NOTE TO USERS

Fault Detection of Rolling Element Bearings

Chao-Shih Liu

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2005

Program Authorized to Offer Degree: Mechanical Engineering

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

University of Washington
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

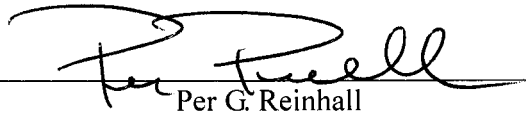
Chao-Shih Liu

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of Supervisory Committee:

Per G. Reinhall

Reading Committee:



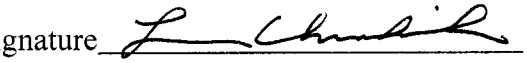
Per G. Reinhall

Les E. Atlas

Wei-Chih Wang

Date: 8-5-05

In presenting this dissertation in partial fulfillment of the requirements for the doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of the dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to Proquest Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature 

Date July 1 2005

University of Washington

Abstract

Fault Detection of Rolling Element Bearings

Chao-Shih Liu

Chairperson of the Supervisory Committee:
Professor Per G. Reinhall
Department of Mechanical Engineering

The early detection of bearing failure requires methods that are sensitive to impulsive signals and their changes. The difficulty of early bearing defect detection is that the signal component generated by early defect is very small compared to other vibration sources. It was found that techniques focus at signal changes in the higher frequency is more sensitive to early bearing defect because the periodic impact generated by early bearing defect will resonant at high frequency regions with less noise interference. The most popular technique is the high frequency resonant technique (HFRT) by demodulating the envelope of the bandpass signal using Hilbert transform. There are two major limitations of using the HFRT. First of all, the HFRT demands on knowledge of the resonance frequency range where the defect generated impulses is more pronounced with respect to normal system vibrations. The second limitation is when there are multiple defects developed at the same time, a traditional HFRT may not be able to identify each defects due to the overlap of each transient component.

A resonant frequency band selection method based on the wavelet packet transformation is developed to search the optimal resonant packet automatically using kurtosis and changes of signal energy.

Two new HFRT algorithms based on wavelet packet transformation and quadratic energy detector are developed. The results from the synthetic bearing defect signal shown that the performance of newly proposed HFRT algorithms are sensitive to early bearing defect and are able to detect multiple defects which cannot be detected by the traditional HFRT method based on Hilbert transform. A series of experiments using recorded signals from both the artificially defected bearings and bearings running from normal till failure are performed. It was found that both methods work well for field applications.

When the signal to noise ration is low, we can pre-process the signal with harmonic analysis to extract the fault related impulse component to improve the performance of HFRT algorithms for bearing defect detection.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Acknowledgements

First of all, I would like to thank my supervisor Professor Reinhall, for his inspiring and encouraging way to guide me to a deeper understanding of knowledge work, and his invaluable comments during the whole work with this dissertation. A special thanks goes to Professor Atlas, who shares many ideas on signal processing techniques. I also like to thank Professor Bassingthwaighte for always being supportive and introducing me the field of fractal and chaos. Thanks to Dr. Fabien to serve in my committee and giving me advices on my research. To Dr. Wei-Chih Wang, I extend my gratitude for being a friend and providing advice on my dissertation. It has also been a pleasure to work with you on teaching ENGR100 class.

I would like to thank all of my lab mates and friends over the years including Joseph Chao, Peter Huang, Chaw-Wu Tseng, Richard Chu etc. who have put up with me being a recluse while I worked on my thesis. Thanks to Jessica Yellin for proof read my dissertation and offering helps in many ways. Thanks to Bill Constantine for pulling me away mid-PhD to do a very interesting beer brewing project. Hope there are many more to come Bill! Thanks to Eric Hofbeck for sharing his bearing data. Thanks to Rey Panergo for helping me to take SEM pictures.

Finally, the most thanks must go to my wife Hui Lan who has shown such great support not only for this PhD, but throughout my whole life. Much of what I am, I owe to you.

Dedication

To my family, for all their love and support.

Chapter 1 Introduction

1.1 Machinery Maintenance Programs

The major dictionaries define *maintain* as *cause to continue* (Oxford) or *keep it an existing state* (Webster). From the engineering point of view, the role of maintenance is to reduce risks due to machine failure in a cost effective manner. In industry, maintenance accounts for 15% to 40% of manufacturing costs [1,48,92]. A minor improvement in a maintenance program can save as much as 50-60% of maintenance costs [49]. Currently, three common methods for performing maintenance on machinery exist. These methods are reactive, time-based preventive and condition-based predictive maintenance [1,2].

1.1.1 Reactive maintenance

Reactive maintenance, also known as known breakdown maintenance or run-to-fail maintenance, is typically the most expensive maintenance method. Using this method, machinery is repaired or replaced after failure occurs. As crude as this method, it is still the most common form of maintenance practiced today because the expenditures for instrumentation and personnel training are minimal. The reactive maintenance method results in obvious inefficiencies with significant costs due to unplanned downtime, overtime cost, and damaged machinery.

1.1.2 Time-based preventive maintenance

Because of the necessity of avoiding unexpected breakdowns, maintenance has logically changed from being reactive to preventive maintenance (PM). When performing preventive maintenance one does not wait until the machine breaks down but schedules regular inspections of the critical components. This often means that the machine is par-

tially or totally disassembled to allow for visual inspection. Even though this method assists in accurate diagnosis of some faults, unnecessary disassembling often induces secondary damage and high down time cost [2]. Additionally, PM might miss the faults that are just beginning to occur, which might lead to a serious situation in the near future.

1.1.3 Condition-based predictive maintenance

In order to avoid unnecessary maintenance work and expense, a condition-based predictive maintenance (PdM) program has been developed. The PdM program has been recognized as the most cost-effective maintenance program because it can [2,6]:

- Minimize the cost of unnecessary and ineffective maintenance;
- Improve personnel and environmental safety;
- Reduce the quantity of unexpected and unplanned breakdowns to maximize production availability.

In order to have the above benefits, an effective PdM program must be able to accurately characterize machinery condition and provide early warning of imminent failure. A typical process of a PdM system includes three steps: data acquisition; extraction of features for monitoring; and classification of these features to diagnose the type and cause of a fault (see Figure 1-1).

There are many kind of measurements can be used for monitoring, and which method is optimal for a given situation depends on the type of machinery, accessibility and purpose of monitoring. Commonly used monitoring measurements are vibration, wear debris, acoustic, current, temperature and pressure [92]. Among them, vibration is considered the most obvious representative of machine operation and the best operating parameter to determine the health condition of machinery [4,50-53].

Feature extraction is a vital step to PdM performance. By analyzing collected measurements, the PdM system must be able must be able to reduce the amount of data, identify

and select the best fault-indicating features for classification and diagnosis. If the features are not well chosen, the PdM may give frequent false alarms and be insensitive to a legitimate failure condition [5].

The ultimate goal of a PdM system is to determine the machinery condition and diagnose the actual problem. Trending is a popular method to determine changing of machinery condition by plotting measurements as a function of time. The change of magnitude as well as the rate of change from the trending plot can provide information about the degradation of machinery condition. At last, to recognize what needs to be repaired is a classification problem. A simple way to implement this is to examine individual spectral components in the spectrum. More advanced techniques that depend on artificial intelligence, such as neural nets [41-45], genetic algorithm and fuzzy logic [46,47], are becoming a standard part of the maintenance world [92].

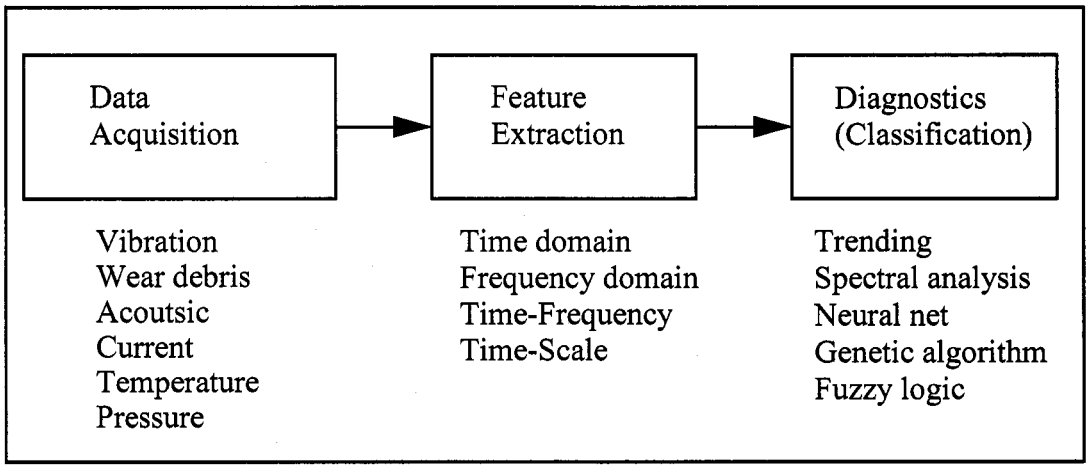


Figure 1-1. Typical process of a predictive maintenance system

1.2 Why Is Bearing Defect Detection Important?

In this dissertation, we will focus on condition monitoring of rolling element bearings because they are the most widely used component in rotating machinery and have long

presented a difficult task for PdM programs [54]. The consequences of bearing failure may cause substantial economic loss and catastrophic failure.

Nearly all pumps and motors of different sizes operate with rolling element bearings and bearing failure accounts for 51% of electric motor failure (see Figure 1-2) [88]. In addition, bearing defects can also be a warning signal for other fault symptoms because imbalance and misalignment are very often the cause of bearing defects. Therefore, condition monitoring and diagnostics of rolling element bearings are an important issue in the maintenance industry.

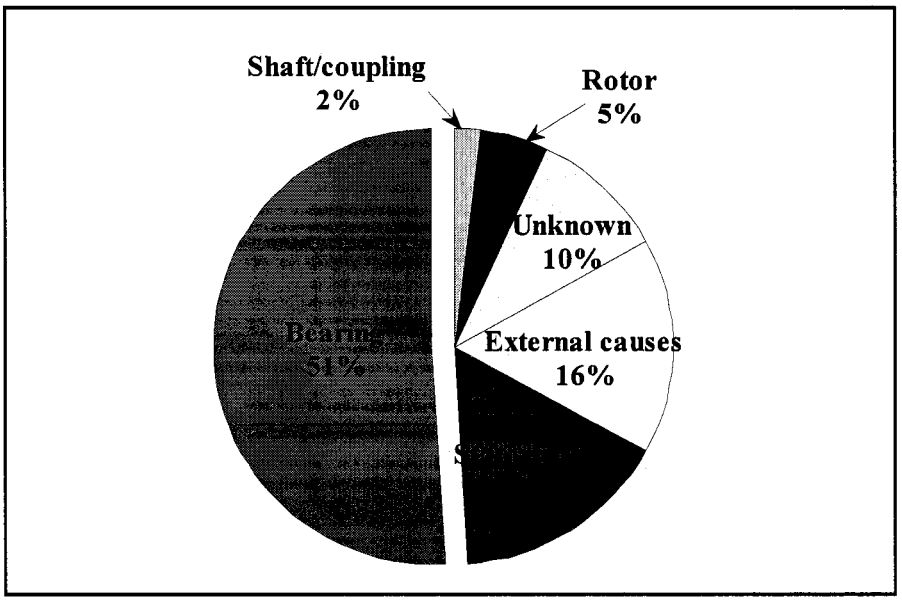


Figure 1-2. Causes of the electric motor failure.

1.3 Vibration Analysis Techniques

For bearing monitoring and diagnosis, vibration is considered the best operating parameter of physical properties such as temperature, pressure, oil analysis to determine bearing health [50-53]. Vibration is the result of internal dynamic forces generated by the rotating and sliding elements in the machinery. Once the machine is placed into service,

the degradation of the machine condition will change these forces and hence produce a unique pattern in its vibration signature.

How the vibration characteristics change as rolling element bearings approach failure has been the subject of many studies [6,74,87]. These studies have shown that short periodic impulsive components are produced by impacts between bearing components and small cracks at the pre-failure stage. The energy of these low-level, impulsive signals spreads across all frequencies on the spectrum and will excite the resonance frequencies of mechanical components and sensors. Harmonics of these defect related frequency and sidebands may appear when a defect is developing. As the damage spreads, the short periodic impulses become larger, making the fault more easily visible at lower frequencies. By the end of the bearing life, the overall vibration amplitude increases very rapidly because the deterioration of the bearing failure will cause other type of damage such as misalignment. It is therefore important to detect rolling element bearing defects at an early stage to prevent unexpected breakdowns.

Many attempts have been made to predict upcoming failure of rolling element bearings using vibration measurements and signal processing. Based on the representation and processing of a signal, a method can be classified as time-domain, frequency domain, or time-frequency domain.

Time domain methods generate a single value to represent machinery condition based on either the amplitude or the shape of the signal. During the early stage of a bearing defect, the bearing generated impulse is usually masked by the background noise and other dominating vibration components which can not be detected using an amplitude indicator. Shape indicators, such as crest factor [9] and kurtosis [8], have found to be more sensitive to impulsive signal. However, these indicators can not provide diagnostic information when machine vibration and failure mechanisms are complex.

Frequency domain methods are the most common method and provide effective tools for a broad range of machine fault monitoring and diagnostics [6-12]. Frequency domain methods assume that the components of vibration signal are periodic and sta-

tionary. When a defective bearing is rotating at a constant speed, the defect will generate a periodic pattern at a specific frequency which will stand out from the spectrum. However, machine diagnostic systems based solely on frequency spectra are often only be useful for detecting of bearing defects in their latter stages because the change in the vibration output due to the initiation of a small crack in a bearing component is often masked by the surrounding noise, making the extraction of salient features problematic [55,63]. These shortcomings led to the development of more advanced techniques for early bearing defect detection such as cepstrum analysis and high frequency resonance techniques (HFRT) [6,15].

Recently, impressive progress has been made in exploring the decomposition of vibration signal using time-frequency [13-22] and time-scale (also known as wavelet transform) [27-31,36,37] analysis. These analysis techniques allow us to investigate non-stationary signals in a two-dimension function. Time-frequency analysis, such as Wigner-Ville distribution and short time Fourier transform (STFT) are popular methods to map a one-dimensional signal to a two-dimensional function of time and frequency. However, each of time-frequency analysis methods have some disadvantages [122]. For example, STFT using fixed window size to perform Fourier transform suffers a trade-off of time-frequency resolution.

Wavelet transform (WT) is an extension of time-frequency analysis to transform signals into a sum of small waves instead of cosine waves in the STFT. Unlike STFT, WT analysis provides a high time resolution at high frequencies and low time resolution at low frequencies using various window sizes at different scales [23,24]. Figure 1-3 shows the time-frequency resolution differences of the time based, the frequency based, the STFT and the WT [26]. A recent review from Peng and Chu summarizes that WT is a useful tool for machine condition monitoring and diagnostics especially for non-stationary signal analysis, fault feature extraction, singularity detection, denoising, localized weak transient components and system identification [122]. For example, Xia uses normalized RMS for all wavelet packets at the last level of decomposition to form

a feature vector for pattern recognition of bearing faults [118]. Eren and Devaney investigate motor bearing damage detection by comparing the RMS of DWT at different scales and DWPT coefficients at different packets [119,120]. Zhang et al. use DWPT to de-noise vibration signals and pick up the fault frequency to enhance bearing defect detection under strong noise [121]. Shao and Nezu calculate the kurtosis of the de-noised wavelet coefficients as an indicator of bearing condition and use an envelope spectrum to diagnosis bearing defect [123]. A fast algorithm is also available for real time application [25].

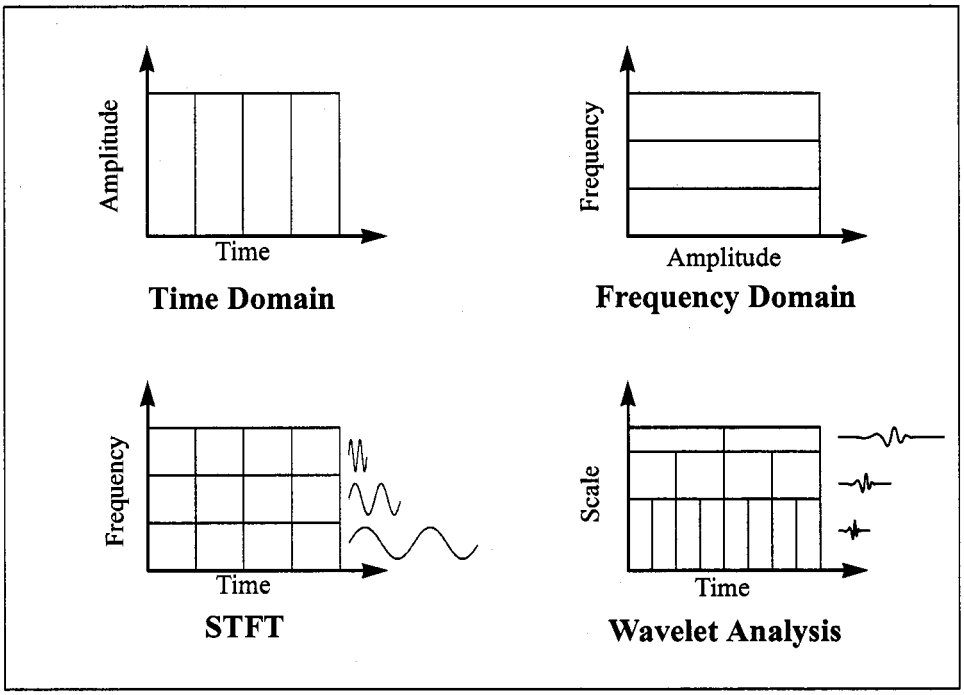


Figure 1-3. Time-frequency resolution of the time based, the frequency based, the short time Fourier transform, and the wavelet analysis.

1.4 Motivations and Objectives

The main purpose of this research is to investigate the fault detection of rolling element bearings using wavelet transforms and to compare the effectiveness of a wavelet based method with traditional techniques in various situations.

It is crucial to detect bearing defect as soon as possible. Early detection of bearing faults provides more time to take corrective action to prevent further damage and extend bearing life. Without an early fault detection technique, maintenance personnel must wait until later stages of failure. By this time, the damage is more extensive than if the fault had been detected earlier. The early detection of bearing failure requires methods that are sensitive to impulsive signals and their changes. The difficulty of early bearing defect detection is that the signal component generated by early defects is very small compared to other vibration sources in the overall system. To improve the early defect detection of rolling element bearings, various signal processing techniques have been developed [6-9,56,57,70,79,99,105,109]. It was found that techniques focusing at signal changes at higher frequencies are more sensitive to early bearing defects because the periodic impact generated by an early bearing defect will resonate at high frequency regions with less noise interference. The most popular technique is the HFRT which works by demodulating the envelope of the bandpass signal. There are two major limitations of using the HFRT [100]. First of all, the HFRT demands knowledge of the resonance frequency range where the impulses generated by the defects are more pronounced with respect to the normal system vibrations. For a complicated system, characterizing these resonant frequencies is a time consuming job and it is impractical to find the resonant frequencies using impact tests. The second limitation is when multiple defects develop at the same time, traditional HFRT may not be able to identify each defect due to the overlap of each transient component [57,100,104]. The first objective of this work is to resolve the limitations and improve the performance of HFRT for early bearing defect detection using wavelet transform. An automatic resonant frequency band selection method based on the wavelet packet transformation that searches the optimal resonant packet automatically will be described. A new HFRT algorithm based on wavelet packet transformation will be presented. Finally, The the performance of the wavelet based methods will be compared to the performance of the existing methods for simulated bearing defect data, experimentally measured data from

bearings with artificially created defects, and experimentally measured data from a series of run-to-fail bearing tests.

Most of the current studies of bearing vibration analysis use bearings that have artificially produced defects in order to demonstrate their effectiveness. However, artificially produced defects cannot simulate the gradual changes that would occur in the bearing vibrations as components goes from normal to failure [67]. In this work, we study the performance of the newly proposed and traditional methods using synthetic signals and experimentally measured signals from both bearings with artificial defects and bearings running from normal till failure. The importance of the synthetic data sets studied in Chapter 6 lies in the fact that everything about the data is known, and thus it is easier to interpret the results. The importance of the recorded data sets studied in Chapter 7 is to show how the algorithm works for real problems in field applications.

1.5 Organization

This thesis begins with an introduction to rolling element bearing that describes their geometry, dynamics and failure mechanisms (see Chapter 2). The chapter then continues with a discussion of bearing vibration models and characteristics. An extensive background on traditional signal processing methods used to analyze and diagnose rolling element bearing defects is given in Chapter 3. The development of the wavelet transform analysis is presented in Chapter 4, which includes the continuous wavelet transform (CWT), discrete wavelet transform (DWT), and discrete wavelet packet analysis (DWPT). This chapter also presents a time invariant DWT (and DWPT) called maximal overlap DWT (and DWPT). Two useful techniques are introduced which include the best basis algorithm for feature extraction and the denoising method for signal enhancement.

Chapter 5 discusses the high frequency resonance technique (HFRT) which is the main tool of early bearing defect detection currently used in industry. An automation resonant frequency band selection based on DWPT is proposed to resolve the difficulty of

the HFRT process. Three alternatives based on the wavelet transform and energy detectors are proposed and derived. This chapter represents the main contribution of original material in this dissertation.

Chapter 6 and 7 use synthetic data and recorded data respectively to assess the proposed HFRT alternatives as a tool for resonant frequency band selection and early bearing defect detection. The simulated signals of single defect and multiple defects are generated and analyzed in Chapter 6. Comparisons are made with other popular methods of bearing defect detection such as the Hilbert transform, direct spectrum and time domain indicators. Chapter 7 extends the analysis to real bearing data in order to evaluate the effectiveness of the proposed techniques in real situations. In Chapter 8, we introduce the idea of harmonic modeling to extract impulses from noisy signals. This pre-processing technique can improve the sensitivity of the proposed HFRT alternatives for detecting bearing defects. Chapter 9 summarizes the findings and proposes future work based on this research.

Chapter 2 Rolling Element Bearings

Rolling element bearings are the most common component in rotating machinery that utilizes the rolling action of rollers to minimizes friction between stationary and moving parts. They are widely used in applications from house appliances, such as electrical tooth brush, to gyroscopes in aerospace. It is estimated that only 10 to 20% of bearings achieve their design life due to adversary operation conditions [87] and that bearing failure accounts for more than 50% of motor failure [88]. The economic loss due to bearing failure is huge compared to the cost of the bearing itself [58-60]. Therefore, bearing condition monitoring and diagnosis has drawn many attention in the last four decades [42-47,50,54-58,77-84,87,94-105,116-124].

Before discussing bearing condition monitoring and diagnosis, it is important to have a general understanding of rolling element bearings. This chapter introduces rolling element bearings, beginning with basic bearing types, their major components and a summary of bearing mechanics. Finally, bearing failure mechanisms important to bearing diagnosis are presented.

2.1 Types of Rolling Element Bearings

There are many forms of rolling element bearings. Depending on the application, rolling element bearing can have various dimensions and design. For example, deep groove ball bearings perform well at high speed under moderate radial loads as well as axial loads. They have low friction and can be produced with high precision and in quiet running variants. Therefore they are preferred for small and medium-sized electric motors. One way to classify rolling element bearings is based on the shape of rolling elements as shown in Figure 2-1 [90]. In our study, we will focus on ball bearings because it is

the most popular rolling element bearing in all machines. Most of the studies based on ball bearings can be extended to other type of rolling element bearings.

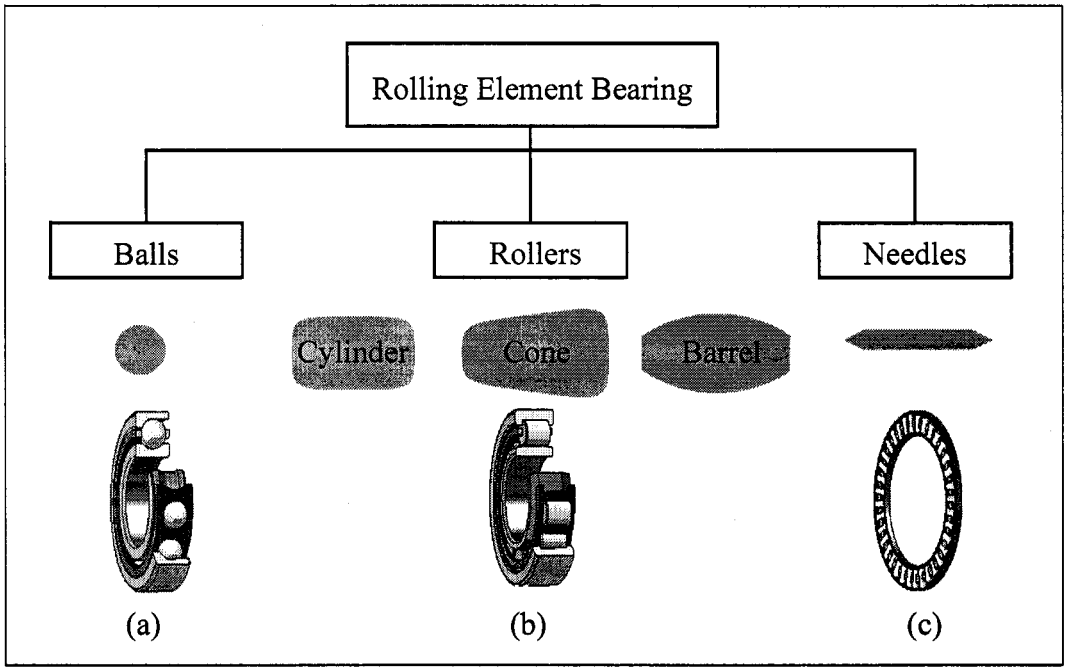


Figure 2-1. Classification of rolling element bearings [90].

2.2 Rolling Element Bearing Components and Geometry

Bearing geometry is a critical factor for diagnosing bearing defects because the geometry of ball bearings determines the dynamics of the bearing components and their vibration characteristics. This section will describe the components and geometry of bearings.

2.2.1 Bearing components

All type of rolling element bearings consist of rolling elements, an inner ring, an outer ring and cage(s). Figure 2-2 is a simple schematic diagram of a ball bearing. The rolling motion of balls decreases friction between components with small contact area. The

raceway in a ball bearing is the circular groove formed in the outside surface of the inner race and in the inside surface of the outer race. These grooves form a circular track to contain the ball set. The cage is used for separating and positioning the balls at equal intervals around bearing's raceway. The most widely used bearing material is ANSI-440C stainless steel with heat treatment to offer good corrosion resistance and hardness. Some of the ball bearings will include shields and/or seals to retain lubricants and prevent contaminants from reaching central work surfaces. A lubricant prevents metal-to-metal contact, carries away local heat generated in the bearing and protects the highly finished working surfaces from corrosion [61].

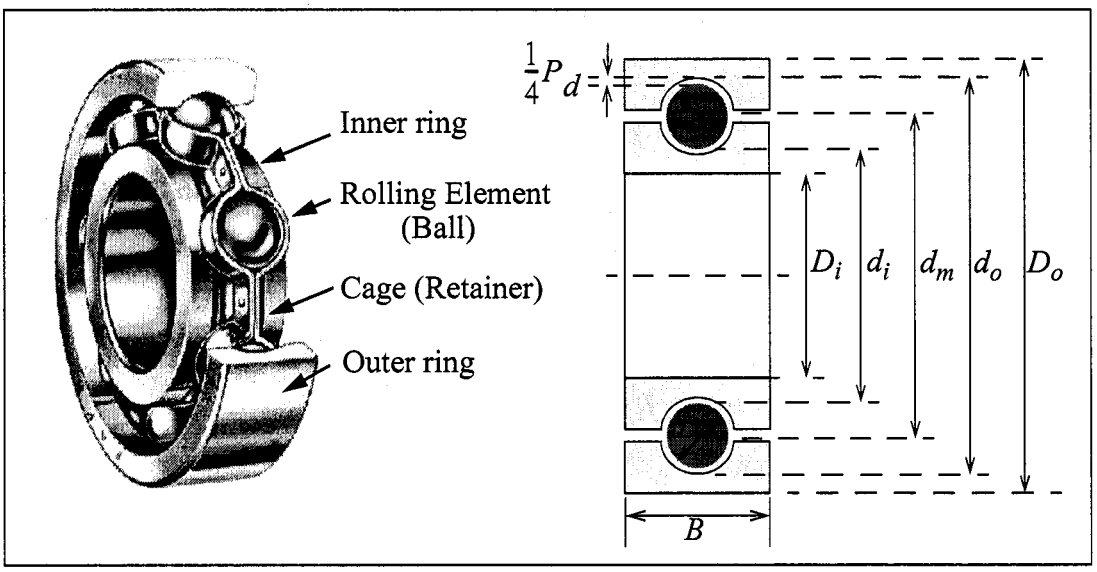


Figure 2-2. Main components and geometry of a ball bearing.

2.2.2 Ball bearing dimensions

The internal geometries of bearing are quite complex and will determine the bearing mechanics under operation. Most bearing manufacturers endorse the standards created by the Antifriction Bearing Manufacturers Association (AFBMA). Following these standards, manufacturers provide dimensions and load ratings for bearing selection.

Table 2-1 is an example of bearing specifications for radial shielded ball bearings. These basic bearing dimensions include the inside diameter (D_i), outside diameter (D_o), ring width (B) and ball diameter (D) as shown in Figure 2-2. Once the required load rating is determined, it is easy to select the best suited bearing size from the table.

Table 2-1. Dimensions and basic load ratings for double shielded ball bearings[a]

Inside Diameter (D_i) in	Outside Diameter (D_o) in	Width (B) in	Ball Number (n)	Ball Diameter (D) in	Dynamic Load Rating lbs
0.4690	0.1562	0.0937	6	1/32	35
0.0781	0.2500	0.1406	6	3/64	75
0.0937	0.1875	0.0937	7	1/32	42
0.1250	0.2500	0.1094	7	1mm	64
0.1562	0.3125	0.1250	7	3/64	88
0.1875	0.3750	0.1250	8	1/16	161
0.2500	0.5000	0.1875	10	1/16	187
0.3125	0.5000	0.1562	11	1/16	198
0.3750	0.8750	0.2812	7	5/32	747
0.5000	0.7500	0.1562	16	1/16	238

a. 400 series miniature bearing from National Precision Bearing, Inc.

In addition to the above dimensions, diametral clearance (P_d), pitch diameter (d_m) and contact angle (α^o) are also important geometrical quantities used to derive bearing mechanics [61]. Their definitions are as following:

- Diametral clearance (P_d)

Most radial ball bearings are designed to have a slight amount of looseness between balls and raceways. This looseness is referred to as diametral clearance P_d and is defined as follows:

$$P_d = d_o - d_i - 2D. \tag{2-1}$$

- Pitch diameter (d_m)

Pitch diameter is approximately equal to the mean of the bore and outside diameter, i.e.,

$$d_m \approx \frac{1}{2}(D_i + D_o). \tag{2-2}$$

- Contact angle (α^o)

When the bearing is under load, removing the diametral clearance in the axial direction causes the ball and raceway to contact at an oblique angle as shown in Figure 2-3. This angle is called contact angle (α^o), which is the angle between a plane perpendicular to the ball bearing axis and the line passing through the points of contact of the ball and both raceways. The magnitude of the free contact angle can be described as follows:

$$\cos\alpha^o = 1 - \frac{P_d}{2A}, \tag{2-3}$$

where A is the distance between the centers of curvature of the inner and outer race ring grooves.

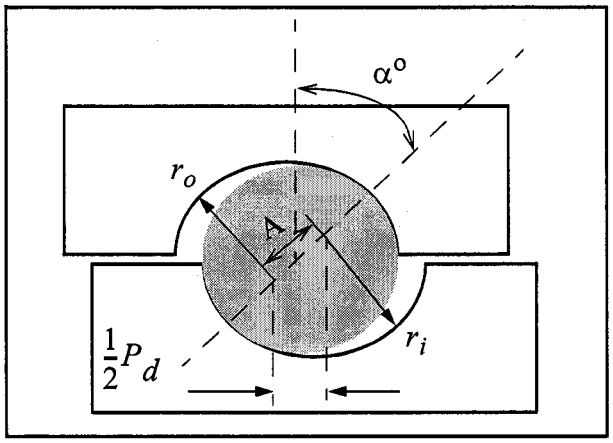


Figure 2-3. Radial ball bearing showing ball-raceway contact due to axial shift of inner and outer rings.

2.3 Mechanics of Rolling Element Bearings

Rolling element bearings are designed to support both radial and thrust loads simultaneously. These loads may be either "static" or "dynamic". The loads carried by bearings are transmitted through rolling elements from one ring to the other and the magnitude of the loading carried by each rolling element depends on the internal geometry of the bearing and the direction of the applied loads. Manufacturers normally provide the static and dynamic load ratings information which are important parameters to estimate the bearing life defined by AFBMA. Additionally, bearing loading condition and their mechanics are closely related to the bearing vibration that is the main focus of this research.

2.3.1 Ball bearing rotation

Consider a ball bearing mounted on a shaft which rotates at n_i RPM as shown in Figure 2-4. The rotational speed of the outer race moves at the speed of n_o RPM. The rolling elements (balls) rotate about their own axes at speeds of n_r RPM [62].

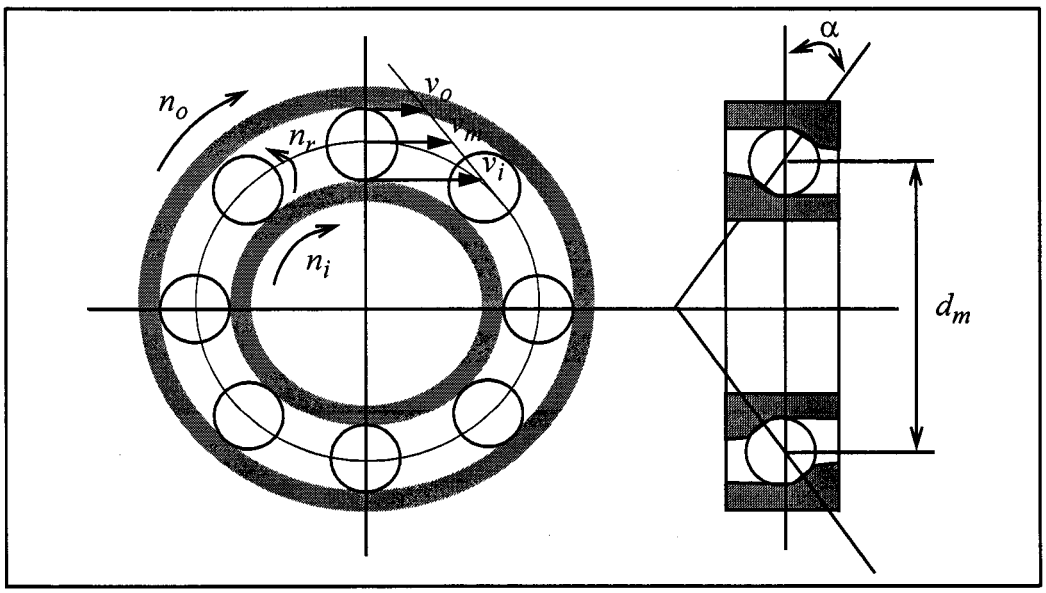


Figure 2-4. Rolling speed and velocity

Let ω be defined as the angular velocity in units of radians per second. The relationship between ω and rotating speed n is

$$\omega = \frac{2\pi n}{60}. \tag{2-4}$$

We know that the linear velocity (v) can be represented as a product of angular velocity (ω) and its rotating radius (r), i.e.,

$$v = \omega r = \frac{2\pi n}{60} r. \tag{2-5}$$

Therefore, the velocity of the inner race (v_i) and outer race (v_o) will be

$$v_i = \frac{1}{2}\omega_i (d_m - D\cos\alpha) = \frac{\pi n_i d_m}{60}\left(1 - \frac{D}{d_m}\cos\alpha\right) \tag{2-6}$$

$$v_o = \frac{1}{2}\omega_o (d_m + D\cos\alpha) = \frac{\pi n_o d_m}{60}\left(1 + \frac{D}{d_m}\cos\alpha\right). \tag{2-7}$$

Assuming there is no slip at the raceway contact, then the velocity of the cage (v_m) is the mean of the inner and outer raceway velocities. Hence,

$$v_m = \frac{1}{2}(v_i + v_o) = \frac{\pi d_m}{120}\left[n_i\left(1 - \frac{D}{d_m}\cos\alpha\right) + n_o\left(1 + \frac{D}{d_m}\cos\alpha\right)\right]. \tag{2-8}$$

2.3.2 Static loading

Both the weight of the machine parts and the shaft to support them are the sources of static loading to the bearing. Consider a ball bearing subjected to a radial load along the line of contact between ball and the raceway as shown in Figure 2-5. The resultant normal load (Q) can be decomposed to radial (Q_r) and thrust (Q_a) components. Their relationships are:

$$Q_r = Q\cos\alpha \tag{2-9}$$

$$Q_a = Q\sin\alpha. \tag{2-10}$$

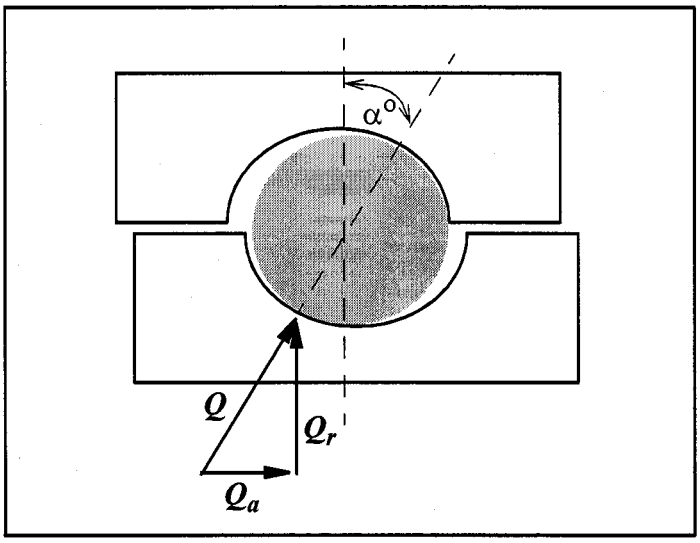


Figure 2-5. Radially loaded ball.

2.3.3 Dynamic loading

The main sources of dynamic loadings for bearings are the centrifugal force and gyroscopic moment. The centrifugal force comes from the rotation of rolling element or external forces acting on the bearing. For rolling element rotation, both rotation about the bearing axis (F_c) and the rotation about an eccentric axis (F_{ce}) can create centrifugal forces, which may be expressed as,

$$F_c = \frac{1}{2} m d_m \omega_m^2 \tag{2-11}$$

$$F_{ce} = mr\omega_e^2. \tag{2-12}$$

where m is the mass of the ball, ω_m is the angular velocity of ball, r is the radius of rotation about an eccentric axis and ω_e is the angular velocity of ball about an eccentric axis. Unbalance is a common cause of external dynamic loading. For example, a rotor running at 6000 rpm with 1 oz. of unbalance on a 9 in. radius can create 575 lbs centrifugal force using Equation 2-12.

Gyroscopic moment is due to the rolling element rotating about nonparallel axes simultaneously. For simplicity, assume the condition of steady-state precession and ignore the pivotal motion due to gyroscopic moment, and the gyroscopic moment for ball bearings becomes

$$M_g = J\omega_R\omega_m \sin\beta, \tag{2-13}$$

where J is the mass moment of inertia of the ball, ω_R is the angular velocity of ball about its own axis and β is the angle of rotation of the ball about its own center.

2.3.4 Bearing load distribution

Bearing load is not distributed uniformly to each rolling elements. The load distribution of a bearing under radial load is a function of angular location (θ) defined approximately by Stribeck as [109]

$$Q(\theta) = Q_{max}\left[1 - \frac{1}{2\varepsilon}(1 - \cos\theta)\right]^n \tag{2-14}$$

where Q_{max} is the maximum load, ε is the load distribution factor and n is the load-deflection exponent; n=1.5 for ball bearings and n=1.11 for roller bearings. Figure 2-6 is an example in which the load distribution for clearance is less than $\pi/2$. Note that for a part of circumference, the load applied is zero with a maximum angular extend of the load zone θ_{max}.

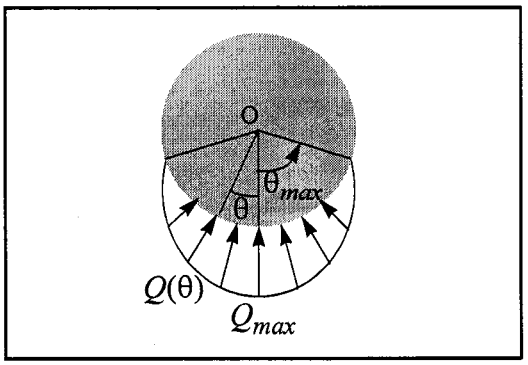


Figure 2-6. The load distribution in a bearing under a radial load.

This load distribution equation implies that the stresses acting on the contact surface of the bearing is also sinusoidally distributed and that stress level is the main reason causing the bearing to fail.

2.4 Rolling Element Bearing Monitoring Method

A bearing will fail over time even under normal operating conditions. Success in detecting imminent bearing failure is important and can be achieved by condition monitoring maintenance. The idea of condition monitoring is to select a measurable parameter on the machines which will change as the health condition deteriorates. Three common methods for monitoring rolling element bearing failures are wear debris analysis, temperature analysis and vibration signal analysis [91].

2.4.1 Wear debris analysis

Wear debris analysis is useful for oil based lubricant bearings by analyzing the size and amount of wear debris in the oil lubricant [84]. Debris can be collected either by magnetic precipitation or filtration. Bearings near failure usually generate larger size particles. However, wear debris analysis usually requires more effort and expensive equipment than temperature analysis or vibration signal analysis [67].

2.4.2 Temperature analysis

Temperature is the easiest measurable parameter for bearing condition using thermocouples or temperature dependent transducers. A rise in bearing cap temperature signals a drastic change in bearing condition. This method is better for the later stages of bearing failure because temperature is not sensitive to early stage bearing failure. It is also easily effected by environmental temperature, loads and running speed.

2.4.3 Vibration signal analysis

Vibration is considered the best operating parameter to determine the health condition of machinery [50-53]. Vibration is the result of internal dynamic forces generated by the rotating and sliding elements in the machinery. Vibration transducers can be easily mounted on the machinery to convert mechanical motion into a dynamic signal. Once the machine is placed into service, the degrading of the machine condition will change these forces and hence produce a unique pattern in its vibration signature. Various techniques have been implemented in current maintenance industry. These methods can be generally classified as time domain and frequency domain methods which will be covered in Chapter 3.

2.5 Bearing Failure

Bearing failure occurs when the bearing no longer fulfills its function of supporting the rotating shaft, or when the degree of damage to the bearing component becomes unacceptable [67]. Bently Nevada Co. found that a rolling element bearing progresses through three failure stages: pre-failure stage, failure stage and near catastrophic stage [74]. As the failure develops, the rate of bearing deterioration tends to grow exponentially as shown in Figure 2-7. Therefore, it is important to detect the onset of bearing flaws as soon as possible.

The major causes of bearing failure include excessive loading, improper lubrication, contamination, and improper installation [74]. Bearing damage can be classified in numerous ways including source, appearances and failure mechanism [83]. Since we are more interested in the damage process, we will briefly introduce four common bearing failure modes during service. They are wear, plastic flow, fracture and fatigue.

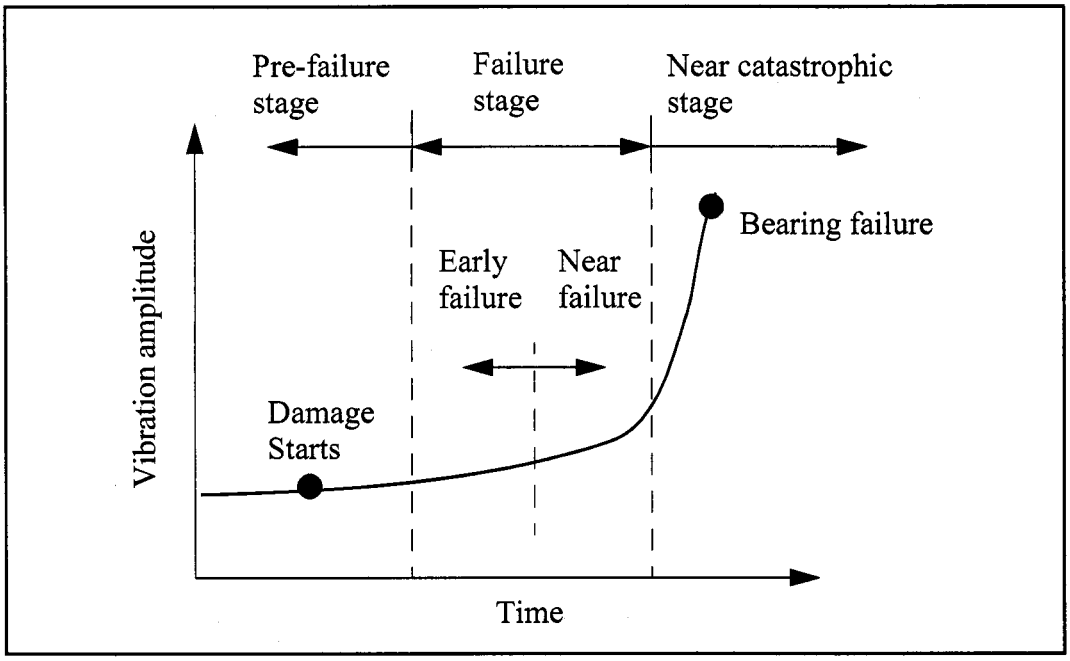


Figure 2-7. Model of bearing failure stage and failure rate

2.5.1 Wear

Mechanical wear occurs when components contact with each other under loads. Material is being removed from the surfaces in small particles due to direct solid-to-solid interaction. A typical evidence of mechanical wear is skid marking. As wear progresses, larger particles often cause more severe mechanical wear named smearing (It is also known as galling, scuffing or scoring). It may cause localized welding and material transfer between the two surfaces in contact. A severe smearing can lead to high vibration levels, seizure or bearing component fracture which is often the cause of catastrophic bearing failure [65].

2.5.2 Plastic flow

Plastic flow arises when due to loss of cooling, increase in load or speed or loss of lubrication, more heat is generated within the bearing than removed. This causes an unstable

temperature rise in the bearing. This sequence of events can cause hot plastic deformation and seizure of the material of the bearing.

2.5.3 Fracture

When bearing has extensive spalling, impact loads rolling over the spall can cause the bulk cracking. In cases of extensive cracking, an entire cross section of a bearing ring or rolling element will separate as a result of fracture.

2.5.4 Fatigue

In practice, 90 percent of bearing failures result from fatigue [1]. A fatigue process is characterized by the sudden removal of material due to alternating stresses. The failure mechanism of fatigue has two stages [75]. The first stage of fatigue is the crack initialization. Fatigue failure always begins at a small dent, or crack on the surface of the raceway. The crack may have been present in the material since its manufacturing, or the crack may have developed over time due to cyclic straining around stress concentrations. Once the cracks form, the second stage of fatigue failure, or crack propagation stage, begins to occur. When the balls roll over these cracks, the loads from these balls will create a stress concentration at the tip of the crack. As a result, the crack grows along planes normal to the maximum tensile stress and spalls follow soon after the propagation [66]. This will make the bearing fail due to fatigue.

Chapter 3 Rolling Element Bearings Vibration and Analysis

The preceding chapter shown the macrogeometry of rolling element bearings and loading under varied conditions of bearing operation. Vibration is one of the most obvious consequences of bearing operation and the best operating parameter to determine the health condition of machinery among various physical properties (such as temperature, pressure, oil analysis) [50-53]. Once the machine is placed into service, the degradation of the machine condition will change these forces and hence produce a unique pattern in its vibration signature. A bearing defect usually starts from a localized dent on one of the bearing components. When a metal surface rolls over this dent, a periodical impact will be generated at a specific frequency determined by the location of the dent, the speed of the shaft and the geometry of bearing. Many vibration analysis techniques have been implemented in current industrial applications. Traditional methods can be classified into time based and frequency based methods [6-12]. Recently, many advanced signal processing techniques have been proposed for bearing defect detection by representing signals in the time-frequency [93,94]or time-scale domains [118-124].

This chapter starts with an introduction of machinery vibration which is related to the basics of condition monitoring and diagnostics. Then, we will focus on bearing vibration characteristics including the bearing defect model, bearing defect frequencies and bearing vibration patterns at each failure stage. As last, we will review the traditional and advanced techniques for bearing vibration analysis methods. Advantages and limitations of these methods are shown with examples.

3.1 Machinery Vibration

Machine operation involves the generation of forces and motions that produce vibrations. Vibration is defined as an oscillation about an equilibrium position. For example,

an unbalanced machine rotor produces a rotating force that appears sinusoidal with respect to a reference position (see Figure 3-1). Since the response amplitude is proportional to the magnitude of the force, the response of the rotor to the force is also a sinusoidal waveform. We can model this harmonic motion by the equation:

$$x(t) = X_0 \sin(2\pi f t) \tag{3-1}$$

where $x(t)$ is the instaneous position of the vibration, X_0 is the maximum amplitude and f is the vibration frequency.

Each component in the machinery exhibits its own characteristic vibration signature. Some characteristic vibration signatures are fixed by the design and construction of the machine, but others may develop during service. A perfect machine usually generates none or light vibration during operation. When machines deteriorate, the vibration level and pattern changes due to wear. This makes vibration the best indicator of overall machine condition and the best diagnostic tool detecting the development of an internal defect.

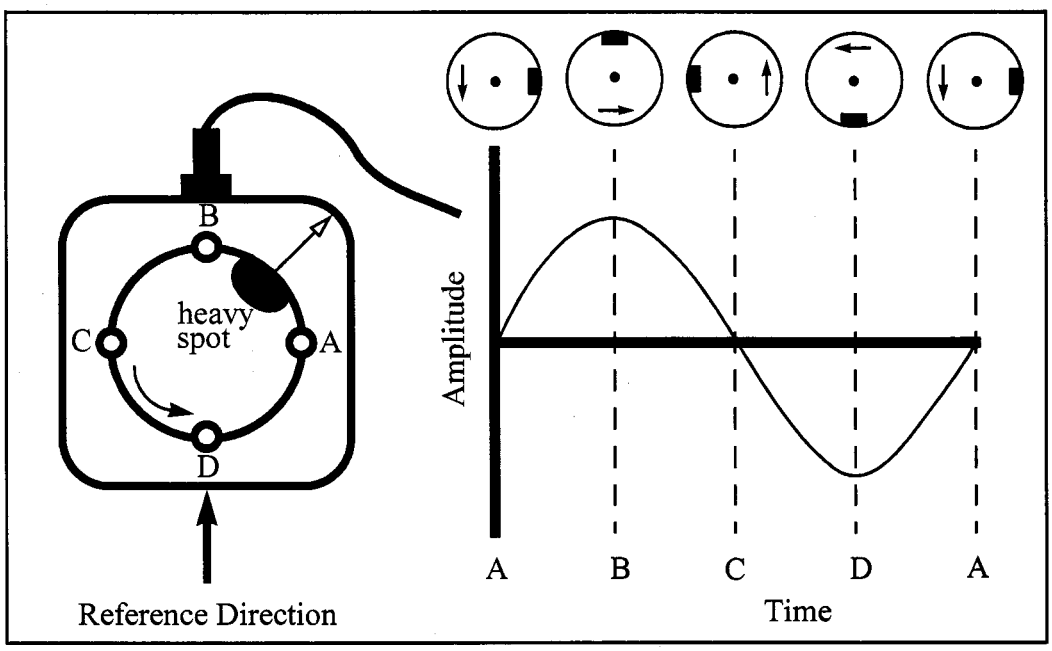


Figure 3-1. An unbalanced rotor generates a sinusoidal force and response waveform.

The most common method for characterizing and representing vibration is the frequency spectrum. A typical frequency spectrum for rotating machinery can be divided into five bands as shown in Figure 3-2[67]. This helps us to identify the sources of vibration. Most rotor related components contribute vibration in a low frequency range. Bearing defect related frequencies are also located in this frequency range. However, the main sources for high frequency vibration are resonances. Table 3-1 is a summary of the vibration characteristics of some common machinery faults [85]. In this table, frequencies at running speed, their harmonics, non synchronous harmonics and high frequency resonances are the most important features for diagnosis.

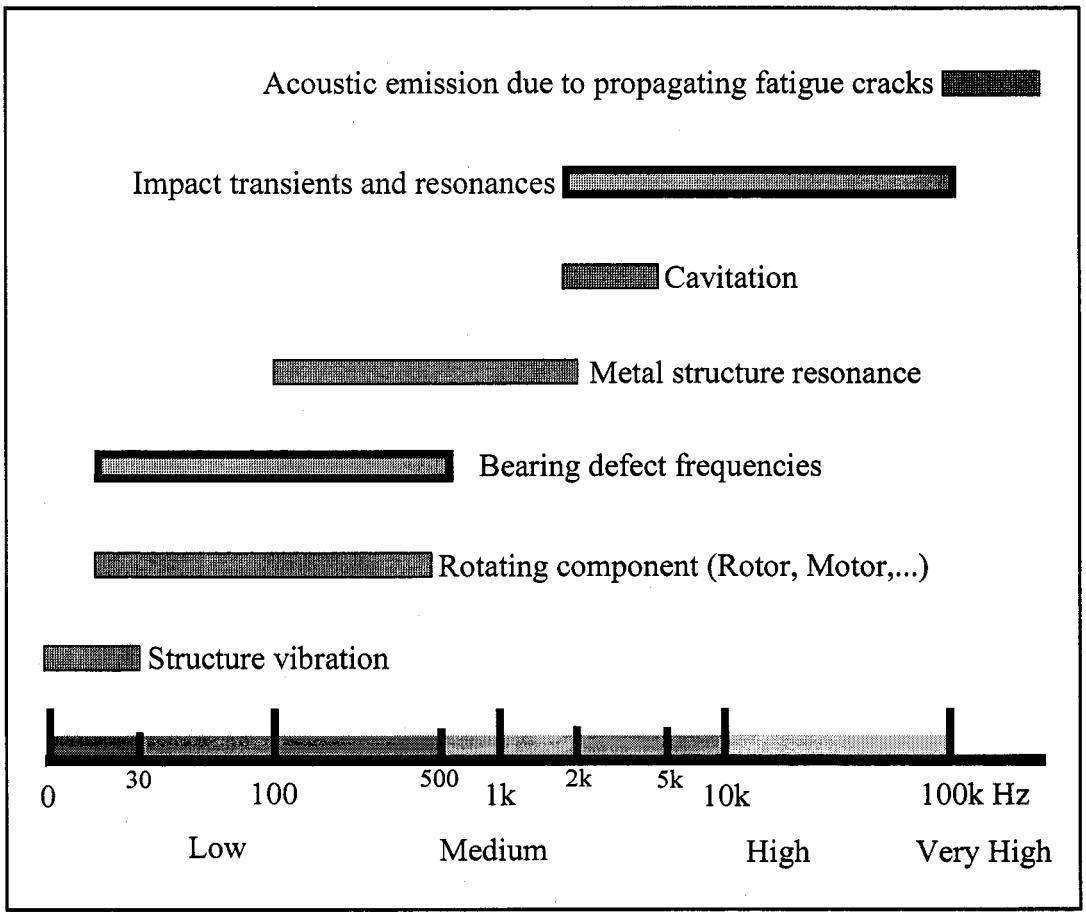


Figure 3-2. Typical vibration sources of rotating machinery in a frequency spectrum.

Table 3-1. Vibration characteristics of common machine faults [85]

Frequency	Possible Cause	Explanation
1 x rpm*	Imbalance	Can be caused by load variation, material buildup, or pump cavitation.
	Misalignment or Bent Shaft	A phase shift of 180 degrees in the axial direction. will exist across the coupling.
	Looseness	Directional-changes with transducer location. Usually high harmonic content and random phase.
	Resonance	Drops off sharply with changes in speed. From attached structures or changes in attitude angle or eccentricity ratio.
2 x rpm	Misalignment or Bent Shaft	High level of axial vibration.
Harmonics	Looseness	Impulsive or truncated time waveform; large number of harmonics.
	Rubs	Shaft contacting machine housing.
Sub-rpm	Bearing cage	See section 3.3.
N x rpm	Rolling element bearings	Non-synchronous. See section 3.3.
	Gears	Gearmesh (teeth x rpm); usually modulated by running speed of bad gear.
	Belts	Belt x running speed and 2 x running speed.
	Blades/Vanes	Blades/vanes x rpm; usually present in normal machine. Harmonics usually indicate that a problem exists.
Resonance	Several sources, including shaft, casing, foundation and attached structures.	

*. rpm is the running speed of shaft.

3.2 Vibration Model of Bearing Defect

Rolling element bearings consist of an inner ring, an outer ring and rollers held by a cage. The bearing failure often starts with a sizable local defect on any of these components. Figure 3-3 is an example of a ball bearing with a localized defect on an outer race. When balls roll over a point defect, it will generate repetitive impacts with a periodicity ($T=1/f$) that depends on the fault location. Many researchers model this periodic impulse train $x(t)$ as a series of impulse functions [60,78,92-93,101-102]

$$x(t) = d_0(t) \sum_{n=1}^{\infty} \delta(t - nT - \tau), \tag{3-2}$$

where τ is the initial delay. The amplitude of the impulse train $d_o(t)$ is determined by the location of the defect, the rotational speed and the severity of the defect. When the defect is on the outer race, $d_o(t)$ tends to be constant because the fault location does not change with respect to the radial load. However, for an inner race defect, $d_o(t)$ will change with respect to its angular position as described in section 2.3.4 [6].

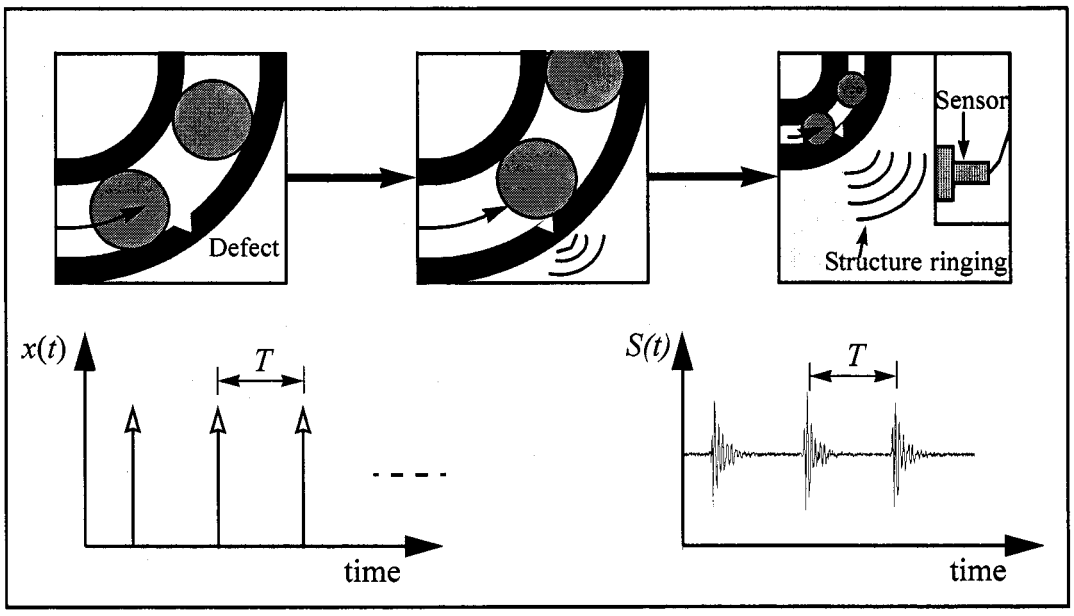


Figure 3-3. Ball bearing with a localized inner race defect.

The response can be measured by mounting an accelerometer on a supporting structure near the bearing. Each impulse excites a short transient vibration in the bearing and machine at their natural frequencies causing it to ring. If we model the system as a viscously damped-mass-spring system, the response to such impulse is an exponentially decreasing oscillation at the natural resonance frequency of f

$$A(t) = A_o e^{-\alpha t} \cos(2\pi f t + \varphi), \tag{3-3}$$

where A_0 is the maximum amplitude, α is the damping (or decay) factor and φ is the initial phase [59]. If we assume that all the pulses are linearly independent, then, the response of bearing with a single defect can be modeled as an impulse train with a repeating period of T:

$$S(t) = \sum_{k=0}^{\infty} A(t - kT)u(t - kT), \tag{3-4}$$

where $u(t)$ is a unit step function.

We can extend this single defect model to the multiple defects case by assuming the response from each defect is linearly independent. Then, the response of multiple defects $S_m(t)$ is a sum of $S(t)$ with different amplitudes which are dependent on the magnitude of the defects. Hence,

$$S_m(t) = \sum_{p=1}^{m} \left(\sum_{k=0}^{\infty} A_p(t - \tau_p - kT_p)u(t - \tau_p - kT_p) \right), \tag{3-5}$$

where m is number of defects and τ is the delay between defects.

This vibration model of bearing defects suggested that an optimal vibration analysis technique should be capable to detect transient oscillations of frequency f and check that these transients occur periodically with a period of T.

3.3 Fundamental Defect Frequencies

The repetition frequency ($1/T$) of this impulse train is determined by the location of the fault. These bearing related frequencies are known as fundamental defect frequencies. Knowing the velocity of each component, we can calculate the fundamental defect frequencies. The calculation of the fundamental defect frequencies of rolling element bearings is useful in machinery condition monitoring, because it tells what has failed in the bearing.

All rolling element bearings have four defect frequencies corresponding to each bearing component. They are called the ball-pass outer race frequency (f_{bpof}), the ball-pass inner race frequency (f_{bpif}), the cage or fundamental train frequency (f_{ftf}), and the ball spin frequency (f_{bsf}) [59].

For most applications, the outer race is fixed in a housing and the inner race rotates at the same speed as the shaft. Define Z as the number of balls or rollers and n_i as the rotational speed of the inner ring unit in revolutions per minute. Assuming no skidding of rolling elements with the stationary outer ring and rotating inner ring, the following bearing frequencies may be calculated:

The cage or fundamental train frequency f_{ftf} is

$$f_{ftf} = \frac{n_i}{120}\left(1 - \frac{D}{d_m}\cos\alpha\right) \qquad Hz. \tag{3-6}$$

The ball pass outer race frequency (f_{bpof}) and inner race frequency (f_{bpif}) are related to the rotational frequency of the inner ring relative to the cage (f_{ci}) which can be easily derived as:

$$f_{ci} = \frac{n_i}{120}\left(1 + \frac{D}{d_m}\cos\alpha\right) \qquad Hz. \tag{3-7}$$

The ball pass outer race frequency (f_{bpof}) and inner race frequency (f_{bpif}) are

$$f_{bpof} = Zf_{ftf} = \frac{Zn_i}{120}\left(1 - \frac{D}{d_m}\cos\alpha\right) \qquad Hz, \tag{3-8}$$

$$f_{bpif} = Zf_{ci} = Z(n_i - f_{ftf}) = \frac{Zn_i}{120}\left(1 + \frac{D}{d_m}\cos\alpha\right) \qquad Hz. \tag{3-9}$$

The ball spin frequency (f_{bsf}) is

$$f_{bsf} = \frac{n_i}{120}\frac{d_m}{D}\left(1 - \left(\frac{D}{d_m}\right)^2\cos^2\alpha\right) \qquad Hz. \tag{3-10}$$

Note that each of the bearing defect frequencies are non-integer multiples of running speed and the cage train frequency is always lower than shaft rotating frequency. When a ball passes over a bearing defect in the races or cages, pulse-like excitation forces are generated that result in one or a combination of these defect frequencies. In practice, these derived defect frequencies are approximate. As a matter of fact, the rolling motions are accompanied by a degree of sliding which occurs in the contact areas. In a ball bearing, substantial amounts of spinning motion occur simultaneously with rolling at high shaft speeds if the contact angles are large. At higher speeds, gyroscopic pivotal motions can occur [78].

3.4 Bearing Vibration Characteristics

Before discussing methods for vibration analysis, it is important to have an understanding of how the vibration signal changes as bearing failures develop [74]. As presented in Chapter 2, rolling element bearing failure progresses through pre-failure, failure and near catastrophic stages. The vibration characteristics of each stage are quite different and complicated. Figure 3-4 presents typical vibration spectra for each failure stage [87]. The frequency axis is divided into three regions: the rotor vibration region; the bearing defect frequency region; and the high frequency region. The rotor vibration region includes periodic signals at low frequencies such as multiples of the shaft speed. The bearing defect frequency region contains the repetition rate of bearing defect frequencies and their harmonics (typically up to 7 times of outer race defect frequency).

The high frequency region ranges approximately from 2k Hz to 120k Hz. Bearing components resonate at these natural frequencies when they are excited by the impacts of rolling elements and flaws on the raceways.

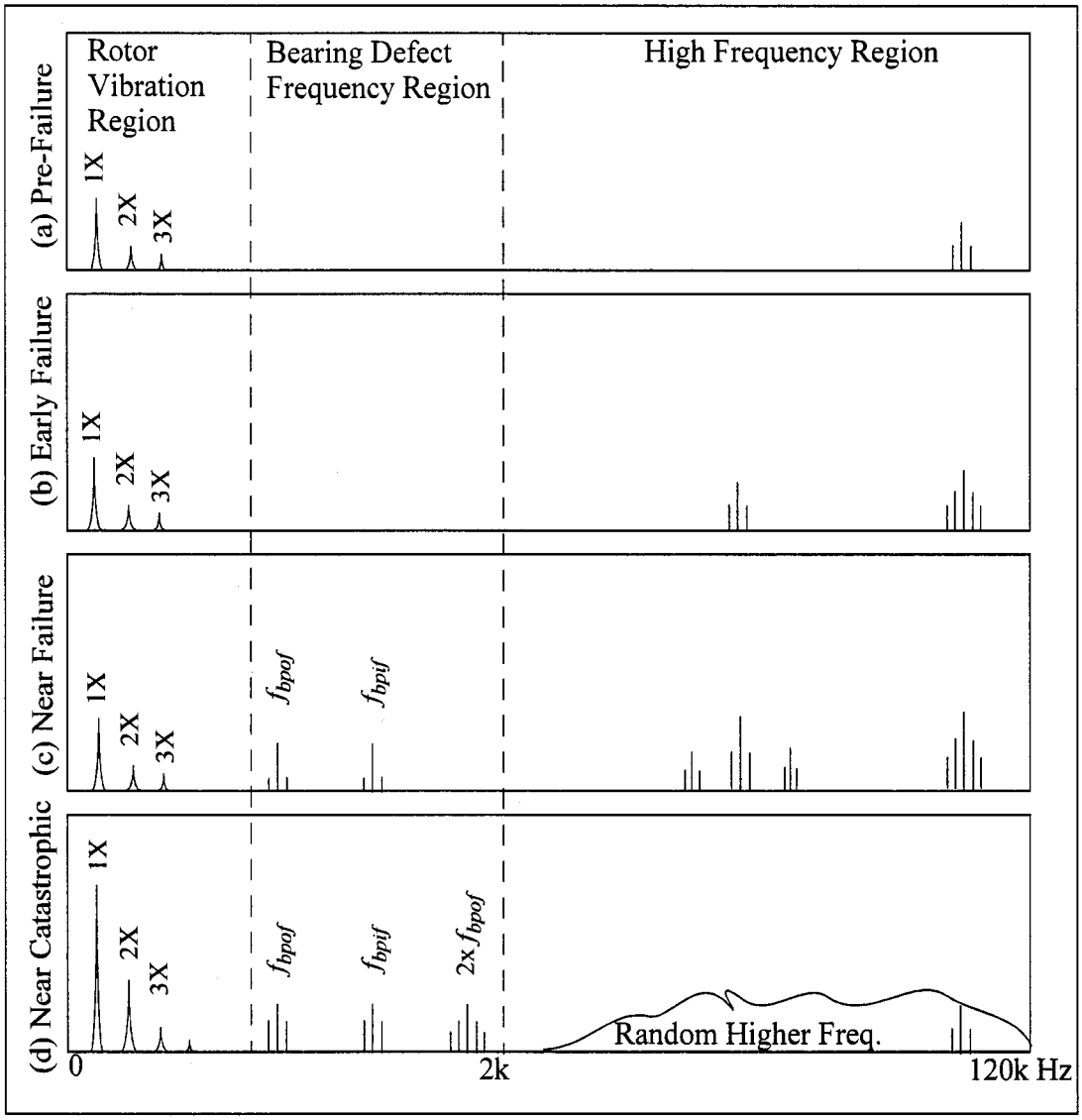


Figure 3-4. Spectra of rolling element bearing at each failure stage

3.4.1 Pre-failure stage

In the pre-failure stage, bearing defects are subsurface or microscopic. The bearing vibration spectrum from the accelerometer looks normal and is dominated by the running speed and its first few harmonics. It is possible to detect flaws at this stage using special sensors to extract impulse trains at very high frequency ranges. For example, acoustic emission (AE) monitoring uses a point contact AE transducer to monitor changes in the stress wave above 100 kHz [80]. Another technique, the shock pulse method, uses an accelerometer with a resonance around 38 kHz to detect shock pulses induced by a small crack [79]. These methods are not considered here because the bearing flaws that occur at this stage are typically not critical. Bearings with pre-failure defects are still capable of functioning at their designed capability.

3.4.2 Failure stage

The failure stage can be sub-divided into two stages: the early failure stage and the near failure stage. At the early failure stage, impulse like excitations cause the installed bearing components, the bearing housing, the machine structure and the transducer to ring periodically at their natural frequencies. Repetition rates of the pulses can be calculated from the bearing geometry (see section 3.3), but are undetectable using the direct spectrum method. This is because the repetition component generated from a small crack is easily masked by other signals in the low frequency range. However, these defect frequencies are usually amplitude modulated with resonances at a higher frequency range.

When the bearing defects progress to the near failure stage, the rate of bearing degrading starts to increase exponentially as shown in Figure 2-7. The size of the cracks become larger and the number of cracks increase during the near failure stage. Therefore, the change in vibration amplitude becomes detectable. Up to 7 harmonics of fundamental bearing frequencies start to appear in the response spectrum. As the degrading continues to the end of the near failure stage, strong components at the fundamental defect frequencies will be visible and the impulsive components in the high frequency

region may begin to move down to the lower frequency region because the area of damage is enlarged. Amplitude modulation of the four frequencies by the speed of the rotating unit causes sidebands. The ball pass frequencies can be modulated by the fundamental train or ball spin frequencies, which can in turn modulate natural frequencies. The frequencies generated by defective bearings are thus combinations of bearing frequencies, natural frequencies, and frequencies of the rotating unit. Bearing failure detection is important at this point and it is necessary to watch the trend of these changes closely. Most of the work in bearing vibration analysis focuses on detecting bearing flaws during this stage as early as possible.

3.4.3 Near catastrophic stage

If the bearing is not replaced at the failure stage, catastrophic bearing failure becomes imminent. At this moment, the bearing material is rapidly worn out which allows the shaft to move significantly. This degradation and the corresponding vibration amplitude begin to grow exponentially. The larger clearance can cause rub impact between rotating and stationary components which leads to accelerated machine failure. Defect frequencies may shift in frequency or disappear due to the sliding motion caused by the increase in clearance. Therefore, the overall vibration energy grows faster than the energy from the impulse signal during the end of bearing life.

Based on the above observations, the choice of bearing failure detection method should depend on how the bearing flaws develop. In the early stage of bearing failure, methods that concentrate on detecting impulses in the high frequency range tend to be more effective for detecting early bearing flaws. Diagnosis of a bearing defect or bearing condition therefore depends on the combination of frequencies present. Raw data is usually pre-processed with a high pass filter to remove low frequency vibration sources. As the flaws progress into the near failure stage, most of the techniques focus on tracking the fundamental defect frequencies. These methods often resulting in false alarms because the vibration signal becomes more random as damage spreads around the bearing. At

the end of bearing life, the overall vibration energy increases rapidly and may be detected very easily. A review of these bearing failure detecting methods and their advantages and disadvantages is given at the following sections.

3.5 Vibration Analysis Techniques

Many attempts have been made to predict upcoming failure of rolling element bearings using vibration measurements and signal processing. Traditional techniques can be broadly divided into two categories: time based and frequency based methods [6-12]. Recently, many techniques had been proposed to process signals into two dimensions (time and frequency/scale) instead of one (time or frequency) [13,86]. This new representation is most suitable for non-stationary signals such as impacts generated by bearing defects.

3.5.1 Time based method

Time based techniques usually produce a single parameter derived from the time history of the vibration signal. This derived parameter represents the current condition and may be compared with the parameter measured when the bearing is in a new and undamaged condition. Using a single number to estimate the bearing condition is desirable for practical applications since most facilities have a large number of rolling element bearings. All of the time based methods are computationally efficient and easy for monitoring. These methods of detecting bearing damage have been somewhat successful [7,8]. However, it is impossible to diagnose all possible faults using only one parameter. It was also found that some of these parameters are sensitive to the presence of noise and to changes in operating condition [77]. These parameters are also unable to identify the cause of the fault when the signal from the machine is complicated with many possible sources other than bearing defect generating impulse-like responses [9]. In this section, we review some of the popular time-based monitoring techniques.

3.5.1.1 Amplitude descriptors

The simplest way of monitoring bearing conditions is to measure the vibration amplitude within a certain frequency band. The range of frequency band should cover the expected vibration from sources of interest. Commonly used amplitude descriptors include the peak value and root mean square (RMS) value. The peak value represents the extreme value of vibration amplitude away from the neutral position. The RMS value measures the overall energy content of the signal. This value is also known as vibration severity in PdM. For a signal X, the definition of RMS of X is

$$X_{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} X_i^2}. \tag{3-11}$$

A comparison of peak-to-peak and RMS values made by Mitchell is shown in Figure 3-5 [6]. This figure illustrates the advantages and disadvantages of these two descriptors for a sinusoidal and nonsinusoidal signal. In Figure 3-5 (A), the peak value of the pure sine wave and the pulse train are the same, however, the RMS value of the pulse train is smaller than the pure sine wave which implies that the vibration energy of the impulse is very small. On the other hand, Figure 3-5 (B), shows that two signals with about the same RMS value may differ significantly in their peak values. This peak value could be misleading if the sudden change was due to some disturbances.

Generally, the RMS value tends to be better for detecting problems such as unbalance and misalignment where damage is mainly a function of energy, or in simple machines where bearings are the main signal source. International standards, such as ISO standards 2372 and 3945 (Table 3-2), provide severity charts for using RMS level to determine machine condition. However, they provide an inadequate assessment of bearing condition at the early stage of failure because the changes in these low frequency bands resulting from bearing damage are small. The peak value is sensitive to impulsive vibrations that are the early symptom of rolling element bearing defects. Using the peak value without examining the signal can be misleading because of signal contamination

from noise. It is important to recognize that the amplitude descriptors depend on not only the condition of the machine but also their load and speed. A trend plot is needed to track the condition of bearings using these values.

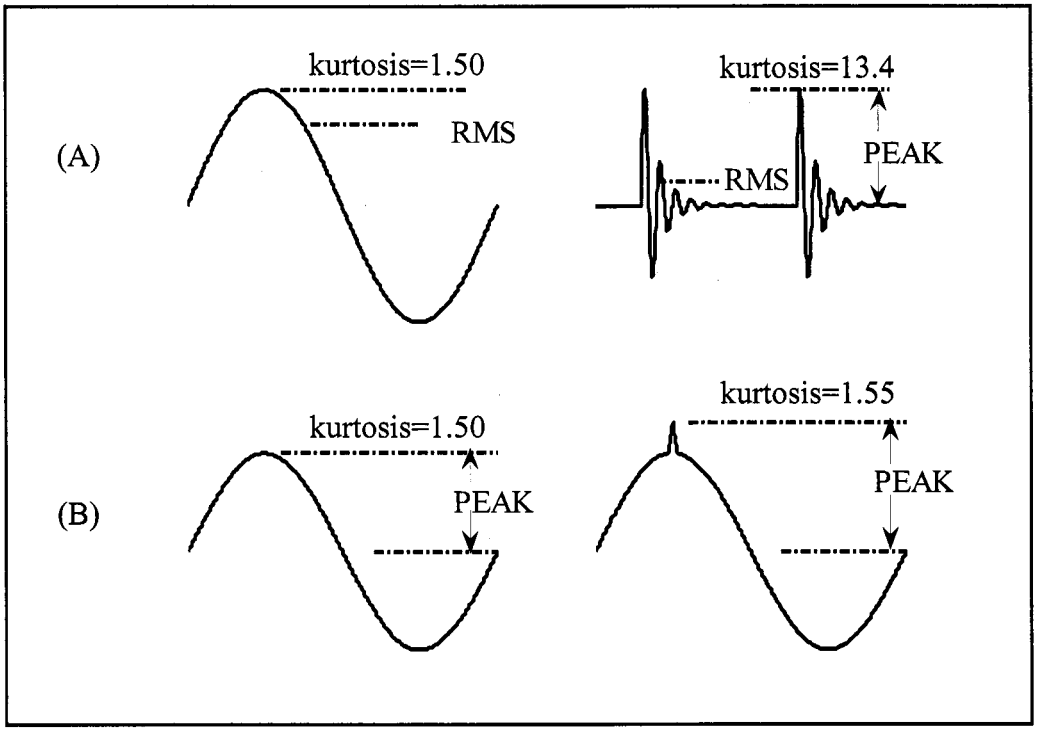


Figure 3-5. Comparison of amplitude descriptors for a sinusoidal and nonsinusoidal signal. (A) Same peak value with different RMS value. (B) Same RMS value with different peak value.

Table 3-2. International Standards ISO 2372 and 3945

Ranges of radial vibration severity			Quality judgement for separate classes of machine			
Range	RMS velocity In the range 10-1000Hz		Class I	Class II	Class III	Class IV
	MM/SEC	IN/SEC				
0.28	0.28	0.011	A	A	A	A
0.45	0.45	0.018				
0.71	0.71	0.018				
1.12	1.12	0.028	B			
1.8	1.8	0.044		B		
2.8	2.8	0.071	C		B	
4.5	4.5	0.11		C		B
7.1	7.1	0.18	D		C	
11.2	11.2	0.28		D		C
18	18	0.44			D	
28	28	0.71				D
45	45	1.1				
71	71	1.8				

Machine Classes

CLASS I Small Machines to 20 HP
CLASS II Medium Size Machines 20 to 100 HP
CLASS III Large Machines 10-200 rev/sec., 400 HP and Larger Mounted on Rigid Support
CLASS IV Large Machines 10-200 rev/sec., 400 HP and Larger Mounted on Flexible Supports

Acceptance Classes

A: Good B: Satisfactory C: Unsatisfactory D: Unacceptable

3.5.1.2 Shape descriptors

Shape descriptors detect the impulsive features of the time domain waveform induced by bearing flaws. They are more sensitive and reliable detectors of early stage bearing failure than amplitude descriptors [6]. The most popular shape descriptors, crest factor and kurtosis, will be presented in this section.

The crest factor is a non-dimensional factor derived from normalizing the peak amplitude of vibration by the RMS value [76].The formulations of crest factor (C) is:

$$C = \frac{\mathrm{MAX}(x(t) - \mu_x)}{\mathrm{RMS}}. \tag{3-12}$$

This ratio is suitable for early stage bearing failure detection because the amplitude of impulsive components increases more rapidly than the RMS level at the beginning of flaw development. For a machine running at a quiet and normal operating condition, the crest factor C, may have values as low as 1.4 to 1.5. Higher values of C often imply that the machine might have complex impact-type components. In a study by Howard, a typical crest factor value and rate of occurrence for a small ball bearing running at 1800 RPM is given in Table 3-3 [79]. In the table, the crest factor and rate of occurrence is related to the bearing condition. The rate of occurrence at a high crest factor value within one revolution of shaft rotation prevents false alarms due to disturbances. Tracking both values provides a more reliable prediction of bearing condition. Note that the crest factor may drop during the later stage bearing failures since the RMS level increases faster than the peak value when defects start to spread.

Table 3-3. Typical crest factor and rate of occurrence for small ball bearing with different bearing condition.

Bearing Condition	Crest Factor	Rate of Occurrence*
Single damage	7 or higher	Greater than 4 per second
Multiple damage	7 or higher	Greater than 10 per second
Acceptable	6 or lower	NA

*. Bearing operating speed is 1800 RPM.

A more sophisticated way to represent the bearing condition is to perform a statistical analysis of the vibration signal. The kurtosis function is a statistical function which emphasizes high amplitude deviations from the normal level. The definition of kurtosis (K) is:

$$K = \frac{\int_0^T [x(t) - \mu_x]^4 dt}{T\sigma^4}, \tag{3-13}$$

where μ_x is the signal mean, σ is signal standard deviation, and T is the duration of the signal.

A perfect random vibration has a kurtosis value of 3, whereas a vibration signal with kurtosis value of greater than 3 implies more impulsiveness. Figure 3-6 illustrates the time waveforms and the change in the probability density function before and after bearing damage occurs in one of our tests. The kurtosis for the normal bearing is very close to 3. As the bearing defect progressed after running for 10 hours, the kurtosis is 4 times larger than the normal condition. This is due to the impulse train as can be seen from the waveform of Figure 3-6 (b).

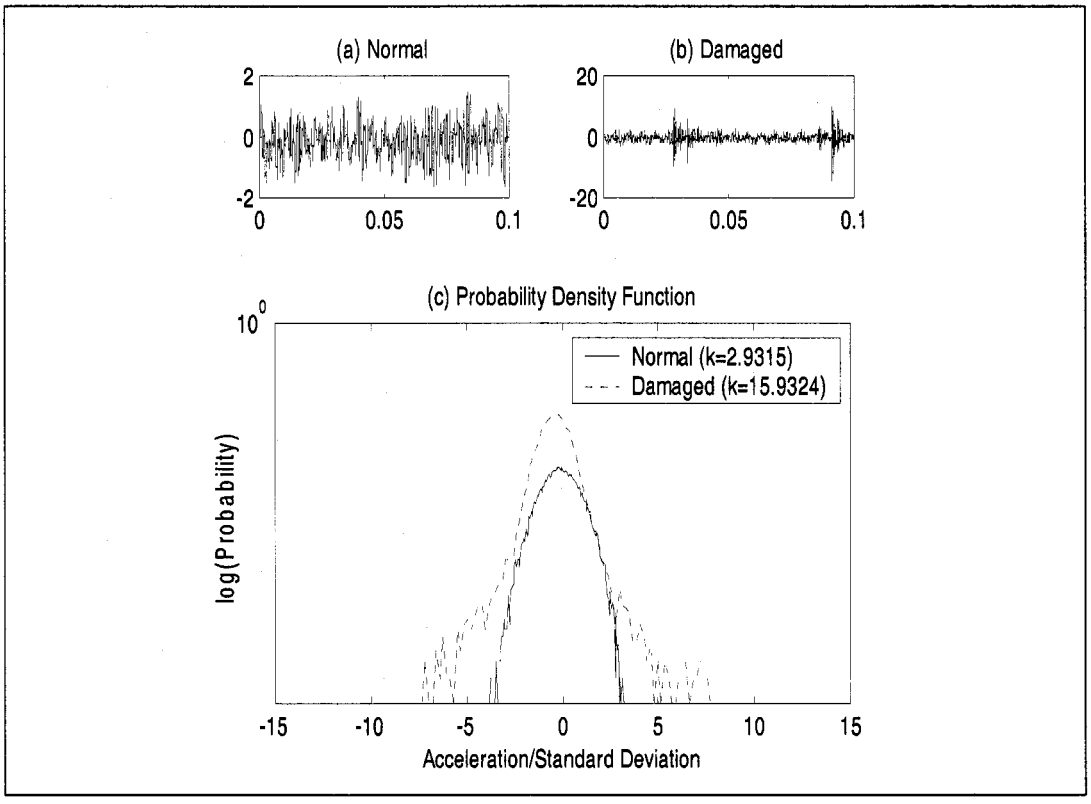


Figure 3-6. The time waveform and probability density function of the normalized acceleration of a ball bearing in normal and damaged condition.

Using the kurtosis value is therefore potentially useful in detecting early stages of failure when impacts are only intermittent. However, when a bearing is in a very damaged condition the signal is dominated by very frequent impulsive peaks, and K falls as impulses become more common. In addition, its high sensitivity to impulsive signals can be deceptive since it may produce false alarms due to noise.

Figure 3-7 shows an example of using time based descriptors for defective and normal bearings. A small dent was introduced on the inner race of the defective bearing. All the time waveforms are pre-processed with a high pass filter to remove low frequency components. The transient part of the signal was discarded because of the filtering. Figure 3-7(a) and (c) show the measurements from defective bearings, and Figure 3-7(b) and (d) show the measurements from normal bearings. For the cases without background noise, all of the time based descriptors for the defective bearing are much higher than those of the normal bearing because of the impulse trains generated by the small dent. However, because the signal is contaminated with background noise in Figure 3-7(c) & (d), RMS is not a suitable indicator for identifying the bearing defect. The other descriptors are more sensitive to the impulse components.

In conclusion, time domain indicators are easy to calculate and implement for automation monitoring. Amplitude based indicators are sensitive to operation conditions. Shape descriptors are independent of loading and speed, which makes them a better diagnosis method for bearings that have no baseline measurements from normal operation [88]. However, shape indicators can cause errors of interpretation in the presence of noise. The raw measurements are usually pre-processed with a high pass filter to remove the effect from low frequency components. Therefore, they are suitable for early stage of bearing failure detection and are easy to calculate. For later stages of failure, their value will drop and should not be used for bearing fault detection.

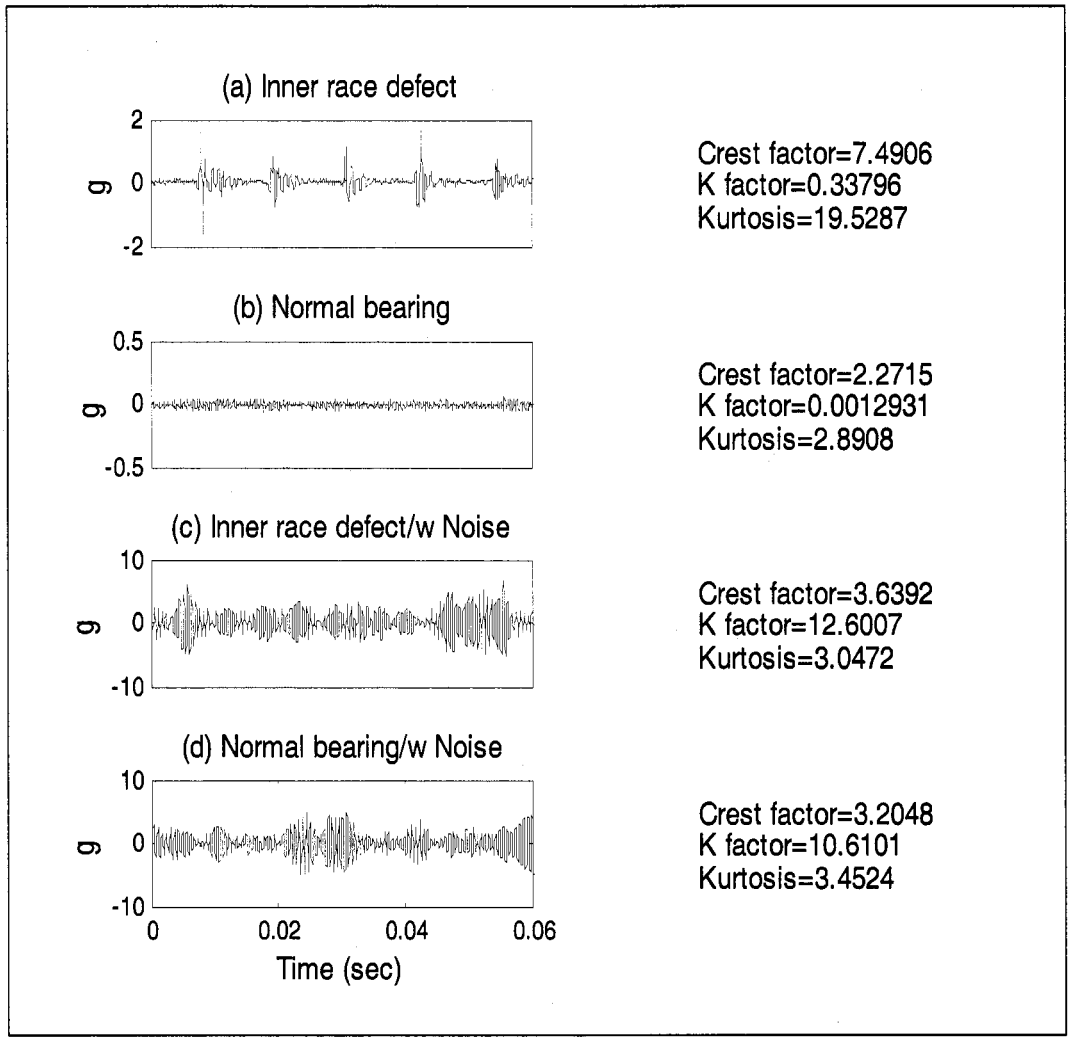


Figure 3-7. Time domain indicators for normal and defective bearing with and without background noise.(a) Defective bearing without background noise, (b) Normal bearing without background noise, (c) Defective bearing with background noise and (d) Normal bearing with background noise. Signal are prepossessed with a high pass filter with a normalized cut of frequency at 0.25.

3.5.2 Frequency based method

Frequency based methods are the most popular techniques for analyzing the vibration signal. Using methods in the frequency domain is computationally efficient and more effective than time based methods. Taking the fast Fourier transform (FFT) of the signal

gives us the spectrum in the frequency domain. The spectrum presents information indicating the level of vibration at a particular frequency. Analyzing the signature of this spectrum can help us to detect and identify most types of faults. The frequency identifies the source of the fault and the amplitude identifies the severity. Commonly used frequency based methods are classified into two categories: direct spectrum analysis and the processes spectrum method [66].

3.5.2.1 Direct spectrum analysis

Direct spectrum analysis uses the result of power spectrums directly to identify bearing failures. Depending on the resolution of the spectrum used, it can be sub-divided into the broadband method and the narrowband method.

- Broadband method

The broadband method sums up all the spectrum amplitudes within the linear range of the accelerometer (typically up to 5 kHz sampling frequency). This method is analogous to the RMS value in a time based method and generates only one number that represents the overall condition of the machinery. This number represents the overall energy of vibration. Similar to the time based method, the broadband method is not a good indicator for bearing condition because it is not sensitive to bearing faults when the background noise is high [15]. This can be shown by comparing the spectrum of a ball bearing with and without the background noise as in Figure 3-8. A small crack is induced at inner race. The broadband level was calculated up to 8 kHz. The difference in broadband level was around 60dB without adding background noise. As we added

background noise to the same artificially defected bearing, the difference in broadband level reduced to 6dB.

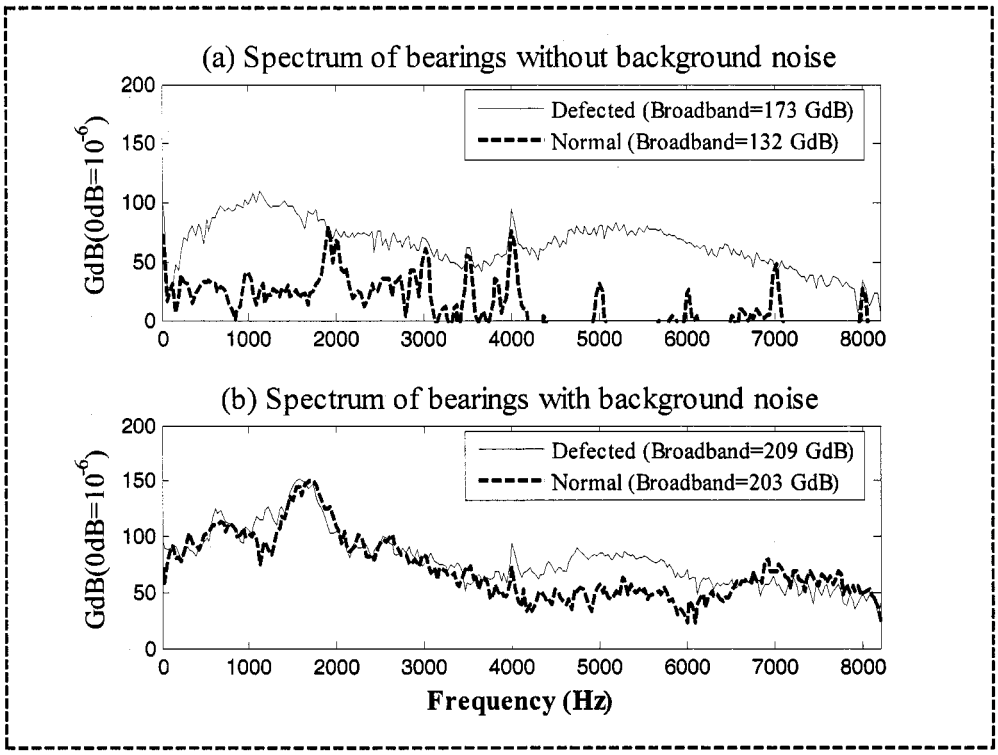


Figure 3-8. Spectrum of normal bearing and defected bearing in inner race (a) without and (b) with background noise.

- Narrowband method

As we know, vibration represents the behavior of the machine's mechanical components as they react to internal or external forces. Each machine defect produces a unique set of vibration components. We can calculate the specific frequencies associated with the components in the machinery for fault identification. The abnormal conditions in machines can be diagnosed by monitoring the trend of these frequency components in the spectrum. This method has been used successfully for most cases of machinery diagnostics. This is also known as the narrowband method. A baseline spectrum is created to represent the normal condition. New and current spectra will be compared with the baseline and the difference spectrum will show the changes in spectrum.

This method requires a prior knowledge of the signal and it is often a time consuming job to track many frequency components. In most of the cases, this method fails to detect the incipient faults since some of the features mask the most important feature.

Figure 3-8 is the spectrum and its broadband level calculated from the bearing data in Figure 3-7. The change of broadband level between a normal bearing and a defective bearing is much smaller in the presence of background noise. In this case, the narrow-band method is better than broadband method because there is an average of 25 dB difference between 4k to 6k Hz even with the background noise.

3.5.2.2 Processed spectrum methods

When a machine is complicated, it will have many sources of vibration. In such a case, the direct spectrum method is not an efficient way to detect a fault. Additional processing is needed in order to identify the source of the defect. Envelope power spectrum and cepstrum are two popular processing methods.

- Envelope power spectrum analysis

Envelope power spectrum analysis, also known as high frequency resonance technique (HFRT) was developed in 1970s for bearing defect analysis [104]. It can magnify the low amplitude, high frequency signals by filtering out low frequency components [89,90]. The vibration signal is first band filtered to remove the major influences from other vibration sources and background noise. The filtered signal is then passed through an amplitude demodulator. The fault indicating frequencies can be determined in the envelope spectrum through a spectrum analyzer. For digital signal processing, the same envelope spectrum can be obtained by using a Hilbert transform [83]. This method can be enhanced by calculating the auto-correlation of the envelope.

The process of enveloping can mask important components resulting from the amplitude modulation. Another limitation of envelope analysis is due to the fact that the damaged surface spreads randomly, causing the response of the vibration to become random

as well [84]. Therefore, the success of using this method for bearing failure detection is mixed [82].

Figure 3-9 shows an example of envelope analysis for a bearing with an inner race defect with and without the presence of background noise. The first column is the case without background noise. The impulse is nicely shown in the envelope and its auto-correlation. When the signal is contaminated with noise as in the second column, we can use a high pass filter to extract low-level impulse train from the large noise as in the third column.

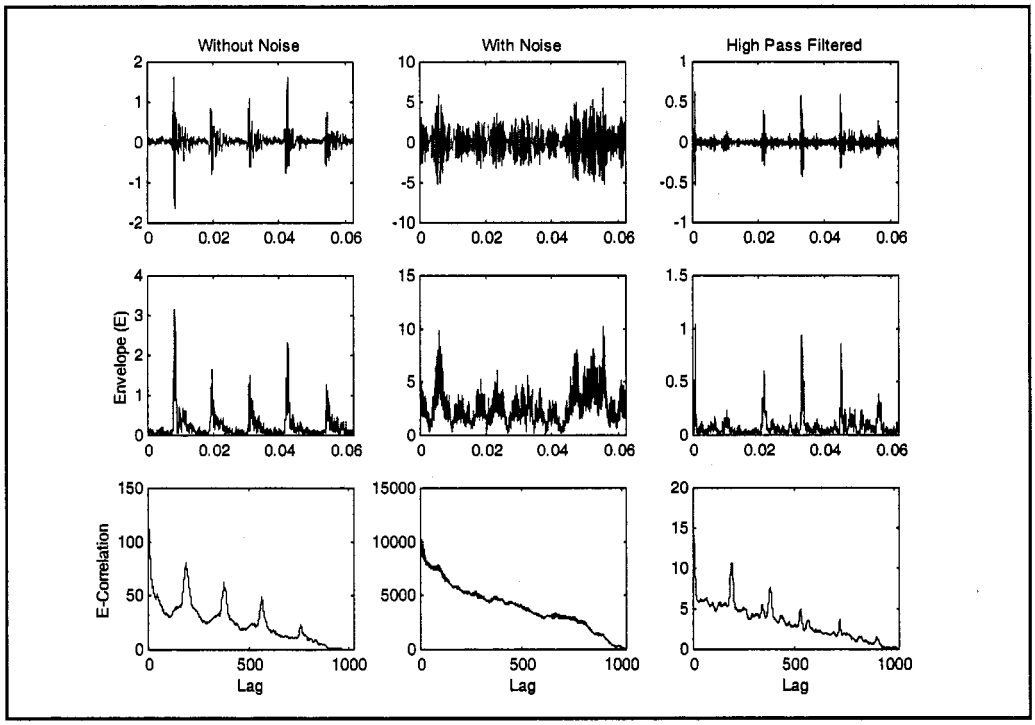


Figure 3-9. Envelope and its auto-correlation for inner race defective bearing with and without background noise.

- Cepstrum analysis

Cepstrum is defined as the power spectrum of the logarithm of the power spectrum. It is simply a spectrum of a spectrum and is in dB scale. It reduces a multiple component harmonic series into a single component. Therefore, cepstrum is ideally suited to the

analysis of complex signals with multiple harmonics. Cepstrum analysis had many applications and is widely used for echo detection, speech processing, bearing fault detection [17], and internal combustion engine diagnosis [18]. A bearing defect generates low energy level impacts making them hard to detect. However, an impact tends to produce a harmonic series in the frequency spectrum with a spacing equal to the frequency of the impact. When there are more than one periodic harmonics, it is more difficult to identify the individual components [92].

3.5.3 Time-Frequency methods

The traditional methods based on the results of Fourier transform are a powerful tool for deterministic and stationary signals. However, for non-stationary signals, Fourier transforms are not sufficient since the time-varying information is missing. For example, the flaws in a bearing produce impulses which are non-stationary transients. The time localization information is important for finding the onset of the bearing faults and the type of bearing failure which cannot be observed with power spectrum. In order to see the temporal information in the spectrum, we can map the signal into a function of time and frequency. The result of this two-dimensional map is called time-frequency representation (TFR).

There are many ways to map a one dimensional signal of time into a two-dimensional function of time and frequency [13,86]. The simplest method to implement TFR is to process the signal frame by frame using a fixed window to give a distribution of signals simultaneously in time and frequency [13]. This is known as short time Fourier transform (STFT). For example, Yazici and Kliman use the spectrogram from a SHFT to form a feature vector for bearing fault detection with very high accuracy [94]. However, STFT suffered a resolution trade-off problem between time and frequency. An alternative method is called Wigner-Ville distribution (WVD), which provides a good resolution both in time and frequency at the expense of interference terms appearing in the TF

plane. Meng and Qu use WVD to observe bearing defect impulses as an energy distribution in both time and frequency domains [93].

Recently, another TFR method is rapidly evolving called wavelet transform. Wavelet transform is also known as time-scale transformation. Wavelet techniques decompose the signal as a sum of orthogonal signals corresponding to different time scales using a multi-scale analysis [23]. It allows us to use a different window size for different scales. Peng and Chu provide a good review and bibliography by applying wavelet transform to machine condition monitoring and fault diagnostics [95].

3.6 Summary

The objective of this study is to detect bearing defect as soon as possible. Therefore, we are more interested in the techniques which can extract the small impulsive components under the presence of noise in the early stage. After reviewing the bearing failure stages and the techniques to detect bearing defects, we will focus on the high frequency resonance techniques (HFRT). In the next chapters, new signal processing tools, such as wavelet transformation and signal energy operator are proposed to improve traditional envelope methods. The results of these new approaches will be compared against the traditional envelope analysis method.

Chapter 4 Wavelet Transform Analysis

In the early stage, a bearing defect usually produces a low-level, non-stationary vibration which can be masked by background noise and other vibration components. Traditional spectral analysis is unable to accurately detect such defect-induced signals because Fourier analysis is based on the assumption of stationarity. To overcome the shortcomings of traditional spectral analysis, non-stationary analysis techniques such as time-frequency analysis and wavelet transform analysis have been developed. We choose wavelet transform analysis because it has the following properties [37]:

- Multi-resolution: Wavelet transform decomposes a signal using different window size at different scale. A large window size for lower scale and a small window size for higher scale. This allows us to capture both the short-time high-frequency information and the long-time low-frequency information of the signal.

- Localization: A wavelet is a function that decreases quickly towards zero at both ends. We can control the shape of the function using dilation and shifting it using translation to capture a localization event. This is especially useful for detecting impulse-like transients in the signal.

- Reconstruction: A signal can be perfectly reconstructed from its WT at any resolution without information loss. It is also possible to reconstruct a filtered version of the raw signal by selecting a subset of WT coefficients. This provides us an opportunity to obtain an optimal resonance frequency region for HFRT where the impulse to noise is highest.

There are many ways to perform WT analysis. Section 4.1 will give the definition of WT in the continuous and discrete from, namely the continuous wavelet transform (CWT) and the discrete wavelet transform (DWT), respectively. In Section 4.2, we introduce a very flexible and useful extension of the conventional DWT known as discrete wavelet packet transform (DWPT). The result of DWPT is a redundant version of DWT in a tree structure. Section 4.3 introduces a best basis algorithm for defining an

optimal collection of packets to represent the original signal. One of the problem for both DWT and DWPT is that the transformations are shift variant. Section 4.4 defines the shift invariant version of DWT and DWPT by the concept of maximal overlap [34]. Section 4.5 presents the denoising techniques which allow us to estimate a signal hidden by background noise using DWPT. Most of the computation algorithms can be found in the references [26,34] and will not be presented in detail here.

4.1 Introduction to Wavelet Transform

It is well known that the Fourier transform (FT) decomposes a square integrable signal (i.e. an energy limited signal) into the frequency domain with the kernel $e^{-i\omega t}$. The assumption of the FT is that the signal $x(t)$ is stationary in time. The transformation pair of the FT is:

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\omega)e^{i\omega t}d\omega, \tag{4-1}$$

$$X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-i\omega t}dt. \tag{4-2}$$

Because FT assumes the observed system is stationary, time localized information cannot be seen. In order to compensate for the missing temporal information, we can divide the signal into several short windows and perform time localized FT in order to incorporate the temporal dynamic feature. This technique is called short time Fourier transform.

Instead of using the $e^{-i\omega t}$, the wavelet transform (WT) uses wavelet bases $\psi_{\lambda,t}(u)$ to decompose signal into the time-scale domain [33]. A wavelet basis is a set of linearly independent functions constructed from a single mother wavelet with dilation and translation, i.e.,

$$\psi_{\lambda,t}(u) = \frac{1}{\sqrt{\lambda}}\psi\left(\frac{u-t}{\lambda}\right), \tag{4-3}$$

where λ is the scale parameter (also known as dilation parameter) and u is the time shift parameter (also called translation parameter) and $\psi(u)$ is the mother wavelet function. A wavelet is any function that integrates to zero and is square integrable, i.e.,

$$\int_{-\infty}^{\infty} \psi(u)du = 0, \tag{4-4}$$

$$\int_{-\infty}^{\infty} \psi^2(u)du < \infty. \tag{4-5}$$

There are many different types of wavelet functions and their families such as Daubechies and Morlet [23]. Figure 4-1 shows an example of Daubechies wavelet family.

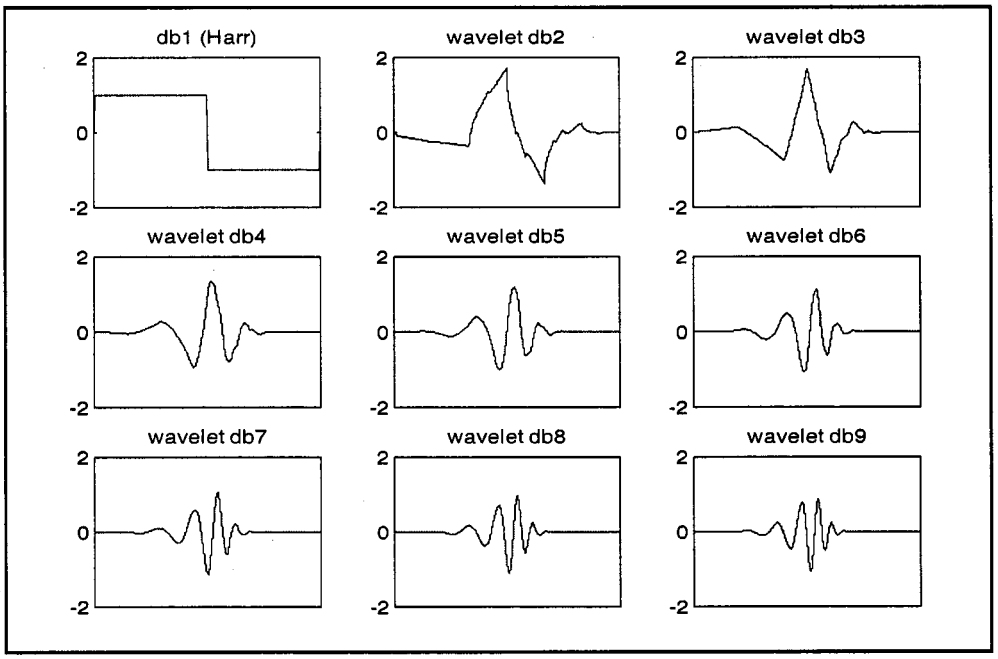


Figure 4-1. Daubechies wavelet family

Form Equation 4-3, a continuous wavelet transform (CWT) is defined as

$$\mathbf{W}(\lambda, t) = \int_{-\infty}^{\infty} x(u)\psi_{\lambda,t}(u)du \tag{4-6}$$

and the inverse CWT is

$$x(t) = \frac{1}{c_\psi}\int_0^\infty \int_{-\infty}^{\infty} \mathbf{W}(\lambda, u)\frac{1}{\sqrt{\lambda}}\psi\left(\frac{t-u}{\lambda}\right)du\frac{d\lambda}{\lambda^2} \tag{4-7}$$

where the constant c_ψ is equal to $\int_0^\infty \frac{|\psi(u)|^2}{u}du$.

In a discrete form, the DWT is a subsampling of $\mathbf{W}(\lambda,t)$ in a dyadic scale (i.e., $\lambda=2^{j-1}$, j=1,2,..). For a signal $x[n]$ with length N=2^J, the DWT can be obtained by

$$\mathbf{W} = wX, \tag{4-8}$$

where $\mathbf{W}$ is an N-by-1 column vector composed by the DWT coefficients W_n, w is an N-by-N transformation matrix and X is an N-by-1 column vector of $x[n]$. Note that the length of $x[n]$ is restricted to a power of 2.

The signal can be recovered by the synthesis DWT using

$$X = w^T\mathbf{W} = \sum_{j=1}^{J} D_j + A_J, \tag{4-9}$$

where j is the decomposition level, D_j is the jth level wavelet detail and A_J is defined as the approximation (or smooth) of the signal.

To implement DWT for digital computation, there is an efficient DWT algorithm that uses the method of multi-resolution analysis (MRA). The idea of MRA is to decompose signal with a two-channel filter bank and a downsampling process recursively [24]. Figure 4-2 is a schematic diagram of a two-channel filter bank where G represents a highpass (or wavelet) filter and H is a lowpass (or scaling) filter. The first stage of MRA uses G to compute the first detail (D_1) of the signal and uses H to obtain the first approximation (A_1). The downsampling process (↓) keeps the total number of coefficients equal to the total length of the original signal. Then, we continue the same decomposition on the approximation part iteratively as shown in Figure 4-3 to a level j. The result is a collection of components with 1 approximation and j details. This algorithm is

called a "pyramid algorithm" and the cost of computation is only $O(N)$ multiplications which is more efficient than FFT using $O(N\log_2 N)$ multiplications.

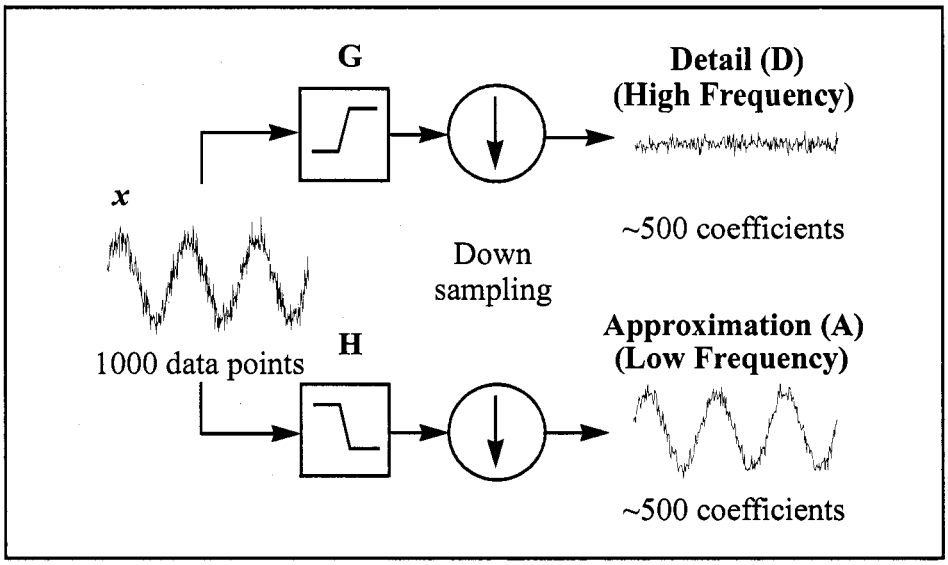


Figure 4-2. Schematic diagram of a two-channel filter bank for discrete wavelet transform.

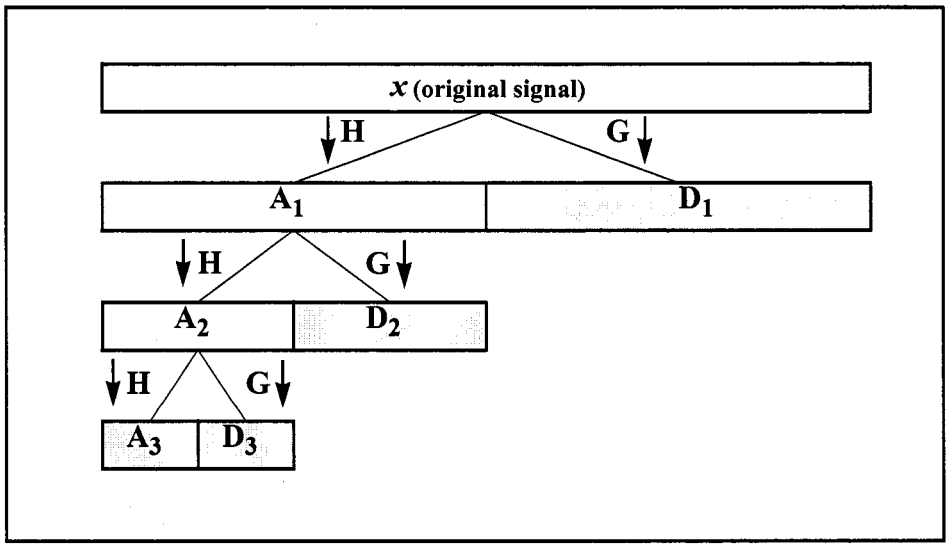


Figure 4-3. A three level discrete wavelet transform.

4.2 Wavelet Packet Transformation

Scale and frequency are related. A low scale using a compressed wavelet is equivalent to a high frequency component in the frequency spectrum [25]. In the original DWT, a signal is decomposed in a dyadic way that is equivalent to dividing the frequencies into octave bands. For J level decomposition, each detail contains information in the normalized frequency interval of $[1/2^{j+1}, 1/2^{j}]$ and the approximation coefficients are associated with the frequency interval of $[0, 1/2^{J+1}]$. Therefore, the first detail coefficients contain the information in the frequency range of [1/4 1/2] of the sampling frequency. For a machinery monitoring and diagnostics problem, this representation is not effective for feature extraction because many fault-indicating components present in a specific high frequency range [39]. This leads to an instinctive thought to apply the same iterative decomposition on both the detail and the approximation subbands at each level. This method of decomposition is defined as the discrete wavelet packet transform (DWPT). The result of the DWPT gives us subbands with even bandwidth at the same scale which allows us to have more flexibility to select the most informative packets with better resolution in the higher-frequency region [32]. Each subband is a square integrable modulated waveform, well localized in both position and frequency [33] that is called the wavelet packet.

The result of the DWPT can be called a time-frequency decomposition since each DWPT coefficients can be localized to a particular band of frequencies and a particular interval of time. Figure 4-4 is the structure of a DWPT and Figure 4-5 depicts the DWPT on a signal use Haar filters in sequence order [33].

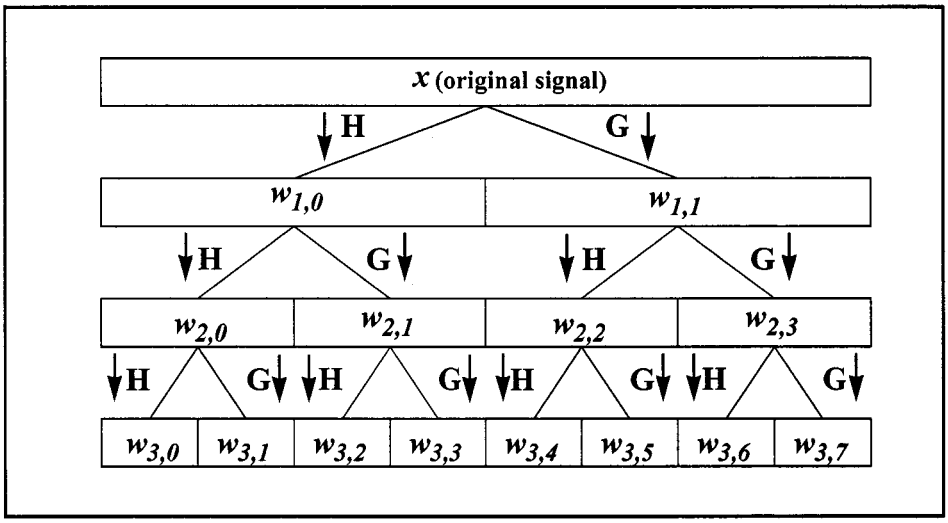


Figure 4-4. Tree structure of a sequence ordered wavelet packets.

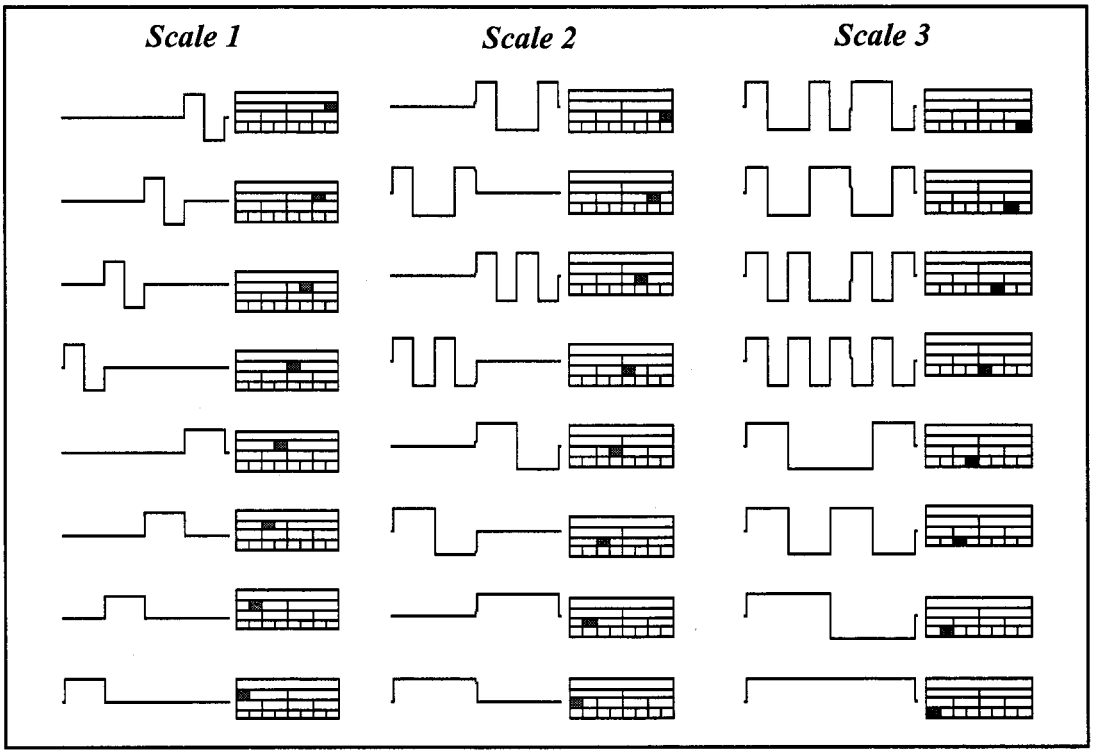


Figure 4-5. Harr wavelet packets.

4.3 Best Basis Algorithm

The result of DWPT is a redundant version of DWT in a dyadic tree structure which offers the freedom and possibilities to select different combinations of packets for representing the signal. For a PdM system, it is desirable to select the most important but the least redundant features for machinery fault diagnostics [39]. This requires a proper measurement to evaluate the information content in each wavelet packet and a criteria to select an optimal collection of packets to represent the machinery condition.

The original way of selecting a collection of packets is an entropy-based best basis algorithm proposed by Coifman and Wickerhauser [32]. The idea is to build a collection with the lowest information cost.

For a given signal x, the amount of information (I), is defined as

$$I(x_i) = -\log S_i^2 . \tag{4-10}$$

We can define a function to measure the cost of describing such signal. The most common used cost function is entropy, $E(S_i)$, that measures the amount of information required in specifying the signal S_i defined as follows:

$$E(S_i) = \sum_i S_i^2 I(S_i) = -\sum_i S_i^2 \log(S_i^2) . \tag{4-11}$$

This means that the cost function should be large when the coefficients are roughly the same size and small when all but a few coefficients are negligible [33]. For example, the cost for describing an impulse train signal is smaller that the cost for describing a periodic sine signal.

The entropy cost function is an additive cost function, which satisfies the additivity condition below:

$$E(S) = \sum E(S_i); \qquad E(0) = 0 \tag{4-12}$$

Therefore, the total cost $E(S)$ is the summation of the costs from each individual packet for the DWPT. The most compact representation is the combination of packets with the minimum cost function and the search for the minimum cost of the for DWPT is called the best basis algorithm.

Figure 4-6 illustrates an example of the best basis algorithm given by Wickerhauser [33]. The first step is to decompose a signal into a 3-level wavelet packet tree. Each rectangular is a subspace of the original signal defined as a node. In the top-down direction, each parent node is the sum of its two children nodes. The information cost is calculated as a number in each node. We start from the lowest level by marking each node with an asterisk as shown in Figure 4-6(a). The total information cost of this level is an initial value which we will try to minimize. In order to find the minimum, we compare the total cost of children nodes with their parent node. Whenever a parent node is of lower information cost than the children, we mark the parent. If the children have lower information cost, we do not mark the parent, but we assign the lower total information cost of the children to the parent. The former information costs of the adjusted nodes are also displayed in parentheses. After the search, we mark all nodes of lower cost as in Figure 4-6(b). The topmost marked nodes, as the shaded boxes in Figure 4-6(c), constitute the best basis.

The entropy-based best basis works well for voice signal and image compression by searching the lowest cost to transmit the information. However, it is not the only way to form a basis. For applications such as condition monitoring and diagnostic systems, we are more interested in the rate of energy change rather than the concentration of energy in the signal. One of the objectives of this research is to select a cost function that is suitable for the early detection of bearing defects. This cost function will be presented in the next chapter.

Figure 4-7 shows three kinds of wavelet packet representations using a signal from a bearing with an artificially seeded inner race defect. The topmost block in each of the tree structure is the original signal. Figure 4-7(a) uses all the 16 packets at the lowest

level to represent the signal which is roughly analogous to a STFT. A wavelet tree is also known as a DWT. Figure 4-7(b) is a wavelet tree which is the same as the DWT by splitting only over the lower frequency bands. The 5 features given by this split are indicated in Figure 4-7(b) with thick box lines. Any other combination corresponding to a disjoint dyadic will also work. An example of best basis tree using entropy criteria is shown in Figure 4-7(c). This gives us 7 features with different bandwidths.

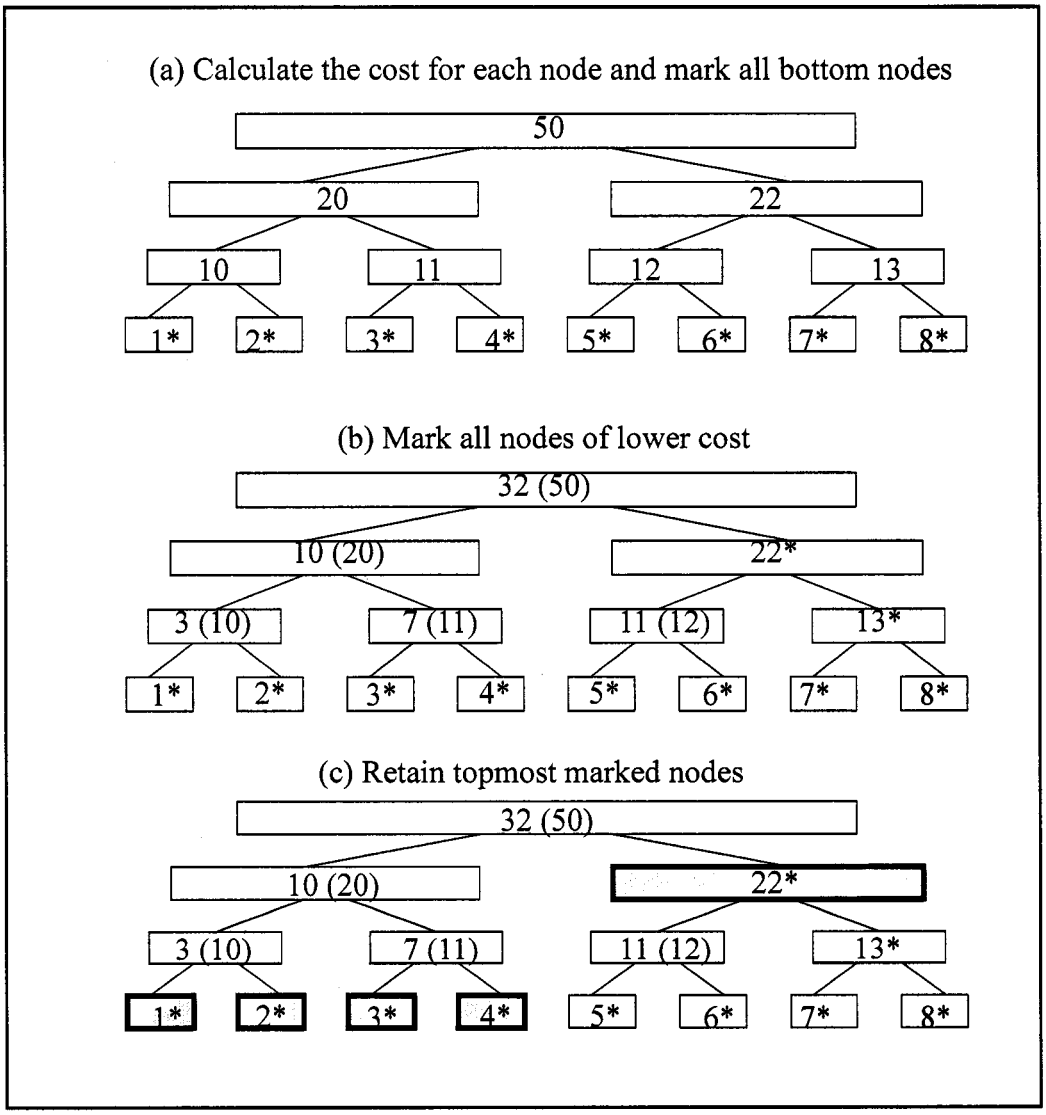


Figure 4-6. Algorithm of searching the best basis in a three level wavelet packet decomposition.

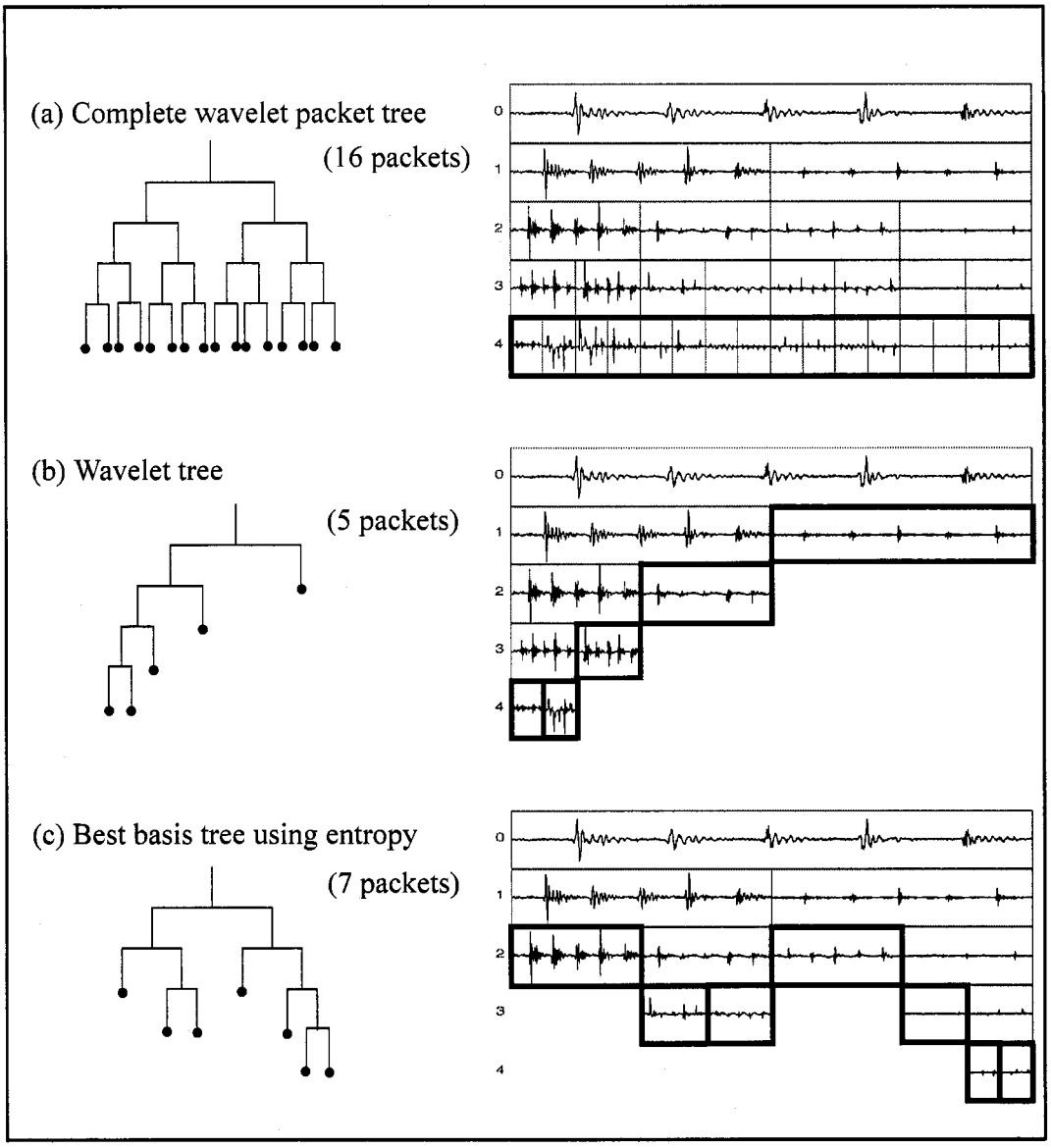


Figure 4-7. Wavelet packet tree structures of a bearing signal with inner race defect using Daubechies wavelets db5.

4.4 The Maximal Overlap DWT and DWPT

Due to the fact that DWT uses downsampling on the outputs of the wavelet and scaling filters at each stage of the pyramid algorithm, the result of DWT is dependent on where

the transformation is performed on the signal. This implies that DWT is a shift invariant transformation which, is misleading.

In order to solve the shift variant problem, many studies proposed a modified version of DWT under different names, such as "undecimated DWT", "shift invariant DWT", "stationary DWT", "time invariant DWT" and "maximal overlap DWT (MODWT)" [34]. In this work, we use MODWT as the acronym for this variation.

The essential concept to get a shift invariant DWT is to apply the DWT pyramid algorithm twice, once to the original signal and once on the one-point delayed signal. Then, the combined coefficients from the two sets of DWT coefficients is MODWT. MODWT is superior than DWT in several ways in additional to the shift invariant perperities. Table 4-1 lists the main differences between DWT and MODWT. The main cost of MODWT is the computation cost which is the same as FFT.

The same idea can be extended to define a maximal overlap DWPT (MODWPT) from DWPT.

Table 4-1. Property comparison between MODWT and DWT.

Property	MODWT	DWT
Computation cost	$O(N \log_2 N)^*$	$O(N)$
Restriction on sample size	None	2^J
Usable for multi-resolution analysis	X	X
Details and approximations are zero phases	X	
Wavelet coefficients usable for analysis of variance	X	X
Scaling coefficients usable for analysis of variance	X	X
Details and approximations usable for analysis of variance		X
Shift invariant transformation	X	

*. Computation cost is the same as FFT [35].

4.5 Denoising

The wavelet denoising method is a very useful technique to extract (or estimate) a significant signal masked by noise. Assume the following corrupted signal model

$$x[n] = u[n] + \sigma z[n], \; n=1,2,3,...,N \tag{4-13}$$

where $u[n]$ is the deterministic signal that interests us, $z[n]$ is a independent and identically distributed (IID) noise with zero mean, and σ is the standard deviation of $z[n]$.

If we apply the DWT to $x[n]$, Equation 4-13 becomes

$$\mathbf{W}_x = \mathbf{W}_D + \sigma \mathbf{W}_Z, \tag{4-14}$$

where $\mathbf{W}$ represents the wavelet coefficient vector of length N. Assuming that some of the wavelet coefficients are small and attributable to noise, we can use M ($M < N$) number of coefficients to estimate the signal as $\hat{x}[n]$.

To determine M, Dohono and Johnstone have developed a wavelet shrinkage method using a threshold function [40]. Hard thresholding and soft thresholding are the most commonly used threshold functions, and they are defined as

$$W_l^{hard} = \begin{cases} 0 & |W_l| \le \delta \\ W_l & \text{otherwise} \end{cases}, \tag{4-15}$$

$$W_l^{soft} = \text{sign}\{W_l\}(|W_l| - \delta), \tag{4-16}$$

where δ is a threshold level.

There are a number of ways to set the level of δ. When z is a IID white noise with zero mean and covariance σ_z^2, a universal threshold δ^u is given by

$$\delta^u = \sqrt{2\sigma_z^2 \log_e N}. \tag{4-17}$$

If σ_z^2 is unknown, the standard deviation estimator can be used to estimate σ_z^2 by

$$\hat{\sigma} = \frac{\text{median}\{|W_1|, |W_2|, \ldots, |W_{N/2-1}|\}}{0.6745}, \tag{4-18}$$

where $\{|W_1|, |W_2|, \ldots, |W_{N/2-1}|\}$ are wavelet coefficients in $\mathbf{W}_1$ at level 1 decomposition [34]. The reason for using only $\mathbf{W}_1$ to estimate the standard deviation is that the high frequency region is usually dominated by noise. The constant 0.6745 rescales the estimator to be a suitable estimator for the white noise case.

Figure 4-8 shows an example of denoising performance on a noised bearing signal with a seeded inner race defect where (b) is the original signal and (c) is the noised signal at SNR level of 2. Figure 4-8 (c) to (e) are the denoised signals at different levels with the universal threshold $\delta^u = 3.7233$. As we can see from the denoised signal, the signal tends to be more smooth as we perform the denoising at higher level.

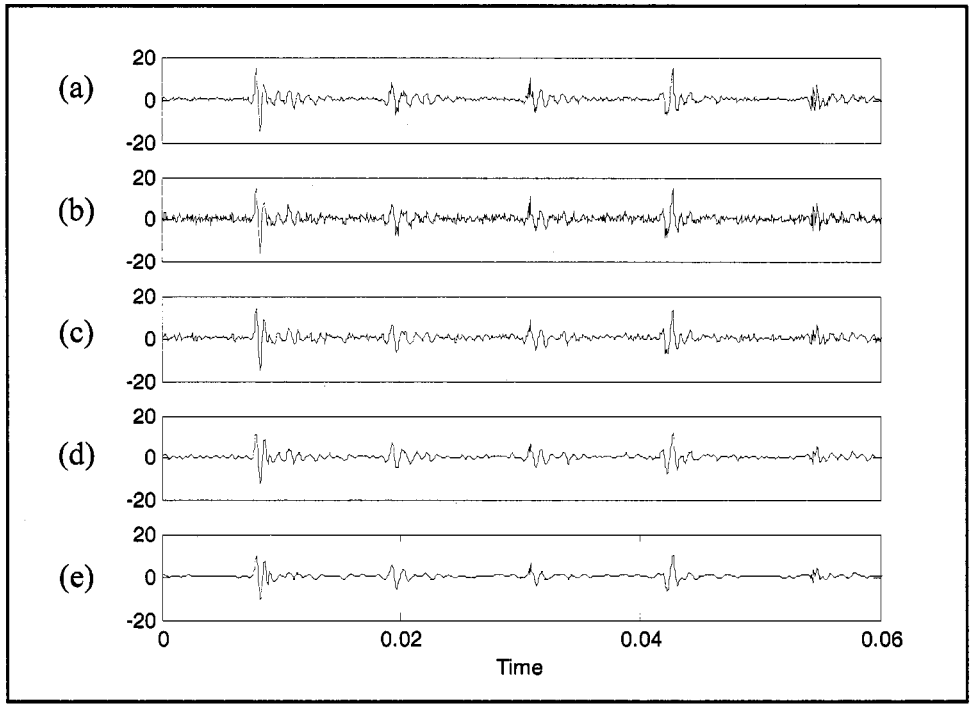


Figure 4-8. Denoising bearing signal with seeded inner race defect. (a) Original signal, (b) Original signal with white noise at SNR=2, (c) Denoised signal at DWT level 1, (d) Denoised signal at DWT level 2, (e) Denoised signal at DWT level 3.

4.6 Summary

In this chapter, we have introduced most of the wavelet transform techniques including: CWT, DWT, DWPT, MODWT, and MODWPT. For bearing defect detection, DWPT and MODWPT might be the best techniques to provide higher resolution in the high frequency region. We also presented the best basis algorithm using the concept of cost function. A cost function that can best describe the bearing defect is desirable, and this is which will be investigated in the following chapter. Finally, the denoising technique is another useful tool which can reduce the noise effect while using wavelet transform techniques. Denoising will be incorporated with our wavelet based high frequency resonance technique in the next chapter.

Chapter 5 High Frequency Resonance Technique

The high frequency resonance technique (HFRT), also known as envelope power spectrum analysis, envelope detection, or demodulated resonance analysis [81], is a traditional method for extracting repetitive low level fault related vibrations in the presence of high background noise levels. The development of HFRT can be traced back to 1970s and was developed for early warning of impending bearing failures [100]. Darlow and Badgley point out the HFRT can also be used for other mechanical components problems, such as gear interaction and clutch slippage, where the low-frequency information of interest is modulated at the higher resonance frequency region [99]. At present, the maintenance industry has adapted HFRT as one of the main tools in the condition monitoring program for early gear and bearing defect detection [57,92,103,108].

The purpose of the HFRT is to extract the bearing defect impulses using a demodulation process. Traditional methods to demodulate the signal use either rectifier circuits or Hilbert transforms. In this study, we introduce alternatives based on energy detectors and wavelet packet transformation.

This chapter starts by an introducing the concept of amplitude modulation in the bearing vibration signal. The basic principles and the conventional implementation of HFRT based on Hilbert transform is presented in Section 5.2. Three alternatives based on energy detectors and wavelet packet transformation are introduced in Section 5.3 and 5.4, respectively. Section 5.5 summarizes the traditional and proposed new HFRT algorithms.

5.1 Amplitude Modulation of Bearing Vibration

The spectrum of an impulse, which is similar to a white noise signal, spreads to all frequencies at low amplitudes. In a defective rolling element bearing, when a dent is struck by a rolling element, a number of high frequency resonances may be excited. These high frequency resonances could be any combination of the ringing of bearing components, bearing housing, machine structure, or transducer. These resonances are excited periodically at the particular bearing defect frequency. As a result, a bearing defect generates an amplitude modulation wave at the structural resonance frequency modulated by the bearing defect frequency.

5.1.1 Amplitude modulation

Amplitude modulation (AM) is defined as the multiplication of two signals at different frequencies [103]. The high frequency component is called the carrier signal and the low frequency component is the modulating signal. The general form of an AM signal (x_{AM}) is given by

$$\begin{aligned} x_{AM}(t) &= m(t)\cos(\omega_c t) + C\cos(\omega_c t) \\ &= [C + m(t)]\cos(\omega_c t), \end{aligned} \tag{5-1}$$

where $m(t)$ is the modulating signal, ω_c is the carrier angular frequency, and C is the amplitude of the carrier signal. The Fourier transform of x_{AM} is

$$\begin{aligned} X_{AM}(\omega) &= \frac{1}{2}[M(\omega + \omega_c) + M(\omega - \omega_c)] \\ &\quad + \pi A[\delta(\omega + \omega_c) + \delta(\omega + \omega_c)] \quad . \end{aligned} \tag{5-2}$$

The instantaneous envelope of the signal is defined as $A(t) = |C + m(t)|$.

If $m(t)$ is a single tone signal with a modulating angular frequency of ω_c, then the AM signal is

$$
\begin{aligned}
x_{AM}(t) &= m(t)\cos(\omega_c t) + C\cos(\omega_c t) \\
&= M\cos(\omega_m t)\cos(\omega_c t) + C\cos(\omega_c t) \\
&= \frac{M}{2}[\cos(\omega_c + \omega_m)t + \cos(\omega_c - \omega_m)t] + C\cos(\omega_c t) \quad .
\end{aligned}
\tag{5-3}
$$

Figure 5-1 shows an example of an amplitude modulation waveform and its spectrum. The modulating signal causes the amplitude of the high frequency signal (1000Hz) to fluctuate at the rate of the modulating frequency (100Hz). The spectrum shows 3 peaks at the carrier frequency (f_c) and two sidebands (f_c-f_m and f_c+f_m) alongside it.

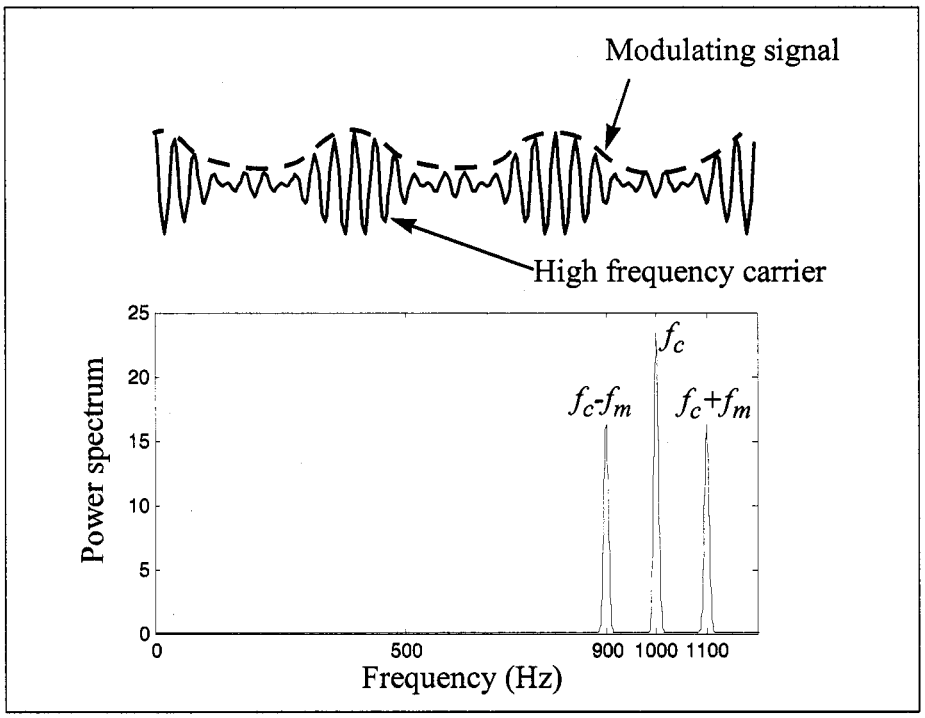


Figure 5-1. Amplitude modulation waveform and its spectrum

5.1.2 Amplitude demodulation

The process to recover the modulating signal from the AM signal is called *demodulation*. There are 3 common methods of AM demodulation: 1) rectifier detector, 2) envelope detector, and 3) square-law detector [107]. The rectifier detector uses a diode and

resistor circuit to suppress the negative part of the AM wave followed by a lowpass filter to remove the higher frequency harmonics. The envelope detector is a peak-hold circuit using a capacitor and a resistor to control the time constant. The capacitor charges up to the peak voltage of the input signal and then discharges through the resistor at a slow rate. The output voltage thus follows the envelope of the modulating signal. An AM signal can also be demodulated by squaring it and then passing the squared signal through a lowpass filter.

5.1.3 The effect of interference in AM signal

In real situation, the signal is always corrupted by interference or noise. Consider an AM signal is with a sinusoidal interference signal $I\cos[(\omega_c + \omega_d)t]$, i.e.,

$$\begin{aligned} r(t) &= [C + m(t)]\cos(\omega_c t) + I\cos[(\omega_c + \omega_d)t] \\ &= \{[C + m(t)] + I\cos(\omega_d t)\}\cos(\omega_c t) - I\sin(\omega_d t)\sin(\omega_c t) \end{aligned} . \quad (5\text{-}4)$$

Hence, the envelope of $r(t)$ is

$$A(t) = \sqrt{[C + m(t) + I\cos(\omega_d t)]^2 + [I\sin(\omega_d t)]^2} . \quad (5\text{-}5)$$

If the interference amplitude is small ($I \ll C$), then $A(t) \cong C + m(t) + I\cos(\omega_d t)$. The spectrum of the envelope will have a line at ω_d in addition to the modulating component. However, if the interference amplitude is greater than carrier amplitude ($I \gg C$), envelope can be approximated as $([C + m(t)]\cos(\omega_d t) + I)$ and the spectrum of the envelope will have sidebands around the interference frequency ω_d.

The noise can be represented by a sum of sinusoids of frequencies $\Delta\omega$ apart ($\Delta\omega \to 0$). If the noise is much smaller than the AM signal, the interference of noise can be neglected. When the interference is large, sidebands and lines of noise make it is impossible to extract the desired modulating signal. Therefore, it is desirable to reduce the interference signal before the demodulation process.

5.2 Conventional Implementation of The HFRT

The essential concept behind using the HFRT is an amplitude demodulation process that extracts the signal of interest from the vibration signal. Of interest in the HFRT analysis is not which system or transducer resonance frequency is excited (the carrier), but the impulse repetition frequency (the modulating component) in the modulated signal that is related to the bearing defect. In this section, we will present the process of the conventional HFRT based on Hilbert transform as well as its limitations.

5.2.1 Process of the HFRT

There are several steps to perform HFRT to extract low-level, impulse train signal generated by bearing local defects. The process of HFRT includes pre-filtering, enveloping and post signal analysis as illustrated in Figure 5-2 [101].

5.2.1.1 Pre-filtering

The first step of applying HFRT is a filtering process. The purpose of pre-filtering is to extract the most representative resonance signal with the best signal to noise ratio by a bandpass or highpass filter. Within industrial environments, a signal recorded by an accelerometer is usually complicated and includes multiple excitations from sources such as low frequency components, background noise and many resonances from other system components. A defect in a rolling element bearing generates an impact for each cycle of rotation. The spectrum of such an impact is a set of lines, separated by the rotation rate, that extend to very high frequencies. The low frequency lines may be masked by other sources, such as imbalance or misalignment, but the high frequency lines, are not so masked. Therefore, proper filter specifications are critical to eliminate undesired interferences in the signal for better demodulation results.

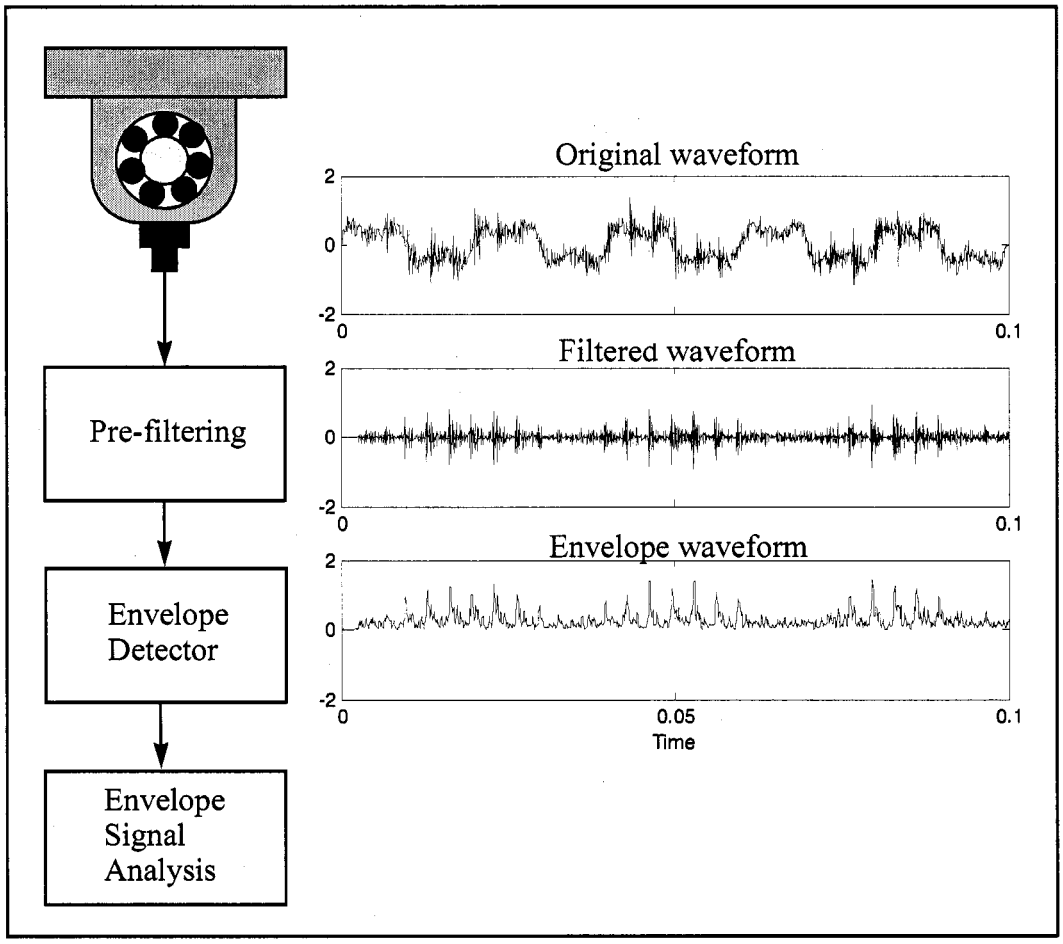


Figure 5-2. The HFRT process and the waveform of the signal produced.

5.2.1.2 Enveloping

The envelope of the filtered signal can be obtained with an analog circuit as described in Section 5.1.2. Alternatively, the signal envelope can also be calculated digitally using a Hilbert transform (HT) [100].

The HT for a real valued signal $x(t)$ extending from $-\infty$ to ∞ is defined by

$$H[x(t)] = \tilde{x}(t) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{x(\tau)}{t-\tau}d\tau, \tag{5-6}$$

where $\tilde{x}(t)$ is the HT of the $x(t)$. The energy of a signal and its HT are equal.

A major use of HT is to obtain the instantaneous magnitude (or envelope), phase and frequency of a real signal by generating an analytic signal [106]. The analytic signal $x_a(t)$ of a real signal $x(t)$ is a complex signal in the form of

$$x_a(t) = x(t) + i\tilde{x}(t). \tag{5-7}$$

The instantaneous envelope $A_h(t)$, phase $\varphi(t)$ and frequency $\omega(t)$ of the signal are calculated by

$$A_h(t) = |x_a(t)| = \sqrt{x^2(t) + i\tilde{x}^2(t)}, \tag{5-8}$$

$$\varphi(t) = \operatorname{atan}\left(\frac{x(t)}{\tilde{x}(t)}\right), \text{ and} \tag{5-9}$$

$$\omega(t) = \frac{d}{dt}\varphi(t). \tag{5-10}$$

5.2.1.3 Signal analysis

The envelope signal can be processed by time domain indicators or the frequency domain method described in Section 3.5 for fault detection and diagnosis. No modulation or a light sinusoidal modulation of the resonance produces a small value for the time domain shape indicator, but the modulation produced by impacts will have a higher shape indicator value. Pachaud, Salvetat and Fray have shown that the time domain shape indictor (i.e., the crest factor or kurtosis) from the envelope signal is more sensitive to early bearing defects because the filtering process reduces the background noise and low frequency harmonics [9].

The most common approach to analyzing the envelope signal is to calculate its frequency spectrum. The repetition impulse frequency generated by a defect will show up clearly in this enveloped spectrum. As the defect becomes more severe, the spectrum of defect frequency will have a number of modulation sidebands spaced at multiples of the shaft running frequency as well as harmonics of the shaft running frequency.

McFadden and Smith deduce that these sideband and harmonics are due to the defect moving into and out of the loaded zone at the shaft running frequency [100].

5.2.2 Limitations of the HFRT

There are several limitations of using Hilbert transform based HFRT for bearing defect diagnosis [57,100,104]. Several studies show that HFRT might not be an effective tool when multiple defects are presented because a considerable overlap will exist between the decay of one impulse to the next. The leading edge of some impulses may become buried in the decay of the previous impact. As the bearing becomes more damaged, a large number of defect peaks will be generated and these peaks will appear to be random. It is not then possible to identify the defect frequency. Therefore, extracting an instantaneous envelope with better time and frequency resolution is needed. In this study, we examine 3 envelope detection alternatives using energy detectors (see Section 5.3) and wavelet packet transformation (see Section 5.4).

Another difficulty of HFRT is to determine which resonance frequency region will give the best results from envelope detection. Until now, the determination of the best resonance frequency is mainly based on an ad hoc approach that consists of identifying a region of resonance within the standard spectrum or a series of experiments using filters with different frequency ranges [6,116,117]. A rule of thumb proposed by Weller is to set the lower corner frequency to be higher than 10 times the running speed of the machine and to set the upper corner frequency at approximately 60 times the ball pass outer race frequency [109]. A more elaborate method is to use impact testing to catch the resonant frequencies within the system. An impact test uses an instrumental hammer to excite the structure by external impacts at a succession of various points near the accelerometers and then taking the FFT of the resultant vibrations to obtain the response spectrum [108]. However, this technique is not only time consuming, but also unrealistic in practice for most industry applications because it requires finding the resonant frequencies for each component of machinery and all alternations of these reso-

nant frequencies resulting from different operational conditions. In addition, the bearing resonant frequencies obtained through this techniques are only estimations because the location of the hammer strikes is on the external casing of the bearing housing rather than on the bearing component itself.

In this study, we proposed a resonant frequency selection algorithm based on wavelet packet transformation. This technique is an automation selection process that compares historical data at each wavelet packet and looks for the packet where the maximum (relative) change of a pre-defined feature function occurs due to the bearing fault.

5.3 Energy Detector Based HFRT

In traditional signal processing, the square of the magnitude of the signal is an estimation of its instantaneous energy. Square-law is one of the standard demodulation methods based on this concept and is also called the linear (or standard) energy detector [110]. In 1980s, Teager derived an energy operator to estimate the energy required to generate a simple harmonic oscillation. Later, Kaiser extended this energy operator to a discrete form and its applications for AM-FM energy separation and detection of transient signals [111,112,113]. Fang and Atlas developed a quadratic energy detector to solve the time and frequency trade-off using linear energy detectors [114]. In this section, we will briefly review the background of these energy operators and detectors. Then, we will apply these energy detectors to the HFRT process as alternative methods for detecting bearing defects.

5.3.1 Energy operator

Consider a simple mass-spring system oscillating about the equilibrium position with mass m and spring constant k. The equation of motion of this system is

$$m\ddot{x} + kx = 0. \tag{5-11}$$

The solution of this second-order differential equation is given by

$$x(t) = A\cos(\omega t + \varphi), \tag{5-12}$$

where A is the amplitude of the oscillation, ω ($=\sqrt{k/m}$) is the frequency of the oscillation and φ is a phase angle resulting from the initial conditions.

Based on the law of conservation of energy, the total energy E of the system is the sum of the potential energy and kinetic energy and maybe expressed as

$$E = \frac{1}{2}m\dot{x}^2 + \frac{1}{2}kx^2 = \frac{1}{2}m\omega^2 A^2. \tag{5-13}$$

Equation 5-13 infers that the total energy of a simple oscillation is proportional to the square of its amplitude and frequency.

Using Equation 5-11, we can replace the kx term in Equation 5-13 and rearrange the energy as

$$E = \frac{1}{2}m(\dot{x}^2 - \ddot{x}x). \tag{5-14}$$

Neglecting the constant factor, Teager's energy operator is defined as

$$E_T(t) = \dot{x}^2(t) - \ddot{x}(t)x(t). \tag{5-15}$$

The discrete form of Teager's energy operator was introduced by Kaiser [111]

$$E_T[n] = x^2[n] - x[n+1]x[n-1]. \tag{5-16}$$

Note that Equation 5-16 is an invariant algorithm with only 2 multiplications and 1 subtraction. It is a computational efficient way to estimate the signal energy. However, this algorithm is sensitive to noise with an error equal to the variance of the noise.

In addition, Teager's energy operator is best for stationary signals with a single component because it is derived from a simple harmonic model. In order to estimate the energy of a non-stationary signal using an energy operator, Fang developed a generalized energy operator based on a mass-damper-spring system [110]. This generalized energy operator is derived as

$$E_G(t) = \begin{cases} E_T(t) & (\dot{x} = 0) \cup (x = 0) \\ E_T(t) + \frac{1}{2}\dot{E}_G(t)\frac{x(t)}{\dot{x}(t)} & (\dot{x} \neq 0) \cap (x \neq 0) \end{cases}, \tag{5-17}$$

where the $\dot{E}_G$ term comes from the energy loss due to damping.

Similar to Equation 5-16, the discrete form of the generalized energy operator can be obtained by approximating the continuous time derivatives

$$\begin{aligned} E_T[n] &= x^2[n] - x[n+1]x[n-1] \\ \dot{E}_T[n] &= \frac{1}{2}(E_T[n+1] - E_T[n-1]) \\ \dot{x}[n] &= \frac{1}{2}(x[n+1] - x[n-1]) \quad . \end{aligned} \tag{5-18}$$

To demodulate an AM signal using Teager's energy detector, we use Equation 5-15 to estimate the signal energy of x_{AM}, i.e.,

$$\begin{aligned} E_T(x_{AM}(t)) &= [C + m(t)]^2(\omega_c)^2 + \cos(\omega_c)^2 E_T(C + m(t)) \\ &= [C + m(t)]^2(\omega_c)^2 + \varepsilon(t) \end{aligned}, \tag{5-19}$$

where $\varepsilon(t)$ is an approximation error. Maragos, Kaiser and Quatieri have shown that the approximation error can be neglected for a band-limited modulating signal with bandwidth $\Delta\omega_m \ll \omega_c$ [112]. Hence, the approximated envelope is $|C + m(t)|\omega_c$. In general, the instantaneous frequency and amplitude of a narrowbanded signal can be approximated by

$$\omega_c(t) = \sqrt{\frac{E_T[\dot{x}(t)]}{E_T[x(t)]}} \text{ and} \tag{5-20}$$

$$|A(t)| = \frac{E_T[x(t)]}{\sqrt{E_T[\dot{x}(t)]}}. \tag{5-21}$$

Lin, Zhu, Wu and Zheng have shown that using energy operator demodulation in HFRT is an effective method for detecting rolling element bearing faults [115]. The amount of

calculations using an energy operator approach is about 1/log(N) compared with the amount of calculation in Hilbert transform demodulation.

5.3.2 Linear and quadratic energy detectors

The square law detector shown in Figure 5-3, which is also known as a standard linear detector, consists of a linear time invariant (LTI) filter, a magnitude squared operation, and a linear smoothing filter [107]. The estimated energy of a signal from the standard linear detector, before the application of a smoothing filter in continuous time is given by

$$E_l(t) = \left| \int_{-\infty}^{\infty} x(t-\tau) h_l(\tau) d\tau \right|^2, \tag{5-22}$$

where $h_l(t)$ is the impulse response of a linear bandpass filter. The discrete form of $E_l(t)$ can be expressed as

$$E_l[n] = \left| \sum_k x[n-k] h_l[k] \right|^2. \tag{5-23}$$

This energy detector using an LTI filter is easy to implement, however, it suffers from the resolution trade off between time and frequency [114]. Also, the performance of this linear detector is degraded when the system is signal dependent or where non-Gaussian noise is present.

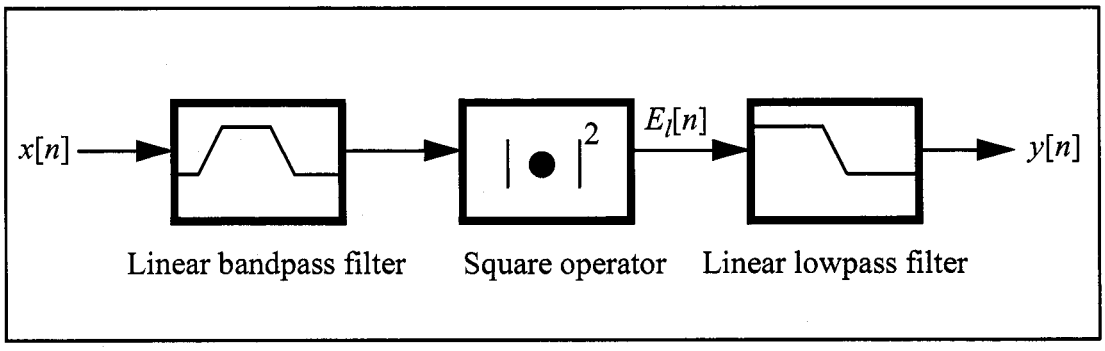


Figure 5-3. A typical linear energy detector.

Jing and Atlas developed a quadratic detector using kernel design methodology to construct a 2-D set of coefficients which integrates frequency filtering with energy detection [114]. Similar to the standard linear detector, a quadratic energy detector consists of a quadratic bandpass filter followed by a cascaded lowpass quadratic filter as shown in Figure 5-4. In general, the quadratic detector in continuous time can be described by

$$E_q(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} q(\tau_1, \tau_2)x(t-\tau_1)x(t-\tau_2)d\tau_1 d\tau_2, \tag{5-24}$$

where $q(\tau_1, \tau_2)$ is the kernel of the quadratic detector.

In this model, it is desirable for a quadratic detector to possess the following properties: 1) real output; 2) time-shift invariance; 3) no nonlinear frequency interference; 4) finite time support; 5) nonnegative output; 6) independent time and frequency resolution; and 7) output proportional to the Teager's energy property. It is impossible to possess all these desired properties due to the inconsistency of the constraints for providing these desired properties. For specific applications, a compromised quadratic detector can be implemented. The general discrete form for a finite-duration, compromised quadratic detector is given as

$$E_q[n] = \sum_{m_1=-M}^{M}\sum_{m_2=-M}^{M} h_q[m_1, m_2]x[n-m_1]x[n-m_2], \tag{5-25}$$

where h_q are quadratic filter coefficients.

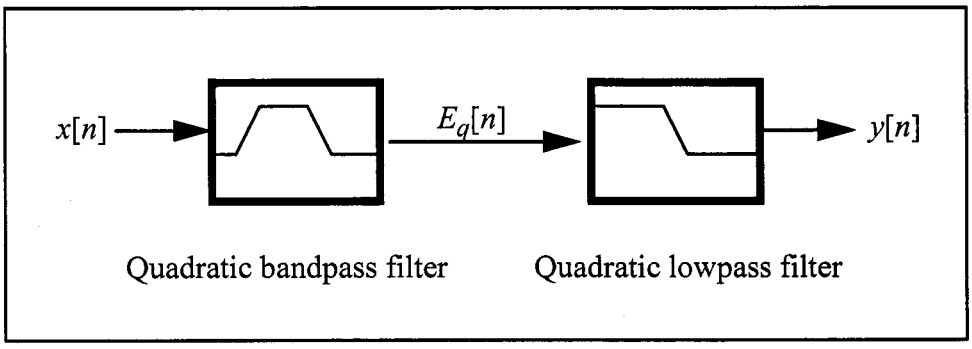


Figure 5-4. A quadratic energy detector.

For a compromised quadratic detector with the properties of frequency interference elimination, finite time support, and independent time-frequency resolution, the designed kernel is a (2*M*+1) by (2*M*+1) matrix and maybe expressed as

$$E_q[n] = \sum_{m_1=-M}^{M} \sum_{\substack{m_2=-M \\ |m_1-m_2|<\frac{L+1}{2}}}^{M} h_q[|m_1-m_2|]x[n-m_1]x[n-m_2] \quad , \tag{5-26}$$

where $2M>L$, L is the number of nonzero bands in the cross-diagonal direction, and $h_q[m]$ is a (4*M*+1) point symmetric set of quadratic detector coefficients. This compromised quadratic detector has been shown to suppress noise and extract AM from a multiple component signal without smoothing in time and frequency.

5.4 Discrete Wavelet Packet Transform Based HFRT

Recently, many studies have used the wavelet transform (WT) and the wavelet packet transform (WPT) for condition monitoring and fault diagnostics [122]. For example, Xia uses normalized RMS for all wavelet packets at the last level of decomposition to form a feature vector for pattern recognition of bearing faults [118]. Eren and Devaney investigate motor bearing damage detection by comparing the RMS of DWT at different scales and DWPT coefficients at different packets [119,120]. Zhang etc. use DWPT to de-noise vibration signals and pick up the fault frequency to enhance bearing defect detection under strong noise [121]. Shao and Nezu calculate the kurtosis of the de-noised wavelet coefficients as an indicator of bearing condition and use an envelope spectrum to diagnosis bearing defect [123].

In this work, we propose a novel algorithm to integrate an automation resonance frequency band selection criteria with the discrete wavelet packet transform (DWPT). The frequency selection criteria is based on the measurement of signal energy using RMS and spikeness using kurtosis.

5.4.1 Automation resonant frequency band selection

The extraction of high-frequency transients due to the bearing impact resonance is vital to HFRT for defect detection. It is desired to select the resonance frequency range which results in the best signal to noise ratio for demodulating bearing defect frequencies. Therefore, the objective of this section is to identify a resonance frequency band with the best signal to noise ratio in order to isolate the bearing transients from other signal components. The selected packet can then be used for subsequent processing.

Our motivation for the development of the proposed technique using DWPT is because a DWPT algorithm is based on the concept of MRA which is equivalent to a multiple bandpass filter bank. Each packet from the decomposition is related to a particular frequency band and a filtered signal can be reconstructed. Therefore, selecting an optimal resonance frequency range for HFRT is the same design problem as choosing which packet(s) from the entire DWPT collection should be used for HFRT.

The most common way to select the best representative packets is to use the best basis algorithm with an information cost function (i.e., least Shannon entropy) [32]. However, a measure of information entropy might not be the best feature to represent a bearing defect signature. For representing and classifying a vibration waveform, both the strength and the shape of the signal is important. A strength of the signal can be defined in terms of the amplitude or the energy of the signal. A shape indicator can be described using its statistical property such as kurtosis and skewness. Most of the studies using DWPT or DWT for bearing defect detection choose RMS as the sole defect-representing feature at different packets or scale. For early stage bearing defect detection, the change of RMS might not be sensitive to the low-level impact component. The packet with the dominant energy change is very likely the result of other vibration sources. Therefore, a measure of signal energy using RMS should be combine with a shape property for more robust representation.

Our automatic resonance frequency band selection algorithm integrates a resonance frequency band selection criteria with the DWPT. The frequency selection criteria is a two step criteria based on the signal energy using RMS and spikeness using kurtosis. The following steps describe the procedure using the above techniques to implement an automatic impulsive resonance packet selection as illustrated in Figure 5-5.

(1) Wavelet packet transformation of signals: The vibration signal measured from an accelerometer is decomposed by DWPT. This step results in a complete wavelet packet tree and the wavelet coefficients for each packet are denoted as $w_{i,j}$.

(2) Screening with kurtosis: For each of the packets, use Equation 3-13 to calculate the kurtosis $k_{i,j}$ of $w_{i,j}$. All the packets with $k_{i,j}$ greater than a threshold value δ are passed to step (3). We use kurtosis as a screening tool because it is more sensitive to impulsive signals. We might have several packets with high kurtosis values because the impulsive signal will excite all the resonance frequencies. It was found $\delta=4$ is a suitable value for the early bearing case [123].

(3) Packet selection from RMS change: For each of the packets from step (2), we calculate the RMS changes with respect to the previous measurement or the measurement of a normal condition. We then claim the packet with maximum RMS change as the best approximate packet for HFRT analysis.

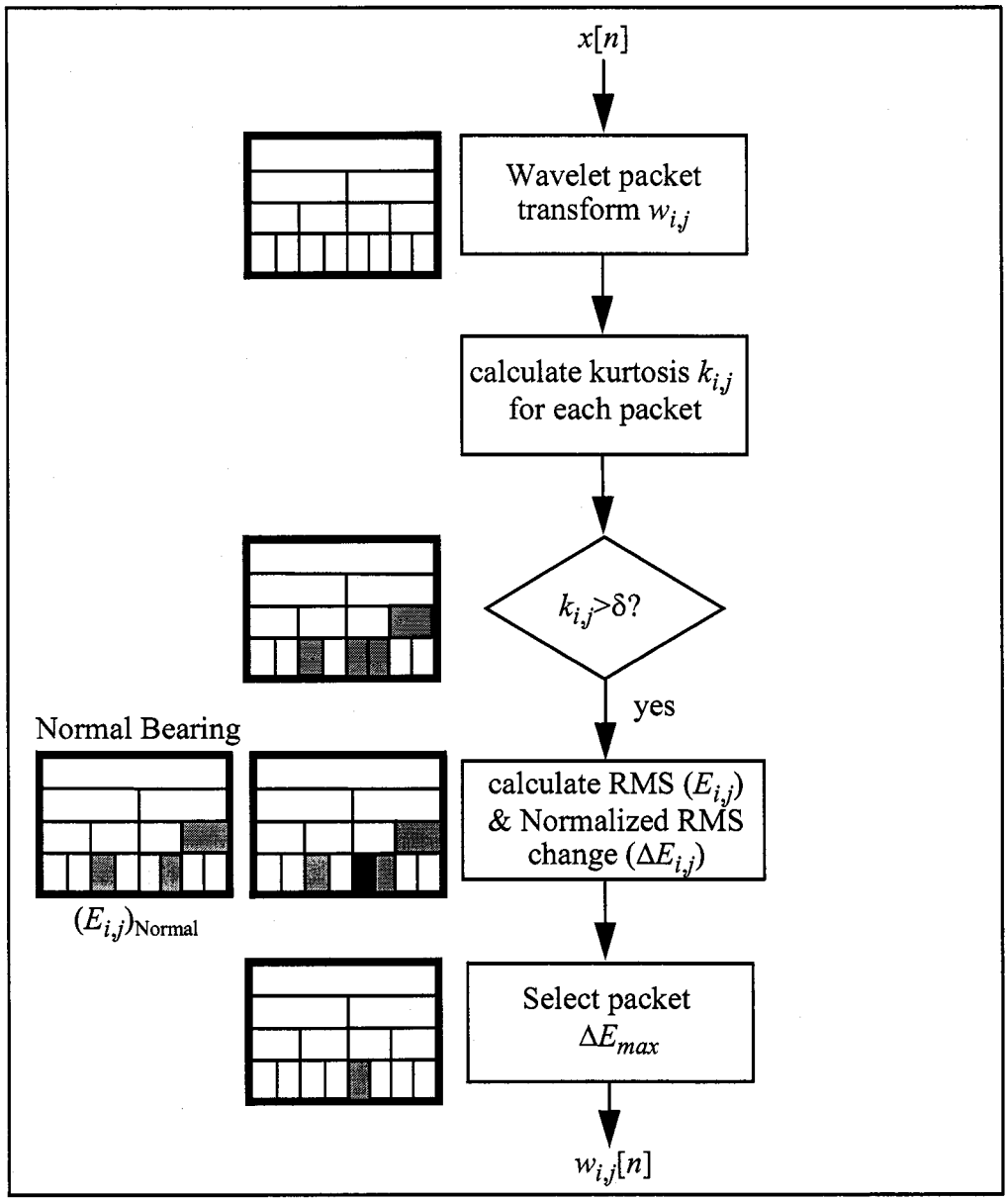


Figure 5-5. The procedure of automatic impulsive resonance packet selection.

5.4.2 DWPT based HFRT

Figure 5-6 is the process of a DWPT based HFRT. Once we can find the packet(s) with the optimal impulse to noise ratio, the wavelet coefficients in the packets can be treated

as a filtered narrow-band signal. We can further reduce the noise level by using denoising process as presented in Section 4.5. Since the denoising process removes of all the small wavelet coefficients, the reconstructed result is a smoothed version of the filtered signal [34]. Since the wavelet coefficients $w_{i,j}$ and the denoised wavelet coefficients $\hat{w}_{i,j}$ are real numbers, we can calculate the envelope of the estimated wavelet waveform $y_w[n]$ using a Hilbert transform, which may be expressed as

$$y_w[n] = \sqrt{\hat{x}^2 + \tilde{\hat{x}}^2}\,. \tag{5-27}$$

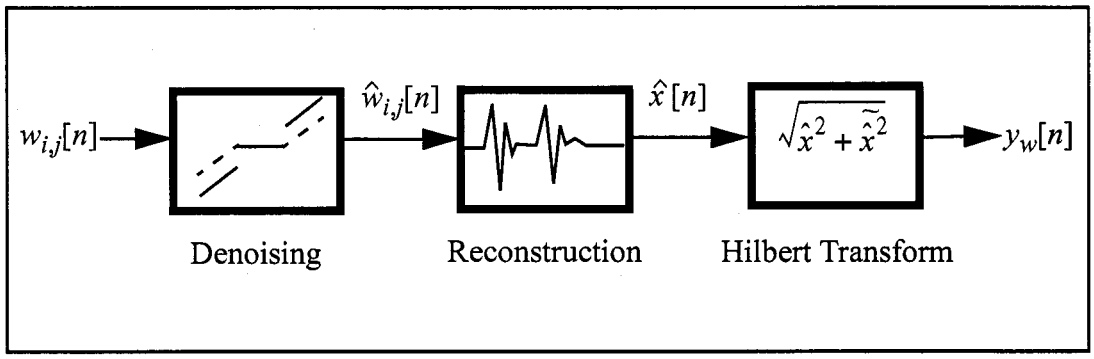


Figure 5-6. DWPT based HFRT.

Note that this DWPT based HFRT can be extended to MODWPT when a shift invariant result is desired.

5.5 Summary

In this chapter, we described 3 HFRT alternatives other than the traditional Hilbert transform based method. Figure 5-7 summarized the process of these alternatives. We define these alternatives as 1) Hilbert transform based HFRT, 2) linear detector based HFRT, 3) Quadratic detector based HFRT and 4) DWPT based HFRT. We also proposed an automation resonance frequency selection process using DWPT. The perfor-

mance of each techniques will be demonstrated in the next chapters using analytic and real bearing defect data.

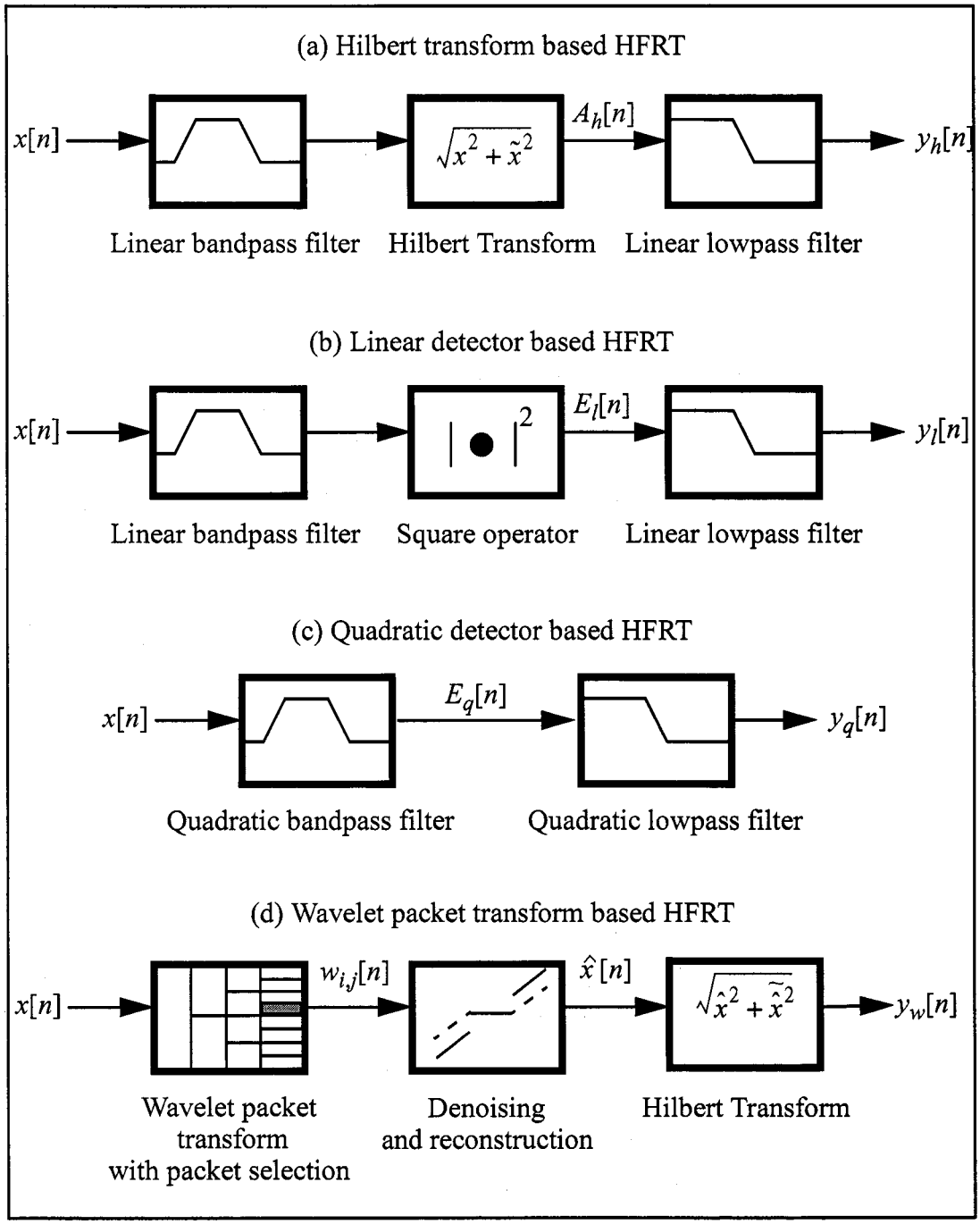


Figure 5-7. The process of (a) Hilbert transform, (b) linear detector, (c) quadratic detector, and (d) wavelet packet transform based HFRT.

Chapter 6 Analysis Results for Synthetic Data

An universally superior algorithm for all bearing defect diagnosis tasks does not currently exist. Some algorithms do very well under some conditions and very poorly under others. This is why it is so important to study a broad range of parameter values as is done in this chapter. Using synthetic data gives complete understanding of the problem domain.

The analysis in this chapter uses the signal model described in Section 6.1 to create synthetic data sets to show the relative strengths and weaknesses of both the traditional methods and each of the HFRT algorithms for detecting the bearing defects described in Chapter 5. This signal model was chosen due to its widespread use and acceptance as a general model for rotating machinery. It allows us to simulate different types of bearing defect vibration. In Section 6.2, both synthetic single defect and multiple defect signals are investigated using various HFRT algorithms. The results of these analytic simulations are summarized in Section 6.3.

6.1 Synthetic Signal Model

In general, the form of signals in real machinery is complicated because real machinery has many vibration sources. In this section, we define a general vibration signal model for simple rotating machinery as the superposition of the vibration responses due to all excitation sources. Typical vibration sources can be classified into 3 categories: harmonic excitations, non-synchronous excitations and noise. Each type of excitation will generate a distinct vibration response of its own.

6.1.1 Generalized signal model

A generalized signal model for simple rotating machinery can be expressed as [105]

$$y(t) = h(t) + n(t) + w(t) \tag{6-1}$$

where $y(t)$ is the response signal measured from the sensor, $h(t)$ represents harmonic signal components which are integer multiples of the fundamental shaft rotating speed, $n(t)$ is the non-synchronous signals from the bearings, and $w(t)$ is noise.

6.1.1.1 Harmonics of rotating frequency

A signal at the shaft rotating frequency will always exist whenever the machine is running. In additional, most of the machinery components generate vibration signal components at frequencies that are synchronous to the fundamental shaft rotating frequency. When they have defects, the response amplitude of these harmonics increases as an indication of abnormal condition (see Table 3-1). These harmonics of rotating frequency are called machine-induced harmonics and their response equation can be expressed as

$$h(t) = \sum_{n=1}^{Nr} A_n \cos(2\pi f_n t + \varphi_n), \tag{6-2}$$

where A_n, f_n, and φ_n are the amplitude, frequency and initial phase of the harmonic components respectively. N_r is the number of harmonic components. Since these components are commensurate with the shaft rotating frequency f_1, frequency f_n can be expressed as,

$$f_n = kf_1 \quad ; \; k \text{ is an integer number}. \tag{6-3}$$

6.1.1.2 Bearing-induced harmonics

There are also non-synchronous vibrations, such as gear and bearing defect frequencies. In order to simplify our analysis, we only consider the non-synchronous vibration from the bearings. Defective bearings produce impulsive signals that are at non-integer multiples of the shaft rotating frequency as described in Section 3.2. These impulses excite

resonances in the bearing and the machine structure that can be approximated as an amplitude modulated response,

$$n(t) = x(t) \sum_{j=1}^{N_j} B_j e^{-\alpha_j t} \cos(2\pi f_j t + \varphi_j), \tag{6-4}$$

where B_j, α_j, f_j, and φ_j are the amplitude, damping factor, resonance frequency and initial phase of resonant components respectively. N_j is the number of resonance components. The modulating component $x(t)$ for a single defect is given as Equation 3-2, i.e.,

$$x(t) = d_0(t) \sum_{n=1}^{\infty} \delta\left(t - \frac{n}{f_b} - \tau\right), \tag{6-5}$$

where $d_o(t)$ is the amplitude of the impulse train, and τ is the initial delay. The repetition frequency, f_b, can be any one of the fundamental defect frequencies depending on where the defect occurs. Note that $d_o(t)$ is a time dependent variable due to the non-uniform loading between the roller and raceway. When a defect is located at the inner race or on one of the rolling elements, $d_o(t)$ is also a function of the shaft rotating frequency because the distance between the defect and the sensor changes with the rotating angle.

If more than one defect exists and they are linearly independent, the modulating component is a summation of several impulse trains with different delays and repetition frequencies, i.e.,

$$x_m(t) = \sum_{p=1}^{m} x_p(t) = \sum_{p=1}^{m} d_p(t) \sum_{n=1}^{\infty} \delta\left(t - \frac{n}{f_{b,p}} - \tau_p\right), \tag{6-6}$$

where m is the number of defects and $f_{b,p}$ is the bearing defect frequency for pth defect.

6.1.1.3 Noise

Causes of noise include environmental disturbances and measurement distortions. Generally, noise is modelled as a Gaussian white noise with zero mean.

6.1.2 Signal to noise ratio

In this Chapter, the signal to noise ratio (SNR) for a given signal is defined as,

$$\text{SNR} = \frac{|w(t)|^2}{|y(t) - w(t)|^2}, \tag{6-7}$$

This is the ratio of the expected bearing-induced harmonics energy to the total noise energy.

6.2 Simulation Results

Based on the generalized signal model (Equation 6-1), we construct synthetic data sets to simulate vibration signals generated by a simple rotating machinery. Depending on the bearing condition, synthetic data of a normal bearing as well as bearings with single and multiple defect are constructed with specific desired properties.

Consider a misaligned rotating machinery running at 1800 rpm (f_0=30Hz) in the presence of a zero mean white noise. With misalignment, an increase of peak amplitudes (A_1 and A_2) at 1X and 2X of the rotating frequency is a common symptom which in turn causes the bearing to fail due to fatigue. For a detective bearing, we consider only one resonance in this simulation because the HFRT will often perform on a single resonance region where the bearing fault signal has increased most above the noise using a pre-filtering process (see Section 5.2). We assume an outer race defect will excite a resonance frequency ar 4.5kHz ($f_{r=1}$) with a constant impulse amplitude (=d_o) and an inner race defect will excite a resonance frequency at 3.4kHz ($f_{r=2}$) with a time variant impulse amplitude (= $d_o|\cos(2\pi f_1 t)|$). The viscous damping factors for an outer race defect and inner race defect are equal to 3000 (α_1) and 3500 (α_2) respectively. The ball pass outer race frequency (f_{bpof}) is 105Hz and the ball pass inner race frequency inner race (f_{bpif}) is 165Hz. The sampling frequency for all the simulations is 20kHz. Based on the above assumptions, our synthetic signal model can be simplified as

$$
\begin{aligned}
y(t) \approx & A_1 \cos(2\pi f_1 t) + A_2 \cos(2\pi f_2 t) \\
& + d_1(t) e^{-\alpha_1 t} \cos(2\pi f_{r=1} t + \varphi_1) \sum_{n=1}^{\infty} \delta\left(t - \frac{n}{f_{b1}} - \tau_1\right) \\
& + d_2(t) e^{-\alpha_2 t} \cos(2\pi f_{r=2} t + \varphi_2) \sum_{n=1}^{\infty} \delta\left(t - \frac{n}{f_{b2}} - \tau_2\right) + w_G(t)
\end{aligned}
\quad , \qquad (6\text{-}8)
$$

where w_G is a Gaussian white noise signal with zero mean.

6.2.1 Simulation of bearing without fault

For a bearing in good condition, no impulse series are present in the signal. The only vibration sources are the shaft rotating frequency and noise. In this case, the white noise is large in comparison to the vibrational signal. Figure 6-1 is the simulated signal and spectrum generated with peak amplitude of shaft fundamental frequency equals 0.1 (A_1) and SNR=0.1. The resulting power spectrum has only one peak at the rotating frequency as shown in Figure 6-1.

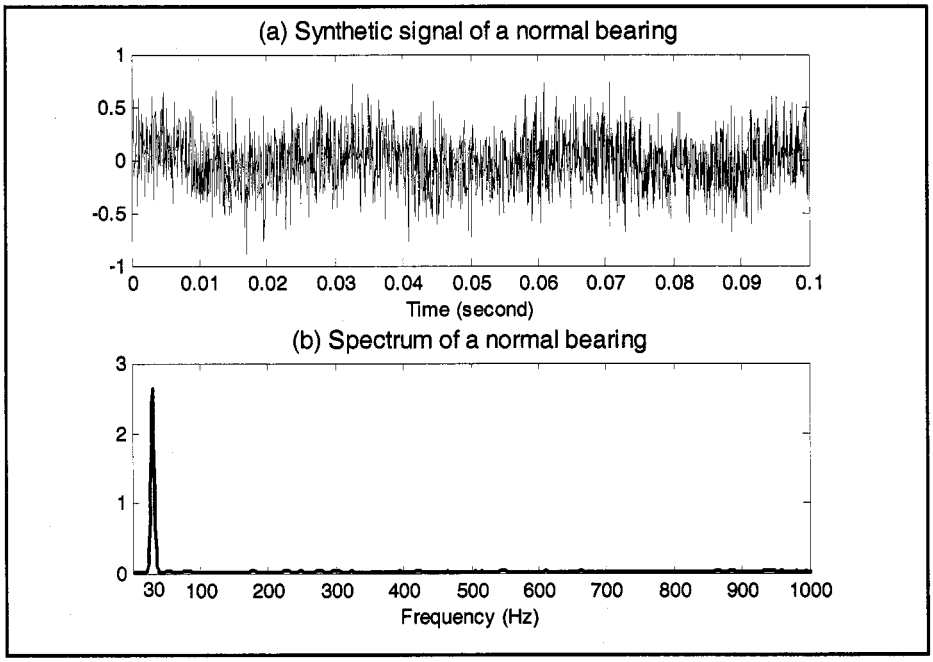


Figure 6-1. Simulated signal of a normal bearing and its spectrum.

6.2.2 Single defect

Assume a shaft is slightly misaligned with a defective ball bearing operating in the presence of a background white noise. The defective ball bearing has a single spall on the outer race. The vibration characteristics of misalignment include an increase of vibration amplitude at 1X rotating frequency and a higher than normal 2X harmonic amplitude. We assume the amplitude of 1X and 2X harmonics are $0.2(A_1)$ and $0.1(A_2)$ respectively. The individual signal components are shown in Figure 6-2 along with their sum as the composite signal at SNR=1. The composite signal is a noised impulse train undulating on a low frequency wave. The impulsive component is the consequence of a single-point defect on the outer race with a constant unit impulse amplitude (d_0=1). The goal of our analysis is to detect the impulsive component from the composite signal and to diagnose the type of bearing defect.

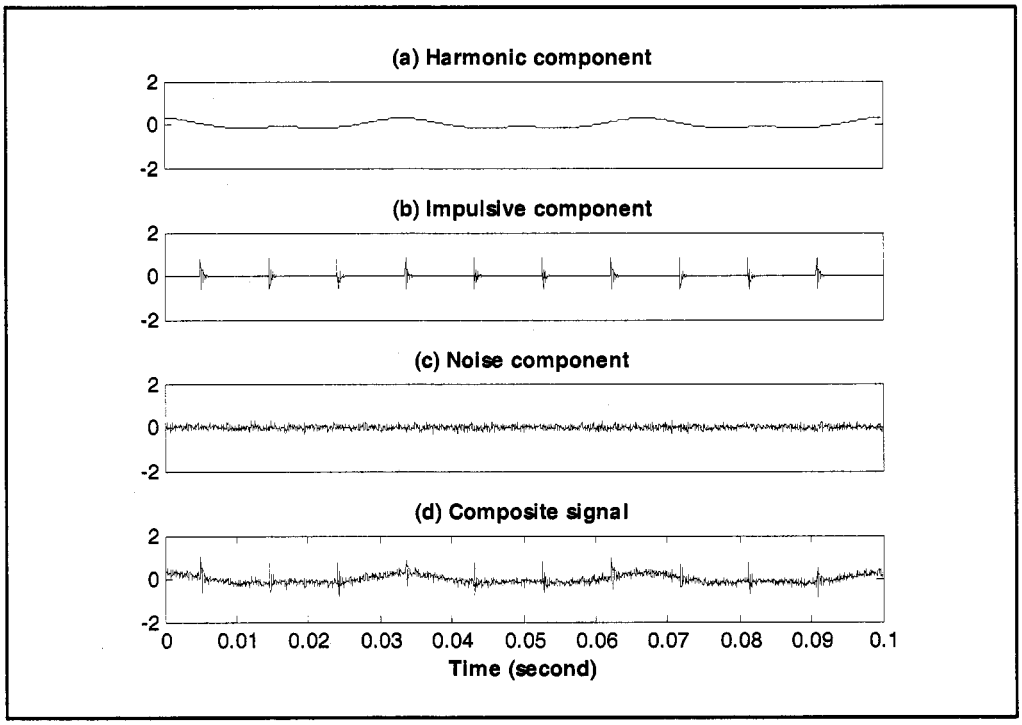


Figure 6-2. Constructing of the synthetic signal with single defect at outer race.

6.2.2.1 Result of MODWPT analysis

A level 4 MODWPT using 16th order Daubechies wavelet is performed to extract impulsive components from the synthetic signal. The synthetic signal and the reconstructed MODWPT signals at level 4 are shown in Figure 6-3. Top row is the synthetic signal and the lower block is the reconstructed MODWPT at level 4 with 16 wavelet packets. The numbers at the left of the reconstructed signals are the corresponding frequency range of each wavelet packet. Their kurtosis values are also listed at the right of the reconstructed signals.

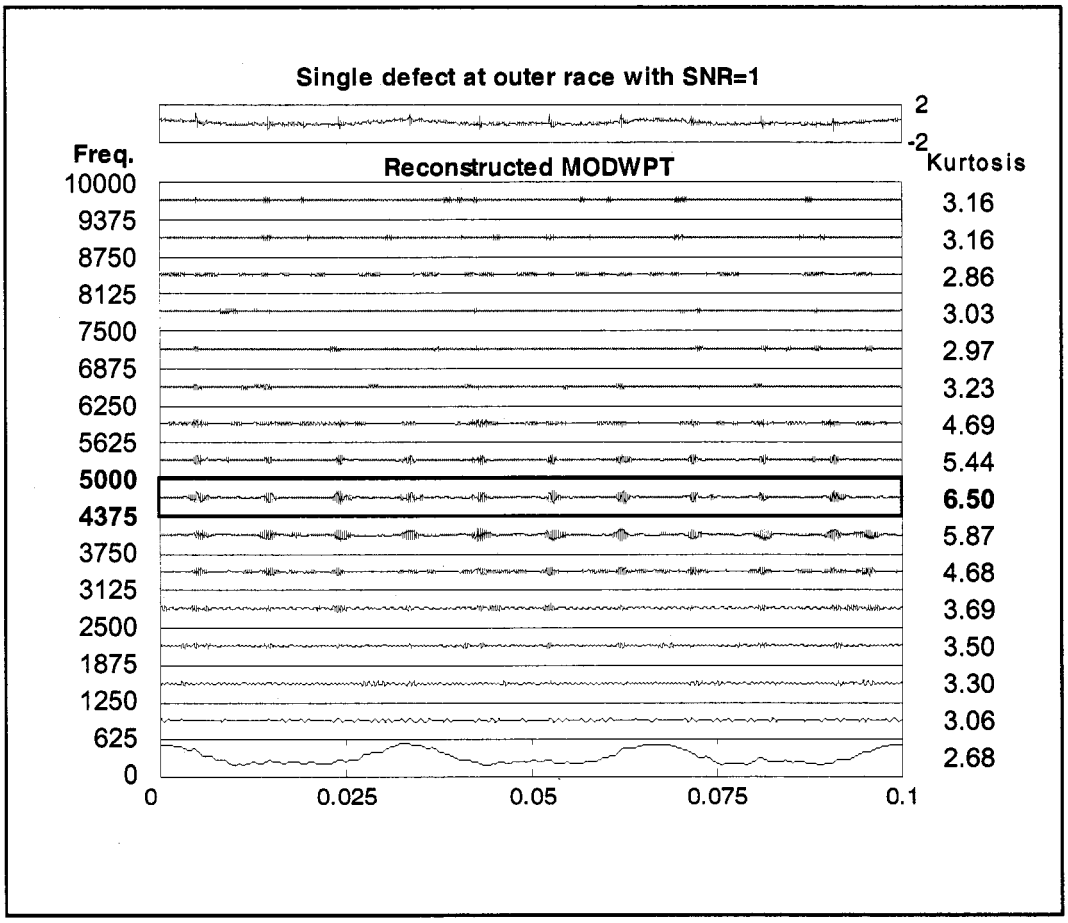


Figure 6-3. The synthetic signal of a single defect bearing on the outer race and its reconstructed MODWPT at level 4 using 16th order Daubechies wavelet.

The first packet at the bottom contains the lowest frequency components reconstructed from the approximation coefficients. The oscillating waveform at the lowest packet contains the energy of rotating frequency and its harmonics. An amplitude increase at this packet implies a possible of misalignment or unbalance fault.

Several packets have a noticeable pulse train that represents the impulsive component generated by bearing defect. To determine which packet is the optimal packet for HFRT analysis, we calculate the kurtosis for each packet. The highest kurtosis is 6.5 at packet 8. The 8th packet contains the signal content of the frequency interval of [4375 5000]Hz. Since the dominant resonant frequency is 4.5kHz, kurtosis of a band filtered signal is a sensitive feature for selecting the optimal frequency interval for extracting the desired impulsive signal from MODWPT.

6.2.2.2 Results of HFRT algorithms

In this section, we compare the performance of MODWPT based HFRT with the traditional Hilbert transform based (HT), the linear detector based (LD) and the quadratic detector based (QD) algorithms as described in Section 5.5.

We use two features to investigate the performance of these HFRT algorithms. The first feature is the time localization of pulses by tracking the rise time and the width of pulses from the smoothed or denoised envelope. The second feature is the effectiveness of demodulating the bearing defect frequency. Comparing the amplitudes of the spectrum at the bearing defect frequency and its harmonics helps us to diagnose the bearing defect type.

The previous analysis demonstrated that kurtosis is a useful indicator to select the best frequency band for extracting fault related impact trains from MODWPT. All of the HFRT algorithms, except the MODWPT based algorithm, are designed with same sets of coefficients. The coefficients are obtained from a 16th order bandpass FIR filter with a 4687.5Hz center frequency and 625Hz bandwidth. The number of nonzero bands (L) in the quadratic detector is 11.

Figure 6-4 compares the instantaneous amplitude results of the different HFRT algorithms before the smooth process. Instantaneous amplitudes were normalized by their maximum amplitudes in order to compare the effectiveness of extracting the impulse train and the noise suppression. Figure 6-4 shows that all of the HFRT algorithms were able to extract fault related impulse peaks for a SNR=1. The background noise level of the HT, LD and QD based algorithms are larger than the MODWPT based algorithms which can cause these results to be misleading. However, The leading edges of the pulses obtained from MODWPT are smeared compared with the other methods. The degree of localization of the MODWPT can be improved by using a smaller width for the wavelet filter [34].

To demodulate the bearing defect frequency from the instantaneous signals, a smooth process is applied to the results of Figure 6-4. For the smooth process, the FIR lowpass filter had a length of 17 coefficients with a cut-off frequency of 1000Hz. The MODWPT used a denoising process to reduce the noise level with a soft threshold level of 0.08262 (see Section 4.5). The smoothed results are shown in Figure 6-5. The signal envelopes of HT and LD based algorithm are similar with a high background noise ranging between 20% to 50% of the peak amplitudes. The QD based algorithm provides a much better time resolution after the smooth processing using a cascaded quadratic detector. The denoising process in the MODWPT algorithm is also an effective technique to suppress the background noise level by eliminating small wavelet coefficients. The QD based algorithm gives a shorter rise time at each pulse.

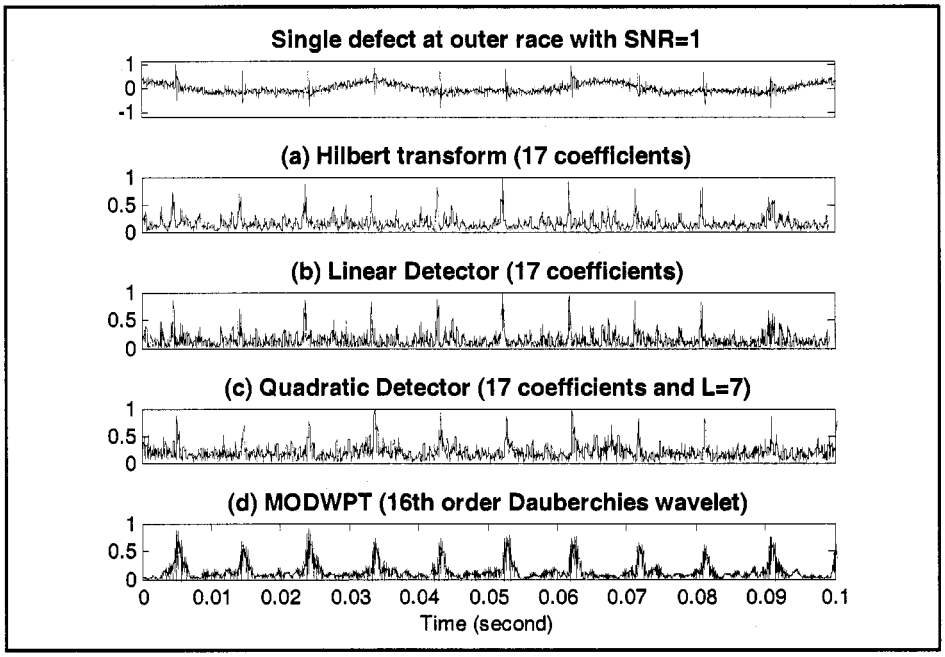


Figure 6-4. Single defect at outer race with SNR=1 and its normalized amplitude from (a) Hilbert transform, (b) linear detector, (c) quadratic detector and (d) MODWPT based algorithm.

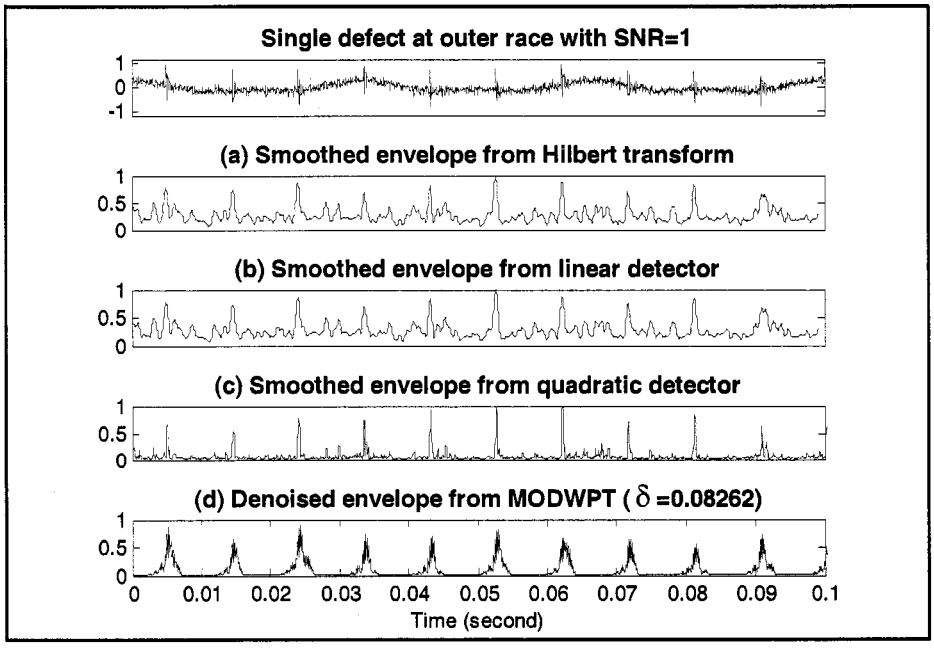


Figure 6-5. Smoothed results of Figure 6-4.

Figure 6-6 shows the normalized cascade spectra of the smoothed envelope and original synthetic signal. The power spectrum was normalized by its total power in order to compare its relative strength. The normalized spectrum of the original signal has peaks at 1X and 2X rotating frequency which results from the misalignment fault. The outer race defect frequency (BPOF) component does not exist in the original spectrum because the strength of BPOF is much smaller than the 1X and 2X component. In order to enhance the BPOF, we apply HFRT to demodulate the BPOF from the high frequency region. Figure 6-6 shows that the demodulation process using the HFRT algorithms successfully demodulated the BPOF and its harmonics. Comparing the normalized spectra, the MODWPT based HFRT gives the highest amplitude at 1X and 2X BPOF. This result implies that the wavelet transform, in which small waves grow and decay in a limited time period, is a better tool to capture localized events than methods based on sinusoidal functions.

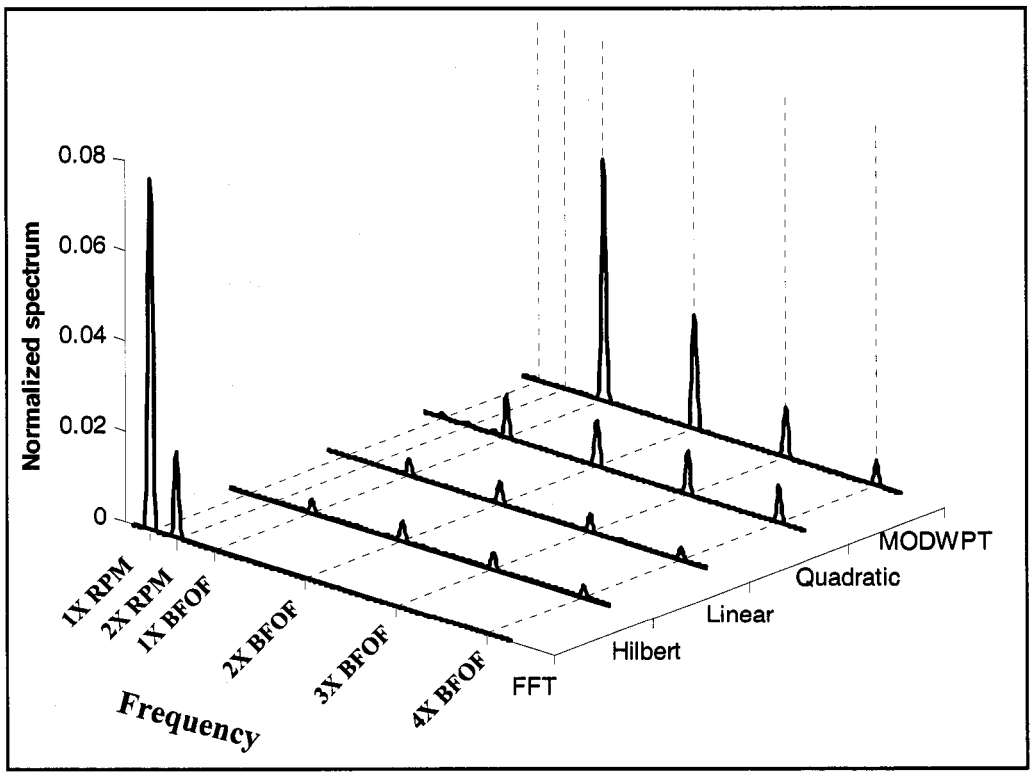


Figure 6-6. Cascade envelope spectra of the synthetic signal for a single defect at the outer race with SNR=1.

6.2.2.3 Results of various SNR level

To simulate the early stage of bearing defect, we use the same model as above with a lower SNR level. Figure 6-7 shows the HFRT analysis results at SNR level of 0.5. As shown in Figure 6-7, the periodic pulses are indistinguishable from both the HT based and LD based algorithms. The quadratic detector was able to extract the underlying pulses with a very sharp rise time and few undesirable spikes. The denoised MODWPT using a 16th order Daubechies wavelet outperforms all the other methods in terms of background noise suppression. The drawback of the MODWPT method is the smearing at each pulse.

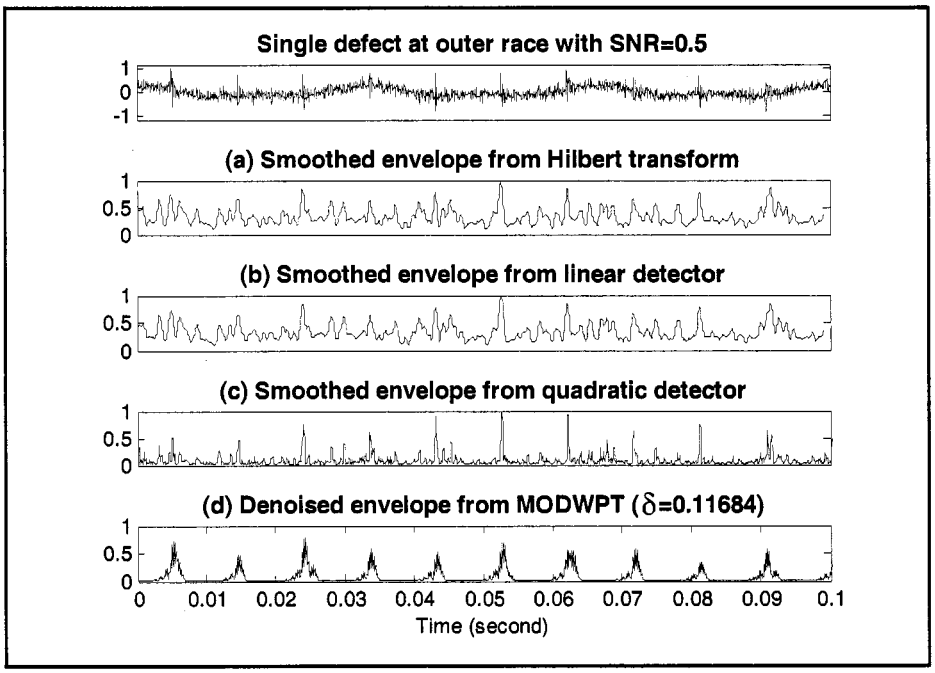


Figure 6-7. Single defect at outer race with SNR=0.5 and its normalized amplitude from (a) Hilbert transform, (b) linear detector, (c) quadratic detector and (d) MODWPT based algorithm.

If we further reduce the SNR to 0.1 for simulating a very early bearing defect, the impulsive component cannot be identified from the smoothed or denoised amplitude envelopes from HFRT algorithms as illustrated in Figure 6-8. However, the envelope

spectra from the QD and MODWPT based methods are still able to capture the spectrum component at the outer race defect frequency and its harmonics as shown in the cascade spectra of Figure 6-9. Therefore, the MODWPT is the best algorithm to detect the bearing defect frequency under high noise level compared with other HFRT algorithms. In the filter-based methods, the frequency components outside the cut-off frequency are set to zero which may smooth some impulses. On the other hand, the wavelet-based method using denoising to set some small wavelet transform coefficients to zero can retain small impulses in signals because they are represented as large wavelet transform coefficients rather than noise. This is why a wavelet-based method is better for extracting impulses under noise than the filter-based methods.

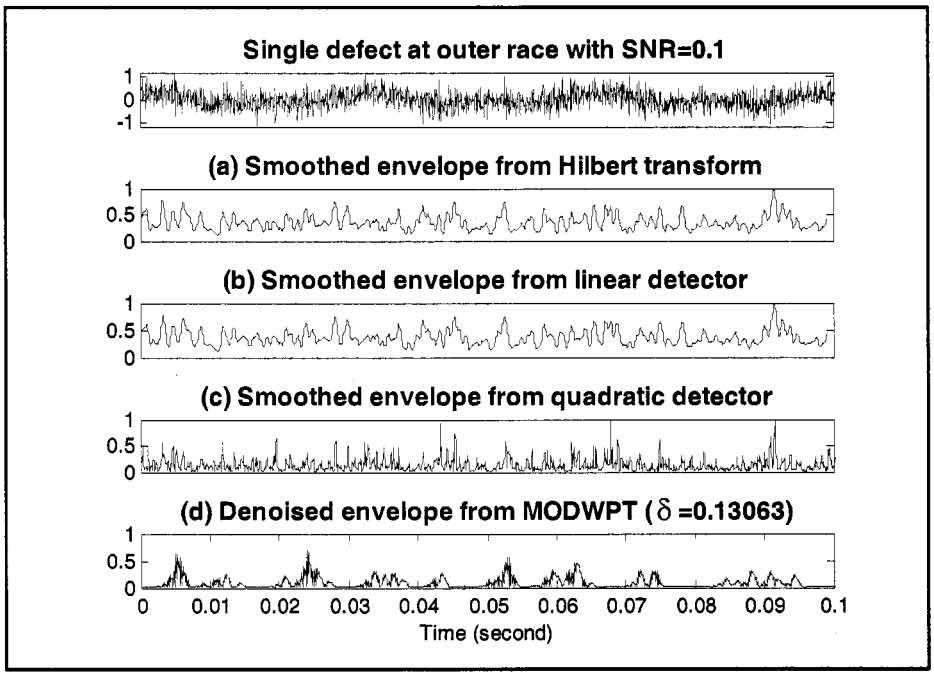


Figure 6-8. Single defect at outer race with SNR=0.1 and its normalized amplitude from (a) Hilbert transform, (b) linear detector, (c) quadratic detector and (d) MODWPT based algorithm.

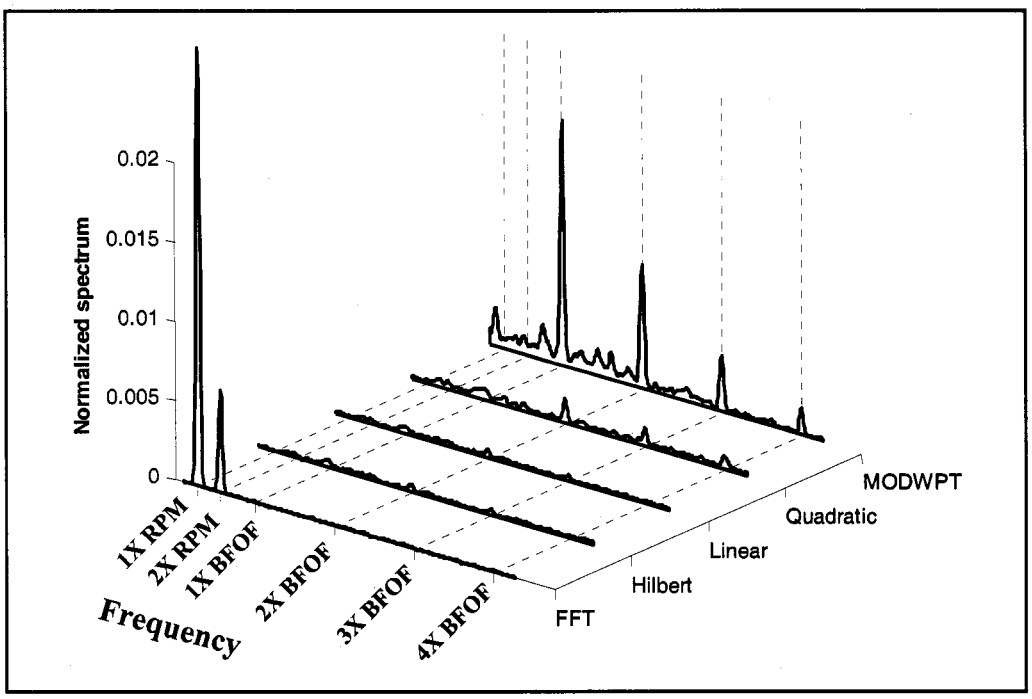


Figure 6-9. Cascade spectra of Figure 6-8 with SNR=0.1.

6.2.3 Multiple defects

Once a defect is generated in a rolling element bearing, flakes falling from the surface will cause additional defects easily. In this section, we investigate a bearing with two defects at both the same bearing component and different bearing components.

6.2.3.1 Two defects at same component

Figure 6-10 shows a schematic diagram of a ball rolling over two point defects at the outer race and the resulting waveform that simulates these two defects. The amplitudes of the impulse trains will depend on the severity of the defects and the phase angles will depend on the position of the defects. In this example, the defect at point A is larger than the defect at point B; the larger defect at point A generates a higher peak than a smaller defect at point B.

The number of defects and their size are closely related to the bearing condition. In order to determine the number of defects and their severities, it is important to find an optimal HFRT algorithm which can extract both impulse trains and track their peak amplitudes.

A simulated signal from a misaligned ball bearing with two outer race defects is generated using Equation 6-8 (see Figure 6-11) with SNR=1. Due to the separation of these two defects, there is a $0.0012(\tau_2)$ second delay between two impulse trains with the same repetition frequency of 105Hz. The peak amplitudes of these two spalls are set to be $1(d_1)$ and $0.75(d_2)$ respectively. Assume that both defects are excited at the same resonant frequency (=4500Hz) with the same viscous damping factor (=3000). The MODWPT uses a 4th order Daubechies wavelet in order to increase the degree of time localization. The rest of the parameters are the same as for the single defect case.

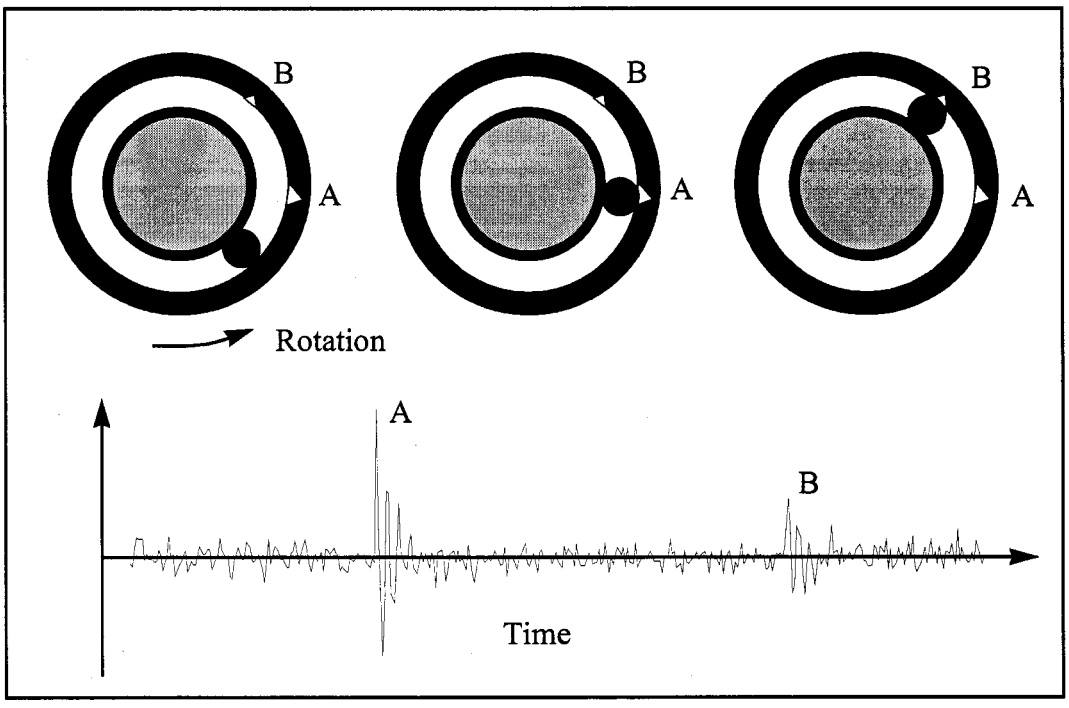


Figure 6-10. Simulated signal of a ball bearing with two defects on the outer race.

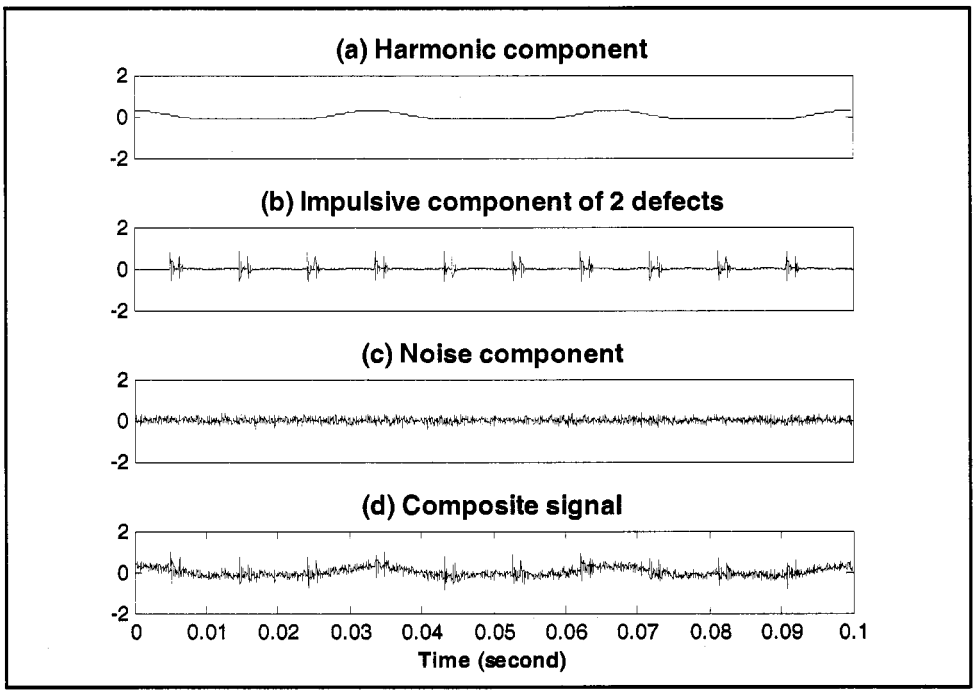


Figure 6-11. Constructing the synthetic signal with two defects at the outer race.

Figure 6-12 shows the results of HFRT algorithms for 2 point defects on the outer race with SNR=1. Vertical dashed lines denote time instances when impacts are generated. In general, all the HFRT algorithms can detect the peaks and the time instances of the impulse trains. However, it is difficult to identify whether the peak is generated by a pulse train or noise from the HT and LD based results because their background noise level is high. The MODWPT method using soft denoising can suppress almost all background noise with a threshold value of 0.1022. However, the leading edge of each pulse is smeared using the MODWPT method. The QD based method with 11 nonzero bands (L) provides a good noise suppression with less smearing in time. Figure 6-13 compares the envelope spectra of the various HFRT algorithms. Referring to the spectrum peaks at bearing defect frequency (BPOF) and their harmonics, the MODWPT is the best algorithm for the two defects at BPOF.

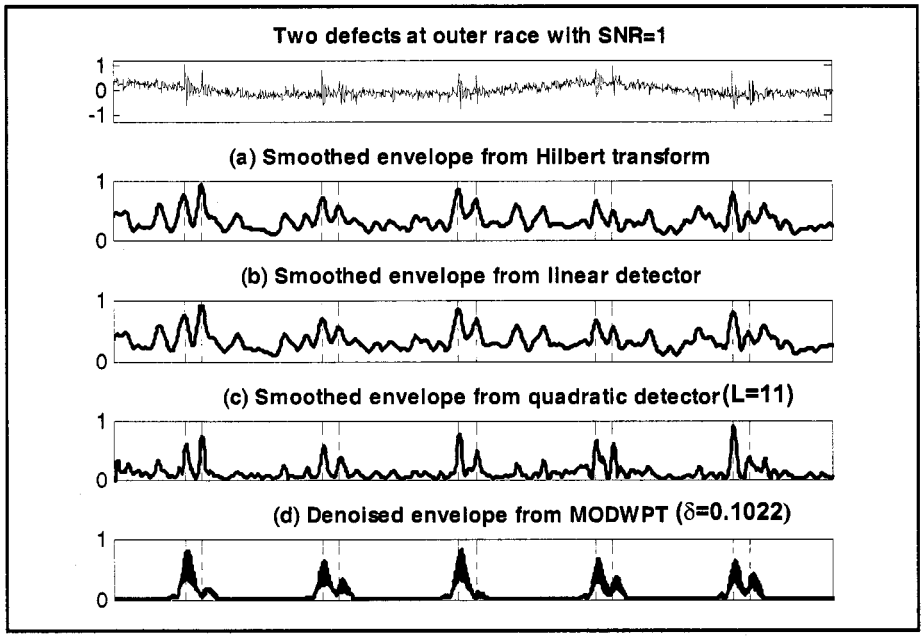


Figure 6-12. Two defects at outer race with SNR=1 and its normalized amplitude from (a) Hilbert transform, (b) linear detector, (c) quadratic detector and (d) MODWPT algorithm.

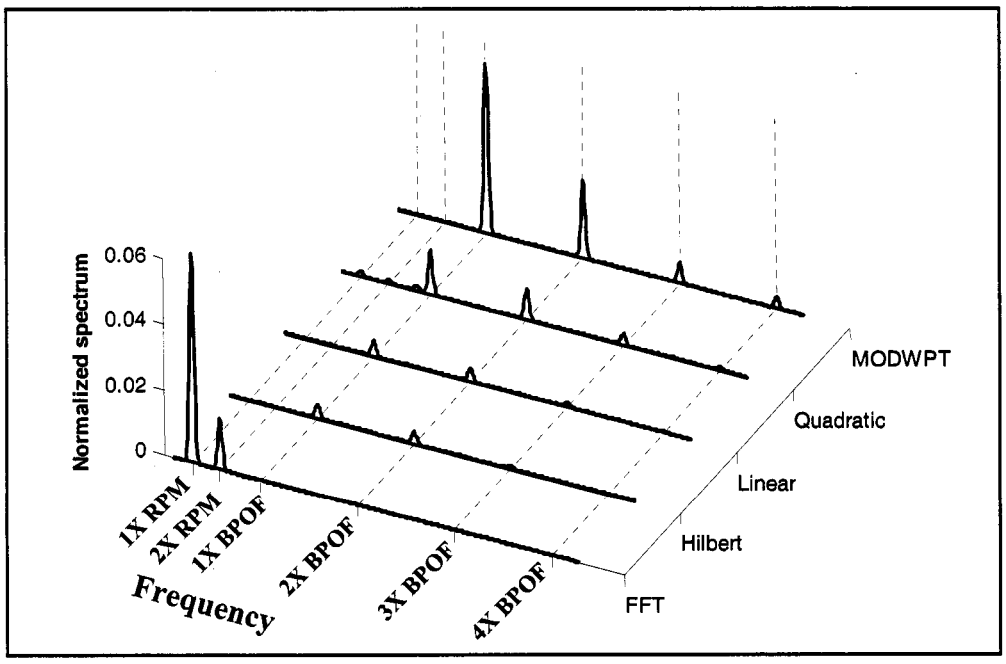


Figure 6-13. Cascade envelope spectra of synthetic signal of two defects at bearing outer race with SNR=1.

6.2.3.2 Two defects at different components

It is also possible to have multiple defects at different components. Figure 6-14 is an example of two point defects in different bearing components. The first impulse train is caused by the spall at the outer race with constant amplitude. The second impulse train is caused by the spall at the inner race for a pulse amplitude varied with the shaft rotating frequency. The time delay between the two impulse trains is 0.0012 second. Since the transmission loss from an inner race defect to the sensor is higher than from an outer race defect to the sensor, the viscous damping factor for the inner race defect (α_1=3500) is higher than that of the outer race (α_2=3000). The resonant frequencies generated by the inner race and outer race are assumed to be 4.5kHz and 3.4kHz respectively. The synthetic signal is a combination of two impulse trains with background noise level of SNR=1. Note that the time instance between the two pulse peaks varies because their repetition frequencies are different.

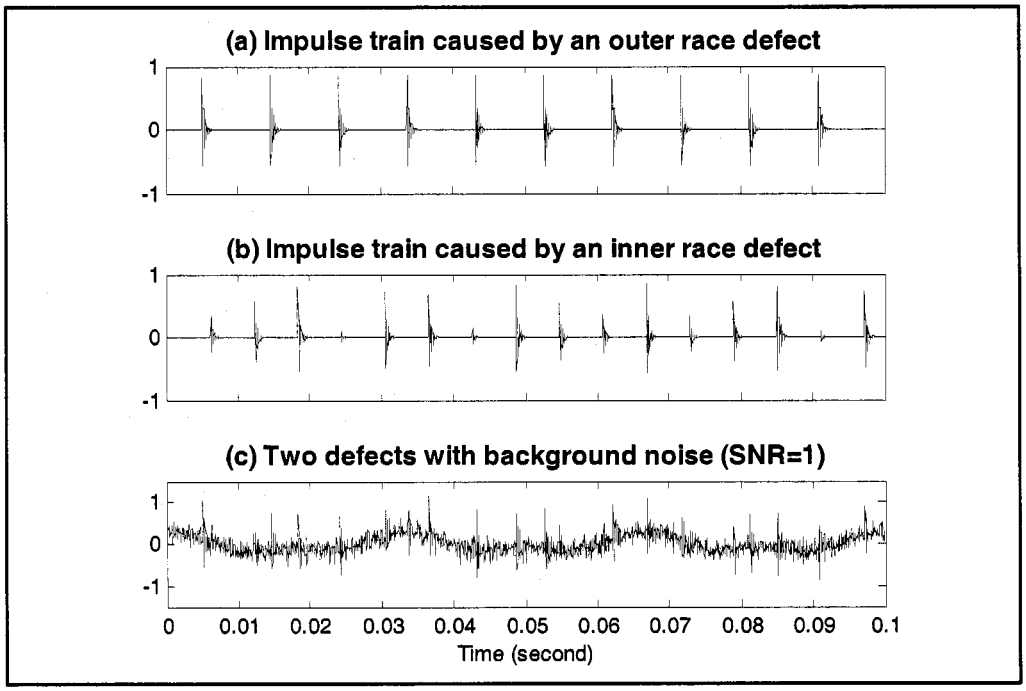


Figure 6-14. Synthetic signal of two defects at different components.

Since there are two resonant frequencies, we need to determine the best resonant frequency band before the HFRT analysis. Figure 6-15 is the reconstructed MODWPT at level 2 using a 16th order Dauberchies wavelet. The highest kurtosis is at the packet with frequency interval of [2500 5000] Hz. This frequency range covers the resonant frequencies of both the inner race defect and the outer race defect.

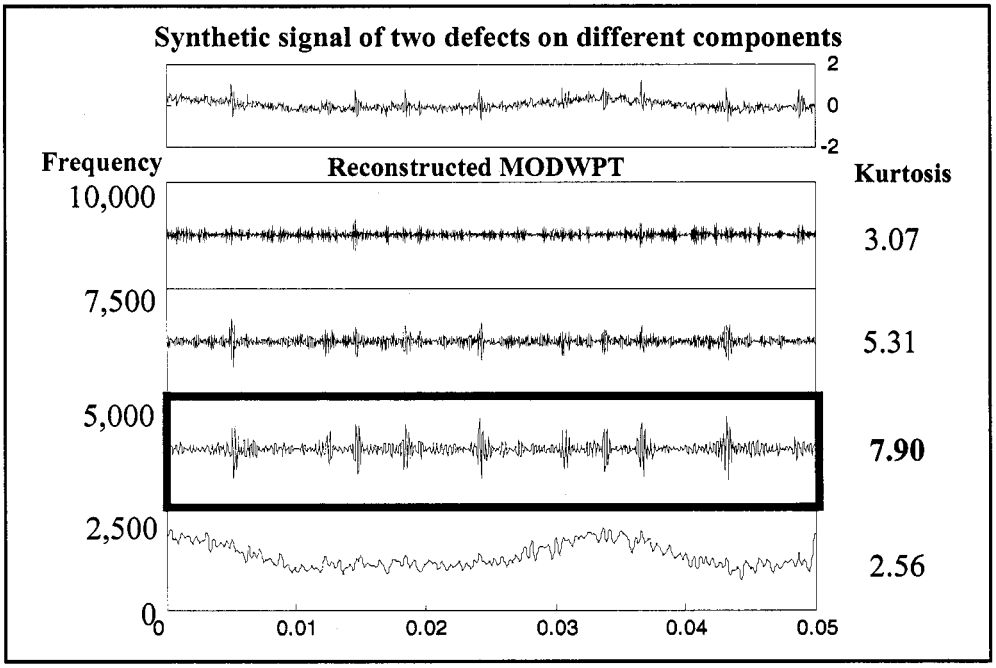


Figure 6-15. The synthetic signal of a bearing with two defects and its reconstructed MODWPT at level 2 using a 16th order Daubechies wavelet.

The HFRT analysis results are shown in Figure 6-16 and Figure 6-17. Figure 6-16 shows the results of the smoothed envelope. The vertical dashed lines denote time instances when impacts are generated. Both the bandpass filter and smooth filter use 17 coefficients. The nonzero bands for cascaded QD are 7 and 1. The MODWPT uses a 16th order Dauberchies wavelet. Again, both the results from HT and LD methods based on conventional FIR filters fail to extract pulses and suffered from the noise interference. The QD based method gives a better result compared with HT based and LD based methods. However, some of the pulse peaks are reduced or masked by back-

ground noise. The MODWPT gives better results for detecting pulses in terms of tracking pulse localization and amplitude. When the adjacent pulses are very close to each other, the MODWPT fails to identify both peaks. In addition, the rise time of the MODWPT result is comparable to the result from the QD method.

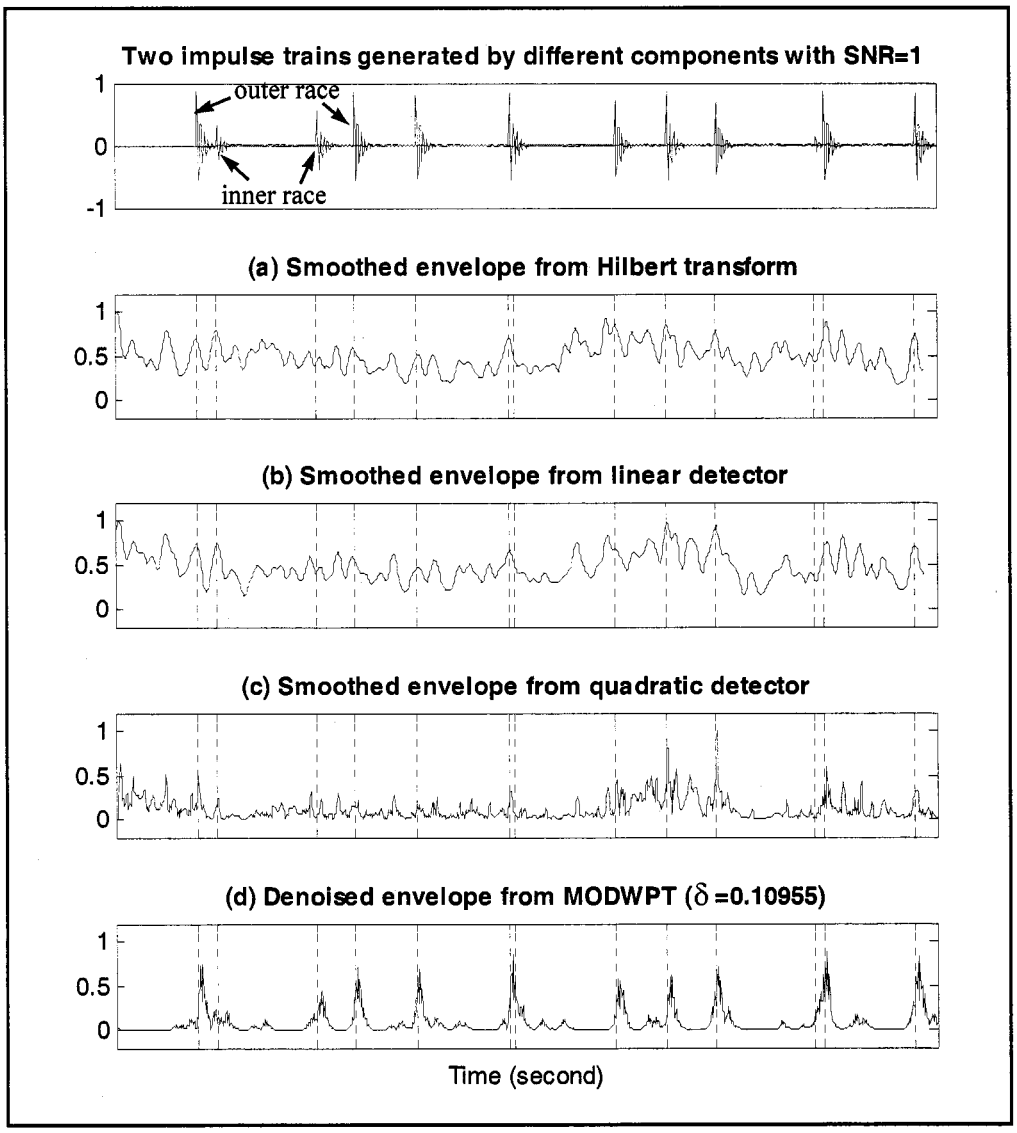


Figure 6-16. HFRT analysis results of a bearing with two defects for a resonant frequency band from 2500Hz to 5000Hz.

The cascade envelope spectra are shown in Figure 6-17. The MODWPT method outperforms all of the filter based methods with distinct peaks at BPIF, BPOF and their sidebands (BPIF+RPM and BPIF-RPM).

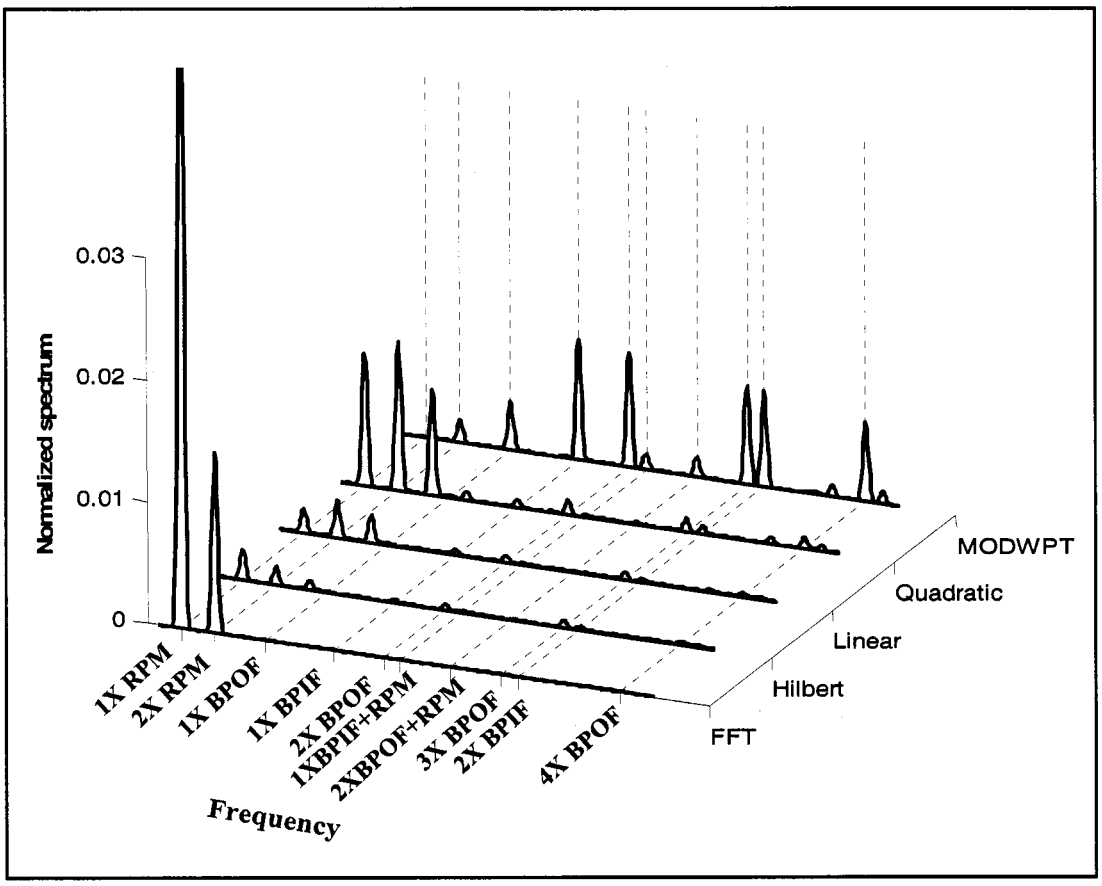


Figure 6-17. Cascade envelope spectra of a synthetic signal with two defects (one at outer race and one at inner race) with SNR=1.

6.3 Summary

Table 6.1 summarized the performance of each HFRT algorithms for bearing defect detection from synthetic data sets. In general, all the HFRT algorithms can extract a single impulse train at high SNR level. Using same number of filter coefficients, the quadratic detector exhibits better noise suppression and pulse detection with less smearing in time. When the SNR level decreases to 0.5, MODWPT using Dauberchies wave-

let outperforms Hilbert transform, linear detector and quadratic methods. This result implies that MODWPT is more sensitive to early bearing defect signatures.

Examples of multiple defects at same or different components show that both the methods of MODWPT and quadratic detector are more effective than Hilbert and linear detector methods. When the adjacent pulses are close to each other, the quadratic detector gives better time resolution than the MODWPT method.

Table 6.1 Performance summary of HFRT algorithms

Performance feature	Hilbert transform	Linear detector	Quadratic detector	MODWPT
Pulse localization tracking	Good	Good	Excellent	Excellent
Pulse amplitude tracking	Good	Good	Fair	Excellent
Peak rise time	Good	Good	Excellent	Fair
Background noise suppression	Fair	Fair	Good	Excellent
Against noise interference	Fair	Fair	Excellent	Excellent
Separation of adjacent pulses	Fair	Fair	Good	Good
Demodulate fault frequency	Fair	Fair	Good	Excellent

We also show that the harmonic analysis can extract the impulse train at low SNR level which can improve the sensitivity of linear and quadratic detector based HFRT alogrithms.

In a real situation, a complex machine will have many components that will generate vibration in addition to noise. We will investigate the effectiveness of these methods using experimentally measured data in the next chapter.

Chapter 7 Analysis Results for Recorded Data

In the previous chapter, we investigated the performance of 4 HFRT algorithms using synthetic data sets generated by a simple bearing defect model. In the real machinery, the form of the signal is more complicated. For example, noise might be colored rather than a Gaussian distributed white noise. In addition, a simple impulse response model will never be able to adequately model all possible signal patterns since the shape of spalls is irregular. To examine how these methods work in real applications, two sets of recorded bearing vibration data are studied in this chapter. The performance of the MODWPT and QD based methods is compared with traditional methods.

Section 7.1 is the analysis results of bearing data experimentally measured with artificially manufactured defects recorded by the Case Western Reserve University bearing data center using a horizontal testing rig. Artificial faults were introduced at the inner and outer races with three different defect sizes. The performance of each of the HFRT techniques are compared with the traditional methods. The application of the proposed automatic impulsive resonance packet selection (see Section 5.4.1) is also illustrated to select the best band for analysis.

Since studies using artificially defected bearings cannot offer insight into the early detection of a naturally propagating crack, we built a bearing test rig to measure the vibration data for bearings operating from an undamaged state through to failure. Section 7.2 describes the design of the experiment and analysis results using both the traditional methods and the proposed HFRT algorithms. The results of these analyses on real bearing data are summarized in Section 7.3.

7.1 Analysis results of artificially defected bearing data

The bearing data center at Case Western Reserve University performed a series of seeded fault bearing testing using a Reliance Electric motor. The test stand consisted of a 2 hp motor (left), a torque transducer/encoder (center), a dynamometer (right), and control electronics (not shown) as shown in Figure 7-1. The SKF (6205-2RS JEM deep groove ball bearing) bearings were used as the test bearings to support the motor shaft at drive end. The specifications of the test bearings, including the bearing geometry and the defect frequencies are tabulated in Table 7-1.

Single point faults were introduced to the SKF test bearings at the inner raceway and outer raceway using electro-discharge machining with fault diameters of 0.007", 0.014", and 0.021". The bearings with these artificially induced faults were reinstalled into the test motor and vibration data was recorded for a motor speed of 1797 rpm. Data was also collected from a normal bearing at the same operating condition. All of the data sets were sampled for 10 seconds at a sampling frequency of 12kHz.

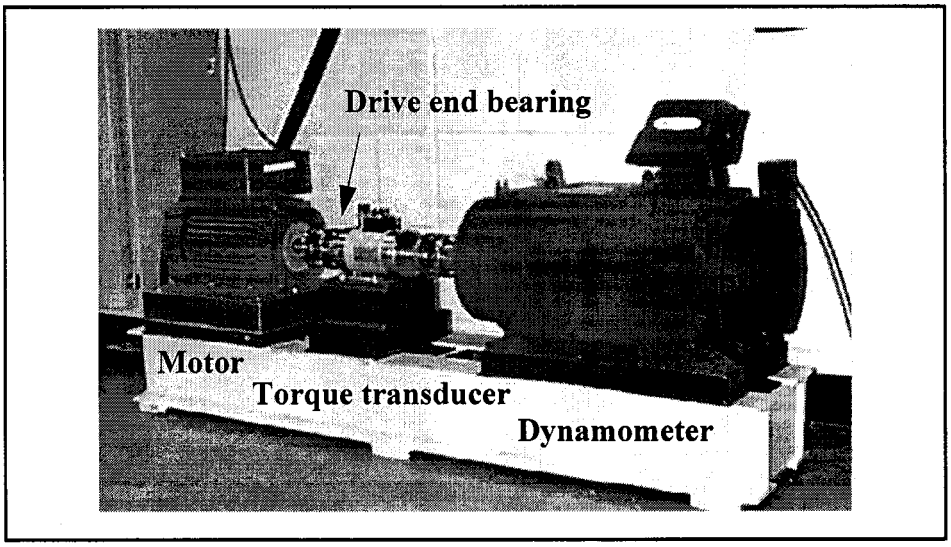


Figure 7-1. Experiment setup of the Case Western Reserve University bearing testing rig.

Table 7-1. Specifications of drive end SKF ball bearing.

Bearing Model Number	6205-2RS JEM SKF deep groove ball bearing
Bore (D_i)	0.9843 inches
Outside diameter (D_o)	2.0472 inches
Width (B)	0.5906 inches
Ball diameter (D)	0.3126 inches
Cage train frequency (f_{ftf})	11.9285 (Hz)
Ball pass outer race frequency (f_{bpof})	107.3648 (Hz)
Ball pass inner race frequency (f_{bpif})	162.1852 (Hz)
Ball spin frequency (f_{bsf})	141.1693 (Hz)

7.1.1 Analysis results of traditional methods

Figure 7-2 and Figure 7-3 illustrate the time history and frequency spectrum of the normal bearing and the defective bearings at the inner race and outer race respectively. Time and frequency domain fault indicators are summarized in Table 7-2 for the different defect sizes. For defective bearings, the impulsive components are easy to discern from the time history and their kurtosis except for the 0.014" outer race defect bearing. The envelope and amplitude of impulsive components are very different for different defect locations and defect sizes. Both the time domain and frequency domain indicators do not increase proportionally with the defect size as can be seen from Table 7-2. In general, the peak amplitude of the outer race defective bearing is larger than the amplitude of the inner race defective bearing for equal size flaws. This is because the resonance due to inner race defect passes through more transitions than an outer race defect; therefore, for some defect sizes, the peak of the enveloped spectrum for the inner race is much smaller than that of the outer race defect. Additionally, the variation of peak amplitudes due to the modulation into and out of the load zone is more obvious for the inner race defects as shown in Figure 7-2.

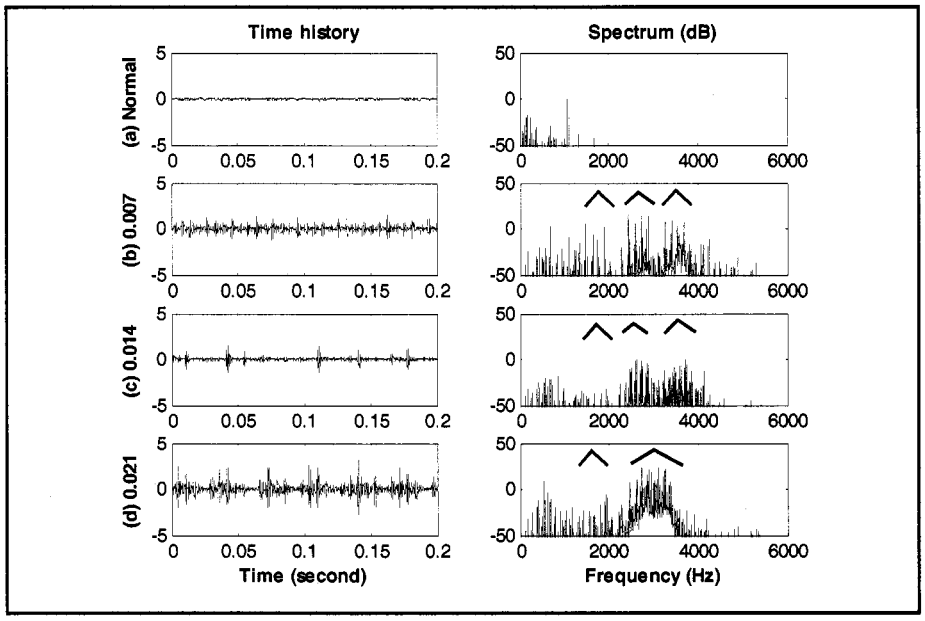


Figure 7-2. Time history and spectrum of the (a) normal, (b) 0.007" inner race defect, (c) 0.014" inner race defect, and (d) 0.021" inner race defect bearing.

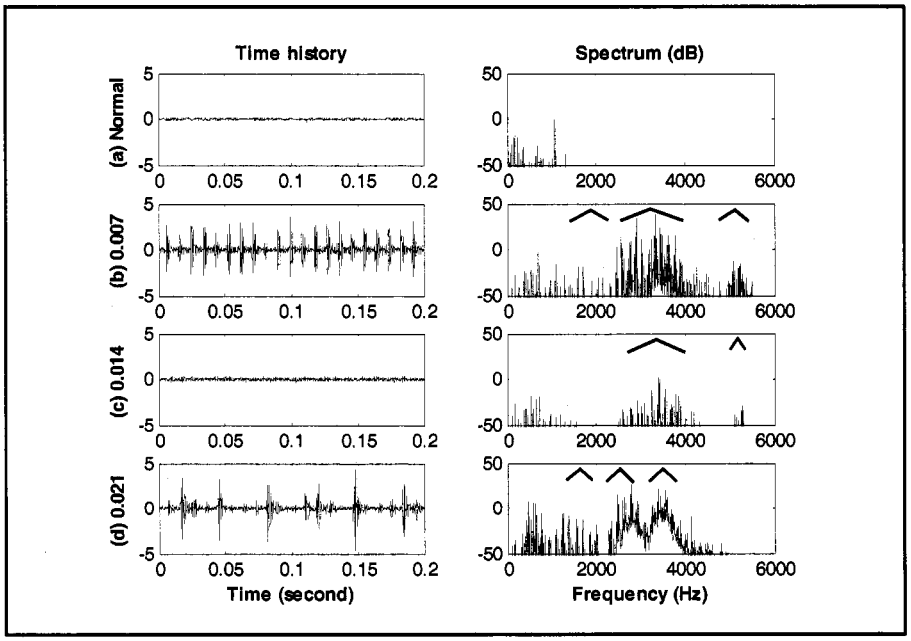


Figure 7-3. Time history and spectrum of the (a) normal, (b) 0.007"outer race defect, (c) 0.014" outer race defect, and (d) 0.021" outer race defect bearing.

Table 7-2. Time and frequency domain indicators for bearing at drive end.

Bearing condition	Normal	Inner race defect			Outer race defect		
		0.007"	0.014"	0.021"	0.007"	0.014"	0.021"
Peak amplitude	0.31	1.74	2.01	3.79	3.63	0.55	6.65
Kurtosis	2.78	5.36	22.14	7.56	7.62	3.05	21.00
Broadband level (dB)	9.70	34.86	29.19	45.08	49.97	17.10	46.97

Comparing the spectra of the normal and defective bearings, the defective bearings generate broader spectra than the normal bearing. The main vibration components of a normal bearing are concentrated at the lower frequency (below 1kHz) region and its overall power spectrum energy (broadband level) is much smaller than that of the defective bearings as shown in Table 7-2. The spectra of the defective bearings spread from the low to high frequency ranges with several resonant frequency regions. Upon a closer examination, the spectra of the defective bearings, the number and the frequency range of the resonant frequency regions (denoted by ^) change with the type of bearing defect and defect size. For example, the spectrum of a 0.007" outer race defect has 3 aggregations at the frequency range of [1k 2k] Hz, [2.5k 4k] Hz and [5k 5.5k] Hz (see Figure 7-3(b)). When the defect size increased to 0.021", the resonant frequency region of [2.5k 4k] Hz splits into 2 humps, the resonant frequency region of [5k 5.5k] Hz disappeared and the change of low frequency (below 1kHz) components becomes stronger. These observations imply that the resonant frequency is a function of defect size and the change of frequency components in the higher frequency region is more sensitive to bearing defects than the frequency components in the lower frequency region.

7.1.2 Resonant frequency band selection

We use the procedure of automatic impulsive resonance packet selection described in Section 5.4.1 to select the dominant resonant frequency band for the HFRT algorithms. The MODWPT decomposes the signal at level 4 using a 16th order Dauberchies wavelet. As an example, the kurtosis and normalized RMS changes for the bearing with a 0.007" inner race defect in a tree structure are tabulated in Table 7-3 and Table 7-4 respectively. The shaded packets in Table 7-3 are the packets with kurtosis greater than 4. We then select the packets with the maximum RMS changes from the normal bearing for each level. For example, the fifth packet at level 3 is selected as the best packet. The equivalent frequency range of this packet is [3000 3750] Hz. A series of calculations were performed and these calculations confirmed that the selected packet gives the best results for detecting the inner race defect.

Table 7-3. Kurtosis of bearing with 0.007" inner race defect in a tree structure.

5.54 (kurtosis of original signal)															
5.38								5.37							
3.9				4.63				5.38				4.61			
2.67		3.82		5.27		5.08		5.02		5.23		3.85		4.59	
2.49	2.62	3.45	3.23	4.78	2.71	2.8	4.15	3.82	3.97	4.55	3.76	3.62	3.24	4.45	4.94

Table 7-4. Normalized RMS changes of bearing with 0.007" inner race defect in a tree structure.

2.90 (Normalized RMS changes with respect to a normal bearing)															
1.58								57.58							
0.52				8.11				81.29				10.47			
0.40		0.53		2.87		50.32		87.70		43.81		9.66		21.17	
0.42	2.87	0.41	6.58	9.70	1.56	33.7	81.4	57.9	129	57.9	25.7	14.3	6.48	21.4	9.06

Except for the case of the 0.014" outer race defect, the best resonant packet selected by our algorithm with a kurtosis threshold value of 4 are the same for the frequency range of [3000 3750] Hz. We found that the kurtosis for all the packets for the 0.014" outer race defect are smaller than 4. By lowering the threshold value to 3, the optimal resonant packet is the same as the other cases.

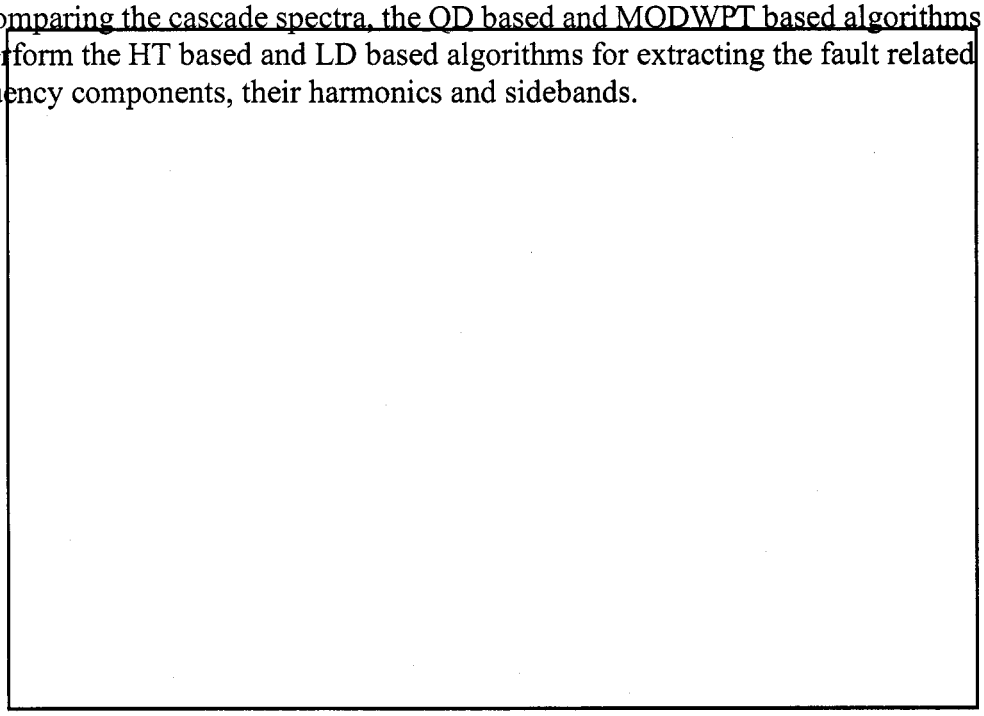
7.1.3 Results of HFRT algorithms

Knowing the optimal resonant packet and its equivalent frequency range, we can compare the performance of the MODWPT based and the energy detector based HFRTs with the traditional HFRT based on Hilbert transform. Figure 7-4 to Figure 7-15 are the results of the smoothed envelope and their envelope spectra. For traditional filter based methods, a bandpass filter uses 17 FIR filter coefficients for pre-filtering with the center frequency of 3375Hz and a bandwidth of 750Hz. For smoothing, a lowpass filter with 9 filter coefficients and a cut-off frequency of 400Hz is used. The non-zero bands for cascade quadratic detectors are set to 7 and 1 respectively. The MODWPT uses a 16th order Dauberchies wavelet in order to have an equivalent comparison with filter based methods. A hard denoising function with an estimated universal threshold from Equation 4-17 and Equation 4-18 are used to suppress the background noise.

The following observations can be concluded from the results (except for the 0.014" outer race defect case):

- The cascade spectra demonstrate that all of the HFRT algorithms can enhance the energy changes at the bearing defect frequencies than the direct FFT spectrum.
- The HT based and LD based algorithms give similar results. This can be seen from the traces of the smoothed envelope and their envelope spectra.
- The QD based and MODWPT based algorithms give the best background noise suppression for the given smooth and denoising parameters than the HT and LD based methods.

- Comparing the cascade spectra, the QD based and MODWPT based algorithms outperform the HT based and LD based algorithms for extracting the fault related frequency components, their harmonics and sidebands.



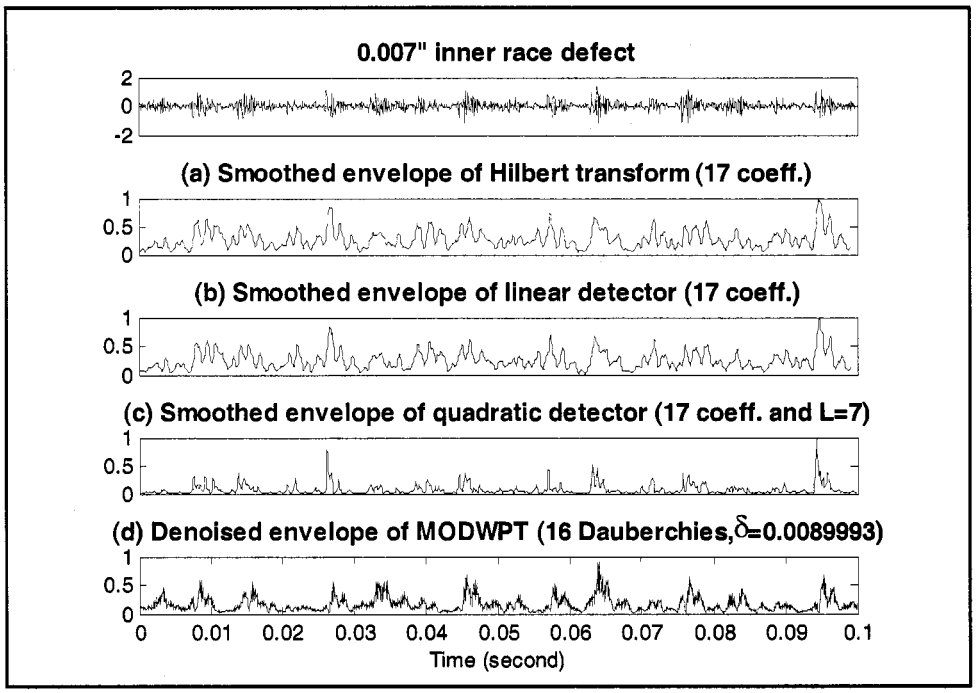


Figure 7-4. Smoothed or denoised envelope of bearing with 0.007" inner race defect using different HFRT algorithms.

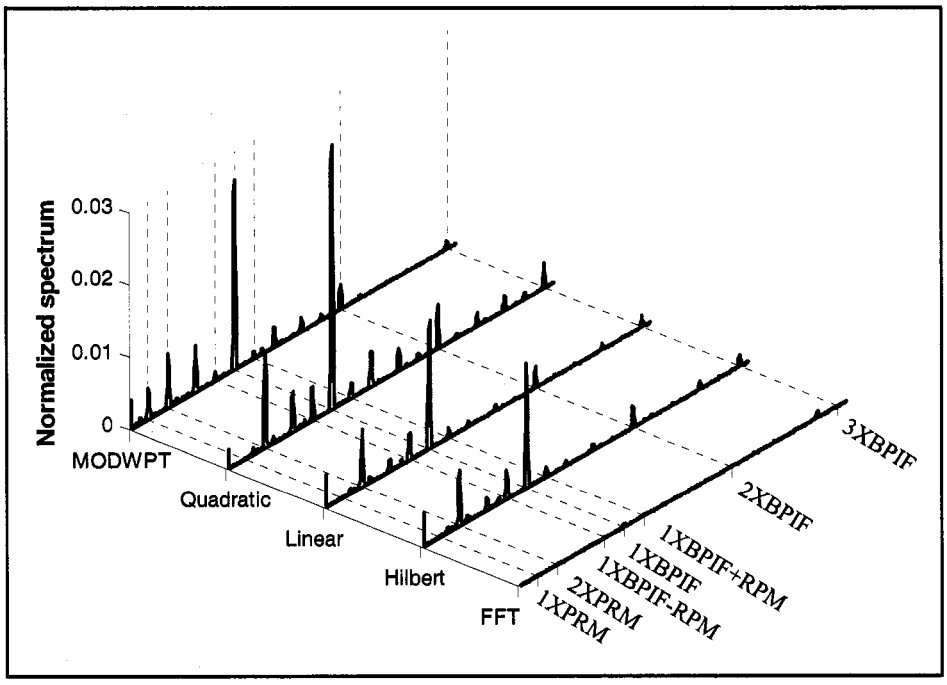


Figure 7-5. Cascade envelope spectra of bearing with 0.007" inner race defect using different HFRT algorithms.

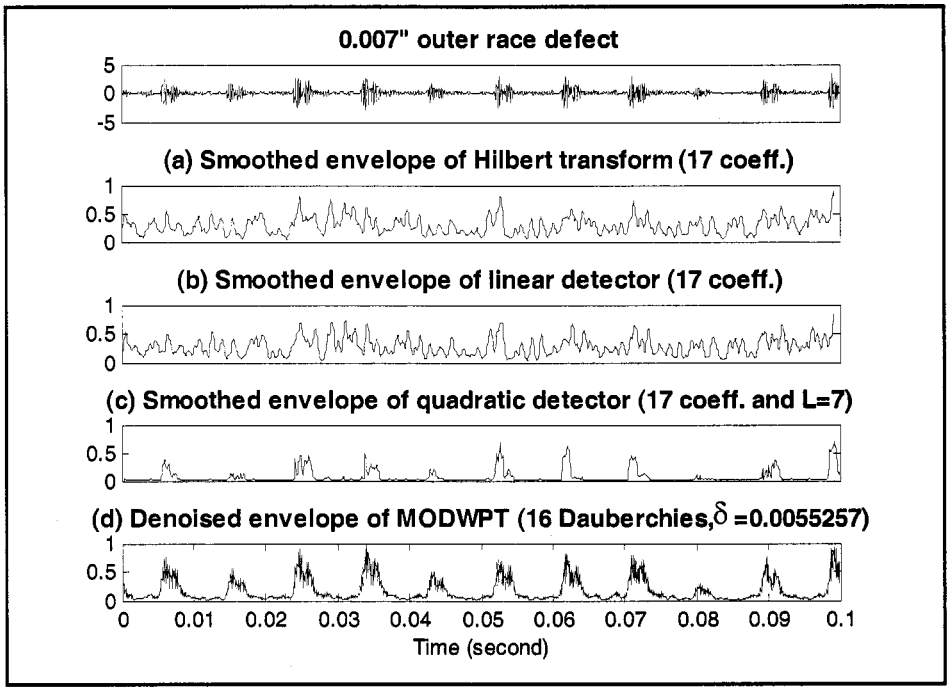


Figure 7-6. Smoothed or denoised envelope of bearing with 0.007" outer race defect using different HFRT algorithms.

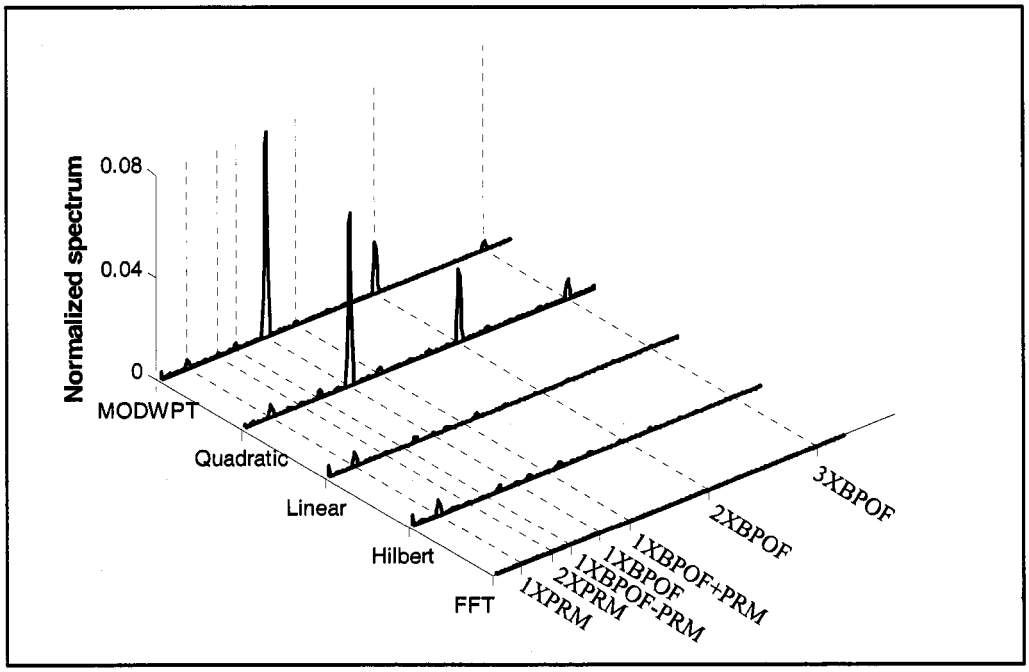


Figure 7-7. Cascade envelope spectra of bearing with 0.007" outer race defect using different HFRT algorithms.

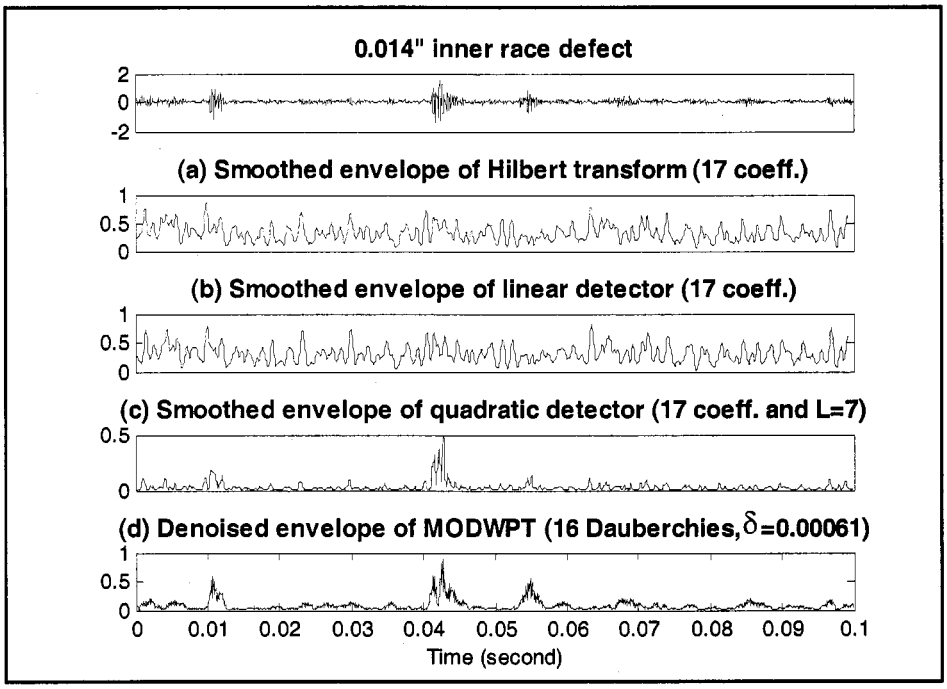


Figure 7-8. Smoothed or denoised envelope of bearing with 0.014" inner race defect using different HFRT algorithms.

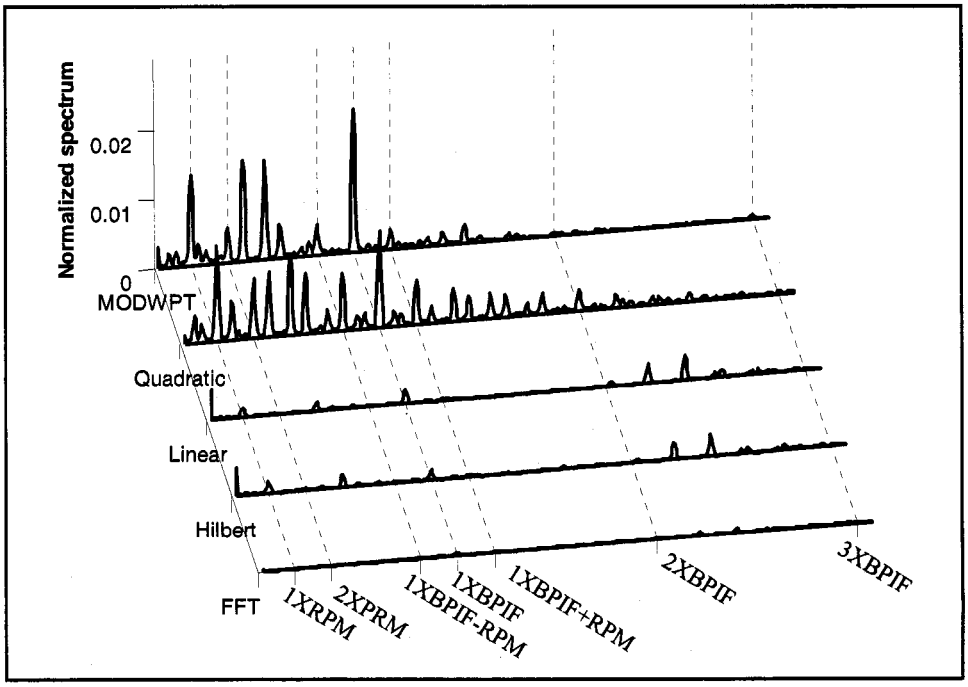


Figure 7-9. Cascade envelope spectra of bearing with 0.014" inner race defect using different HFRT algorithms.

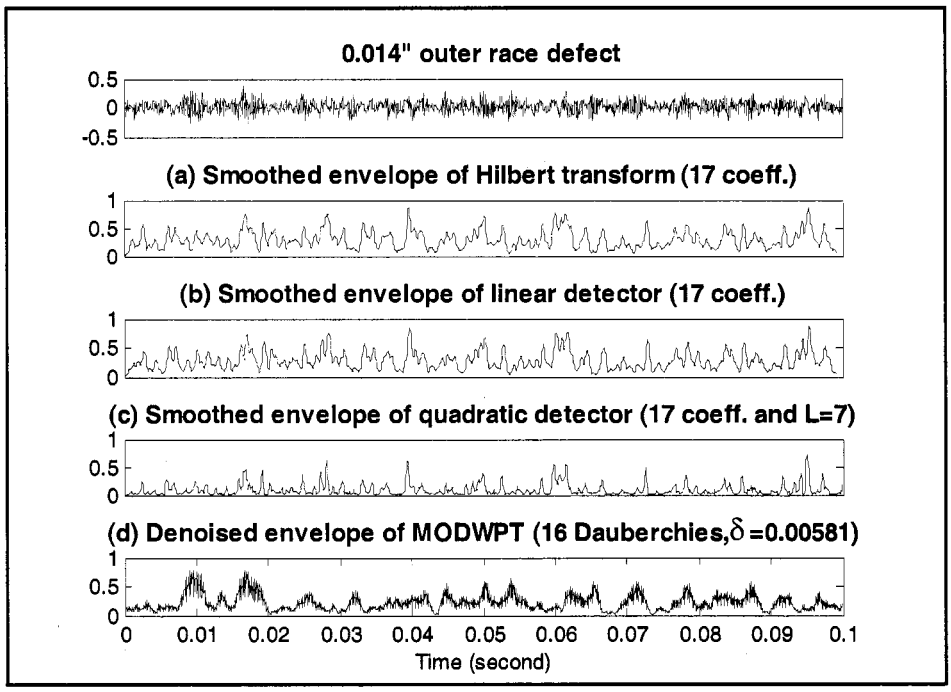


Figure 7-10. Smoothed or denoised envelope of bearing with 0.014" outer race defect using different HFRT algorithms.

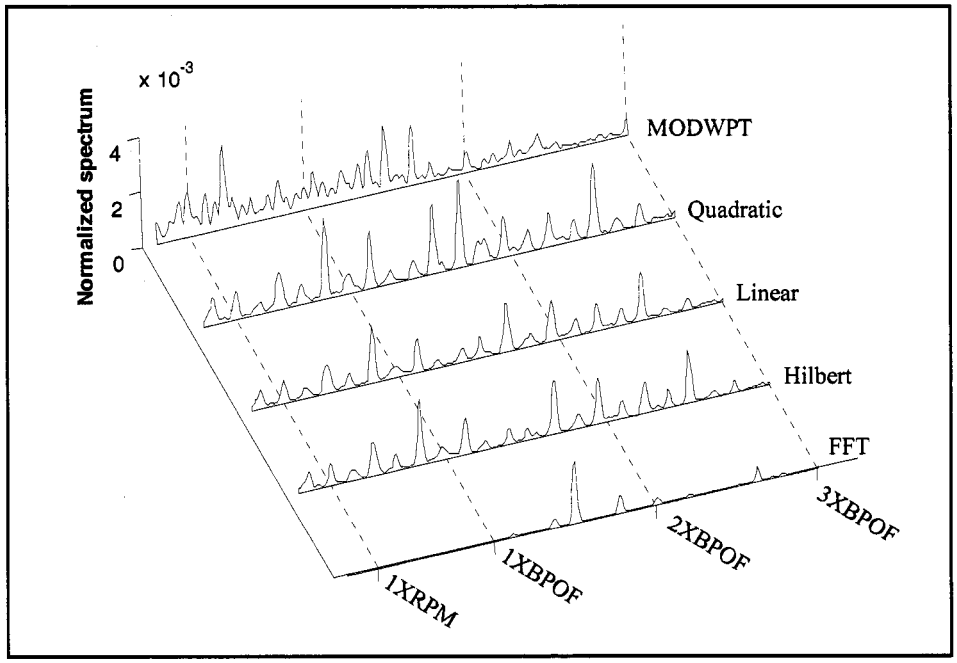


Figure 7-11. Cascade envelope spectra of bearing with 0.014" outer race defect using different HFRT algorithms.

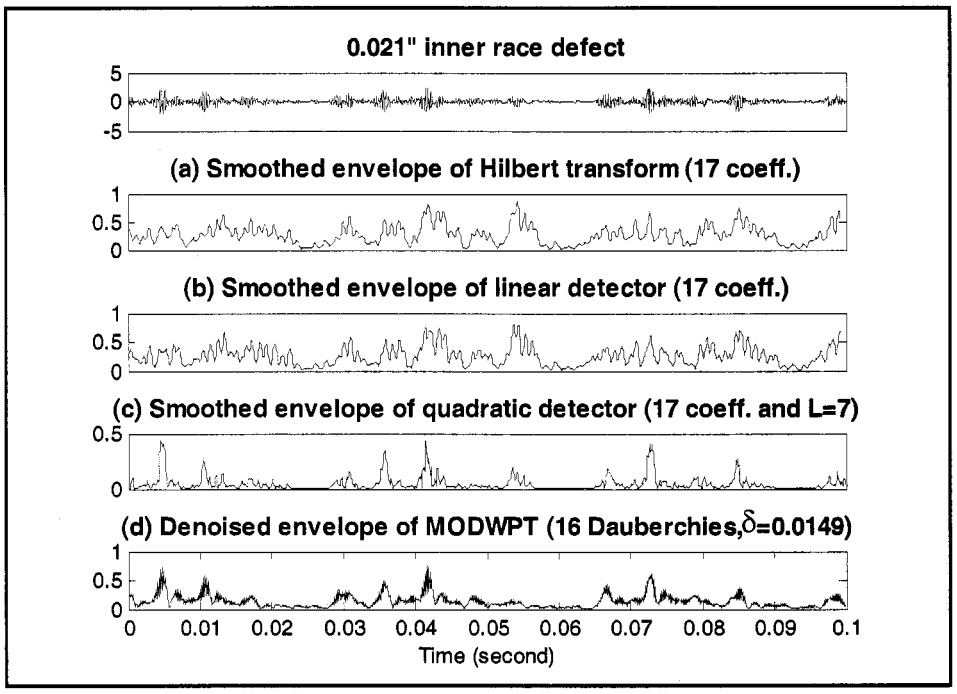


Figure 7-12. Smoothed or denoised envelope of bearing with 0.021" inner race defect using different HFRT algorithms.

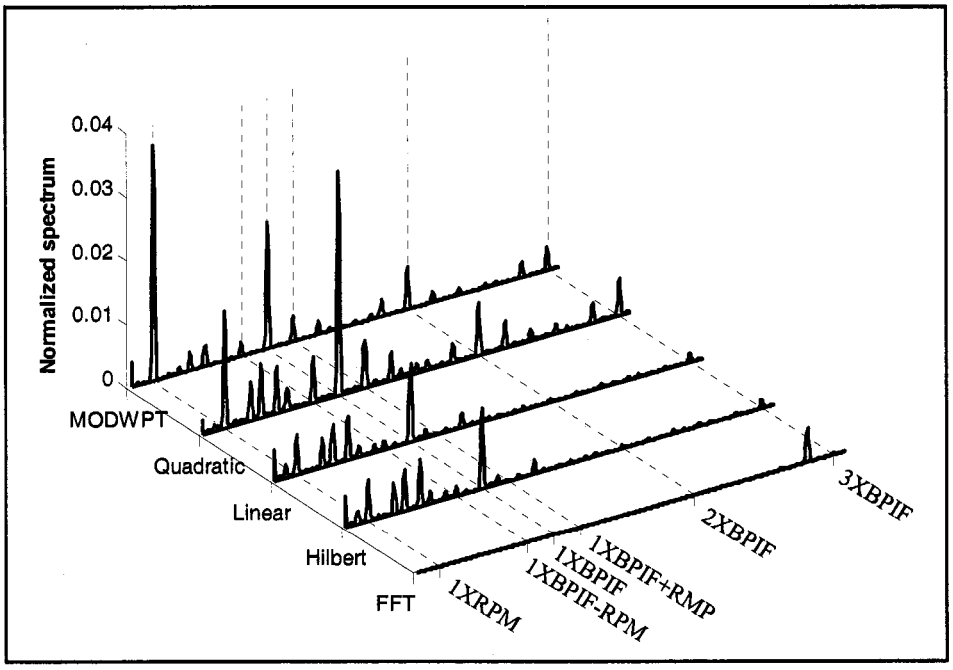


Figure 7-13. Cascade envelope spectra of bearing with 0.021" inner race defect using different HFRT algorithms.

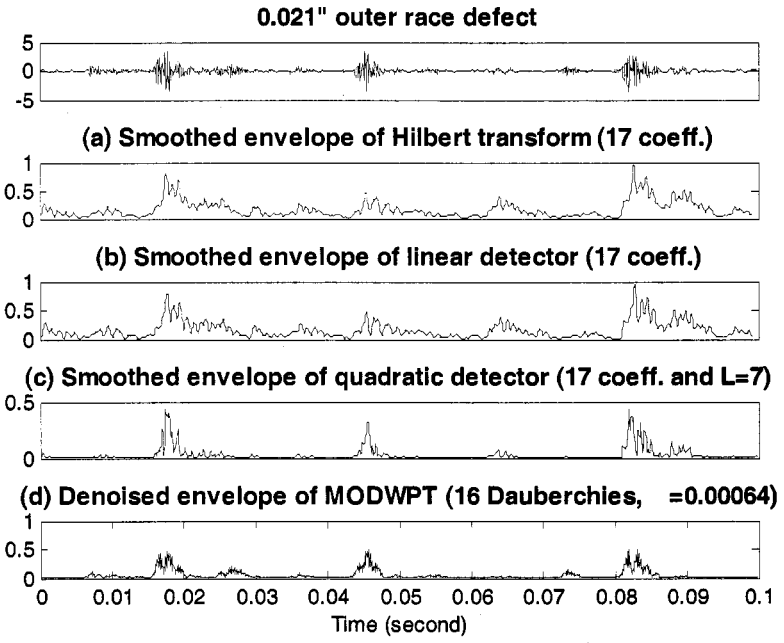


Figure 7-14. Smoothed or denoised envelope of bearing with 0.021" outer race defect using different HFRT algorithms.

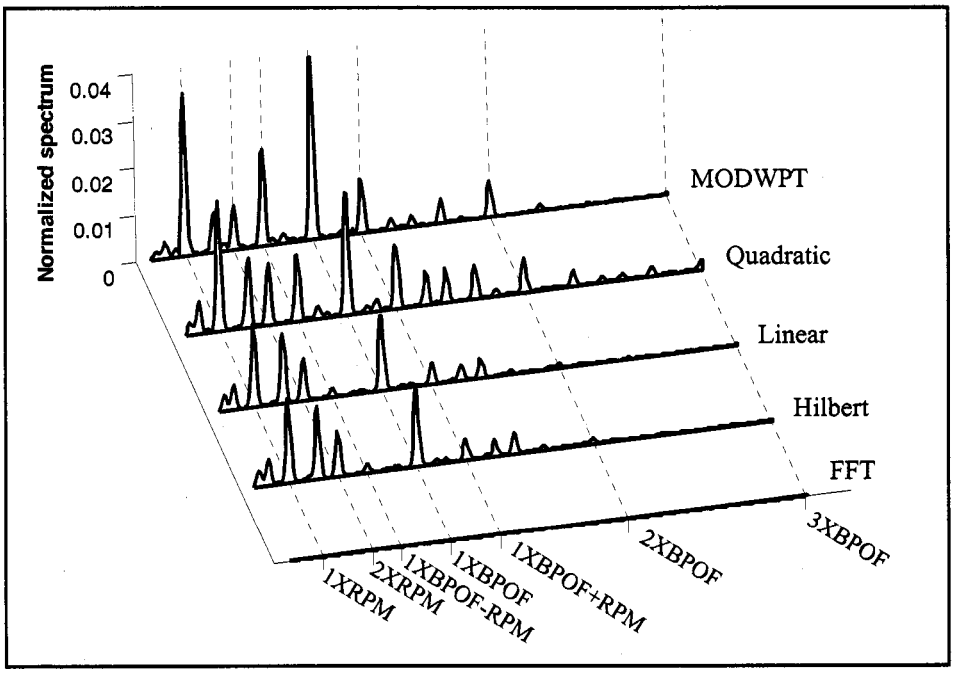


Figure 7-15. Cascade envelope spectra of bearing with 0.021" outer race defect using different HFRT algorithms.

7.2 Analysis Results of Run-to-Fail Bearing Data

Until now, most of the research on bearing defect detection has used artificially induced flaws either by scratching the surface, introducing debris into the lubricant or machining a defect with a drill or EMD [50,97]. Since studies using artificially defective bearings cannot offer insight into the early detection of a naturally propagating crack, it is necessary to study the effectiveness of the newly developed techniques with real bearing data. We perform both the traditional techniques and the proposed HFRT algorithms on bearing lifetime data which was experimentally measured from normal bearings running through to failure.

Section 7.2.1 introduces the design of the experiment and data collection. The forensic analysis of the defective bearing is presented in Section 7.2.2. Based on the observation from the failure analysis, we perform trending plots to compare the performance of traditional time based and frequency based techniques with proposed HFRT algorithms in Section 7.2.3. Analysis results are summarized in Section 7.3.

7.2.1 Description of experimental setup

Figure 7-16 shows a schematic diagram of the experimental setup used in this study. There are three sub-systems in this setup: a test rig system, a data acquisition system and a real time monitoring system.

7.2.1.1 Test rig system

The test rig consists of a 1/4" shaft supported by two ball bearings, an unbalanced flywheel, a stepper motor to drive the system, a flexible bellows coupling to connect shaft and motor, and a supporting structure to hold the entire system.

Both the testing and supporting bearings are single row, double metallic shielded miniature bearings manufactured from NMB Corporation. The supporting bearing is a larger bearing with higher load rating (about 3 times larger) than the testing bearing. The dimensions of both bearings and the bearing defect frequencies are tabulated in

Table 7-5. Miniature ball bearings were chosen for this study because they are widely used in computer equipment (such as spindle motors for HDD's and fan motors), household electrical appliances and medical instruments (such as dental drills and heart pump machines) [59]. The construction of miniature bearings enables them to carry moderate radial and thrust loads at low operating torque while providing high performance within a small envelope design. However, small size and thin raceways make miniature bearings vulnerable to damage from contamination and high handling forces.

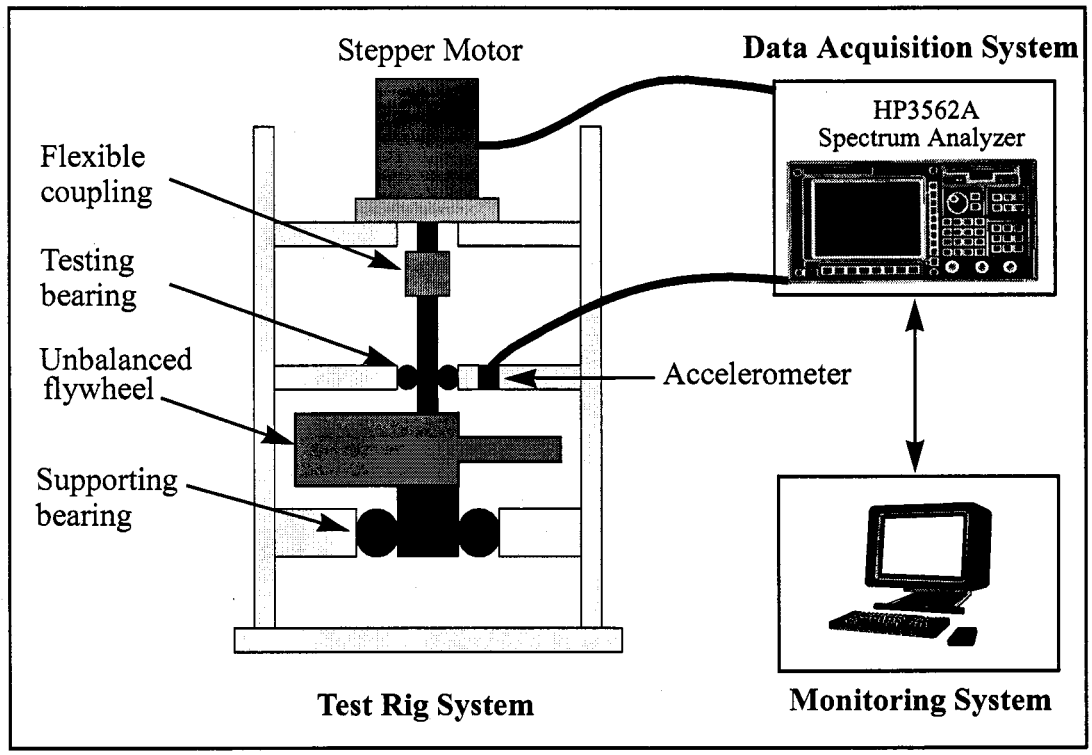


Figure 7-16. Schematic drawing of bearing testing setup.

Table 7-5. Bearing specifications and defect frequencies of run-to-fail experiment.

Bearing Model Number	Testing Bearing DDRI-418	Support Bearing DDRI-814
Bore (D_i)	0.1250"	0.2500"
O.D. (D_o)	0.2500"	0.5000"
Width (B)	0.1094"	0.1875"
Number of Balls (Z)	7	10
Ball Size (D)	1 MM (0.0394")	1/16"
Contact angle (α)	14.5°	12°
Cage type	Crown type	Ribbon type
Static load rating	22 lbs	84 lbs
Dynamic load rating	64 lbs	187 lbs
Enclosures	Double metallic shield-removable	
Bearing material	AISI440C Martensitic stainless steel (heat treated)	
Bearing defect frequency (running speed=17Hz)	f_{ftf} = 6.77Hz f_{bpof} = 47.40Hz f_{bpif} = 71.60Hz f_{bsf} = 77.55Hz	f_{ftf} = 7.11Hz f_{bpof} = 71.14Hz f_{bpif} = 98.86Hz f_{bsf} = 99.29Hz

Figure 7-17 shows the top view of the test rig system. The testing bearing was mounted midway between the supporting bearing and the stepper motor. In order to accelerate the bearing failure process, testing commenced by subjecting the bearings to a radial load using an unbalanced flywheel. Unbalance is a common cause of bearing failure due to non-symmetrical reduction of the bearing clearance from wear. In this test rig, a centrifugal force acts radially outward from the center of rotation which creates a server environment when the bearing is in operation. Using Equation 2-12, the centrifugal force is 22.54lbs for flywheel running at 17Hz (1020RPM).

The stepper motor is a synchronous motor manufactured by Superior Electrics Company (model number: M061-FD-301) which can drive the system in steps in response to electrical pulses from a signal generator. Using conventional 60Hz current, it rotates at exactly 200 steps per revolution.

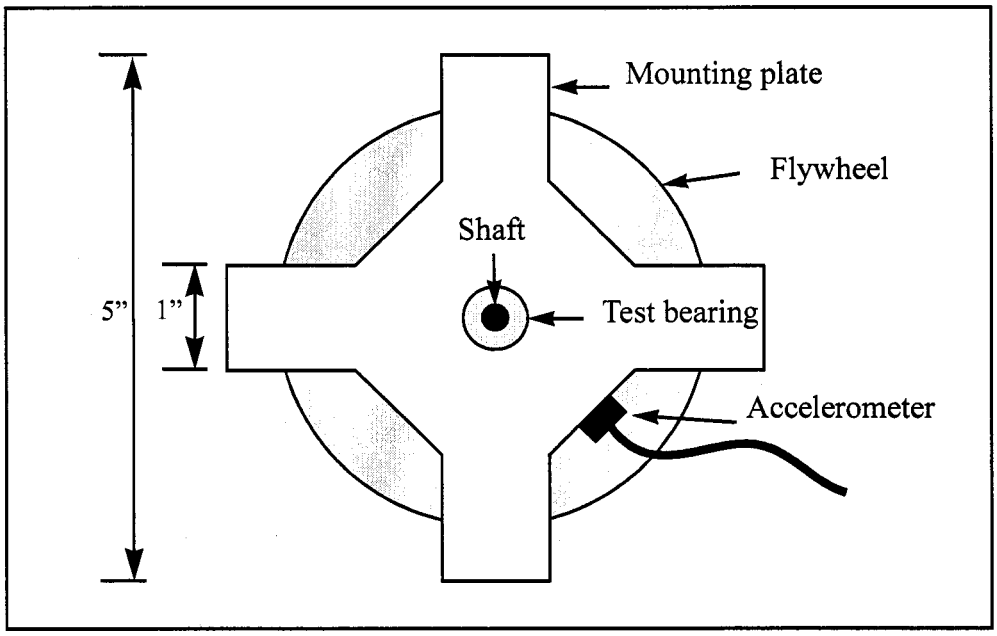


Figure 7-17. Top view of the bearing test rig.

7.2.1.2 Data acquisition system

Data were collected using a HP 3562A spectrum analyzer from a PCB piezotronics accelerometer mounted near the test bearing. The accelerometer was mounted to the bearing in a radial direction using petro wax. The linear frequency range of the accelerometer is 1-10 kHz. The stepper motor ran continuously at 17 Hz. The vibration time history was continuously sampled every 10 minutes. The number of data points collected were 20480 at the sampling rate of 25.6 kHz.

7.2.1.3 Real time monitoring system

A graphical user interface MATLAB program was coded to perform signal analysis. Figure 7-18 is a screen shot of the graphical user interface system. The trending of time domain indicators, the short time Fourier transform and the wavelet packet transform were processed in real time. A base level was set based on an average of spectrum energy within 50Hz to 400Hz from the first 10 records. When the band energy was found to be 3 times greater than the base level or decreased to 20% of the base level (bearing seizure), the program sent a signal to stop the stepping motor, thus ending the test run.

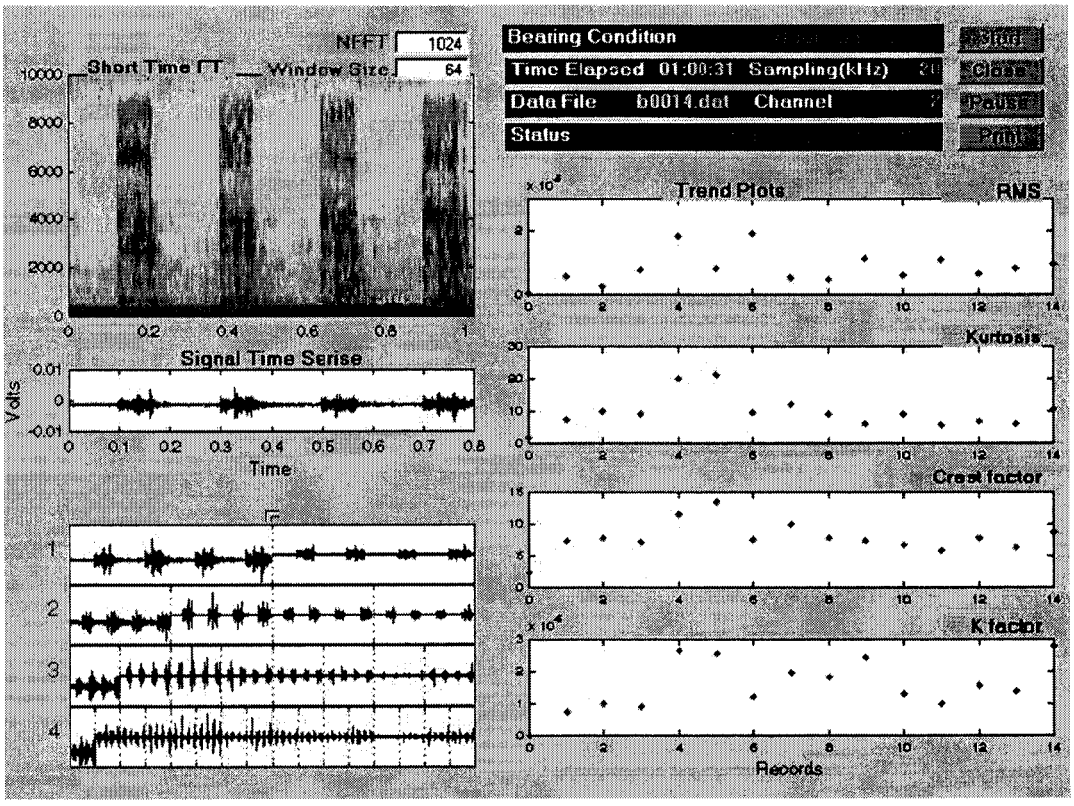


Figure 7-18. A graphical user interface of real time monitoring system.

7.2.2 Failure Analysis

A total number of six tests were performed at a constant running speed of 17Hz. The life of the bearings varied from 3 hours to 25 hours. After removing the test bearing and inspecting the races, cage and balls, the damaged areas in the bearing were inspected using a scanning electron microscope (SEM). Table 7-6 summarizes the resulting damage generated on each of the bearing components at the end of the run-to-fail testing. All of the tests have multiple defects in two or more components by the end of the bearing life. Only test R24 shows damage on the cage. Most of the damage was located in the inner race, outer race and balls. The SEM pictures are shown in Appendix A. The bearings in tests number R24, R25, R26 and R27 were the most severely damaged.

Table 7-6. Summary of run-to-fail tests.

Test no.	No. of records	Test duration (hour)	Failure mode			
			Cage	Inner race	Outer race	Ball
R24	84	14.00	tear out at crown edge and spalls at root	subsurface spalls (<10μm)	small spalls (<10μm)	NA (missing)
R25	122	20.33	normal	normal	spalls (100μm)	spall (35μm)
R26	45	7.50	normal	spall (90μm)	wear marks and spalls (<10μm)	scratch (50μm)
R29	20	3.33	normal	spalls (50μm)	normal	small pits (<3μm)
R30	149	24.83	normal	wearing marks with flakes	wearing marks and rough surface	normal
R31	47	7.83	normal	wearing marks with flakes	wearing marks	scratch (100μm)

7.2.3 Bearing signal processing

In order to compare the performance of the bearing defect detection techniques, the time domain indicators, spectrum of frequency domain and features from HFRT algorithms were evaluated by trending.

Trending is a popular method in the condition monitoring industry to determine whether the condition of the measured machinery or component indicates that it has become damaged. The idea of trending detection is to select a fault-representing indicator and plot this indicator as a function of time. A warning signal is triggered when the current amplitude of the trend curve exceeds a preset alarm level or the percent change from the normal condition exceeds a preset limit. Experience in industrial settings shows that setting the alarm level at 2 times the mean amplitude of the normal condition provides optimum sensitivity to small changes combined with maximum immunity to false alarms [6]. We will use this warning criteria for the following analysis.

7.2.4 Time domain indicators

Figure 7-19 shows the vibration waveforms of all bearings at the beginning and the end of their lifetime testing. All of the waveforms recorded in the first sample were similar to a random noise by visual inspection and the peak amplitudes are around 2. At the end of the run, the shape of waveforms were different. Test R26, R29 and R31 remained random with approximately the same peak amplitude as they had at the beginning of the test run. However, R24 and R25 vibrated at much higher amplitudes. In addition, R25 had a distinct impulse with a maximum amplitude of 15. On the other hand, near the time of failure, the vibration of R30 became very small because of bearing seizure.

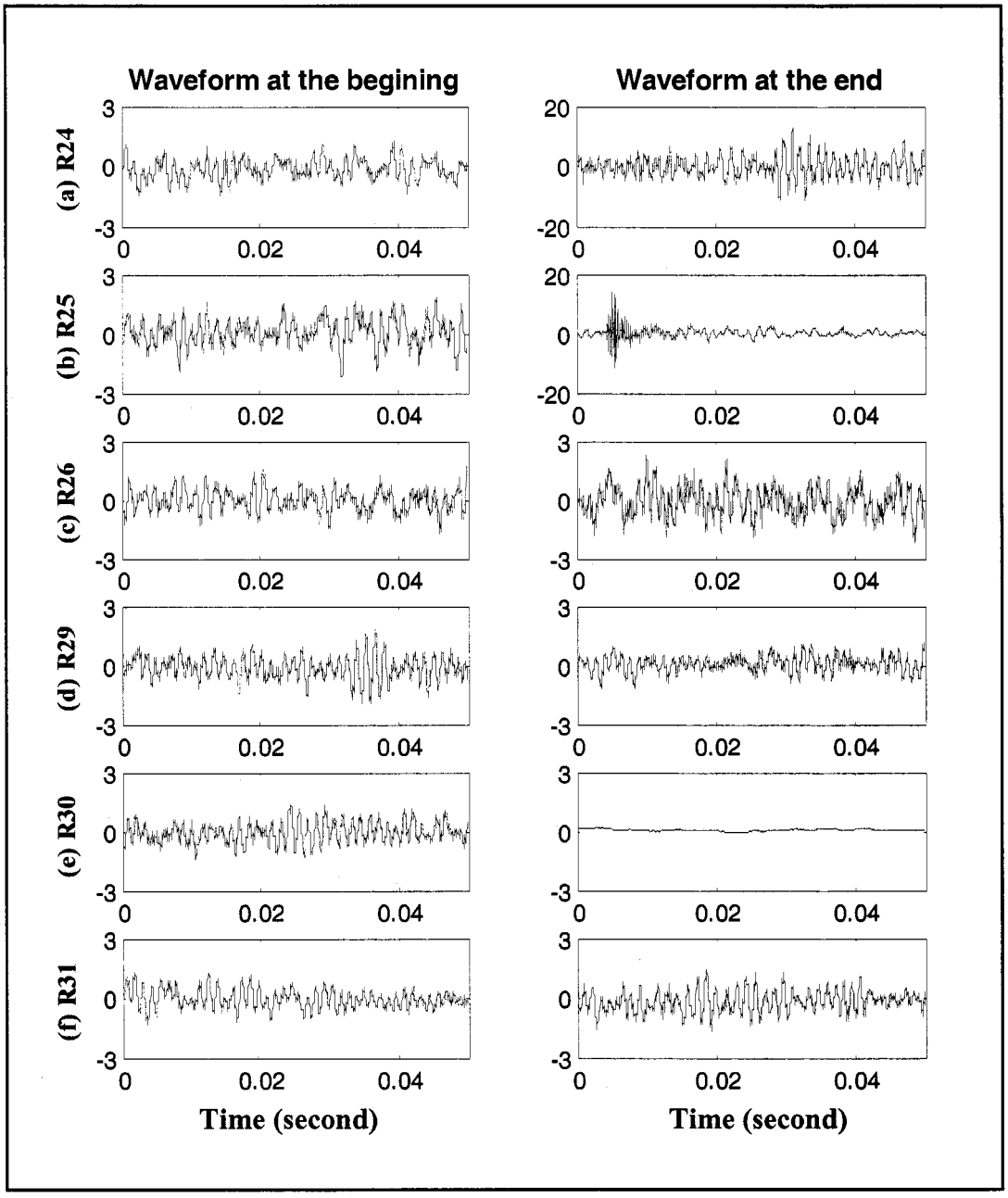


Figure 7-19. The vibration waveforms of bearings at the beginning and the end of life testing.

Figure 7-20 shows the trending plot of the normalized RMS and kurtosis for all of the bearings tested in the run-to-fail tests. In general, an increase of the normalized RMS to 2 or kurtosis of greater than 4 was considered a possible indicator of bearing defects. Based on this assumption, kurtosis is a more sensitive indicator than RMS as can be seen from the trending plots of R24, R25, R26 and R30. However, both indicators cannot detect the bearing defect for the run-to-fail tests of R29 and R31. Table 7-7 summarizes the sensitivity of the time domain indicators based on the above assumptions.

Table 7-7. First triggered record number using time domain indicators.

Test Number	First triggered record number	
	Normalized RMS>2	Kurtosis>4
R24	55	54
R25	NA	27
R26	NA	16
R29	NA	NA
R30	NA	144
R31	NA	NA

We can improve the sensitivity of time domain indicators by using a filtered signal. For example, Figure 7-21 is the trending plot of R25 for a filtered signal at a different frequency band. Except for the lowest frequency band (0-1600Hz), all the filtered signals triggered the warning at earlier records compared with the original signal as listed in Table 7-8. This is because the time domain indicators are dominated by the lower frequency components.

Table 7-8. First triggered record number of the test R25 at different frequency band using time domain indicators.

Frequency band (Hz)	First triggered record number	
	Normalized RMS>2	Kurtosis>4
1600-3200	81	11
3200-4800	81	9
4800-6400	81	11
6400-8000	NA	11
8000-9600	NA	11

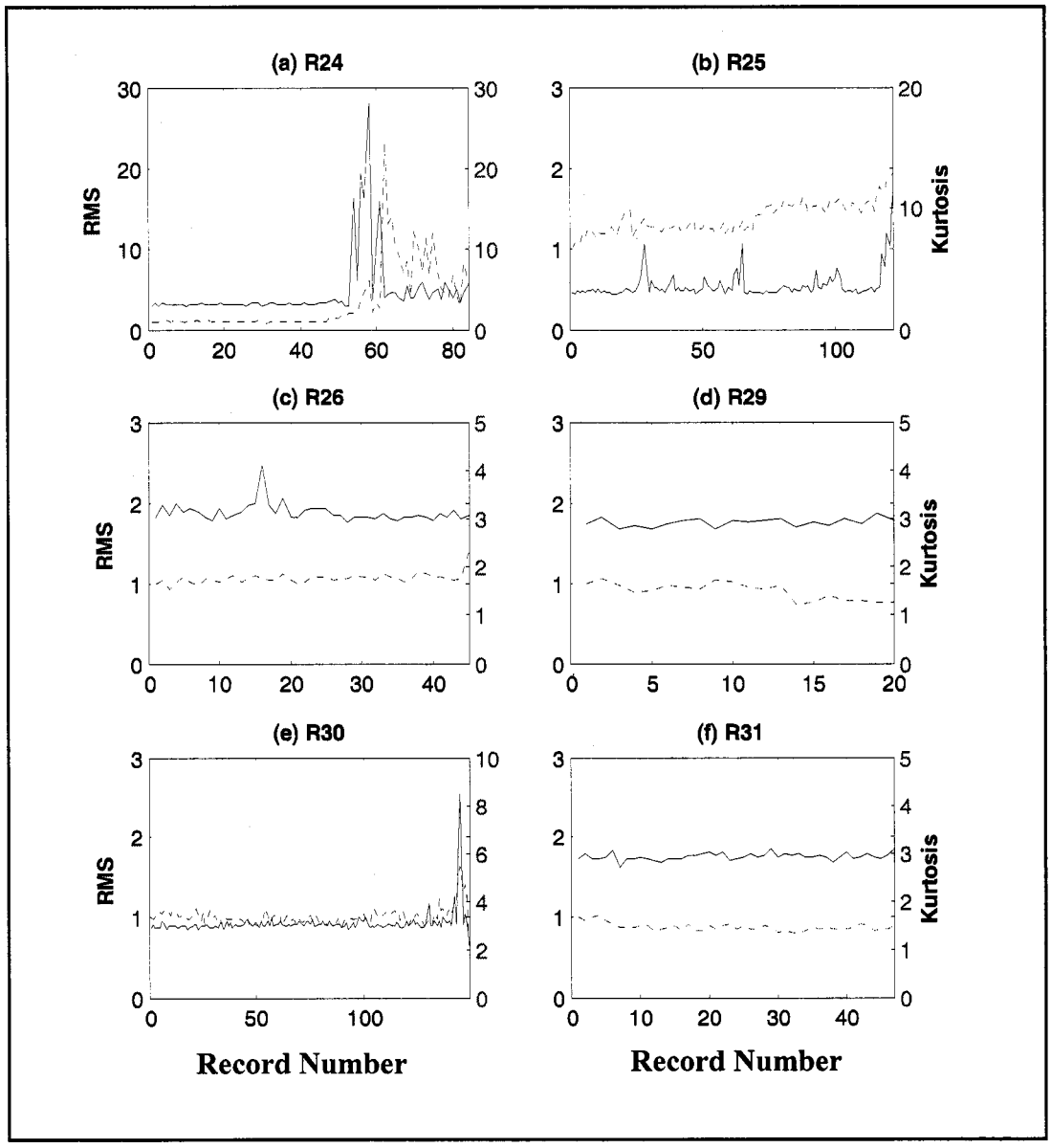


Figure 7-20. The normalized RMS (dotted line) and kurtosis (solid line) trending of lifetime testing bearings.

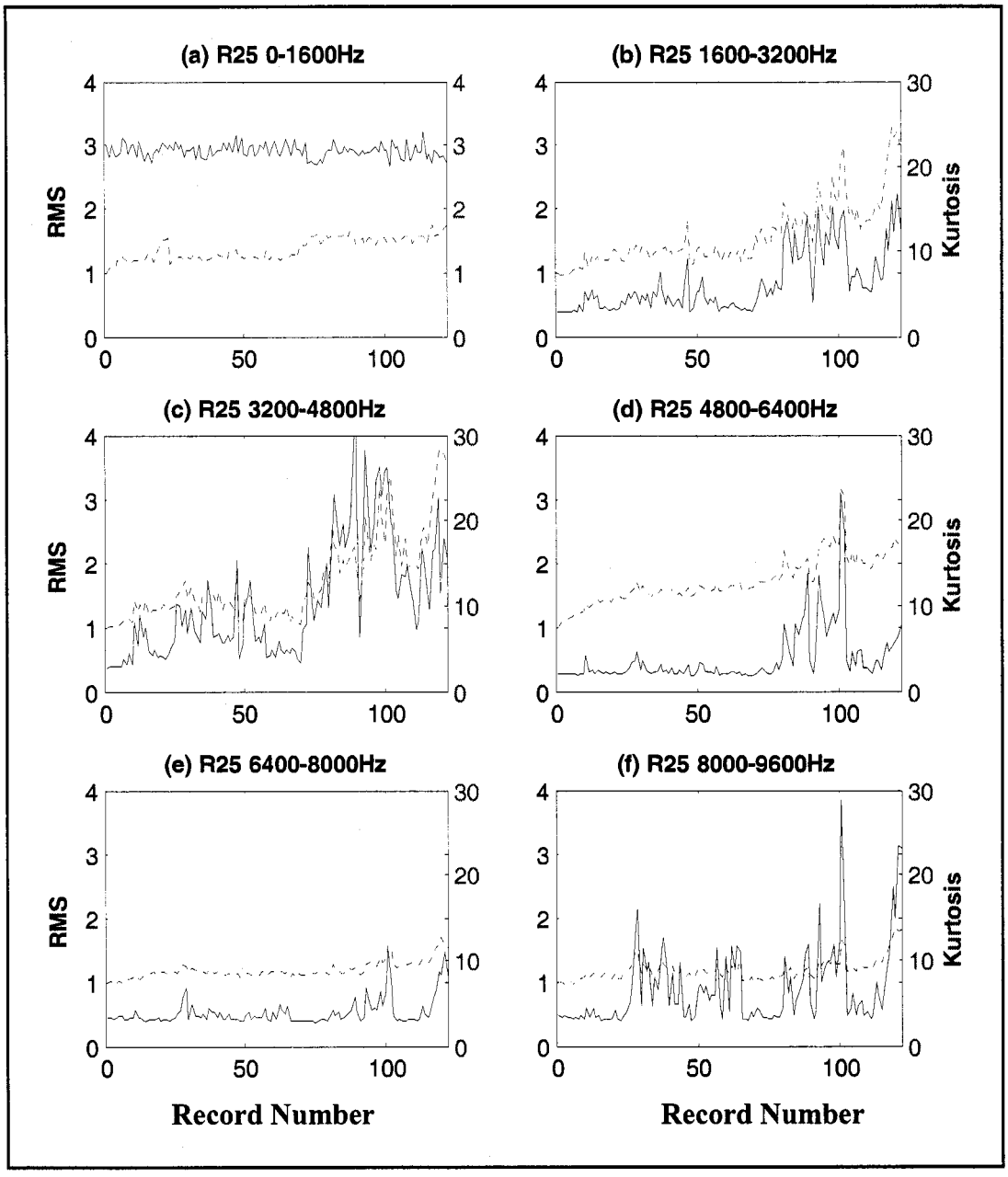


Figure 7-21. The normalized RMS (dotted line) and kurtosis (solid line) trending of testing bearing R25 at different frequency band.

7.2.5 Frequency domain trending

In the frequency domain, the spectrum represents information indicating the level of vibration at a particular frequency. Tracking the changes of the spectrum can help us to detect and identify bearing defects. Figure 7-22 to Figure 7-27 show the trending maps of the bearing spectra in dB scale in which x axis is the time (or record number) and the y axis is the frequency. In general, the highest energy was concentrated in the lower frequency range (below 1000Hz) which is due to the unbalanced load applied by the flywheel. The lowest energy is in the range of 4000Hz to 6000Hz. During the bearing testing, a bearing defect may cause hot spots to be generated at a higher frequency. For example, after running for 540 minutes, a hot spot shows up at 9500Hz for testing R24. Some trending maps do not have obvious changes in the spectrum map such as the test data for R26 and R31.

Traditionally, spectrum analysis uses a narrowband method to monitor the energy changes at bearing defect frequencies and their harmonics. When the energy level at the defect frequency and its harmonics becomes greater than a preset alarm level, the bearing is diagnosed as a defective bearing. Figure 7-28 illustrates the zoomed trending maps of bearing spectra from 10Hz to 100Hz. The running speed and bearing defect frequencies are denoted as horizontal dotted lines. All the trending maps are dominated by the running speed which is due to the unbalanced load. Except for the R24 test, none of the obvious spectrum changes at the bearing defect frequencies can be detected from the zoomed maps. The change of bearing defect frequency is very likely to be masked by the dominant spectrum which is a common deficiency of traditional spectrum analysis.

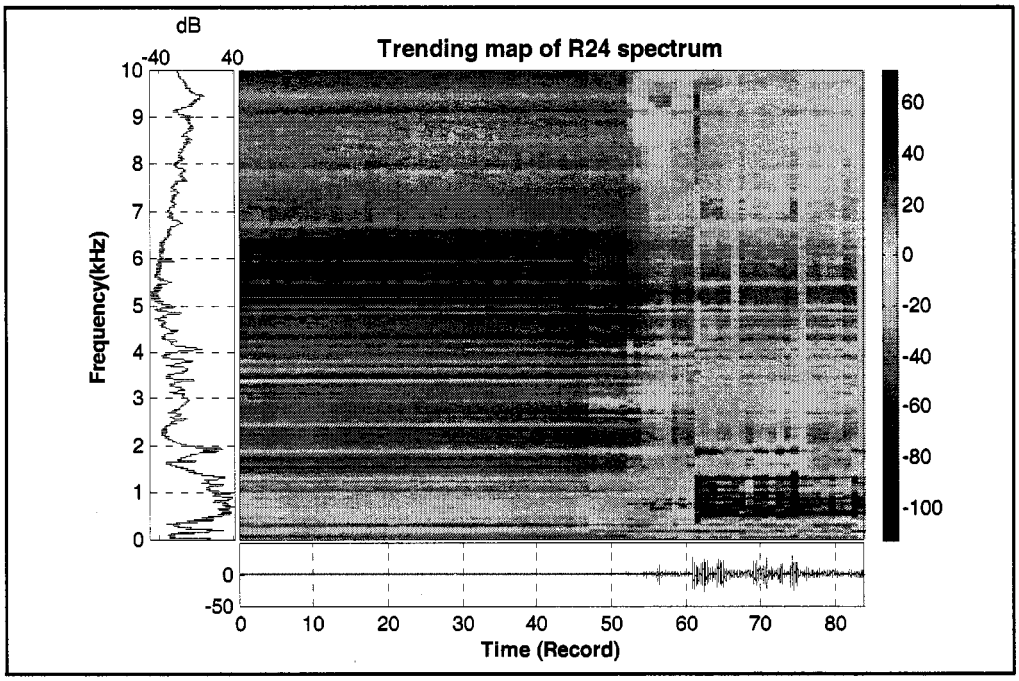


Figure 7-22. Trending map of R24 spectrum.

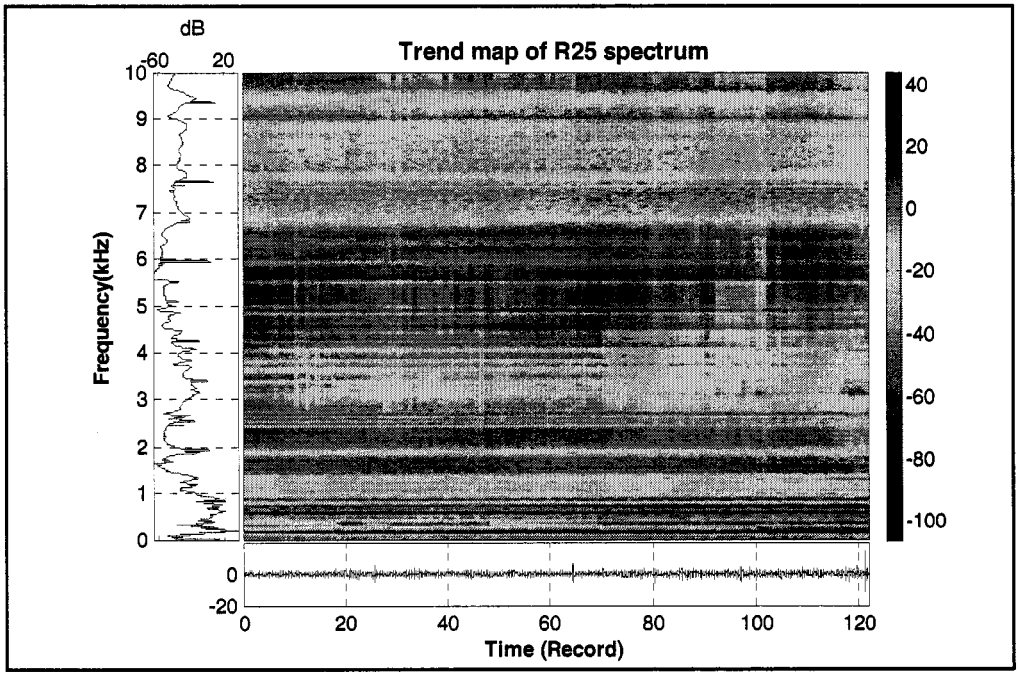


Figure 7-23. Trending map of R25 spectrum

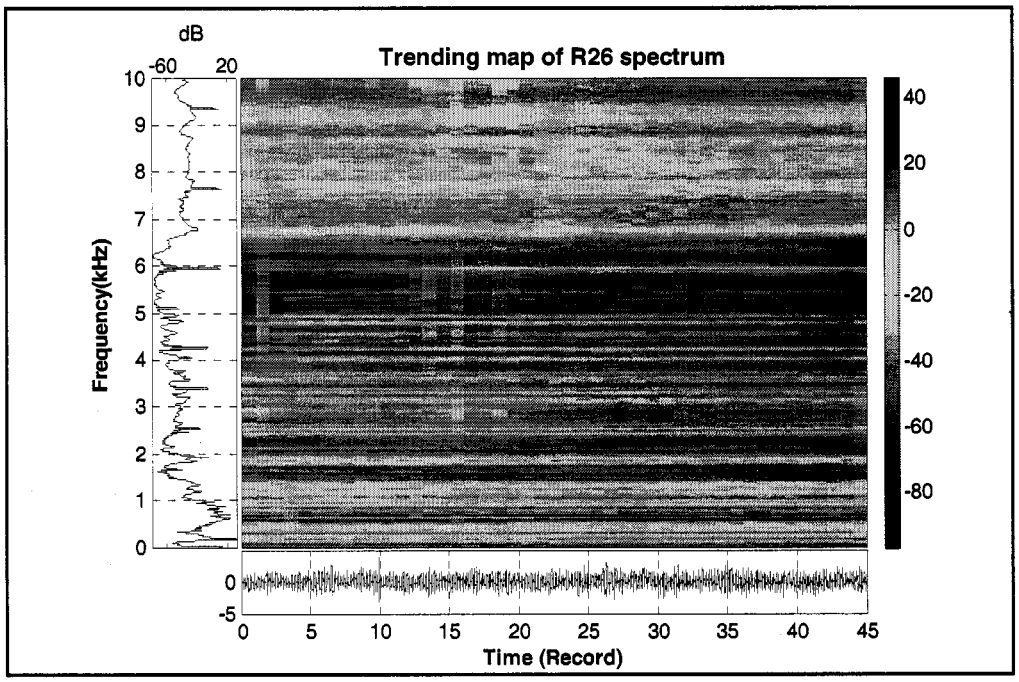


Figure 7-24. Trending map of R26 spectrum.

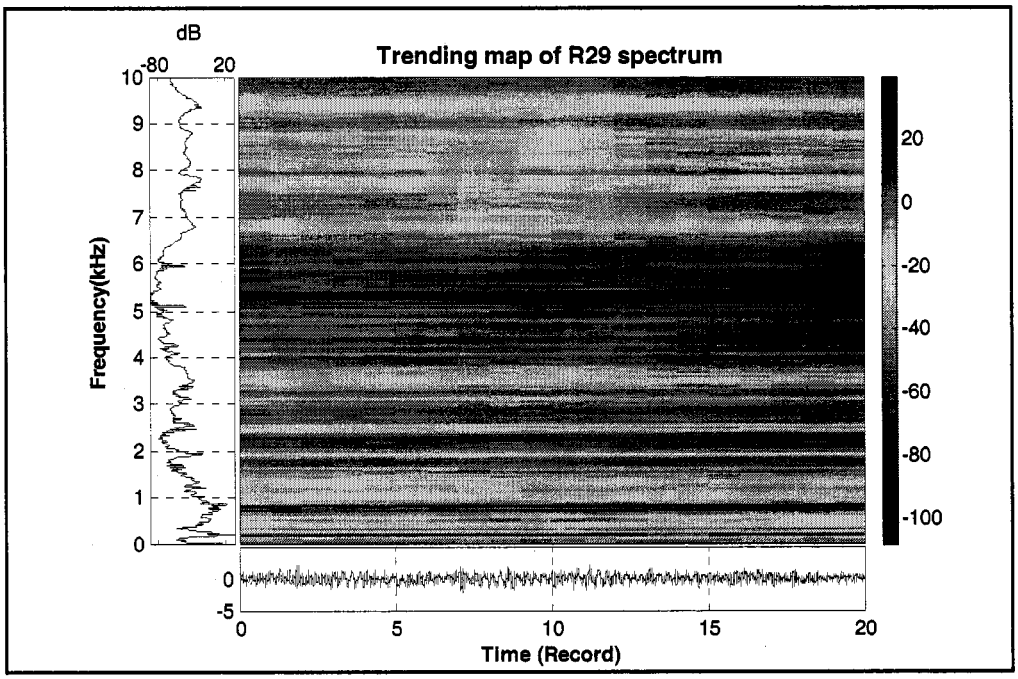


Figure 7-25. Trending map of R29 spectrum.

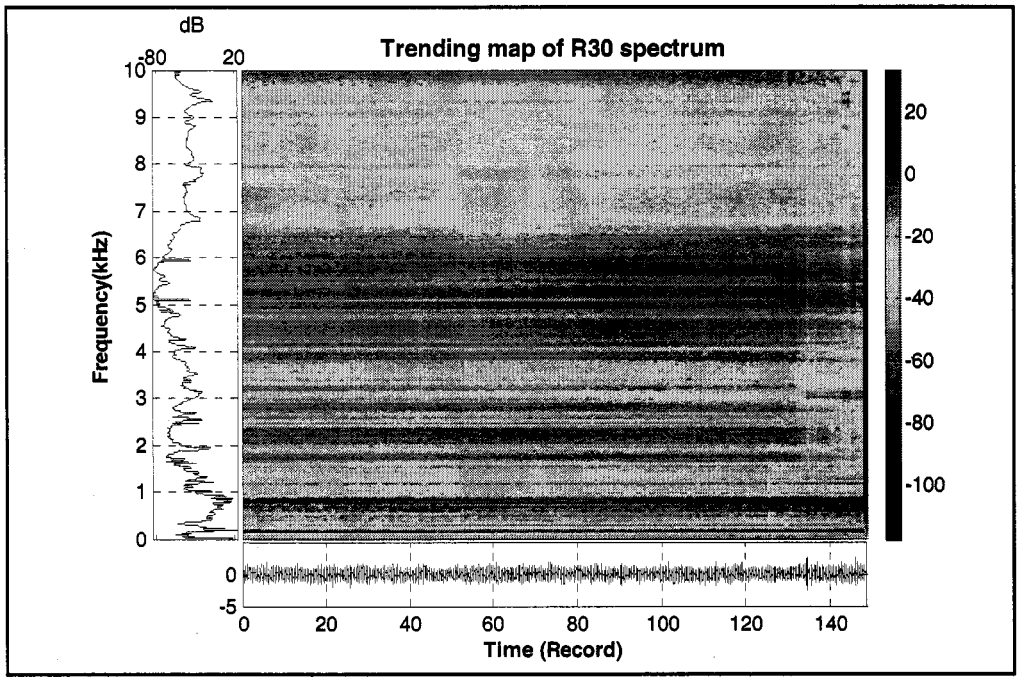


Figure 7-26. Trending map of R30 spectrum.

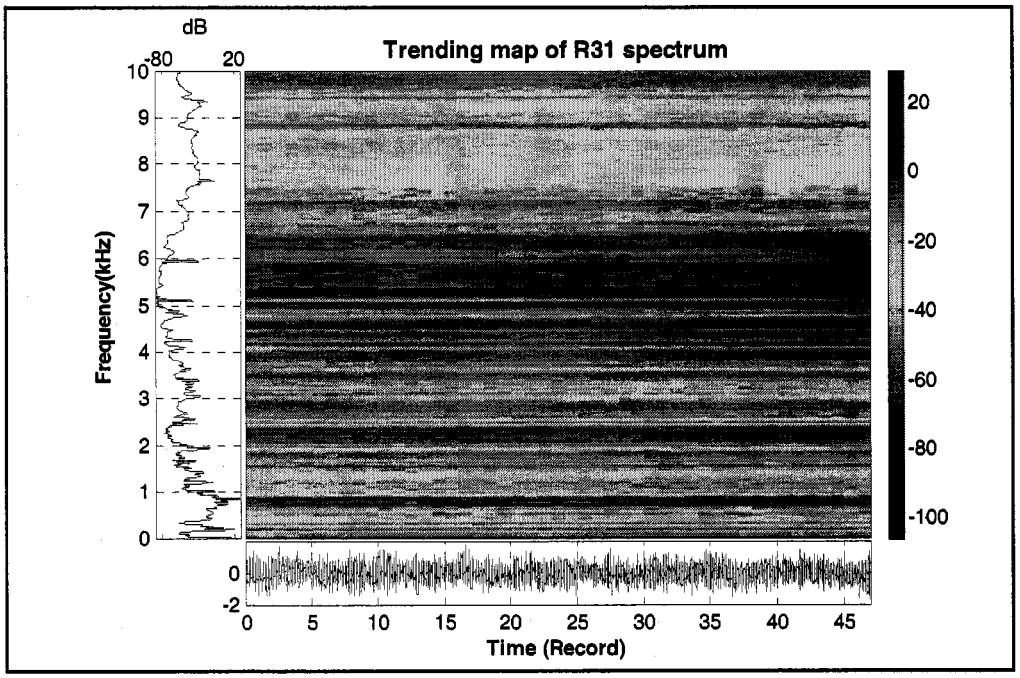


Figure 7-27. Trending map of R31 spectrum.

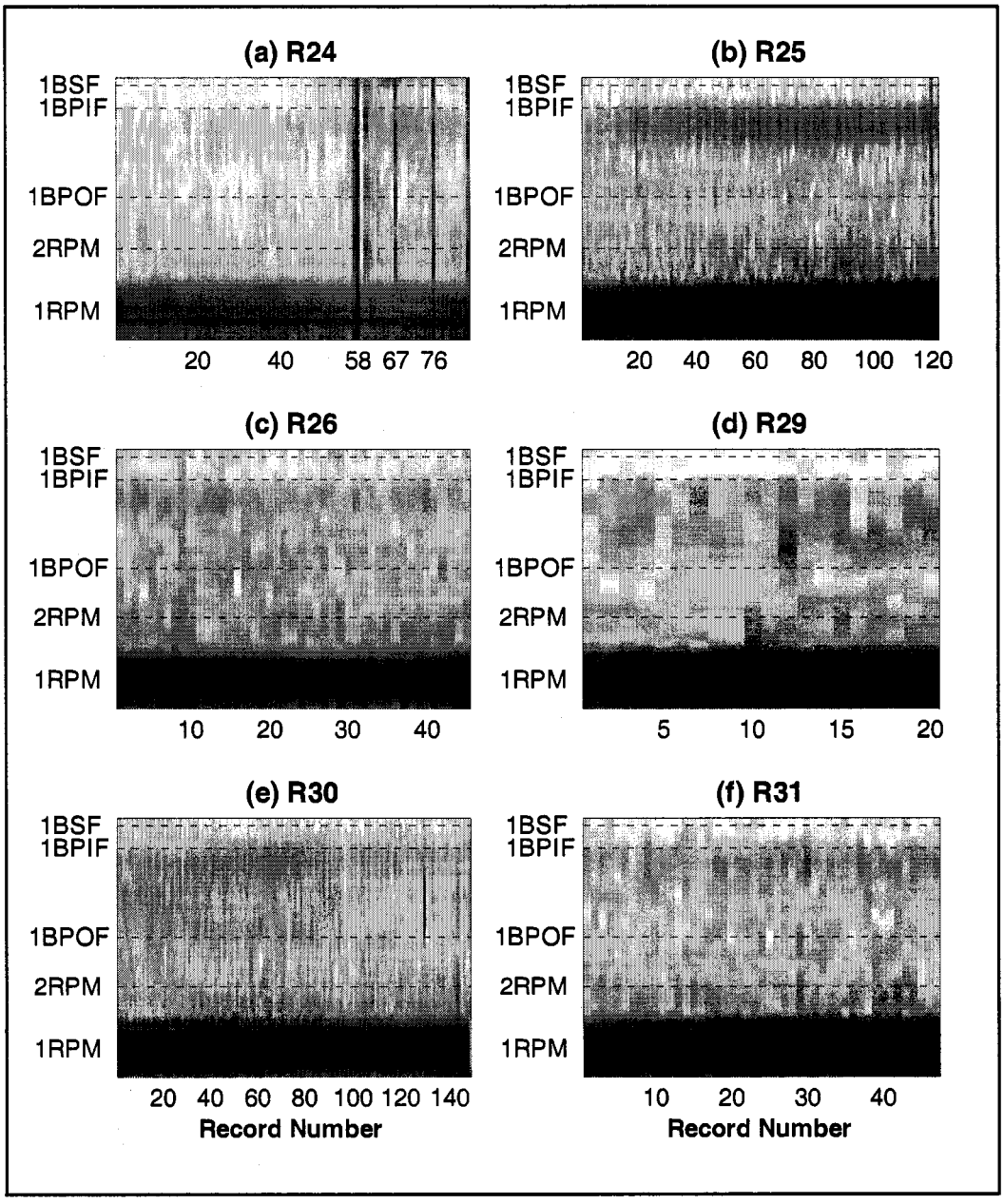


Figure 7-28. The zoomed trending maps of bearing spectra in dB scale from 10Hz to 100Hz.

7.2.6 Results of automatic resonant packet selection

The same procedure of resonant packet selection is applied to the data sets of lifetime bearings. The first step is to decompose the signal at level 4 using 16th order Daubechies wavelet. A threshold value of kurtosis is set to 4 to extract packets with high peakness. The optimal resonant packet is the packet with maximum RMS change with respect to the RMS average of the first 5 records.

The trending of the optimal resonant packet for all of the testing is shown in Figure 7-29. The corresponding resonant frequency range is denoted as the y axis. Most of the resonant frequency range is located within 6.8kHz to 9.6kHz. For testing R24 and R25, at the end of the run, the resonant frequency shifted to below 4kHz, which implies that the bearing is severely damaged with multiple defects. The SEM pictures of R24 and R25 confirmed this observation.

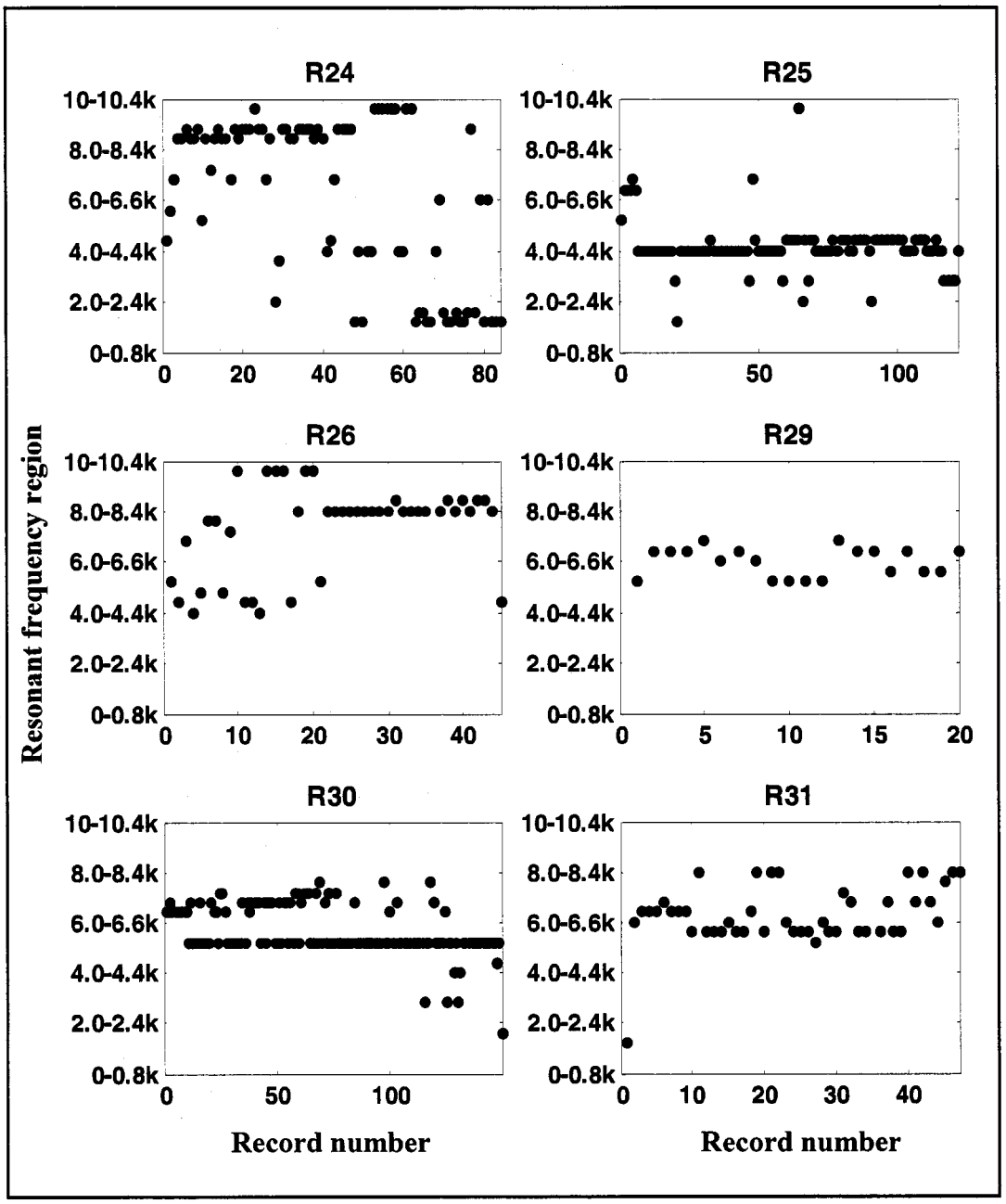


Figure 7-29. Trending of resonant frequency region of lifetime bearing testing.

7.2.7 Results of HFRTs

Using the optimal resonant packet from the previous section, we apply the HFRT algorithms to the selected packet or the filtered signal with an equivalent frequency band. By examining the envelope spectrum, we can diagnose the type of bearing defect through the existence of bearing defect frequencies in the spectrum. The zoomed trending maps of the envelope spectra in linear scale for all the bearing testing are shown in Figure 7-30 to Figure 7-35. The strong tones are denoted by darker spots in the map. Comparing with the zoomed trending maps using direct FFT (see Figure 7-28), the dominant component at the running frequency is suppressed and the components related to the bearing defect are enhanced using HFRT algorithms.

The results demonstrate that all the HFRT algorithms are better techniques for early defect detection as compared with the traditional direct FFT method. For example, the trending maps of testing R24 illustrated that all the HFRT algorithms were able to detect the inner race defect (1XBPIF) at various records.

Comparing the performance between different HFRT algorithms, we found that the QD based and the MODWPT based algorithms outperforms the Hilbert based and the LD based algorithms for noise suppression as can be seen from Figure 7-30 and Figure 7-32. The maps of Hilbert based and LD based have more dark strips, which are undesirable interference than the maps of the QD based and MODWPT based algorithms. The QD based algorithm using a compromised kernel can reduce interferences both in time and frequency. For the MODWPT based algorithm, interference is reduced in two ways. First, the basis function of the wavelet decreases quickly to zero at both ends, which is the same characteristics of the impulse signal. Secondly, the denoising function removes small wavelet coefficients, which can further reduce the background noise level.

For the case of bearings with uniform wear out or very small spalls, such as R30 and R31, none of the methods were able to detect the defect frequency components because

the HFRT depends on a higher impulse to noise ratio as presented in the synthetic signal simulations.

It is also found that the components of bearing defect frequency disappeared toward the end of running (see Figure 7-30, Figure 7-31 and Figure 7-35). It was found the bearings became severely damaged by the end of the run. The spectra generated by different flaws might be canceled out or amplified [102]. The signal resulting from the addition of all the impulse trains appears more like a random noise signal rather than a impulse train. Therefore, the HFRT algorithms are no longer able to extract the defect frequency when the bearing has many flaws.

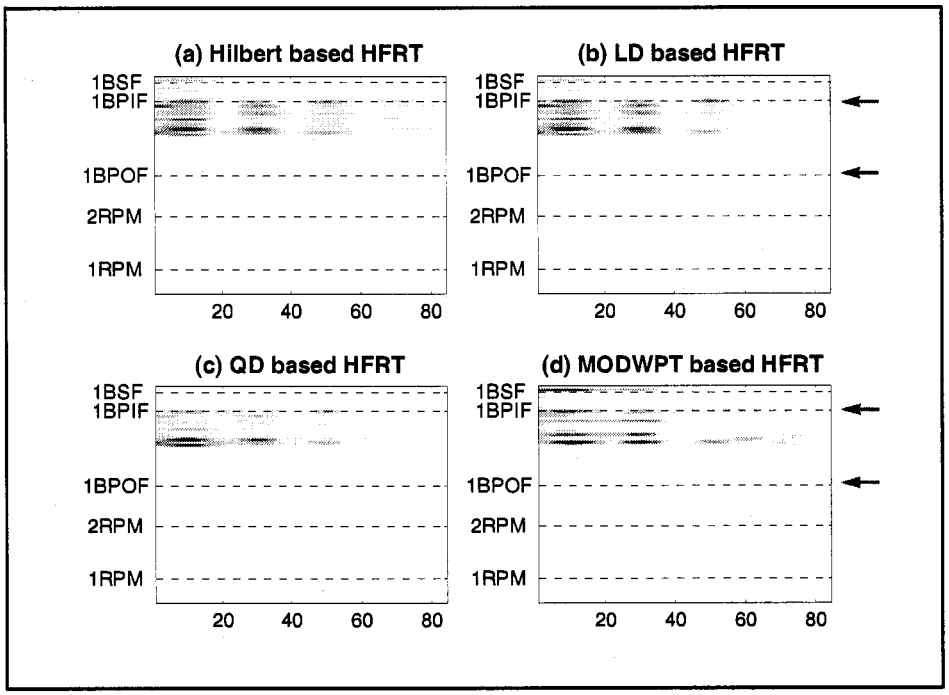


Figure 7-30. Trending maps of envelope spectra in linear scale for bearing R24 from 10Hz to 100Hz.

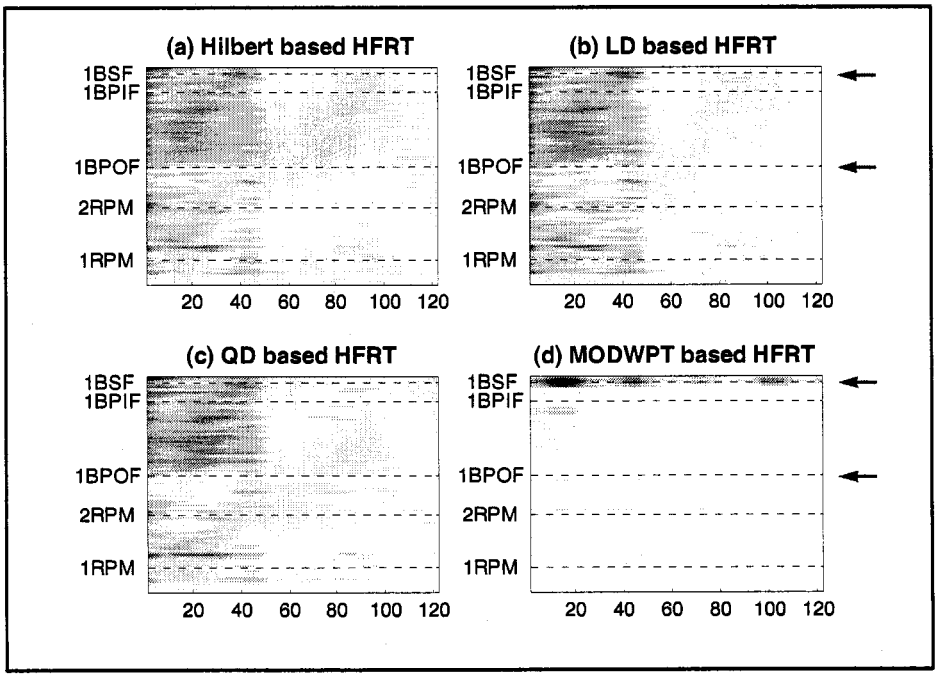


Figure 7-31. Trending maps of envelope spectra in linear scale for bearing R25 from 10Hz to 100Hz.

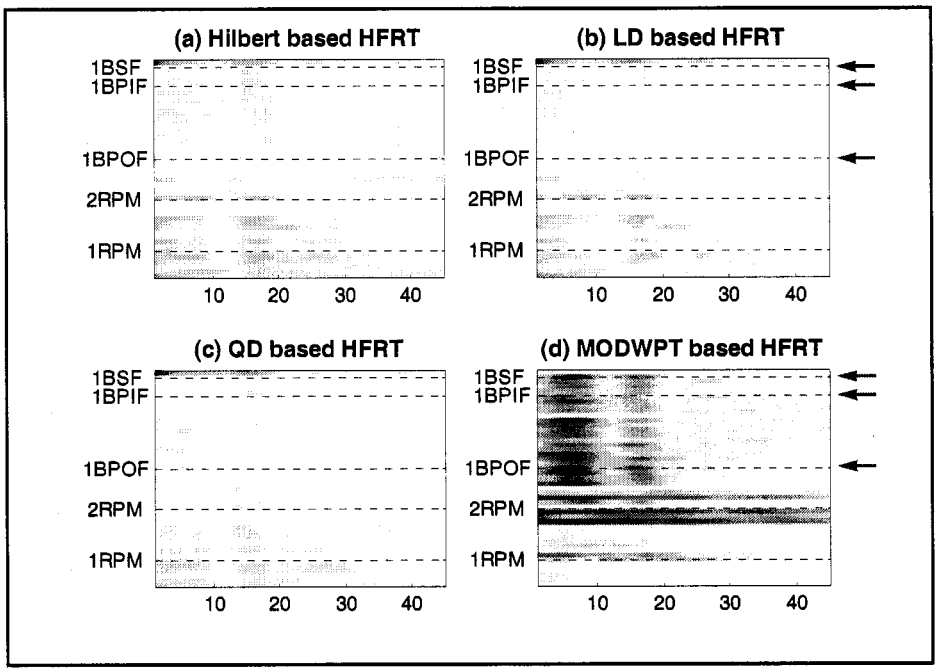


Figure 7-32. Trending maps of envelope spectra in linear scale for bearing R26 from 10Hz to 100Hz.

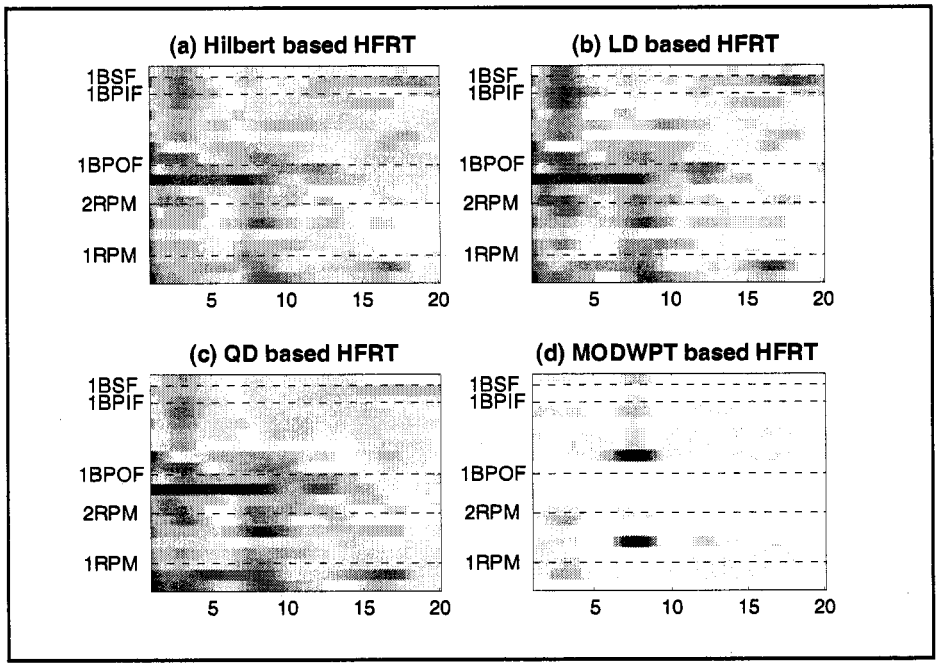


Figure 7-33. Trending maps of envelope spectra in linear scale for bearing R29 from 10Hz to 100Hz.

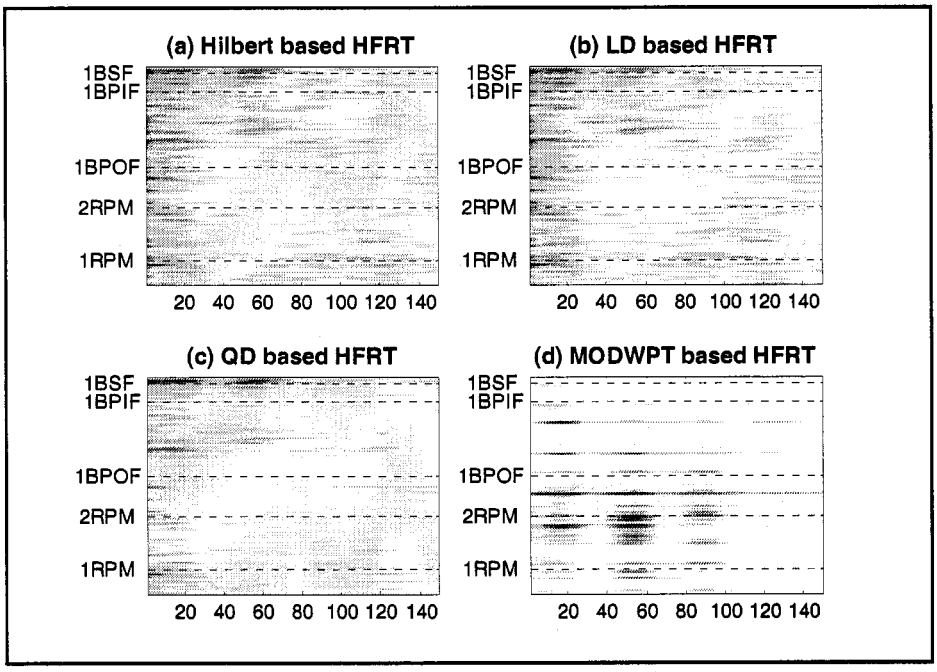


Figure 7-34. Trending maps of envelope spectra in linear scale for bearing R30 from 10Hz to 100Hz.

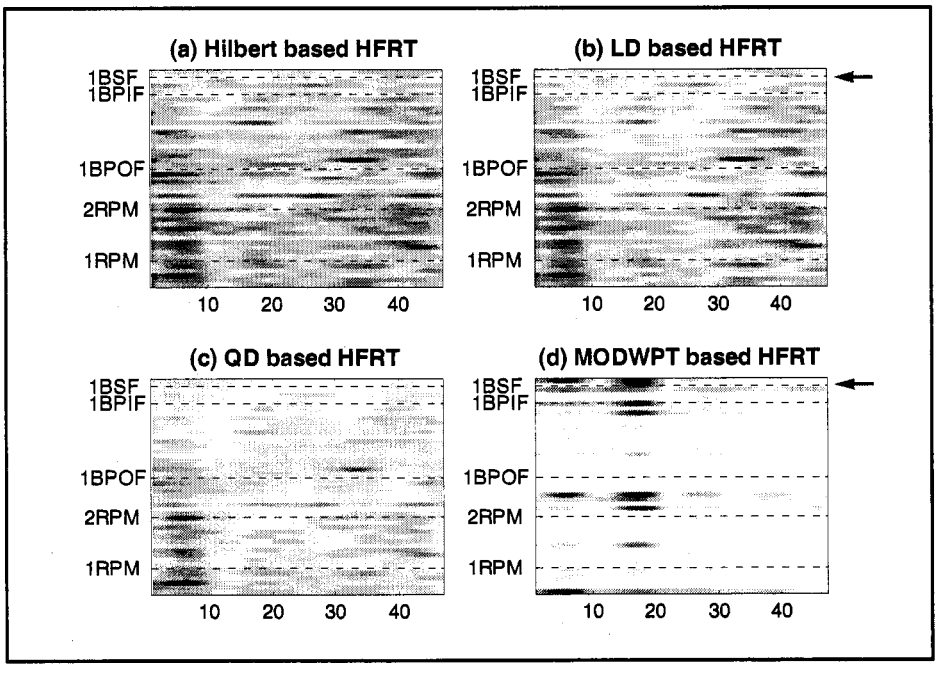


Figure 7-35. Trending maps of envelope spectra in linear scale for bearing R31 from 10Hz to 100Hz.

7.3 Summary

In this chapter, we investigated the bearing defect detection from experimentally measured vibration data using the traditional time domain and frequency domain methods as well as the proposed wavelet based and energy detector based methods. We started by using data from bearings with artificially manufactured. We also used bearings data from run-to-fail bearing tests using a simple testing rig. The following conclusions can be drawn from the results:

1. In general, the HFRT algorithm is a better technique for early bearing detection than the traditional methods. When the bearing is severely damaged with multiple defects, the HFRT algorithm is no longer able to identify the component by defect frequency.
2. The automatic resonant packet selection algorithm based on the MODWPT using kurtosis and changes of RMS can successfully identify the optimal frequency range with an optimal signal to noise ratio for HFRT algorithms on artificially defected bearing data.
3. The QD based and MODWPT based algorithms outperform the Hilbert based and LD based algorithms in terms of background noise suppression. The QD based algorithm uses a compromised kernel to reduce interferences both in time and frequency. The MODWPT based algorithm reduces interferences in two ways. First is the basis function of wavelet decreases quickly toward to zero at both ends which is the same characteristic of the impulse signal. Secondly, the denoising function removes small wavelet coefficients and can further reduce the background noise level.
4. In detecting damage to rolling bearings, there are appreciable differences between artificially defecting test data and run-to-fail test data. An artificially defective bearing generates distinctive signals of various kinds which can be conveniently

sampled and examined to detect failure. On the other hand, there are usually other sources of conflicting signals and obstacles to bearings in a run-to-fail testing. It is more difficult to separate the signal from background noise for a run-to-fail testing.

5. The progression of bearing degradation can be different for a run-to-fail testing. The HFRT algorithm is best for single defect with fewer noise and other vibration sources. For complicated machinery, the HFRT algorithm may fail to detect bearing failure.

Chapter 8 Harmonic Analysis

In Section 6.2, we have shown that the performance of all HFRT algorithms deteriorate at low SNR level. When the SNR level decreased to 0.1, the individual pulses can not be identified from the envelope of the signal (see Figure 6-8). In this Chapter, we introduce a method called harmonic analysis which extend the applications of the HFRT methods to lower SNR levels.

In Section 8.1, we will present the background of harmonic modeling as introduced by McAulay and Quatieri [125] and a Least Squares Harmonic (LSH) algorithm developed by Abu-Shikhah and Deriche [126]. The LSH algorithm can be a pre-processing tool to extract impulses buried in the high level of background noise. In Section 8.2 and 8.3, we compare the results of HFRT algorithms with and without the LSH pre-processing for synthetic data and experimentally measured data respectively. Section 8.4 summarizes the advantages and disadvantages of harmonic analysis for bearing defect detection.

8.1 Harmonic Modeling

Harmonic modeling is also know as a sinusoidal representation, was first introduced by McAulay and Quatieri to model a signal by a finite combination of sinusoidal components [125]. Harmonic analysis has been widely applied in speech coding, compression, enhancement, synthesis, epoch detection, and pitch estimation [126].

The idea of the harmonic modeling is to represent a signal by two parts: a quasi-periodic (harmonic component) and a non-periodic part (noise component). The discrete form can be expressed as

$$s[k] = h[k] + n[k], \tag{8-1}$$

where $h[k]$ is the harmonic component and $n[k]$ is the noise component.

For most application, the key of harmonic modeling is to find the parameters of harmonic components, e.g., their amplitude, frequency and phase. Abu-Shikhah and Deriche develop a Least Squares Harmonic (LSH) algorithm for harmonic analysis of speech [127]. Their technique shows a very accurate harmonic representation at very low SNR. The algorithm starts by assuming a length N signal with M harmonic components, i.e.,

$$h[k] = \sum_{l=1}^{M} C_l \cos(l\omega_0 k + \phi_l) \qquad k = 0, 1, \ldots, N-1, \tag{8-2}$$

where C_l, ω_l and ϕ_l are the amplitude, frequency and phase angle of each sinusoid. The above equation can be expanded and rearranged as a combination of sine and cosine functions as

$$\begin{aligned} h[k] &= \sum_{l=1}^{M} C_l\{\cos(l\omega_0[k])\cos(\phi_l) - \sin(l\omega_0[k])\sin(\phi_l)\} \\ &= \sum_{l=1}^{M} A_l \cos(l\omega_0[k]) - B_l \sin(l\omega_0[k]) \end{aligned}, \tag{8-3}$$

where, $A_l = C_l\cos(\phi_l)$ and $B_l = C_l\sin(\phi_l)$. The mean square error (MSE) between the signal and its harmonics can be obtained by

$$\text{MSE} = E = \frac{1}{N}\sum_{k=0}^{N-1} [(S[k]) - (h[k])]^2, \tag{8-4}$$

Therefore, for a known frequency, ω_0, the minimum MSE can be found by setting

$$\frac{\partial E}{\partial A_j} = 0, \quad \text{and} \quad \frac{\partial E}{\partial B_j} = 0, \quad \text{for } j = 1, 2, \ldots, P. \tag{8-5}$$

The above equations can be expressed in a matrix form

$$\begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{R} \\ \boldsymbol{S} & \boldsymbol{T} \end{bmatrix} \begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{B} \end{bmatrix}, \tag{8-6}$$

where A and B are Px1 unknown vectors to be estimated and where

$$\boldsymbol{Q}(i,j) = \sum_k \cos(i\omega_o k)\cos(j\omega_o k); \tag{8-7}$$

$$\boldsymbol{R}(i,j) = -\sum_k \sin(i\omega_o k)\cos(j\omega_o k); \tag{8-8}$$

$$\boldsymbol{S}(i,j) = \sum_k \cos(i\omega_o k)\sin(j\omega_o k); \tag{8-9}$$

$$\boldsymbol{T}(i,j) = -\sum_k \sin(i\omega_o k)\sin(j\omega_o k); \tag{8-10}$$

$$\boldsymbol{Y}_1(j) = \sum_k s(k)\cos(j\omega_o k); \tag{8-11}$$

$$\boldsymbol{Y}_2(j) = \sum_k s(k)\sin(j\omega_o k). \tag{8-12}$$

Therefore, $\boldsymbol{A}$ and $\boldsymbol{B}$ can be obtained by solving the following equations,

$$\boldsymbol{A} = (\boldsymbol{S} - \boldsymbol{T}\boldsymbol{R}^{-1}\boldsymbol{Q})^{-1}(\boldsymbol{Y}_2 - \boldsymbol{T}\boldsymbol{R}^{-1}\boldsymbol{Y}_1); \tag{8-13}$$

$$\boldsymbol{B} = \boldsymbol{R}^{-1}(\boldsymbol{Y}_1 - \boldsymbol{Q}\boldsymbol{A}). \tag{8-14}$$

The LSH algorithm can solve for $\boldsymbol{A}$ and $\boldsymbol{B}$ over a frequency range of interest as shown in Figure 8-1 [127].

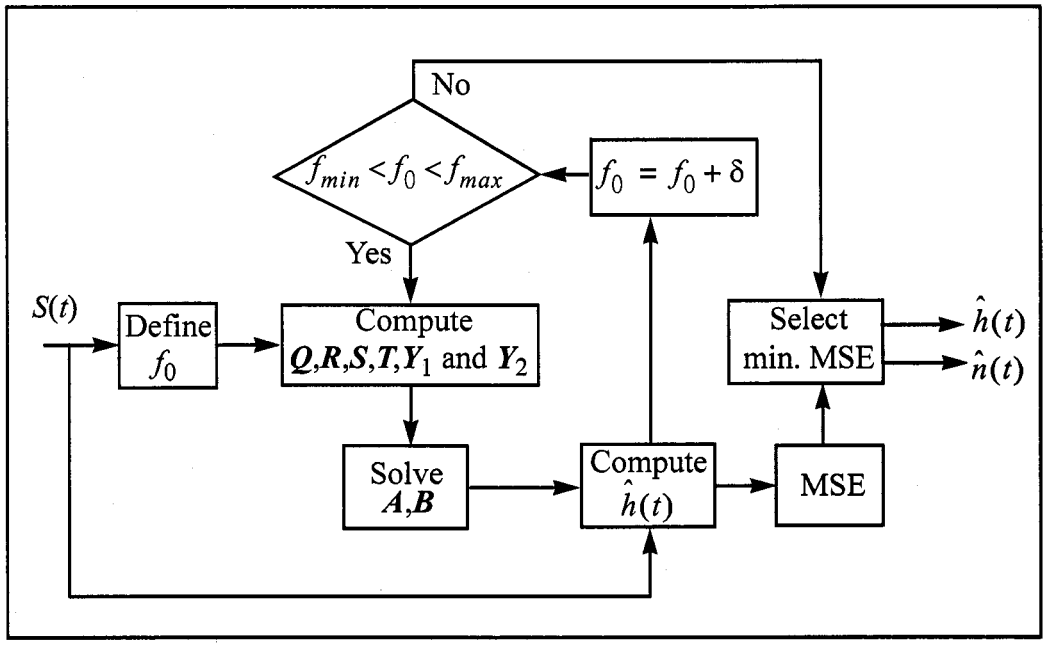


Figure 8-1. The LSH algorithm

8.2 Harmonic Analysis of Synthetic Bearing Data

Since an impulse train can be approximated as a finite number of harmonics with a fundamental frequency at the bearing defect frequency, we can use harmonic analysis as a pre-processing tool to extract the impulse component generated by bearing defects from a noisy signal. In this section, we present the harmonic analysis results using the synthetic bearing data as described in Section 6.1. The sampling frequency for all the analysis in this section is 20k Hz.

8.2.1 Extraction of periodic signal

We start with a simple case to show the performance of LSH technique by extracting a periodic component from a noisy signal. Assume the signal of interest is a periodic signal with 3 harmonics at a fundamental frequency of 500 Hz generated by

$$h[k] = \cos(f \cdot 2\pi k) + 0.75\cos(2f \cdot 2\pi k) + 0.5\cos(3f \cdot 2\pi k). \qquad (8\text{-}15)$$

When this periodic signal is contaminated with a white noise at a SNR level of 0.1, the harmonic component can not be recognized from the time history as shown in Figure 8-2(a). Suppose we only know the approximate range of this fundamental frequency is between 490 to 510 Hz and we use LSH algorithm to search the fundamental frequency for an increment of 0.5Hz. The true fundamental frequency was successfully estimated by LSH algorithm using 3 harmonics. Figure 8-2 shows the synthetic signal extracted by LSH techniques and its spectrum.

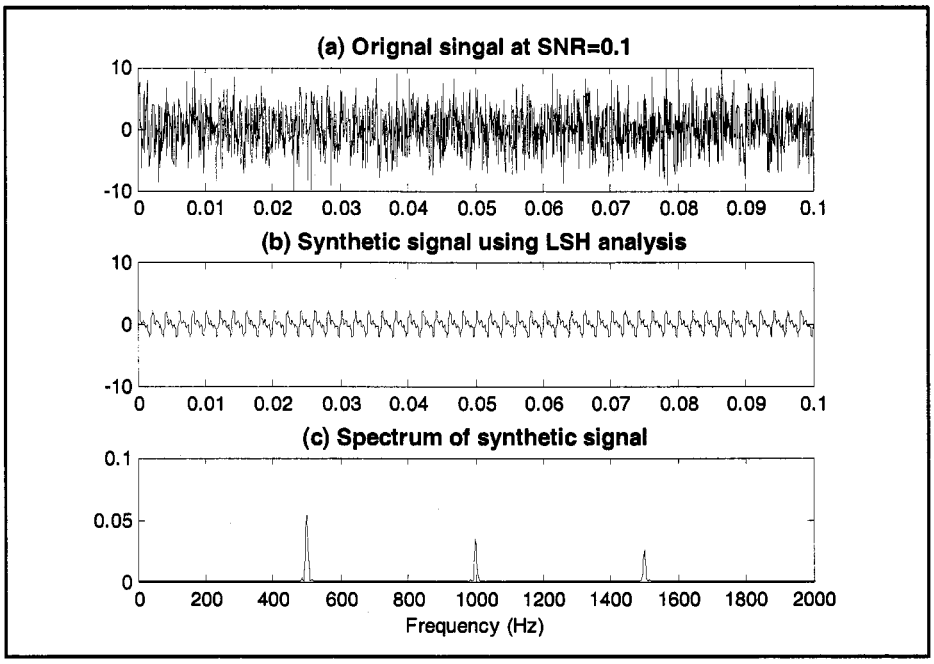


Figure 8-2. Results of LSH analysis for a periodic single with SNR level at 0.1.

An accurate estimation of fundamental frequency is important for bearing defect detection because the bearing defect frequencies obtained from equation 3-8, 3-9 and 3-10 are approximations. The true defect frequency can be deviated from the derived frequency due to sliding or spinning motion of the ball [78].

8.2.2 Extraction of impulse train generated by single defect

Using the same model that we tested in Section 6.2.2.3, the response of a defective rolling element bearing is an impulse train with a repeat ion frequency at bearing defect frequency. Since an impulse signal can be approximated by a finite number of harmonics, we can use the harmonic analysis to extract bearing defect information from a noisy signal. Let the resonant frequency and the damping factor of the system be 6000 Hz and 3000 respectively. When the SNR level is as low as 0.1, none of the HFRT algorithms can extract the envelope of the fault related impulse train (see Figure 6-8) and their envelope spectra interfere with random peaks (see Figure 6-9). The outer race defect frequency at 105 Hz is not obvious using the Hilbert transform and energy detector based HFRT algorithms. However, the impulse train can be extracted after pre-processing the original defect signal (top row of Figure 8-3) with LSH analysis. Searching from 104.5Hz to 105.5Hz, the true defect frequency was obtained at 105Hz with 0.1Hz increment. Figure 8-3 also shows the HFRT results using the synthetic signal from the LSH analysis using 60 harmonics. Since the impulse component of interest was extracted from the noisy signal, the performance of the HFRT methods are greatly improved. All the HFRT algorithms are able to detect the fault related frequency as shown in Figure 8-4. If we further reduce the SNR level to 0.01, the smoothed envelope results and their spectra look like random as shown in Figure 8-5 and Figure 8-6. However, all the HFRT algorithms are able to detect bearing defect frequency after pre-processing with LSH analysis as illustrated in Figure 8-7 and Figure 8-8.

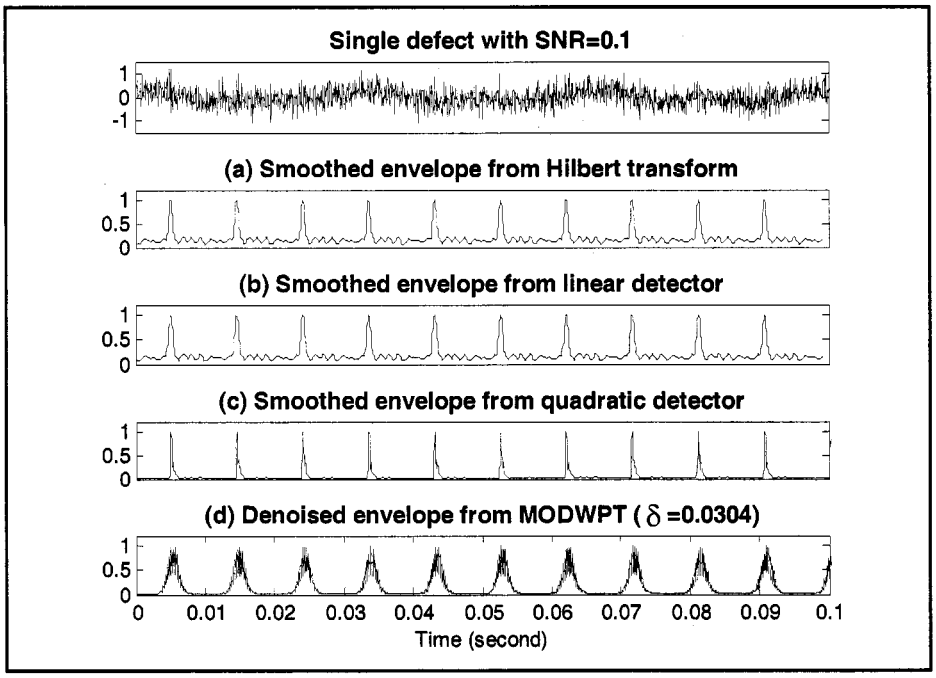


Figure 8-3. HFRT results of synthetic data with single defect at SNR=0.1 after pre-processing with LSH analysis.

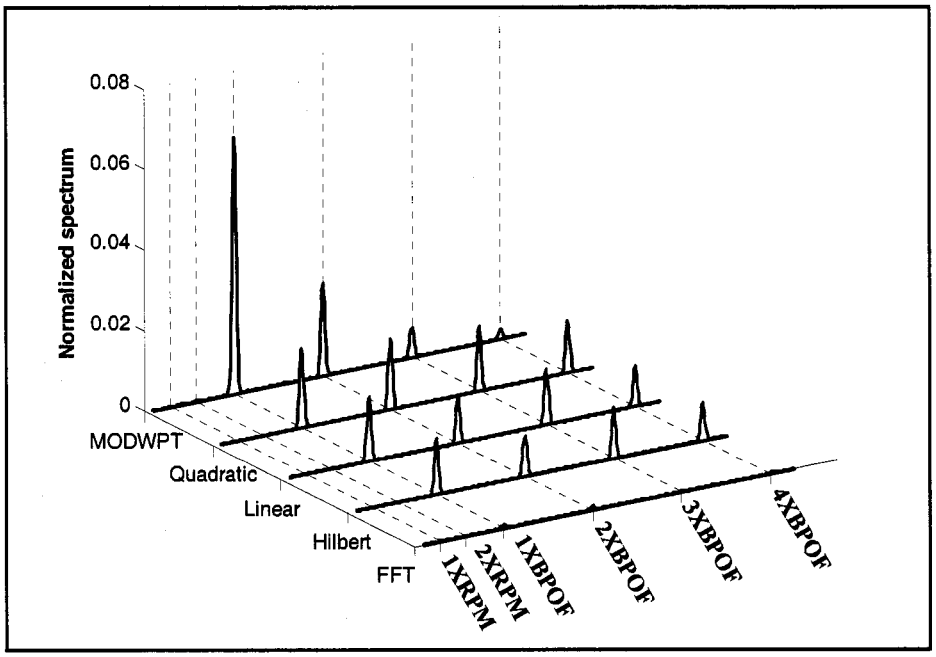


Figure 8-4. Cascade spectra of Figure 8-3 with SNR=0.1.

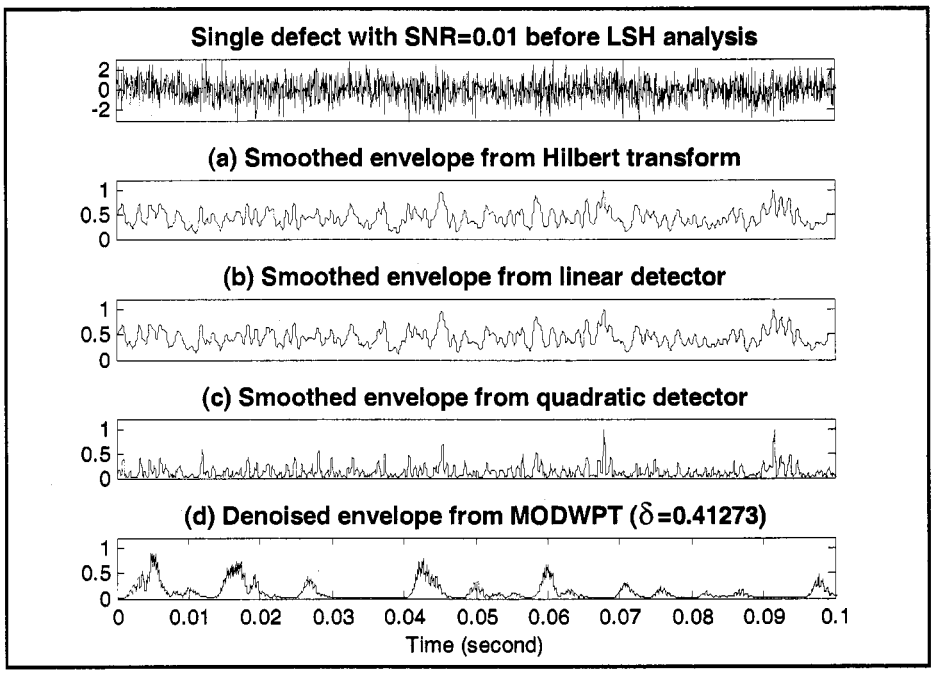


Figure 8-5. HFRT results of synthetic data with single defect at SNR=0.01 before pre-processing with LSH analysis.

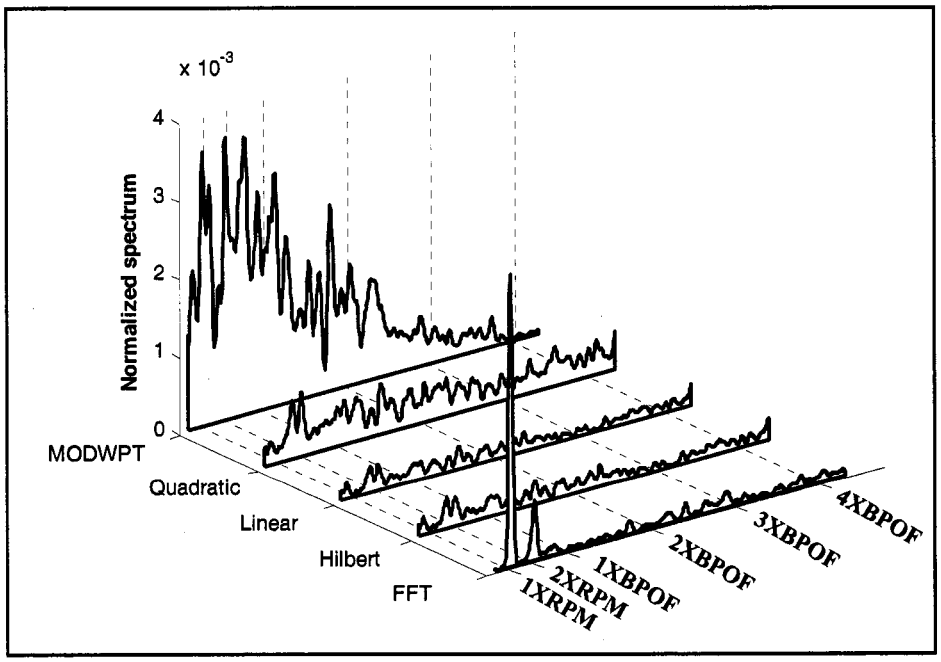


Figure 8-6. Cascade spectra of Figure 8-5 with SNR=0.01.

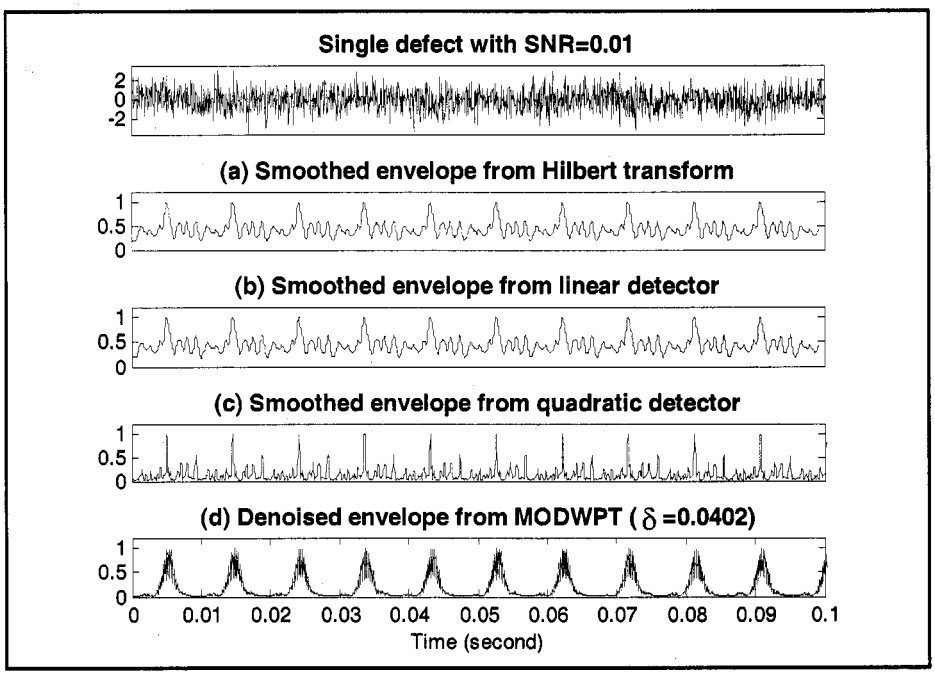


Figure 8-7. HFRT results of synthetic data with single defect at SNR=0.01 after pre-processing with LSH analysis.

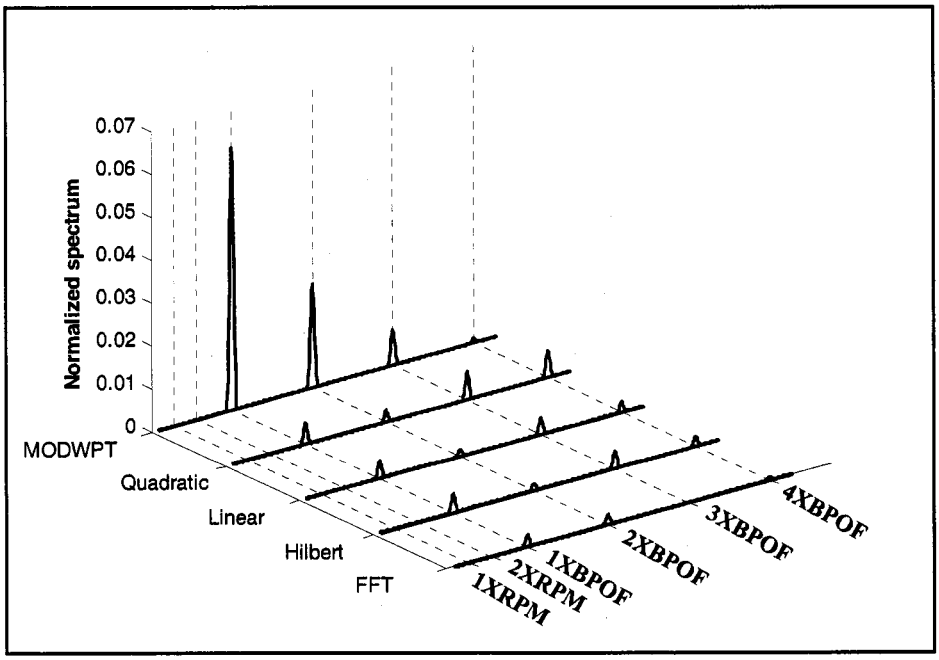


Figure 8-8. Cascade spectra of Figure 8-7 with SNR=0.01.

8.3 Harmonic Analysis of Recorded Bearing Data

In this section, we demonstrate the performance of LSH analysis on the real bearing data using both the seeded defect and the run-to-failure time series as described in Chapter 7.

8.3.1 Analysis results of artificially defected bearing data

This section uses bearing data with artificially created defects at the inner race to investigate how harmonic analysis can improve the sensitivity of HFRT algorithms for detecting impulse components.

As before, the sizes of defects are 0.007", 0.014" and 0.021". From the geometry and the running speed of the shaft, the approximate inner race defect frequency is 162 Hz (see Table 7-1). If we let the search frequency range from 161.5Hz to 162.5Hz and the searching increment to be equal to 0.1Hz, the estimated defect frequencies for 0.007", 0.014" and 0.021" defect size become 161.7Hz, 162.2Hz and 162.3Hz respectively. These frequencies correspond to a minimum value of the MSE as indicated on Figure 8-1. The variation of inner race defect frequencies with respect to defect size implies that the effective bearing defect frequencies are a function of bearing size.

For each defect size, we pre-process the bearing data with the LSH algorithm using 35 harmonics at the estimated bearing defect frequency. Figure 8-9 to Figure 8-14 are the results of the smoothed envelope and their spectra for each defect size. Figure 8-10, Figure 8-12 and Figure 8-14 show that the peaks at inner race defect frequency (1XBPIF) and its harmonics (2XBPIF and 3XBPIF) with the LSH analysis are enhanced as compared with results using no LSH analysis(see Section 7.1). In addition, periodic pulses are visible from envelope signal for each defect size case. However, the changes of each pulse amplitude due to the modulation of shaft rotating frequency is lost because the model of harmonic analysis use only one fundamental frequency and their harmonics to synthesis the impulse signal. Therefore, all the pulses have the same shape in the envelope signal as can be seen from Figure 8-9, Figure 8-11 and Figure 8-

13. This observation suggests that for a bearing with multiple defects it is necessary to perform the harmonic analysis at each fault related defect frequency.

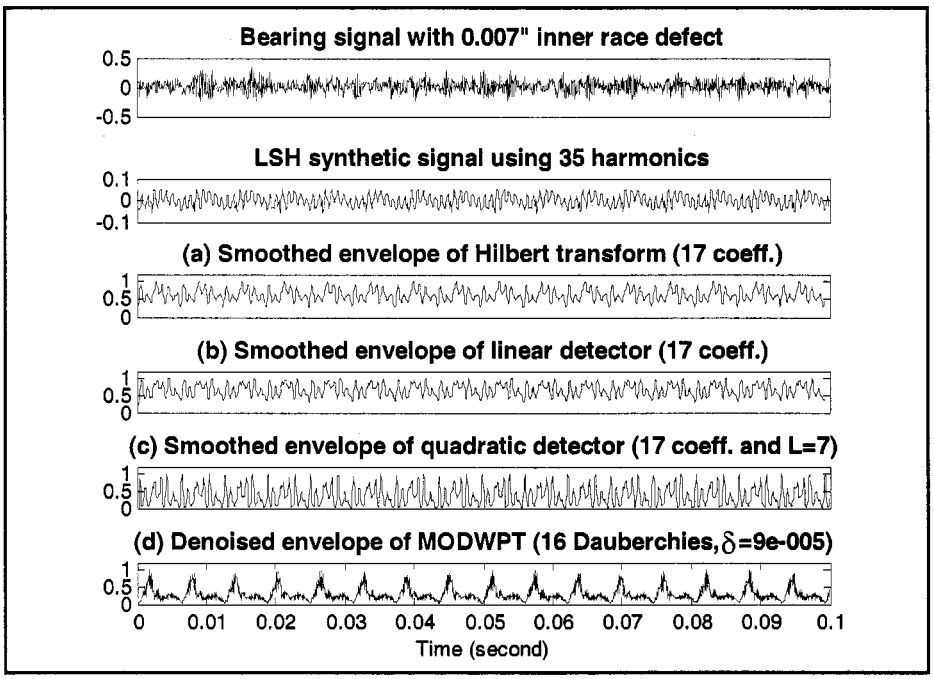


Figure 8-9. Smoothed or denoised envelope of bearing data with 0.007" inner race defect after LSH pre-processing.

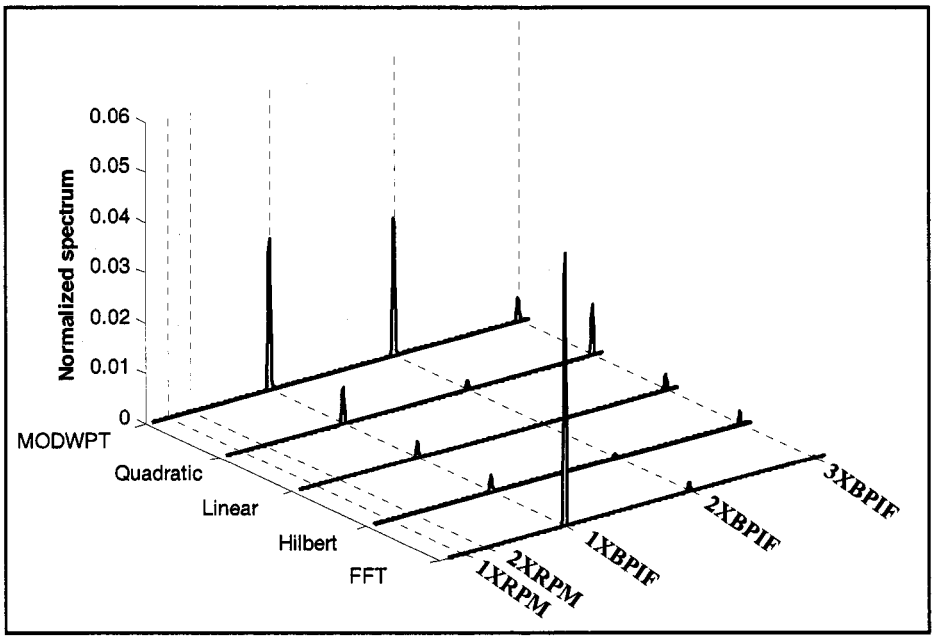


Figure 8-10. Cascade envelope spectra of bearing data with 0.007" inner race defect after LSH pre-processing.

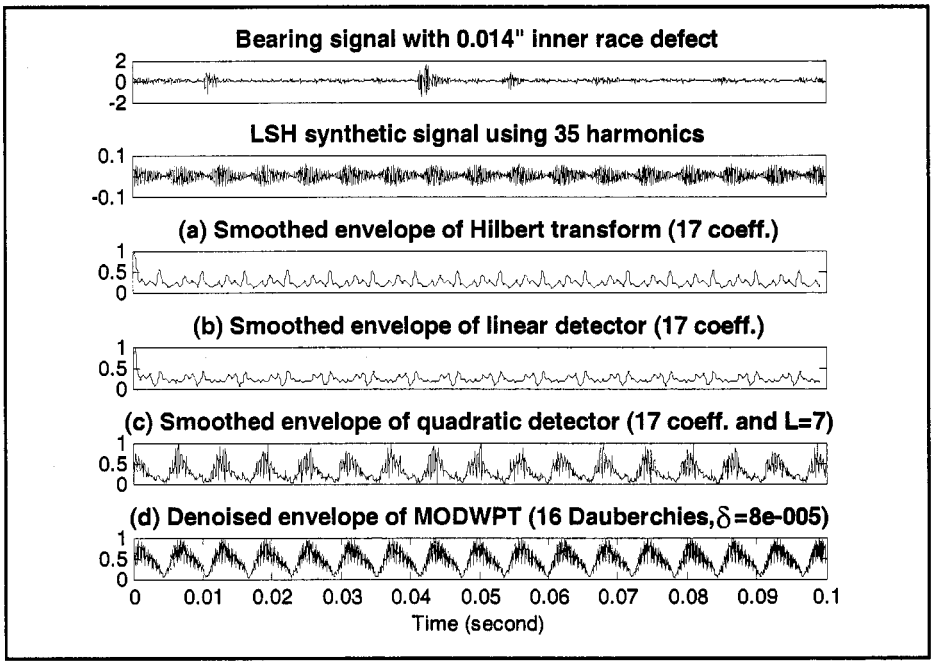


Figure 8-11. Smoothed or denoised envelope of bearing data with 0.014" inner race defect after LSH pre-processing.

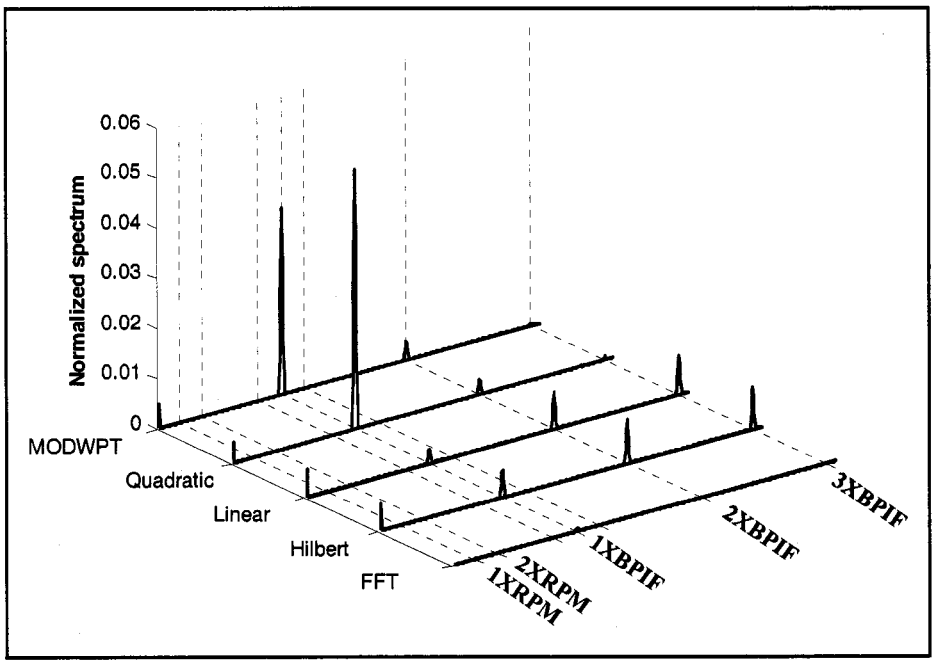


Figure 8-12. Cascade envelope spectra of bearing data with 0.014" inner race defect after LSH pre-processing.

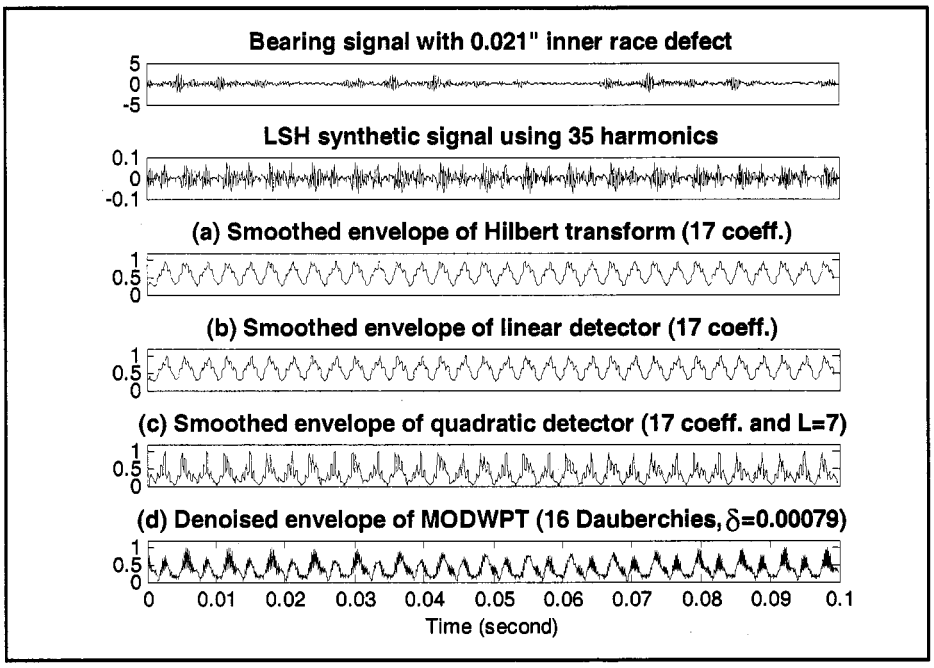


Figure 8-13. Smoothed or denoised envelope of bearing data with 0.021" inner race defect after LSH pre-processing.

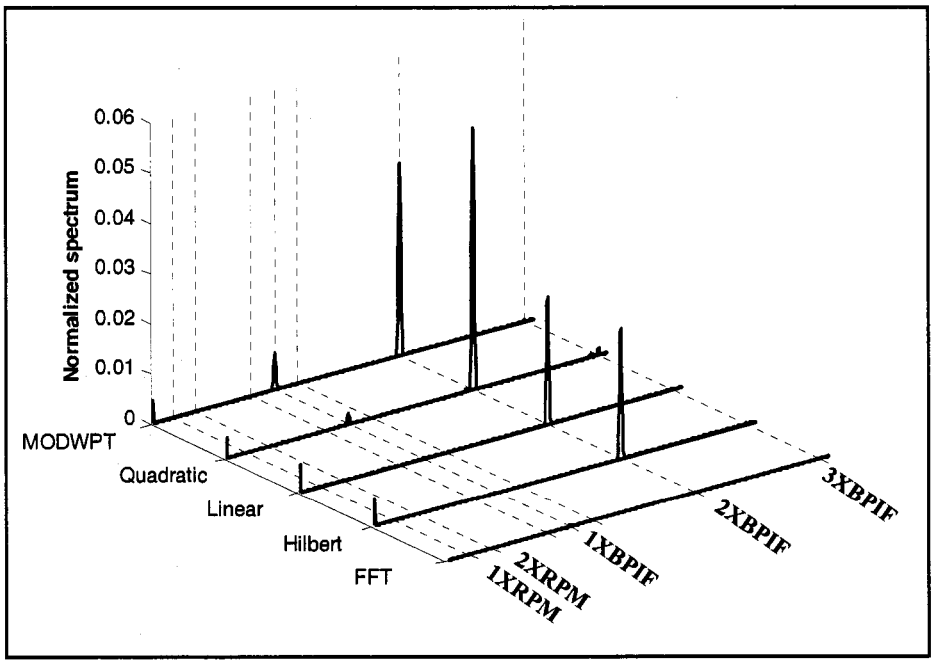


Figure 8-14. Cascade envelope spectra of bearing data with 0.021" inner race defect after LSH pre-processing.

8.3.2 Analysis results of run-to-fail bearing data

Since the real bearing running from normal to fail can have defects initiated from any of the bearing components, it is necessary to track the estimated signal from the LSH analysis at each of the bearing defect frequency. In this section, we use testing R25 to show how a harmonic analysis can be implemented on a run-to-failure bearing data.

The first step of using LSH analysis for run-to-failure bearing data is to search the fundamental defect frequencies for the ball, the inner race and the outer race as shown in Figure 8-1. These estimated bearing defect frequencies and the parameters to obtain these frequencies are listed in Table 8.1. Figure 8-15 is the trending maps of the bearing defect frequencies. In general, the estimated frequencies are very close to stationary with a standard deviation from 0.22 Hz to 0.31 Hz.

Table 8.1 Bearing defect frequencies and parameters for using LSH

Component	Ball	Inner Race	Outer Race
Derived defect frequency from equation 3-8, 3-9 and 3-10	77.6167 Hz	71.5955 Hz	47.4045 Hz
Lower bond searching frequency (f_{min})	77.1167 Hz	71.0955 Hz	46.9045 Hz
Upper bond searching frequency (f_{max})	78.1167 Hz	72.0955 Hz	47.9045 Hz
Frequency increment (δ)	0.1 Hz	0.1 Hz	0.1 Hz
Number of harmonics	100	100	100
Mean of the trend of the defect frequency using LSH (μ)	77.39 Hz	71.48 Hz	47.11 Hz
Standard deviation of the trend of the defect frequency using LSH (σ)	0.22 Hz	0.28 Hz	0.31 Hz

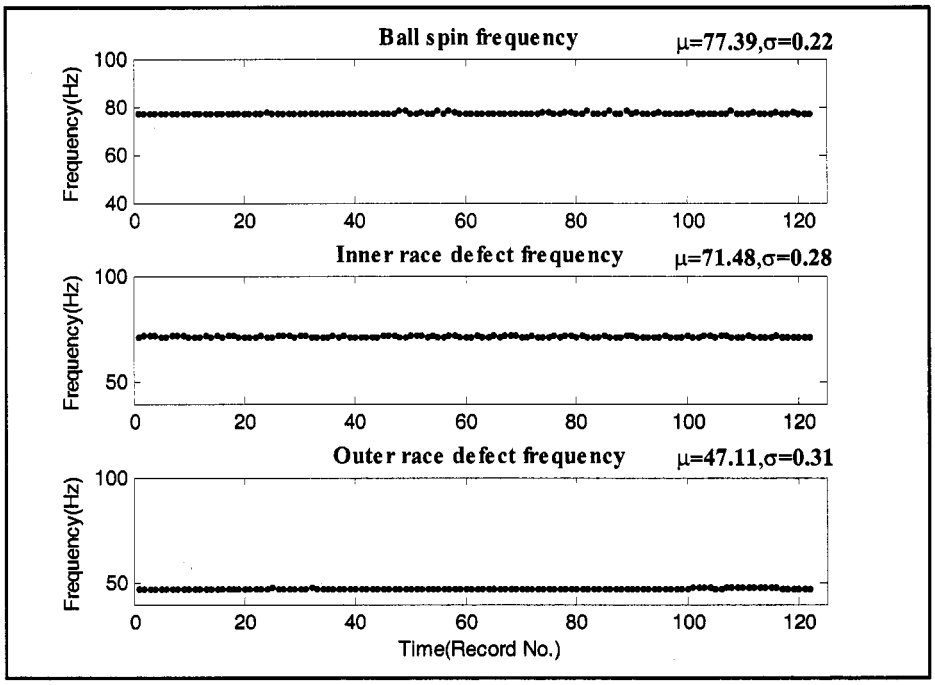


Figure 8-15. The trend of bearing defect frequency for R25.

After we determine the defect frequency for each bearing component, we can extract their synthetic signals for further analysis. Figure 8-15 shows the trend of the sum of the 1X~3X harmonics spectra energy at each bearing component. The energy changes at the outer race is much more noticeable than the inner race and ball. The SEM pictures (see Figure A-2) show that the bearing R25 has defects at both the outer race and the ball. As presented in Chapter 7, none of these defects were able to be detected by using the HFRT methods except the ball defect was detected using the MODWPT based HFRT as shown in Figure 7-31. After pre-processing with the LSH analysis, the energy changes of the sum of the first 3 harmonics at outer race defect is discern after running for 230 minutes. Figure 8-17 shows the trend of the sum of the 1X~3X harmonics spectra energy for the test R30. After pre-processing, the inner race defect is the first detected flaw after running for 990 minutes.

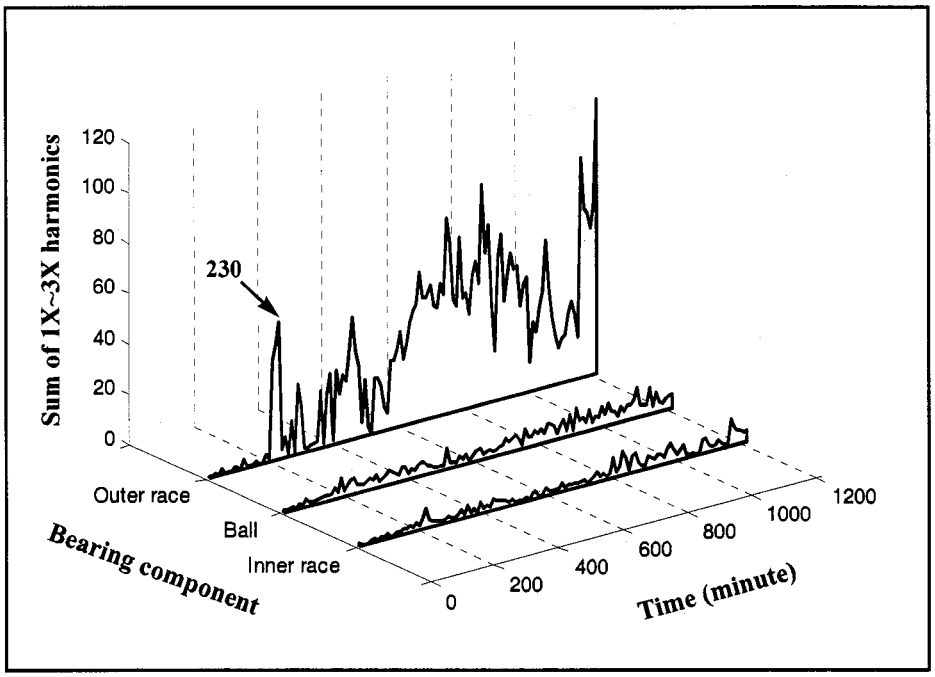


Figure 8-16. Trending of the sum of the 1X~3X harmonics spectra at defect frequencies for R25.

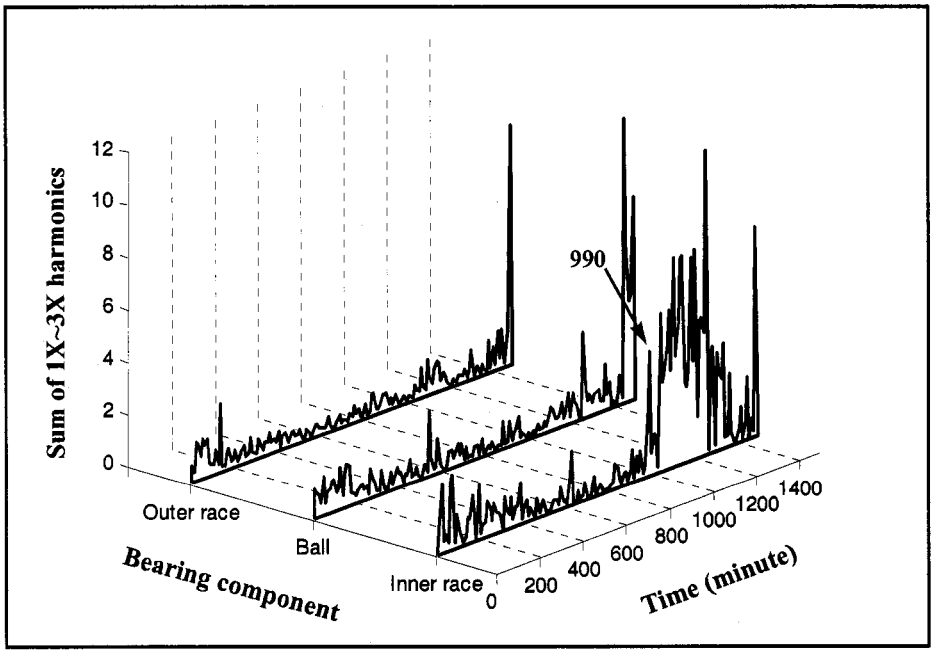


Figure 8-17. Trending of the sum of the 1X~3X harmonics spectra at defect frequencies for R30.

8.4 Summary

In this chapter, we investigated LSH analysis as a pre-processing tool for bearing defect detection. Comparing the results with and without the LSH analysis, we make the following observations:

1. For the synthetic one defect bearing signal, the impulse component can be extracted form a very low SNR signal using LSH analysis.
2. For the seeded defect and run-to-failure bearing data, the sensitivity of the HFRT algorithms to the bearing defect are improved after pre-processing with LSH analysis.
3. For the run-to-fail bearing data, the change of the spectra energy at the bearing defect frequencies can detect bearing defects earlier than the HFRT algorithms after pre-processing with the LSH analysis.
4. For the real bearing data, harmonic analysis must be applied at different defect frequencies in order to track the condition of each bearing component. The application of harmonic analysis for a bearing condition monitoring system is possible in real time when the period of sampling is long enough to allow the search for the defect frequency.

Chapter 9 Conclusions and Future Work

9.1 Conclusions

We have developed two new HFRT algorithms which are more sensitive to impulse signals than traditional HFRT algorithm based on the Hilbert transform. The first one is based on the shift invariant wavelet packet transform (MODWPT) and the second one is based on a quadratic energy detector (QD).

The MODWPT based HFRT takes advantage of wavelet packet signal processing with denoising process to efficiently demodulate fault related frequency components from sampled bearing spectra for the purpose of early bearing defect detection. The quadratic energy detector based HFRT possesses the properties of frequency interference elimination, finite time support, and independent time-frequency resolution which gives the best rise time and pulse localization tracking for impulse detection.

In a series of extensive analysis on real and synthetic data, it was shown that the HFRT algorithms perform as well or better than traditional time domain techniques (such as the peak value and kurtosis) and frequency domain methods using FFT for early defect detection. The following observations can be drawn from the experiments:

1. For the testing results using synthetic failure data, the MODWPT based HFRT outperformed other HFRT algorithms when SNR was below 0.1. The MODWPT based HFRT was the only method of these method that was able to detect two defects at different bearing components using the envelope spectrum.
2. For the recorded data from seeded defect bearings, all HFRT algorithms were able to detect bearings with one small inner race defect. However, only the QD and MODWPT based algorithms were able to detect large inner race defects.

3. For the run-to-fail bearing data, defects caused by spalls (R24, R25, R26, and R31) were partially detected using HFRT. The MODWPT based algorithm indicates failure earlier than other HFRT algorithms. When the bearing failure was due to uniform wear out (R29 and R30), none of the HFRT algorithms were able to detect bearing failure.

Until now, almost all the studies used either synthetic data from an impulse train model or bearings with artificially created defects for early bearing defect detection. Our study shows that there are appreciable differences between artificially defect bearing test data and run-to-fail bearing test data. Run-to-fail bearing test data is complicated because many vibrating sources can developed during the development of bearing failure. In the run-to-fail testing, there are various bearing failure mechanisms. Therefore, it is not possible to use only one technique to detect for all possible failure modes. In our study, the HFRT algorithms proved to be a useful tool for detecting early bearing defect when single or few defects developed during operation. Unfortunately, we can not use HFRT alone because we do not have enough evidence to claim that the absolute level of a demodulated signal provides an indication of bearing fault severity. The reason is that demodulation looks for the rate at which the continuous ringing of the bearing housing gets louder. If this frequency is essentially random due to numerous faults, then the demodulated signal also appears to be noise. In fact, as the bearing fault gets worse, modulation of defect frequency can actually disappear. In addition, for different type of bearing failure, such as the case of uniform wear out, a simple RMS measurement can be a reliable indicator. Hence, it is important to choose a monitoring technique to suit the failure modes of greatest interest and importance. With complex machines, it is better to use a combination of techniques, and extensive analysis and interpretation of the results.

The automatic resonant packet selection algorithm performs well in the recorded bearing data, which indicates that this method can be implemented for a real time monitoring applications. From the result of using seeded defect data, it is found our resonant

packet selection algorithm gives a fair good estimation when the measurement of resonant frequency due to bearing defects is not easy or possible.

Pre-processing with LSH analysis can extract the impulse component form a very low SNR signal. The results from recorded bearing data show that the sensitivity of the HFRT algorithms to the bearing defect is improved after pre-processing with LSH analysis. However, the modulating information of impulse peaks is lost because the harmonic modeling uses only one fundamental frequency to extract the synthetic signal. In addition, LSH analysis is computational expensive for real application.

In summary, the HFRT algorithms developed in this thesis, and in particular the wavelet packet transformation based and quadratic energy detector based methods perform very well for early detection of bearing defects with few spalls. Using wavelet packet signal processing techniques allowed this algorithm to be implemented very efficiently. Manufactured defect data should be used carefully because it may not model real failure mode properly.

9.2 Future Work

The following are a number of possible improvements that are yet to be investigated after this study:

- Spalling has been detected somewhat earlier in test rigs. There are clearly uncertainties which preclude accurate prediction. One of the purposes of a run-to-fail bearing testing is to find out the progression of bearing failure. However, our testing rig can only provide the end results of the bearing failure. A revision of the testing rig with the capability to examine the bearing defect during the testing can give a better understanding of the failure mechanism and their relation to the vibration signal. Most bearing can be considered to have reached the end of their useful life when spalling can be detected.

- Analysis of bearings with multiple defects is complicated. A simple impact model no longer suited for large number of spalls. Demodulation looks for the rate at which bearing defects hit by rolling elements. When there are multiple defects, this frequency is essentially random, then, the demodulated signal from HFRT also appears to be noise. A model based on statistical observation may be more appropriate for multiple defects case. More testing is needed to establish a statistical model.

- Although the resonant packet from our automation resonance frequency band selection algorithm agrees with our analytic simulation, it might not be the optimal frequency band for a complex machine. The actual resonant frequencies can be estimated with impact testing. Even though the measured resonance frequency using impact testing deviates from the actual resonance frequency, estimation from impact testing can be very close to the actual number.

- Currently, there is no a standard to select the wavelet family for defect detection and to determine the level of decomposition used in the wavelet packet transformation [122]. We select Dauberchies wavelet family because the similarity of the wavelet function and the impact response function. It is needed to establish general rules for wavelet based method using real bearing data.

Bibliography

[1] John Moubray, *Reliability-centered maintenance*, Industrial Press Inc., New York, 1992.

[2] Computational Systems Incorporated, "Reliability based maintenance", 1992

[3] Irem Y. Tumer, Foundations of condition monitoring for manufacturing and design, Ph.D. dissertation, The University of Texas at Austin, 1998.

[4] Dave Kelly, "The fundamentals of a successful vibration monitoring/analysis program", P/PM Technology, Vol.12, No.5, pp. 30-33, Oct. 1999.

[5] S. Spiewak, "A predictive monitoring and diagnosis system for manufacturing", Annals of the CIRP:manufacturing technology, Vol.40, No.1, pp.401-404, 1991.

[6] John S. Mitchell, *Introduction to machinery analysis and monitoring*, 2nd edition, PennWell Books, Oklahoma, 1993.

[7] D. Dyer & R. M. Stewart, "Detection of rolling element bearing damage by statistical vibration analysis", Journal of Mechanical Design, Vol. 100, pp.229-235, 1978.

[8] A.A. Rush, "Kurtosis, a crystal ball for maintenance engineers", Iron and Steel International, 1979.

[9] C. Pachaud, R. Salvetat & C. Fray, "Crest factor and kurtosis contributions to identify defects inducing periodical impulse forces", Mechanical Systems and Signal Processing, Vol.11, No.6, pp.903-916, 1997.

[10] S. Braun, *Mechanical signature analysis: theory and applications*, Academic Press, London, 1986.

[11] N. Tandon, "A comparison of some vibration parameters for condition monitoring of rolling element bearings", Measurement, Vol.12, No.3, pp.285-289, 1998.

[12] G. Massouros, L. Petrou, P. Drakatos, "Interpretation of the vibration spectrums of vehicle engine by cepstrum analysis", Proceedings of the IMACS European Simulation Meeting on Simulation of Distributed-Parameter and Large-Scale Systems. North-Holland, Amsterdam, Netherlands, pp.355-358, 1980.

[13] Leon Cohen, "Time-frequency distributions-a review", Proceedings of the IEEE, Vol. 77, No. 7, July, 1989.

[14] J. Pitton and L. Atlas, "Discrete-time implementation of the cone-kernel time-frequency representation", IEEE Transactions on Signal Processing, vol. 43, pp. 1996-8, August 1995.

[15] J.M. Dawling, "Application of non-stationary analysis to machinery monitoring", Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 93, pp.59-62, 1993.

[16] H. Oehlmann, D. Brie, M. Tomczak, A. Richard, "A method for analysing gearbox faults using time-frequency representations", Mechanical Systems and Signal Processing, Vol. 11, No. 4, Jul 1997, pp. 529-545.

[17] S. K. Lee, P. R. White, "Higher-order time-frequency analysis and its application to fault detection in rotating machinery", Mechanical Systems and Signal Processing, Vol. 11, No. 4, Jul 1997, pp. 637-650.

[18] P. J. Loughlin, G. D. Bernard, "Cohen-Posch (positive) time-frequency distributions and their application to machine vibration analysis", Mechanical Systems and Signal Processing, Vol. 11, No. 4, pp. 561-576, Jul 1997.

[19] W. J. Staszewski, G. R. Tomlinson, "Local tooth fault detection in gearboxes using a moving window procedure", Mechanical Systems and Signal Processing, Vol. 11, No. 3, pp. 331-350, May 1997.

[20] I. Bucher, D. J. Ewins, "Multidimensional decomposition of time-varying vibration response signals in rotating machinery", Mechanical Systems and Signal Processing, Vol. 11, No. 4, Jul 1997, pp. 577-601.

[21] W. J. Wang, P. D. McFadden, "Early detection of gear failure by vibration analysis i. calculation of the time-frequency distribution", Mechanical Systems and Signal Processing, Vol. 7, No. 3, May 1993, pp. 193-203.

[22] W. J. Wang, P. D. McFadden, "Early detection of gear failure by vibration analysis--ii. interpretation of the time-frequency distribution using image processing techniques", Mechanical Systems and Signal Processing, Vol. 7, No. 3, May 1993, pp. 205-215.

[23] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis", IEEE Transactions on Information Theory, Vol.36, pp.961-1005, 1990.

[24] S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation", IEEE Pattern Anal. and Machine Intell., Vol.11, No.7, pp.674-693, 1989.

[25] O. Rioul & P. Flandrin, "Time-scale energy distributions: a general class extending wavelet transforms,", IEEE Trans. Sig. Proc., July 1992.

[26] M. Misiti, Y. Misiti, G. Oppenheim, J. Poggi, *Wavelet toolbox user's guide*, Version 3, MathWorks, 2004.

[27] D. Boulahbal, M. Farid Golnaraghi, F. Ismail, "Amplitude and phase wavelet maps for the detection of cracks in geared systems", Mechanical Systems and Signal Processing, Vol. 13, No. 3, May 1999, pp. 423-436.

[28] S. T. Lin, P. D. McFadden, "Gear vibration analysis by b-spline wavelet-based linear wavelet transform", Mechanical Systems and Signal Processing, Vol. 11, No. 4, Jul 1997, pp. 603-609.

[29] M. Kotani, Y. Ueda, H. Matsumoto, T. Kanagawa, "Acoustic diagnosis for blower with wavelet transform and neural networks", IEEE International Conference on Neural Networks Proceedings, New York, USA, pp.718-723, 1995.

[30] P. Samuel, D. Pines, "Health monitoring/damage detection of a rotor craft planetary gear train system using piezoelectric sensors", Proceedings of the SPIE, Vol. 3041, pp.44-53, 1997.

[31] L. Ziye, S. Akishita, S. Tokumoto, T. Kato, "Failure diagnosis system by sound signal for automobile engine", Proceedings of the Japan-USA Symposium on Flexible Automation, ASME, New York, USA, pp.427-432, 1996.

[32] R. Coifman & V. Wickerhauser, "Entropy-based algorithms for best-basis selection", IEEE Trans. on Inf. Theory Vol.38, No.2, pp.713-718, 1992.

[33] M.V. Wickerhauser, "Lectures on wavelet packet algorithms", Tech. Report, Department of Math., Yale University, March 1991.

[34] D.B. Percival & A.T. Walden, *Wavelet methods for time series analysis*, Cambridge University Press, 2000.

[35] A.V. Oppenheim, R.W. Schafer and J.R. Buck, *Discrete-time signal processing*, 2nd edition, Prentica Hall International, 1999.

[36] Zhongxing Geng & Liangsheng Qu, "Vibrational diagnosis of machine parts using the wavelet packet technique", British Journal of NDT, pp.11-15.

[37] Ya Wu & R. Du, "Feature extraction and assessment using wavelet packets for monitoring of machining processes", Mechanical Systems and Signal Processing, Vol.10, No.1, pp.29-53, 1996.

[38] R.R. Coifman & M.V. Wickerhauser, "Adapted waveform analysis as a tool for modeling, feature extraction, and denoising", Optical Engineering, Vol 33, No.7, pp.2170-2174, July 1994.

[39] B. Liu & S.F. Ling, "On the selection of informative wavelets for machinery diagnosis", Mechanical Systems and Signal Processing, Vol.13, No. 1, pp.145-162, 1999.

[40] D.L. Dohono and I.M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage", Biometrika, 81, 1994, pp.425-455.

[41] W. J. Stevenson, D.L. Brown, R.W. Rost, and T.A. Grogan, "The use of neural nets in rotor balancing", Proceedings of the 10th International Model Analysis Conference, San Diego, CA, pp54 - 59, Feb. 1992.

[42] Y-S Chiou, M.S. Tavakoli, and Steven Liang, "Bearing fault detection based on multiple signal features using neural network analysis", Proceedings of the 10th International Model Analysis Conference, San Diego, CA, pp60 - 64, Feb. 1992.

[43] T.I. Liu & J.M. Mengel, "Intelligent monitoring of ball bearing conditions", Mechanical Systems and Signal Processing, Vol.6, No.5, pp.419-431, 1992.

[44] H.C. Choe, Y. Wan & A.K. Chan, "Neural pattern identification of railroad wheel-bearing faults from audible acoustic signals: comparison of FFT, CWT, and DWT features", SPIE, Vol. 3078, pp.480-496, 1997.

[45] B. A. Paya, I. I. Esat, M. N. M. Badi, "Artificial neural network based fault diagnostics of rotating machinery using wavelet transforms as a preprocessor", Mechanical Systems and Signal Processing, Vol. 11, No. 5, Sep 1997, pp. 751-765.

[46] J. Strackeljan & D. Behr, "Condition monitoring of rotating machinery using fuzzy pattern recognition algorithm", IMechE, pp.507-516, 1996.

[47] H. Guo, L.B. Jack, A.K. Nandi, "Feature generation using genetic programming with application to fault classification", Systems, Man and Cybernetics, Part B, IEEE Transactions on Vol. 35, No. 1, Feb. 2005, pp.89 - 99.

[48] Muggier, M.,"Effective Predictive Maintenance Can Save Industry Millions of $'s", Proc. 16th Mtg. Vib. Inst., Williamsburg, VA, June 9-11, 1992, pp37-41.

[49] Pardue, Forrest, Ken Piety and Ron Moore, Elements of Reliability Based Maintenance, Knoxville, Computational System, Inc., 1992.

[50] Mathew, J., and Alfredson, R. J., "The condition monitoring of rolling element bearings using vibration analysis", Transactions of the ASME, Journal of Vibration, Acoustics, Stress, and Reliability in Design, V. 106, July 1984, pp.447-453.

[51] Braun, S. "State-of-the-Art Review: MSA-Mechanical Signature Analysis", Journal of Vibration, Acoustics, Stress, and Reliability in Design, Vol 106(1), Jan. 1984, pp.1-3.

[52] C.A.W. Glew, "The effectiveness of Vibration analysis as a maintenance tool", Transactions of The Institute of Marine Engineers, Vol 86, 1974, pp.29-50.

[53] T. Carmody, "The measurement of vibration as a diagnostic tool", Transactions of The Institute of Marine Engineers, Vol 84, 1972, pp.147-159.

[54] Donald E. Bently, "Monitoring rolling element bearings", Orbit, November 1982, pp.2-3.

[55] M.W. Hawman and W.S. Galinaitis, "Acoustic Emission monitoring of rolling element bearings", Ultrasonic Symposium, 1988. Proceedings, IEEE , Vol. 2, 1988, pp. 885-889.

[56] Louis E. Morando, "Bearing condition monitoring by the shock pulse method", Iron and Steel Engineer, Vol.75, No 12, Dec. 1998, pp40-43.

[57] Peter F.J. Burgess, "Antifriction bearing fault detection using envelope detection", Transactions of the In stitute of Professional Engineers New Zealand - Electrical/Mechanical Chemical Engineering Section, Vol. 15, No.2, July 1988, p.77-82.

[58] Y. Li, T.R. Kurfess and S.Y. Liang, "Stochastic prognostics for rolling element bearings", Mechanical Systems and Signal Processing, 14(5), PP747-762, 2000.

[59] Tedric A. Harris, "Rolling Bearing Analysis",3rd Edition, John Wiley & Sons, Inc., 1991.

[60] Y.-F. Wang and P. J. Kootsookos, "Modeling of low shaft speed bearing faults for condition monitoring", Mechanical Syatems and Signal Processing, 12(3), pp415-426, 1998.

[61] T.S. Nisbet, "Rolling Bearings", Published for the Design Council, the British Standards Institution, and the Council of Engineering Institutions by Oxford University Press. 1974.

[62] Hudson T. Morton, Anti-friction bearings, 2nd edition, 1965.

[63] M.C.S. Young and S.W. McMahon, "DSP applications in the condition monitoring of bearings", IEE Colloquium on Digest No. 009, 1992, pp.3/1-3/4.

[64] T.E. Tallian, G.H. Baile, H. Dalal, & O.G. Gustafsson, Rolling Bearing Damage Atlas, SKF Industries, 1974.

[65] R.A. Collacott, Mechanical Falut Diagnosis and Condition Monitoring, 1977, John Wiely and Sons, Inc.

[66] D. Scott, "Effects of environment in rolling element bearings", Rolling element bearings, Proceedings of a seminar organized by the institute of mechanical engineers on 22 Feb 1983., London.

[67] L.G. Hampson, "Diagnostic checks for rolling bearing", Rolling element bearing, Proceedings of a Seminar organized by the Institution of Mechanical Engineers, IMechE, London, Feb. 1983.

[68] Fredric F. Ehrich, *Handbook of Rotordynamics*, McGraw-Hill, 1992.

[69] H.R. Martin, F. Ismail & A. Sakuta, "Algorithms for statistical moment evaluation for machine health monitoring", Mechanical Systems and Signal Processing, 6(4), pp317-327, 1992.

[70] Jones, Robert M., "Enveloping for bearing analysis", SKF Condition Monitoring.

[71] Jones, Robert M., "Vibrations help find faulty components", Evolution, March 1999.

[72] J. Rasmussen and W.B. Rouse, "Human detection and diagnosis of system failures", Plenum Press, New York, pp.9-18, 1981.

[73] N. Praetorius and K.D. Duncan, "Flow representation of plant processes for fault diagnosis", Behaviour and Information Technology, 10(1), pp.41-52, 1991.

[74] Bently Nevada Co.,"Predictive Maintenance through the monitoring and diagnostics of rolling element bearings", ORBIT, Aug. 1987, pp.5-9.

[75] William D. Callister, Jr., Materials Science and Engineering: An Introduction, 2nd edition, John Wiley & Sons, 1991.

[76] B. Weichbrodt and J. Bowden, "Instrument for predicting bearing damage", General Electric Company Report, March 1970, S-70-1021 AD 869, pp.633.

[77] J. Mathew and R.J. Alfredson, "The condition monitoring of rolling element bearing using vibration analysis", Journal of Vibration, Acoustics, Stress, and Reliability in Design, July 1984, Vol. 106, pp.447-453.

[78] David Brie, "Modelling of the spalled rolling element bearing vibration signal: an overview and some new results", Mechanicla Systems and Signal Processing, 2000, 14(3), pp.353-369.

[79] P.L. Howard, "Application of shock pulse technology and vibration analysis of rolling bearing condition monitoring", Proceeding of the 20th International Instrument Symposium, Vol. 20, Pittsburgh, PA, USA, 1974, pp. 231-238.

[80] M.W. Hawman and W.S. Galinaitis, "Acoustic emission monitoring of rolling element bearings", Proceedings of Ultrasonics Symposium, 1988. IEEE,Vol. 2, pp.885-889.

[81] D. R. Harting, "Incipient failure detection by demodulation response analysis", Instrumentation Technology, Sept. 1977.

[82] Bently Nevada, "Acceleration enveloping, avoid it!", Orbit, First Quarter, 1999.

[83] R.B. Randall and Y. Gao, "Masking effect in digital envelope analysis of faulty bearing signals", IMechE, 1996, pp.351-359.

[84] Peter F. J. Burgess, "Antifrication bearing fault detection using envelope detection", Transactions of the Institution of Professional Engineers New Zealand, Electronical/Mechanical Chemical Engineering Section, Vol. 15, No.2. July 1988, pp.77-82.

[85] Hewlett Packet, Effective Machinery Measurements using Dynamic Signal Analyzers, Application Note 243-1, 1990, pp.37.

[86] F. Hlawatsch and G.F. Boudreaux-Bartels, "Linear and quadratic time-frequency signal representations", IEEE SP magazine, April 1992, pp.21-68.

[87] James E. Berry, "How to track rolling element bearing health with vibration signature analysis", Sound and Vibration, Nov. 1991, pp.24-35.

[88] Chuck Yung, Austin Bonnett, "Repair or replace?", IEEE Industry Applications Magazine, Vol. 10, No.5, September/October 2004, pp.48-58.

[89] R. Stribeck, "Ball Bearings for Various Loads", Trans. ASME, 29, 1907, pp.420-463.

[90] Uffe Hindhede et. cl., *Machine Design Fundamentals A Pretical Approach*, Prentice-Hall, New Jersey, 1983, Chapter 17.

[91] Czeslaw Cempel, Vibroacoustic Condition Monitoring, Ellis Horwood, 1991.

[92] B.K.N. Rao, "Handbook of Condition Monitoring", 1996, Elsevier Advanced Technology, pp. 58-59.

[93] Qingfeng Meng and Liangsheng Qu, "Rotating machinery fault diagnosis using Wigner distribution", Mechanical Systemsa and Signal Processing 5(3), 1991, pp.155-166.

[94] Birsen Yazici and Gerald B. Kliman, "An adaptive statistical time-frequency method for detection of broken bars and bearing faults in motors using stator current", IEEE Transactions on Industry Applications, Vol. 35, No.2, 1999, pp.442-452.

[95] Z.K. Peng and F.L. Chu, "Application of the wavelet transform in machine condition monitoring and fault diagnostics: a review with bibliography", Mechanical Systems and Signal Processing 18, 2004, pp.199-221.

[96] J. Shiroishi, Y. Li, S. Liang, T. Kurfess and S. Danyluk, "Bearing condition diagnostics via vibration and acoustic emission measurements", Mechanical Systems and Signal Processing 11(5), 1997, pp.693-705.

[97] T. Williams, X. Ribadeneira, S. Billington and T. Kurfess, "Rolling element bearing diagnostics in run-to-failure lifetime testing", Mechanical Systems and Signal Processing 15(5), 2001, pp.979-993.

[98] Tao Meng, Ming-Fu Liao, Hui Li, "Detection and diagnosis of the gear fault by the delayed correlation-envelope technique", Hangkong Dongli Xuebao, Journal of Aerospace Power, v 18, n 1, February, 2003, pp. 109-113.

[99] M.S. Darlow, "Applications for early detection of rolling-element bearing failure bearing failure using the high-frequency resonance technique", ASME paper 75-DET-46, 1975.

[100] P.D. McFadden and J.D. Smith, "Vibration monitoring of rolling element bearings by the high frequency resonance technique - a review", International Journal of Tribology, 1984, pp.3-10.

[101] P.D. McFadden and J.D. Smith, "Model for the vibration produced by a single point defect in a rolling element bearing", Journal of Sound and Vibration, 96(1), 1984, pp.69-82.

[102] P.D. McFadden and J.D. Smith, "The vibration produced by multiple point defects in a rolling element bearing", Journal of Sound and Vibration, 98(2), 1985, pp.263-273.

[103] Glenn White, "Amplitude demodulation - a new tool for predictive maintenance", Sound and Vibration, Sept. 1991, pp.14-19.

[104] X.F. Wang and P.D. McFadden, "Simulation models for bearing vibration generation and fault detection via neural networks", IMechE 1996, pp.441-450.

[105] S.K. Lee and P.R. White, "The enhancement of impulsive noise and vibration signals for fault detection in rotating and reciprocating machinery", Journal of Sound and Vibration, 217(3), 1998, pp.485-505.

[106] R.H. Lyon, *Machinery noise and diagnostics*, Butterworths, 1987.

[107] B.P. Lathi, Modern digital and analog communication systems, Holt, Rinehart and Winston, 1983, Chapter 4.

[108] S.J. DiMaggio and B.H. Sako, "Basic system identification for condition monitoring of turbopumps", Aerospace Conference, 2001, IEEE Proceedings, Vol.7, 10-17 March 2001, pp. 3189-3200.

[109] Nathan Weller, "Acceleration enveloping - higher sensitivity, early detection", Orbit, 2004, pp.10-19.

[110] Jing Fang, Energy Estimation: Theory and Application, PhD dissertation, University of Washington, 1994.

[111] J.F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal", Proc. ICASSP, 1990, pp.381-384.

[112] P. Maragos, J.F. Kaiser and T.F. Quatieri, "On amplitude and frequency demodulation using energy operators", IEEE Transactions on Signal Processing. Vol. 41, No. 4, April 1993, pp.1532-1550.

[113] R.B. Dunn, T.F. Quatieri and J.F. Kaiser, "Detection of transient signals using the energy operator", Proc. ICASSP, 1993, Vol. III, pp. 145-148.

[114] J.Fang and L.E. Atlas, "Quadratic detectors for energy estimation", IEEE Transactions on Signal Processing, Vol. 43, No. 11, Nov. 1995, pp.2582-2594.

[115] R. Lin, S. Zhu, H. Wu, and J Zheng, "Rolling bearings fault diagnosis based on energy operator demodulation approach", Proc. of the 4th World Congress on Intelligent Control and Automation, June 10-14, 2002, pp.2723-2727.

[116] D. Ho and R.B. Randall, "Optimisation of bearing diagnostic techniques using simulated and actual bearing fault signals", Mechanical Systems and Signal Processing 14(5), 2000, pp.763-788.

[117] J. Altmann and J. Mathew, "Multiple band-pass autoregressive demodulation for rolling-element bearing fault diagnosis", Mechanical Systems and Signal Processing 15(5), 2001, pp.963-977.

[118] Xia Limin, "Wavelet packets analysis of rolling bearing vibration signal and fault testing", Proc. of the 4th World Congress on Intellignet Control and Automation, June 10-14, 2002, pp. 3005-3008.

[119] L. Eren and M.J. Devaney, "Motor bearing damage detection via wavelet analysis of the starting current transient", IEEE Instrumentation and Measurement Technology Conference, May 21023, 2003, pp.1797-1800.

[120] L. Eren and M.J. Devaney, "Bearing damage detection via wavelet packet decomposition of the stator current", IEEE Transactions on Instrumentation and Measurement, Vol. 53, No. 2, April 2004, pp.431-436.

[121] H. Zhang, S. Wang, Q. Zhang and G. Zhai, "The research on rolling element bearing faullt diagnosis based on wavelet packets transform", IEEE, 2003, pp.1745-1749.

[122] Z.K. Peng and F.L. Chu, "Application of the wavelet transform in machine condition monitoring and fault diagnostics: a review with bibliography", Mechanical Systems and Signal Processing 18, 2004, pp.199-221.

[123] Y. Shao and K Nezu, "Extracting symptom of bearing faults in the wavelet domain", Proc. Instn Mech. Engrs, Vol.218, Part I: J. Systems and Control Engineering, 2004, pp. 39-51.

[124] C. Li, Z. Song and P. Li, "Bearing fault detetion via wavelet packet transform and rough set theory", Proc. of the 5th World Congress on Intelligent Control and Automation, June 15-19, 2004, pp. 1663-1666.

[125] R.J. McAulay and T.F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation", IEEE Transactions on Acoustics, Speech and Signal Precessing, Vol. 34, pp. 744-754, 1986.

[126] Qin Li, Time-varying harmonic modeling for pitch estimation and epoch detection, Thesis, University of Washington, 2003.

[127] N. Abu-Shikhah and M. Deriche, "A robust technique for harmonic analysis of speech", Proc. IEEE ICASSP, Piscataway, NJ, 2001.

Appendix A. SEM Pictures of Run-to-Fail Bearings

Visible examination of the dismantled bearing by a naked eye cannot reveal details of bearing condition. When magnified with scanning electron microscope (SEM), small metal chips, finishing marks, rough surface and spalls became clearly visible. The following atlas (Figure A-1 to Figure A-6) present some evidences of the bearing failure and their severity.

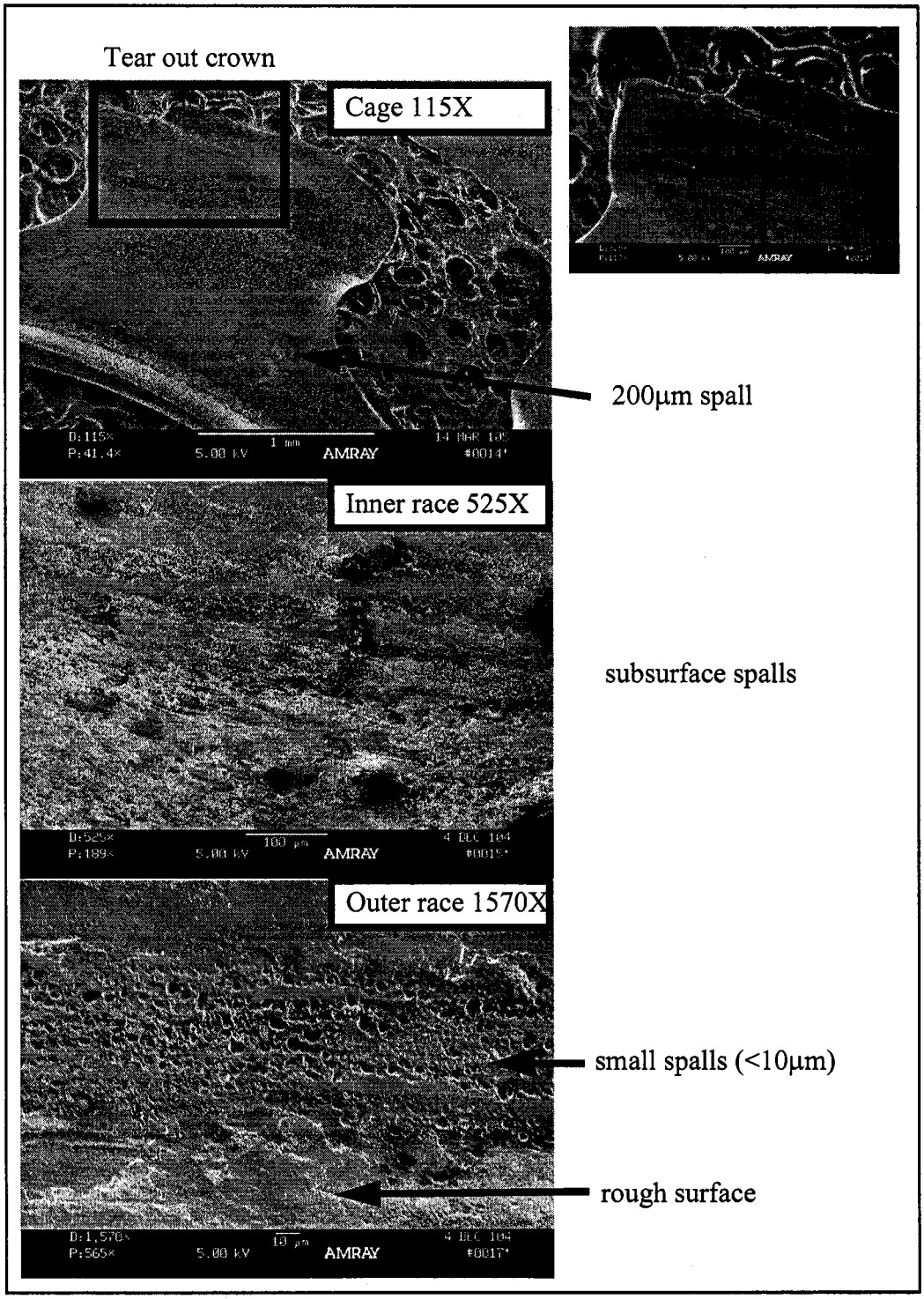


Figure A-1. SEM pictures of bearing R24.

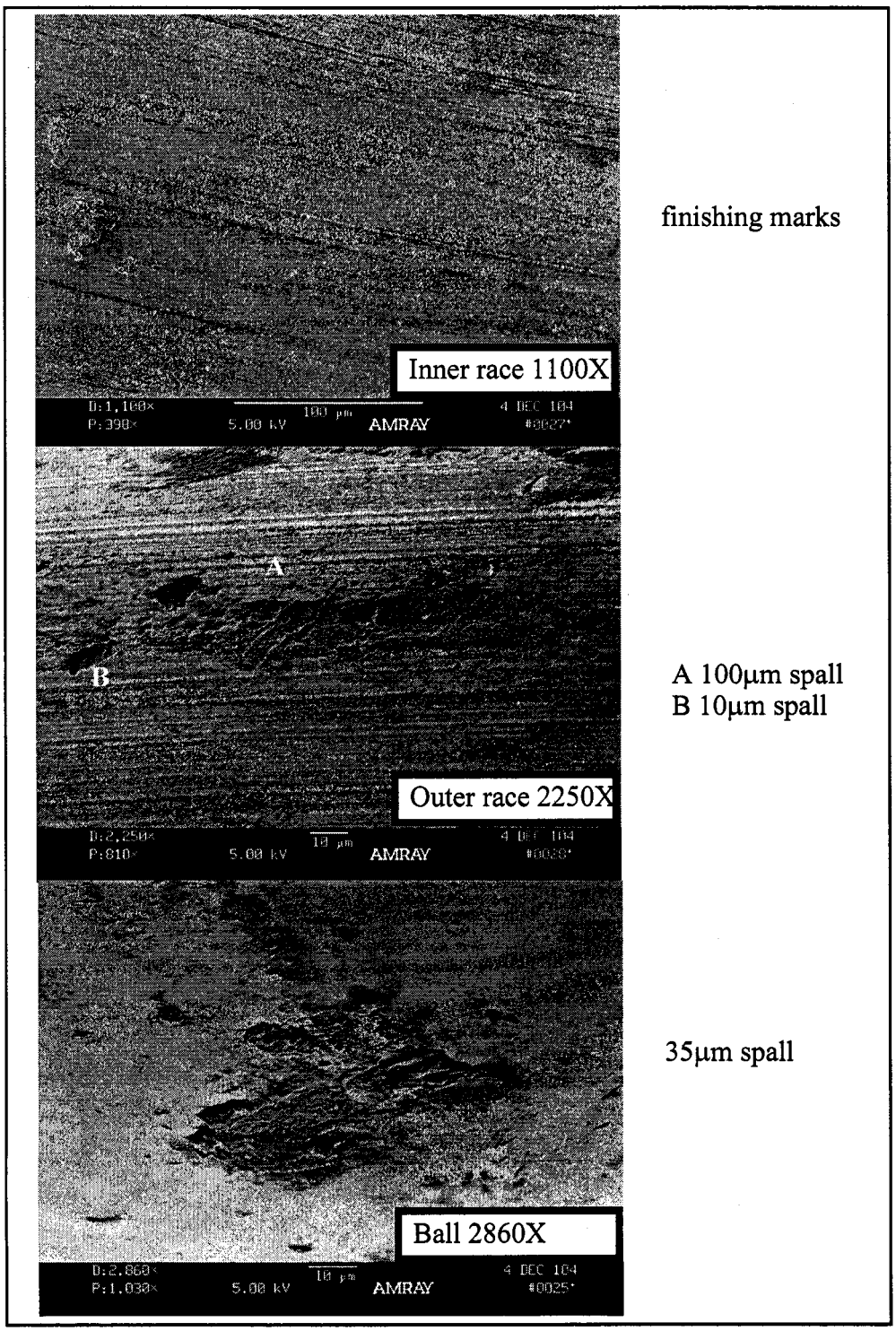


Figure A-2. SEM pictures of bearing R25.

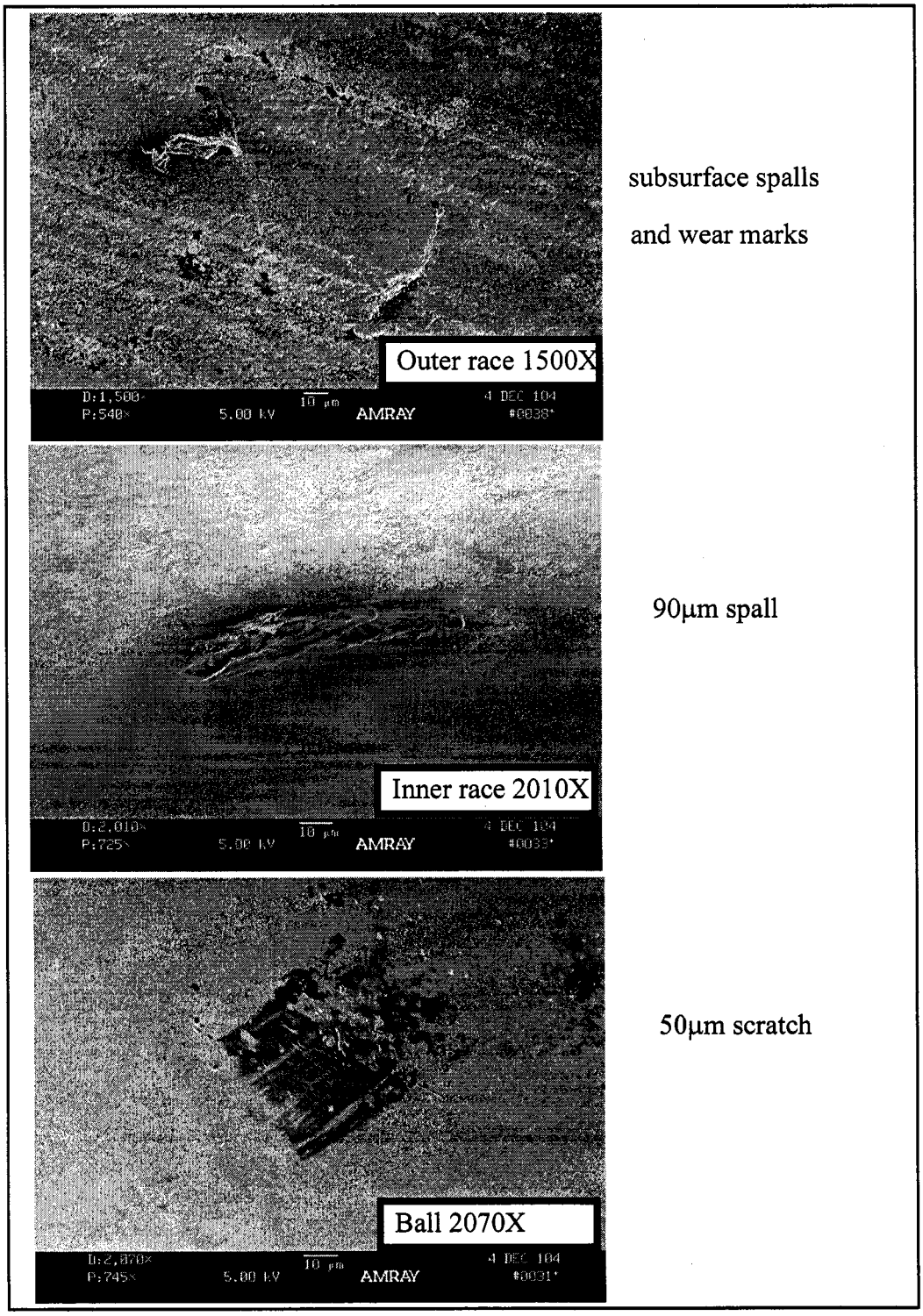


Figure A-3. SEM pictures of bearing R26.

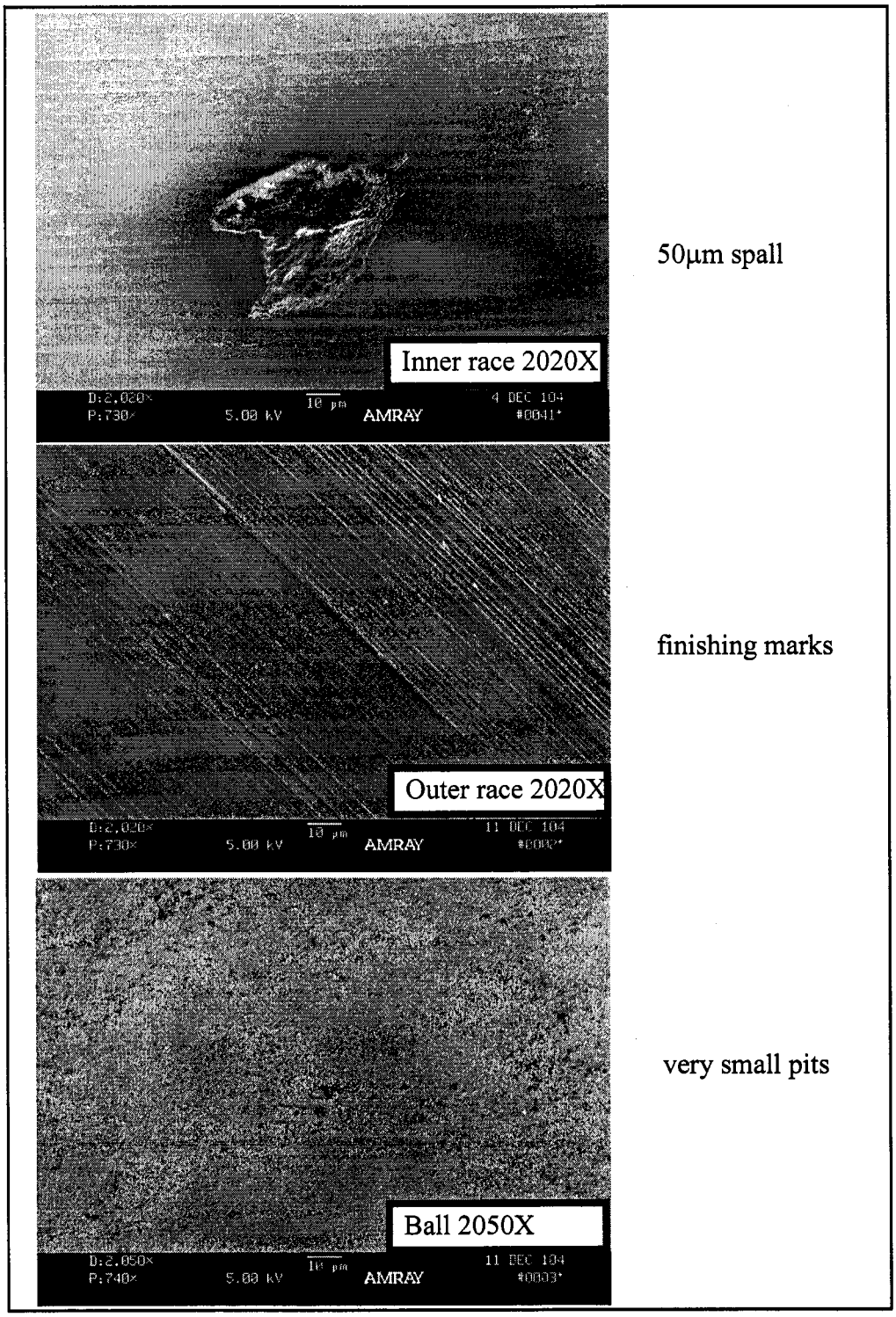


Figure A-4. SEM pictures of bearing R29.

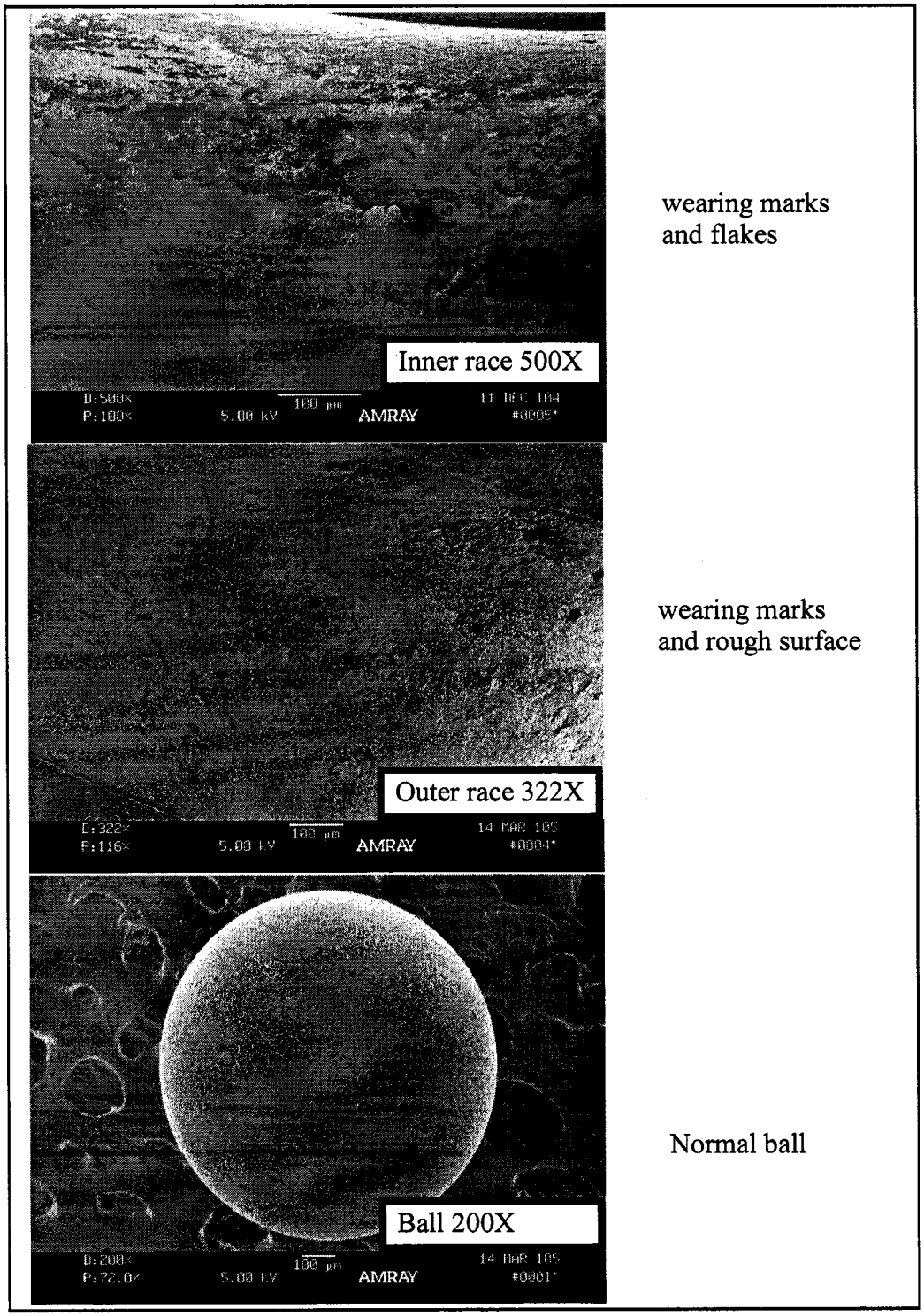


Figure A-5. SEM pictures of bearing R30.

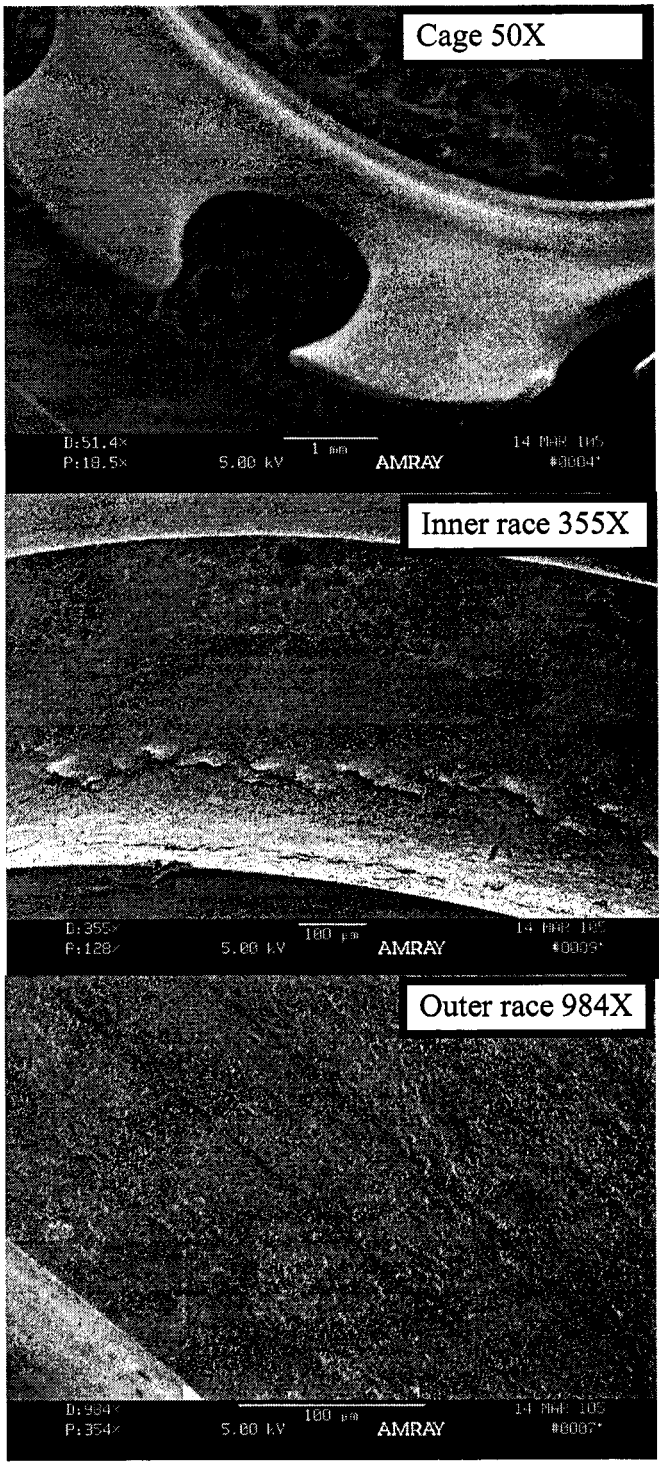


Figure A-6. SEM pictures of bearing R31.

Vita

Chao-Shih Liu was born in Taipei, Taiwan, on September 27th, 1964. He earned a Bachelor of Science degree in Chemical Engineering from Chung-Cheng Institute of Technology (CCIT) in 1987, where he started his military career. After completing his bachelor degree, he became a First Lieutenant of the Army of the Republic of China in Taiwan. In 1987-1991, he worked in the Department of Applied Chemistry of the CCIT for four years as a teaching and research assistant. In 1991-1993, he earned a Master degree in the field of Mechanical Engineering from Naval Postgraduate School (NPGS) at Monterey, CA. After graduated from NPGS, he returned to CCIT as a lecturer in the Department of System Engineering from 1993 to 1996. In 1996, he entered the Department of Mechanical Engineering at University of Washington to pursue his Ph.D. and earned the degree in 2005. Currently, he is a Lieutenant Colonel in the Division of Research and Continuing Education of CCIT as a senior research officer.