed in the next chapter.

Chapter 4

K-Hybrid: A Kurtosis Based Hybrid Thresholding Method for Mechanical Signal De-noising[1]

This chapter presents a kurtosis based hybrid thresholding method, *K-hybrid*, for de-noising mechanical fault signals. The threshold used in the hybrid thresholding method is determined based on kurtosis which is an important indicator of the signal-to-noise ratio (SNR) of a signal. This together with its sensitivity to outliers and data-driven nature makes a kurtosis-based thresholding particularly suitable for on-line detection of mechanical faults featuring impulsive signals. To better reflect the signal composition, the proposed hybrid thresholding rule divides the wavelet transformed input signals into four zones associated with different de-noising actions. This alleviates the difficulties present in the simple *keep-or-remove* and *shrink-or-remove* approaches adopted by the hard- and soft-thresholding rules. The boundaries of the four zones are on-line adjusted in response to the kurtosis change of the signal. Simulation results suggest that the mean squared error (MSE) is unable to distinguish the results in terms of the amount of falsely identified impulses. It is therefore inappropriate to use MSE alone for evaluating the de-noising results of mechanical signals. As such, a combined criterion incorporating both

[1]Appeared in *Transactions of ASME Journal of Vibration and Acoustics*, vol.129, iss.4, pp.458-470. See Appendix B.4 for MATLAB script.

MSE and false identification power (P_{false}) is proposed. An analysis has shown that the proposed K-hybrid approach outperforms the soft, hard, and BayesShrink thresholding methods in terms of the combined criterion. It also compares favorably with the MAP thresholding method for signals with low kurtosis or low SNR. The proposed approach has been successfully applied to noise reduction and fault feature extraction of bearing signals.

4.1 Overview of Wavelet-Thresholding De-noising

The fundamentals of the wavelet thresholding de-noising are briefly introduced in this section. Let $x(t)$ be the signal measured from a sensor at time t. The measured signal can be considered as a combination of impulses (by faults) and noise (from Eq. (3.1)):

$$x(t) = s(t) + n(t)$$

where $s(t)$ denotes the impulses generated by faults at time t and $n(t)$ is the Gaussian noise with zero mean and standard deviation σ_n. The impulses and the noise components are assumed independent of each other. The task of de-noising is to find an estimated impulse signal $\hat{s}(t)$.

The overall procedure of the wavelet de-noising is shown in Fig. 4.1. The wavelet threshold de-noising is normally performed in three steps: wavelet transform, thresholding, and reconstruction. In the first step, the measured signal, $x(t)$, is transformed into wavelet domain by either continuous wavelet transform (CWT) or discrete wavelet transform (DWT). If CWT is applied, one obtains (from Eq. (3.2))

$$W(a,b) = \int x(t)\, w^*_{a,b}(t)\, dt$$

where "$*$" indicates the complex conjugate, $W(a,b)$ is the wavelet coefficient, and $w_{a,b}(t)$ is the daughter wavelet defined as (from Eq. (3.3))

$$w_{a,b}(t) = \frac{1}{|a|^{1/2}} w\left(\frac{t-b}{a}\right)$$

where a is the scale factor, b is the translation factor in time, and $w(\cdot)$ is the mother wavelet function or wavelet basis function. The factor $|a|^{-1/2}$ is used to ensure energy preservation (Goupillaud et al., 1984).

Then, in the second step, a thresholding rule is applied to the wavelet coefficients to remove the noise. The key is to find a proper threshold for the given input wavelet coefficient and to develop an effective thresholding rule to remove noise and retain the impulse as much as possible. The universal threshold, $T_{univ} = \sigma_n\sqrt{2\log N}$ (Donoho, 1995), which is based on the standard deviation of noise and the number of data points, has been widely used in combination with soft- and hard-thresholding as defined in the following (Donoho, 1995):

$$Out_s = \begin{cases} \text{sign}(Inp)\,(|Inp| - T_{thr}) & , |Inp| \geq T_{thr} \\ 0 & , |Inp| < T_{thr} \end{cases} \tag{4.1}$$

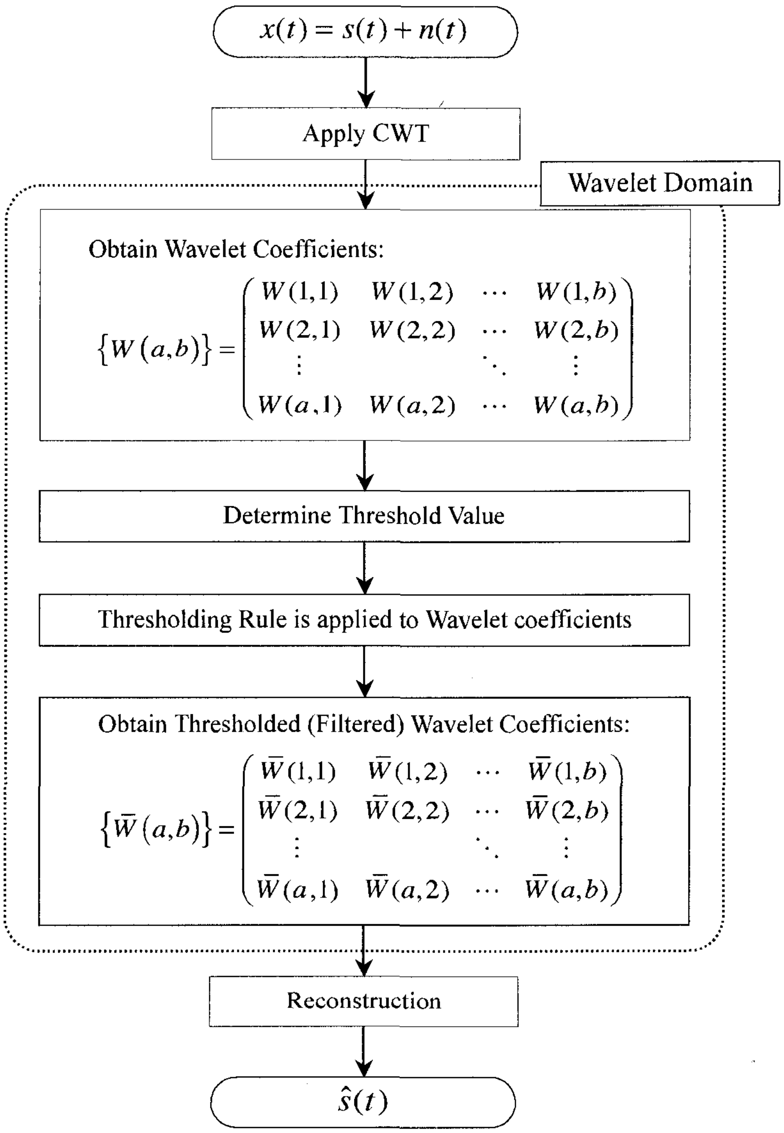


Figure 4.1. Flowchart of the wavelet de-noising method

$$Out_h = \begin{cases} Inp & , |Inp| \geq T_{thr} \\ 0 & , |Inp| < T_{thr} \end{cases} \tag{4.2}$$

where T_{thr} is the threshold value, Out_s and Out_h are respectively the outputs of soft- and hard-thresholding rules for a given input Inp (a wavelet coefficient of a scale). From the above equations, it is obvious that the soft-thresholding rule either removes the input if it is below the threshold T_{thr} or shrinks it by T_{thr} if the input is greater than T_{thr}. The hard-thresholding rule, on the other hand, takes a keep-or-remove approach depending on the magnitude of the input.

In step three, the filtered wavelet coefficients are inversely transformed into time domain to reconstruct the impulses $\hat{s}(t)$, the approximation of the true impulses. The Morlets formula (Rioul and Duhamel, 1992) can be applied to the thresholding results, i.e., the filtered wavelet coefficients, as follows:

$$\hat{s}(t) = \frac{1}{C_{1w}} \int \overline{W}(a,b) a^{-3/2} da \tag{4.3}$$

$$C_{1w} = \int_{-\infty}^{\infty} \frac{\hat{w}^*(\omega)}{|\omega|} d\omega \tag{4.4}$$

where $\overline{W}(a,b)$ is filtered (or thresholded) wavelet coefficient, a is the scale level, ω is the angular frequency, and $\hat{w}^*(\omega)$ is the complex conjugate of Fourier transform of wavelet basis function.

4.2 Hybrid-Thresholding and Threshold Selection

4.2.1 Hybrid-thresholding rule

The hard-thresholding rule would provide perfect results if the threshold represented the true boundary between the noise and the impulse from the machine fault. However, in reality, the "boundary" between noise and impulse may vary due to the randomness and dynamics of the process, making it inappropriate to use a single valued threshold to distinguish the noise and true signal. Hence, the overly simplified keep-or-remove approach adopted by the hard-thresholding rule is inadequate: it may falsely classify a true impulse as noise or a noise as impulse for the inputs around the threshold. The soft-thresholding rule, on the other hand, goes to another extreme — it keeps only the portion that is above the threshold. This will lead to serious distortion of the signal for heavily contaminated signal with low SNR. Since the reconstruction of thresholded wavelet coefficients will be based on heavily truncated or shrunk values for signals with low SNR, the final de-noised result cannot reflect the true nature of the impulses and may leave the mechanical fault undetected due to the large artificial reduction of signal strength.

The above difficulties in applying soft- or hard-thresholding motivate the development of a hybrid thresholding approach. Using a properly chosen thresholding rule can reduce the effect of uncertainty of the threshold selection. A customized thresholding rule proposed by Yoon and Vaidyanathan (2004) could be considered as an example of hybrid-thresholding rule for reducing the uncertainty of threshold. It is shown that a properly chosen thresholding rule (adjusted between soft- and hard-thresholding rules) can improve de-noising performance.

This chapter proposes a different hybrid-thresholding rule with four zones that can be adjusted based on the statistical characteristics of the input. The objective is to reduce the effect of uncertainty of threshold by regulating these four zones. With the hybrid approach, the inputs according to their magnitudes are classified into four groups associated with the four zones as shown in Fig. 4.2: a reduced "remove" zone (zone 1) with a upper bound t_{1a}, two transient zones (zones 2 and 3) beside the original threshold t_{2a}, and a reduced "keep" zone (zone 4). Any inputs below t_{1a} are small enough to be considered as pure noise and are removed. The inputs above t_{3a} are large enough to be classified as true impulses and kept with their full magnitudes. The inputs in the two transient zones are considered to consist of both noise and impulse components. However, it is reasonable to assume that the inputs in zone 2 contain stronger noise components with weaker impulse contents and hence larger shrinkage rates are applied. It is also logical to expect that the inputs in zone 3 contain larger impulse components with less noise ingredient, which warrants the use of smaller shrinkage rates. The two transient zones will alleviate the difficulties present in the soft- and hard-thresholding

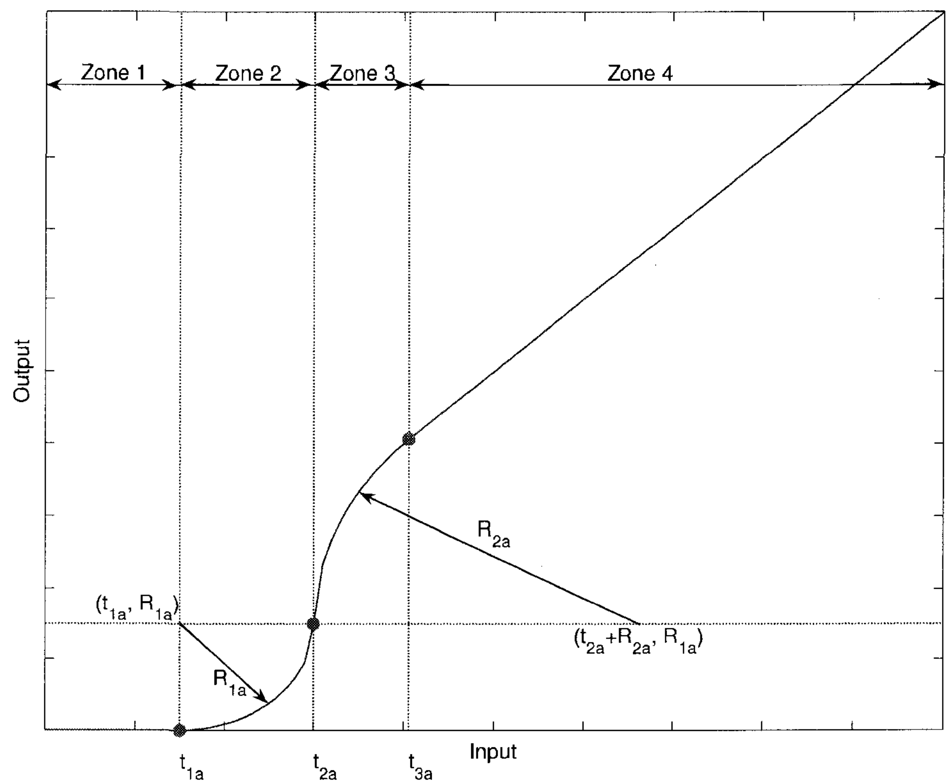


Figure 4.2. K-hybrid thresholding rule

rules and lead to smoother transition from zone 1 to zone 4. Better de-noising results can then be expected.

The hybrid thresholding rule reflecting the above scheme can be written as (using continuous wavelet transform, i.e., $Inp = W(a,b)$):

$$Out_{k_a} = \begin{cases} 0 & , |W(a,b)| < t_{1a} \\ sign\left(W(a,b)\right) \cdot \left[R_{1a} - \sqrt{C_{H_1}}\right] & , t_{1a} \leq |W(a,b)| < t_{2a} \\ sign\left(W(a,b)\right) \cdot \left[R_{1a} + \sqrt{C_{H_2}}\right] & , t_{2a} \leq |W(a,b)| < t_{3a} \\ W(a,b) & , t_{3a} \leq |W(a,b)| \end{cases} \tag{4.5}$$

$(a = 1,\, 2,\, \cdots,\, L)$

where

$$C_{H_1} = \left(R_{1a}^2 - t_{1a}^2\right) + \left(2t_{1a} + |W(a,b)|\right)|W(a,b)|,$$
$$C_{H_2} = -t_{2a}\left(2R_{2a} + t_{2a}\right) + 2\left(R_{2a} + t_{2a}\right)|W(a,b)| - |W(a,b)|^2.$$

$W(a,b)$ is the continuous wavelet transform coefficient, t_{2a} is the threshold which will be determined by a kurtosis-based approach individually for every scale and t_{1a}, t_{3a}, R_{1a}, and R_{2a} will be specified according to the t_{2a} value in each scale as follows.

$$\left.\begin{aligned} t_{1a} &= \frac{t_{2a}}{k_a} \\ t_{3a} &= \left(1 + \frac{1}{\sqrt{2}k_a}\right)t_{2a} \\ R_{1a} &= \left(1 - \frac{1}{k_a}\right)t_{2a} \\ R_{2a} &= \frac{1+\sqrt{2}}{k_a}t_{2a} \end{aligned}\right. \qquad a = 1,\, 2,\, \cdots,\, L. \tag{4.6}$$

With the above two radii R_{1a} and R_{2a}, two arcs can be obtained to make connecting curves tangential at the threshold value. This ensures a continuous smooth transition throughout the input range. It should be pointed out that other smoothing transition curves could also be used if better performance can be achieved. The details of k_a and t_{2a} will be explained in sections 4.2.2 and 4.2.3.

4.2.2 Kurtosis-based threshold

The universal threshold is commonly used to determine the threshold value. It is dependent on the number of data (N) that is not the statistical characteristic of the signal and hence leads to different results when different N values are selected. The effect of N may not be very significant and the universal threshold can remove the noise very well if $N \to \infty$ according to the studies reported in (Donoho, 1995; Donoho and Johnstone, 1994). However, this is suitable only for the orthogonal wavelet transform since the independent and identically distributed (IID) random noises are still IID after orthogonal transform (Donoho, 1995). The variables are not IID when the signal is transformed using a non-orthogonal wavelet function, e.g. the Morlet wavelet, as a basis wavelet function. Therefore, the thresholding performance may deteriorate when the universal threshold is used with the Morlet wavelet function (Lin et al., 2004). In addition, machine fault detection requires quick real time responses and hence the threshold should be data-driven to reflect the dynamics of the data. The universal threshold is hence not suitable for machine fault detection due to the large N requirement.

For the above reason, the thesis proposes a new threshold selection method based on kurtosis since it reflects the statistic nature of the signal and represents the trend similar to that of SNR. The idea is that signals can be classified in terms of kurtosis: a signal is classified as *sub-Gaussian* (kurtosis < 3.0, no impulse signal of interest), *super-Gaussian* (kurtosis > 3.0, impulses become stronger), and *Gaussian* (kurtosis $=$ 3.0, the noise dominates). It is therefore logical to specify threshold value based on the signal-to-kurtosis ratio. The new threshold is determined as

$$t_{2a} = \frac{\max\left(|W(a,b)|\right)}{Kurt\left(W(a,b)\right)/3} \qquad a = 1,\, 2,\, \cdots,\, L \tag{4.7}$$

where $W(a,b)$ is the wavelet coefficient and $Kurt\left(W(a,b)\right)$ is the kurtosis of the wavelet-transformed signal for a given scale level a. This threshold value is calculated separately for each scale. The wavelet-transformed signal in each scale is related to one of the following three cases depending on the gaussianity, which require different treatments:

Case 1: $Kurt\left(W(a,b)\right) = 3.0$.

The signal (wavelet transformed) is Gaussian. This signal does not show any impulses distinguishable from noise components and hence a high threshold, $t_{2a} = \max\left(|W(a,b)|\right)$, should be applied to remove heavy noise.

Case 2: $Kurt\left(W(a,b)\right) < 3.0$.

The signal is sub-Gaussian and is not our concern in this application since we are trying to identify impulses generated by mechanical fault. In this case, $t_{2a} > \max\left(|W(a,b)|\right)$, leading to zero output.

Case 3: $Kurt\left(W(a,b)\right) > 3.0$.

The signal is super-Gaussian and is our main concern since it contains outliers (or impulses) generated by the mechanical fault. Now, the threshold is less than $\max(|W(a,b)|)$ in order to preserve the impulse signal. Naturally, the threshold will decrease further as the kurtosis increases.

Since the kurtosis represents the trend similar to that of SNR, this approach is consistent with the commonly accepted guideline: use a high threshold if the SNR is low and choose a low threshold otherwise (Chang et al., 2000a; Lin and Qu, 2000). In addition, kurtosis is sensitive to outliers and is therefore especially suitable for processing mechanical fault signals mainly consisting of impulses.

4.2.3 Specification of parameter k_a

Since a higher kurtosis value indicates stronger impulse, a higher k_a is desirable to preserve the impulse components (See Eq. (4.6)). The opposite is true for signals with lower kurtosis values. The parameter k_a is therefore specified by

$$k_a = 1 + \max\left(Kurt\left(W(a,b)\right) - 3,\ 0\right) \tag{4.8}$$

From Eqs. (4.7) and (4.8), it can be seen that, if the signal is sub-Gaussian or Gaussian, the kurtosis is less than or equal to 3.0 and k_a is set to 1.0. In this case, $t_{1a} = t_{2a}$. This is because the makeup of the measured signal is mainly noise and hence all inputs up to this level can be confidently removed. As kurtosis increases, our confidence in removing all signal components declines. This, in the hybrid rule, is reflected by a reduced t_{1a} that is achieved by a larger k_a. If the signal is super-Gaussian, that is, the impulses are stronger than noise components, both kurtosis and k_a further increase. In the meantime, all the three boundary points t_{1a}, t_{2a}, and t_{3a} decrease. This is desirable since a larger kurtosis value indicates that the measured signal is mixed with weaker noise and stronger impulse. Therefore, it is reasonable to use a lower threshold. As kurtosis and k_a further increase, t_{3a} merges with t_{2a}, which is in consistence with our increased confidence that any input above this point is impulse and should be kept in full magnitude. The above phenomena are depicted in Fig. 4.3.

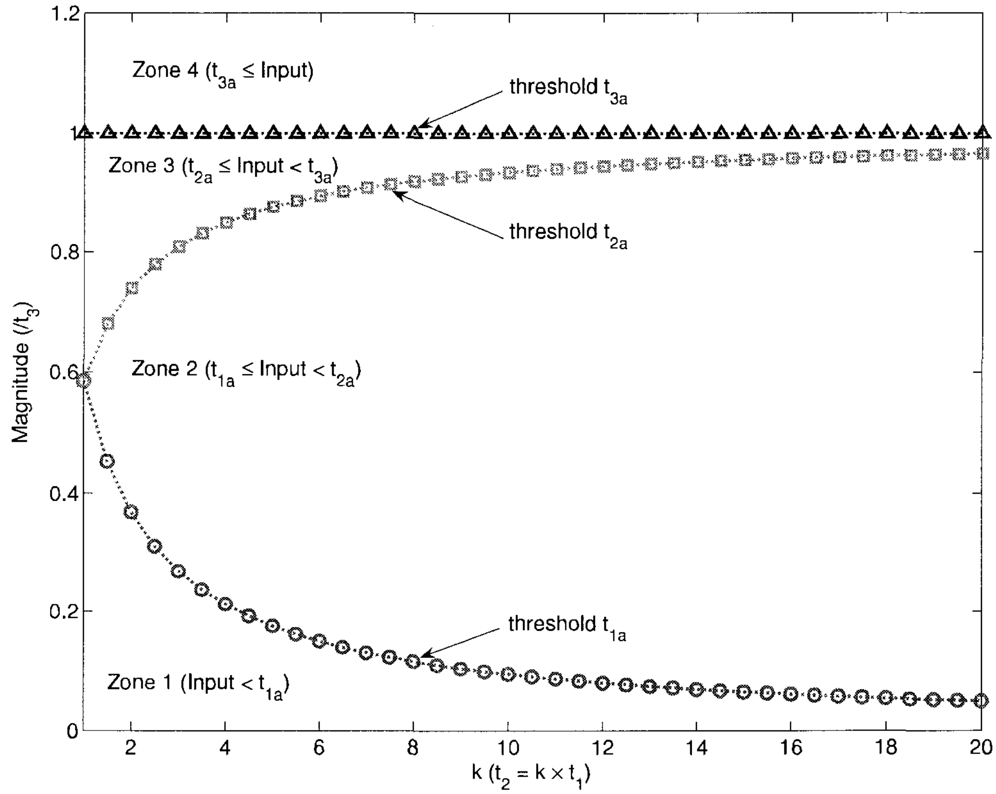


Figure 4.3. K-hybrid thresholding with four different thresholding zones (Parameter k_a determines the width of each zone)

4.3 Performance Evaluation

4.3.1 Evaluation criterion

The de-noising performance is generally evaluated by the mean squared error (MSE) defined as

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(s_i - \hat{s}_i)^2 \tag{4.9}$$

where s_i is a data point sampled from an actual impulse due to a mechanical fault, $\hat{s}_i$ is a data point from an estimated impulse (or de-noised signal), and N is the number of data points. s_i and $\hat{s}_i$ are discretized data samples of continuous time domain signal

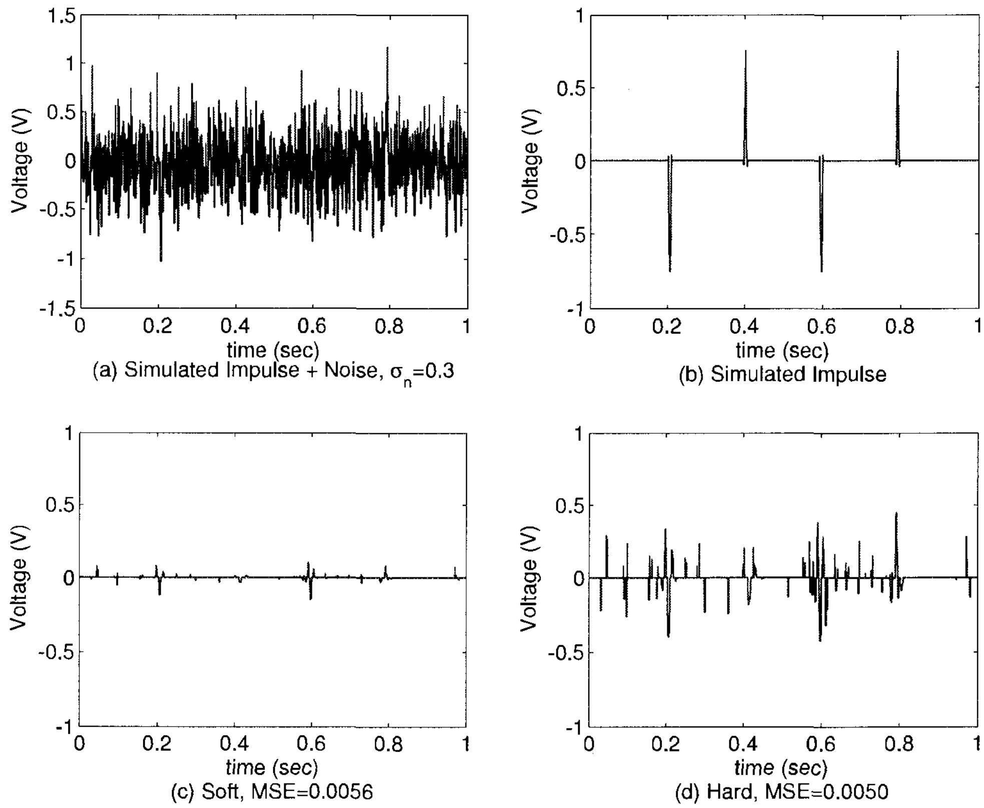


Figure 4.4. Limitation of MSE as a criterion for evaluating de-noising performance

$s(t)$ and $\hat{s}(t)$, respectively. Though the MSE has been used as a criterion to evaluate de-noising methods by many researchers, it is inadequate to use it alone. This is illustrated by Fig. 4.4. A simulated signal (Fig. 4.4a) composed of impulses (Fig. 4.4b) and heavy Gaussian noise is de-noised by the typical soft (Fig. 4.4c) and hard (Fig. 4.4d) thresholding rules. As shown in the figure, the de-noised result contains both true and falsely identified impulses. The hard-thresholding rule yields a smaller MSE (0.0050) (Fig. 4.4d) than the soft-thresholding rule does (Fig. 4.4c, MSE=0.0056). However, it is almost impossible to identify the original impulses from the result of hard-thresholding rule because of the large number of falsely identified fake impulses. On the other hand, the soft thresholding (Fig. 4.4c) reveals impulses with reduced magnitude even though there are fewer fake impulses (or spurious impulses). The reduced magnitude makes it difficult to identify true impulses as well. Obviously, a good de-noising result should contain as few fake impulses as possible while preserving the original magnitude as much as possible. This, however, as indicated in Fig. 4.4, is not always reflected by a low MSE value.

To compensate for the limitation of the MSE criterion, the false identification of the impulses should be considered as part of the performance measure. In this study, the false identification is quantified using the "power" of the fake impulses, P_{false}, given by

$$\begin{aligned} P_{false} &= \frac{1}{N}\sum_{i=1}^{N} \tilde{s}_i^2 \\ \tilde{s}_i &= \begin{cases} \hat{s}_i & \text{if } s_i = 0 \\ 0 & \text{elsewhere} \end{cases} \end{aligned} \tag{4.10}$$

where $\hat{s}_i$ is a data point sampled from a fake impulse and hence P_{false} is calculated only based on the data points from the fake impulses. To illustrate, consider Fig. 4.4b. A true impulse appears after every 0.2 seconds. Any non-zero values between these simulated impulses in the de-noising result are classified as false identification as shown in Fig. 4.4c and Fig. 4.4d since they do not exist in the real signal (Fig. 4.4b). Obviously, a good de-noising method should yield as few such fake non-zero values as possible, revealed by a low P_{false}. To evaluate the performance of a de-noising method more accurately, a new criterion expressed as a "cost" function incorporating both MSE and P_{false} is defined as follows:

$$C(T) = \arg\min_{T} \left(MSE + P_{false}\right) \tag{4.11}$$

4.3.2 Simulation tests and comparison with existing methods

The performance tests of the proposed threshold and thresholding rule were carried out using MATLAB with simulated signals. The Morlet wavelet is selected as a basis function for continuous wavelet transform which is defined as (from Eq. (3.4))

$$w(t) = \exp\left(-\frac{\gamma^2 t^2}{2}\right)\cos(\omega_0 t)$$

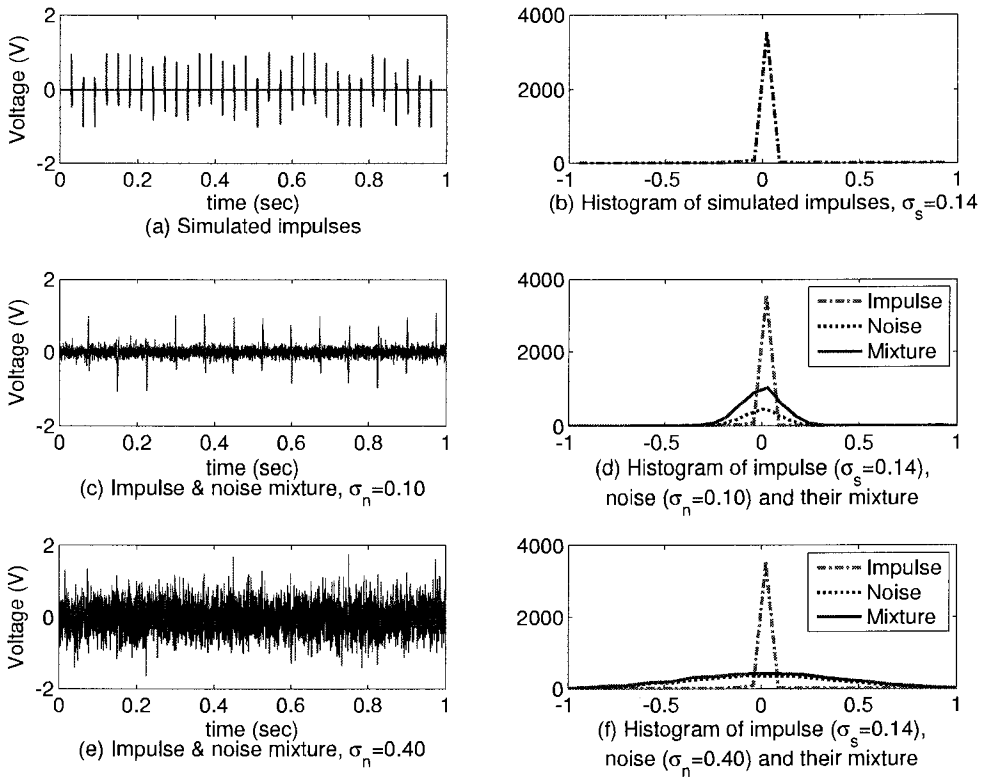


Figure 4.5. Simulated impulses (mean = 0), impulse and noise mixtures, and comparisons between histograms of impulse, noise and their mixture

where $\gamma = 2$ and $\omega_0 = 6$ are selected arbitrarily.

The simulated signals are composed of impulses and Gaussian noise with various intensities (or standard deviations of noise). The generated impulse signal shows super-Gaussian probability distribution as shown in Fig. 4.5a and Fig. 4.5b with standard deviation $\sigma_s = 0.14$ and zero mean. The Gaussian noise is generated by changing its standard deviation, σ_n.

As the noise intensity increases, the sparseness of the signal impulse is gradually submerged in the noise and is becoming less observable. This can be illustrated by comparing Fig. 4.5c through Fig. 4.5f. In Fig. 4.5c, the noise intensity is relatively low ($\sigma_n = 0.1$) and the histogram of the associated signal-noise mixture has very thin tails outside the $[-0.5, 0.5]$ region as shown in Fig. 4.5d. Such a distribution is super-gaussian and is a clear indication of the existence of outliers or impulses. However, when the noise intensity increases to a level of $\sigma_n = 0.4$ as shown in Fig. 4.5e, the tails of the histogram of the signal-noise mixture become much thicker outside the same region (Fig. 4.5f). In fact, the histogram of the signal-noise mixture shown in Fig. 4.5f is very close to Gaussian distribution and hence the impulses or outliers are swamped by noise. Therefore, de-noising technique is required to reveal the hidden impulses from the signal mixture contaminated by high-intensity noise. In the following sub-sections, the proposed K-hybrid method is compared with two sets of de-noising methods. First, it is compared with the typical soft- and hard-thresholding rules with the universal threshold. Second, it is compared with *BayesShrink* (Chang et al., 2000a,b), a near optimal soft-thresholding method in terms of MSE, and MAP (maximum *a posteriori*) thresholding (Hyvärinen, 1999), a method based on maximization of the posterior probability of impulses. The de-noising performance is evaluated in terms of the combined measure, $C(T) = \arg\min_T (MSE + P_{false})$.

A. Comparison of K-hybrid, soft- and hard-thresholding rules

As shown in Fig. 4.6, K-hybrid provides the best de-noising result in terms of the proposed cost function for the low-intensity noise ($\sigma_n = 0.1$). The hard-thresholding rule shows many falsely identified impulses between true impulses. For the signal with low-intensity noise, the impulses can be seen easily with bare eyes. Hence, a de-noising process may not be very crucial to identify the fault features (impulses). However, for the signal with high-intensity noise, the selection of threshold value is not an easy task because of the uncertainty of the exact boundary between impulse and noise. For this reason, the hard-thresholding rule with the universal threshold results in many fake impulses in the de-noised signal although it gives relatively larger magnitudes. Soft-thresholding provides much better results compared to the hard-thresholding rule in terms of the amount of fake impulses (or P_{false}). Nevertheless, it reduces impulse magnitude, which may mislead the decisions in fault detection.

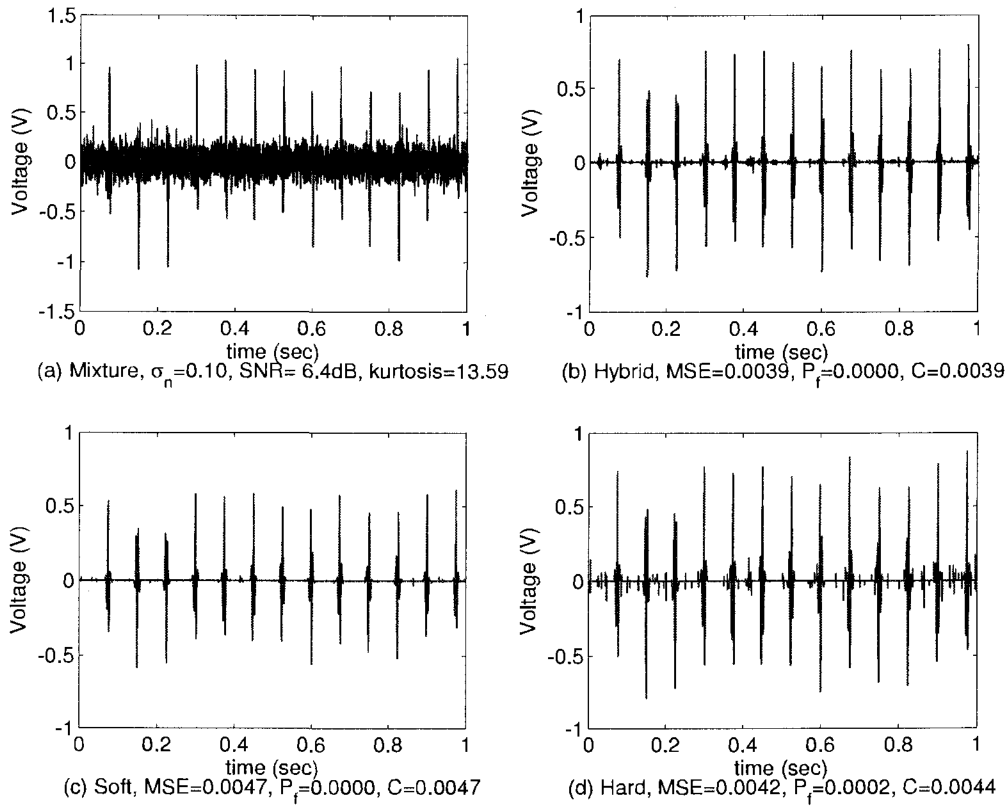


Figure 4.6. Comparison of de-noising results (K-hybrid, Soft and Hard, $\sigma_n = 0.1$)

As noise intensity increases (Fig. 4.7), the performance of the soft and hard rules deteriorates substantially. The uncertainty of the threshold value causes too many fake impulses in the hard-thresholding result, making it almost impossible to identify true impulses hidden in the noise. With the soft-thresholding rule, the magnitudes of impulses are drastically truncated (Fig. 4.7c) and hence it becomes very difficult to distinguish impulses from noise components. On the other hand, K-hybrid provides much better results comparing to both the hard and soft rules. It preserves much of the signal magnitude and makes it much easier to identify the true impulses. Only a few distinguishable fake impulses are seen in the de-noising results by K-hybrid as displayed in Fig. 4.7b. As shown in Fig. 4.6 and Fig. 4.7, the C values of K-hybrid results (0.0039, 0.0055) are much lower than those of the soft rule (0.0047, 0.0115) and hard rule (0.0044, 0.0078),

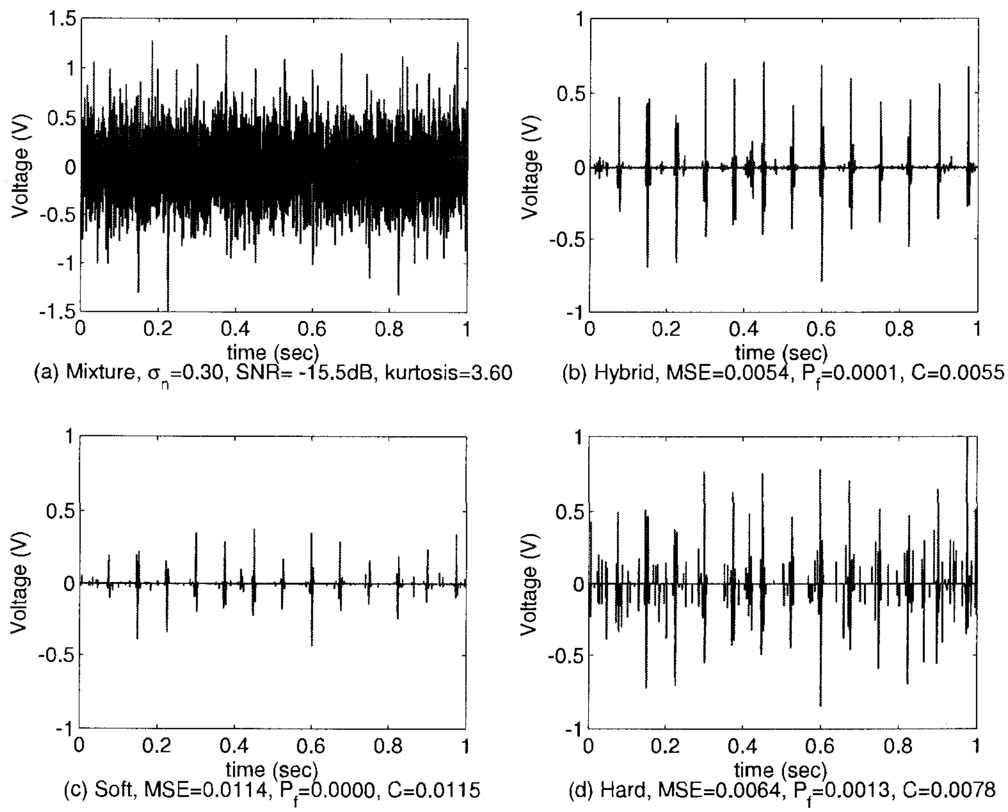


Figure 4.7. Comparison of de-noising results (K-hybrid, Soft and Hard, $\sigma_n = 0.3$)

which further demonstrates the appropriateness of using C value as a criterion. It is also interesting to note that the gap between the C values obtained by K-hybrid, soft and hard rules widens as the noise intensity level increases, reflecting the superiority of K-hybrid in handling severely contaminated signals.

Additional simulations were conducted to compare the three thresholding rules in a broader range of noise intensity spanning from Gaussian to super-Gaussian. The results are summarized in Table 4.1. As shown in the table, K-hybrid outperforms both the hard and soft thresholding methods throughout the entire test range.

Table 4.1. Comparison of de-noising results in terms of three performance measures

Noise (σ)	SNR (dB)	Kurtosis	K-hybrid			Soft			Hard		
			MSE	P_{false}	C	MSE	P_{false}	C	MSE	P_{false}	C
0.1000	6.4277	13.5949	0.0039	0.0000	0.0039	0.0047	0.0000	0.0047	0.0042	0.0002	0.0044
0.1500	-1.6816	8.0462	0.0040	0.0001	0.0041	0.0062	0.0000	0.0062	0.0044	0.0003	0.0048
0.2000	-7.4353	5.3862	0.0042	0.0001	0.0043	0.0079	0.0000	0.0079	0.0048	0.0006	0.0054
0.2200	-9.3414	4.7861	0.0043	0.0001	0.0045	0.0086	0.0000	0.0086	0.0051	0.0007	0.0058
0.2400	-11.0816	4.3467	0.0045	0.0001	0.0046	0.0093	0.0000	0.0093	0.0053	0.0008	0.0062
0.2600	-12.6825	4.0230	0.0047	0.0001	0.0048	0.0100	0.0000	0.0100	0.0056	0.0010	0.0066
0.2800	-14.1646	3.7827	0.0050	0.0001	0.0051	0.0108	0.0000	0.0108	0.0060	0.0012	0.0071
0.3000	-15.5445	3.6029	0.0054	0.0001	0.0055	0.0114	0.0000	0.0115	0.0064	0.0013	0.0078
0.3200	-16.5815	3.4531	0.0059	0.0001	0.0059	0.0127	0.0000	0.0127	0.0072	0.0015	0.0087
0.3400	-17.7941	3.3467	0.0064	0.0001	0.0065	0.0134	0.0000	0.0134	0.0079	0.0017	0.0095
0.3600	-18.9373	3.2640	0.0071	0.0001	0.0072	0.0140	0.0000	0.0141	0.0086	0.0019	0.0104
0.3800	-20.0186	3.1991	0.0079	0.0000	0.0080	0.0146	0.0000	0.0147	0.0093	0.0021	0.0114
0.4000	-21.0444	3.1478	0.0088	0.0000	0.0089	0.0152	0.0000	0.0152	0.0101	0.0023	0.0124

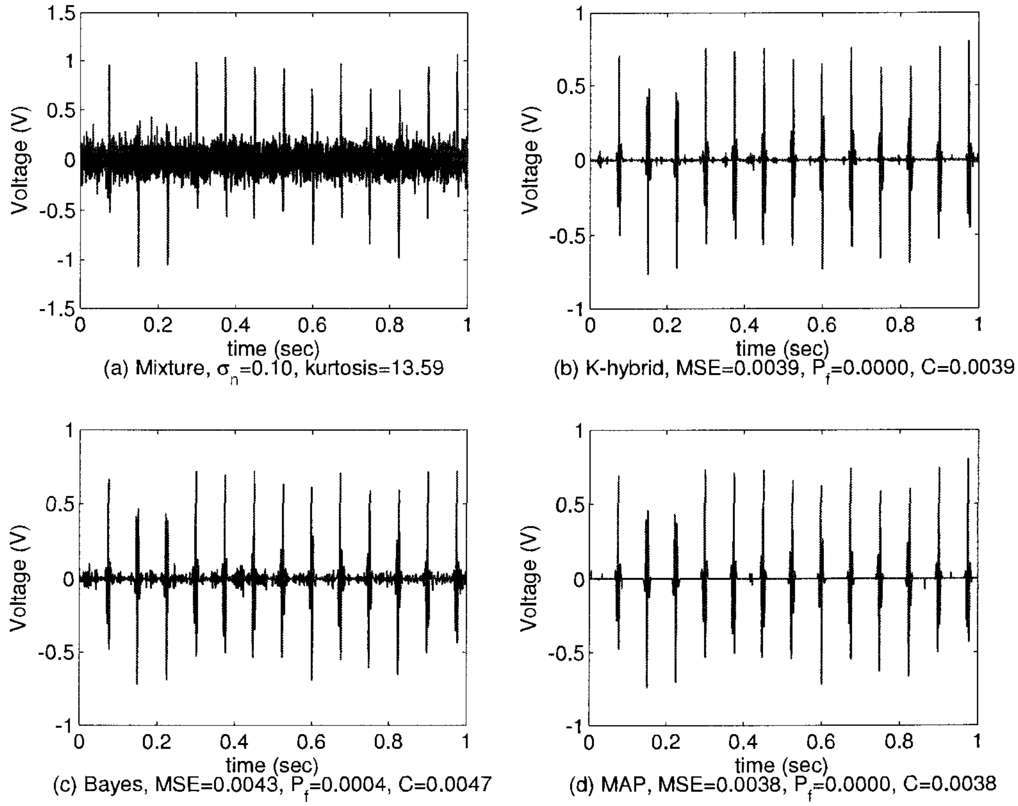


Figure 4.8. Comparison of K-hybrid, BayesShrink and MAP de-noising results ($\sigma_n = 0.1$)

B. Comparison of K-hybrid, BayesShrink and MAP thresholding rules

To further examine the performance of the proposed approach, comparisons are made with two more sophisticated de-noising methods, BayesShrink and MAP thresholding rule. The BayesShrink method uses the soft thresholding rule with an adaptive, data-driven threshold to minimize the *Bayesian risk.* Chang et al. (2000a,b) has shown that this de-noising method outperforms several other popular de-noising methods such as *OracleShrink* and *SureShrink* in image de-noising.

The thresholding rule based on the MAP estimator was proposed by Hyvärinen (1999) to reduce noise of the non-Gaussian data. It has been applied to de-noise mechanical signal and to detect faults by Lin et al. (2004). As there is no commonly recognized

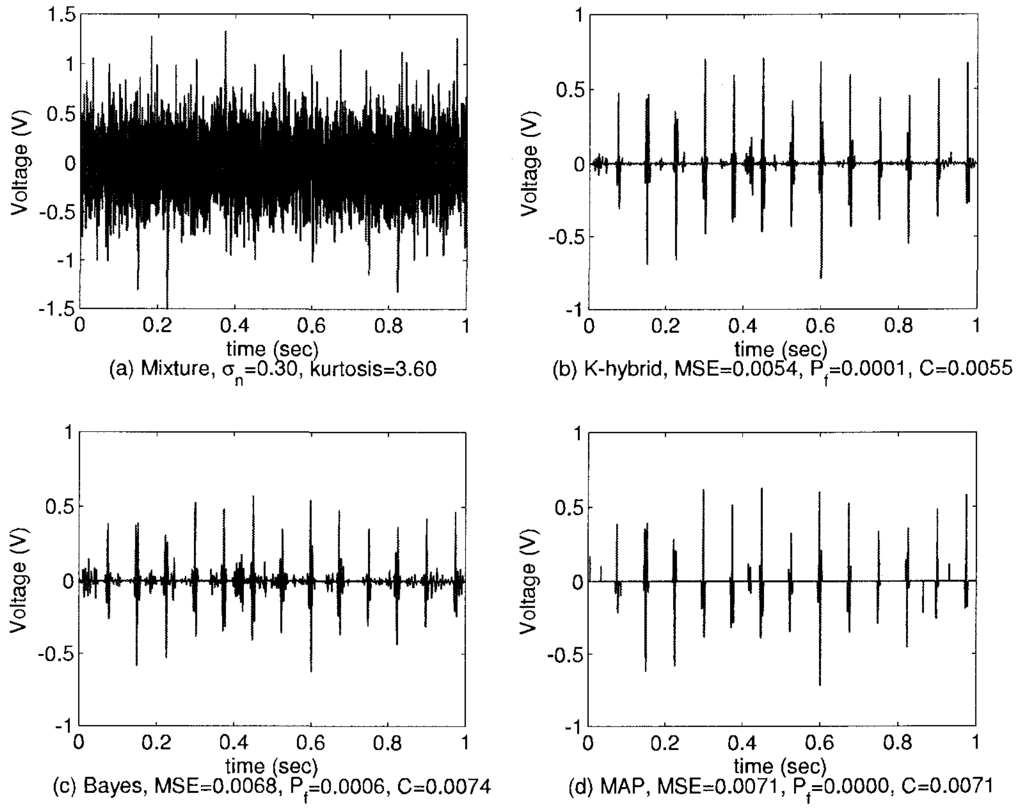


Figure 4.9. Comparison of K-hybrid, BayesShrink and MAP de-noising results ($\sigma_n = 0.3$)

name for this thresholding method, the author calls it *MAP-thresholding rule* in this study. According to Lin et al. (2004), the MAP-thresholding rule performs better than the optimal minimax hard- and soft-thresholding rules.

Here, K-hybrid is compared with these two de-noising methods by simulation. The simulated impulses and noise signals are the same as those used in the previous sections. The de-noising results for several typical cases are plotted in Fig. 4.8 through Fig. 4.10 associated with different noise intensity levels (σ_n values). The comparisons in terms of MSE, P_{false} and C with respect to kurtosis are displayed in Fig. 4.11. All the three methods show fairly good de-noising results in terms of C value when the kurtosis value (or SNR) is high. However, BayesShrink does not remove fake impulses well in a wide

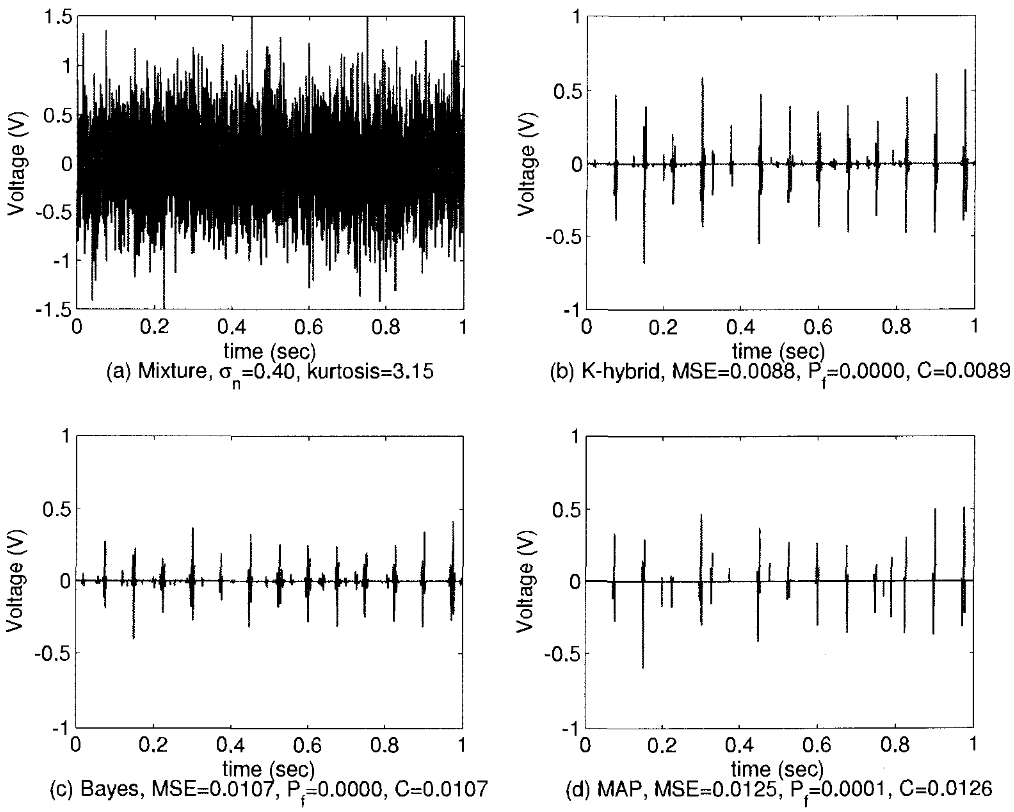


Figure 4.10. Comparison of K-hybrid, BayesShrink and MAP de-noising results ($\sigma_n = 0.4$)

range of kurtosis values (from 3.6 upwards), which is demonstrated by the high P_{false} values in Fig.4.11b. Though it performs reasonably well for low kurtosis values with no significant fake impulses as shown in Fig. 4.10, the impulse magnitudes are reduced substantially. This may restrict the detectability of the weak impulses generated by a mechanical fault.

The MAP-thresholding rule shows the lowest P_{false} in large part of the kurtosis range as shown in Fig. 4.11b. However, in the lower kurtosis range, the MAP-thresholding rule performs poorly in terms of all the three criteria: MSE, P_{false} and C (Fig. 4.11 a, b, and c). This is a serious drawback because this region represents signals with most severe

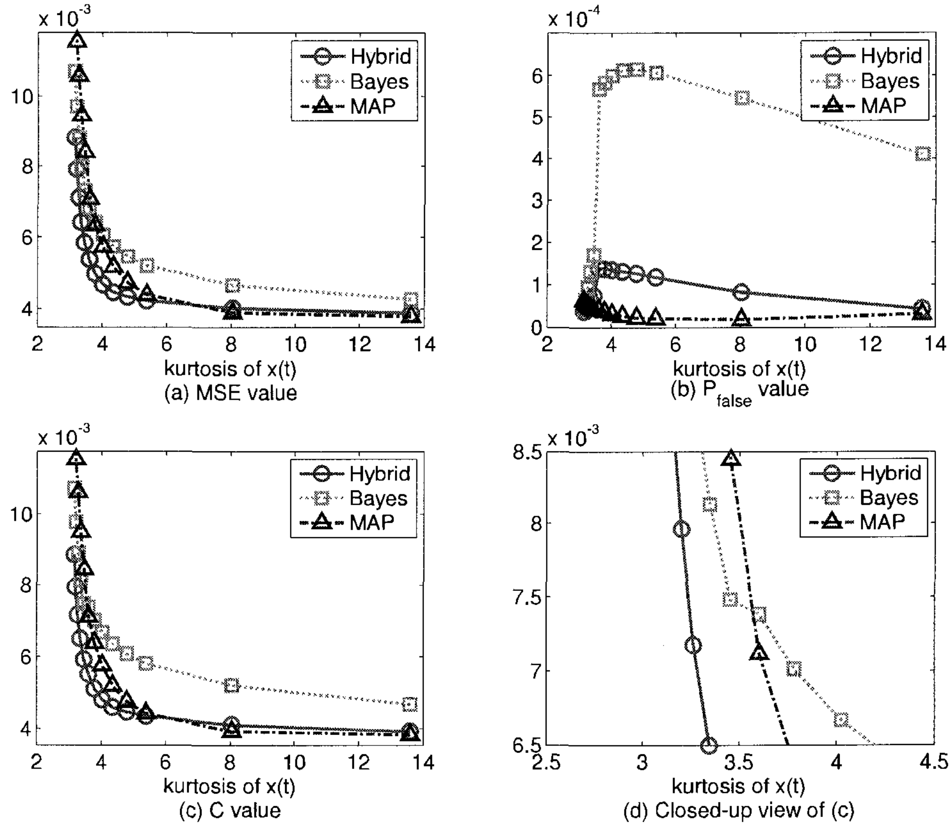


Figure 4.11. Comparison of K-hybrid, BayesShrink and MAP de-noising results in terms of MSE, P_{false}, and C values

noise contamination and hence is the very region where de-noising is particularly needed. The MAP-rule apparently does not have the needed filtering power in this kurtosis range and would more likely to provide misleading results. This is also evidenced by comparing Fig 4.8d, Fig. 4.9d, and Fig. 4.10d: the MAP-thresholding outputs deteriorate as the noise intensity (reflected by σ_n) increases.

There is a small bump in Figs. 4.11 c and d in the BayesShrink plot when kurtosis value is at around 3.5. This is mainly due to the drastic change of P_{false} shown in Fig. 4.11b. Such a sudden change in P_{false} is caused by the smoothing effect of soft-thresholding rule and large threshold values. Since the threshold is determined by $T_{bayes} = \sigma_n^2/\sigma_s$ (Chang et al., 2000a), threshold value increases drastically as the noise intensity increases. This leads to a drastic change of P_{false} at the kurtosis value around 3.5 as shown in Fig. 4.11b, which is mainly caused by the inaccurate estimate of σ_n. All noise components are shrunken at that moment and do not show significant magnitude afterwards in the lower kurtosis region.

K-hybrid leads to the lowest C value when the kurtosis value is below six (Fig. 4.11c). The C value obtained by K-hybrid is up to 18% and 30% lower (when kurtosis = 3.15) than the ones yielded by the BayesShrink and MAP-thresholding methods, respectively. Hence, the proposed approach is the best of the three in dealing with the signals that are heavily polluted by noise. Since de-noising is necessary especially in the region of lower kurtosis values, the ability to reveal the deeply hidden impulses makes the proposed K-hybrid more attractive for purifying *weak* mechanical fault signals. Moreover, K-hybrid outperforms BayesShrink in the entire simulated kurtosis range (3.15 to 13.59) in terms of C value.

As illustrated above, K-hybrid demonstrates better performance than the typical soft- and hard-thresholding rules with universal threshold. It also outmatches the two more sophisticated approaches, BayesShrink and MAP-thresholding methods in the low kurtosis values ($3 \leq K \leq 6$), representing signals with heavy noise components. This is a very important advantage because the mechanical signals are often contaminated by high intensity noise components. The probability density functions of such contaminated mechanical signals are similar to the Gaussian distribution with a kurtosis value of 3.0. In addition, the sensitivity of kurtosis to the outlier makes K-hybrid more suitable for detecting impulses generated by mechanical faults.

4.4 Application to Mechanical Fault Detection

4.4.1 Bearing Fault

The previous tests have clearly shown the advantages of the proposed K-hybrid over soft- and hard-thresholding rules, as well as the BayesShrink and the MAP-thresholding methods in identification of hidden impulses. In this section, the proposed method will be applied for bearing fault detection. The envelope analysis is a commonly used technique in bearing fault detection. To carry out envelop analysis, the resonance frequency of the bearing system is needed. In doing so, the repeated (for more reliable results) impact is often applied to excite the bearing system and hence to obtain its resonance frequency. The cut-off frequency of a bandpass filter is then determined accordingly. Envelope analysis includes the following steps:

Step 1. Apply the bandpass filter to the measured signal.

Step 2. Obtain the envelope signal.

Step 3. Obtain the envelope spectrum by FFT.

Tse et al. (2001) successfully used the above approach to identify the bearing fault frequencies. However, it may be difficult to obtain the resonance frequency of a bearing system in practice. In particular, bearings are often enclosed in a mechanical system and are not easily accessible for impact test. The inaccurate resonance frequency may lead to an inappropriate bandpass filter and hence to false identification of the fault frequency. This is a major hurdle of using envelope analysis.

For the above reason, this study applies K-hybrid as a pre-processing tool for bearing fault detection using envelope analysis. This, if successful, would avoid the use of resonance frequency in the envelope analysis and make it possible to detect faults for bearings that are in an enclosed environment. The signal processing steps are:

Step 1. De-noise the measured signal using K-hybrid.

Step 2. Extract the envelope signal from the purified signal obtained in Step 1.

Step 3. Identify the frequency spectrum of the envelope signal from Step 2 by FFT and then detect faults based on the FFT result.

The above procedure is applied to the measured bearing data from the Case Western Reserve University (2007). The bearing is located at the drive end of a shaft. The signals were monitored by accelerometers near the bearing at a sampling frequency of 12,000Hz. Three cases, i.e., healthy, inner-race fault, and outer-race fault, with original data plotted in the top figures of Figs. 4.12 and 4.13 are tested. The bearing is of type 6205-2RS JEM SKF (deep groove ball bearing). The geometrical information of bearing and characteristic frequencies is shown in Table 4.2. In Table 4.2, the bearing characteristic frequencies (BCFs), ball pass frequency of inner-race (BPFI) and ball pass frequency of outer-race (BPFO), are shown as functions of shaft rotational speed f_s in hertz (Hz). They are obtained using the following equations (Harris and Piersol, 2002):

$$BPFI(\text{Hz}) = \frac{n_e}{2} f_s \left(1 + \frac{BD}{PD} \cos\beta\right) \tag{4.12}$$

$$BPFO(\text{Hz}) = \frac{n_e}{2} f_s \left(1 - \frac{BD}{PD} \cos\beta\right) \tag{4.13}$$

where n_e is the number of rolling elements, BD is the ball diameter, PD is the pitch diameter, and β is the contact angle of the balls on the races.

The driving shaft rotating speed is 1,772RPM (≈29.53Hz). Therefore, the inner- and outer-race fault frequencies are 159Hz and 105Hz, respectively. The single point fault of 0.011 inches in depth and 0.007 inches in diameter is introduced to the bearing inner and outer races.

The healthy bearing shows no significant impulses in the time domain signal (Fig. 4.12a). Consequently, the wavelet-transformed signals yield kurtosis values around 3.0 over all the 16 selected scales. This results in a larger threshold which removes most input values. As expected, no impulses are observed in the de-noised signal (Fig. 4.12b). For the signals of inner and outer races with defects, the SNR or kurtosis values are relatively

Table 4.2. Bearing specification and characteristic frequencies (6205-2RS JEM SKF)

Inner Dia. (in)	0.9843	Fault depth (in)	0.0110
Outer Dia. (in)	2.0472	Fault Dia. (in)	0.0070
Thickness (in)	0.5906	BPFI (Hz)	$5.4152 f_s$
Pitch Dia. (in)	1.5370	BPFO (Hz)	$3.5848 f_s$
Shaft load (HP)	1		

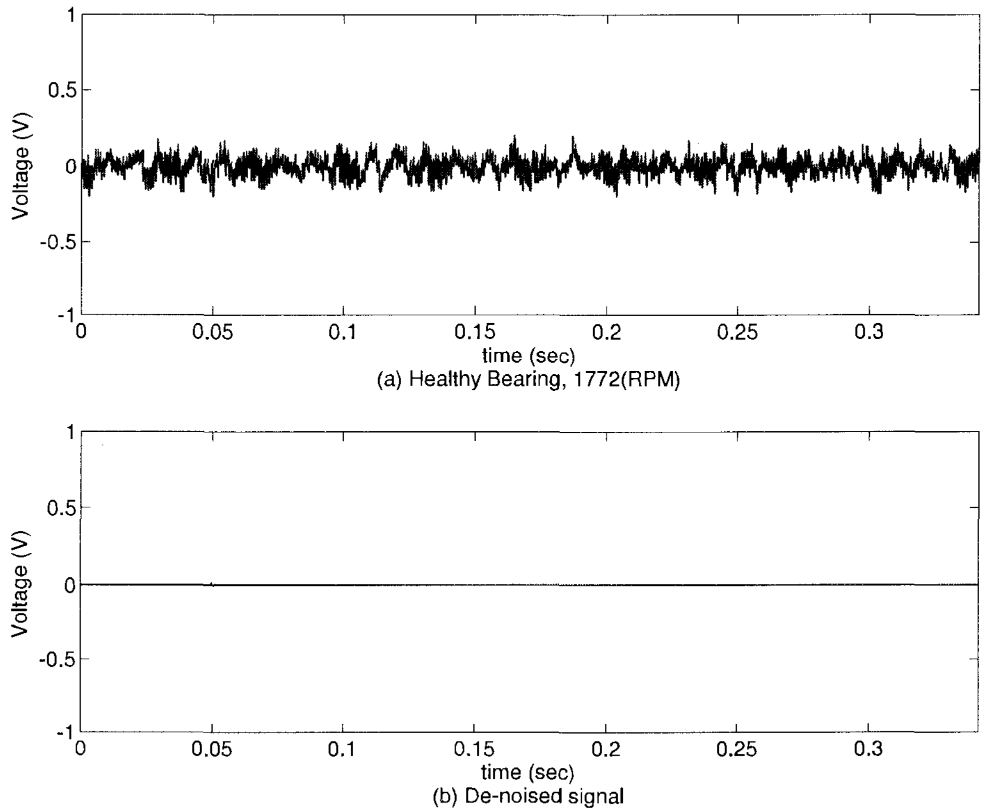


Figure 4.12. Healthy bearing (SKF 6205-2RS)

high, which warrants the use of smaller threshold values. The time domain signals are shown in the top figures (Fig. 4.13a and Fig. 4.13b) and are de-noised using K-hybrid. The de-noised results are displayed in the middle figures (Fig. 4.13c and Fig. 4.13d). To identify the frequencies of the faults, the envelope signals are obtained from the de-noised signals and the envelope spectra shown in the bottom figures (Fig. 4.13e and Fig. 4.13f) are obtained by Fourier transform. The BCFs can be clearly seen from the envelope spectra without being clouded by other peaks around them and the faults can be detected accordingly: 159Hz for inner-race fault and 105Hz for outer-race fault.

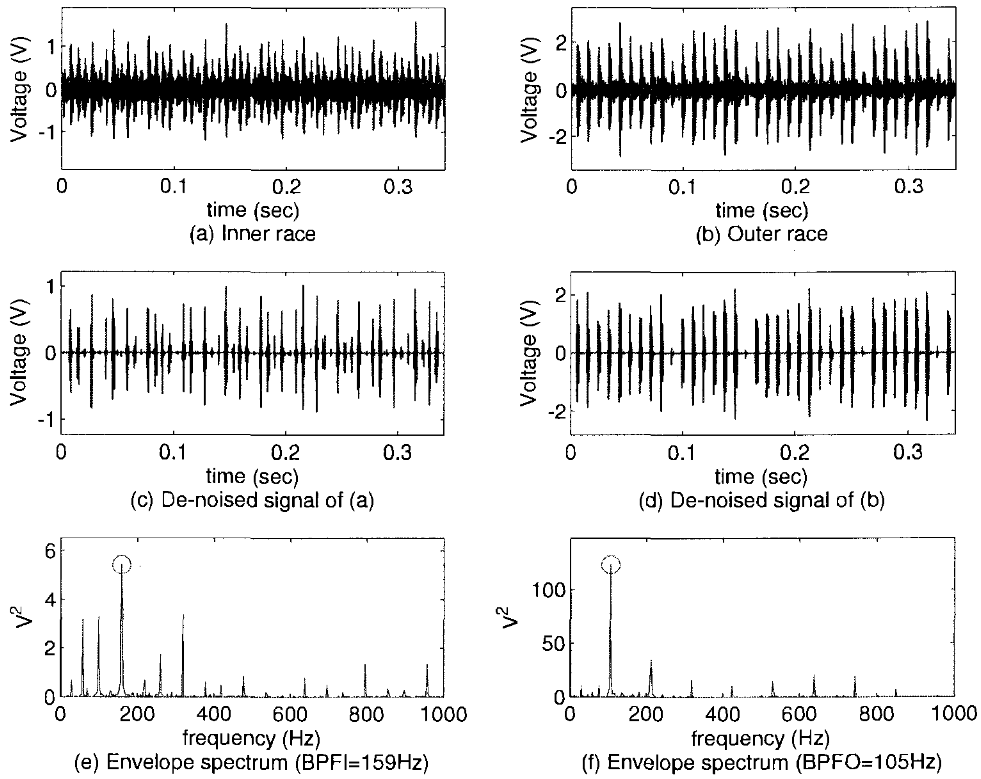


Figure 4.13. De-noising results of Inner- and outer-race faults (fault size: 0.007 inches in diameter and 0.011 inches in depth, BPFI=159Hz, BPFO=105Hz)

4.4.2 Gear Fault

The proposed method is applied to the gear fault signal. The test setup of gearbox simulator was same as shown in Fig. 3.10 (Case 1: Sensor is on the gearbox.). The faulty gear signal is acquired by accelerometer (Montronix VS100-100 with 100mV/g) attached on the gearbox case at the sampling frequency of 12,000Hz, motor shaft speed of 24Hz and the gearbox input shaft speed of about 9.4Hz. Two cases are considered: healthy gear and a gear with a missing tooth. The acquired acceleration data is sent to DAQ board (NI PCI-6040E with 12bits) for signal processing. The proposed de-noising technique is applied to the measured vibration signals and their results are shown in Fig. 4.14 and Fig. 4.15 for the healthy gear and the faulty gear, respectively. As expected,

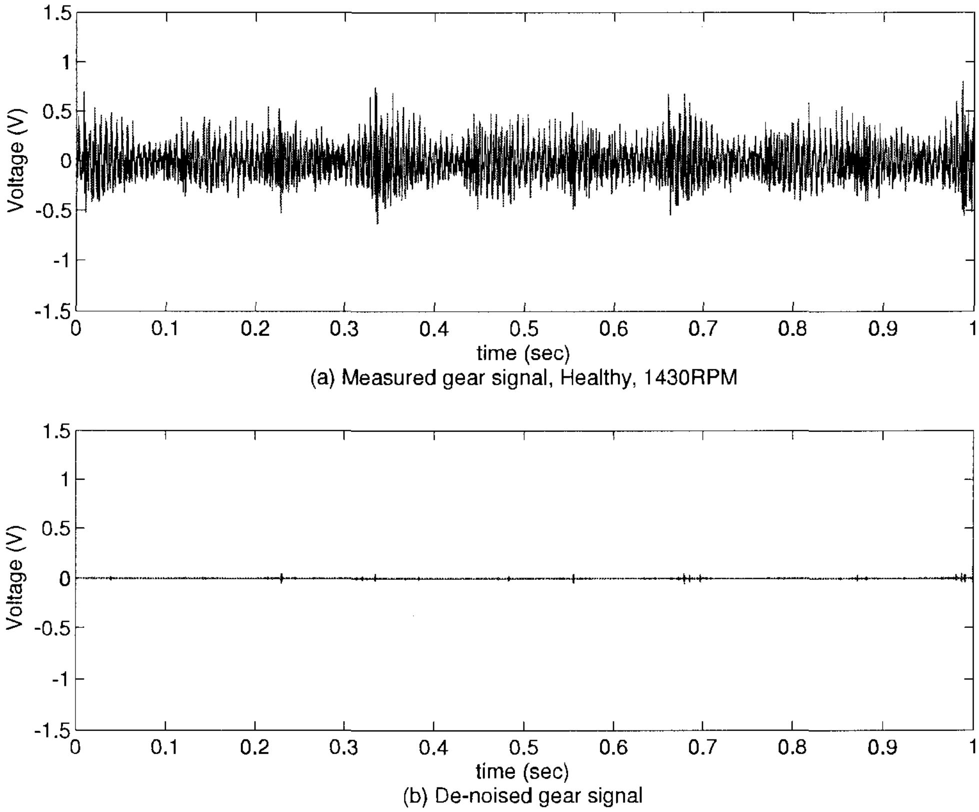
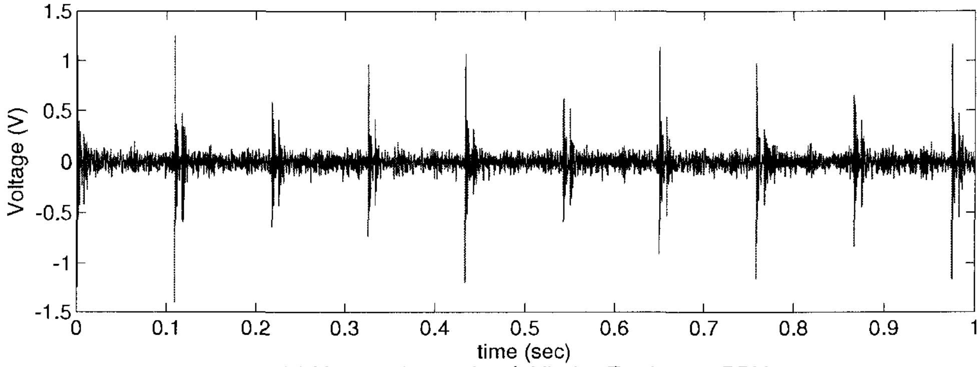


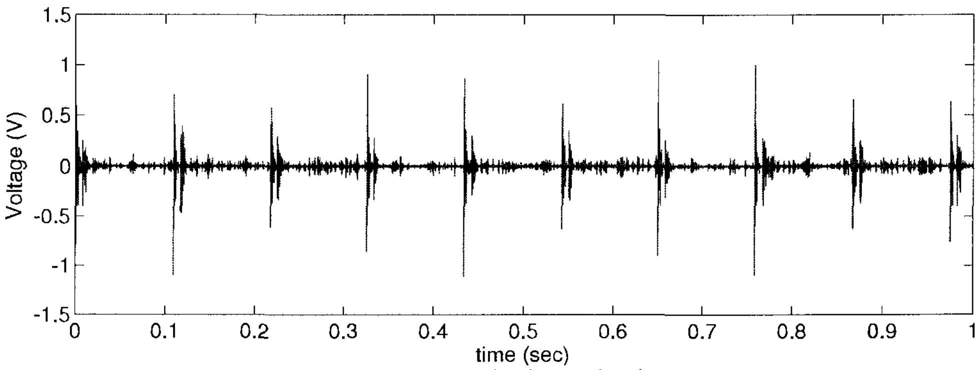
Figure 4.14. De-noising result of health gear signal

the proposed de-noising method removes noise and the de-noised result practically shows no impulses for the healthy gear. For the gear with a missing tooth, the noise is trimmed down substantially whereas the impulsive fault signature is well preserved.

To evaluate the de-noising performance of the proposed method, another test is performed for the same gear with missing tooth as used in Fig. 4.15; the sensor is located far away from the gearbox. The experimental setup is shown in Fig. 4.16 (Case 2: Sensor is on the test bed.). As shown in the figure, the sensor is installed on the baseboard of the test bed, which results in the low SNR (or heavy noise) of the measured signal. The motor speed is 1,510RPM and sampling frequency is 12,000Hz. All other conditions are the same as the ones in Fig. 4.15. Both the measured and de-noised signals are shown



(a) Measured gear signal, Missing Tooth, 1430RPM



(b) De-noised gear signal

Figure 4.15. De-noising result of gear with a missing tooth (Case 1)

in Fig. 4.17. The upper figure (Fig. 4.17a) represents the low SNR signal and hence the impulsive fault feature is not as distinguishable as in Fig. 4.15. The proposed de-noising method removes and extracts the impulsive fault feature as shown in Fig. 4.17b. This result further demonstrates the excellent performance of the proposed method.

4.5 Discussion

The proposed K-hybrid method has yielded reasonably good results for both simulated and measured dada. The examples shown in the previous sections are for the periodic signal under stationary operating conditions. In this section, the possibility of de-noising non-stationary signal is explored through simulated non-stationary signal. The input

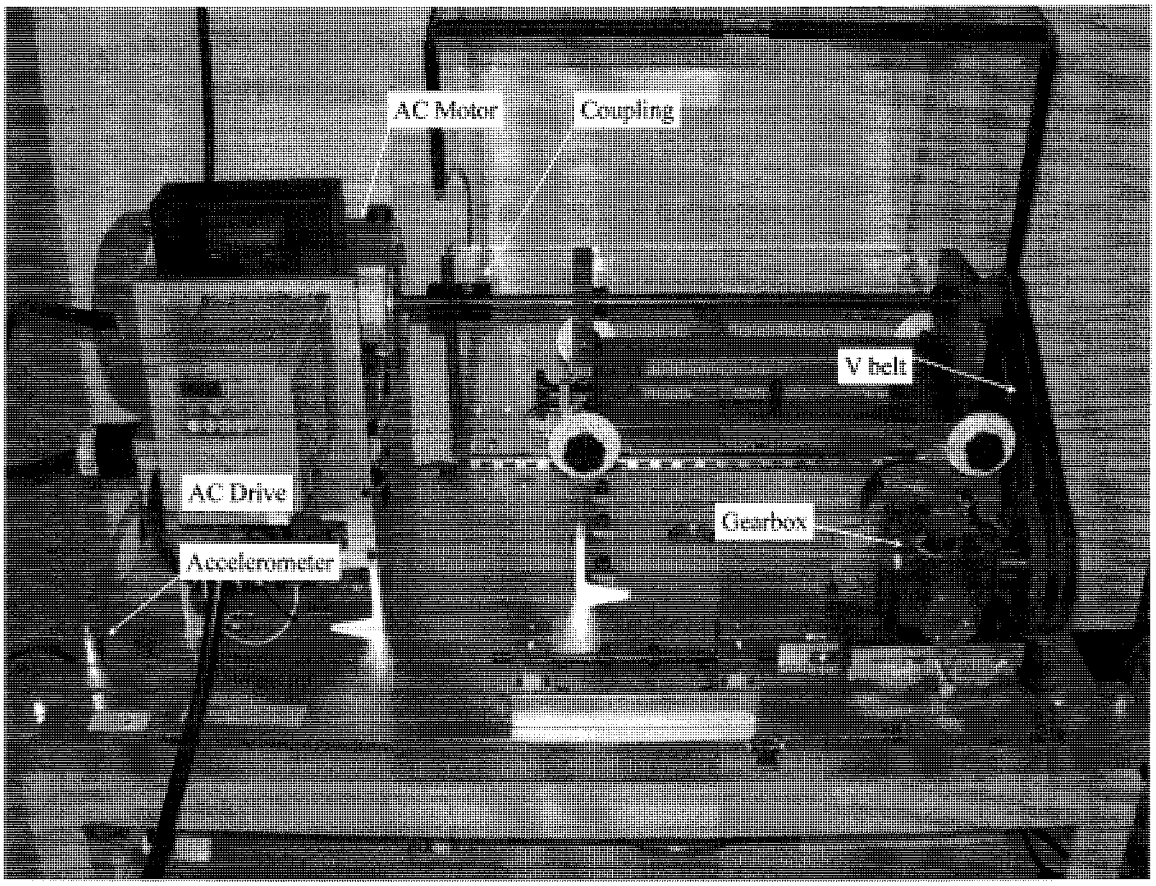


Figure 4.16. Experimental setup for Case 2

signal with eight impulses is created as follows:

$$x(t) = \sum_{i=1}^{8} \left\{ \exp\left(-\frac{(t-p_i)^2}{100}\right) \cdot \cos\left(\frac{\pi t}{10 q_i}\right) \right\} \tag{4.14}$$

where p_i and q_i are randomly generated from the uniform distribution ranging [0, 4000] and [0, 1], respectively. As the amplitudes and time gaps between adjacent impulses are randomly generated, the signal is non-stationary (Fig. 4.18b). It also displays a super-Gaussian distribution pattern (Fig. 4.18a). After the Gaussian noises of different strength are added to the simulated impulses (Fig. 4.18c and Fig. 4.18e), the K-hybrid de-noising method is applied. The de-noising results are shown in Fig. 4.18d and Fig. 4.18f with very low C values (0.0067 and 0.0061 respectively). In particular,

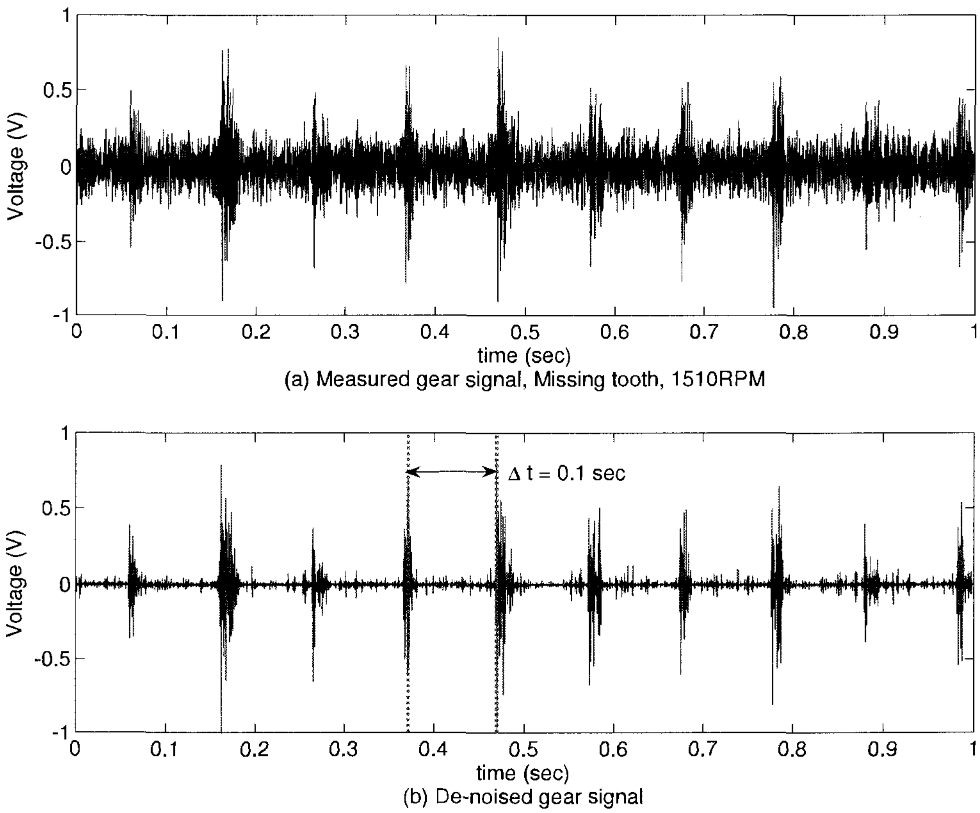


Figure 4.17. De-noising result of missing tooth gear signal (Case 2)

as illustrated in Fig. 4.18f, the proposed method can extract the super-Gaussian signal from heavily contaminated signal mixture. The example shows that K-hybrid could potentially be applied to non-stationary signals. However, more systematic investigations of its behavior in de-noising non-stationary signals are required. It should be noted that Hadjileontiadis and Rekanos (2003) have recently proposed an iterative kurtosis-based method using time-domain sliding window to enhance the explosive bowel sound (EBS), a transient and non-stationary signal with super-Gaussian distribution. This method has shown to be quite robust in extracting EBS impulses from the background noise without losing structural integrity of signal.

In addition, it should be noted that the computation time is critical for many on-line applications. In this study, all the computations are carried out using MATLAB on a

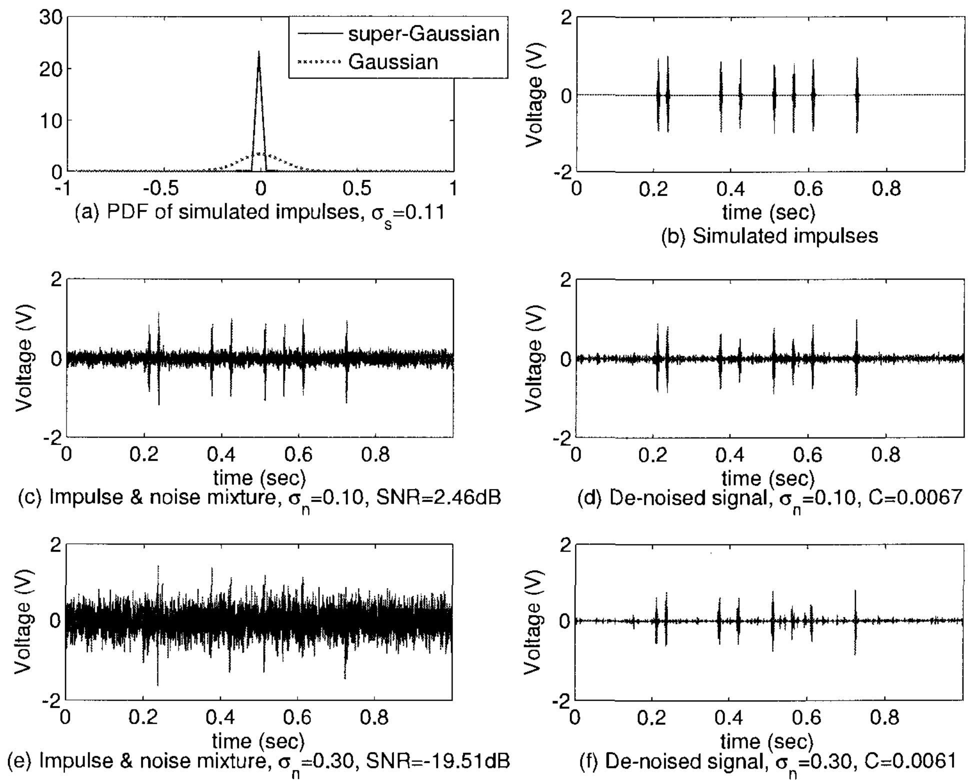


Figure 4.18. De-noising result of non-stationary signal. (Note: In (a), solid line stands for PDF of simulated impulses with super-Gaussian behavior, dotted line for PDF of Gaussian distribution with the same standard deviation, i.e., 0.12)

Pentium 4 PC with 3.4GHz CPU and 1GB RAM. The average computing times for the entire process including CWT, de-noising and reconstruction of the signal with 4,000 data points in all conditions listed in Table 4.1, are 0.4447, 0.3654, and 0.3546 seconds for K-hybrid, hard, and soft thresholding rules, respectively. Similarly, for the simulated signals shown in Fig. 4.11, the average computing times are 0.4447, 0.3714 and 0.3846 seconds for K-hybrid, BayesShrink and MAP, respectively. Such a computing speed should be sufficient for many on-line applications.

Chapter 5

Extracting In-line ODM (Oil Debris Monitoring) Sensor Signature Using Generalized Differentiation[1]

A particle signature is similar to a single period sine wave featuring a positive and a negative peak. The existing method for capturing these signatures is based on threshold (Miller and Kitaljevich, 2000), which has certain drawbacks (to be elaborated later). To avoid the difficulties present in the thresholding method, this thesis proposes that the particle is extracted by capturing the sine-like waveform and then be confirmed by identifying the three inflexion points in each particle signature. Based on this idea, a new feature extraction method is proposed for in-line oil debris monitoring (ODM) sensor signal analysis. This new method consists of two detectors developed using generalized differentiation. The first detector is used to extract the waveform similar to the particle signature. However, if the signature signal is weak, the resulting signature obtained by this detector alone cannot definitely be classified as that of a particle. Hence, the second detector is used to detect the existence of the three inflexion points. The passage of a particle can be confirmed only when all the three inflexion points exist in the waveform extracted by the first detector. Otherwise, the existence of a particle is excluded. As the proposed method does not require any threshold, it can avoid the difficulty in specifying

[1]See Appendix B.5 for MATLAB script.

suitable threshold values for various applications. Both simulated and experimental results have shown that the proposed method is robust in a noisy environment and can effectively detect small particles, a sign of early flaws. It should also be noted that this method does not require any additional filtering or de-noising steps and therefore signal processing can be simplified considerably.

5.1 Introduction to ODM (oil debris monitoring) sensor

In-line ODM sensor measures the disturbances in electromagnetic field created by particles, either ferromagnetic or non-ferromagnetic material (Miller and Kitaljevich, 2000), passing through the sensor. Since an ODM sensor does not trap metallic debris, it does not require regular inspection or cleaning to prevent false alarm. The ODM sensor consists of three coils surrounding a magnetically inert section of tubing of oil passage, which creates unique particle signature whenever metallic debris passes through the sensing region determined by the three coils (Howe and Muir, 1998). A schematic diagram of the ODM sensor is shown in Fig. 5.1. The amplitude and the phase of the particle signature are dependent on the size of oil debris and the material type. The amplitude of the signal is proportional to the mass of a particle for ferromagnetic material and to the surface area for non-ferromagnetic material. The phase of ferromagnetic material is opposite to that of non-ferromagnetic material, allowing distinction between two mate-

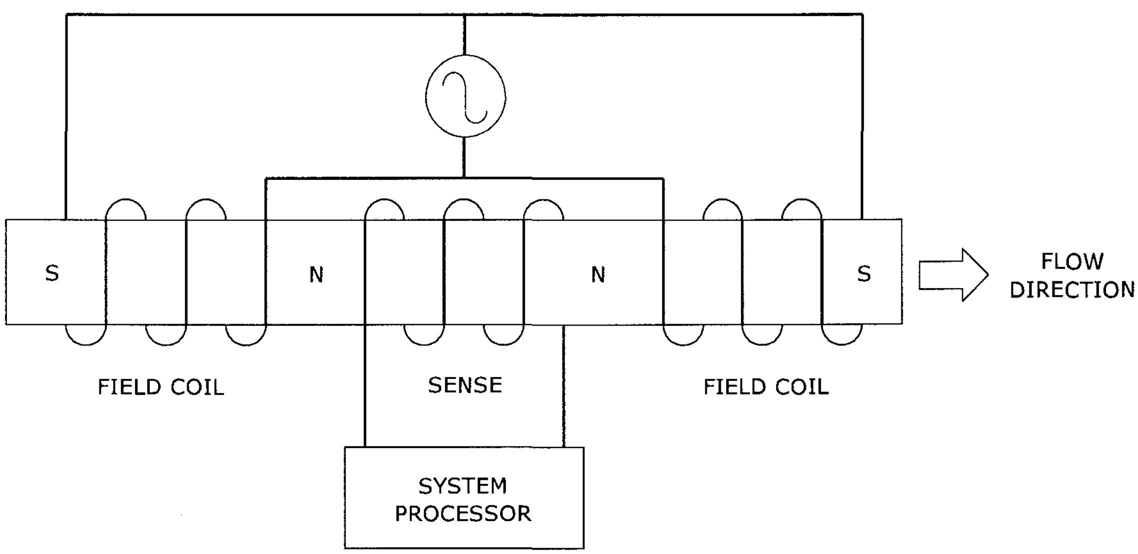


Figure 5.1. Schematic diagram of in-line ODM sensor (Howe and Muir, 1998)

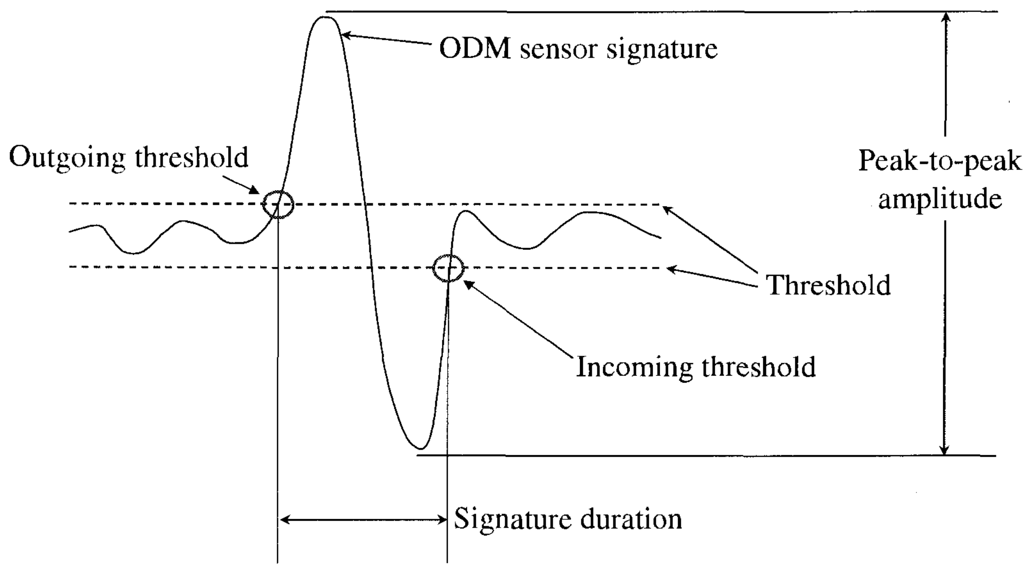


Figure 5.2. Illustration of the ODM sensor signal detection in time domain (modified from Miller and Kitaljevich (2000))

rials. The minimum size of a particle detectable by the ODM sensor is dependent on the sensor bore size. For example, the 1/2-inch ODM sensor is able to detect spherical ferromagnetic particles of 124 microns in size (Miller and Kitaljevich, 2000).

Three parameters — the number of particles, the approximate size, and the accumulated mass — are measured from the particle signature extracted from the noisy signal by using the threshold-based detection method as shown in Fig. 5.2. As mentioned earlier, deciding a threshold value is not always an easy task because the measured signal is generally contaminated by noise. Therefore, a threshold-based algorithm may not be able to detect small particle signatures correctly and hence early fault detection is unlikely to be reliable.

5.2 The proposed method

A simplified ideal particle signal is illustrated in Fig. 5.3a. As shown in the figure, the oil debris signature shows three inflexion points at t_1, t_2, and t_3 respectively, which together with the sine-like waveform can be used to confirm the existence of a particle. The first and third inflexion points in the same phase are the counterparts of outgoing and incoming threshold values shown in a threshold-based method in Fig. 5.2. Though the three inflexion points can be located by the three associated peak points appeared in the plot of the first derivative of the ideal waveform as shown in Fig. 5.3b, the first derivative is too sensitive to noise and cannot be used for this purpose in an industrial setting. This can

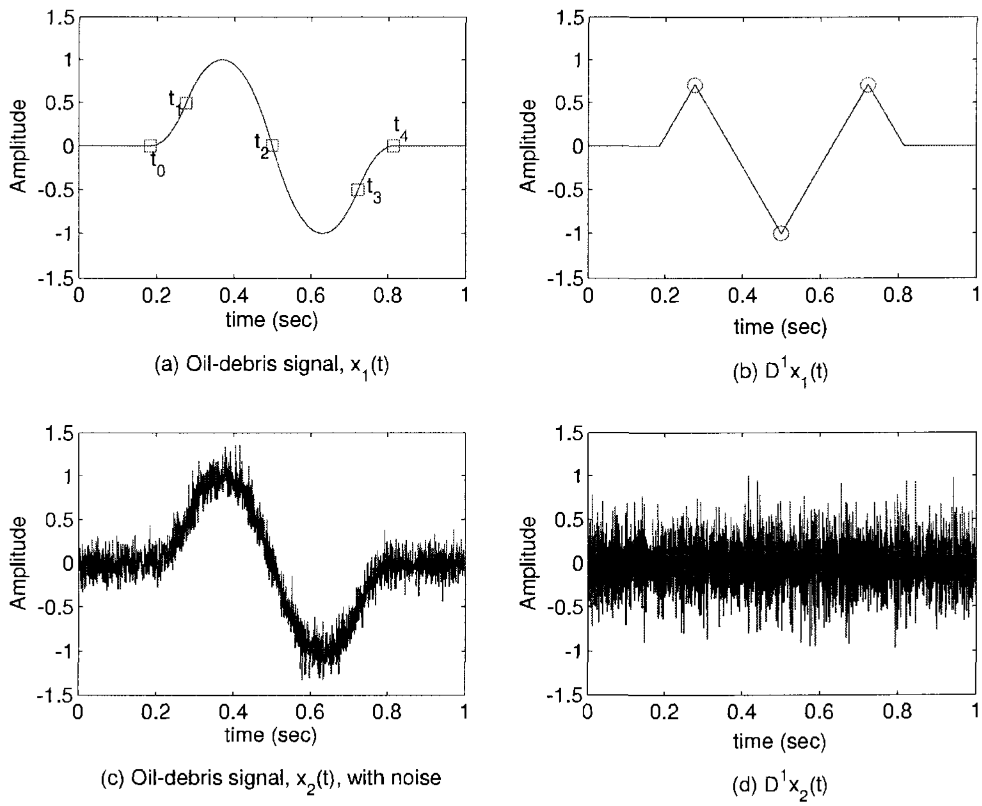


Figure 5.3. Illustration of the first-order differentiation of oil-debris signature

be illustrated by Fig. 5.3c and Fig. 5.3d. Fig. 5.3c shows a more realistic particle signal obtained by adding some Gaussian noise to the signal in Fig. 5.3a. The first derivative of this noisy signal is plotted in Fig. 5.3d, which obviously cannot reveal any inflexion points. The same can be said for other higher integer order derivatives. Therefore, it is impossible to employ integer order differentiation for ODM signal analysis.

Generalized differentiation (including both non-integer order differentiation and integration) may overcome the above problem and will be exploited in the following. The most commonly adopted definition of generalized differentiation is an integral transform called the Riemann-Liouville integral (Oldham and Spanier, 2006) given as

$$\left[\frac{d^n f}{[d(t-a_0)]^n}\right]_{R-L} = \frac{1}{\Gamma(-n)}\int_{a_0}^{t} [t-y]^{-n-1} f(y)\, dy \tag{5.1}$$

where the symbol $[\ ]_{R-L}$ designates the Riemann-Liouville fractional integral, n is the non-integer order and the function f is differentiable in $a_0 \leq y \leq t$. In the above equation, a positive non-integer n implies fractional differentiation, a negative non-integer n indicates fractional integration, and a zero n value represents identity transformation. More details can be found in Ortigueira (2006), Kleinz and Osler (2000), Miller (1995), Doyle (1995), Lavoie et al. (1976), and Post (1930).

To obtain analytical expressions of the generalized differentiation, the oil-debris signature shown in Fig. 5.3a can be approximated by (the power function is used as an example):

$$\begin{aligned}
x(t) &= 0 && ,\ t \leq t_0 \\
x(t) &= at^2 && ,\ t_0 \leq t \leq t_1 \\
x(t) &= -at^2 + 4at_0t - 2at_0^2 && ,\ t_1 \leq t \leq t_2 \\
x(t) &= a\,(t-t_3)^2 - 2at_0^2 && ,\ t_2 \leq t \leq t_3 \\
x(t) &= -a\,(t-t_4)^2 && ,\ t_3 \leq t \leq t_4 \\
x(t) &= 0 && ,\ t_4 \leq t
\end{aligned} \tag{5.2}$$

where a is a constant.

The generalized differentiation of Eq. (5.2) can be obtained using the Riemann-Liouville definition. However, the definition given by Eq. (5.1) limits its application to causal function determined by the lower limit a_0 and, as Oustaloup et al. (1993) pointed out, the application of the definition equation is impractical in general because of the integral form. For a power function, its generalized differentiation can be expressed as a gamma function (Oldham and Spanier, 2006):

$$\frac{d^n[t-a_0]^p}{[d(t-a_0)]^n} = \frac{\Gamma(p+1)[t-a_0]^{p-n}}{\Gamma(p-n+1)} \tag{5.3}$$

According to Eq. (5.3), the generalized differentiation of Eq. (5.2) is:

$$D^n x(t) = \begin{cases} 0 & , \ t \le t_0 \\ \dfrac{2a}{\Gamma(3-n)} t^{2-n} & , \ t_0 \le t \le t_1 \\ \dfrac{2a}{\Gamma(3-n)} [-2(t-t_1)^{2-n} + t^{2-n}] & , \ t_1 \le t \le t_2 \\ \dfrac{2a}{\Gamma(3-n)} [2(t-t_2)^{2-n} - 2(t-t_1)^{2-n} + t^{2-n}] & , \ t_2 \le t \le t_3 \\ \dfrac{2a}{\Gamma(3-n)} [-2(t-t_4)^{2-n} + 2(t-t_2)^{2-n} - 2(t-t_1)^{2-n} + t^{2-n}] & , \ t_3 \le t \le t_4 \\ 0 & , \ t_4 \le t \end{cases} \tag{5.4}$$

Although the generalized differentiation of analytical functions can be performed using the Riemann-Liouville integral, it is impractical to use it in signal processing such as oil-debris signal analysis because the analytical expression of the data mixture of signal and noise cannot be obtained and cannot be easily approximated using Eq. (5.2). Consequently, the analytical calculus expressed by Eq. (5.4) can no longer be applied. Therefore, a discretized form of generalized differentiation is required to deal with discrete data measured from a sensor.

A discretized form of generalized differentiation can be obtained by generalizing the ordinary integer-order derivation (Mathieu et al., 2003):

$$D_R^1 x(t) = \frac{x(t) - x(t-h)}{h} \qquad (h \to 0). \tag{5.5}$$

By introducing shift operator $q^{-1}x(t) = x(t-h)$,

$$D_R^1 x(t) = \frac{1-q^1}{h} x(t) \tag{5.6}$$

the associated operator for any order n is then (See Appendix A.3 for derivation.),

$$D_R^n = \left(\frac{1-q^{-1}}{h}\right)^n. \tag{5.7}$$

Expanding $(1-q^{-1})^n$ based on Newton's binomial formula, one obtains

$$D_R^n = \frac{1}{h^n}\sum_{k=0}^{\infty}(-1)^k \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} q^{-k}. \tag{5.8}$$

Therefore, the generalized differentiation of $x(t)$ is (See Appendix A.4 for proofs of Eqs. (5.8) and (5.9).),

$$\begin{aligned} D_R^n x(t) &= \frac{1}{h^n}\sum_{k=0}^{\infty}(-1)^k \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} x(t-kh) \\ &= \frac{1}{h^n}\sum_{k=0}^{\infty}(-1)^k \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} x(t-kh). \end{aligned} \tag{5.9}$$

Similarly, for the definition of

$$D_L^1 x(t) = \frac{x(t)-x(t+h)}{h} \qquad (h \to 0) \tag{5.10}$$

the generalized differentiation of $x(t)$ is calculated by,

$$\begin{aligned} D_L^n x(t) &= \frac{1}{h^n}\sum_{k=0}^{\infty}(-1)^k \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} x(t+kh) \\ &= \frac{1}{h^n}\sum_{k=0}^{\infty}(-1)^k \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} x(t+kh). \end{aligned} \tag{5.11}$$

The following subsections examine the application potential of positive and negative fractional n generalized differentiation based on the simulated oil debris signal shown in Fig. 5.3a and Fig. 5.3c.

5.2.1 Fractional differentiation with positive order n

A fractional differentiation is obtained for the positive non-integer order $n > 0$. Though a fractional differentiation reveals inflexion points, the accuracy is dependent on n value;

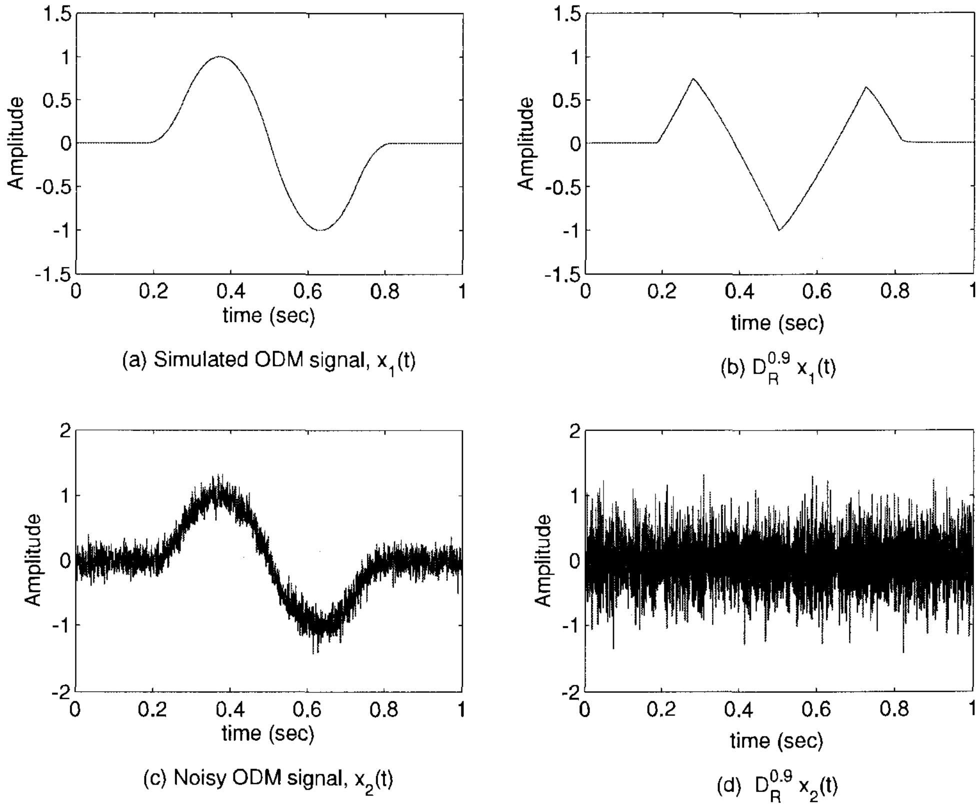
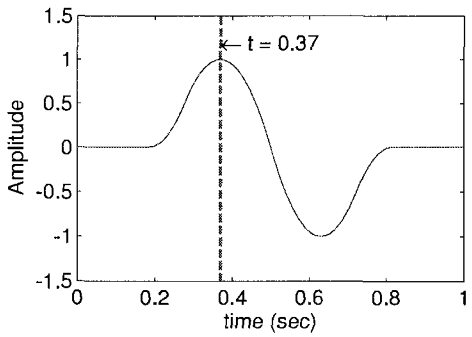
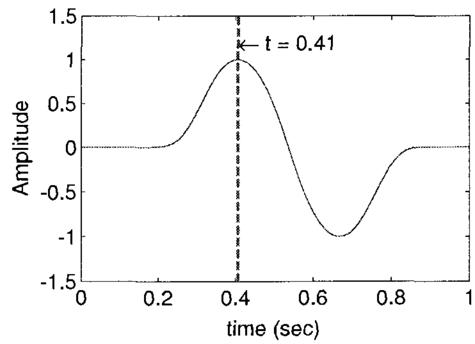


Figure 5.4. Simplified oil debris signatures and their fractional differentiations: (a, b) without noise, (c, d) with noise

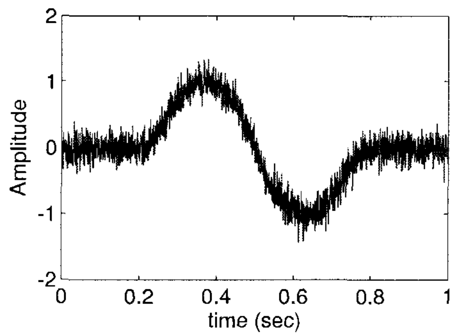
the closer to 2.0, the more accurate inflexion location can be detected (Mathieu et al., 2003). Fig. 5.4 shows fractional differentiations (Fig. 5.4b and Fig. 5.4d) with a positive non-integer order, $n = 0.9$ is used for illustration, for a simulated clean signal (Fig. 5.4a) and a mixture of the signal with noise (Fig. 5.4c), respectively. The three peaks in Fig. 5.4b indicate the approximate locations of the three inflexion points in Fig. 5.4a. However, when the simulated signal is mixed with noise (Fig. 5.4c), inflexion points cannot be detected by the fractional differentiation as shown in Fig. 5.4d. This indicates that a fractional differentiation (with positive n) is sensitive to noise and is not suitable for extracting particle signatures.



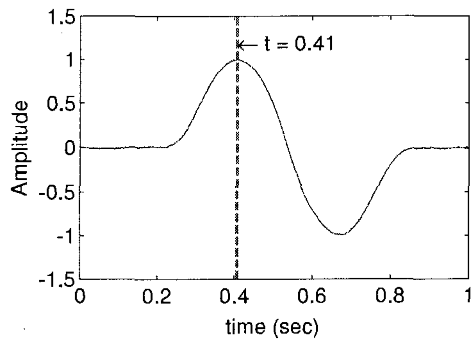
(a) Simulated ODM signal, $x_1(t)$



(b) $D_R^{-0.9}\ x_1(t)$



(c) Noisy ODM signal, $x_2(t)$



(d) $D_R^{-0.9}\ x_2(t)$

Figure 5.5. Simplified oil debris signatures and their fractional integrations: (a, b) without noise, (c, d) with noise

5.2.2 Fractional integration with negative order n

As shown in the previous section, a fractional differentiation is not suitable for noisy particle signature. Then the fractional integration ($n < 0$) is considered. As illustrated in Fig. 5.5, the fractional integration is much less susceptible to noise. Fig. 5.5b and Fig. 5.5d show the fractional integrations ($n = -0.9$) of the same simulated and contaminated signals described in Fig. 5.4a and Fig. 5.4c, respectively. Interestingly, a fractional integration has demonstrated a very good de-noising effect as shown in Fig. 5.5d. The particle signature can be clearly revealed by fractional integration (with negative n).

However, the inflexion points cannot be located by using the above simple fractional integration. Furthermore, comparing Fig. 5.5a and Fig. 5.5b, as well as Fig. 5.5c and Fig. 5.5d, indicates that the locations of the signature signals processed by fractional integration have been right-shifted.

In view of the effect and drawbacks of the simple fractional integration, two new detectors will be developed based on fractional integration to: a) extract the oil-debris signature without location shifting, and b) locate the inflexion points. Before developing the two detectors, it is important to further examine the robustness of fractional integration in dealing with noisy signals.

5.2.3 Robustness to noise

Generalized differentiation with negative non-integer order has de-noising effect as shown in Fig. 5.5. This de-noising effect is due to the low-pass filter property of the generalized differentiation with negative non-integer order. Such a low-pass filter characteristic can be shown by taking Laplace transform of the fractional derivative of a unit step function as follows. The generalized differentiation of a unit step function can be expressed as (Oldham and Spanier, 2006):

$$
\begin{aligned}
D_R^n u(t) &= \frac{t^{-n}}{\Gamma(1-n)} \\
u(t) &= \begin{cases} 1 & , \ t \geq 0 \\ 0 & , \ \text{elsewhere} \end{cases}
\end{aligned}
\tag{5.12}
$$

where $u(t)$ is a unit step function. The Laplace transform of the above equation is (See Appendix A.5 for derivation.)

$$
\begin{aligned}
L\left\{D_R^n u(t)\right\} &= L\left\{\frac{t^{-n}}{\Gamma(1-n)}\right\} \\
&= \frac{1}{s^{-n+1}} \qquad (n < 0)
\end{aligned}
\tag{5.13}
$$

As shown in Eq. (5.13), the Laplace transform of the generalized differentiation of unit step function has the low-pass filter characteristic $(1/s^{-n+1})$ when a negative non-integer order n is selected $(n < 0)$. This property makes the generalized differentiation with negative non-integer order work as a low-pass filter and hence less susceptible at least to high frequency noise as shown in Fig. 5.5d.

5.2.4 Development of two detectors

Two discretized forms of generalized differentiation shown in Eqs. (5.9) and (5.11) can be applied to the data measured from a sensor without knowledge of analytical equations. However, there is a translation effect when Eqs. (5.9) and (5.11) (right and left translation, respectively) are directly applied to the measured data. For example, the fractional integral ($n < 0$) obtained by Eq. (5.9) shifts the signal towards right, which can be easily observed by comparing Fig. 5.5a and Fig. 5.5b and Fig. 5.5d. This is also illustrated in Fig. 5.6a. It should be noted that the time-domain translation is caused by the finite number of data. When the signal mixture is severely contaminated by noise or when the oil-particle signature is too small, it would be very difficult to distinguish the particle signature. In this case, locating three inflexion points of the extracted particle signature helps identify the oil particle signature. For this reason, maintaining the particle location, i.e., extracting signature without translation in time, is necessary. To extract particle signatures and detect the inflexion points without shifting signal locations, two detectors, one for extracting particle signatures and the other for locating the inflexion points, are developed as follows.

a. *Particle signature extractor*

This detector is expressed as

$$H^n x(t) = D_L^n x(t) + D_R^n x(t) \qquad (n < 0). \tag{5.14}$$

As shown in Fig. 5.6a, both D_R^n and D_L^n can reproduce the signature. However, applying each of them alone leads to either right or left shift. Adding the two together would cancel out the shifting effect and the correct location of the signal can be preserved. However, though $H^n x(t)$ can preserve the signature location, it also changes the magnitude of the signature as suggested by Eqs. (5.9), (5.11), and (5.14). For this reason, the particle signal magnitude, $\hat{s}(t)$, is approximately recovered by applying:

$$\hat{s}(t) \approx \hat{H}^n x(t) = X \cdot \tilde{H}^n x(t) \tag{5.15}$$

where $\tilde{H}^n x(t)$ is given by

$$\tilde{H}^n x(t) = \frac{H^n x(t)}{\max\left(|H^n x(t)|\right)} \tag{5.16}$$

and

$$X = \max\left(|x(t)|\right) \tag{5.17}$$

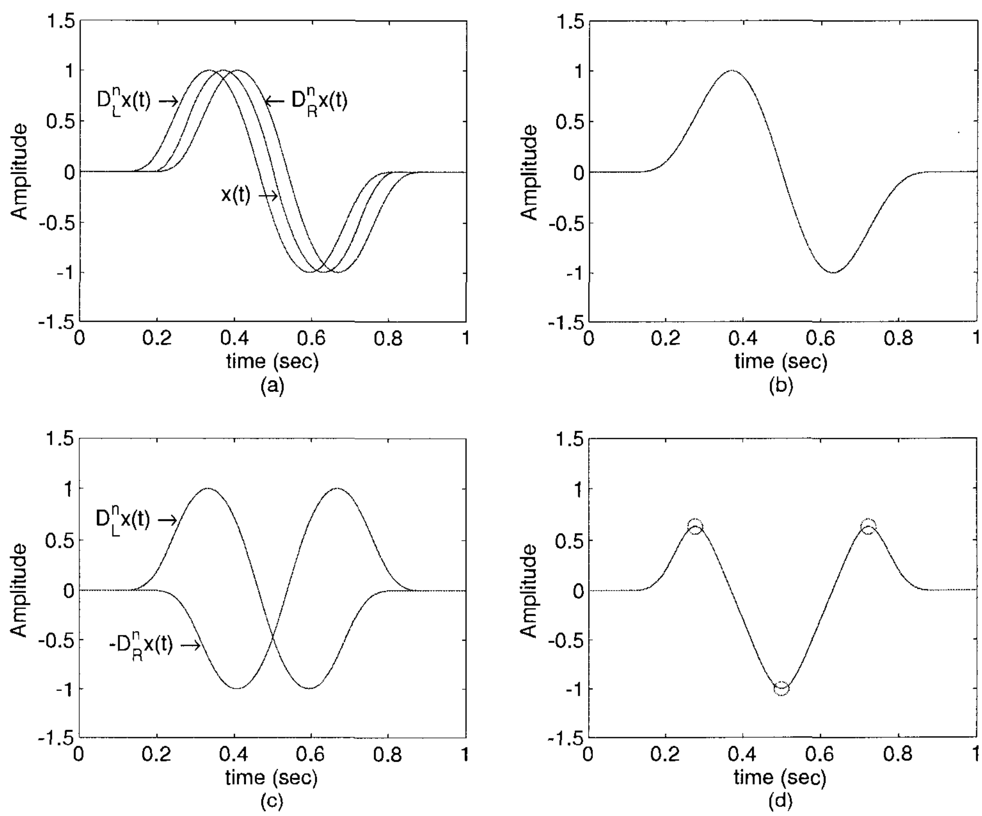


Figure 5.6. Illustration of (for $n = -0.9$): (a) simulated signal $x(t)$, and its D_R^n and D_L^n, (b) recovered signature ($\approx \hat{H}^n x(t)$) by the signature extractor, (c) applying D_L^n and $-D_R^n$ to the signal, and (d) the three inflexion points (marked with circle) located by the inflexion-point locator

b. *Inflexion-point locator*

Similarly, an inflexion-point detector can be defined as

$$I^n x(t) = D_L^n x(t) - D_R^n x(t) \qquad (n < 0). \tag{5.18}$$

Subtracting D_R^n from D_L^n can identify inflexion points as peaks in the plot of $I^n x(t)$ and preserve the signal phase information that is crucial to distinguish ferromagnetic and non-ferromagnetic materials (The mathematical verification of the proposed method is given in the Appendix A.6.). To illustrate, Eq. (5.18) is applied to the simulated signal $x(t)$ (Fig. 5.6a). The result clearly reveals the three inflexion points (associated with the three peak locations in Fig. 5.6d) with correct phase information. For general application purpose, the $I^n x(t)$ is normalized as follows

$$\tilde{I}^n x(t) = \frac{I^n x(t)}{\max\left(\left|I^n x(t)\right|\right)} \tag{5.19}$$

The overall procedure of applying the two detectors to extract the particle signature from a measured signal is shown in Fig. 5.7 and outlined as follows:

Step 1: Select a negative non-integer order n and find the maximum amplitude, X, of the measured signal from the measured signal $x(t)$. In this thesis, n is selected as "-0.9". Generally, for better de-noising result, the value of n should be as close to "-1" as possible, but not equal to "-1".

Step 2: Calculate $D_L^n x(t)$, $D_R^n x(t)$ and associated $H^n x(t)$ and $I^n x(t)$.

Step 3: Extract particle signature, $\hat{H}^n x(t)$, and locate the inflexion points, $\tilde{I}^n x(t)$, using the two detectors.

The two proposed detectors do not require any threshold values and are robust to noise. In addition, their "built-in" de-noising capability has streamlined oil-debris signal processing. In the following sections, the performance of the two detectors will be examined using simulated signal and data measured from an ODM sensor.

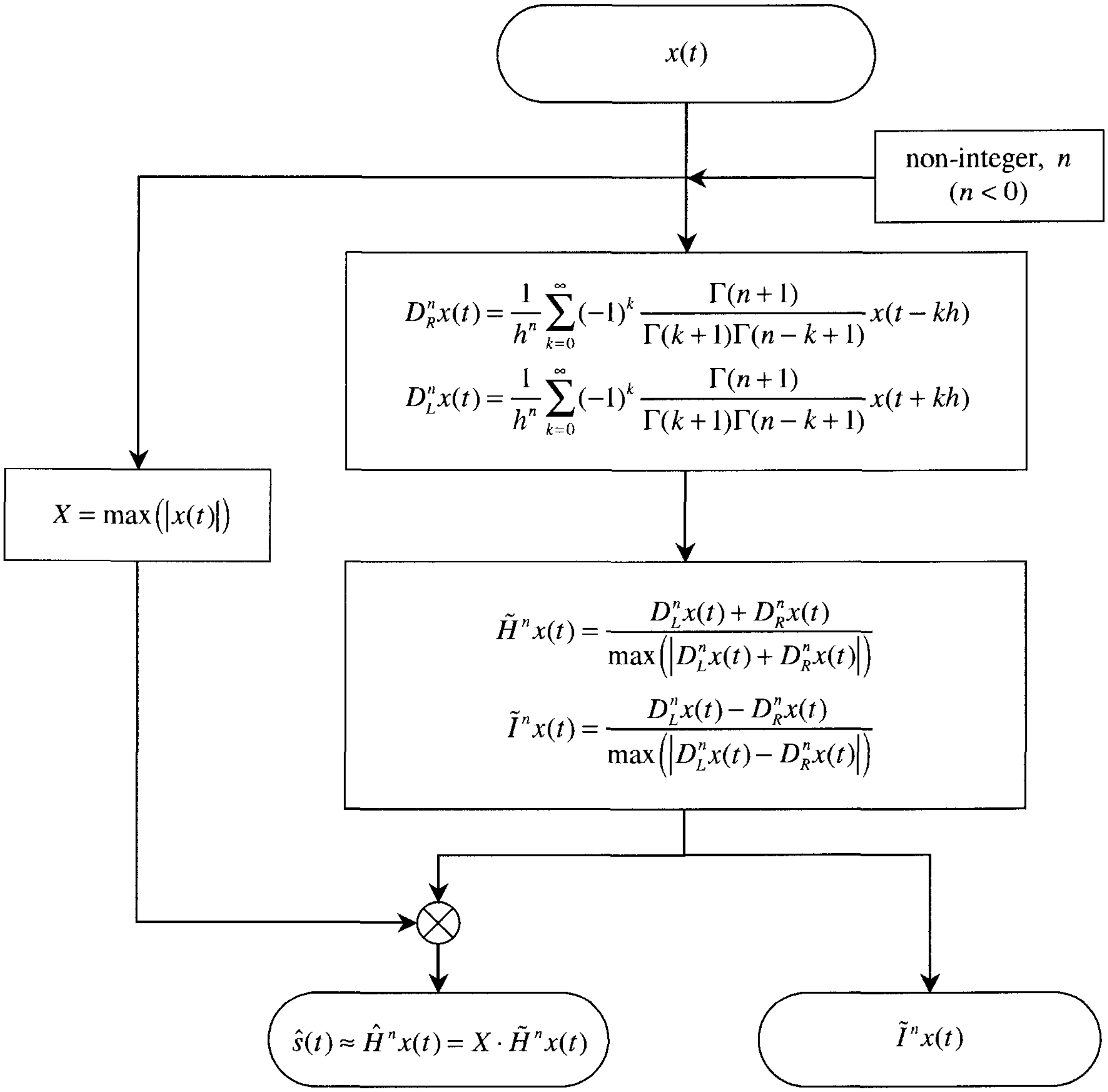
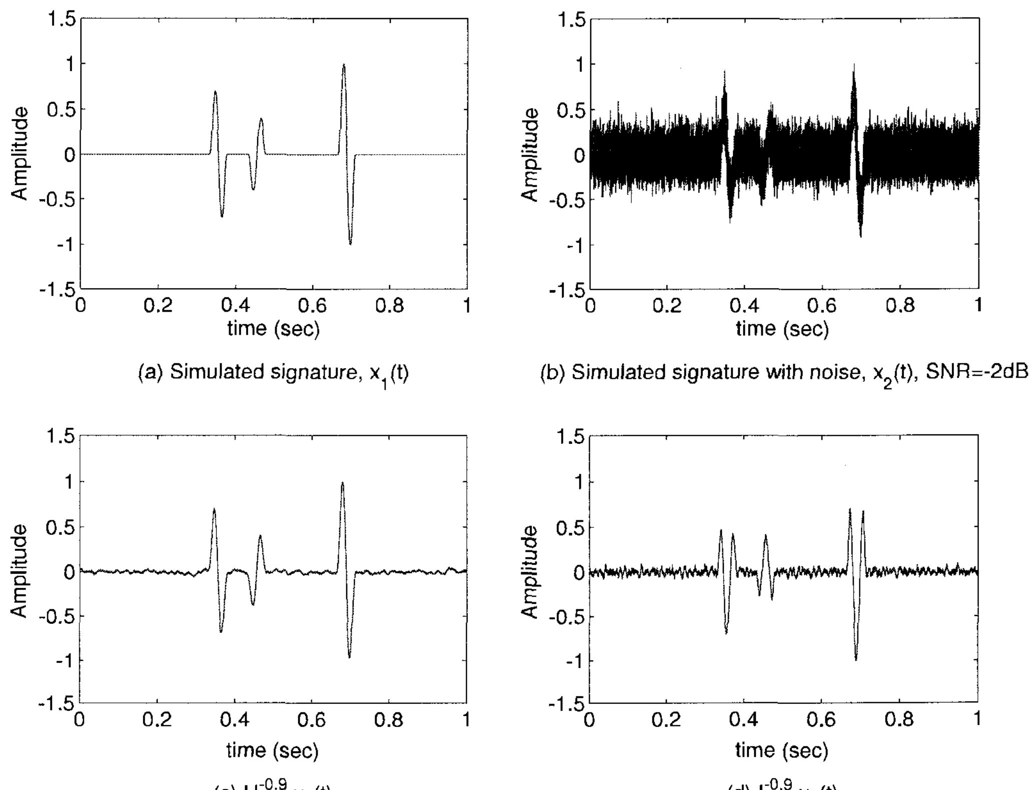


Figure 5.7. Block diagram of the proposed oil debris signature extraction method

5.3 Simulated examples

In this section, the two detectors defined in the previous section are evaluated using simulated particle signatures with different phase information and various levels of noise. A simulated signal consisting of three particle signatures with different magnitudes and phase information is shown in Fig. 5.8a and Fig. 5.9a. To better reflect the real monitoring environment, the simulated signature is mixed with different levels of Gaussian noise. The associated signal-to-noise ratio's (SNR's) of the resulting data mixtures are −2dB (Fig. 5.8b) and −5dB (Fig. 5.9b), respectively. The detection results, i.e., waveform and inflexion points, using the two detectors are shown in parts (c) and (d) of Fig. 5.8 and Fig. 5.9. The simulated particle signatures are correctly recovered from the noisy signal mixture for both cases as shown in Fig. 5.8c and Fig. 5.9c. Applying detector H^n correctly identifies phase information and peak locations without being confused by spurious signatures. It is also observed that, because of its low-pass filter characteristic, detector H^n can remove noise and recover the particle signature without resorting to additional filters. It is also shown that the smallest signature that has been submerged



(a) Simulated signature, $x_1(t)$

(b) Simulated signature with noise, $x_2(t)$, SNR=-2dB

(c) $H^{-0.9}\ x_2(t)$

(d) $I^{-0.9}\ x_2(t)$

Figure 5.8. Numerical evaluation 1 (SNR=−2dB)

in the noise (Fig. 5.9b) is well recovered by H^n in Fig. 5.9c. Such a heavily contaminated small signature may not be detected using a threshold-based method because there is no clear distinction between the noise and signal. The authenticity of the recovered signatures is further confirmed by detector I^n by revealing their inflexion points. For both SNR levels, the locations of the three inflexion points of each of three signatures are clearly identified by the peak locations shown in Fig. 5.8d and Fig. 5.9d.

5.4 Experimental work

In this section, the two detectors are applied to the data measured by an in-line ODM sensor. A commercially available in-line ODM sensor (GasTOPS, Ltd.) is used in this

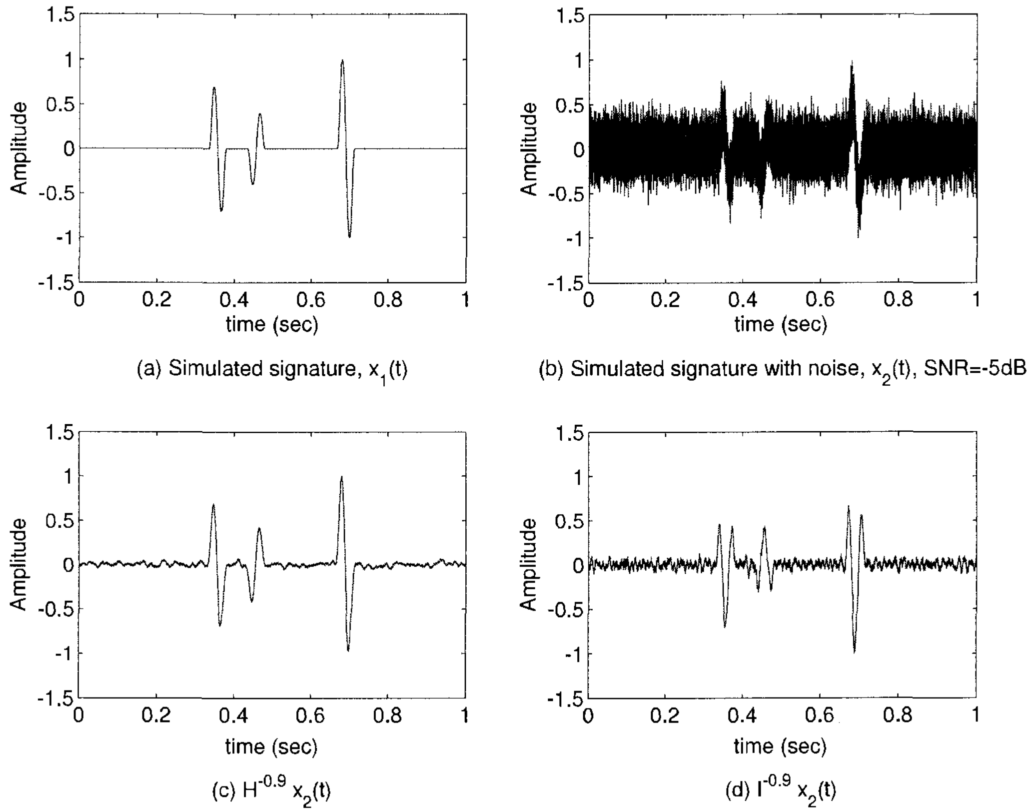


Figure 5.9. Numerical evaluation 2 (SNR=−5dB)

experiment. The data is collected using an NI DAQ board (PCI-6040E) with 500kS/s reading capacity (for single channel) and 12-bit resolution. The analogue signal from the ODM sensor is digitized at 8,000 samples/s and processed using a Pentium 4 PC with 3.4GHz CPU and 1GB RAM. The proposed algorithm is implemented using MATLAB. Two manufactured spherical particles (one ferromagnetic and one non-ferromagnetic) are used to test the algorithm. The diameters are $175\mu m$ and $508\mu m$ for the ferromagnetic and non-ferromagnetic particles, respectively. The data are acquired by moving the particles through the ODM sensor. It should be mentioned that the ferromagnetic and non-ferromagnetic particles have opposite signature phases. This is useful for distinguishing debris materials.

The signals of the non-ferromagnetic and ferromagnetic particles measured by the ODM sensor are plotted in Fig. 5.10a. The phase difference of the two signatures, non-ferromagnetic followed by ferromagnetic, is clearly shown in the figure. The outputs of detectors H^n and I^n are displayed in Fig. 5.10b and Fig. 5.10c, respectively. It is obvious that detector H^n has removed much of the noise components from the measured signal as shown in Fig. 5.10b. In addition, the inflexion points of two particle signatures are located in Fig. 5.10c, which further evidence the existence of the two particles.

The measured signals in Fig. 5.10a are strong and obvious. This is because the data are acquired in a relatively "quiet" environment. In practice, the sensor is often mounted near an operating engine and the data are acquired in a noisy environment. Under this circumstance, the measured signal generally contains much more severe noise and its SNR is much lower. To simulate such a condition, the ferromagnetic and non-ferromagnetic particles used in the previous experiment are replaced by two paper balls to generate much weaker signals. The paper balls used in this experiment are 4mm and 6mm in diameter. Fig. 5.11a shows the measured in-line oil-debris signal containing two particle signatures. The second particle signature (generated by the 6-mm paper ball) is easily identified because of its distinguishable magnitude. However, by looking at Fig. 5.11a alone, it is difficult to identify the first one (generated by the 4-mm paper ball) because of the smaller magnitude hidden in the noise. The two detectors are then applied to the

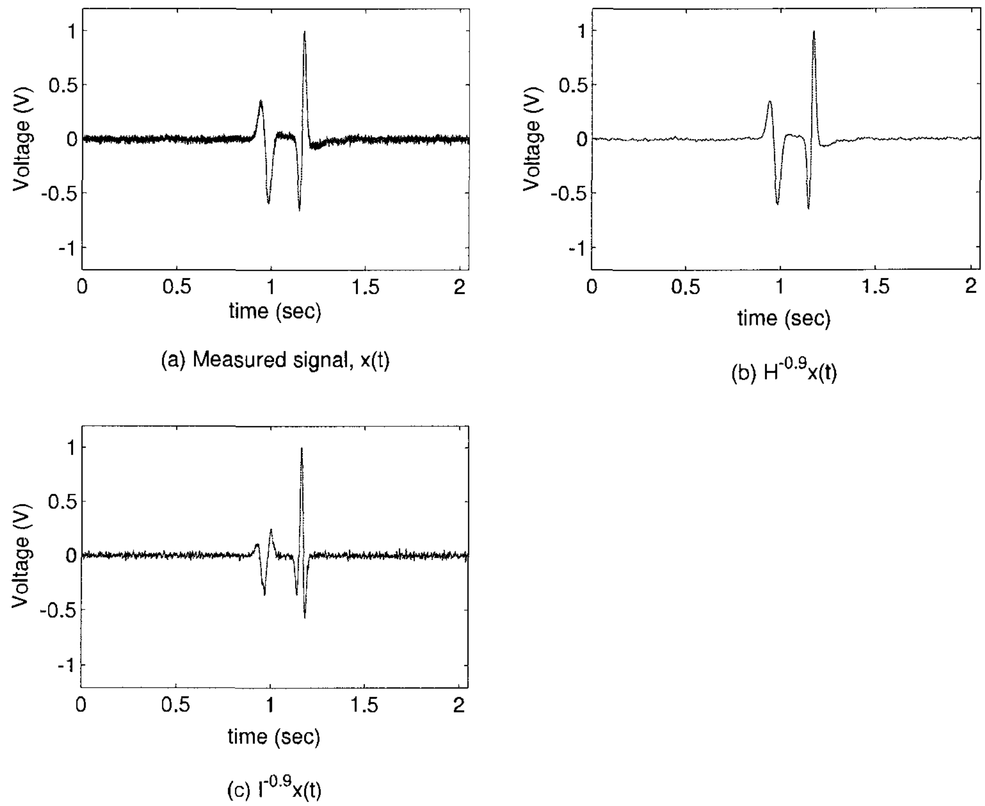


Figure 5.10. Extraction of non-ferromagnetic particle signature (the first signature) and ferromagnetic particle signature (the second one)

raw data. The signatures and inflexion points of both paper balls can be clearly identified as shown in Fig. 5.11b and Fig. 5.11c. This application demonstrates the robustness of two detectors in extracting particle signatures under noisy environment.

5.5 Discussion

Two detectors defined by generalized differentiation have been proposed to detect oil debris signatures. As shown in the simulation and experimental studies, they do not require any threshold information and are effective in extracting particle signatures even in a very noisy environment. Therefore, the two issues mentioned in section 2.1.2, i.e.,

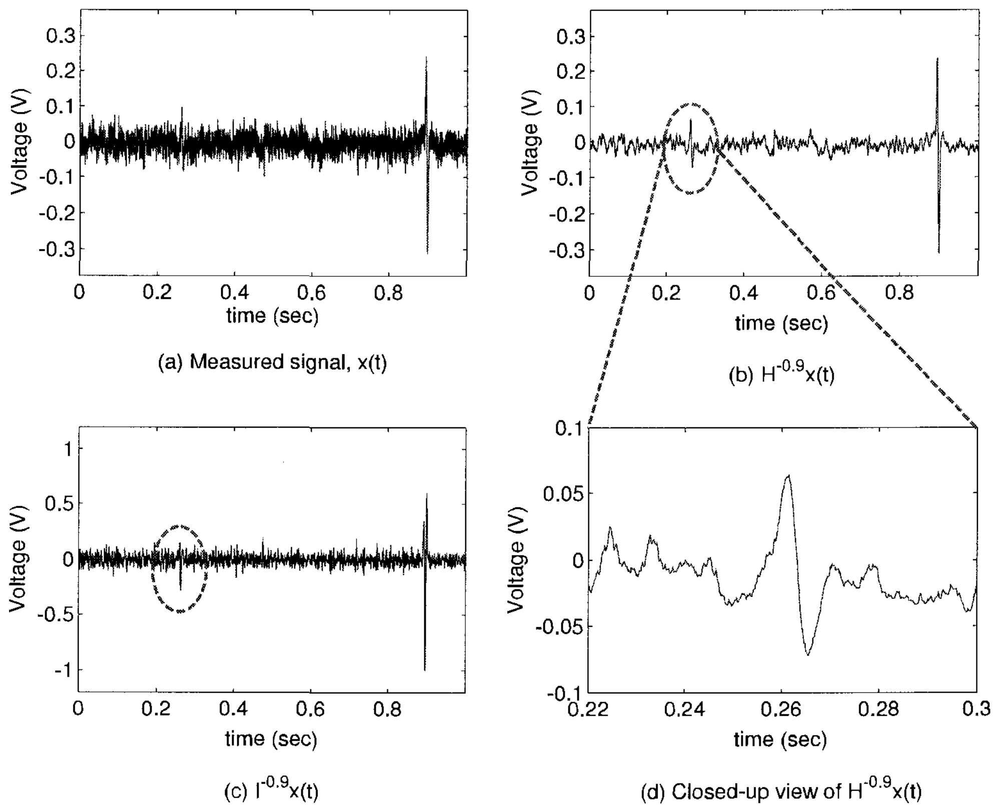


Figure 5.11. Extraction of in-line paper ball signatures (spherical paper balls, 4mm and 6mm in diameter associated with the first and the second signatures)

the selection of threshold values and the need for digital filtering, have been resolved accordingly. As a result, a detection algorithm can be substantially simplified.

Part II

Single-Channel Signal Separation

Chapter 6

Separation of Fault Features from a Single-Channel Mechanical Signal Mixture using Wavelet Decomposition[1]

This chapter reports a joint wavelet decomposition and Fourier transform approach to the separation of periodic mechanical source signals from a single-channel signal mixture. With this method, the signal mixture is first decomposed into certain wavelet scales. The resulting wavelet coefficients are then Fourier transformed to extract the information pertinent to each signal source from these scales. Next, the number of signal sources is determined and the wavelet coefficients for each signal are constructed in all scale levels. Finally, the source signals can be reconstructed using these wavelet coefficients. Since this method does not require the number of sources to be known *a priori*, it is particularly suitable for mechanical fault signal separation as the number of source signals varies with time and is unpredictable. It is also important to point out that the number of sources is determined without the commonly adopted sequential extraction/learning process and hence the proposed method can be used for on-line fault detection due to the reduced computing burden. The application of this method has been demonstrated using mixed

[1] Appeared in *Mechanical Systems and Signal Processing*, vol.21, no.5, 2007, pp.2025-2040. See Appendix B.6 for MATLAB script.

bearing data containing both inner- and outer-race fault signals.

6.1 Proposed Single-Channel BSS (Blind Signal Separation) Method

At a time instant t, the monitored single-channel (sensor) signal $x(t)$ is a blend of unknown independent source signals, i.e.,

$$x(t) = \sum_i p_i s_i(t) \tag{6.1}$$

where $s_i(t)$ is the signal from source i at time t and p_i is the weight of source signal i. The value of p_i is constant but unknown. The purpose is to a) determine the number of sources, and b) separate the source signals. Even if the number of sources is known, the problem is still too ill-conditioned to tackle mathematically (Jang and Lee, 2003).

This chapter proposes a joint wavelet decomposition and Fourier transform method. The main idea is to decompose the signal mixture to certain wavelet scales and then extract the information pertinent to each signal source from these scales. The wavelet coefficient at a different scale contains a different amount of information of a signal and hence would contribute to the construction of this signal differently. Therefore, if the levels of contributions, or contribution ratios, of all the wavelet coefficients at all scales to a source signal can be identified, the signal can be constructed. If the number of sources is known or can be determined, all the source signals can be constructed in the same way. Accordingly the proposed method involves the following three phases:

1. Decompose the single-channel signal mixture by wavelet transform.
2. Determine the number of source signals and contribution ratios of all wavelet coefficients to all source signals.
3. Reconstruct source signals based on the above results.

It should be noted that, unlike many sequential approaches, the proposed method does not require the "learning" process that may restrict the application in dynamic situations involving new and unpredictable fault features. The overall process is shown in Fig. 6.1 and the details of the method are described in the following sections.

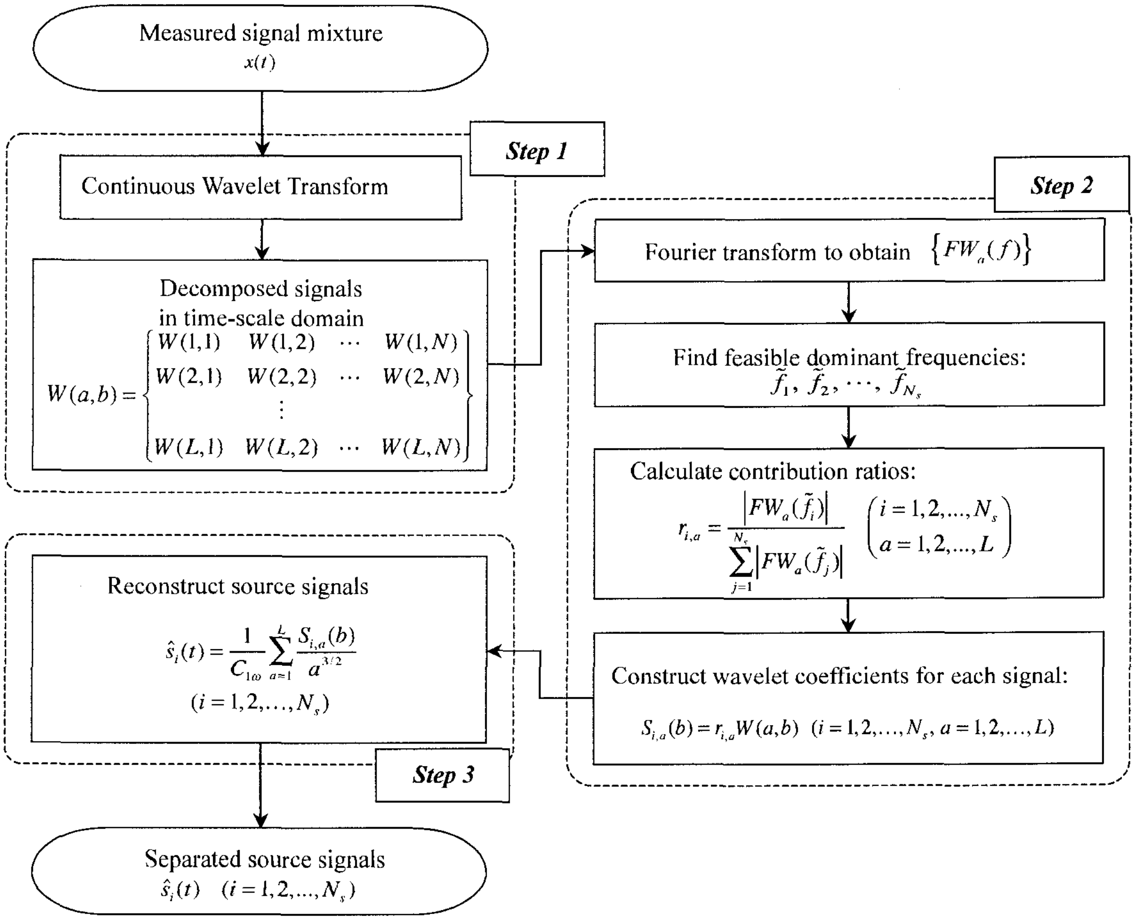


Figure 6.1. Block diagram of the single-channel signal separation process

6.1.1 Decomposition of the single-channel signal mixture

A. Wavelet transform

Since a single-channel signal mixture does not provide enough information for separation of source signals, additional information will be extracted in this phase by decomposing the signal mixture in different wavelet scales. In this study, the continuous wavelet transform (CWT), given as follows (from Eq. (3.2)), is used:

$$W(a,b) = \int x(t)w^*_{a,b}(t)dt \quad (a = 1,\, 2,\, \ldots,\, L)$$

where $W(a,b)$ is the wavelet coefficient, L is the largest scale level, asterisk $(*)$ indicates complex conjugate and $w_{a,b}(t)$ is the daughter wavelet given by (from Eq. (3.3)):

$$w_{a,b}(t) = \frac{1}{|a|^{1/2}} w\left(\frac{t-b}{a}\right)$$

where a is the dilation (scale) factor and b is the translation factor in time. $|a|^{-1/2}$ is used to ensure energy preservation (Goupillaud et al., 1984). Since each scale is associated with one aspect of the signal composition, additional information can be extracted from the decomposed signal in each scale. In this study, the Morlet wavelet is used because it is similar to that of the impulsive mechanical signals (Lin and Qu, 2000; Lin et al., 2004). These wavelet transformed subband signals can be considered as the bandpass filtered signals, where the bandwidth of each bandpass filter is determined by the scale level of the wavelet transformation.

B. Determine number of wavelet scale levels

The selection of number of scales L is important in signal reconstruction. The bandpass filter resulting from a small L may not provide sufficient frequency coverage to include the frequency components related to the mechanical faults. In general, the more scale levels we use, the more "authentic" the reconstructed signals would be. However, the use of a large number of scales inevitably leads to heavy computing burden, which undermines the usefulness of the CWT approach. An appropriate scale level can be determined by considering fault feature frequencies of interest and frequency corresponding to the scale level based on the concept of *pseudo frequency.* Each scale level has its own pseudo frequency which is determined by the center frequency of the wavelet function, sampling period and scale level as follows (Mathworks, 2007).

$$f_a^{ps} = \frac{f^c}{a \cdot 1/F_s} \tag{6.2}$$

where f_a^{ps} is the pseudo frequency corresponding to scale a, f^c is the center frequency of the selected wavelet function, and F_s is the sampling freqeuncy. For the pseudo frequency to cover the entire frequency range of interest, the following condition must be satisfied:

$$f_{a=L}^{ps} \leq f_{\text{interest}} \leq f_{a=1}^{ps}. \tag{6.3}$$

Therefore, the minimum frequency of interest ($\min f_{\text{interest}}$) should be greater than or equal to the pseudo frequency of the maximum scale, i.e.,

$$f^{ps}_{a=L} \leq \min f_{\text{interest}}. \tag{6.4}$$

In the context of machine fault detection, the frequencies of interest may be gear mesh frequencies for gear faults or bearing characteristic frequencies. Hence, the smallest a that satisfies Eq. (6.4) can be used as the maximum scale level such that all possible frequencies of interest can be covered with the minimum possible computational effort. Consider an example with $f^c = 5\text{Hz}$, $1/F_s = 1\text{ms}$, and $\min f_{\text{interest}} = 200\text{Hz}$. From Eqs. (6.2) and (6.4), we have

$$f^{ps}_{a=L} = \frac{5}{0.001L} \text{ or } L \geq \frac{5000}{200} = 25.$$

To reduce computing time, we choose $L = 25$.

The computing time can be further reduced by specifying a lower limit of a, denoted L^*, based on the maximum frequency of interest, i.e.,

$$\max f_{\text{interest}} \leq f^{ps}_{a=L^*}. \tag{6.5}$$

For example, suppose the maximum frequency of interest is 1,000Hz and other information remains unchanged. Then the lower limit of a is:

$$\max f_{\text{interest}} = 1000 \leq f^{ps}_{a=L^*} = \frac{5}{0.001L^*} \text{ or } L^* \leq 5.$$

Hence, the wavelet decomposition can be done from scale 5 to 25. However, since most of the time saving is achieved by specifying a tighter upper bound of scale levels and L^* is often close to 1 in the context of machine fault detection, we assume $L^* = 1$ in the following discussions.

6.1.2 Determine number of source signals and calculate contribution ratios

Once the single-channel signal mixture is decomposed, i.e., $\{W(a,b)\}$ is determined, the fast Fourier transform (FFT) will be applied to the decomposed results in all scale levels to obtain $\{FW_a(f)\}$ which will then be used to estimate the number of source signals and calculate contribution ratios as detailed in the following.

A. Number of source signals

The FFT spectrum in a wavelet scale a, i.e., $\{FW_a(f)\}$, may contain several peaks, each corresponding to a frequency. The frequency associated with the highest peak is most likely to be related to one of the source signals and is called *dominant frequency* here, denoted $\hat{f}_a\,(a = 1, 2, \ldots, L)$. To determine the number of sources, the dominant frequencies for all scales should be identified.

However, a dominant frequency of a certain scale could be a harmonic of another dominant frequency in a different scale and, if so, it should not be considered as an independent signal. Furthermore, some scales may cover certain frequency ranges that are not of our interest. Such redundant frequencies have to be removed. In other words, we are looking for dominant frequencies, $\tilde{f}_1, \tilde{f}_2, \ldots, \tilde{f}_{N_s}$, such that:

1. $\tilde{f}_i \neq n\tilde{f}_a, \; (i = 1, 2, \ldots, N_s, \; a = 1, 2, \ldots, L, \; n = 2, 3, \ldots)$
2. $\min f_{\text{interest}} \leq \tilde{f}_i \leq \max f_{\text{interest}}$.

The dominant frequencies that satisfy the above conditions are called *feasible dominant frequencies*. The number of feasible dominant frequencies N_s is then considered as the number of source signals. It should be noted that the same $\tilde{f}_i$ may appear in several scales but this does not increase signal count.

B. Contribution ratios

The contribution ratio represents the weight of the wavelet coefficient of a scale in a source signal. For completeness, we include all scales for every source in the calculation even though only the scales with non-zero peaks at a source frequency, e.g. $\tilde{f}_i$, actually contribute to that source signal, say, i. Defining $\left|FW_a(\tilde{f}_i)\right|$ as the peak magnitude at feasible frequency $\tilde{f}_i$ in scale a, its contribution ratio in signal source i is given by:

$$r_{i,a} = \frac{\left|FW_a(\tilde{f}_i)\right|}{\sum_{j=1}^{N_s}\left|FW_a(\tilde{f}_j)\right|} \qquad (i = 1, 2, \ldots, N_s,\ a = 1, 2, \ldots, L). \tag{6.6}$$

With the above contribution ratio, the wavelet coefficients for each signal source can be constructed as:

$$S_{i,a}(b) = r_{i,a}W(a,b) \qquad (i = 1, 2, \ldots, N_s,\ a = 1, 2, \ldots, L) \tag{6.7}$$

where $S_{i,a}(b)$ is the constructed wavelet coefficient or wavelet representation of signal i in scale a at time b. In a sense, vector $\boldsymbol{S}_i(b) = (S_{i,1}(b),\ S_{i,2}(b),\ \ldots,\ S_{i,L}(b))$ represents the make-up or "spectrum" of signal i in the wavelet domain. These $S_{i,a}(b)$ values will be used for constructing the source signals.

6.1.3 Construct source signals

Once the number of source signals is determined and the wavelet coefficients for each signal are obtained, the source signals can be constructed. Based on the Morlet formula (Rioul and Duhamel, 1992), the source signals can be calculated as:

$$\begin{aligned} \hat{s}_i(b) &= \frac{1}{C_{1\omega}}\int S_{i,a}(b)a^{-3/2}da \\ C_{1\omega} &= \int_{-\infty}^{\infty}\frac{\hat{w}^*(\omega)}{|\omega|}d\omega \end{aligned} \qquad (i = 1, 2, \ldots, N_s). \tag{6.8}$$

Considering only the selected scales, we have

$$\hat{s}_i(b) = \frac{1}{C_1\omega}\sum_{a=1}^{L}\frac{S_{i,a}(b)}{a^{3/2}} \qquad (i = 1, 2, \ldots, N_s). \tag{6.9}$$

6.1.4 Summary of single-channel BSS procedure and illustration

Based on the above results, the overall single-channel BSS procedure is detailed in Fig. 6.1 and can be summarized as follows:

Step 1. Apply CWT to the single-channel signal mixture $x(t)$ to obtain the wavelet decomposed signals $W(a, b)$ using Eq. (3.2). The number of scale levels is determined by Eqs. (6.2), (6.4) and (6.5).

Step 2. Perform Fourier transform of $W(a, b)$ and obtain frequency spectrum, FW_a, for each scale.

Step 3. Find the dominant frequency, $\hat{f}_a$, in each scale.

Step 4. Check the feasibility conditions for all $\hat{f}_a$ and find $\tilde{f}_i$. The number of $\tilde{f}_i$'s, i.e., N_s, is considered as the number of source signals.

Step 5. Calculate contribution ratios and construct wavelet coefficients for each signal using Eqs. (6.6) and (6.7).

Step 6. Reconstruct source signals, $\hat{s}_i(b)$, using Eq. (6.9).

To illustrate the BSS process, consider three sinusoidal signals (Fig. 6.2). After applying wavelet transform in 30 scale levels (step 1) and FFT to the wavelet coefficients in each of the 30 scales (step 2), we obtain the frequency spectrum for every scale. Fig. 6.3 displays the wavelet transform results and frequency spectra in scales 4, 8, 16, and 30. One scale may contain more than one peak but only the highest is of interest for determining number of sources. For example, there are two peaks corresponding to 120Hz and 300Hz respectively in scale 8 (Fig. 6.3d). In this case, the highest peak is at 120Hz and hence the dominant frequency is 120Hz for this scale. Repeating this process for all scales yields three dominant frequencies, i.e., $\hat{f}_{a=4}$ = 300Hz, $\hat{f}_{a=8} = \hat{f}_{a=16}$ = 120Hz, and $\hat{f}_{a=30}$ = 45Hz (step 3). Since all the three frequencies satisfy the non-harmonic and frequency coverage conditions, they are all feasible dominant frequencies, denoted respectively $\tilde{f}_1$(=45Hz), $\tilde{f}_2$(=120Hz), and $\tilde{f}_3$(=300Hz) (step 4). Accordingly the number of source signals is determined as three (N_s=3). To construct the three source signals, the contribution ratios and wavelet coefficients for each signal have to be calculated using Eqs. (6.6) and (6.7) in each scale (step 5). For instance, the peak heights at the three identified signal frequencies in scale 8 are $\left|FW_8(\tilde{f}_1)\right| = |FW_8(45)| = 0$, $\left|FW_8(\tilde{f}_2)\right| =$

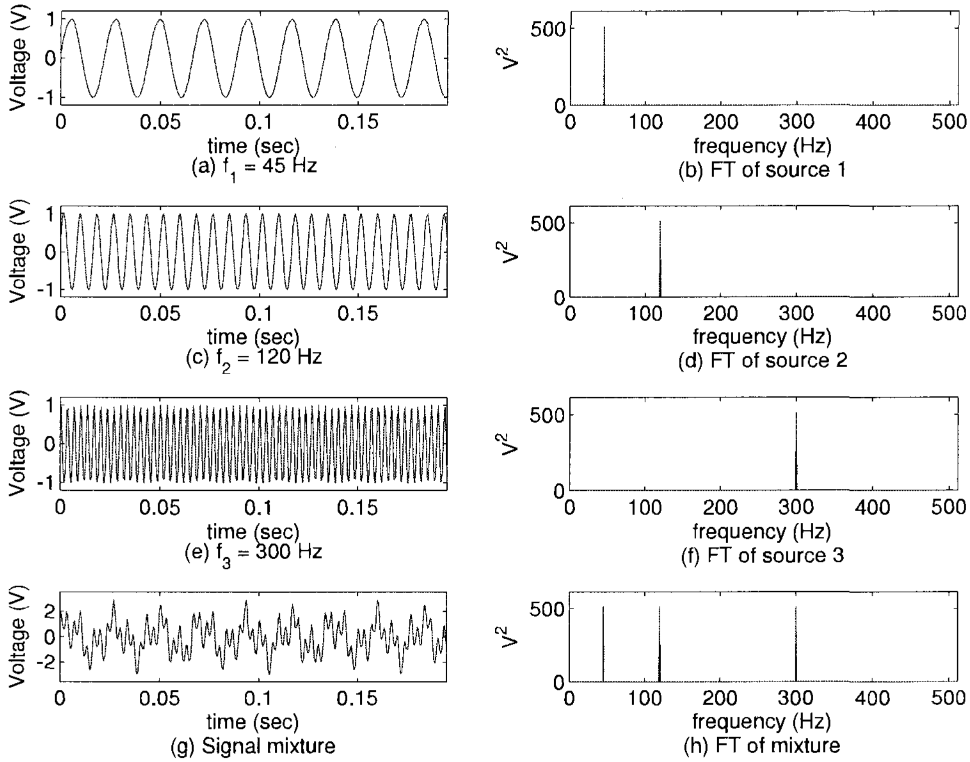


Figure 6.2. Time and frequency domain representations of source signals and signal mixture

$|FW_8(120)| = 92.07$, and $\left|FW_8(\tilde{f}_3)\right| = |FW_8(300)| = 23.42$, respectively. Hence, the associated contribution ratios are

$$r_{1,8} = \frac{\left|FW_8(\tilde{f}_1)\right|}{\sum_{j=1}^{3}\left|FW_8(\tilde{f}_j)\right|} = \frac{0}{0 + 92.07 + 23.42} = 0.00$$

$$r_{2,8} = \frac{\left|FW_8(\tilde{f}_2)\right|}{\sum_{j=1}^{3}\left|FW_8(\tilde{f}_j)\right|} = \frac{92.07}{0 + 92.07 + 23.42} = 0.80$$

$$r_{3,8} = \frac{\left|FW_8(\tilde{f}_3)\right|}{\sum_{j=1}^{3}\left|FW_8(\tilde{f}_j)\right|} = \frac{23.42}{0 + 92.07 + 23.42} = 0.20.$$

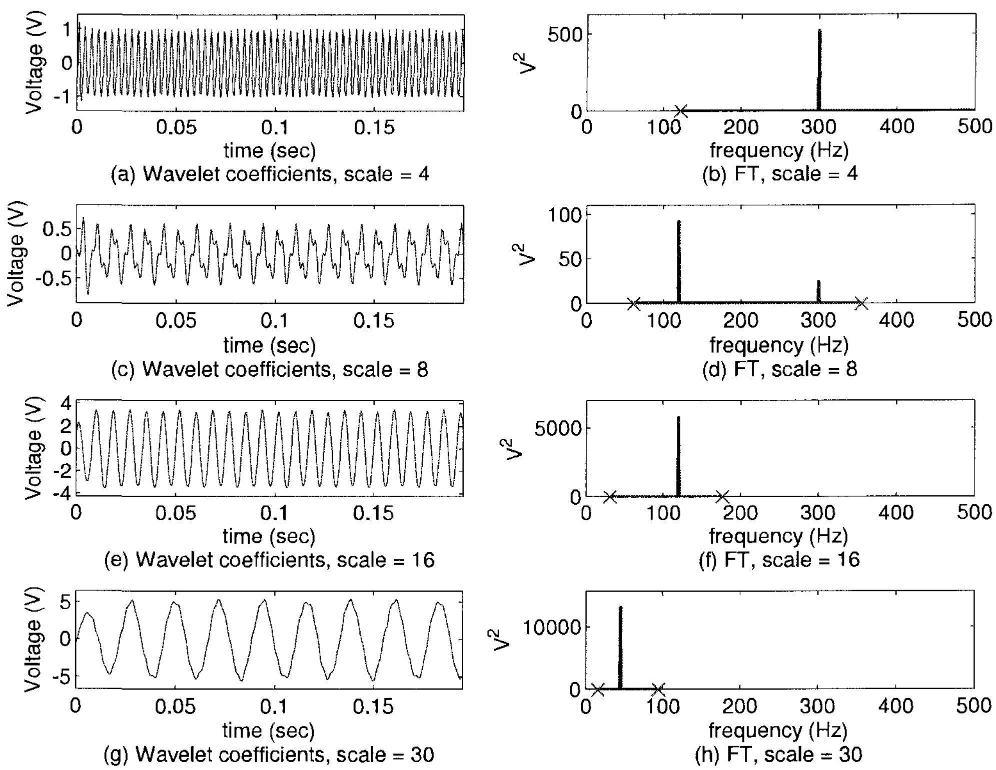
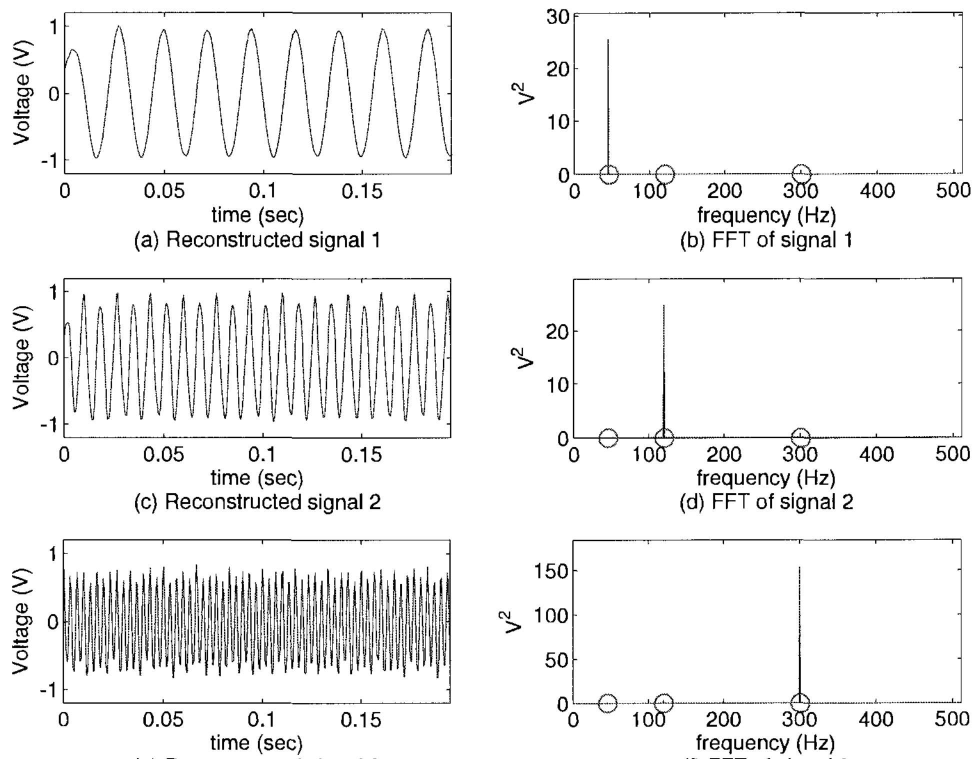


Figure 6.3. Wavelet transformed signals and their Fourier transforms for four selected scale levels (Note: "×" specifies the frequency range of the associated scale in the FT plots)

The contribution ratios for other scales can be calculated similarly. Then the source signals can be constructed using Eq. (6.9) (step 6). The constructed source signals are shown in Fig. 6.4, which is consistent with the original three signals.

6.1.5 Quality evaluation of separated signals

If the original signals were known, then the quality of the separated or reconstructed signals could be evaluated by simply comparing the reconstructed and the original signals. However, in the context of BSS, the original signals are unknown and hence it is impossible to make such comparison. For this reason, this thesis proposes to evaluate the



(a) Reconstructed signal 1
(b) FFT of signal 1
(c) Reconstructed signal 2
(d) FFT of signal 2
(e) Reconstructed signal 3
(f) FFT of signal 3

Figure 6.4. Separated source signals and their frequency representations (Note: the circles represent the source signal frequencies)

quality of a reconstructed signal by examining the power spectral density (PSD) of all the separated signals in the "home" frequency (or envelope) spectrum of the concerned signal. The rationale is that the PSD of a signal in its own frequency (or envelope) spectrum should dominate others in the same spectrum. The level of dominance can be considered as the quality of the reconstructed signal. This study adopts the signal-to-noise ratio (SNR) concept and rename it signal-to-signal ratio (SSR) to quantify the level of dominance. More specifically, the quality of a reconstructed signal i, is quantified by

$$SSR_i = \min_{\substack{1 \le j \le N_s \\ i \ne j}} \left(10 \cdot \log \frac{PSD_i}{PSD_j} \right) \tag{6.10}$$

where PSD_i and PSD_j are the power spectral densities of signals i and j in the "home" spectrum of signal i, respectively.

6.2 Application

The evaluation of the proposed method is done by using two different signal mixtures: i) synthesized single-channel signal mixture obtained by mixing two independently measured outer- and inner-race bearing fault signals, and ii) single-channel signal mixture of outer- and inner-race bearing fault signals measured from a machine fault simulator.

6.2.1 Application to synthesized signal mixture

The bearing fault data were obtained from Case Western Reserve University (2007). Both the inner and outer races have a single point fault created by electro-discharge machining. The fault size is 0.007 inch in diameter and 0.011 inch in depth for both inner and outer races. The shaft rotating speed is 1,797RPM (29.95Hz). The bearing characteristic frequencies (BCFs) of inner-race fault, BPFI (ball pass frequency of inner race), and outer-race fault, BPFO (ball pass frequency of outer race), can be calculated theoretically as (Harris and Piersol, 2002), from Eqs. (4.12) and (4.13), respectively:

$$BPFI = \frac{n_e}{2} f_s \left[1 + \frac{BD}{PD}\cos\beta\right] \text{ (Hz)}$$
$$BPFO = \frac{n_e}{2} f_s \left[1 - \frac{BD}{PD}\cos\beta\right] \text{ (Hz)}$$

where PD is the bearing pitch diameter, BD is the ball diameter, β is the contact angle of the balls on the races, n_e is the number of rolling elements, and f_s stands for the shaft rotational frequency. The BCFs are determined by the geometry of bearing and the shaft rotational speed. In this case, a deep groove ball bearing is used (6205-2RS JEM SKF). The other data are $BD = 0.3126$ in, $PD = 1.5370$ in, $n_e = 9$, and $\beta = 0°$, which leads to the following BCFs:

$$BPFI = 5.4152 f_s = 162.2\text{Hz}$$
$$BPFO = 3.5848 f_s = 107.4\text{Hz}.$$

The sampling rate was 12,000 samples/sec and 4,000 data points were taken from the reference data source. The original data from CWRU were collected for single point fault, i.e., the inner and outer race fault signals were monitored separately. Hence, to obtain the single-channel signal mixture, the inner- and outer-race fault signal data are mixed using mixing matrix [1 1]. The signal separation method is then applied to this signal mixture.

Fig. 6.5 shows the time and frequency domain (envelope spectra) plots of the source signals and their mixture. From the signal mixture (Fig. 6.5e and Fig. 6.5f), it is difficult to conclude how many source signals are involved. It should also be noted that, in

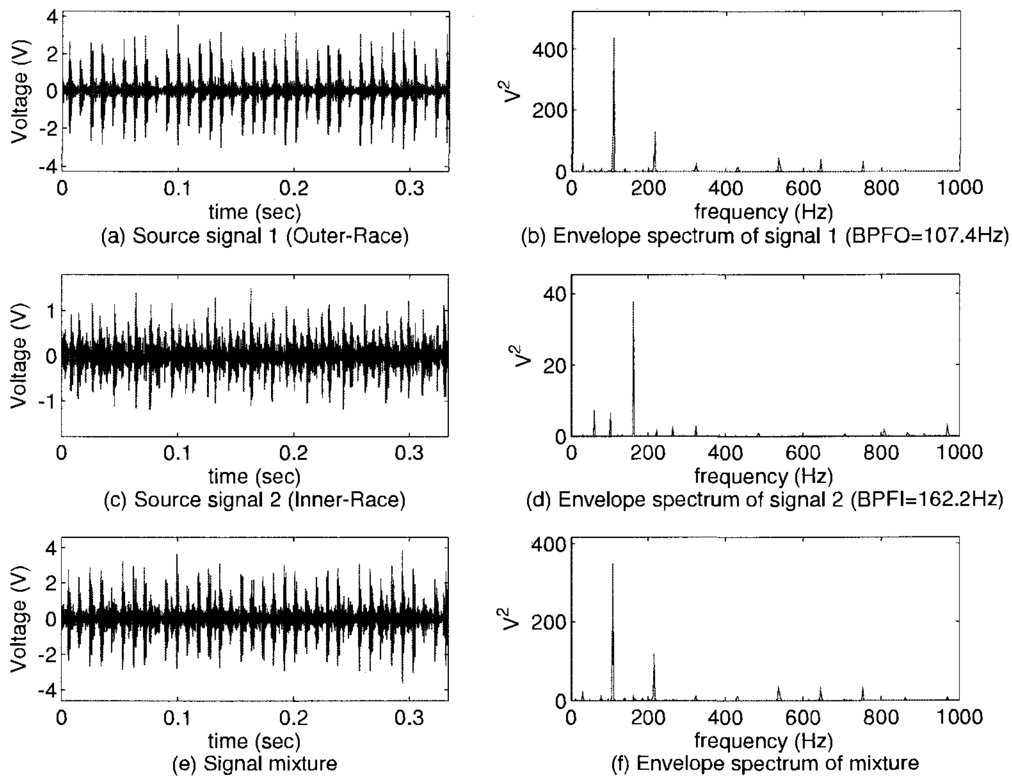


Figure 6.5. Source signals, signal mixture and their envelope spectra: (a, b) outer-race fault signal, (c, d) inner-race fault signal, and (e, f) signal mixture

Fig. 6.5f, the peak associated with signal 2 at 162.2Hz is very small as compared to that of signal 1.

The proposed separation method is applied as follows:
Steps 1 and 2: Carry out wavelet decomposition and Fourier transform. After applying CWT (12 scale levels) to the signal mixture and FFT to the each of the scales, the frequency spectra are obtained for all scales.

Step 3: Three dominant frequencies are found at 12, 108 and 162Hz.

Step 4: Check the feasibility conditions and determine number of sources. In this case, the minimum frequency of interest is BPFO, i.e., 107.4Hz and the maximum frequency is set to 1,000Hz (about 6 times of the BPFI). Therefore, the dominant frequency at 12Hz is not in the frequency range of interest and is excluded. The other two dominant frequencies, 108Hz and 162Hz, satisfy the two conditions, i.e., they are both within the frequency range of interest and are not harmonics to each other. It is hence concluded that there are two source signals at 108Hz and 162Hz, respectively.

Steps 5 and 6: Construct wavelet coefficients for each of the two signals using Eq. (6.7) and reconstruct the two source signals with Eq. (6.9). The constructed signals are plotted in Fig. 6.6 which clearly shows the two separated frequency components.
From Eq. (6.10), the signal-to-signal ratios for the two reconstructed signals are:

$$SSR_{108\text{Hz}} = 10 \cdot \log \frac{PSD_{108\text{Hz}}}{PSD_{162\text{Hz}}} = 10 \cdot \log \frac{100.47V^2}{4.43V^2} = 13.55\text{dB}$$
$$SSR_{162\text{Hz}} = 10 \cdot \log \frac{PSD_{162\text{Hz}}}{PSD_{108\text{Hz}}} = 10 \cdot \log \frac{11.82V^2}{0.24V^2} = 16.92\text{dB}.$$

The above results indicate that the level of dominance of a signal in its own frequency (or envelope) spectrum dominates the other substantially and hence the proposed method can separate the source signals from the single-channel signal mixture reasonably well.

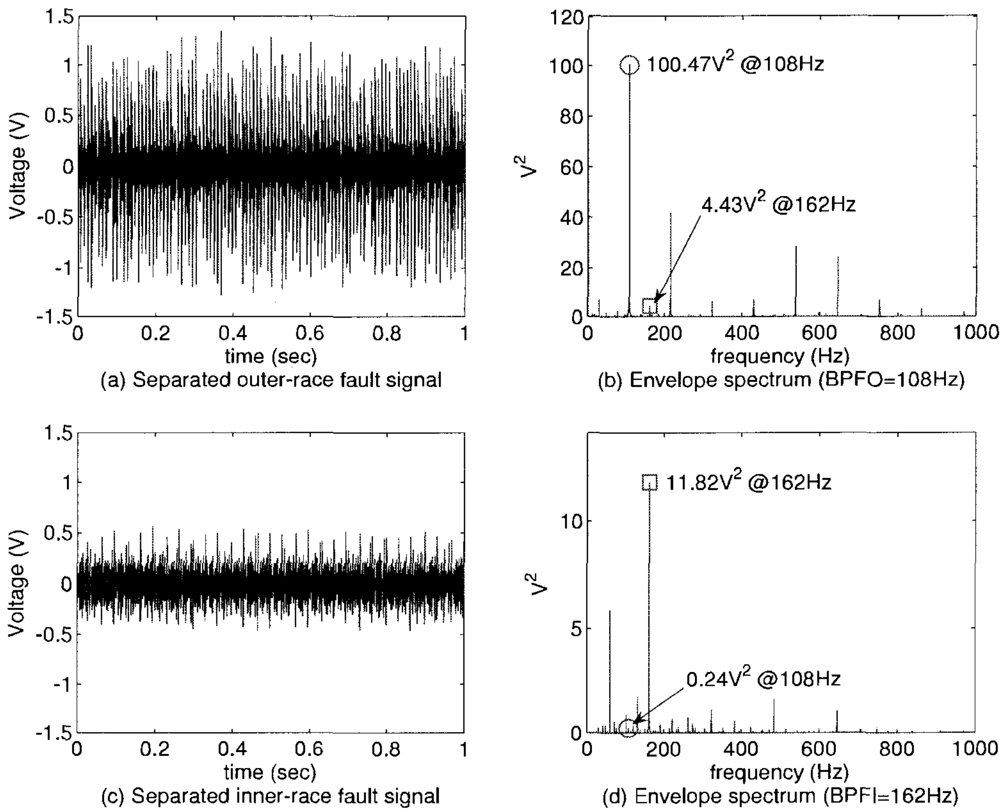


Figure 6.6. Separated source signals and their frequency representations

6.2.2 Application to real signal mixture

The proposed method was applied to the signal mixture of outer- and inner-race fault signals (for single-point fault on the race). The experiment was carried out using a SpectraQuest machinery fault simulator (MFK-PK5M). The bearing faults were pre-created by SpectraQuest — the supplier of the machinery fault simulator. The machine fault simulator and experimental setup are shown in Fig. 6.7. A Montronix accelerometer (Model VS100-100), with a sensitivity range of 1–12 kHz and a sensitivity of 100mV/g, was mounted on top of the housing of a bearing with inner-race fault to measure the mixture of the signals from the two bearings. The two bearings were used to support a two mass rotor test kit. The two well balanced mass rotors are 2-inch thick, 4-inch in

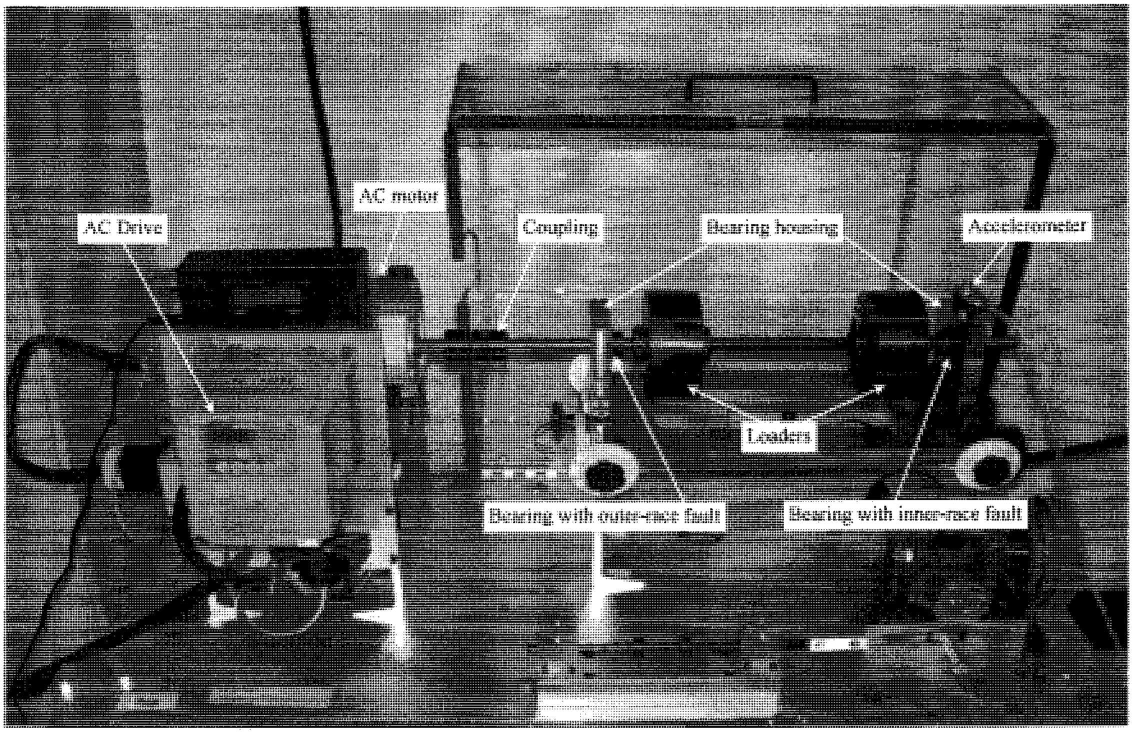


Figure 6.7. Experimental setup

diameter and 11.1 lbs each, fitting into a 5/8-inch steel shaft. The test kit was driven by a 3-hp AC motor controlled by a Hitachi controller (SJ200-022NFU). The signal was sent to an NI DAQ card (PCI-6040E) with 500kS/s reading capacity (for single channel) and 12-bit resolution. The analogue signal from the accelerometer amplifier was digitized at 6,000 samples/s and processed using a Pentium 4 PC. The proposed single-channel source separation algorithm was implemented using MATLAB. Two motor shaft rotating speeds were considered: 1,660RPM (27.67Hz) and 1,402RPM (23.37Hz).

A. *Application 1: Shaft rotating speed at 1,660RPM*

The associated bearing characteristic frequencies for shaft rotating speed at 1,660RPM are as follows:

$$BPFI = 5.002f_s = 138.4\text{Hz}$$

$$BPFO = 2.998f_s = 82.9\text{Hz}.$$

The measured signal mixture and the separated source signals using the proposed method are illustrated in Fig. 6.8 in both time and frequency domains. The signal-to-signal ratios of the two reconstructed signals are respectively:

$$SSR_{83\text{Hz}} = 10 \cdot \log \frac{PSD_{83\text{Hz}}}{PSD_{138\text{Hz}}} = 10 \cdot \log \frac{0.21}{0.06} = 5.44\text{dB}$$
$$SSR_{138\text{Hz}} = 10 \cdot \log \frac{PSD_{138\text{Hz}}}{PSD_{83\text{Hz}}} = 10 \cdot \log \frac{1.60}{0.34} = 6.73\text{dB}.$$

The outer-race fault signal (Fig. 6.8d) contains many low frequency components that are mostly caused by the transmission noise. As the accelerometer is installed near the inner-race fault bearing, the outer-race fault signal may be contaminated during trans-

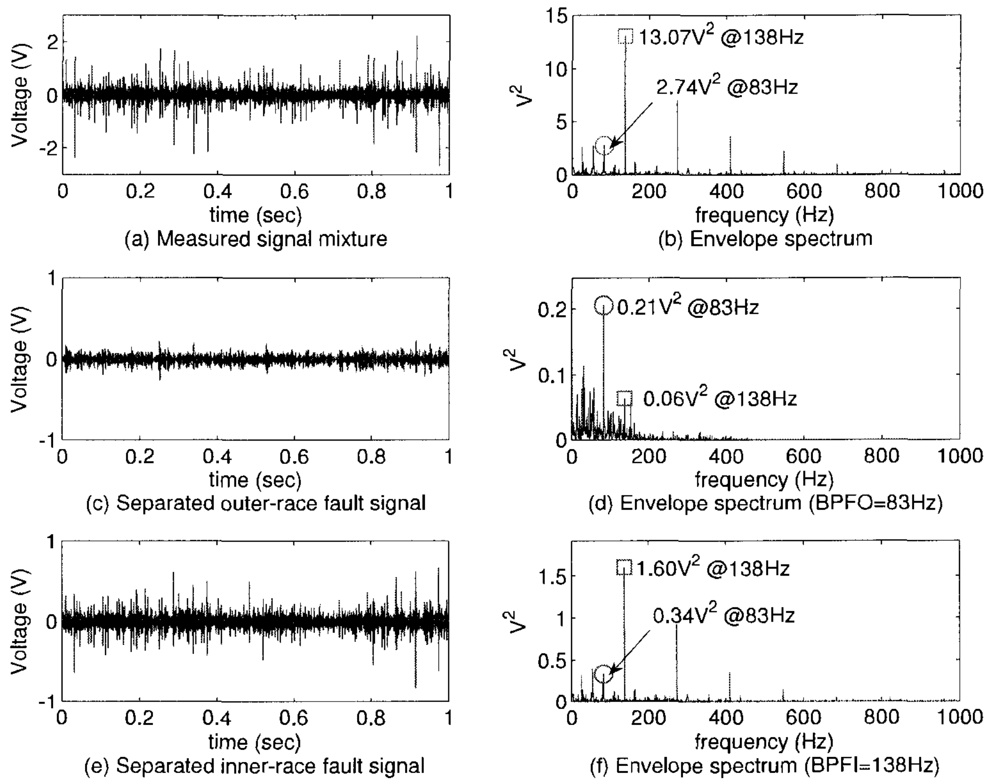


Figure 6.8. Measured signal mixture from the machinery fault simulator and separated source signals in time and frequency domains (shaft rotating speed of 1,660RPM)

mission from the fault source to the sensor. Even though there is a noise effect, the peak frequency component of BPFO is clearly shown in the envelope spectrum.

B. *Application 2: Shaft rotating speed at 1,402RPM*

For the motor shaft rotating speed at 1,402RPM (23.37Hz), the associated bearing characteristic frequencies are:

$$BPFI = 5.002 f_s = 117.0\text{Hz}$$
$$BPFO = 2.998 f_s = 70.0\text{Hz}.$$

The measured signal is plotted in Fig. 6.9a and its envelope spectrum is displayed in

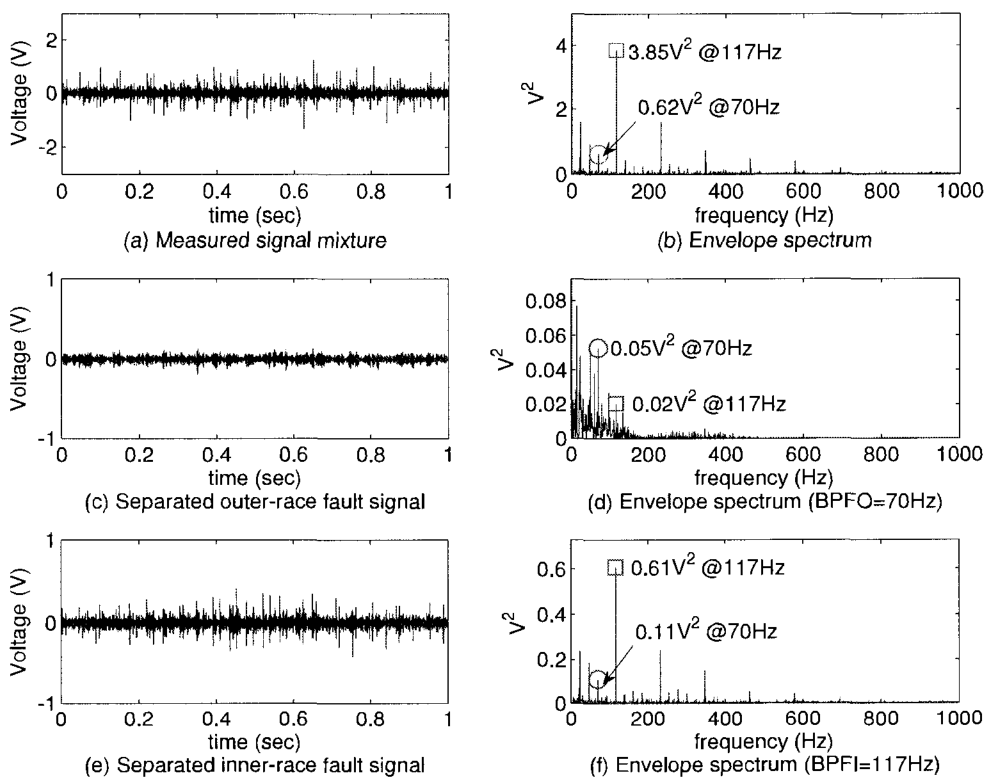


Figure 6.9. Measured signal mixture from the machinery fault simulator and separated source signals in time and frequency domains (shaft rotating speed of 1,402RPM)

Fig. 6.9b. Similar to Application 1, the outer-race fault signal (70Hz) cannot be observed clearly in Fig. 6.9b. The magnitude of outer-race fault signal in envelope spectrum is much smaller than that of inner-race fault signal and hence it is difficult to conclude whether the measured signal contains two fault signals.

The proposed method is applied to the data. The separated two source signals as plotted in Fig. 6.9 (c and e) and the associated envelope spectra are shown in Fig. 6.9 (d and f). In this case, the signal-to-signal ratios of the two reconstructed signals are respectively:

$$SSR_{\mathrm{70Hz}} = 10 \cdot \log \frac{PSD_{\mathrm{70Hz}}}{PSD_{\mathrm{117Hz}}} = 10 \cdot \log \frac{0.05}{0.02} = 3.98\mathrm{dB}$$
$$SSR_{\mathrm{117Hz}} = 10 \cdot \log \frac{PSD_{\mathrm{117Hz}}}{PSD_{\mathrm{70Hz}}} = 10 \cdot \log \frac{0.61}{0.11} = 7.44\mathrm{dB}.$$

The separation result of outer-race fault signal (70Hz in Fig. 6.9d) is not as good as inner-race fault signal (117Hz in Fig. 6.9f). There are a lot of interfering low-frequency components and the 117Hz signal appears with relatively large magnitude. Nevertheless, the signal-to-signal ratio of outer-race fault signal (70Hz) increases from −7.93dB (Fig. 6.9b) to 3.98dB (Fig. 6.9d), as shown below, by the proposed method. This obviously enhances the detectability of such a weak fault.

$$SSR_{\mathrm{70Hz,\,(b)}} = 10 \cdot \log \frac{PSD_{\mathrm{70Hz}}}{PSD_{\mathrm{117Hz}}} = 10 \cdot \log \frac{0.62}{3.85} = -7.93\mathrm{dB}$$
$$SSR_{\mathrm{70Hz,\,(d)}} = 10 \cdot \log \frac{PSD_{\mathrm{70Hz}}}{PSD_{\mathrm{117Hz}}} = 10 \cdot \log \frac{0.05}{0.02} = 3.98\mathrm{dB}.$$

6.3 Discussion

The proposed method is developed assuming that the source signals are periodic. However, the signal may not always be exactly periodic. For instance, the sliding of rolling elements may disturb the periodicity pattern. Some studies (Ho and Randall, 2000; Randall et al., 2001) have indicated that certain source signals such as bearing faults display a continuous rather than a line spectrum. Though our experimental results (Fig. 6.8 and Fig. 6.9) do not show visible impact of such disturbance to periodicity, it, if severe enough, could affect the accuracy of the result. Another potential limitation of the proposed method is that it may not effectively handle signals with small fluctuations of rotating speed. The reason is that the small speed fluctuation may smear the frequency spectrum, hence making it difficult to detect and separate frequency peaks. Since the device used in this work is not capable of on-line speed change, the actual impact of speed fluctuation on separation result remains to be investigated. It should also be pointed out that the signal mixture was not de-noised before separation. The separated signals may contain heavy noise as shown in Fig. 6.8d and Fig. 6.9d. Therefore, it is reasonable to expect that the implementation of pre-denoising process should enhance the quality of the separation results.

Part III

Fault Severity Evaluation

Chapter 7

Analysis of Fault Severity of Rolling Element Bearing by Lempel-Ziv Complexity[1]

As reviewed in Chapter 2, most fault evaluation methods rely on the comparison of two different data sets: normal and faulty conditions or less severe and more severe conditions. There are three major difficulties in the implementation of such methods:

1. A fault of the same severity may display quite different signal strengths of magnitudes under different operating conditions or for bearings of different sizes. Therefore, the signal may not be directly used for assessing fault severity.
2. The data sets may not be available especially for new machines and new working conditions. Therefore, the comparison based methods cannot be applied under such circumstances.
3. Even if the data can be obtained, it would be very time-consuming to collect, store, manage, and retrieve them for on-line applications.

It is therefore highly desirable to develop a severity measure that is independent of working conditions and does not require comparison between data sets. For this

[1] See Appendix B.1 and B.7 for MATLAB scripts.

reason, Yan and Gao (2004) proposed to evaluate fault deterioration using the Lempel-Ziv complexity. A normalized Lempel-Ziv complexity yields a non-dimensional value between zero and one [0, 1], close to zero for a pure sinusoidal signal and one for a white Gaussian noise. Yan and Gao (2004) have shown that the Lempel-Ziv complexity value increases as the size of the outer race bearing fault grows. Their results also indicate that the increase of bearing speed tends to increase the complexity value and the bearing load has no appreciable effect on the complexity. With the Lempel-Ziv complexity based fault evaluation measure, it would be possible to assess fault severity without comparing with reference data. However, as reviewed in Chapter 2, several issues require further analysis. For example, there is no clear distinction between a mechanical fault and noise — a relatively high complexity could signify either existence of a bearing fault or the presence of heavy noise. In addition, the method as reported in Yan and Gao (2004) yields quite different complexity levels for different data sets of the same bearing. Therefore, their method requires using many data sets to calculate the average of the complexity values of the data sets. The average complexity value is then considered a reliable indication of the fault severity. This could become an obstacle for real applications. Furthermore, a theoretical explanation based on the cause and mechanism of bearing faults is required as for why the complexity value can be used for measuring fault severity.

In view of the above, this chapter proposes a new version of the Lempel-Ziv complexity as a bearing fault severity measure based on the continuous wavelet transform (CWT) results, and attempts to address the issues present in the current version of the Lempel-Ziv complexity measure. To establish the relationship between the Lempel-Ziv complexity and bearing fault severity, an analytical model for a single-point defective bearing is adopted and the factors contributing to the complexity value are explained. To avoid the ambiguity between fault and noise, the Lempel-Ziv complexity is jointly applied with the CWT. The CWT is used to identify the best scale where the fault resides and to eliminate the interferences of noise and irrelevant signal components as much as possible. Then, the Lempel-Ziv complexity values are calculated for both the envelope and high-frequency carrier signal obtained from wavelet coefficients at the best scale level. As the noise and other un-related signal components have been largely removed, the Lempel-Ziv complexity value will be mostly contributed by the bearing system and hence can be reliably used as a bearing fault measure. The applications to the bearing inner- and outer-race fault signals have demonstrated that the revised Lempel-Ziv complexity can effectively measure the severity of both inner- and outer-race faults. The principal advantage of the proposed method is that a single set of measured data is sufficient for severity assessment and it is not necessary to compare the result with reference data.

7.1 A model for single-point defective bearing

For a single-point defect of rolling element bearing, a signal model consists of several simplified terms as follows (McFadden and Smith, 1984; Wang and Kootsookos, 1998).

$$x(t) = x_f(t) \cdot x_l(t) \cdot x_b(t) + n(t) \tag{7.1}$$

where $x(t)$ is the measured vibration signal, $x_f(t)$ is the impulse series produced by the fault, $x_l(t)$ is the modulation effect caused by the non-uniform load distribution, $x_b(t)$ is the bearing-induced vibration including decay, and $n(t)$ is the mixture of noise and unnecessary signal components by other parts of machine. Each vibration term can be modeled as a sinusoidal function as described in the following subsections.

7.1.1 Impulses generated by fault

A mechanical fault can be modeled based on a series of impulses with frequency f_f or period $T_f = 1/f_f$ as follows (Braun and Datner, 1979; Braun, 1980; McFadden and Smith, 1984; Wang and Kootsookos, 1998):

$$x_f(t) = d_f \sum_{k=-\infty}^{\infty} \delta\left(t - kT_f\right) \approx A_f \sum_{i=1}^{N_1} \cos\left(2\pi i f_f t + \phi_f^i\right) \tag{7.2}$$

where $A_f = d_f f_f$ is the amplitude, $\delta(\cdot)$ is the unit impulse function, ϕ_f^i is the phase of i-th harmonic, and N_1 is the number of harmonics induced by impulse series. The fault frequency can be one of bearing characteristic frequencies of inner-race fault, outer-race fault, or rolling element fault depending on the location of fault.

7.1.2 Non-uniform load distribution

The following Stribeck equation explains the vibration caused by non-uniform load distribution around the rolling element bearing (Harris, 1984).

$$x_l(\theta) = P_{\max}\left[1 - \frac{1}{2\epsilon}\left(1 - \cos\theta\right)\right]^m \tag{7.3}$$

where $P_{\max}$ is the maximum contact pressure at a circumferential angle θ around shaft as shown in Fig. 7.1, ϵ is the load distribution factor, and $m = 1.5$ for ball bearings and

$m = 10/9$ for other rolling element bearings. The terms P_{max}, θ_{max} and ϵ are dependent on the bearing clearance and the applied load. For bearing with positive clearance, $\epsilon < 0.5$ and $\theta_{\text{max}} < \pi/2$ (McFadden and Smith, 1984). In general, the inner race of the bearing is attached to a shaft rotating at a constant frequency f_s. Therefore, one can rewrite Eq. (7.3) in terms of time and rotational frequency of shaft by using $\theta(t) = 2\pi f_s t$ in the following format:

$$x_l(t) = \begin{cases} P_{\text{max}}\left[1 - \dfrac{1}{2\epsilon}\Big(1 - \cos\theta(t)\Big)\right]^m & \text{, for } |\theta(t)| < \theta_{\text{max}} \\ 0 & \text{, elsewhere.} \end{cases} \tag{7.4}$$

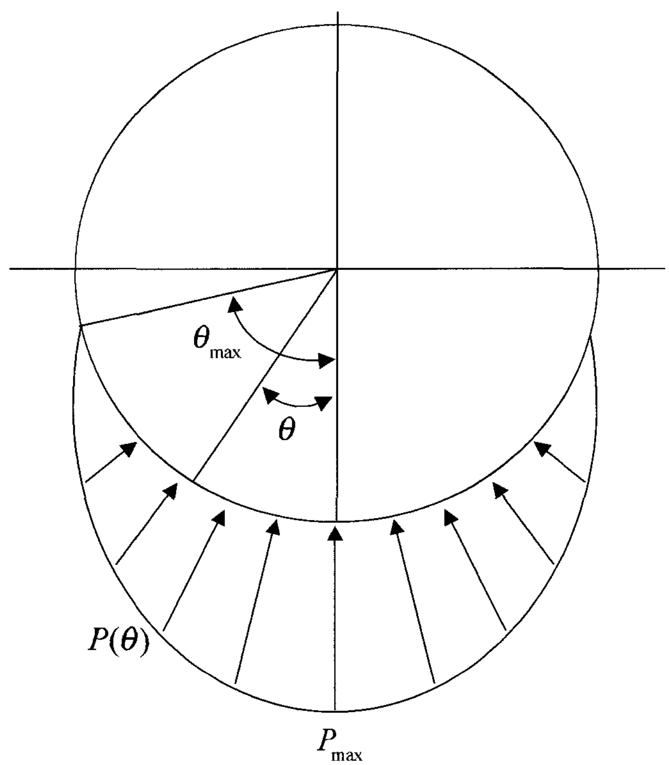


Figure 7.1. Load distribution in a bearing under a radial load

7.1.3 Bearing-induced vibration

The high-frequency resonance vibration with decay can be modeled as a superposition of all vibration modes (Wang and Kootsookos, 1998).

$$x_b(t) = \sum_{j=1}^{N_2} A_b^j e^{-\alpha_b^j t} \cos\left(2\pi f_b^j t + \phi_b^j\right) \tag{7.5}$$

where A_b^j, f_b^j, ϕ_b^j and α_b^j are the amplitude, the resonance frequency, the phase and the damping factor of the exponential decay of j-th vibration mode of the bearing, respectively, and N_2 is the modal order of bearing vibration.

7.2 Complexity as a measure of fault severity: the rationale

The complexity of a signal increases when more frequency components exist. For mechanical parts such as gears and bearings, contact pressure between two mating parts changes when a fault exists. The contact pressure variation results in amplitude and frequency modulations and hence more frequency peaks will appear in the frequency domain with their harmonics (Randall, 1982). Therefore, the complexity of a signal increases as a fault develops because a larger fault intensifies variation of contact pressure between two mating parts.

The signal models of different vibration sources described in the previous section are adopted to evaluate complexity of inner- and outer-race fault signals. This study focuses on a bearing with a fixed outer race. In this case, the rotating speed of the inner race is the same as that of the shaft. It is further assumed: a) the initial phase angles are zero, and b) the modulation frequencies caused by non-uniform load and the fault frequencies of impulses are lower than the resonance frequencies of bearing, i.e.,

$$f_b^j \gg f_f > f_s \qquad (j = 1, 2, \ldots, N_2) \tag{7.6}$$

where f_b^j, f_f and f_s are the resonance frequency of bearing, the frequency of impulses generated by fault and the shaft rotating frequency, respectively.

7.2.1 Outer-race fault

For the outer-race fault, the non-uniform load distribution does not have a strong effect on the outer-race fault signal since the relative location of the fault is not affected by the non-uniform load distribution. Therefore, the non-uniform load term in Eq. (7.1) is negligible, i.e., $x_l(t) \approx 1$. Then, the signal model of outer-race fault is simplified as

$$\begin{aligned} x_{OR}(t) &= x_f(t) \cdot x_b(t) + x_n(t) \\ &= A_f \sum_{i=1}^{N_1} \cos\left(2\pi i f_f t + \phi_f^i\right) \cdot \sum_{j=1}^{N_2} A_b^j e^{-\alpha_b^j t} \cos\left(2\pi f_b^j t + \phi_b^j\right) + n(t). \end{aligned} \tag{7.7}$$

Based on the assumptions mentioned above, Eq. (7.7) can be written as (See the Appendix A.7 for derivation.)

$$\begin{aligned} x_{OR}(t) &= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} A_f A_b^j e^{-\alpha_b^j t} \cos\left(2\pi i f_f t\right) \cdot \cos\left(2\pi f_b^j t\right) + n(t) \\ &= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} A_f A_b^j e^{-\alpha_b^j t} \cdot \frac{1}{2}\left(X_1 + X_2\right) + n(t) \end{aligned} \tag{7.8}$$

where

$$\begin{aligned} X_1 &= \cos\left[2\pi\left(f_b^j + i f_f\right) t\right] = \cos\left[2\pi f_b^j \left(1 + \frac{i f_f}{f_b^j}\right) t\right] \\ X_2 &= \cos\left[2\pi\left(f_b^j - i f_f\right) t\right] = \cos\left[2\pi f_b^j \left(1 - \frac{i f_f}{f_b^j}\right) t\right]. \end{aligned}$$

It should be noted that the two terms, X_1 and X_2, are expressed with the frequency modulation effect.

The complexity of the outer-race fault is explained as follows. Consider one of rolling elements and the bearing outer race. For a healthy bearing, the contact pressure between the ball and the outer race of bearing does not change much with respect to the angular position of ball. However, for a faulty bearing, the contact pressure increases when the ball hits the fault on the surface of the outer race as the contact area becomes smaller because of the fault. Hence, the contact pressure becomes variable when a fault exists on the outer race of bearing. The fluctuating contact pressure results in frequency variation that leads to the frequency modulation effect (Randall, 1982) as shown in Eq. (7.8). As the size of fault on the outer race increases, fluctuation of the contact pressure intensifies, which reinforces the frequency modulation effect. The stronger frequency modulation effect is evidenced by the increased number of frequency peaks in the frequency domain such as harmonics. The larger number of frequency peaks renders higher complexity. Therefore, one can expect that, for the outer-race single-point defect, the complexity will increase as the defect develops.

7.2.2 Inner-race fault

The radial load applied to the rotating shaft causes non-uniform load distribution around the inner race of the bearing as illustrated in Fig. 7.1. Non-uniform load distribution produces low-frequency amplitude modulation with a frequency of f_s as shown in Eq. (7.4). For simplicity, consider a bearing with $\epsilon = 0.5$ and $\theta_{\max} = \pi/2$. The non-uniform load distribution term for $|\theta(t)| < \theta_{\max}$ is accordingly

$$\begin{aligned} x_l(t) &= \lim_{\epsilon \to 0.5} P_{\max}\left[1 - \frac{1}{2\epsilon}\Big(1 - \cos\left(2\pi f_s t\right)\Big)\right]^m \\ &= P_{\max}\Big[\cos\left(2\pi f_s t\right)\Big]^m \end{aligned} \tag{7.9}$$

where $m = 1.5$ (for a ball bearing) is negligible since it does not have any effect on the frequency component of $x_l(t)$. Eq. (7.9) can be rewritten as follows (Tandon and Choudhury, 1997; Wang and Kootsookos, 1998)

$$x_l(t) \approx \sum_{k=1}^{N_3} A_l^k \cos\left(2\pi k f_s t + \phi_l^k\right) \tag{7.10}$$

where A_l^k, ϕ_l^k and N_3 are the amplitude, the phase of k-th harmonic and the number of harmonics, respectively. Then, the signal model for inner-race fault is

$$\begin{aligned} x_{IR}(t) &= x_f(t) \cdot x_b(t) \cdot x_l(t) + x_n(t) \\ &= A_f \sum_{i=1}^{N_1} \cos\left(2\pi i f_f t + \phi_f^i\right) \cdot \sum_{j=1}^{N_2} A_b^j e^{-\alpha_b^j t} \cos\left(2\pi f_b^j t + \phi_b^j\right) \\ &\quad \cdot \sum_{k=1}^{N_3} A_l^k \cos\left(2\pi k f_s t + \phi_l^k\right) + n(t). \end{aligned} \tag{7.11}$$

The above equation can be rewritten as (See the Appendix A.8 for derivation.)

$$
\begin{aligned}
x_{IR}(t) &= \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{k=1}^{N_3} A_f A_l^k A_b^j e^{-\alpha_b^j t} \cdot \cos\left(2\pi i f_f t\right) \cdot \cos\left(2\pi f_b^j t\right) \cdot \cos\left(2\pi k f_s t\right) \\
&\quad + n(t) \qquad (7.12) \\
&= \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{k=1}^{N_3} A_f A_l^k A_b^j e^{-\alpha_b^j t} \cdot \frac{1}{4}\left(X_3 + X_4 + X_5 + X_6\right) + n(t)
\end{aligned}
$$

where

$$
\begin{aligned}
X_3 &= \cos\left[2\pi\left(f_b^j + i f_f + k f_s\right) t\right] \\
X_4 &= \cos\left[2\pi\left(f_b^j - i f_f - k f_s\right) t\right] \\
X_5 &= \cos\left[2\pi\left(f_b^j - i f_f + k f_s\right) t\right] \\
X_6 &= \cos\left[2\pi\left(f_b^j + i f_f - k f_s\right) t\right].
\end{aligned}
$$

The inner-race fault signal creates three major frequency components as shown in Eq. (7.12).

According to Howard (1994), there are three dominant factors in impact action, i.e., bearing geometry, rotational speed and fault size. As the effects of the geometry and the rotational speed are constant for the same bearing under stationary operating conditions, the fault size is the most important factor for the impact action. Consider each term in Eq. (7.11). As fault size increases, the impulse associated with $x_f(t)$ gets dull since the edges of the fault are flattened by repeated impacts (Rubini and Meneghetti, 2001). The dull edges can no longer produce sharp impulses and the strength of sharp impulses decreases as a fault grows. Such a phenomenon has been confirmed by the shrinking peak magnitudes in the frequency spectrum as the inner race fault expands (Rubini and Meneghetti, 2001). Therefore, the contribution of $x_f(t)$, in terms of amplitude, in Eq. (7.11) diminishes as the fault size increases.

The resonance frequency is constant for the same bearing since it is the system's unique characteristic. Therefore, the bearing resonance vibration term, $x_b(t)$, is considered to have less contribution in Eq. (7.11) as fault size increases.

The vibration term corresponding to the non-uniform load distribution, $x_l(t)$, creates low-frequency amplitude modulation. The amplitude of this term is dependent on the maximum contact pressure $P_{\max}$. For a ball bearing, the surfaces of a ball and the inner race contact at a single point, which is called *point contact.* The contact point expands to an ellipse when the load is applied (Harris, 1984; Brandlein et al., 1999). The maximum contact pressure is located at the center of the ellipse and its magnitude is (Brandlein et al., 1999)

$$P_{\max} = 1.5\frac{F}{A} \tag{7.13}$$

where F is the force by which the components (ball and inner race) are pressed against each other and A is the nominal contact area without defect. If we approximate the defect

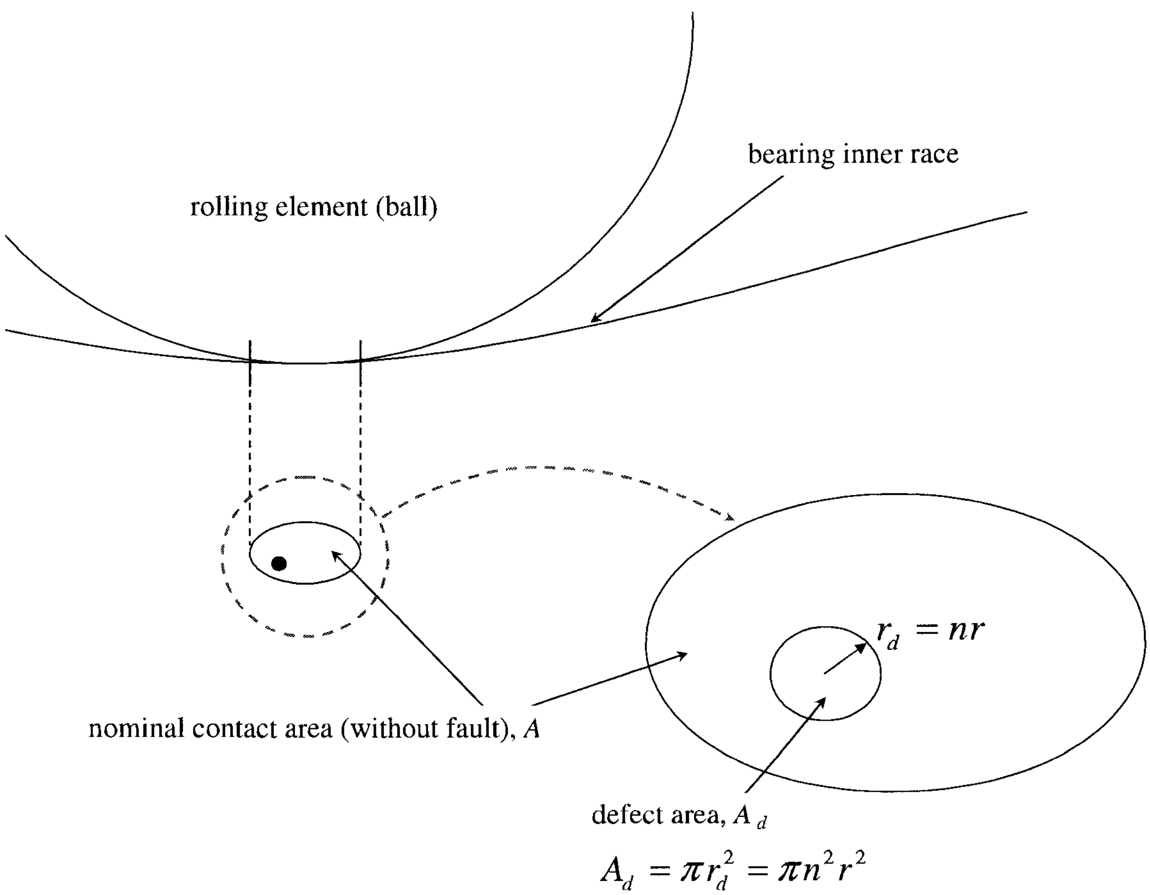


Figure 7.2. Contact pressure between ball and race

area (the cavity caused by a fault) using a circle with radius r, the area corresponding to the fault would swell in proportion to n^2r^2, as the fault size (diameter) increases to $nr\,(n > 1)$ as shown in Fig. 7.2. Then, the maximum contact pressure of defective bearing is

$$\begin{aligned} P_{\max,d} &= 1.5\frac{F}{A - A_d} \\ A_d &= \pi r_d^2 = \pi n^2 r^2 \end{aligned} \tag{7.14}$$

Therefore, $P_{\max,d} > P_{\max}$ and the actual area of contact between the ball and the inner race decreases as the fault size grows. This leads to rapid increase of the maximum contact pressure, as suggested by Eq. (7.14).

7.2.3 Lempel-Ziv complexity

Lempel and Ziv introduced an easily calculable measure of complexity of finite sequences, which adopted two basic processes: copy and insert (Lempel and Ziv, 1976). Fig. 7.3 explains the procedure of calculating Lempel-Ziv complexity with finite sequence with length N. Consider a string $S_N = \{s_1s_2\cdots s_N\}$. Assume that a string up to $s_r\,(1 < r < N)$ can be reconstructed by simply copying and inserting some of the existing vocabulary of $S_{v,r-1} = \{s_1s_2\cdots s_v\}\ (v < r)$. To check the rest string $S_{N-r} = \{s_{r+1}\ \cdots\ s_N\}$ can be re-produced by the same approach. Lempel and Ziv introduced the following steps:

Step 1. Take $Q_{r+1} = \{s_{r+1}\}$ and ask if this term belongs to the vocabulary $S_{v,r} = \{S_{v,r-1}s_r\}$. If so, string $Q_{r+1} = \{s_{r+1}\}$ is a simple repetition of the portion of existing string $S_{v,r}$ (or a simple "copy" of existing vocabulary can restore it).

Step 2. Read the next character and take $Q_{r+2} = \{s_{r+1}s_{r+2}\}$. Check if $Q_{r+2} = \{s_{r+1}s_{r+2}\}$ belongs to $S_{v,r+1} = \{S_{v,r-1}s_rs_{r+1}\}$ (obtained by augmenting S_v with s_{r+1}).

Step 3. If the term Q_{r+2} does not belong to $S_{v,r+1}$, increase the complexity by one, i.e., $c_N(r+2) = c_N(r+1) + 1$, nullify $Q_{r+2} = \{\,\}$, read the next character and take $Q_{r+3} = \{s_{r+3}\}$.

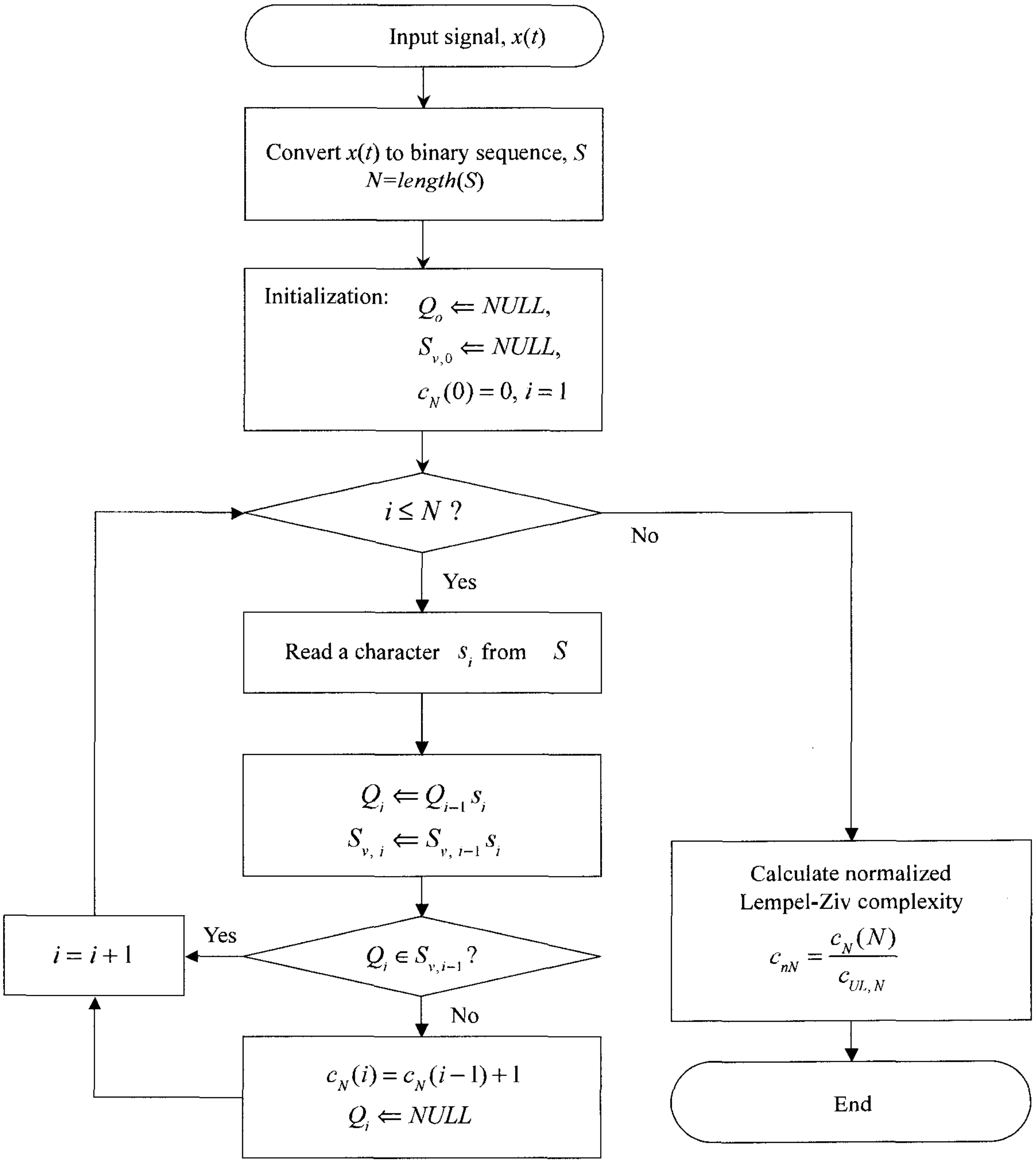


Figure 7.3. Block diagram for calculating Lempel-Ziv complexity

Repeat the above procedure until $S_N = \{s_1 s_2 \cdots s_N\}$ is completely covered. Then, the resulting $c_N(N)$ is a complexity measure of a given string.

Consider an example of a finite string $S = \{0001101001000101\}$ with length $N = 16$, the above procedure can be illustrated as follows:

Initialization: $i = 0$, $S_{v,0} = \{\ \}$, $Q_0 = \{\ \}$, $c_N(0) = 0$.

$i = 1$, $S_{v,1} = \{0\}$, $Q_1 = \{0\}$, $Q_1 \notin S_{v,0} \rightarrow c_N(1) = 1$, $Q_1 = \{\ \}$.

$i = 2$, $S_{v,2} = \{00\}$, $Q_2 = \{0\}$, $Q_2 \in S_{v,1} \rightarrow c_N(2) = 1$.

$i = 3$, $S_{v,3} = \{000\}$, $Q_3 = \{00\}$, $Q_3 \in S_{v,2} \rightarrow c_N(3) = 1$.

$i = 4$, $S_{v,4} = \{0001\}$, $Q_4 = \{001\}$, $Q_4 \notin S_{v,3} \rightarrow c_N(4) = 2$, $Q_4 = \{\ \}$.

$i = 5$, $S_{v,5} = \{00011\}$, $Q_5 = \{1\}$, $Q_5 \in S_{v,4} \rightarrow c_N(5) = 2$.

$i = 6$, $S_{v,6} = \{000110\}$, $Q_6 = \{10\}$, $Q_6 \notin S_{v,5} \rightarrow c_N(6) = 3$, $Q_6 = \{\ \}$.

$i = 7$, $S_{v,7} = \{0001101\}$, $Q_7 = \{1\}$, $Q_7 \in S_{v,6} \rightarrow c_N(7) = 3$.

$i = 8$, $S_{v,8} = \{00011010\}$, $Q_8 = \{10\}$, $Q_8 \in S_{v,7} \rightarrow c_N(8) = 3$.

$i = 9$, $S_{v,9} = \{000110100\}$, $Q_9 = \{100\}$, $Q_9 \notin S_{v,8} \rightarrow c_N(9) = 4$, $Q_9 = \{\ \}$.

$i = 10$, $S_{v,10} = \{0001101001\}$, $Q_{10} = \{1\}$, $Q_{10} \in S_{v,9} \rightarrow c_N(10) = 4$.

$i = 11$, $S_{v,11} = \{00011010010\}$, $Q_{11} = \{10\}$, $Q_{11} \in S_{v,10} \rightarrow c_N(11) = 4$.

$i = 12$, $S_{v,12} = \{000110100100\}$, $Q_{12} = \{100\}$, $Q_{12} \in S_{v,11} \rightarrow c_N(12) = 4$.

$i = 13$, $S_{v,13} = \{0001101001000\}$, $Q_{13} = \{1000\}$, $Q_{13} \notin S_{v,12}$
$\rightarrow c_N(13) = 5$, $Q = \{\ \}$.

$i = 14$, $S_{v,14} = \{00011010010001\}$, $Q_{14} = \{1\}$, $Q_{14} \in S_{v,13} \rightarrow c_N(14) = 5$.

$i = 15$, $S_{v,15} = \{000110100100010\}$, $Q_{15} = \{10\}$, $Q_{15} \in S_{v,14} \rightarrow c_N(15) = 5$.

$i = 16$, $S_{v,16} = \{0001101001000101\}$, $Q_{16} = \{101\}$, $Q_{16} \in S_{v,15} \rightarrow c_N(16) = 5$.

Now, the string shown in this example has a complexity of $c_N(16) = 5$.

The complexity obtained above is equal to the number of nullifications of Q. This is, to certain extent, affected by the length of the string, or the number of data samples N. For general applications, a complexity measure should be independent of the number of data samples N. To find such a complexity measure, Lempel and Ziv (1976) suggested a normalized complexity measure after their names c_{nN} defined by:

$$0 \leq c_{nN} = \frac{c_N(N)}{c_{UL,N}} \leq 1 \tag{7.15}$$

where

$$c_{UL,N} = \lim_{N\to\infty} c_N(N) = \lim_{N\to\infty} \frac{N}{(1-\alpha)\log_* N} \approx \frac{N}{\log_* N} \tag{7.16}$$

where "$*$" is the size of the alphabets (or distinct elements) used in the string S (for binary string $*$=2), $\alpha \to 0$ as $N \to \infty$ (Lempel and Ziv, 1976; Cover and Thomas, 1991).

However, in spite of their effort, the "normalized" c_{nN} is still affected by N if N is not sufficiently large. This can be illustrated as follows. To examine the effect of the number of data samples on the normalized complexity, a series of binary strings of different length is randomly generated based on the Gaussian distribution. Their normalized complexities are calculated following the procedure shown in Fig. 7.3 and are plotted in Fig. 7.4. Since the data is random, the normalized complexity approaches 1.0 as the string length increases. The normalized complexity drops below 1.05 for $N \geq 3,600$ which is very close to the upper bound, 1, of the "normalized" complexity. This procedure therefore could be used for selecting the number of data samples to obtain a complexity index ranging roughly between [0, 1].

7.3 Calculating the complexity of bearing fault signal

Yan and Gao (2004) applied Lempel-Ziv complexity to the cylindrical roller bearing and obtained different complexity values for different fault sizes. According to their study, the normalized Lempel-Ziv complexity approaches the upper limit, around 1.0, as fault develops. They calculated the Lempel-Ziv complexity from the measured acceleration signal. However, the measured signal is contaminated by noise. This may increase the complexity value and it could lead to inaccurate diagnosis. As shown in Fig. 7.4, the pure random noise has a normalized complexity value around 1.0. According to the work by Yan and Gao (2004), the Lempel-Ziv complexity of the signal of severe outer-race

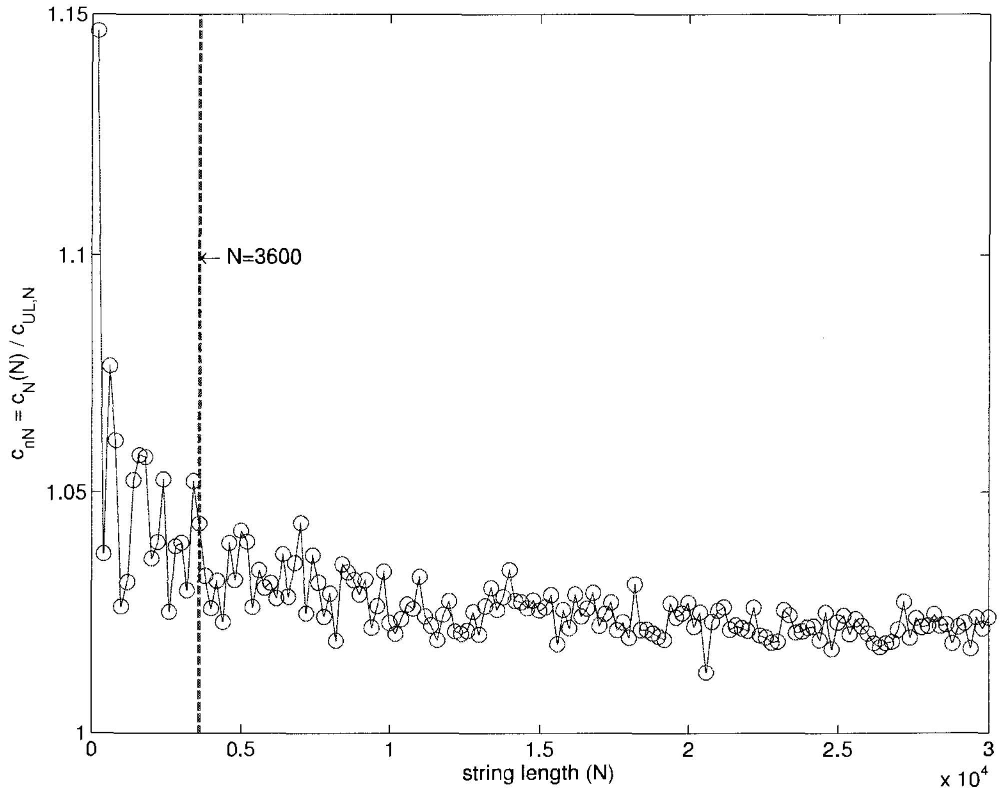


Figure 7.4. The effect of data length on the normalized complexity for randomly generated data

fault is close to 1.0. Therefore, it is difficult to distinguish the sources (noise or fault) of the complexity value. To alleviate this difficulty, the following procedure is proposed, which is further detailed in Fig. 7.5:

Step 1. Apply continuous wavelet transform (CWT) to the measured signal.

Step 2. Obtain the energy and kurtosis in each scale level and find the best scale level by maximizing energy and kurtosis distributions over scales used in step 1.

Step 3. Calculate normalized complexity of the wavelet coefficients at the best scale level.

The rationale of finding the best scale in step 2 is to extract the fault feature and exclude the interferences coming from other frequency ranges. Then the normalized complexity is calculated only based on the extracted fault feature in step 3. In this way, much of the noise components have been eliminated and hence the ambiguity between noise and severe faults can be cleared.

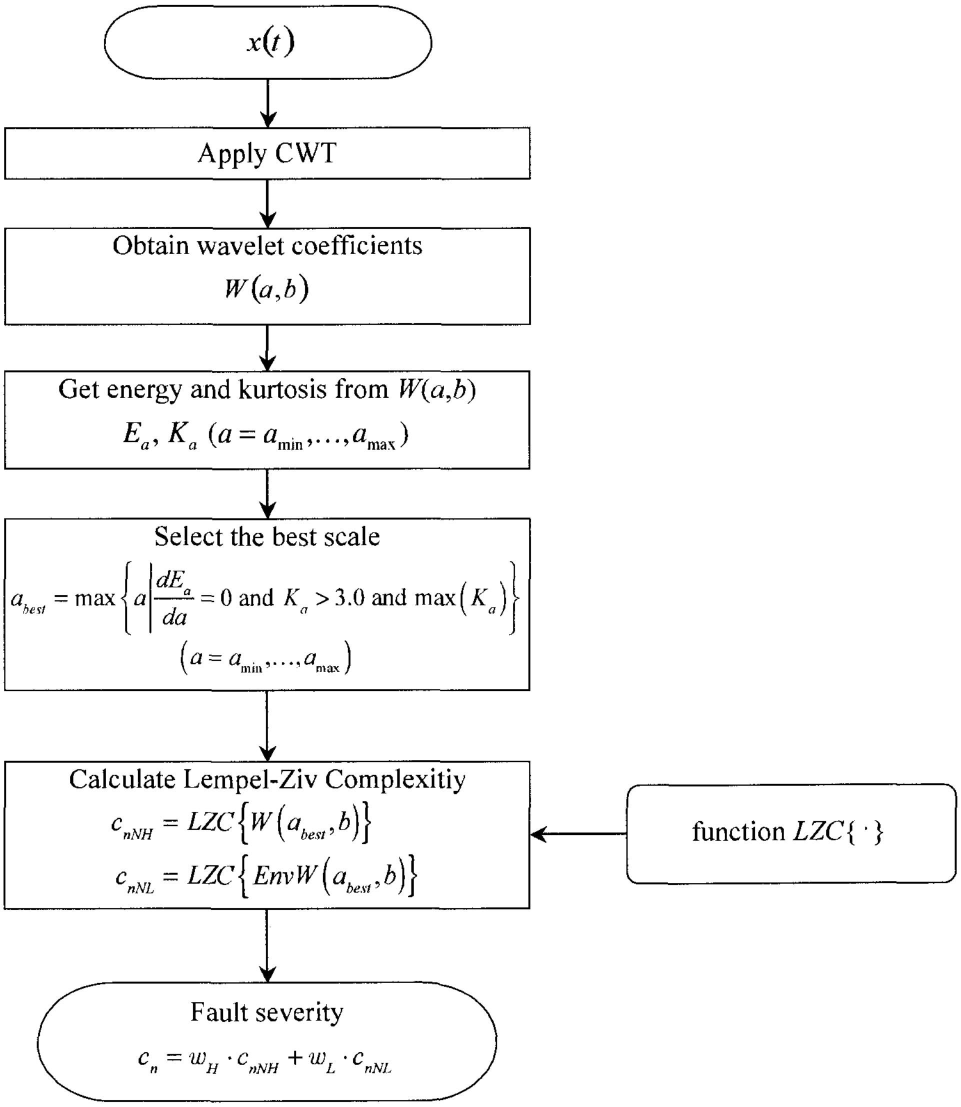


Figure 7.5. Block diagram of Lempel-Ziv complexity as a measure of fault severity

7.3.1 Continuous wavelet transform (CWT)

As shown in Fig. 7.5, the procedure starts with mapping the measured vibration signal into the scale-time domain by CWT. The CWT is selected since it provides finer resolution in the scale-time plane than discrete wavelet transform (DWT) does. The CWT is expressed as shown in Chapter 3 (from Eq. (3.2)):

$$W(a,b) = \int x(t)w^*_{a,b}(t)dt \quad (a = a_{\min}, \dots, a_{\max})$$

where $a_{\min}$ and $a_{\max}$ are the lowest and the highest scale levels of interest, asterisk (*) indicates complex conjugate and $w_{a,b}(t)$ is daughter wavelet defined as (from Eq. (3.3))

$$w_{a,b} = \frac{1}{|a|^{1/2}} w\left(\frac{t-b}{a}\right)$$

where a is dilation (scale) factor and b is translation factor in time. The factor $|a|^{-1/2}$ is used to ensure energy preservation (Goupillaud et al., 1984). In this study, the Morlet wavelet as shown below is used as a mother wavelet function because of its similarity to the impulsive mechanical fault feature in shape (Lin and Qu, 2000; Lin et al., 2004) (from Eq. (3.4)):

$$w(t) = \exp\left(-\frac{\gamma^2 t^2}{2}\right)\cos(\omega_0 t)$$

with $\gamma = 1$ and $\omega_0 = 5$. A measured signal normally contains many unnecessary signal components with Gaussian behavior such as noise and interfering vibration generated by other machine elements. As mentioned earlier, the normalized complexity of Gaussian noise approaches the upper limit of 1.0. Therefore, the complexity of the signal contaminated by noise does not truly represent the severity of fault features. By applying wavelet transform, one can select a scale level that reveals fault features with less interference of unwanted signal components. This ensures relevant and reliable complexity values for the fault features under noisy environment.

7.3.2 Selection of the best scale

To select the best scale level, two measures, i.e., peak energy and kurtosis, are considered. The reasons for considering the two measures, as stated in Chapter 3, are: a) the energy level is a measure of the shape similarity between the measured signal and the daughter wavelet, b) the energy criterion cannot be used alone in this study with CWT because higher energy level is also corresponding to higher scale level, and c) kurtosis is a good indicator of impulsiveness. The use of two criteria would take care of both the signal-wavelet similarity and the fault signature impulsiveness. As presented in Chapter 3, the energy of wavelet coefficients for a scale level can be obtained as (from Eq. (3.7))

$$E_a = \sum_b \{W(a,b)\}^2 \qquad a = a_{\min},\, a_{\min}+1,\, \ldots,\, a_{\max}-1,\, a_{\max}$$

where E_a is the energy of wavelet coefficients in scale a. A scale with peak energy value is considered as a candidate for the best scale. The best scale, as detailed in Chapter 3, can be found by (from Eq. (3.10))

$$a_{best} = \left\{ a \middle| \frac{dE_a}{da} = 0 \text{ and } \max(K_a) \text{ and } K_a > 3.0 \right\} \qquad a \in [a_{\min}, \ldots, a_{\max}]$$

where K_a is the kurtosis at scale a and a_{best} is the selected best scale. The scale selected by Eq. (3.10) is considered to contain the signal nature closest to the impulses produced by the mechanical fault. This selection method can thus reduce the side effects of noise and un-related signals.

7.3.3 Calculation of the Lempel-Ziv complexity

The Lempel-Ziv complexity values are calculated for the wavelet transformed signal and its envelope at the best scale a_{best}. As shown with the signal models in the previous section, the bearing fault signal consists of higher-carrier and lower-modulating frequencies. Therefore, considering the Lempel-Ziv complexity for both high- and low-frequency signals is suitable for measuring bearing fault severity. The overall procedure is as follow:

Step 1. Obtain the high- and low-frequency signals from the wavelet coefficients at the best scale.

Step 2. Convert the high- and low-frequency signals into binary sequences.

Step 3. Calculate the Lempel-Ziv complexity of the binary sequences based on the procedure shown in Fig. 7.3.

More details are given as follows. In step 1, the wavelet coefficients at the best scale a_{best} are demodulated to obtain the high-frequency carrier signal $W(a_{best}, b)$ and low-frequency envelope signal $EnvW(a_{best}, b)$. The Hilbert transform is applied to the wavelet coefficients to find envelope signal of the wavelet coefficients, i.e., the low-frequency signal.

In the second step, both high- and low-frequency signals are converted into binary sequences. Any values greater than or equal to the mean value of the given signal are replaced with one and others with zero.

In step 3, the normalized Lempel-Ziv complexity values for the high- and low-frequency signals are calculated by:

$$\begin{aligned} c_{nNH} &= \left\{ \frac{c_N(N)}{c_{UL,N}} \right\}_{\text{high frequency}} = LZC\left\{W(a_{best},\, b)\right\} \\ c_{nNL} &= \left\{ \frac{c_N(N)}{c_{UL,N}} \right\}_{\text{low frequency}} = LZC\left\{EnvW(a_{best},\, b)\right\} \end{aligned} \tag{7.17}$$

where $EnvW(a_{best},\, b)$ and $LZC\{\cdot\}$ indicate the envelope signal of wavelet coefficients at the best scale a_{best} and function calculating Lempel-Ziv complexity, respectively. An indicator of fault severity is then obtained based on the two normalized complexity values as follows:

$$c_n = w_H \cdot c_{nNH} + w_L \cdot c_{nNL} \tag{7.18}$$

where w_H and w_L are the weights of high-frequency carrier signal and low-frequency envelope signal. Equal weight is assigned to each dominant signal components. The outer-race fault signal, as shown in Eqs. (7.7) and (7.8), contains two dominant signal components, i.e., bearing resonance and impulses generated by fault as shown in Eqs. (7.7) and (7.8). The two signal components are corresponding to the high-frequency carrier signal and low-frequency envelope signal respectively and each receives 50% of the weight. However, as suggested by Eqs. (7.11) and (7.12), there exist three dominant signal components in the inner-race signal contributed by bearing resonance, fault impulse and non-uniform load, respectively. Since the resonance frequency is usually much higher than the frequencies of the non-uniform load and fault impulses, it acts as the carrier frequency and the other two are classified as low frequencies. In summary, the weights are allocated as follows

$$\begin{cases} w_H &= 1/2 \\ w_L &= 1/2 \end{cases} \quad \text{for outer race}$$

$$\begin{cases} w_H &= 1/3 \\ w_L &= 2/3 \end{cases} \quad \text{for inner race}$$

7.4 Application to bearing fault data

The proposed method is applied to the bearing fault signal. The bearing fault signals are obtained from the Case Western Reserve University (2007). Single-point defects of different sizes are created on the surfaces of inner and outer race of a deep groove ball bearing (SKF6205-2RS JEM). The fault sizes are 0.007, 0.014, 0.021 inches in diameter and 0.011 inches in depth. The sampling frequency is 12,000Hz and 12,000 data samples are used. The calculated Lempel-Ziv complexity values are summarized in Fig. 7.6. The complexity of inner-race fault decreases as fault size increases because of the amplitude modulation effect of the non-uniform load distribution. As a fault grows, the edge of the fault cavity is smoothened out, leading to the decrease in the strength of sharp impulses. As such, the contribution of the associated frequency component to the low frequency envelope signal diminishes. On the other hand, the actual contact area is reduced due to the growing fault area, which in turn strengthens the non-uniform load distribution effect. Because of the above, the vibration caused by the non-uniform load distribution is becoming a dominant contributor to the low-frequency envelope signal. The diminished contribution of the fault impulse and the increased effect of the non-uniform load distribution lead to the lower complexity which means more order or less randomness (Kaspar and Schuster, 1987; Wolfman, 1984, 1985). The complexity of outer-race fault, on the other hand, shows an increasing trend. This opposite trend is also validated by the work of Rubini and Meneghetti (2001). The outer-race fault for 0.014-inch diameter is not considered, as the data is not available. Fig. 7.6 clearly indicates that the Lempel-Ziv complexity is proportional (for outer-race fault) or inversely proportional (for inner-race

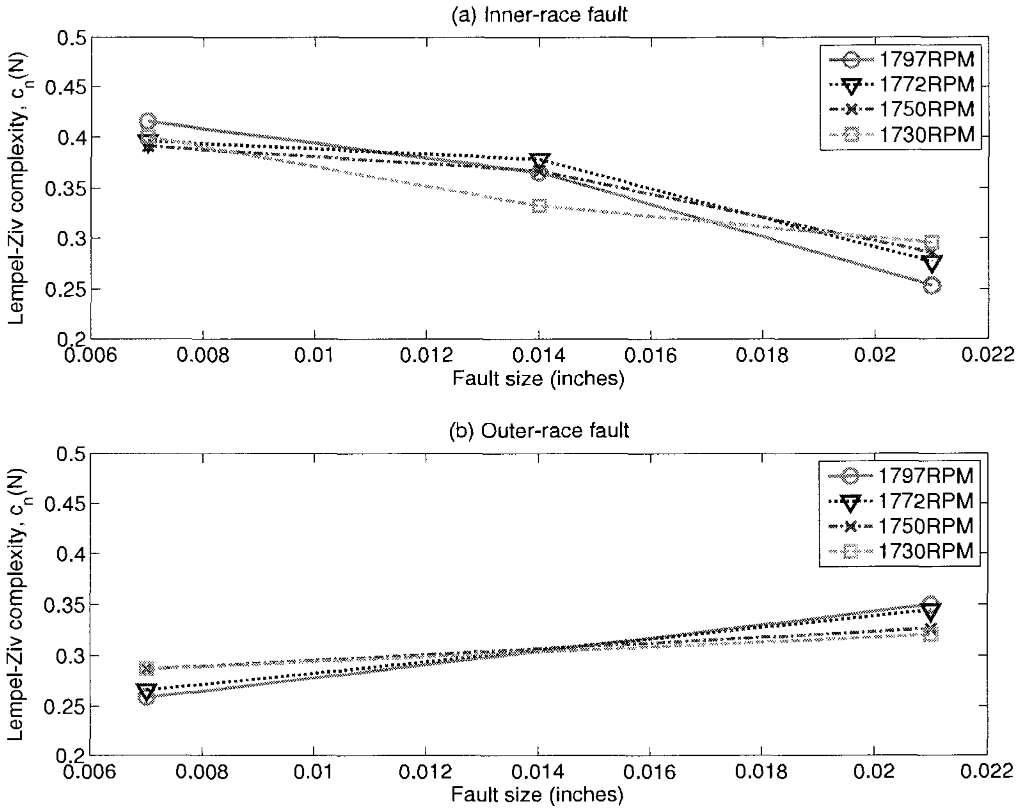


Figure 7.6. Lempel-Ziv complexity vs. fault size: (a) inner-race fault and (b) outer-race fault

fault) to fault size (severity) for all rotational speeds. The complexity differences for different operating speeds are relatively small. Since the Lempel-Ziv complexity is a normalized value between zero and one, it can be used as a non-dimensional severity measure of bearing faults. Unlike most other methods (e.g., the spectral analysis), the Lempel-Ziv complexity based fault severity measure does not require comparison with reference data to evaluate the fault severity. This is the principal advantage of the Lempel-Ziv complexity.

A closer look at the plots in Fig. 7.6 indicates that the complexity value is somewhat affected by shaft rotating speed. This is in agreement with the observation of Yan and Gao (2004). To further enhance the proposed method, the speed factor may be incorporated in future.

7.5 Discussion

Measuring fault severity is one of the main challenges in fault diagnosis. So far, the severity has been mostly evaluated based on comparison methods, e.g., comparing faulty signal with healthy signal. However, the magnitudes of the signals are dependent on many factors such as bearing size, load, speed, and working conditions. Hence, the same signal magnitude does not necessarily indicate the same fault severity and the comparison based methods may not provide a reliable indication of fault severity. To reliably measure fault severity, many different reference data would have to be collected for different conditions if comparison based methods are used. However, such data are often unavailable particularly for new bearings and new working conditions. Even if the data can be acquired, it would be very tedious and time-consuming to collect, store, maintain and retrieve them especially for variable and unpredictable working conditions. To avoid this difficulty, the Lempel-Ziv complexity, a non-dimensional measure is mainly affected by fault size (severity) and should be much less susceptible to other factors. As a result, the proposed severity measure can be used for ball bearings of different fault sizes under different operation environments without using reference data. In addition, it is important to point out that the Lempel-Ziv complexity is calculated based on the best CWT scale level, and hence the noise and many un-related signal components have been screened out. This makes the diagnosis result much more reliable.

Chapter 8

Conclusions and Future Work Directions

8.1 Conclusions

Fault diagnosis can be classified into three categories: fault detection, fault isolation, and fault identification. Associated methodologies and tools are developed by examining three related topics: fault feature extraction, single-channel signal separation, and fault severity. Accordingly, the thesis is divided into three parts. In Part I, fault feature extraction methods are developed for three types of signals, i.e., periodic signal, non-stationary signal, and oil-debris signal. A single-channel signal separation method is proposed for fault isolation in Part II. Finally, a fault severity measure based on the Lempel-Ziv complexity and continuous wavelet transform (CWT) is devised for rolling element bearings in Part III. The conclusions are summarized as follows.

1. Correlation-based feature extraction

A new method is proposed to extract periodic fault features based on the identification of the time instants of fault feature occurrences. An autocorrelation coefficient function of the wavelet-transformed signal at the best scale level is developed to reveal the time instants of fault occurrences. The best scale level of CWT is selected by considering both energy and kurtosis. Though the energy has been commonly used to select the best node in wavelet packet analysis, it does not guarantee that the scale with the maximum energy always reflects the nature of impulsive fault features generated by mechanical fault. By adopting kurtosis value as a supplementary measure, the best scale level containing information on the fault features can be selected more effectively. It is also shown that considering both energy and kurtosis can provide information on the machine conditions. The simulation studies indicate that the proposed method can significantly reduce spurious impulses, one of the major obstacles to accurate mechanical fault detection. The proposed method is experimentally validated for gear fault detection in our lab. The results have demonstrated that the proposed method is suitable for mechanical fault detection of rotating machinery as it can accurately identify periodic fault features with minimum spurious impulse interference.

2. K-Hybrid thresholding de-noising

The kurtosis based hybrid thresholding method is especially suitable for on-line de-noising mechanical fault signals due to its data-driven nature and its ability to detect outliers. The four input signal zones divided by the hybrid thresholding method provides a better representation of the signal makeup and hence leads to better de-noising results compared to the simple keep-or-remove and shrink-or-remove actions taken by the soft- and hard-thresholding rules. In addition, the four zones adapt themselves to the change of the kurtosis value and hence are able to reflect the signal dynamics. An analysis also indicates that the MSE should not be used alone for evaluating the performance of a thresholding method since it does not take into account fake (or spurious) impulses. The false identification power (P_{false}) is hence proposed as part of a combined performance measure. The simulation study has shown clear superiority of K-hybrid over soft, hard, and BayesShrink thresholding rules in terms of the combined evaluation criterion. It also performs better than the MAP thresholding rule for heavily contaminated signals. The application of K-hybrid has been demonstrated by de-noising bearing signals. It is important to note that the application of K-hybrid has replaced the bandpass filter and hence avoided the use of resonance frequency required to specify the bandpass filter in the envelope analysis.

3. ODM sensor signal extraction

A new feature extraction method was proposed for in-line ODM sensor signal analysis. Two detectors, a particle signature extractor and an inflexion-point locator, have been developed based on the generalized differentiation. The particle signature extractor is used to identify any signature that resembles that of a metal particle. As the particle signals feature three inflexion points, the inflexion-point locator is used to help confirm the authenticity of the signature. Because of its low-pass filter property, the signature extractor can extract signatures under a noisy environment. Unlike the existing approaches, the two detectors do not require threshold values to extract oil-debris signatures, which alleviates the difficulty in specifying threshold values for various applications. The experimental tests have shown that the two detectors can reliably extract particle signature of small amplitudes that are heavily clouded by noise. The ability to detect small particles would facilitate early detection of incipient flaws and hence help protect a mechanical system from disastrous damage. In addition, the proposed method has shown satisfactory results without applying additional filters or any other de-noising measures. This would simplify the signal processing effort in real applications.

4. Single-channel signal separation

A single-channel signal separation method has been developed using continuous wavelet transform and Fourier transform. The proposed method is suitable for rotating machinery fault signal separation and detection because: (1) the number of sources can be automatically and dynamically determined, which can effectively handle variable and unpredictable number of faults, and (2) the number of sources is determined directly without going through the commonly adopted sequential extraction and learning process, which requires less computing time and hence facilitates on-line machine condition monitoring. The single-channel signal separation method has been applied to separate: a) a bearing signal mixture synthesized using inner and outer race fault signals obtained from literature, and b) a real bearing signal mixture containing inner and outer race faults obtained in our lab. However, it is noted that the accuracy of the separation results may be influenced by speed fluctuation and non-periodicity. Such effects remain to be investigated.

5. *Fault severity evaluation by Lempel-Ziv complexity*

As described in Chapter 7, the non-dimensional fault severity measure based on the Lempel-Ziv complexity has many advantages over the comparison-based methods. However, the existing Lempel-Ziv complexity based methods have three deficiencies. They are unable to provide: a) a clear distinction between noise and severe fault, b) reliable fault severity assessment based on a single set of measured data; and c) theoretical explanation as for why the complexity value can be used for fault severity assessment. In view of the above, a new non-dimensional fault severity measure is devised. The proposed fault severity merasure has effectively overcome the difficulties present in the current Lempel-Ziv complexity based approaches. The ambiguity between fault and noise is avoided by jointly applying the Lempel-Ziv complexity and CWT. Adopting the best scale of the CWT not only improves distinction between fault and noise, but also increases the accuracy of calculated complexity value by reducing noise. It is shown that single data measurement provides satisfactory result through application to the bearing inner- and outer-race fault signals. To establish the relationship between the Lempel-Ziv complexity and the bearing fault severity, the factors contributing to the complexity value are explained by an analytical model for a single-point defective bearing. The proposed method features the following: a) it is developed based on the mechanism of bearing faults, b) it does not need multiple measurements for the single fault, c) it does not require reference data set, and d) it provides a non-dimensional severity value between 0 and 1. Therefore, it can be used for more reliable and quicker on-line fault detection and diagnosis in comparison with the existing methods.

8.2 Future work

The following work may be further investigated.

1. Removal of the fake impulses or spurious impulses from the non-periodic signal

In Chapter 3, the correlation function and the continuous wavelet transform were jointly used to estimate the time instants of the periodic impulses and to remove spurious impulses. The correlation function reveals the hidden periodicity of signal, however, it may be difficult to apply the proposed method to non-periodic signals. In industrial applications, the machine may be operated in non-stationary conditions. Hence, further research may be necessary to estimate the time instants of the non-periodic impulses and to remove spurious impulses from the non-periodic signal.

2. Further validation of the applicability of K-hybrid method for non-stationary signals

As shown in Chapter 4, the K-hybrid de-noising method is effective for de-noising gear and bearing signals under stationary operating conditions. The possible application of the method to the non-stationary signal was shown through simulation. It is suggested that the applicability to non-stationary problems be validated using experimental data. In addition, the application of the K-hybrid method to extraction of impulsive biomedical signatures, e.g., EBS (Explosive Bowel Sound) , could also an interesting research topic.

3. Improvement of the low-pass filter characteristic of the oil-particle signature extractor

The two detectors devised based on the fractional integration have shown some low-pass filter characteristic in extracting the particle signature (in Chapter 5). It is also shown that the two detectors can remove the high-frequency noise without applying additional filtering process. However, low-frequency noise or low frequency interfering component could remain in the signal. To obtain more reliable results, such low-frequency noise or interference should also be removed. Therefore, it is recommended to improve the proposed algorithm by adding a feature to remove the low-frequency noise and/or interference.

4. Blind source separation of non-stationary signals

The single-channel signal separation method proposed in Chapter 6 is intended to be used for separation of stationary signals. Such a restriction is mainly because of the use of the Fourier transform in the algorithm. The Fourier transform can handle only stationary periodic signals obtained from rotating machinery. However, the rotating machine can also run in a non-stationary operating condition. In this case, the proposed method may not provide reliable results. Further research in this direction would be a useful topic.

5. Enhancement of the fault severity measure

In Chapter 7, a fault severity measure for rolling element bearings was developed based on the Lempel-Ziv complexity and continuous wavelet transform. The non-dimensional fault severity measure has shown to be useful for assessing the fault severity without comparing two different machine states. However, the result is somewhat affected by shaft rotating speed. In addition, the applicability of the new severity measure to other rotating components such as gears is yet to be evaluated. Therefore, it will be a very important topic to enhance the severity measure that is not susceptible to operating conditions (such as speed and load changes) and can be used for a wider range of rotating components.

Bibliography

Abbate, A., DeCusatis, C., and Das, P. (2002). *Wavelets and subbands - fundamentals and applications.* (Boston: Birkhäuser).

Akochi-Koble, E., Pelchat, M., Pinchuk, D., Ismail, A., Van de Voort, F., Dong, J., Dwight, S., and Davies, M. (1998). Validation of FTIR spectroscopy method for measuring and monitoring fuel dilution in lubricating oils. In *Proceedings of a Joint International Condition Monitoring Conference*, pages 177–185.

Antoni, J. (2005). Blind separation of vibration components: Principles and demonstrations. *Mechanical Systems and Signal Processing*, 19(6):1166–1180.

Antoni, J. and Braun, S. (2005). Editorial. *Mechanical Systems and Signal Processing*, 19(6):1163–1165.

Baydar, N. and Ball, A. (2001). A comparative study of acoustic and vibration signals in detection of gear failures using wigner-ville distribution. *Mechanical Systems and Signal Processing*, 15(6):1091–1107.

Baydar, N. and Ball, A. (2003). Detection of gear failures via vibration and acoustic signals using wavelet transform. *Mechanical Systems and Signal Processing*, 17(4):787–804.

Belcastro, C. and Allen, C. (2007). Integrated vehicle health management-technical plan summary. http://www.aerospace.nasa.gov/nra_pdf/ivhm_tech_plan_c1.pdf.

Bendat, J. and Piersol, A. (1991). *Random data - analysis and measurement procedures.* (Singapore: John Wiley & Sons, Inc.), 2 edition.

Bonnardot, F., Randall, R., and Guillet, F. (2005). Extraction of second-order cyclostationary sources - application to vibration analysis. *Mechanical Systems and Signal Processing*, 19(6):1230–1244.

Bouguerriou, N., Haritopoulos, M., Capdessus, C., and Allam, L. (2005). Novel cyclostationarity-based blind source separation algorithm using second order statistical properties: Theory and application to the bearing defect diagnosis. *Mechanical Systems and Signal Processing*, 19(6):1260–1281.

Boulahbal, D., Golnaraghi, M., and Ismail, F. (1999). Amplitude and phase wavelet maps for the detection of cracks in geared systems. *Mechanical Systems and Signal Processing*, 13(3):423–436.

Boustany, R. and Antoni, J. (2005). A subspace method for the blind extraction of a cyclostationary source: Application to rolling element bearing diagnostics. *Mechanical Systems and Signal Processing*, 19(6):1245–1259.

Brandlein, J., Eschmann, P., Hasbargen, L., and Weigand, K. (1999). *Ball and roller bearings - theory, design and aplication.* (New York: John Wiley & Sons, Inc.), 3 edition.

Braun, S. (1980). The signature analysis of sonic bearing vibrations. *IEEE Transactions on Sonics and Ultrasonics*, SU-27(6):317–328.

Braun, S. and Datner, B. (1979). Analysis of roller/ball bearing vibrations. *Transactions of the ASME Journal of Mechanical Design*, 101:118–125.

Cao, M. and Liu, R. (1996). General approach to blind source separation. *IEEE Transactions on Signal Processing*, 44:562–571.

Case Western Reserve University (2007). Bearing datacenter - seeded fault test data. http://www.eecs.cwru.edu/laboratory/bearing/welcome_overview.htm.

Chang, S., Yu, B., and Vetterli, M. (2000a). Adaptive wavelet thresholding for image denoising and compression. *IEEE Transactions on Image Processing*, 9(9):1532–1546.

Chang, S., Yu, B., and Vetterli, M. (2000b). Spatially adaptive wavelet thresholding with context modeling for image denoising. *IEEE Transactions on Image Processing*, 9(9):1522–1531.

Cichocki, A., Thawonmas, R., and Amari, S. (1997). Sequential blind signal extraction in order specified by stochastic properties. *Electronics Letters*, 33:64–65.

Covacece, M. and Introni, A. (2002). Analysis of damage of ball bearings of aeronautical transmissions by auto-pwer spectrum and cross-power spectrum. *Transactions of the ASME Journal of Vibration and Acoustics*, 124:180–185.

Cover, T. and Thomas, J. (1991). *Elements of information theory.* (New York: John Wiley & Sons, Inc.).

Dalpiaz, G. and Rivola, A. (1997). Condition monitoring and diagnostics in automatic machines: comparison of vibration analysis techniques. *Mechanical Systems and Signal Processing*, 11(1):53–73.

Dalpiaz, G., Rivola, A., and Rubini, R. (2000). Effectiveness and sensitivity of vibration processing techniques for local fault detection in gears. *Mechanical Systems and Signal Processing*, 14(3):387–412.

Davies, A. (1990). *Management guide to condition monitoring.* (London: The Institute of Production Engineers).

Dempsey, P. (2000). A comparison of vibration and oil debris gear damage detection methods applied to pitting damage. Technical Report TM-2000-210371, NASA.

Dempsey, P. (2001). Gear damage detection using oil debris analysis. Technical Report TM-2001-210936, NASA.

Donoho, D. (1995). De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627.

Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

Doyle, J. (1995). Some thoughts on fractional calculus and the abolition of integration formulae. *Teaching Mathematics and its Applications*, 15(1):16–19.

Ertunc, H., Loparo, K., and Ocak, H. (2001). Tool wear condition monitoring in drilling operations using hidden markov models. *International Journals of Machine Tools and Manufacture*, 41:1363–1384.

Garry, M. and Doll, D. (1998). Fourier transform infrared analysis applied to maintenance management - making the most of this powerful tool. In *Proceedings of a Joint International Condition Monitoring Conference*, pages 162–169.

Gelle, G., Colas, M., and Delaunay, G. (2000). Blind sources separation applied to rotating machines monitoring by acoustical and vibrations analysis. *Mechanical Systems and Signal Processing*, 14:427–442.

Goupillaud, P., Grossmann, A., and Morlet, J. (1984). Cycle-octave and related transforms in seismic signal analysis. *Geoexploration*, 23:85–102.

Gradshteyn, I. and Ryzhik, I. (1965). *Table of integrals, series, and products.* (New York: Academic Press Inc.), 4 edition.

Hadjileontiadis, L. and Rekanos, I. (2003). Enhancement of explosive bowel sounds using kurtosis-based filtering. In *Proceedings of the 25th Annual International Conference of the IEEE EMBS*, pages 2479–2482, Cancun, Mexico.

Harris, C. and Piersol, A. (2002). *Harris' shock and vibration handbook.* (New York: McGraw-Hill Book Company), 5 edition.

Harris, T. (1984). *Rolling bearing analysis.* (New York: John Wiley & Sons, Inc.), 2 edition.

Ho, D. and Randall, R. (2000). Optimisation of bearing diagnostic techniques using simulated and actual bearing fault signals. *Mechanical Systems and Signal Processing*, 14(5):763–788.

Howard, I. (1994). A review of rolling element bearing vibration - detection, diagnosis and prognosis. Research Report DSTR-RR-0013, DSTO Aeronautical and Maritime Research Laboratory, Australia.

Howe, B. and Muir, D. (1998). In-line oil debris monitor (odm) for helicopter gearbox condition assessment. In *Proceedings of a Joint International Condition Monitoring Conference*, pages 101–109, Mobile, Alabama, USA.

Humphrey, G., Whitlock, R., Little, D., and Godin, R. (1998). Energy dispersive X-ray fluorescence evaluation of debris from F-18 engine oil filters. In *Proceedings of a Joint International Condition Monitoring Conference*, pages 236–249.

Hyvärinen, A. (1999). Sparse code shrinkage de-noising of non-gaussian data by maximum likelihood estimation. *Neural Computation*, 11:739–768.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent component analysis.* Adaptive and learning systems for signal processing, communications, and control. John Wiley & Sons, Inc.

Iseler, L. and Maio, J. (2002). An analysis of U.S. civil rotorcraft accidents by cost and injury (1990-1996). Technical Report NASA TM-2002-209615, NASA.

Jang, G. and Lee, T. (2003). A maximum likelihood approach to single-channel source separation. *Journal of Machine Learning Research*, 4:1365–1392.

Jansen, M. (2001). *Noise reduction by wavelet thresholding*, volume 161 of *Lecure Notes in Statistics.* (New York: Springer-Verlag).

Jardine, A., Lin, D., and Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*, 20:1483–1510.

Kaspar, F. and Schuster, H. (1987). Easily calculable measure for the complexity of spatiotemporal patterns. *The American Physical Society Physical Review A.*, 36(2):842–848.

Kleinz, M. and Osler, T. (2000). A child's garden of fractional derivatives. *The College Mathematics Journal*, 31(2):82–88.

Kreyszig, E. (1993). *Advanced engineering mathematics.* (New York: John Wiley & Sons, Inc.).

Latino, K. (1996). Failure modes and effect analysis - a modified approach. In *Proceedings of NPRA refiery and petrochemical plant maintenance conference*, number MC-96-89, Nashville, Tennessee. National Petroleum Refineries Association.

Lavoie, J., Osler, T., and Tremblay, R. (1976). Fractional derivatives and special functions. *SIAM Review*, 18(2):240–268.

Learmount, D. (2000). Rotary woes. *Flight International*, 157:34–44.

Lempel, A. and Ziv, J. (1976). On the complexity of finite sequences. *IEEE Transactions on Information Theory*, IT-22(1):75–81.

Li, W., Gu, F., Ball, A., Leung, A., and Phipps, C. (2001). A study of the noise from diesel engines using the independent component analysis. *Mechanical Systems and Signal Processing*, 15:1165–1184.

Li, W. and Wang, J. (2002). Sequential blind extraction of instantaneously mixed sources. *IEEE Transactions on Signal Processing*, 50:997–1006.

Lin, J. (2001). Feature extraction of machine sound using wavelet and its application in fault diagnosis. *NDT & E International*, 34:25–30.

Lin, J. and Qu, L. (2000). Feature extraction based on morlet wavelet and its application for mechanical fault diagnosis. *Journal of Sound and Vibration*, 234(1):135–148.

Lin, J., Zuo, M., and Fyfe, K. (2004). Mechanical fault detection based on the wavelet de-noising technique. *Transactions of the ASME Journal of Vibration and Acoustics*, 126:9–16.

Loutridis, S. (2004a). Damage detection in gear systems using empirical mode decomposition. *Engineering Structures*, 26:1833–1841.

Loutridis, S. (2004b). A local energy density methodology for monitoring the evolution of gear faults. *NDT & E International*, 37:447–453.

Martin, K. (1994). A review by discussion of condition monitoring and fault diagnosis in machine tools. *International Journals of Machine Tools and Manufacture*, 34(4):527–551.

Mathieu, B., Melchior, P., Oustaloup, A., and Ceyral, C. (2003). Fractional differentiation for edge detection. *Signal Processing*, 83:2421–2432.

Mathworks (2007). MATLAB on-line help. http://www.mathworks.com/access/helpdesk /help/tecdoc.

McFadden, P. and Smith, J. (1984). Model for the vibration produced by a single point defect in a rolling element bearing. *Journal of Sound and Vibration*, 96(1):69–82.

Meltzer, G. and Dien, N. (2004). Fault diagnosis in gears operating under non-stationary rotational speed using polar wavelet amplitude maps. *Mechanical Systems and Signal Processing*, 18:985–992.

Miller, J. and Kitaljevich, D. (2000). In-line oil debris monitoring for aircraft engine condition assessment. In *Proceedings of IEEE Aerospace Conference*, volume 6, pages 49–56, Big Sky, MT, USA.

Miller, K. (1995). Derivatives of non-integer order. *Mathematics Magazine*, 68(3):183–192.

Myers, L., Baer-Riedhart, J., and Maxwell, M. (1985). Fault detection and accommodation testing on an F100 engine in an F-15 airplane. Technical Report NASA-TM-86735, NASA.

Nikolaou, N. and Antoniadis, I. (2002). Rolling element bearing fault diagnosis using wavelet packets. *NDT & E International*, 35:197–205.

Oldham, K. and Spanier, J. (2006). *The fractional calculus - theory and applications of differentiation and integration to arbitrary order.* (New York: Dover Publications, Inc.).

Ortigueira, M. (2006). A coherent approach to non-integer order derivatives. *Signal Processing*, 86:2505–2515.

Oustaloup, A., Levron, F., Bansard, M., and Nanot, F. (1993). Generalized differential equation and generalized characteristic equation: solving and application to engineering sciences. In *Proceedings of IEEE International Conference on Systems, Man and Cybernetics, Systems Engineering in the Science of Humans*, volume 2, pages 156–165, Le Touquet, France.

Papoulis, A. and Pillai, S. (1998). *Probability, random variables and stochastic processes.* (Boston: McGraw-Hill Book Company), 4 edition.

Peled, R., Braun, S., and Zacksenhouse, M. (2005). A blind deconvolution separation of multiple sources with application to bearing diagnostics. *Mechanical Systems and Signal Processing*, 19(6):1181–1195.

Peterborough, N. (1965). A detailed failure analysis of the ball bearings from a model 1631 tape recorder. Technical Report NASA-CR-71026, NASA.

Pipe, K. (2003). Measuring the performance of a hum system - the features that count. Research Report DSTO-GD-0348, Defence Science and Technology Organisation.

Post, E. (1930). Generalized differentiation. *Transactions of the American Mathematical Society*, 32(4):723–781.

Powell, J. (1998). Molecular condition monitoring in the commercial world: Objectives and application of FT-IR analysis. In *Proceedings of a Joint International Condition Monitoring Conference*, pages 186–193.

Proakis, J. and Manolakis, D. (1996). *Digital signal processing - principles, algorithms, and applications.* (Upper Saddle River: Prentice Hall), 3 edition.

Puigt, M. and Deville, Y. (2005). Time-frequency ratio-based blind separation methods for attenuated and time-delayed sources. *Mechanical Systems and Signal Processing*, 19(6):1348–1379.

Randall, R. (1982). A new method of modeling gear faults. *Transactions of the ASME Journal of Mechanical Design*, 104:259–267.

Randall, R., Antoni, J., and Chosbaard, S. (2001). The relationship between special correlation and envelope analysis in the diagnostics of bearing faults and other cyclostationary machine signals. *Mechanical Systems and Signal Processing*, 15(5):945–962.

Reintjes, J., Mahon, R., Duncan, M., McClelland, T., Tankersley, L., Schultz, A., Lu, C., Howard, P., and Stevens, C. (1998). Lasernet optical oil debris monitor. In *Proceedings of a Joint International Condition Monitoring Conference*, pages 110–116.

Rioul, O. and Duhamel, P. (1992). Fast algorithms for discrete and continuous wavelet transforms. *IEEE Transactions on Information Theory*, 38(2):569–586.

Roweis, S. (2001). One microphone source separation. *Advances in Neural Information Processing Systems*, 13:793–799.

Rubini, R. and Meneghetti, U. (2001). Application of the envelope and wavelet transform analyses for the diagnosis of incipient faults in ball bearing. *Mechanical Systems and Signal Processing*, 15(2):287–302.

Servière, C. and Fabry, P. (2004). Blind source separation of noisy harmonic signals for rotating machine diagnosis. *Journal of Sound and Vibration*, 272:317–339.

Servière, C. and Fabry, P. (2005). Principal component analysis and blind source separation of modulated sources for electro-mechanical systems diagnostics. *Mechanical Systems and Signal Processing*, 19(6):1293–1311.

Spiegel, M. (1963). *Schaum's outline of Theory and Problems of Advanced Calculus.* Schaum's Outline Series. (New York: McGraw-Hill Book Company).

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9(6):1135–1151.

Tandon, N. and Choudhury, A. (1997). An analytical model for the prediction of the vibration response of rolling element bearings due to a localized defect. *Journal of Sound and Vibration*, 205(3):275–292.

Toms, A. and MacIsaac, B. (2005). On-line oil condition sensor for industrial gas turbine condition assessment. Technical Report Paper No. 05-IAGT-2.3, Industrial Application of Gas Turbines Committee.

Toms, A., Powell, J., and Dixon, J. (1998a). The utilization of FT-IR for Army oil condition monitoring. In *Proceedings of a Joint International Condition Monitoring Conference*, pages 170–176.

Toms, A., Rookey, M., and Fitzgerald, R. (1998b). Comparison of 100 micron transmission cell and the 3M IR card for FT-IR analysis of military fluids. In *Proceedings of a Joint International Condition Monitoring Conference*, pages 194–202.

Tse, P., Peng, Y., and Yam, R. (2001). Wavelet analysis and envelope detection for rolling element bearing fault diagnosis - their effectiveness and flexibilities. *Transactions of the ASME Journal of Vibration and Acoustics*, 123:303–310.

Tucker, J., Reintjes, J., Duncan, M., McClelland, T., Tankersley, L., Lu, C., Howard, P., Sebok, T., Holloway, C., and Fockler, S. (1998). Lasernet fines optical oil debris monitor. In *Proceedings of a Joint International Condition Monitoring Conference*, pages 117–124.

Tumer, I. and Bajwa, A. (1999). A survey of aircraft engine health monitoring systems. In *The 35th AIAA/ASME/SAE/SAEE Joint Propulsion Conference*, number AIAA-99-2528.

Volponi, A. and Wood, B. (2005). Engine health management for aircraft propulsion systems. In *Proceedings of the International Forum on Integrated System Health Engineering and Management in Aerospace*, Napa, California, USA.

Wang, Y. and Kootsookos, P. (1998). Modeling of low shaft speed bearing faults for condition monitoring. *Mechanical Systems and Signal Processing*, 12(3):415–426.

Whitlock, R., Humphrey, G., and Churchill, D. (1998). The path to affordable long term failure warning: the XRF-wear monitor. In *Proceedings of a Joint International Condition Monitoring Conference*, pages 250–259.

Wilson, E., Sutter, D., Berkovitz, D., Betts, B., Del-Mundo, R., Kong, E., Lages, C., Mah, R., and Papasin, R. (2003). Motion-based system identification and fault detection and isolation technologies for thruster controlled spacecraft. In *Proceedings of the JANNAF 3rd Modeling and Simulation Joint Subcommittee Meeting*, Colorado Springs, CO, USA.

Wolfman, S. (1984). Cellular automata as models of complexity. *Nature*, 311(4):419–424.

Wolfman, S. (1985). Origins of randomness in physical systems. *The American Physical Society Physical Review Letters*, 55(5):449–452.

Yan, R. and Gao, R. (2004). Complexity as a measure for machine health evaluation. *IEEE Transactions on Instrumentation and Measurement*, 53(4):1327–1334.

Ye, J., Zhu, X., and Zhang, X. (2004). Adaptive blind separation with an unknown number of sources. *Neural Computation*, 16:1641–1660.

Yoon, B. and Vaidyanathan, P. (2004). Wavelet-based de-noising by customized thresholding. In *Proceedings of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, volume 2, pages 925–928, Motreal, Canada.

Zakrajsek, J. (1989). An investigation of gear mesh failure prediction techniques. Technical Report NASA-TM-102340, NASA.

Zakrajsek, J., Townsend, D., and Decker, H. (1992). An analysis of gear fault detection methods as applied to pitting fatigue failure data. Technical Report NASA TM-105950, NASA.

Appendices

Appendix A

Derivations and proofs

A.1 Continuous Wavelet Transform, Eq. (3.6)

By substituting Eqs. (3.4) and (3.5) into Eq. (3.2), the wavelet coefficients $W(a,\, b)$ can be written as follows:

$$\begin{aligned} W(a,\, b) &= \int x(t) w^*_{a,b}(t)\, dt \\ &\approx \frac{1}{\sqrt{a}} \int \cos(\omega_1 t) \cdot \exp\left(-\frac{1}{2}\left(\frac{t-b}{a}\right)^2\right) \cdot \cos\left(\frac{\omega_0(t-b)}{2a}\right) dt + n'(t) \end{aligned}$$

where $n'(t)$ is CWT of noise $n(t)$.

Letting $p = t - b$, $dt = dp$ leads to

$$\begin{aligned} W(a,\, b) &\approx \frac{1}{\sqrt{a}} \int \exp\left(-\frac{p^2}{2a^2}\right) \cdot \cos(\omega_1 p + \omega_1 b) \cdot \cos\left(\frac{\omega_0 p}{2a}\right) dp + n'(t) \\ &\approx \frac{1}{\sqrt{a}} \int \exp\left(-\frac{p^2}{2a^2}\right) \cdot \left[\cos(\omega_1 p)\cos(\omega_1 b) - \sin(\omega_1 p)\sin(\omega_1 b)\right] \cdot \cos\left(\frac{\omega_0 p}{2a}\right) dp \\ &\quad + n'(t) \\ &\approx \frac{1}{\sqrt{a}} \cos(\omega_1 b) \int \exp\left(-\frac{p^2}{2a^2}\right) \cdot \cos(\omega_1 p) \cos\left(\frac{\omega_0 p}{2a}\right) dp \\ &\quad - \frac{1}{\sqrt{a}} \sin(\omega_1 b) \int \exp\left(-\frac{p^2}{2a^2}\right) \cdot \sin(\omega_1 p) \cos\left(\frac{\omega_0 p}{2a}\right) dp + n'(t). \end{aligned}$$

Solving the indefinite integral of the first term and rewriting the sine function in the second term result in the following:

$$W(a,\,b) \approx \frac{1}{\sqrt{a}} \cos(\omega_1 b) \cdot \frac{1}{4}\sqrt{\frac{\pi}{1/2a^2}} \left\{ \exp\left(-\frac{\left(\omega_1 - \frac{\omega_0}{2a}\right)^2}{4\,(1/2a^2)} \right) + \exp\left(-\frac{\left(\omega_1 + \frac{\omega_0}{2a}\right)^2}{4\,(1/2a^2)} \right) \right\}$$

(Eq. (3.898-2), pp.480 of Gradshteyn and Ryzhik (1965))

$$-\frac{1}{\sqrt{a}} \sin(\omega_1 b) \int \exp\left(-\frac{p^2}{2a^2}\right) \cdot \frac{1-\cos(2\omega_1 p)}{2} \cdot \cos\left(\frac{\omega_0 p}{2a}\right)\, dp + n'(t)$$

or

$$\begin{aligned} W(a,\,b) \approx \frac{1}{\sqrt{a}} \Bigg\{ & \cos(\omega_1 b) \cdot \frac{1}{4}\sqrt{2\pi a^2} \cdot \exp\left(-\left(\omega_1 - \frac{\omega_0}{2a}\right)^2 \Big/ \left(\frac{2}{a^2}\right) \right) \\ & + \cos(\omega_1 b) \cdot \frac{1}{4}\sqrt{2\pi a^2} \cdot \exp\left(-\left(\omega_1 + \frac{\omega_0}{2a}\right)^2 \Big/ \left(\frac{2}{a^2}\right) \right) \\ & + \frac{1}{2} \sin(\omega_1 b) \int \exp\left(-\frac{p^2}{2a^2}\right) \cdot \cos\left(\frac{\omega_0 p}{2a}\right)\, dp \\ & - \frac{1}{2} \sin(\omega_1 b) \int \exp\left(-\frac{p^2}{2a^2}\right) \cdot \cos(2\omega_1 p) \cdot \cos\left(\frac{\omega_0 p}{2a}\right)\, dp \Bigg\} + n'(t). \end{aligned}$$

The two indefinite integrals are solved as follows:

$$\begin{aligned} W(a,\,b) \approx \frac{1}{\sqrt{a}} \Bigg\{ & \cos(\omega_1 b) \cdot \frac{1}{4}\sqrt{2\pi a^2} \cdot \exp\left(-\left(\omega_1 - \frac{\omega_0}{2a}\right)^2 \Big/ \left(\frac{2}{a^2}\right) \right) \\ & + \cos(\omega_1 b) \cdot \frac{1}{4}\sqrt{2\pi a^2} \cdot \exp\left(-\left(\omega_1 + \frac{\omega_0}{2a}\right)^2 \Big/ \left(\frac{2}{a^2}\right) \right) \\ & + \frac{1}{2} \sin(\omega_1 b) \cdot \frac{1}{2}\sqrt{\frac{\pi}{(1/2a^2)}} \cdot \exp\left(-\frac{(\omega_0/2a)^2}{4\,(1/2a^2)} \right) \end{aligned}$$

(Eq. (3.896-4), pp.480 of Gradshteyn and Ryzhik (1965))

$$- \frac{1}{2} \sin(\omega_1 b) \cdot \frac{1}{2}\sqrt{\frac{\pi}{1/2a^2}} \cdot \left\{ \exp\left(-\frac{\left(\omega_1 - \frac{\omega_0}{2a}\right)^2}{4\,(1/2a^2)} \right) + \exp\left(-\frac{\left(\omega_1 + \frac{\omega_0}{2a}\right)^2}{4\,(1/2a^2)} \right) \right\} \Bigg\}$$

(Eq. (3.898-2), pp.480 of Gradshteyn and Ryzhik (1965))

$$+ n'(t).$$

Hence, the continuous wavelet transform of the periodic signal mixture can be expressed as,

$$
\begin{aligned}
W(a,\, b) \approx \frac{1}{\sqrt{a}} \Bigg\{ & \cos(\omega_1 b) \cdot \frac{1}{4}\sqrt{2\pi a^2} \cdot \exp\left(-\left(\omega_1 - \frac{\omega_0}{2a}\right)^2 \Big/ \left(\frac{2}{a^2}\right)\right) \\
& + \cos(\omega_1 b) \cdot \frac{1}{4}\sqrt{2\pi a^2} \cdot \exp\left(-\left(\omega_1 + \frac{\omega_0}{2a}\right)^2 \Big/ \left(\frac{2}{a^2}\right)\right) \\
& + \frac{1}{4}\sqrt{2\pi a^2} \sin(\omega_1 b) \cdot \exp\left(-\frac{\omega_0^2}{8}\right) \\
& - \frac{1}{4}\sqrt{2\pi a^2} \sin(\omega_1 b) \cdot \exp\left(-\left(\omega_1 - \frac{\omega_0}{2a}\right)^2 \Big/ \left(\frac{2}{a^2}\right)\right) \\
& + \frac{1}{4}\sqrt{2\pi a^2} \sin(\omega_1 b) \cdot \exp\left(-\left(\omega_1 + \frac{\omega_0}{2a}\right)^2 \Big/ \left(\frac{2}{a^2}\right)\right) \Bigg\} \\
& + n'(t). \qquad \blacksquare
\end{aligned}
$$

A.2 Autocorrelation function of wavelet coefficients, Eq. (3.13)

At the best scale a_{best}, the autocorrelation coefficient function can be obtained as follows:

$$\begin{aligned}
\rho_{WW}(\tau) &\approx \frac{1}{\sigma_W^2} E\left[|W(a,\, b)| \cdot |W(a,\, b+\tau)|\right] \\
&\approx \frac{1}{\sigma_W^2} \lim_{M\to\infty} \frac{1}{2M} \int_{-M}^{M} \Big| \left[C_1 \cos(\omega_1 b) - C_2 \sin(\omega_1 b)\right] \Big| \\
&\quad \cdot \Big| \left[C_1 \cos(\omega_1 (b+\tau)) - C_2 \sin(\omega_1 (b+\tau))\right] \Big|\, db \qquad \text{(from Eq. (3.11))} \\
&\approx \frac{1}{\sigma_W^2} \lim_{M\to\infty} \frac{1}{2M} \int_{-M}^{M} \Big| \{ C_1^2 \cos(\omega_1 b) \left[\cos(\omega_1 b)\cos(\omega_1 \tau) - \sin(\omega_1 b)\sin(\omega_1 \tau)\right] \\
&\quad - C_1 C_2 \cos(\omega_1 b) \left[\sin(\omega_1 b)\cos(\omega_1 \tau) + \cos(\omega_1 b)\sin(\omega_1 \tau)\right] \\
&\quad - C_1 C_2 \sin(\omega_1 b) \left[\cos(\omega_1 b)\cos(\omega_1 \tau) - \sin(\omega_1 b)\sin(\omega_1 \tau)\right] \\
&\quad - C_2^2 \sin(\omega_1 b) \left[\sin(\omega_1 b)\cos(\omega_1 \tau) + \cos(\omega_1 b)\sin(\omega_1 \tau)\right] \} \Big|\, db\,.
\end{aligned}$$

Integration can be carried out below:

$$\begin{aligned}
\rho_{WW}(\tau) \approx \frac{1}{\sigma_W^2} \lim_{M\to\infty} \frac{1}{2M} \int_{-M}^{M} \Big| \{ &\left[C_1^2 \cos(\omega_1 \tau) - C_1 C_2 \sin(\omega_1 \tau)\right] \cos^2(\omega_1 b) \\
&+ \left[C_1 C_2 \sin(\omega_1 \tau) + C_2^2 \cos(\omega_1 \tau)\right] \sin^2(\omega_1 b) \\
&+ \left[(C_2^2 - C_1^2)\sin(\omega_1 \tau) - 2C_1 C_2 \cos(\omega_1 \tau)\right] \sin(\omega_1 \tau)\cos(\omega_1 \tau) \} \Big|\, db
\end{aligned}$$

using the following relations

$$\int_{-M}^{M} \cos^2(\omega_1 b)\, db = \left[\frac{b}{2} + \frac{\sin(2\omega_1 b)}{4\omega_1}\right]_{-M}^{M} = M + \frac{\sin(2\omega_1 M)}{2\omega_1}$$

(Eq. (2.513-11), pp.132 Gradshteyn and Ryzhik (1965))

$$\int_{-M}^{M} \sin^2(\omega_1 b)\, db = \left[\frac{b}{2} - \frac{\sin(2\omega_1 b)}{4\omega_1}\right]_{-M}^{M} = M - \frac{\sin(2\omega_1 M)}{2\omega_1}$$

(Eq. (2.513-5), pp.132 Gradshteyn and Ryzhik (1965))

$$\int_{-M}^{M} \sin(\omega_1 b)\cos(\omega_1 b)\, db = \left[-\frac{\cos(2\omega b)}{4\omega_1}\right]_{-M}^{M} = 0$$

(Eq. (2.532-11), pp.140 Gradshteyn and Ryzhik (1965))

The autocorrelation coefficient function is simplified as

$$\rho_{WW}(\tau) \approx \frac{1}{\sigma_W^2} \lim_{M\to\infty} \frac{1}{2M} \left\{ \left| \left[C_1^2 \cos(\omega_1 \tau) - C_1 C_2 \sin(\omega_1 \tau) \right] \right| \left(M + \frac{\sin(2\omega_1 M)}{2\omega_1} \right) + \left| \left[C_1 C_2 \sin(\omega_1 \tau) + C_2^2 \cos(\omega_1 \tau) \right] \right| \left(M - \frac{\sin(2\omega_1 M)}{2\omega_1} \right) \right\}.$$

After rearranging terms, one obtains

$$\begin{aligned} \rho_{WW}(\tau) &\approx \frac{1}{\sigma_W^2} \lim_{M\to\infty} \frac{1}{2M} \left| \cos(\omega_1 \tau) \cdot \left[M\left(C_1^2 + C_2^2\right) + \left(C_1^2 - C_2^2\right) \frac{\sin(2\omega_1 M)}{2\omega_1} \right] \right. \\ &\quad \left. - \sin(\omega_1 \tau) \left[2C_1 C_2 \frac{\sin(2\omega_1 M)}{2\omega_1} \right] \right| \\ &\approx \frac{1}{\sigma_W^2} \lim_{M\to\infty} \frac{1}{2M} \left| C_3 \cos(\omega_1 \tau) - C_4 \sin(\omega_1 \tau) \right| \\ &\approx \frac{1}{\sigma_W^2} \lim_{M\to\infty} \frac{1}{2M} \left| C_3 \cos\left(2\pi \frac{\tau}{T_1}\right) - C_4 \sin\left(2\pi \frac{\tau}{T_1}\right) \right| \\ C_3 &= M\left(C_1^2 + C_2^2\right) + \left(C_1^2 - C_2^2\right) \frac{\sin(2\omega_1 M)}{2\omega_1} \\ C_4 &= 2C_1 C_2 \frac{\sin(2\omega_1 M)}{2\omega_1} \end{aligned}$$

■

A.3 n-th order derivative operator D_R^n, Eq. (5.7)

The first order differentiation is

$$\begin{aligned} D_R^1 x(t) &= \frac{x(t) - x(t-h)}{h} && (h \to 0) \\ &= \frac{1-q^{-1}}{h} x(t) && \big(q^{-1}x(t) = x(t-h)\big) \\ &= \dot{x}(t) \end{aligned}$$

The second and third order differentiations are

$$\begin{aligned} D_R^2 x(t) &= \frac{\dot{x}(t) - \dot{x}(t-h)}{h} \\ &= \frac{(1-q^{-1})}{h} \dot{x}(t) \\ &= \frac{(1-q^{-1})}{h} \cdot \frac{(1-q^{-1})}{h} x(t) \\ &= \left(\frac{1-q^{-1}}{h}\right)^2 x(t) \\ &= \ddot{x}(t) \end{aligned}$$

and

$$\begin{aligned} D_R^3 x(t) &= \frac{\ddot{x}(t) - \ddot{x}(t-h)}{h} \\ &= \frac{(1-q^{-1})}{h} \ddot{x}(t) \\ &= \frac{(1-q^{-1})}{h} \cdot \frac{(1-q^{-1})}{h} \dot{x}(t) \\ &= \left(\frac{1-q^{-1}}{h}\right)^2 \cdot \frac{(1-q^{-1})}{h} x(t) \\ &= \left(\frac{1-q^{-1}}{h}\right)^3 x(t) \\ &= \dddot{x}(t) \end{aligned}$$

Therefore, for any order n, one can write

$$D_R^n x(t) = \left(\frac{1-q^{-1}}{h}\right) x(t) \qquad \blacksquare$$

A.4 Generalized differentiation of shift operator, Eqs. (5.8) and (5.9)

Based on the Newtons' binomial formula, Eq. (5.7), i.e.,

$$D_R^n = \left(\frac{1-q^{-1}}{h}\right)^n$$

can be expanded as (Spiegel, 1963)

$$\begin{aligned}
\left(1-q^{-1}\right) &= 1 + n\left(-q^{-1}\right) + \frac{n(n-1)}{2!}\left(-q^{-1}\right)^2 + \cdots \\
&\quad + \frac{n(n-1)\cdots(n-k+1)}{k!}\left(-q^{-1}\right)^k + \cdots \\
&= \sum_{k=0}^{\infty} \frac{n(n-1)\cdots(n-k+1)}{k!}\left(-q^{-1}\right)^k \\
&= \sum_{k=0}^{\infty}(-1)^k \cdot \frac{n(n-1)\cdots(n-k+1)}{k!}\left(q^{-k}\right) .
\end{aligned}$$

Therefore, Eq. (5.8) can be obtained, i.e.,

$$\begin{aligned}
D_R^n &= \left(\frac{1-q^{-1}}{h}\right)^n \\
&= \frac{1}{h^n}\sum_{k=0}^{\infty}(-1)^k \cdot \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}\left(q^{-k}\right) .
\end{aligned}$$

Accordingly, the generalized differentiation of $x(t)$, Eq. (5.9), is

$$\begin{aligned}
D_R^n x(t) &= \frac{1}{h^n}\sum_{k=0}^{\infty}(-1)^k \cdot \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}\left(q^{-k}\right)\cdot x(t) \\
&= \frac{1}{h^n}\sum_{k=0}^{\infty}(-1)^k \cdot \frac{n!}{k!(n-k)!}\left(q^{-k}\right)\cdot x(t) \\
&\quad \left(\because \binom{n}{k} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} = \frac{n!}{k!(n-k)!}\right) \\
&\quad \text{(pp.1162 of Kreyszig (1993))} \\
&= \frac{1}{h^n}\sum_{k=0}^{\infty}(-1)^k \cdot \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)}\left(q^{-k}\right)\cdot x(t) \\
&= \frac{1}{h^n}\sum_{k=0}^{\infty}(-1)^k \cdot \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} x(t-kh) . \qquad \blacksquare
\end{aligned}$$

A.5 Laplace transform of the generalized differentiation of unit step function, Eq. (5.13)

$$
\begin{aligned}
L\left\{D_R^n u(t)\right\} &= L\left\{\frac{t^{-n}}{\Gamma(1-n)}\right\} \\
&= \frac{1}{\Gamma(m+1)} L\left\{t^m\right\} \qquad (m=-n) \\
&= \frac{1}{\Gamma(m+1)} \int_0^\infty t^m e^{-st} dt \\
&= \frac{1}{\Gamma(m+1)} \int_0^\infty \left(\frac{p}{s}\right)^m e^{-p} \frac{dp}{s} \qquad (p=st) \\
&= \frac{1}{\Gamma(m+1)s^{m+1}} \int_0^\infty p^m e^{-p} dp \\
&= \frac{1}{\Gamma(m+1)s^{m+1}} \Gamma(m+1) \qquad \left(\Gamma(m) = \int_0^\infty x^{m-1} e^{-x} dx\right) \\
&= \frac{1}{s^{m+1}} \\
&= \frac{1}{s^{-n+1}} \qquad (n<0) \qquad \blacksquare
\end{aligned}
$$

A.6 Proof: The three points yielded by the inflexion-point locator are true inflexion points

Proof: the three points yielded by the inflexion-point locator are true inflexion points.

At inflexion points, the following two conditions are true:

1. the second derivative is equal to zero, and
2. the second derivative has opposite signs around the inflexion point.

The following proves that the above two conditions are satisfied and the three points yielded by the inflexion-point locator are true inflexion points.

Providing the above two conditions are satisfied is to prove that, at any inflexion point, say, $t = t_1$, the particle signature $x(t)$ satisfies:

- Condition 1:

$$\left.\frac{d^2}{dt^2}x(t)\right|_{t=t_1} = \ddot{x}(t_1) = 0 \tag{A.1}$$

- Condition 2:
  For a sufficiently small $\epsilon > 0$,

$$\left\{\begin{array}{l} \lim\limits_{t\to t_1-\varepsilon} \ddot{x}(t) > 0 \\ \lim\limits_{t\to t_1+\varepsilon} \ddot{x}(t) < 0 \end{array}\right\} \quad \text{or} \quad \left\{\begin{array}{l} \lim\limits_{t\to t_1-\varepsilon} \ddot{x}(t) < 0 \\ \lim\limits_{t\to t_1+\varepsilon} \ddot{x}(t) > 0 \end{array}\right\} \quad (\varepsilon > 0) \tag{A.2}$$

To facilitate the proof, consider Eqs. (5.9) and (5.11),

$$D_R^n x(t) = \frac{1}{h^n}\sum_{k=0}^{\infty}(-1)^k \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} x(t-kh) \qquad \text{Eq. (5.9)}$$

$$D_L^n x(t) = \frac{1}{h^n}\sum_{k=0}^{\infty}(-1)^k \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} x(t+kh) \qquad \text{Eq. (5.11)}$$

and re-write $I^n x(t)$ as

$$\begin{aligned}
I^n x(t) &= D_L^n x(t) - D_R^n x(t) \qquad (n < 0) \\
&= \frac{1}{h^n}\sum_{k=0}^{\infty}(-1)^k \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} x(t+kh) \\
&\quad - \frac{1}{h^n}\sum_{k=0}^{\infty}(-1)^k \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} x(t-kh) \\
&= \frac{1}{h^n}\sum_{k=0}^{\infty}(-1)^k \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} \left[x(t+kh) - x(t-kh)\right] \qquad \text{(A.3)}
\end{aligned}$$

By Taylor expansion (Spiegel, 1963), one obtains

$$f(t+t_0) = f(t) + \frac{t_0}{1!}\dot{f}(t) + \frac{t_0^2}{2!}\ddot{f}(t) + \frac{t_0^3}{3!}\dddot{f}(t) + \cdots$$

The two terms, $x(t+kh)$ and $x(t-kh)$, in Eq. (A.3) can be written as follows,

$$x(t+kh) \approx x(t) + \frac{kh}{1!}\dot{x}(t) + \frac{(kh)^2}{2!}\ddot{x}(t) \qquad \text{(A.4)}$$

$$x(t-kh) \approx x(t) - \frac{kh}{1!}\dot{x}(t) + \frac{(kh)^2}{2!}\ddot{x}(t) \qquad \text{(A.5)}$$

Substituting Eqs. (A.4) and (A.5) into the inflexion-point locator Eq. (A.3) yields,

$$\begin{aligned}
I^n x(t) &= \frac{1}{h^n}\sum_{k=0}^{\infty}(-1)^k \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} \left[x(t+kh) - x(t-kh)\right] \\
&= \frac{1}{h^n}\sum_{k=0}^{\infty}(-1)^k \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} \cdot \\
&\qquad \left[x(t) + \frac{kh}{1!}\dot{x}(t) + \frac{(kh)^2}{2!}\ddot{x}(t) - \left(x(t) - \frac{kh}{1!}\dot{x}(t) + \frac{(kh)^2}{2!}\ddot{x}(t)\right)\right] \\
&= \frac{1}{h^n}\sum_{k=0}^{\infty}(-1)^k \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} \cdot 2kh\dot{x}(t) \qquad \text{(A.6)}
\end{aligned}$$

With the above results, the proof can be carried out as follows.

a. Proof 1: Condition 1 is satisfied.

Taking the first derivative of Eq. (A.6) leads to

$$\frac{d}{dt}I^n x(t) = \left[\frac{1}{h^n}\sum_{k=0}^{\infty}(-1)^k \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} \cdot 2kh\right]\left[\frac{d^2}{dt^2}x(t)\right] \tag{A.7}$$

Since $\frac{1}{h^n}\sum_{k=0}^{\infty}(-1)^k \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} \cdot 2kh$ is independent of time t, condition 1 is equivalent to $\left.\frac{d}{dt}I^n x(t)\right|_{t=t_1} = 0$.

Substituting $t = t_1$ into the above equation, Eq. (A.7), yields:

$$\begin{aligned}
\left.\frac{d}{dt}I^n x(t)\right|_{t=t_1} &= \frac{1}{h^n}\sum_{k=0}^{\infty}(-1)^k \frac{2kh\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} \cdot \ddot{x}(t)\Bigg|_{t=t_1} \\
&= \frac{2}{h^{n-1}}\sum_{k=0}^{\infty}(-1)^k \frac{kn\Gamma(n)}{k\Gamma(k)\Gamma(n-k+1)} \cdot \ddot{x}(t)\Bigg|_{t=t_1} \\
&= \frac{2}{h^{n-1}}\sum_{k=0}^{\infty}(-1)^k \frac{n\Gamma(n)}{\Gamma(k)\Gamma(n-k+1)} \cdot \ddot{x}(t)\Bigg|_{t=t_1} \\
&= 0
\end{aligned}$$

Hence, $\left.\frac{d}{dt}I^n x(t)\right|_{t=t_1} = \left.\frac{d^2}{dt^2}x(t)\right|_{t=t_1} = 0$ and Condition 1 is satisfied.

b. Proof 2: Condition 2 is satisfied.

Again, as $\frac{1}{h^n}\sum_{k=0}^{\infty}(-1)^k \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} \cdot 2kh$ is independent of time t, Eq. (A.2) is equivalent to (for $n < 0$)

$$\left\{\begin{array}{l} \lim\limits_{t\to t_1-\varepsilon}\frac{d}{dt}I^n x(t) > 0 \\ \lim\limits_{t\to t_1+\varepsilon}\frac{d}{dt}I^n x(t) < 0 \end{array}\right\} \quad (\varepsilon > 0) \tag{A.8}$$

or

$$\left\{\begin{array}{l} \lim\limits_{t\to t_1-\varepsilon}\frac{d}{dt}I^n x(t) < 0 \\ \lim\limits_{t\to t_1+\varepsilon}\frac{d}{dt}I^n x(t) > 0 \end{array}\right\} \quad (\varepsilon > 0) \tag{A.9}$$

Eq. (A.8) requires (from Eq. (A.6)):

$$\left\{\begin{array}{l} \lim\limits_{t\to t_1-\varepsilon} \dfrac{d}{dt} I^n x(t) = \lim\limits_{t\to t_1-\varepsilon} \dfrac{2}{h^{n-1}} \sum\limits_{k=0}^{\infty} (-1)^k \dfrac{n\Gamma(n)}{\Gamma(k)\Gamma(n-k+1)} \ddot{x}(t) > 0 \\ \lim\limits_{t\to t_1+\varepsilon} \dfrac{d}{dt} I^n x(t) = \lim\limits_{t\to t_1+\varepsilon} \dfrac{2}{h^{n-1}} \sum\limits_{k=0}^{\infty} (-1)^k \dfrac{n\Gamma(n)}{\Gamma(k)\Gamma(n-k+1)} \ddot{x}(t) < 0 \end{array}\right\} \tag{A.10}$$

As

$$\left\{\begin{array}{l} \lim\limits_{t\to t_1-\varepsilon} \ddot{x}(t) > 0 \\ \lim\limits_{t\to t_1+\varepsilon} \ddot{x}(t) < 0 \end{array}\right\} \quad (\varepsilon > 0),$$

the above equation, Eq. (A.10), is true if and only if

$$Y(k) = \frac{2}{h^{n-1}} \sum_{k=0}^{\infty} (-1)^k \frac{n\Gamma(n)}{\Gamma(k)\Gamma(n-k+1)} > 0.$$

The following proves $Y(k)$ is indeed greater than zero.

$$\begin{aligned} Y(k) &= \frac{2}{h^{n-1}} \sum_{k=0}^{\infty} (-1)^k \frac{n\Gamma(n)}{\Gamma(k)\Gamma(n-k+1)} \\ &= \frac{2}{h^{n-1}} \sum_{k=0}^{\infty} (-1)^k \frac{n\Gamma(n)}{\Gamma(k)\Gamma(n-k)(n-k)} \\ &= \frac{2}{h^{n-1}} \left\{ \frac{(-1)^0 n\Gamma(n)}{\Gamma(0)\Gamma(n-0)(n-0)} + \frac{(-1)^1 n\Gamma(n)}{\Gamma(1)\Gamma(n-1)(n-1)} \right. \\ &\quad + \frac{(-1)^2 n\Gamma(n)}{\Gamma(2)\Gamma(n-2)(n-2)} + \frac{(-1)^3 n\Gamma(n)}{\Gamma(3)\Gamma(n-3)(n-3)} \\ &\quad \left. + \cdots + \frac{(-1)^k n\Gamma(n)}{\Gamma(k)\Gamma(n-k)(n-k)} + \cdots \right\} \\ &= \frac{2}{h^{n-1}} \left\{ 1 + \frac{-n(n-1)\Gamma(n-1)}{\Gamma(n-1)(n-1)} + \frac{n(n-1)(n-2)\Gamma(n-2)}{\Gamma(2)\Gamma(n-2)(n-2)} \right. \\ &\quad + \frac{-n(n-1)(n-2)(n-3)\Gamma(n-3)}{\Gamma(3)\Gamma(n-3)(n-3)} + \cdots \\ &\quad \left. + \frac{(-1)^k n(n-1)(n-2)\cdots(n-k+1)(n-k)\Gamma(n-k)}{\Gamma(k)\Gamma(n-k)(n-k)} + \cdots \right\} \\ &= \frac{2}{h^{n-1}} \left\{ 1 - n + \frac{n(n-1)}{\Gamma(2)} - \frac{n(n-1)(n-2)}{\Gamma(3)} + \cdots \right. \end{aligned}$$

$$+\frac{(-1)^k n(n-1)(n-2)\cdots(n-k+1)}{\Gamma(k)}+\cdots\Big\} \tag{A.11}$$

The sign of $Y(k)$ can be determined by examining each term of Eq. (A.11). If every k-th term is greater than zero, then $Y(k) > 0$. The sign of k-th term (represented as y_k) is dependent on the value of k: even k or odd k.

- Even k,

  For even k, $(-1)^k = 1$ and hence the k-th term is written as,

  $$y_{k(\text{even})} = \frac{1}{\Gamma(k)} n(n-1)(n-2)\cdots(n-k+1) \tag{A.12}$$

  Since Eq. (A.12) has even number of terms and $n < 0$ for the inflexion-point locator $I^n x(t)$, then $y_{k(\text{cvcn})} > 0$ for every even k.

- Odd k,

  For odd k, $(-1)^k = -1$ and hence the k-th term is written as,

  $$y_{k(\text{odd})} = \frac{-1}{\Gamma(k)} n(n-1)(n-2)\cdots(n-k+1) \tag{A.13}$$

  Eq. (A.13) has odd number of terms with $n < 0$ and hence the k-th term is greater than zero, i.e., $y_{k(\text{odd})} > 0$, for every odd k.

Now the the k-th term is greater than zero for both even and odd k's. Consequently, the sum of all the terms must also be greater than zero, i.e., $Y(k) > 0$, for any k.

For Eq. (A.9), as

$$\left\{ \begin{array}{l} \lim\limits_{t\to t_1-\varepsilon} \ddot{x}(t) < 0 \\ \lim\limits_{t\to t_1+\varepsilon} \ddot{x}(t) > 0 \end{array} \right\} \quad (\varepsilon > 0)$$

and

$$Y(k) = \frac{2}{h^{n-1}} \sum_{k=0}^{\infty} (-1)^k \frac{n\Gamma(n)}{\Gamma(k)\Gamma(n-k+1)} > 0,$$

this leads to

$$\left\{ \begin{array}{l} \lim\limits_{t\to t_1-\varepsilon} \frac{d}{dt} I^n x(t) = \lim\limits_{t\to t_1-\varepsilon} \frac{2}{h^{n-1}} \sum\limits_{k=0}^{\infty} (-1)^k \frac{n\Gamma(n)}{\Gamma(k)\Gamma(n-k+1)} \ddot{x}(t) < 0 \\ \lim\limits_{t\to t_1+\varepsilon} \frac{d}{dt} I^n x(t) = \lim\limits_{t\to t_1+\varepsilon} \frac{2}{h^{n-1}} \sum\limits_{k=0}^{\infty} (-1)^k \frac{n\Gamma(n)}{\Gamma(k)\Gamma(n-k+1)} \ddot{x}(t) > 0 \end{array} \right\}$$

Hence Condition 2 as expressed by both Eqs. (A.8) and (A.9) is satisfied.

Therefore, the inflexion-point locator can indeed locate the inflexion points and this completes the proof.

■

A.7 Bearing Model: Outer-race fault, Eq. (7.8)

$$
\begin{aligned}
x_{OR}(t) &= \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} A_f A_b^j e^{-\alpha_b^j t} \cos\left(2\pi i f_f t\right) \cdot \cos\left(2\pi f_b^j t\right) \\
&\quad \big(\cos(A+B) = \cos A \cos B - \sin A \sin B\big) \\
&= \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} A_f A_b^j e^{-\alpha_b^j t} \left\{\cos\left[2\pi\left(f_b^j + i f_f\right) t\right] + \sin\left(2\pi f_b^j t\right) \cdot \sin\left(2\pi i f_f t\right)\right\} \\
&\quad \left(X_1 = \cos\left[2\pi\left(f_b^j + i f_f\right) t\right]\right) \\
&\quad \big(\cos(A-B) = \cos A \cos B + \sin A \sin B\big) \\
&= \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} A_f A_b^j e^{-\alpha_b^j t} \left\{X_1 + \cos\left[2\pi\left(f_b^j - i f_f\right) t\right] - \cos\left(2\pi f_b^j t\right) \cdot \cos\left(2\pi i f_f t\right)\right\} \\
&\quad \left(X_2 = \cos\left[2\pi\left(f_b^j - i f_f\right) t\right]\right) \\
&= \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} A_f A_b^j e^{-\alpha_b^j t} \left\{X_1 + X_2 - \cos\left(2\pi f_b^j t\right) \cdot \cos\left(2\pi i f_f t\right)\right\} .
\end{aligned}
$$

From the first and the last lines of the above equation,

$$
\cos\left(2\pi i f_f t\right) \cdot \cos\left(2\pi f_b^j t\right) = \frac{1}{2}\left(X_1 + X_2\right)
$$

Therefore,

$$
\begin{aligned}
x_{OR}(t) &= \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} A_f A_b^j e^{-\alpha_b^j t} \cdot \frac{1}{2}\left(X_1 + X_2\right) \\
X_1 &= \cos\left[2\pi\left(f_b^j + i f_f\right) t\right] = \cos\left[2\pi f_b^j \left(1 + \frac{i f_f}{f_b^j}\right) t\right] \\
X_2 &= \cos\left[2\pi\left(f_b^j - i f_f\right) t\right] = \cos\left[2\pi f_b^j \left(1 - \frac{i f_f}{f_b^j}\right) t\right] \qquad \blacksquare
\end{aligned}
$$

A.8 Bearing Model: Inner-race fault, Eq. (7.12)

$$x_{IR}(t) = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{k=1}^{N_3} A_f A_l^k A_b^j e^{-\alpha_b^j t} \cos\left(2\pi k f_s t\right) \cdot \cos\left(2\pi f_b^j t\right) \cdot \cos\left(2\pi i f_f t\right)$$

The above equation can be expanded by using the following identities.

$$\cos(A+B) = \cos A \cos B - \sin A \sin B$$
$$\cos(A-B) = \cos A \cos B + \sin A \sin B$$

Then,

$$\begin{aligned}
x_{IR}(t) &= \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{k=1}^{N_3} A_f A_l^k A_b^j e^{-\alpha_b^j t} \cos\left(2\pi k f_s t\right) \\
&\qquad \cdot \left\{ \cos\left[2\pi\left(f_b^j + i f_f\right)t\right] + \sin\left(2\pi f_b^j t\right) \cdot \sin\left(2\pi i f_f t\right) \right\} \\
&= \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{k=1}^{N_3} A_f A_l^k A_b^j e^{-\alpha_b^j t} \left\{ \cos\left[2\pi\left(f_b^j + i f_f\right)t\right] \cdot \cos\left(2\pi k f_s t\right) \right. \\
&\qquad \left. + \sin\left(2\pi f_b^j t\right) \cdot \sin\left(2\pi i f_f t\right) \cdot \cos\left(2\pi k f_s t\right) \right\} \\
&= \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{k=1}^{N_3} A_f A_l^k A_b^j e^{-\alpha_b^j t} \Big\{ \cos\left[2\pi\left(f_b^j + i f_f + k f_s\right)t\right] \\
&\qquad + \sin\left[2\pi\left(f_b^j + i f_f\right)t\right] \cdot \sin\left(2\pi k f_s t\right) \\
&\qquad + \left\{ \cos\left[2\pi\left(f_b^j - i f_f\right)t\right] - \cos\left(2\pi f_b^j t\right) \cdot \cos\left(2\pi i f_f t\right) \right\} \cdot \cos\left(2\pi k f_s t\right) \Big\}.
\end{aligned}$$

By the assumption $f_b \gg f_f > f_s$ and $X_3 = \cos\left[2\pi(f_b^j + i f_f + k f_s)t\right]$,

$$\begin{aligned}
x_{IR}(t) = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{k=1}^{N_3} A_f A_l^k A_b^j e^{-\alpha_b^j t} & \left\{ X_3 + \cos\left[2\pi\left(f_b^j + i f_f - k f_s\right)t\right] \right. \\
& - \cos\left[2\pi\left(f_b^j + i f_f\right)t\right] \cdot \cos\left(2\pi k f_s t\right) \\
& + \cos\left[2\pi\left(f_b^j - i f_f\right)t\right] \cdot \cos\left(2\pi k f_s t\right) \\
& \left. - \cos\left(2\pi f_b^j t\right) \cdot \cos\left(2\pi i f_f t\right) \cdot \cos\left(2\pi k f_s t\right) \right\}.
\end{aligned}$$

Letting $X_6 = \cos\left[2\pi(f_b^j + if_f - kf_s)t\right]$, one obtains

$$x_{IR}(t) = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{k=1}^{N_3} A_f A_l^k A_b^j e^{-\alpha_b^j t}\left\{X_3 + X_6 - \cos\left[2\pi\left(f_b^j + if_f\right)t\right]\cdot\cos\left(2\pi kf_s t\right)\right.$$
$$+ \cos\left[2\pi\left(f_b^j - if_f + kf_s\right)t\right] + \sin\left[2\pi\left(f_b^j - if_f\right)t\right]\cdot\sin\left(2\pi kf_s t\right)$$
$$\left. - \cos\left(2\pi f_b^j t\right)\cdot\cos\left(2\pi if_f t\right)\cdot\cos\left(2\pi kf_s t\right)\right\}\ .$$

Similarly, letting $X_5 = \cos\left[2\pi\left(f_b^j - if_f + kf_s\right)t\right]$ yields

$$x_{IR}(t) = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{k=1}^{N_3} A_f A_l^k A_b^j e^{-\alpha_b^j t}\Big\{X_3 + X_5 + X_6 - \cos\left[2\pi\left(f_b^j + if_f\right)t\right]\cdot\cos\left(2\pi kf_s t\right)$$
$$+ \sin\left[2\pi\left(f_b^j - if_f\right)t\right]\cdot\sin\left(2\pi kf_s t\right)$$
$$- \cos\left(2\pi f_b^j t\right)\cdot\cos\left(2\pi if_f t\right)\cdot\cos\left(2\pi kf_s t\right)\Big\}$$
$$= \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{k=1}^{N_3} A_f A_l^k A_b^j e^{-\alpha_b^j t}\Big\{X_3 + X_5 + X_6$$
$$- \cos\left[2\pi\left(f_b^j + if_f\right)t\right]\cdot\cos\left(2\pi kf_s t\right)$$
$$+ \sin\left[2\pi\left(f_b^j - if_f\right)t\right]\cdot\sin\left(2\pi kf_s t\right)$$
$$- \left\{\cos\left[2\pi\left(f_b^j - if_f\right)t\right] - \sin\left(2\pi f_b^j t\right)\cdot\sin\left(2\pi if_f t\right)\right\}\cdot\cos\left(2\pi kf_s t\right)\Big\}$$
$$= \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{k=1}^{N_3} A_f A_l^k A_b^j e^{-\alpha_b^j t}\Big\{X_3 + X_5 + X_6$$
$$- \cos\left[2\pi\left(f_b^j + if_f\right)t\right]\cdot\sin\left(2\pi kf_s t\right)$$
$$- \cos\left[2\pi\left(f_b^j - if_f\right)t\right]\cdot\cos\left(2\pi kf_s t\right)$$
$$+ \sin\left[2\pi\left(f_b^j - if_f\right)t\right]\cdot\cos\left(2\pi kf_s t\right)$$
$$+ \sin\left(2\pi f_b^j t\right)\cdot\sin\left(2\pi if_f t\right)\cdot\cos\left(2\pi kf_s t\right)\Big\}$$

where the last terms can be expanded as follows.

$$Y(t) = \sin(2\pi f_b^j t)\cdot\sin(2\pi if_f t)\cdot\cos(2\pi kf_s t)$$
$$= \Big\{\cos\left[2\pi(f_b^j - if_f)t\right] - \cos(2\pi f_b^j t)\cdot\cos(2\pi if_f t)\Big\}\cdot\cos(2\pi kf_s t)$$
$$= \cos\left[2\pi(f_b^j - if_f)t\right]\cdot\cos(2\pi kf_s t)$$

$$
\begin{aligned}
&\quad - \cos(2\pi f_b^j t) \cdot \cos(2\pi i f_f t) \cdot \cos(2\pi k f_s t) \\
&= \cos\left[2\pi(f_b^j - i f_f t - k f_s)t\right] - \sin\left[2\pi(f_b^j - i f_f)t\right] \cdot \sin(2\pi k f_s t) \\
&\quad - \cos(2\pi f_b^j t) \cdot \cos(2\pi i f_f t) \cdot \cos(2\pi k f_s t) \\
&= X_4 - \sin\left[2\pi(f_b^j - i f_f)t\right] \cdot \sin(2\pi k f_s t) \\
&\quad - \cos(2\pi f_b^j t) \cdot \cos(2\pi i f_f t) \cdot \cos(2\pi k f_s t) \\
&\Big(X_4 = \cos\left[2\pi(f_b^j - i f_f - k f_s)t\right]\Big).
\end{aligned}
$$

Substituting $Y(t)$ into $x_{IR}(t)$ leads to

$$
\begin{aligned}
x_{IR}(t) &= \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{k=1}^{N_3} A_f A_l^k A_b^j e^{-\alpha_b^j t} \{X_3 + X_4 + X_5 + X_6 \\
&\quad - \cos\left[2\pi\left(f_b^j + i f_f\right)t\right] \cdot \cos\left(2\pi k f_s\right) \\
&\quad - \cos\left[2\pi\left(f_b^j - i f_f\right)t\right] \cdot \cos\left(2\pi k f_s\right) \\
&\quad + \sin\left[2\pi\left(f_b^j - i f_f\right)t\right] \cdot \sin\left(2\pi k f_s\right) \\
&\quad - \sin\left[2\pi\left(f_b^j - i f_f\right)t\right] \cdot \sin\left(2\pi k f_s\right) \\
&\quad - \cos\left(2\pi f_b^j t\right) \cdot \cos\left(2\pi i f_f t\right) \cdot \cos\left(2\pi k f_s t\right)\} \\
&= \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{k=1}^{N_3} A_f A_l^k A_b^j e^{-\alpha_b^j t} \Big\{X_3 + X_4 + X_5 + X_6 \\
&\quad - \cos\left(2\pi k f_s t\right) \left\{ \cos\left[2\pi\left(f_b^j - i f_f\right)t\right] + \cos\left[2\pi\left(f_b^j + i f_f\right)t\right] \right\} \\
&\quad - \cos\left(2\pi f_b^j t\right) \cdot \cos\left(2\pi i f_f t\right) \cdot \cos\left(2\pi k f_s t\right) \Big\} \\
&= \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{k=1}^{N_3} A_f A_l^k A_b^j e^{-\alpha_b^j t} \Big\{X_3 + X_4 + X_5 + X_6 \\
&\quad - \cos\left(2\pi k f_s t\right) \cdot \left\{ \cos\left(2\pi f_b^j t\right) \cdot \cos\left(2\pi i f_f t\right) + \sin\left(2\pi f_b^j t\right) \cdot \sin\left(2\pi i f_f t\right) \right. \\
&\qquad \left. + \cos\left(2\pi f_b^j t\right) \cdot \cos\left(2\pi i f_f t\right) - \sin\left(2\pi f_b^j t\right) \cdot \sin\left(2\pi i f_f t\right) \right\} \\
&\quad - \cos\left(2\pi f_b^j t\right) \cdot \cos\left(2\pi i f_f t\right) \cdot \cos\left(2\pi k f_s t\right) \Big\} \\
&= \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{k=1}^{N_3} A_f A_l^k A_b^j e^{-\alpha_b^j t} \Big\{X_3 + X_4 + X_5 + X_6 \\
&\quad - 3 \cdot \cos\left(2\pi f_b^j t\right) \cdot \cos\left(2\pi i f_f t\right) \cdot \cos\left(2\pi k f_s t\right) \Big\}.
\end{aligned}
$$

From the above result, one has

$$\cos(2\pi f_b^j t) \cdot \cos(2\pi i f_f t) \cdot \cos(2\pi k f_s t) = \frac{1}{4}\big(X_3 + X_4 + X_5 + X_6\big).$$

Therefore,

$$\begin{aligned}
x_{IR}(t) &= \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{k=1}^{N_3} A_f A_l^k A_b^j e^{-\alpha_b^j t} \cos\left(2\pi k f_s t\right) \cdot \cos\left(2\pi f_b^j t\right) \cdot \cos\left(2\pi i f_f t\right) \\
&= \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{k=1}^{N_3} A_f A_l^k A_b^j e^{-\alpha_b^j t} \cdot \frac{1}{4}\left(X_3 + X_4 + X_5 + X_6\right) \\
X_3 &= \cos\left[2\pi\left(f_b^j + i f_f + k f_s\right) t\right] \\
X_4 &= \cos\left[2\pi\left(f_b^j - i f_f - k f_s\right) t\right] \\
X_5 &= \cos\left[2\pi\left(f_b^j - i f_f + k f_s\right) t\right] \\
X_6 &= \cos\left[2\pi\left(f_b^j + i f_f - k f_s\right) t\right]\ . \qquad \blacksquare
\end{aligned}$$

Appendix B

MATLAB programs

B.1 Finding the best scale level from CWT

```
function [aBest] = func_findBestScaleCWT(kurt, Ecwtx, Kurt_Threshold)

    % Inputs:
    %   kurt :  kurtosis distribution over scales
    %   Ecwtx:  energy distribution over scales
    %   Kurt_Threshold: threshold value, e.g. 4.0
    %
    % Outputs:
    %   aBest:  the best scale level selected by Energy & Kurtosis
    %
    % Get the Energy Peaks by using:
    %   1) Get envelope of energy curve to minimize
    %      the adverse effect of ripples caused by mostly noise
    %   2) Check the sign of derivative and find positive peaks
    %      changing sign from "+1" to "-1".
    %   3) Check the kurtosis values > 4.0
    %
    % -----------------------------------------------------------------
    % Example,
    %
    %   for i = 1:10
    %        kurt(i) = kurtosis(cwtx(i,:));
    %        Ecwtx(i) = sum(cwtx(i,:).^2);
    %   end
    %   Kurt_Threshold = 4.0;
```

```
    %   aBest = func_findBestScaleCWT(kurt, Ecwtx, Kurt_Threshold);
    % ----------------------------------------------------------------

    % Condition 1: Energy Peak
    % Step 1-a: Find scales with energy peaks
    %       (may be single/multiple numbers)
    eX = Ecwtx;
    dX = sign(diff(eX));    % Get sign of the first derivative of x(t).
    dX = [dX(1) dX];        % Since the derivative is one datum less than
                            % the original signal, add zero in the first cell.

    % Step 1-b: Differentiate dX
    %   where dX is the sign function of the first derivative of x(t).
    ddX = -1*diff(dX)/max(abs(diff(dX)));
    %   where "-1" is multiplied as "diff(dX)" gives negative value
    %   at the positive peak as shown in the Figure (c).
    ddX = [ddX, 0];

    % Step 1-c: Get the scales with energy peak,
    %   i.e., index with "ddX == 1" is peak location.
    %   See Figure (d)
    aPeak = find(ddX == 1);

    % Condition 2: Kurtosis Value
    % Step 2-a: Select the scales with "kurtosis > Kurt_Threshold"
    %   where "Kurt_Threshold" could be any value greater than 3.0.
    %   In this example, Kurt_Threshold = 4.0
    aPidx = find(kurt(aPeak) > Kurt_Threshold);
    %   where aPidx are scales with (energy peak) & (kurtosis > 4.0)

    % Step 2-b: Find maximum kurtosis
    %   As multiple number of scales may have kurtosis value greater than
    %   Kurt_Threshold, find the scales with the highest kurtosis value
    %   among them.
    maxKidx = find(max(kurt(aPidx))==kurt(aPidx));
    %   maxKidx is index for scales with the highest kurtosis value.
    %
    aBest = aPeak(maxKidx);
    %   aBest is the best scale with (Energy Peak) & (Kurtosis > 4.0)
% End of function
%
% --------------------------------------------------------------------
```

```
% To illustrate the concept, run the following script.
%     clear;
%     N = 1024;
%     for i = 1:N
%         x(i) = sin(2*pi*2*(i-1)/N);
%     end
%
%         eX = x;
%         % Select peaks of "eX"
%         % Step 1: Get sign of the first derivative of x(t)
%         dX = sign(diff(eX));
%             % where function "sign( )" is MATLAB buit-in function
%         dX = [dX(1) dX];     % Since the derivative is one datum less than
%                              % the original signal, add zero in the first cell.
%
%         % Step 2: Differentiate dX
%         %   where dX is the sign function of the first derivative of x(t)
%         ddX = -1*diff(dX)/max(abs(diff(dX)));
%         %   where "-1" is multiplied as "diff(dX)" gives negative value
%         %   at the positive peak as shown in the Figure (c).
%         ddX = [ddX, 0];
%
%     figure(1);
%         subplot(411), plot(x);
%             title('(a) Sinusoidal function x(t)');
%             axis([0 N -1.5 1.5]);
%         subplot(412), plot(dX);
%             title('(b) Sign of the first derivative of x(t)');
%             axis([0 N -1.5 1.5]);
%         subplot(413), plot(-ddX);
%             title('(c) Differentiation of (b)');
%             axis([0 N -1.5 1.5]);
%         subplot(414), plot(ddX);
%             title('(d) [-1 x (c)], Peak location');
%             axis([0 N -1.5 1.5]);
% End
```

B.2 Extracting periodic fault feature by CWT & Autocorrelation

```
function [xFiltered] = func_CorrExtract(aBest, cwtx)

    %   Inputs:
    %       aBest:      best scale level
    %       cwtx:       wavelet coefficients
    %   Output:
    %       xFiltered:  extracted periodic feature in time-scale domain
    % ---------------------------------------------------------------

    N = length(cwtx);

    % Calculate Auto-correlation of wavelet tranformed signal
    %   at scale level "aBest" and get Envelope.
    Rxx = xcorr(abs(cwtx), abs(cwtx), 'coeff');

    % get envelope of autocorrelation coefficient function
    [RHL, dnEnv] = func_GetEnvelope(-N+1:N-1, Rxx, 'spline');


    % Remove inclined baselines
    [env1, medFilter] = func_GetEnvelope(-N+1:N-1, RHL, 'linear');
    %   where "env1" is the upper envelope (connecting upper peaks),
    %         "medFilter" is the lower envelope
    %         (obtained by connecting lower peaks).

    % medFilter is the inclined baseline
    medRHL = (RHL - medFilter);      % medRHL is PUSH-DOWNed signal
    medRHL = medRHL / max(medRHL);   % normalization for max. vlaue = 1


    % Select peaks from auto-correlation function "medRHL"
    [IDX, yPeak] = func_findCorrPeak(medRHL, length(medRHL));
    %   where IDX is the vector of index of peaks
    %         yPeak is the vector of peaks.

    % Filter based on Normal Distribution with
    %    standard deviation "stdIDX"
    defaultWidth = 4;        % initialization of filter width
```

```
dIDX = diff(IDX);        % gap between selected IDXs
meanIDX = mean(dIDX);    % mean value of dIDX
stdIDX = std(dIDX);      % standard deviation of dIDX
if stdIDX <= defaultWidth    % For ideal situation.
    stdIDX = defaultWidth;   % Does NOT happen in reality.
end
CFilt = zeros(1, N);     % initialization of Comb Filter
if mod(round(stdIDX), 2) ~= 0
    M = 3 * (round(stdIDX)+1);      % Half of the filter size
else                                % 3 times of standard deviation.
    M = 3 * round(stdIDX);
end
% generate bandpass filter following normal distribution
%   with size of "2*M" for both sizes (- & + time lags)
%   size "M" for negative time lag and
%   size "M" for positive time lag.
j = 1;
for i = -M:M
    % normal distribution "yfil" with yfil(0) = 1.0
    % (No amplification or shrinkage)
    yfil(j) = exp( -i^2/(2*stdIDX^2));
    j = j + 1;
end

% Find the location of the first impulse.
% Since the correlations of x(t) and x(t+dt) are same,
% the location of the first peak should be considered.
% Check the translation in time.
ytmp = abs(cwtx(IDX(1):IDX(2)));
[eytmp1, eytmp2] = func_GetEnvelope(IDX(1):IDX(2), ytmp, 'linear');
maxidx1arry = find(max(eytmp1)==eytmp1);
if length(maxidx1arry) > 1
    maxidx1 = maxidx1arry(1);
else
    maxidx1 = maxidx1arry;
end
idxFirstPeak = maxidx1;      % index for the 1st impulse

% Compose Comb filter
% for the first impulse
midM = M+1;
if idxFirstPeak >= M
```

```
        % the first peak is not inside the filter width (2*M).
        CFilt(idxFirstPeak-M:idxFirstPeak+M) = yfil;
    else
        % the first peak is inside the filter width (2*M).
        CFilt(1:idxFirstPeak+M) = yfil(midM-idxFirstPeak:M);
    end
    % from the second impulse to the last
    for j = 1:length(IDX)-1
        i = idxFirstPeak + IDX(j);
        CFilt(i-M:i+M) = yfil;
    end

    % Comb filtering of wavelet coefficients at the scale "aBest"
    xFiltered = cwtx .* CFilt;
% End of function
```

B.3 Finding peaks from autocorrelation coefficient function

```
function [xpeak, ypeak] = func_findCorrPeak(X, N)

    % Inputs:
    %   X:  auto-correlation function
    %   N:  length of X
    % Outputs:
    %   xpeak:  index for peak
    %   ypeak:  peak value at the index xpeak
    % ----------------------------------------------------------------

    % Select peaks of "X"
    X = X.^2;          % square X to make the peaks more distinctive.
    RATIO = 0.15;      % used to select peaks of auto-correlation
                       % with appropriate magnitude decay, e.g.
                       % the next peak should be 15% less of the current peak
                       % in magnitude (RATIO could be in [0.1, 0.3]).

    mid = (N-1)/2 + 1;  % index for (time lag) = 0.

    % Search peaks for "positive time lag" part
    %   since the correlation function is symmetric about (time lag) = 0.
    tmp(1) = mid;
    k = 2;
    for j = mid+1:N-1
        if X(j) >= X(j-1) & X(j) >= X(j+1) & X(j) >= RATIO*X(tmp(k-1))
            % where the peak X(j) should be greater than X(j-1) and X(j+1).
            tmp(k) = j;
            k = k + 1;
        end
    end

% Check the distance between indexes for double checking.
% Since the input signal is periodic, the distance between indexes
% should be same in ideal situation. However, it is natural that
% the distance obtained from the correlation coefficient function has
% little error caused by the reduced number of data samples (due to time lag)
% involved in the calculation of correlation function.
% Find avgGap from the highest probability of tmp
```

```
    [a2, b2] = hist(abs(diff(tmp)));          % positive side
    %   where function "hist()" and "diff() are MATLAB built-in functions.
    avgGap2 = mean(b2(find(a2==max(a2))));    % most probable value for tmp
    %   where "mean()", "find()", and "max()" are MATLAB buit-in functions.

    Factor = 0.1;
        % Used for checking distance between peaks
        % if distance difference is greater than this Factor,
        %    the peak is not considered as the one corresponding to impulse.
    % positive time-delay part
    IDX(1) = tmp(1);
    %   the first one is always the peak with correlation coeff=1.0.
    k = 2;
    for j = 2:length(tmp)
        Gap2(j-1) = abs(tmp(j) - IDX(k-1));
        GapR2(j-1) = abs(Gap2(j-1) / avgGap2 -1);
        if GapR2(j-1) <= Factor
            IDX(k) = tmp(j);
            k = k + 1;
        end
    end

    % Check the number of impulses
    % if there is time left greater than avgGap2,
    %    there must be another impulses.
    % Since the auto-correlation function gets worse as the time lag
    % increases, this "time-gap check" can compensate for the difficulty
    % finding peaks of auto-correlation function when the time lag is large.
    marginX = N - max(IDX);
    LenIDX = length(IDX);
    if marginX > avgGap2
        for i = 1:floor(marginX/avgGap2)
            IDX(LenIDX+i) = IDX(LenIDX) + i*floor(avgGap2);
        end
    end

    xpeak = IDX(2:length(IDX)) - (N+1)/2;
    ypeak = X(xpeak);

    xvalue = -(N+1)/2+1:(N+1)/2-1;

% End of function
```

B.4 K-hybrid: kurtosis based de-noising

```
function [srec] = func_Khybrid(x, LVL)

    % Inputs:
    %   x:          measured signal mixture
    %   LVL:        max scale level used in CWT, e.g. scale=[1, LVL]
    %
    % Outputs:
    %   srec:       de-noised signal by K-hybrid
    % ----------------------------------------------------------------

% Apply CWT & calculate kurtosis distribution
    cwtns = cwt(x, 1:LVL, 'morl');
    %   where "cwt()" is a MATLAB built-in function.
    %   'morl' -> morlet wavelet is used as a basis wavelet function.
    for i = 1:LVL
        kurt(i) = kurtosis(cwtns(i,:));
        %   calculate kurtosis at each scale by using
        %   a MATLAB built-in function, "kurtosis()".
    end
    M = max(size(cwtns));
    %   where "max()" is a MATLAB built-in function.

% Reconstruction Coefficient for Morlet wavelet
omega0 = 5;    % Frequency of Morlet Wavelet Function
beta = 1;      % Decay Parameter of Morlet Wavelet Function
                   % Morlet Wavelet function
                   %    = exp(- beta^2 * t^2 / 2) * exp(i*omega0*t)
sumc = 0;
sumd = 0;
dw = 0.02;
for i = 1:dw:100      % from zero to +infinite
        omega = i;
        morl_fft = (exp(-(omega-omega0)^2/(2*beta^2))/beta);
        %   Fourier Transform of Morlet Wavelet
        sumc = sumc + morl_fft/omega*dw;
end
for i = -100:dw:-1    % from -infinite to zero
        omega = i;
        morl_fft = (exp(-(omega-omega0)^2/(2*beta^2))/beta);
        %   Fourier Transform of Morlet Wavelet
```

```
        sumd = sumd+morl_fft/abs(omega)*dw;
    end

CC = sumc + sumd;       % Reconstruction Coefficient

% Start Thresholding
    for i = 1:LVL
        Kt = 1 + max(kurt(i)-3, 0);
        %   Kt is parameter used to adjust zone width.
        t2(i) = max(abs(cwtns(i,:))) / (kurt(i)/3);
        %   basis threshold.
        t1(i) = t2(i) / Kt;
        t3(i) = (Kt + 1/sqrt(2))*t1(i);

        r1(i) = (Kt-1)*t1(i);
        r2(i) = (1+sqrt(2))*t1(i);
        %   r1() and r2() are two radii used in zone 2 & zone 3.

        for k = 1:M
            ct = abs(cwtns(i,k));
            sgn = sign(cwtns(i,k));
            %   "sign()" is a MATLAB built-in function

            % Region 4 : cwtns >= t3
            if ct >= t3(i)
                cwtns_fil(i,k) = cwtns(i,k);

            % Region 3 : t2 <= cwtns < t3
            elseif ct >= t2(i) & ct<t3(i)
                cwtns_fil(i,k) = sgn*(r1(i) ...
                    + sqrt(abs(-ct^2 + 2*ct*(r2(i)+t2(i)) ...
                    - t2(i)*(2*r2(i)+t2(i)))));

            % Region 2 : t1 <= cwtns < t2
            elseif ct >= t1(i) & ct < t2(i)
                cwtns_fil(i,k) = sgn*(r1(i) ...
                    - sqrt(((r1(i)^2-t1(i)^2) + (2*t1(i)-ct)*ct)));

            % Region 1 : cwtns < t1
            else
                cwtns_fil(i,k) = 0;
            end
```

```
        end
    end

% Reconstruction
for k = 1:M
        sum1 = 0;
        for i = 1:LVL
            sum1 = sum1 + cwtns_fil(i,k)*i^(-3/2);
            %   Morlet formula
        end
        srec(k) = sum1/CC;
    end
% End of function
```

B.5 Oil-debris signature extraction by fractional integration

```
function [HnX, InX] = func_fractCalculus(x, n, Kmax)

    % Inputs:
    %   Kmax:   number of terms used for summation
    %   n:      non-integer order (n<0)
    %   x:      input signal ### (Should be N-by-1 Column Vector) ###
    % Outputs:
    %   HnX:    output of detector H^n{x(t)}, particle signature
    %   InX:    output of detector I^n{x(t)}, inflexion points
    % -----------------------------------------------------------------

% Check the input arguments
    if nargin < 3
        Kmax = 200;     % Data length used for summation
    elseif nargin < 2
        error('Check Input variables!!\n');
    end
% Start the fractional integration with negative non-integer order n
    % maximum amplitude of signal x(t).
    %   this value is used in the signature extractor, H^n, later.
    xGain = max(abs(x));
    x = x / xGain;
    N = length(x);

    % Calculate the summation term used in Eq. (5.9) and (5.11).
    %
    k = 0:Kmax-1;
    ak = ( (-1).^k * gamma(n+1) ./ ( gamma(k+1).*gamma(n-k+1) ) )';
    %   where "ak" is
    %   ak = 1/h^n * \sum_{k=0}^{\infinity} [ TERM1 ]
    %   TERM1 = (-1)^k * Gamma(n+1) / {Gamma(k+1)*Gamma(n-k+1)}
    %   See Eqs. (5.9) and (5.11) of thesis.

    for i = 1:N
        sumDpI = 0;
        sumDnI = 0;
        sumInD = 0;
```

```
        sumI = 0;
        sumD = 0;

        % Use flipud(Y2) if ak and Y2 are N-by-1 row vector
        %     fliplr(Y2) if ak and Y2 are 1-by-N col. vector
        %     "flipud()" and "fliplr()" are MATLAB built-in functions.
        if i <= Kmax
            sumGen = sum( ak(1:i) .* flipud(x(1:i)) );
            sumOpp = sum( ak(1:Kmax) .* x(i:i+Kmax-1) );
        elseif Kmax < i & i < N-Kmax
            sumGen = sum( ak(1:Kmax) .* flipud(x(i-Kmax+1:i)) );
            sumOpp = sum( ak(1:Kmax) .* x(i:i+Kmax-1) );
        elseif N-Kmax <= i
            sumGen = sum( ak(1:Kmax) .* flipud(x(i-Kmax+1:i)) );
            sumOpp = sum( ak(1:N-i) .* x(i:N-1) );
        end
        dnGen(i) = sumGen;
        dnOpp(i) = sumOpp;

    end

    dnGen = xGain * dnGen / max(abs(dnGen));
    %   dnGen = [ x(t) - x(t-h) ] / h
    dnOpp = xGain * dnOpp / max(abs(dnOpp));
    %   dnOpp = [ x(t) - x(t+h) ] / h
    dnOpG = xGain * ( dnOpp + dnGen ) / max(abs( dnOpp + dnGen ));
    dnOnG = ( dnOpp - dnGen ) / max(abs(dnOpp - dnGen));

    HnX = dnOpG;    % Detector H^n: output of Signature extractor
    InX = dnOnG;    % Detector I^n: output of Inflexion-point locator

% End of function
```

B.6 Single-channel signal separation

```
function [S, Ns, feasF] = func_SingleChBSS(x, aMax, Fs, minF, maxF)

    %   Inputs:
    %       x:      measured signal mixture
    %       aMax:   maximum scale level used in CWT, [1,2,...,LVL]
    %       Fs:     sampling freqeuncy in Hz
    %       minF:   min. frequency of interest
    %       maxF:   max. frequency of interest
    %   Outputs:
    %       S:      separated source signals in time domain
    %       Ns:     number of sources
    %       feasF:  frequency index of the source signal
    % ----------------------------------------------------------------

    N = length(x);

    % STEP 1: CWT
        cwtx = cwt(x, 1:aMax, 'morl');

    % STEP 2: Frequency domain
        % 2-a: FFT and Power spectral density (Pxwtfft) at each scale
        f = Fs * (0:N/2-1)/N;
        for i = 1:aMax
            up(i,:) = abs(hilbert(cwtx(i,:)));
            Xwft(i,:) = fft(up(i,:), N);
            Pxwtfft(i,:) = Xwft(i,:) .* conj(Xwft(i,:)) / N;
        end

        % 2-b: Find feasible frequencies
        % Condition 1:
        %   frequency should be between min. and max. frequencies of
        %   interest.
        % min. and max. frequencies of interest
        % Eqs. (6.4) and (6.5) of Thesis.
        % See Section 6.1.2.1 of Thesis
        indx1 = find(f==minF) - 5;  % "-5" is to add little margin
        indx2 = find(f==maxF) + 5;  % "+5" is to add little margin
                                    % to surely cover all range of
                                    %   frequency of interest.
```

```
% Select frequency component which gives maximum magnitude
% Dominant frequencies. see Section 6.1.2.1 of Thesis
% max. magnitude and its frequency index
for i=1:aMax
    maxAmp(i) = max(Pxwtfft(i, indx1:indx2));
    ind_max(i) = find(maxAmp(i) == Pxwtfft(i, 1:N/2));
end
% sort in ascending order for the next calaculation
ind_max = sort(ind_max');

% Condition 2:
%   repeated frequencies and harmonics should be removed.
% Remove frequencies with single occurence
[npeak,xpeak] = hist(ind_max, minF:1:maxF);
j = 1;
for i = 1:length(minF:1:maxF)
    if npeak(i) > 1
        Xpeak(j) = xpeak(i);
        j = j + 1;
    end
end
% Remove repeated peaks
XpeakRepeat(1) = Xpeak(1);
j = 1;
for i = 2:length(Xpeak)
    if Xpeak(i) ~= XpeakRepeat(j)
        j = j + 1;
        XpeakRepeat(j) = Xpeak(i);
    end
end
% Remove harmonics
k = 1;
feasF(k) = XpeakRepeat(1);
for i = 1:length(XpeakRepeat)
    fTmp = XpeakRepeat(i);     % index for feasible frequency
    for j = i:length(XpeakRepeat)
        Remnt = mod(f(XpeakRepeat(j)), f(fTmp));
        if Remnt ~= 0
            k = k + 1;
            feasF(k) = XpeakRepeat(j);  % index for Source
        end
    end
```

```
    end
    % Number of sources
    Ns = length(feasF);

    % 2-c: Contribution ratio
    %   calculate sum of PSD at feasible dominant frequencies
    %   in each scale
    for i = 1:aMax
        sumPscale(i) = sum(Pxwtfft(i, feasF));
        for j = 1:Ns
            Ps(i, j) = Pxwtfft(i, feasF(j));
            rMM(i, j) = Ps(i, j) / sumPscale(i);
            % Ps(i, j) is the magnitude of frequency spectrum
            %   of source j (frequency) at scale i
            % rMM(i, j) is contribution ratio of scale i on to source j
        end
    end

    % 2-d: Reconstruct the source signals
    % by using s(1) = MM(1,1)*Xwt(1) + MM(2,1)*Xwt(2) +
    %               .... + MM(aMax,1)*Xwt(aMax)
    %           s(2) = MM(1,2)*Xwt(1) + MM(2,2)*Xwt(2) +
    %               .... + MM(aMax,2)*Xwt(aMax)
    % where, s, MM, Xwt are time function.
    %
    % S_est(Scale, Time, Source)
    for i = 1:Ns
        for j = 1:aMax
            if rMM(j, i) >= 0.5
                S_est(j,:,i) = rMM(j,i)*cwtx(j,:);
            else
                % This part is used to remove scales with small
                % contribution. This action reduces not only the
                % magnitude of the source signal but also strength
                % of un-related signal.
                S_est(j,:,i) = 0.0*cwtx(j,:);
            end
        end
    end

% STEP 3: Inverse CWT
    % Reconstruction Coefficient for Inverse CWT
```

```
        omega0 = 5;    % Frequency of Morlet Wavelet Function
        beta = 1;      % Decay Parameter of Morlet Wavelet Function
        %              % Morlet Wavelet = exp(- beta^2 * t^2 / 2) *
        %              % exp(i*omega0*t)
        sumc = 0;
        sumd = 0;
        dw = 0.02;
        for i = 1:dw:100     % from 0 to +infinite
            omega = i;
            % Fourier Transform of Morlet Wavelet
            morl_fft = (exp(-(omega-omega0)^2/(2*beta^2))/beta);
            sumc = sumc+morl_fft/omega*dw;
        end
        for i = -100:dw:-1  % from -infinite to 0
            omega = i;
            % Fourier Transform of Morlet Wavelet
            morl_fft = (exp(-(omega-omega0)^2/(2*beta^2))/beta);
            sumd = sumd + morl_fft/(omega)*dw;
        end
        CC = sumc + sumd;     % Reconstruction Coefficient

        % Reconstruction or Inverse CWT
        %  Each column of matrix S_est is
        %   S_est(Scale, Time, Source)
        dk = 1;
        for i = 1:Ns
            for j = 1:N
                SumS = 0;
                for k = 1:dk:aMax
                    SumS = SumS + S_est(k,j,i) * k^(-3/2) * dk;
                end
                S(i,j) = SumS/CC;
            end
        end
% End of function
```

B.7 Calculating the normalized Lempel-Ziv complexity

```
function [CN] = func_Complexity(N, x, Tag)

    % Inputs:
    %   N:      number of input data samples
    %   x:      input signal
    %   Tag:    Conversion to binary? ('Y' or 'N')
    % Outputs:
    %   CN:     normalized Lempel-Ziv complexity
    % -----------------------------------------------------------------

% Convert the input signal into binary signal (0 or 1)
    if Tag == 'Y'
    % need to convert the signal into binary
        m = mean(x);
        for i = 1:N
            if x(i) >= m
                S(i) = 1;
            else
                S(i) = 0;
            end
        end
    elseif Tag == 'N'
    % already binary signal
        S = x;
    end

%  Start calculating Lempel-Ziv complexity
    c = 1;  % initialization of complexity index
    SQ = num2str(S(1));
    idx2 = 'NC';

    for i = 2:N
        SS = num2str(S(i));
        if idx2 == 'NC'
            Q = SS;
        elseif idx2 == 'CC'
            Q = cat(2, Q, SS); % Attach S(i) to Q(i-1)
                               % to get the new Q vector
```

```
        end

        idx = strfind(SQ, Q);
            % if idx = empty, No Match - Q(i) does not belong to SQ(j)
            %   As Q(i) is new, increase complexity c = c + 1.
            % if idx ~= empty, Q(i) belongs to SQ(j)
        if isempty(idx) == 1
            c = c + 1;
            idx2 = 'NC';
        else
            idx2 = 'CC';
        end
        SQ = cat(2, SQ, SS);
            % SQ is vocabulary or strings already read and checked.
    end

    CN = c / (N/log2(N));    % normalized Lempel-Ziv complexity
% End of function
```

B.8 Get envelope signal by interpolation

This function is adopted from *MATLAB Central, File Exchange* accessible at

http://www.mathworks.com/matlabcentral/fileexchange/
loadFile.do?objectId=3142&objectType=file.

```
function [upEnv, dnEnv] = func_GetEnvelope(x, y, interpMethod)

    %   Inputs:
    %    x                the abscissa of the given data
    %    y                the ordinate of the given data
    %    interpMethod     the interpolation method
    %                     default interpolation method : 'Spline'
    %   Outputs:
    %    upEnv      the upper envelope with the same length as x.
    %    dnEnv      the lower envelope with the same length as x.
    %
    %   Designed by: Lei Wang, <WangLeiBox@hotmail.com>, 11-Mar-2003.
    %   Last Revision: 21-Mar-2003.
    %   Dept. Mechanical & Aerospace Engineering, NC State University.
    %   $Revision: 1.1 $  $Date: 3/21/2003 10:33 AM $
    % ---------------------------------------------------------------

    if length(x) ~= length(y)
        error('Two input data (Horz. & Vert.) ...
               should have the same length.');
    end

    if (nargin <2) | (nargin > 3)
        error('Please Check the INPUT ARGUMENTS.');
    elseif nargin == 2
        interpMethod = 'spline';  % default interpolation method
    end

    % Find the Positive Extrema and the corresponding indexes
    %   for UPPER Envelope.
    extrPosIndex = find(diff(sign(diff(y))) == -2) + 1;
    extrPosValue = y(extrPosIndex);

    % Find the Negative Extrema and the corresponding indexes
    %   for LOWER Envelope.
```

```
    extrNegIndex = find(diff(sign(diff(y))) == 2) + 1;
    extrNegValue = y(extrNegIndex);

    % Include the Fist and Last Data Points
    nmax = length(x);
    up_x = [x(1) x(extrPosIndex) x(nmax)];
    up_y = [y(1) extrPosValue y(nmax)];
    dn_x = [x(1) x(extrNegIndex) x(nmax)];
    dn_y = [y(1) extrNegValue y(nmax)];

    % Interpolation of the upper/Lower envelope data
    %    between Extrema obtained above.
    upEnv = interp1(up_x, up_y, x, interpMethod);
    dnEnv = interp1(dn_x, dn_y, x, interpMethod);
% End of function
```