WAVELET PACKET FEATURE EXTRACTION

FOR VIBRATION MONITORING

By

KUO-CHUNG LIN

Bachelor of Science
Tamkang University
Taipei, Taiwan, R. O. C.
1990

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
December, 1998

WAVELET PACKET FEATURE EXTRACTION

FOR VIBRATION MONITORING

Thesis Approved:

Thesis Adviser

Dean of the graduate College

ACKNOWLEDGMENTS

I wish to express my sincere appreciation to my major advisor, Dr. Gary Y. Yen for his intelligent supervision, inspiration and friendship throughout this research. My honest appreciation extends to my other committee members, Dr. Martin T. Hagan and Dr. Scott. T. Acton, for their valuable assistance. I would like to thank the Department of Computer and Electrical Engineering for providing the laboratory facilities which made this study possible.

I want to thank my parents, Chau-Jang Lin, Shiou-Luan Weng, for their support in whatever endeavor I have decided to undertake. I also want to thank my sister, Sheng-Tan Lin for her financial support. I also want to thank my friend, Meihua Koo, and her husband Hsuan-Tsung Hsieh for countless lunches, dinners and rescues from many of my troubles.

TABLE OF CONTENTS

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Condition Monitoring

Any major piece of industrial machinery equipment requires a certain degree of maintenance to assure successful operation over a long period of time. To achieve this objective, an automated condition monitoring system is needed. This system allows early detection of potentially catastrophic faults which could be extremely expensive to repair. It also allows for implementation of condition based maintenance, and significant savings can be made by delaying scheduled maintenance until convenient or necessary. Generally, a simple condition monitoring system is approached from a pattern classification perspective. It can be decomposed into three general tasks: data acquisition, feature extraction and condition classification [1] as briefly described next.

1.2 Data Acquisition - Vibration Signal

The most common family of monitoring methods is based on the vibration measurements using multiple sensors [2, 3, 4, 5, 6, 7, 8]. The general principle behind using vibration signals for monitoring is that components in mechanical systems generate vibration during operation. When faults develop, some of the system dynamics change.

This results in significant deviations in vibration pattern. By employing appropriate data analysis algorithms, it is feasible to detect changes in vibration signals caused by fault components, and to make decisions about the machinery health status.

1.3 Condition Classification

In many of classification systems currently used, neural networks in particular, the process of feature extraction is inherently embedded in the classification technique rather than being apparent as a separate process. If a multi-layer neural network is used to classify unprocessed data, the input layer, which learns from examples, will essentially become a feature extractor. However, in problems such as vibration time series data, the input dimensionality becomes an impediment to classification. Even neural networks are limited by the problem of parameter estimation – as the number of parameters increase, the number of the data required to train the neural networks increases for satisfactory performance. For a complex problem, obtaining the necessary data may be expensive or even impossible. The feature extraction is needed to reduce the dimensionality of the data before performing classification. This is based on the assumption that the important structure in the data actually lies in a much lower dimensional space.

1.4 Feature Extraction

Feature extraction involves preliminary processing of sensor measurements to obtain suitable parameters that reveal whether an interesting pattern is present. It is generally not possible to classify machine conditions based upon an individual sample of the vibration, therefore, a feature extraction technique is needed for preliminary

processing of recorded time-series vibrations to obtain suitable parameters that, in linear and/or nonlinear combination, reveal whether a fault is developing. This, in general, requires windowing of the time-series vibration signals to form signal segments on which linear, bilinear, or nonlinear transformations are applied. The aim of feature extraction is to devise a transformation that extracts the signal features hidden in the original domain. Corresponding to different characteristics of signals, transformations should be properly selected such that specific signal structure could be enhanced in its transformation domain. This might make the following fault classification easier.

1.5 Problem Background

Usually, the vibration signals of defective components are highly structured and could be grouped into two categories: sustained defects and intermittent defects [9]. For sustained defects, the signal is sinusoidal. Fourier based analysis, which uses sinewave functions as basis functions, provides an ideal candidate for extraction of these narrowband signals. For intermittent defects, features reflecting machinery faults in the pick-up (windowed) time series vibration signals neither appear in a repetitive manner nor consist of regular frequency components with the evolution of time. Instead, these signals often demonstrate a nonstationary and transient nature, and carry small yet informative components embedded in larger repetitive signals. In this case, the Short Time Fourier Transorm (STFT) can be employed to detect the local transient. Unfortunately, fixed windowing implies fixed time-frequency resolution in the time-frequency plane [10, 11]. The difficulty is that the accuracy of extracting frequency information is limited by the length of the window relative to the duration of the

interesting signal. For example, in helicopter transmissions, the important information concerning bearings can be on the order of tens of hundreds of Hertz, whereas mesh frequencies and important fundamentals associated with gearing of the engine input, can be on the order of tens of thousands of Hertz. To overcome the fixed time-frequency resolution problems, recently developed wavelet based analysis [10], which provides flexible time frequency resolution, becomes an efficient alternative in dealing with this type of machinery transient process.

Nonetheless, linear expansions in a single basis, whether Fourier or wavelet, is not flexible enough. Fourier basis provided a poor representation of signals localized in time. Wavelet bases are not well adapted to represent signals whose Fourier transform have a narrow "high" frequency support because of poor resolution in high frequency. In both cases, it is difficult to detect and identify the signal pattern from their expansion coefficients because information is diluted across the whole basis. The Wavelet Packet transform [12], on the other hand, uses a rich library of redundant bases with arbitrary time frequency resolution. Therefore it enables the extraction of features of signals that combine non-stationary and stationary characteristics.

1.6 Motivation of the Study

The collection of all wavelet packet coefficients contains far too many elements to efficiently represent a signal. Care must be taken in choosing a subset of this collection in order to be really useful in practical situations. For classification applications, a natural direction is to address the issue of finding a wavelet packet based feature set that offers maximum feature separability due to class-specific characteristics. Our study explores the

feasibility of the wavelet packet transform as a tool in the search for features that may be used in the detection and classification of machinery vibration signals. In particular, we formulate a systematic method of determining wavelet packet based features that exploit class specific differences among interested signals. This would avoid human interaction. One could simply input a sample data set that represents the signals of interest, and receive as output the dominant features that are suitable for classification purposes. In this thesis, we introduce a novel methodology for classifying vibration signals based on wavelet packet analysis. We suggest that such analysis can provide more effective method to achieve robust classification than traditional single resolution techniques.

1.7 Thesis Outline

The thesis investigates the use of the wavelet-packet-based feature in the classification of vibration signals. In chapter 2 we discuss the inefficiency of Fourier based analysis for transient signal analysis, and lead the reader to the wavelet based analysis - wavelet transform and its generalization the wavelet packet transform. Chapter 3 presents an overview of the proposed classification system based on wavelet packet features. The feature measure used throughout the thesis is first described. Then we present two feature selection methodologies that aim to reduce the input dimension for the classifier. In chapter 4 the feasibility of the proposed wavelet-packet-based feature extraction technique is examined through numerical simulations of seed faults in the Westland helicopter transmission data set. We present our results and discuss the performance with respect to parameters considered in our investigation. We conclude our study in chapter 5.

CHAPTER II

WAVELET THEROY

2.1 Introduction

The aim of signal analysis is to devise a transformation that extracts the signal features hidden in the original domain. Corresponding to characteristics of signals, different transformations should be properly selected such that specific signal structure, which is hidden in its original domain, can be revealed on its transformation side. This might make the subsequent processing easier (in our case the vibration signal classification application). In following sections we briefly discuss the Fourier based analysis and its inefficiency in dealing with non-stationary signals. This naturally leads to the Wavelet analysis, which is more efficient than Fourier based analysis for non-stationary signals. Then the Wavelet analysis leads at last to the Wavelet Packet Transform, the generalization of the Wavelet Transform.

2.2 Fourier Based Analysis

Vibration signal classification generally requires windowing of the time-series vibration signals to form signal segments on which linear, bilinear, or nonlinear transformations are applied. The Fourier based methods, in particular the Short Time

Fourier Transform (STFT), are usually employed for the extraction of narrow band frequency content in signals. The difficulty with STFT is that the accuracy for extracting frequency information is limited by the length of this window relative to the duration of the signal. Specifically, the STFT is defined as:

$$G(f,\tau) = \int x(t)g^*(t-\tau)e^{-j2\pi ft}dt\,, \tag{2.1}$$

where $g(t)$ is a window function. The STFT decomposes a signal in time domain into a two-dimensional function in a time frequency plane (f,τ). At a given frequency f, Eq. (2.1) is equivalent to filtering a signal at all times with a bandpass filter having as impulse response the window function modulated to that frequency f. Alternatively, given a segment of signal windowed around time instant τ, one computes all frequencies of the STFT. Now consider the ability of the STFT to discriminate between two pure sinusoids. Given a window function $g(t)$ and its Fourier transform $G(f)$, define the bandwidth Δf of the filter as

$$\Delta f^2 = \frac{\int f^2 |G(f)|^2 \, df}{\int |G(f)|^2 \, df}. \tag{2.2}$$

Then two sinusoids will be discriminated only if they are more than Δf apart. Similarly, the spread in time is given by Δt defined as:

$$\Delta t^2 = \frac{\int t^2 |g(t)|^2 \, dt}{\int |g(t)|^2 \, dt}. \tag{2.3}$$

So, two pulses in time can be discriminated only if they are more than Δt apart. Thus, the resolution in frequency of the STFT analysis is given by Δf, and the resolution in time is

given by Δt. One important property, according to the uncertainty principle [13], is that for any suitably chosen window function, the time-bandwidth product of the window function has lower bound given by

$$\Delta t \Delta f = c \geq 1/4\pi \ . \tag{2.4}$$

Here c is a constant dependent on the choice of $g(t)$. Note that once the window function $g(t)$ is defined, the area (time-bandwidth product) of the window function in the time frequency plane remains fixed. It means we cannot increase the time and frequency resolutions simultaneously. If we choose a window function with small Δt (good time resolution), then the corresponding frequency resolution will be poor (Δf will be large). Figure 2-1 shows how the STFT decomposes a signal into the time-frequency plane using two different window functions. We can use a shorter duration window function to get a better time resolution, at the cost of losing frequency resolution or vice versa.

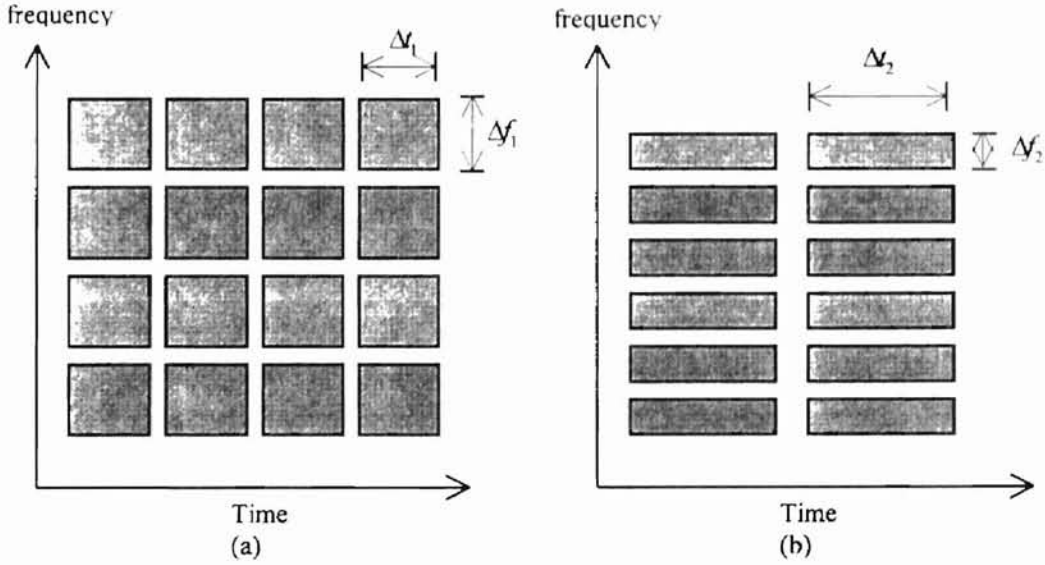

Figure 2-1: Decomposition of signal using STFT (a) long analysis window function, (b) short analysis window function.

Consider analyzing various signals with different analysis window functions to demonstrate the possible drawbacks of fixed time-frequency resolution associated with the STFT. In Figure 2-2, the signal, $x(t)$, contains a high frequency component, so the window function, $g(t)$, has captured enough numbers of cycles for extracting accurate frequency information in the signal. Whereas in Figure. 2-3, the duration of $g(t)$ is too long for the high frequency burst, therefore it will also capture other components of the signal in that time duration. This may cause the short-time pulse to be buried and remain undetected. For Figure 2-4, $g(t)$ is too short to capture the low frequency signal. Consequently, if we are analyzing the low frequency content of a signal, we might desire a wide window function in time. Conversely, if we were interested in high frequency phenomena, a short duration window function would be preferred. The STFT does not allow this desired flexibility, but, as we will see in next section, wavelets give a framework for which this is automatic.

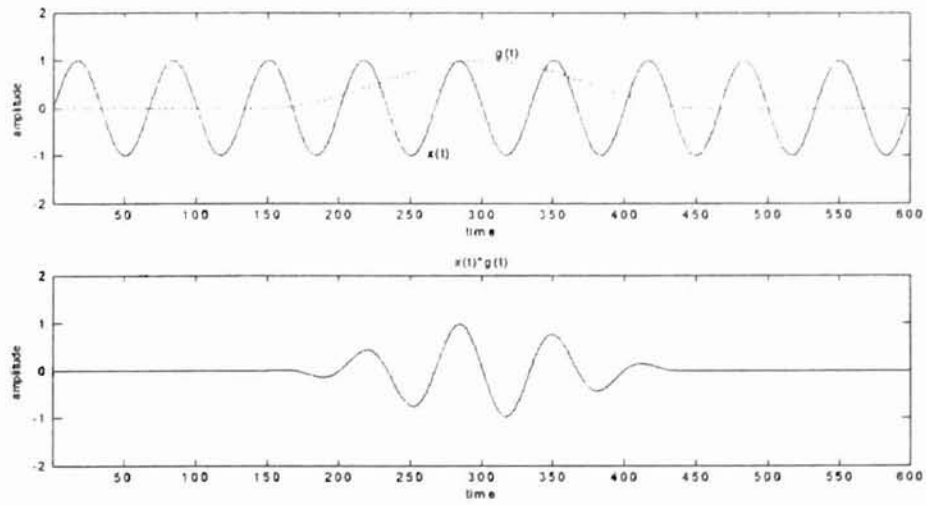

Figure 2-2: Proper analysis window function

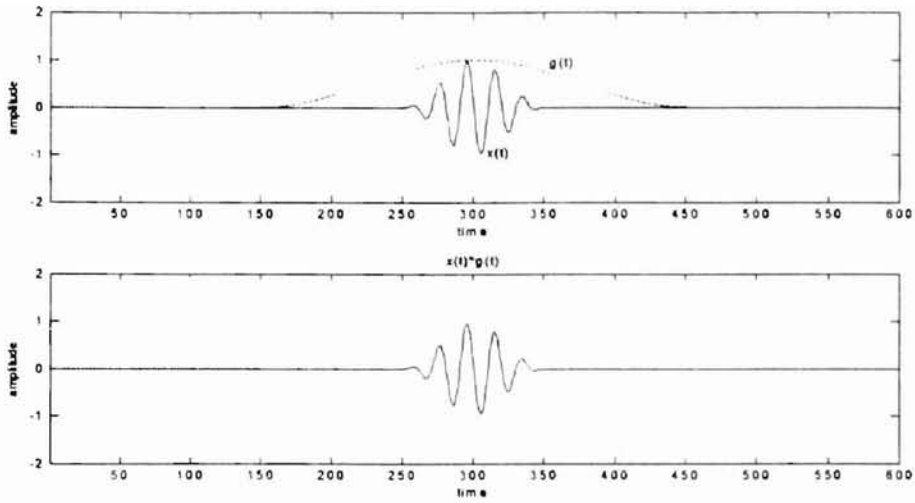

Figure 2-3: Too long analysis window function

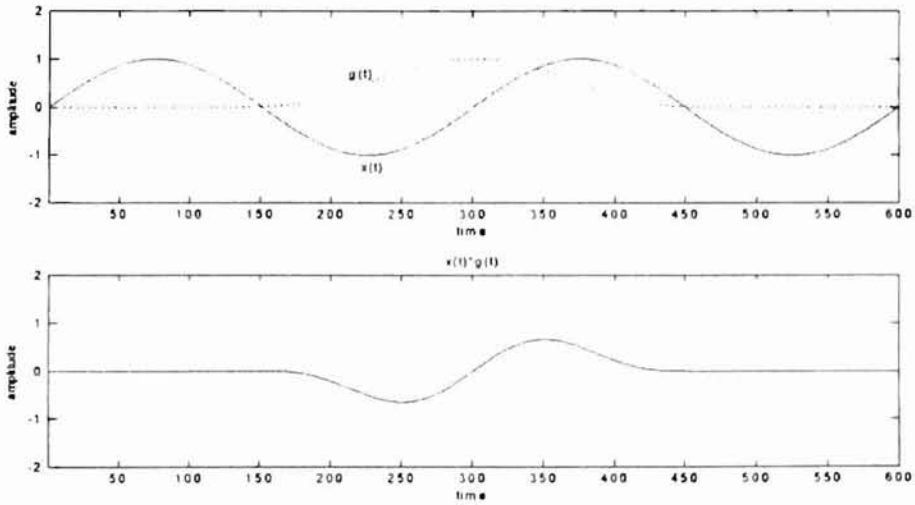

Figure 2-4: Too short analysis window function

2.3 Wavelet Based Analysis

Whereas Fourier based analysis is based on sinusoidal functions of various frequencies, the wavelet analysis, on the other hand, is founded on basis functions formed by dilation and translation of a prototype function $\psi(t)$, also known as a mother wavelet. A typical wavelet function is shown in Figure 2-5. One could note that the wavelet function is localized in both time and frequency domains.

The wavelet basis function, $\psi_{a,\tau}(t)$, is a family of short-duration, high frequency and long-duration, low frequency functions defined as [14]:

$$\psi_{a,\tau}(t) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{t-\tau}{a}\right),\ a>0, \tau \in \Re. \tag{2.5}$$

The parameter τ indicates the translation in time, and the parameter a is the scale parameter.

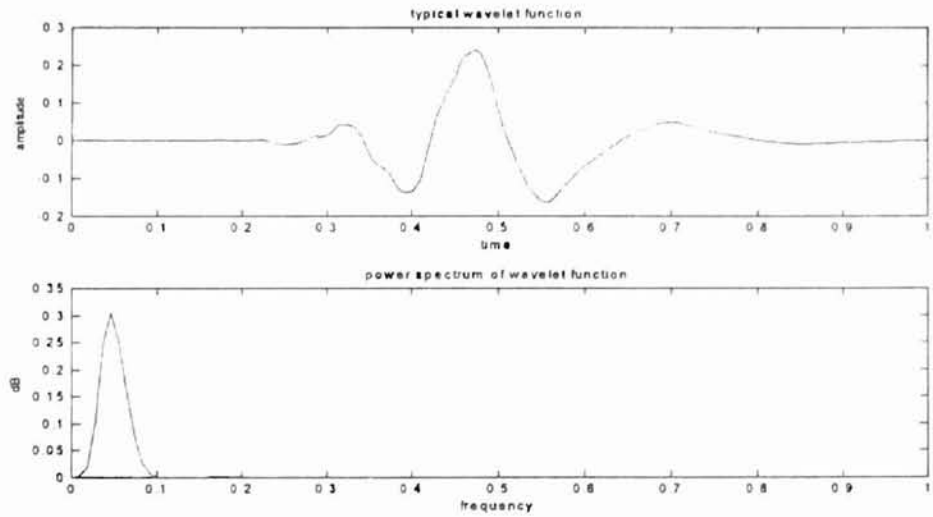

Figure 2-5: A typical wavelet function and its spectra. The frequency axis is in units of $\pi \times$ radians.

From the scaling property of Fourier transforms, if

$$\psi(t) \leftrightarrow \Psi(\Omega) \tag{2.6}$$

formed a Fourier transform pair, then

$$\frac{1}{\sqrt{a}}\psi\left(\frac{t}{a}\right) \leftrightarrow \sqrt{a}\Psi(a\Omega) \tag{2.7}$$

where $a > 0$ is a continuous variable. Thus a contraction in one domain is accompanied by an expansion in the other, but in a nonuniform way over the time-frequency plane. Depending on the dilation parameter, a, the wavelet function dilates or contracts in time causing the corresponding contraction or dilation in the frequency domain. Figure 2-6 displays a set of wavelet functions and their corresponding Fourier transform for different dilation parameters. When a is large ($a > 1$), the basis function becomes a stretched version of the mother wavelet ($a = 1$) and demonstrates a low-frequency characteristic. When a is small ($a < 1$), this basis function is a contracted version of the mother wavelet function and demonstrates a high frequency characteristic. Note, however, that each scale parameter represents a frequency band, not pure frequency information.

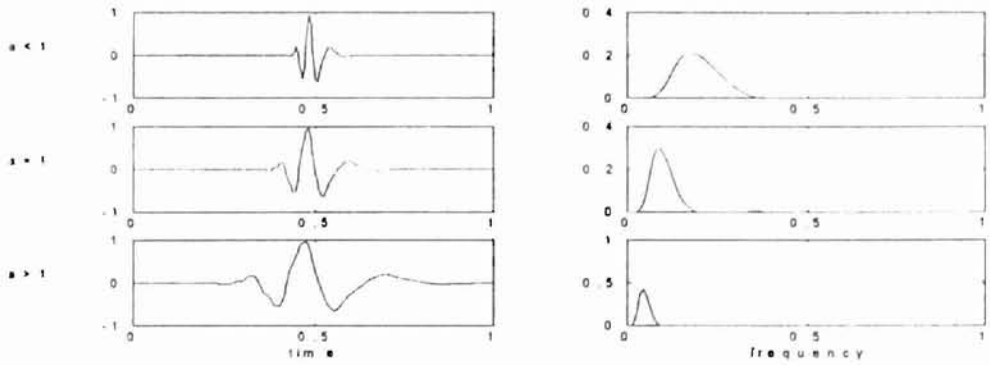

Figure 2-6: Wavelet basis function and corresponding frequency spectrum. The frequency axis is in units of π × radians.

Similar to the STFT, one can analyze a signal with continuous wavelet transform (CWT) which decomposes a signal in time domain into a two-dimensional function in *time-scale* plane (a,τ):

$$\Psi(a,\tau) = \int x(t)\psi_{a,\tau}(t)dt\,. \tag{2.8}$$

The wavelet coefficient $\Psi(a,\tau)$ measures the *time-frequency* content in a signal indexed by the scale parameters and translation parameters. The term *frequency* instead of *scale* has been used in order to aid in understanding since a wavelet with large scale parameter is related to high frequency content component, and vice versa. Similar to Figure 2-1, we can construct a picture giving some idea of the simultaneous time-frequency localization that takes place when applying CWT. Figure 2-7 gives a rough idea of the time-frequency localization corresponding to the CWT. Thus we see that the CWT corrects the noted deficiencies of the Fourier analysis as described in the previous section. That is, the CWT analyzes the low frequency content of a signal with a wide duration function and conversely, analyzes high frequency phenomena with a short duration function.

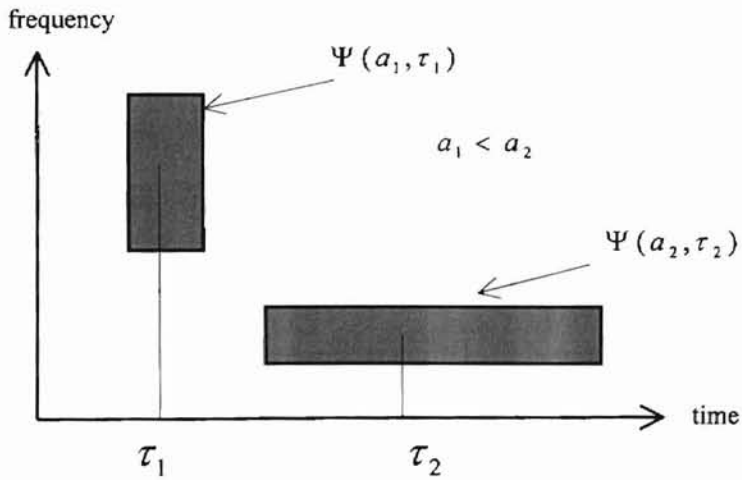

Figure 2-7: Decomposition of signal using continuous wavelet transform

2.4 Fast Wavelet Transform

In practice, calculating wavelet coefficients at every possible scale using Equation (2.8) is a fair amount of work and it generates a lot of redundant data. It turns out that if we limit the choice of a and τ in Equation (2.5) to a discrete number then our analysis will be sufficiently accurate. In particular, if we choose scale and translation parameters based on power of two, then there exists $\psi(t)$ with good time-frequency localization properties. The set of functions

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k), \quad j,k \in Z \tag{2.9}$$

constitutes an orthonormal basis for $L^2(\Re)$ [15]. Here Z denotes the set of integers, and $L^2(\Re)$ denotes the class of measurable functions, $x(t)$, in $\Re$ satisfying:

$$\int_{\Re} |x(t)|^2 \, dt < \infty . \tag{2.10}$$

Any signal $x(t)$ in $L^2(\Re)$ can then be expressed as

$$x(t) = \sum_{j,k} < f, \psi_{j,k} > \psi_{j,k}(t) . \tag{2.11}$$

This is called a discrete wavelet transform (DWT). In practice, the implementation of the DWT suitable for finite length discrete time signals is based upon the multiresolution analysis (MRA) introduced by S. Mallat [16] which leads to a highly efficient algorithm known as the Fast Wavelet Transfom (FWT). By introducing a new function, the scaling function, the orthogonal wavelets could be constructed and incorporated into a system that uses a cascade of filters to decompose a signal. This practical filtering algorithm is in fact a classical scheme known as a two-channel subband coding using quadrature mirror filters (QMF) [17]. A consequence of multiresolution is that we can transform a signal

into wavelets without using wavelets or scaling functions. In general these functions do not exist as explicit functions; they are limits of iterations. To compute the wavelet transform all we need are filters. Rather than taking the scalar product of the scaling function or the wavelet with the signal, we convolve the signal with these filters.

Specifically, MRA consists of a sequence of closed subspaces $\{V_j\}_{j\in Z}$ of $L^2(\Re)$ which have the following properties :

(a.1) $\{0\} \subset \ldots \subset V_{-1} \subset V_0 \subset V_1 \subset \ldots \subset L^2(\Re)$

(a.2) $\bigcup_{j\in Z} V_j = L^2(\Re)$ and $\bigcap_{j\in Z} V_j = \{0\}$

(a.3) $f(x) \in V_0 \Leftrightarrow f(2^j x) \in V_j$ for $j \in Z$

(a.4) There exists a scaling function, $\phi(t) \in V_0$, such that $\forall j \in Z$, the set

$\{\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k)\}_{k\in Z}$ constructs an orthonormal basis for V_j.

Let W_j to be the orthogonal complement of V_j in V_{j+1}, i.e. $V_j \perp W_j$ and

$$V_{j+1} = V_j \oplus W_j, \tag{2.12}$$

where $\oplus$ denotes the direct sum of vector spaces. Then the $L^2(\Re)$ can be decomposed as an infinite direct sum of W_j :

$$\oplus_{j=-\infty}^{j=\infty} W_j = L^2(\Re) \tag{2.13}$$

It is shown that there exists a function , $\psi(t) \in W_0$, such that $\{\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)\}_{k\in Z}$ is an orthonormal base for W_j [17]. This function, $\psi(t)$, is the mother wavelet function associated with the multiscale analysis.

Since $V_0 \subset V_1$, any function in V_0 can be expanded in terms of basis functions of V_1, i.e. $\{\phi_{1,k}\}_{k\in Z}$. In particular, $\phi(t) \in V_0$ so

$$\phi(t) = \sqrt{2}\sum_k h(k)\phi(2t-k) \tag{2.14}$$

In analogy, since $\psi(t) \in W_0$ and $W_0 \in V_1$ we can expand $\psi(t)$ as

$$\psi(t) = \sqrt{2}\sum_k g(k)\psi(2t-k) \tag{2.15}$$

Now suppose that the function $x(t)$ is in V_0 so that:

$$x(t) = \sum_n a_{0,n}\phi_{0,n}(t) \tag{2.16}$$

Since $V_0 = V_{-1} \oplus W_{-1}$, we can express the function $x(t) \in V_0$ as the sum of two functions, one lying in V_{-1} and the other in the W_{-1} :

$$x(t) = f_v^{-1}(t) + f_w^{-1}(t) \tag{2.17}$$

where

$$f_v^{-1}(t) = \sum_k a_{-1,k}\phi_{-1,k}(t) \tag{2.18}$$

$$f_w^{-1}(t) = \sum_k d_{-1,k}\psi_{-1,k}(t) \tag{2.19}$$

Multiplying both sides of Eq. (2.17) by $\phi_{-1,n}(t)$ and integrating yields

$$\langle x(t), \phi_{-1,n}(t)\rangle = \langle f_v^{-1}(t), \phi_{-1,n}(t)\rangle + \langle f_w^{-1}(t), \phi_{-1,n}(t)\rangle \tag{2.20}$$

Since $f_w^{-1}(t)$ is a linear combination of $\{\psi_{-1,k}(t)\}$, each component of which is orthogonal to $\phi_{-1,n}(t)$, the second inner product in Eq. (2.20) is zero. Also, since $\{\phi_{-1,k}(t)\}$

is the orthonormal basis for V_{-1}, all components are mutually orthogonal to each other such that :

$$\left\langle f_v^{-1}(t), \phi_{-1,n}(t) \right\rangle = \sum_k \int a_{-1,k}\phi_{-1,k}(t)\phi_{-1,n}(t)dt = a_{-1,n} \tag{2.21}$$

Therefore we have:

$$\left\langle x(t), \phi_{-1,n}(t) \right\rangle = \left\langle f_v^{-1}(t), \phi_{-1,n}(t) \right\rangle = a_{-1,n} \tag{2.22}$$

by substituting $\phi_{-1,n}(t) = \frac{1}{\sqrt{2}}\phi\left(\frac{t}{2} - n\right)$ into Eq. (2.20), we have

$$\left\langle x(t), \phi_{-1,n}(t) \right\rangle = a_{-1,n} = \int x(t)\frac{1}{\sqrt{2}}\phi\left(\frac{t}{2} - n\right)dt \tag{2.23}$$

From Eq. (2.14)

$$\phi\left(\frac{t}{2} - n\right) = 2\sum_k h(k)\phi(t - 2n - k) \tag{2.24}$$

therefore

$$a_{-1,n} = \sqrt{2}\int x(t)\sum_k h(k)\phi(t - 2n - k)dt$$

$$= \sqrt{2}\sum_k h(k)\int x(t)\phi(t - 2n - k)dt$$

$$= \sqrt{2}\sum_k h(k)a_{0,2n+k} = \sqrt{2}\sum_k h(k - 2n)a_{0,k} \tag{2.25}$$

In a similar way, we can arrive at

$$d_{-1,n} = \sqrt{2}\sum_k g(k - 2n)a_{0,k} \tag{2.26}$$

Following the same process, we have

$$a_{m-1,n} = \sqrt{2}\sum_k h(k-2n)a_{m,k} \tag{2.27}$$

$$d_{m-1,n} = \sqrt{2}\sum_k g(k-2n)a_{m,k} \tag{2.28}$$

Therefore, the wavelet coefficients at coarse level $d_{m-1,n}$ can be computed by filtering $a_{m,k}$ using $g(k)$ as filter coefficients and discarding every other point. I. Daubechies [15] has developed a procedure to solve Eq. (2.14) and Eq. (2.15) such that the sequences $h(k)$ and $g(k)$ have only finite nonzero coefficients, which leads to a very efficient algorithm for computing wavelet coefficients. In general, $h(k)$ is the coefficient of the low pass filter, whereas $g(k)$ represents a high pass filter. Figure 2-8 shows the filter sequences, $h(k)$ and $g(k)$, associated Daubechies 8 point wavelets and their respective frequency spectra.

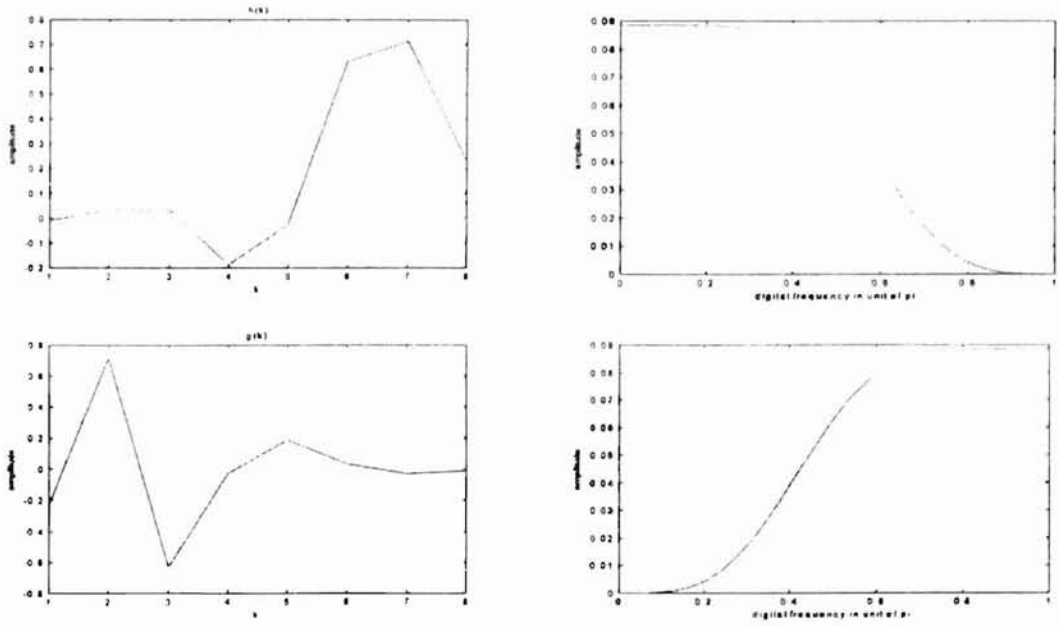

Figure 2-8: Daubechies 8 point filters and corresponding spectra. The frequency axis is in units of $\pi \times$ radians.

In practice, a discrete time signal can only represent a continuous time signal at finite time resolution dependent on the sampling frequency. The wavelet decomposition of a discrete time signal could be implemented by regarding the data values $[x_1, x_2, ..., x_n]$ as the finest resolution scaling function coefficients, i.e. $a_{0,k}$ in Equation (2.16), and from which all coarse-level coefficients are recursively computed using Eq. (2.27) and Eq. (2.28). This decomposition procedure is illustrated in Figure 2-9.

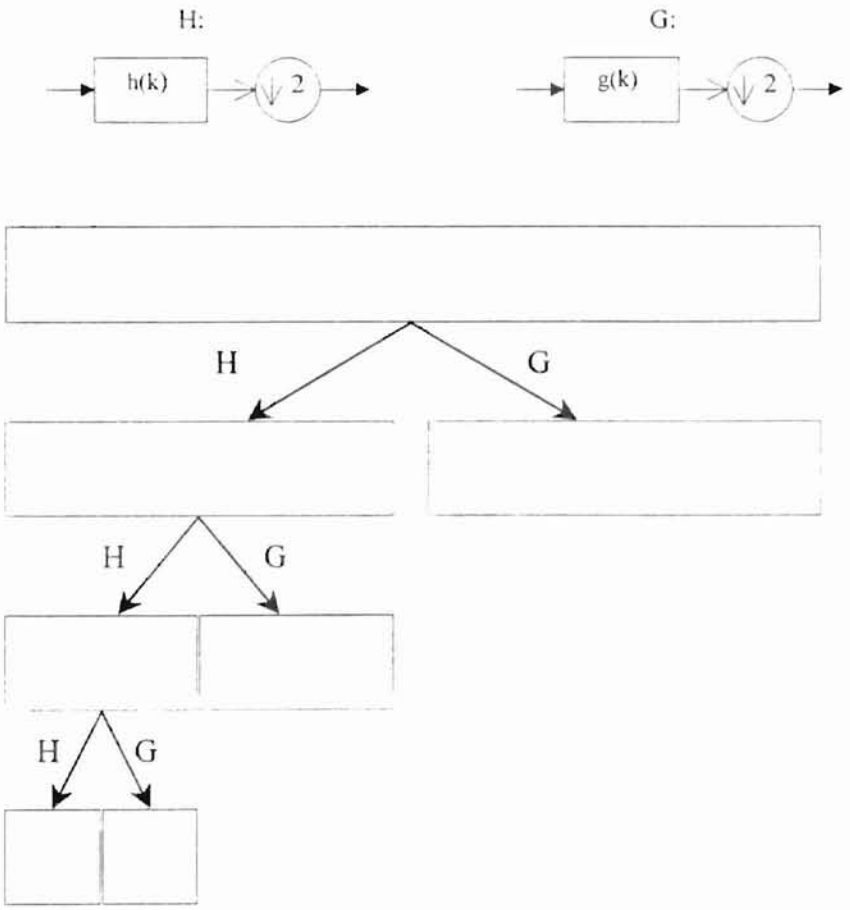

Figure. 2-9: Implementation of fast wavelet transform

The procedure just described indicates that the wavelet transform is equivalent to two step filtering of the signal. The filter bandwidth is successively changed by decimation. Figure 2-10 shows the time frequency plane corresponding to a wavelet decomposition. In contrast with STFT, the time resolution becomes arbitrarily good at high frequency, while the frequency resolution becomes arbitrarily good at low frequencies. Note that in FWT, the number of points is gradually decreased through successive decimation. Thus if we start with a signal of 2^J points, then in the following level we have 2^{J-1} wavelet coefficients. Therefore, the maximum decomposition level is equal to J.

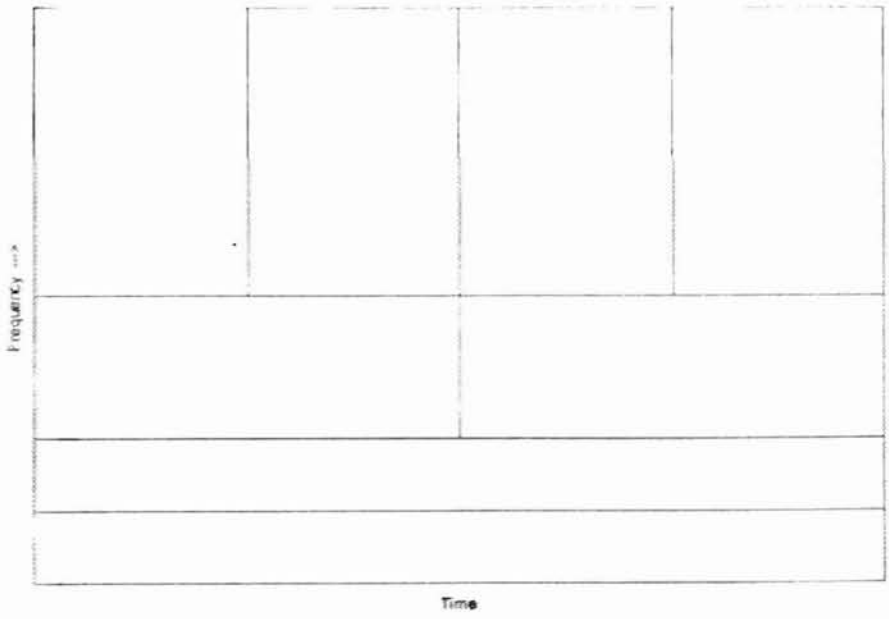

Figure 2-10: Time frequency plane of FWT

2.5 Wavelet Packet Decomposition

Whereas the wavelet transform provides one with more flexible time-frequency resolution properties as described, one possible drawback is that the frequency resolution is rather poor in high frequency region. Therefore, it faces some difficulties for discrimination between signals having close high frequency components.

Wavelet packets, a generalization of wavelet bases, are alternative bases that are formed by taking linear combinations of the usual wavelet functions [12] [18]. These bases inherit the properties such as orthonormality and time frequency localization from their corresponding wavelet functions. A wavelet packet function is a function with three indices: $W_{j,k}^{n}(t)$. As with usual wavelets, integers j and k are index scale and translation operations respectively:

$$W_{j,k}^{n}(t) = 2^{j/2} W^{n}(2^{j}t - k) \tag{2.29}$$

The index n =0,1,... is called *the modulation parameter* or the *oscillation parameter*. The first two wavelet packet functions are the usual scaling function and mother wavelet function respectively:

$$W_{0,0}^{0}(t) = \phi(t) \tag{2.30}$$

$$W_{0,0}^{1}(t) = \psi(t) \tag{2.31}$$

Wavelet packet functions for n =2,3,... are then defined by the following recursive relationships:

$$W_{0,0}^{2n}(t) = \sqrt{2}\sum_{k} h(k) W_{1,k}^{n}(t) \tag{2.32}$$

and

$$W_{0,0}^{2n+1}(t) = \sqrt{2}\sum_{k} g(k) W_{1,k}^{n}(2t-k) \tag{2.33}$$

where $h(k)$ and $g(k)$ are the QMF associated with predefined scaling function and mother wavelet function. To measure a specific time-frequency information in a signal, we simply take the inner product of the signal and that particular basis function. The wavelet packet coefficients of a function f can be computed via

$$w_{j,n,k} = < f, W_{j,k}^{n} > = \int f(t) W_{j,k}^{n}(t) dt \tag{2.34}$$

The idea of the usual wavelet decomposition as shown in Figure 2-9 is generalized to describe the calculation of wavelet packet coefficients $w_{j,n,k}$ of a discrete time signal. Computing the full wavelet packet decomposition of a discrete time signal involves applying both filters to the discrete time signal $[x_1, x_2, ..., x_n]$ and then recursively to each intermediate signal. The procedure is illustrated in Figure 2-11.

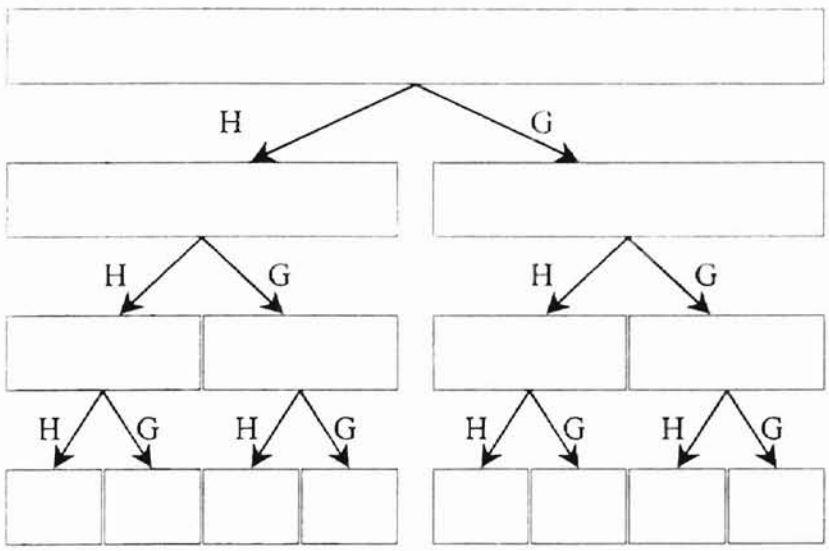

Figure. 2-11: Implementation of discrete wavelet packet decomposition

Note that the method of decomposition described above does not result in a WPT tree displayed in increasing frequency order. This is because aliasing occurs, which exchanges the frequency ordering of some nodes of the tree. A simple swapping of the appropriate nodes results in the increasing frequency ordering referred as the Paley ordering [18] of the tree as shown in Figure 2-12. Here the differences with Figure 2-11 is highlighted by the dashed lines. In this way, the leftmost node at each level will correspond to the lowest frequency band. In following section, we will use this representation for easier interpretation.

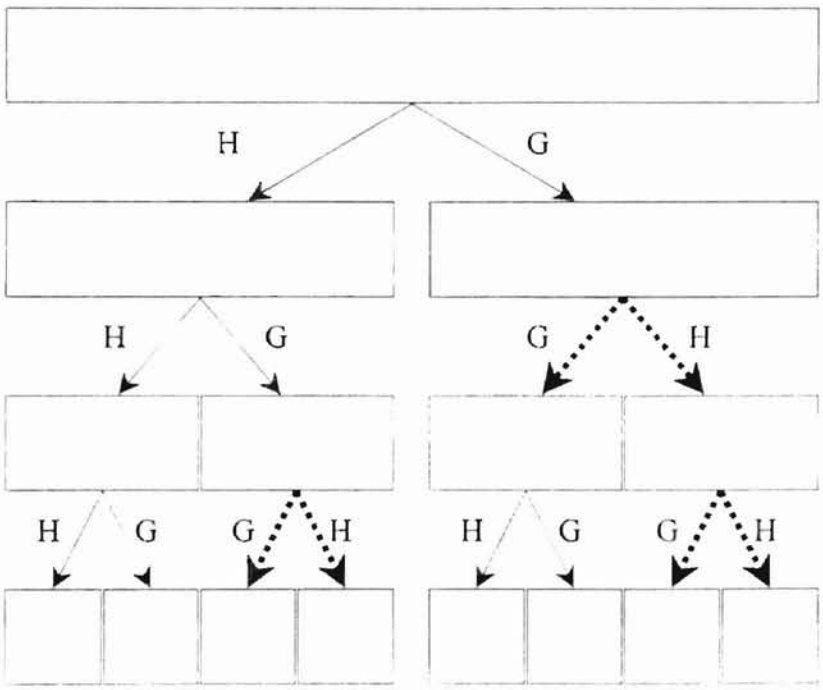

Figure 2-12: The WPD tree displayed in Paley order

Whereas the FWT decomposes only the low frequency components, WPT decompose the signal utilizing both the low frequency components and the high frequency components. This flexibility of a rich collection of abundant information with

arbitrary time-frequency resolution can allow extraction of the features that combine non-stationary and stationary characteristics.

2.6 Example of Wavelet Packet Decomposition

Through the thesis, the image representation as shown in Figure 2-13 will be employed to represent the full wavelet packet decomposition tree as shown in Figure 2-10 for interpretation purpose. For example, cell $w(0,0)$ refers to the root node in the decomposition tree, which corresponds to the time domain signal.

<table>
<tr><td colspan="8">$w(0,0)$</td></tr>
<tr><td colspan="4">$w(1,0)$</td><td colspan="4">$w(1,1)$</td></tr>
<tr><td colspan="2">$w(2,0)$</td><td colspan="2">$w(2,1)$</td><td colspan="2">$w(2,2)$</td><td colspan="2">$w(2,3)$</td></tr>
<tr><td>$w(3,0)$</td><td>$w(3,1)$</td><td>$w(3,2)$</td><td>$w(3,3)$</td><td>$w(3,4)$</td><td>$w(3,5)$</td><td>$w(3,6)$</td><td>$w(3,7)$</td></tr>
</table>

Figure 2-13: Image representation of WPD

Figure 2-14 shows a pulse which is extremely localized in time. The image representation of WPD for the signal is displayed in Figure 2-15 where the darker color corresponds to the higher coefficient value. The level 1 of the WPD image represents the time domain signal. In this level, the signal representation provides the best time resolution while no frequency information is available. Level 2 contains 2 nodes. The left-most node displays the WPD coefficient vector obtained from a lowpass-downsampling operation (H) on the time domain signal. In this level, one has two degrees

of frequency resolution, but due to the down-sampling, each node contains only half of the time resolution that exists in level 1. As one proceeds down to the bottom level, a tradeoff between time resolution and frequency resolution can be observed.

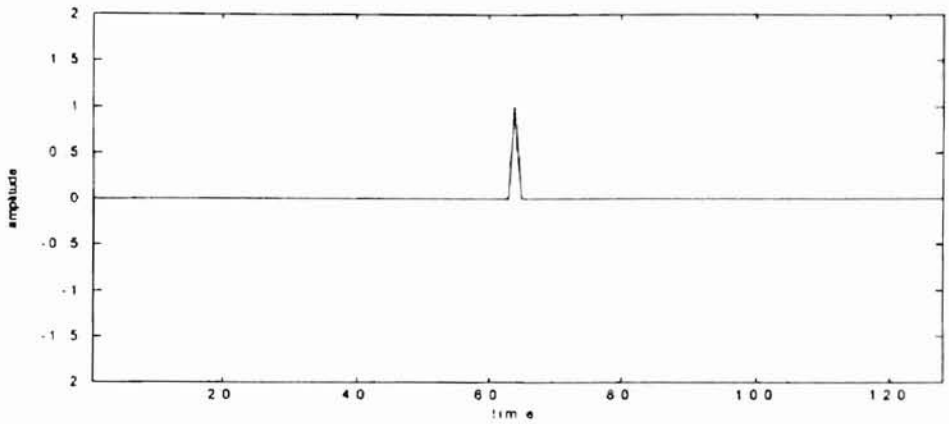

Figure 2-14: A signal localized in time domain

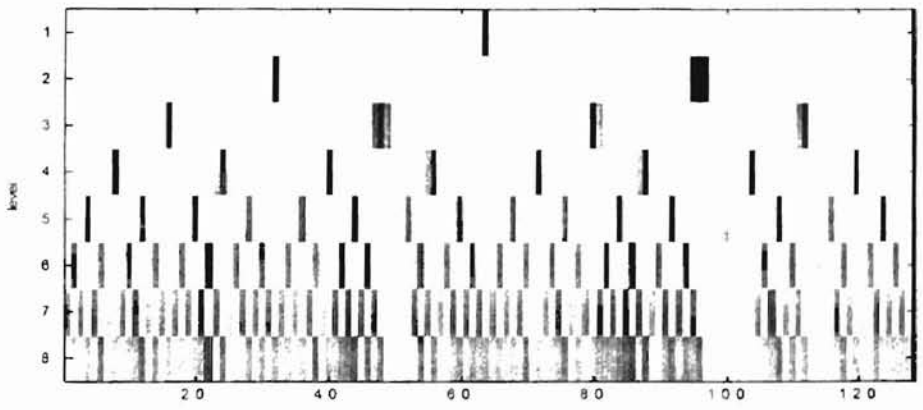

Figure 2-15: The WPD image representation of a time localized signal

Figure 2-16 shows a signal which is extremely localized in frequency domain. From Figure 2-17 we see that at each successive level, the information is gradually distributed into fewer and fewer wavelet packet coefficients.

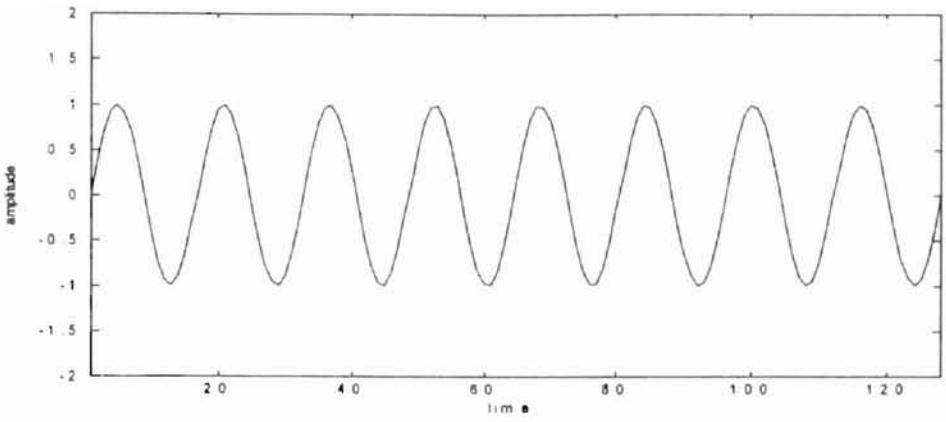

Figure 2-16: A signal localized in frequency domain

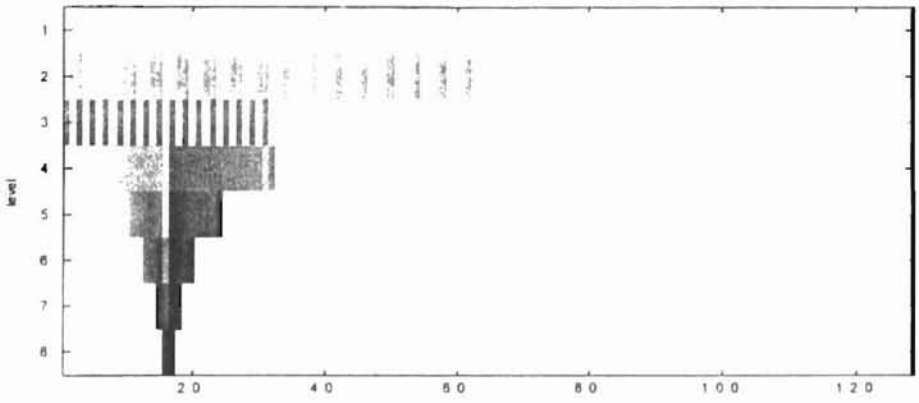

Figure 2-17: The WPD image representation of a frequency localized signal

The next example (Figure 2-18) shows a signal that is both localized in time and frequency domain. As can be seen from Figure 2-19, the information is effectively extracted at level 3. This information is, however, less focused either at the top or bottom level.

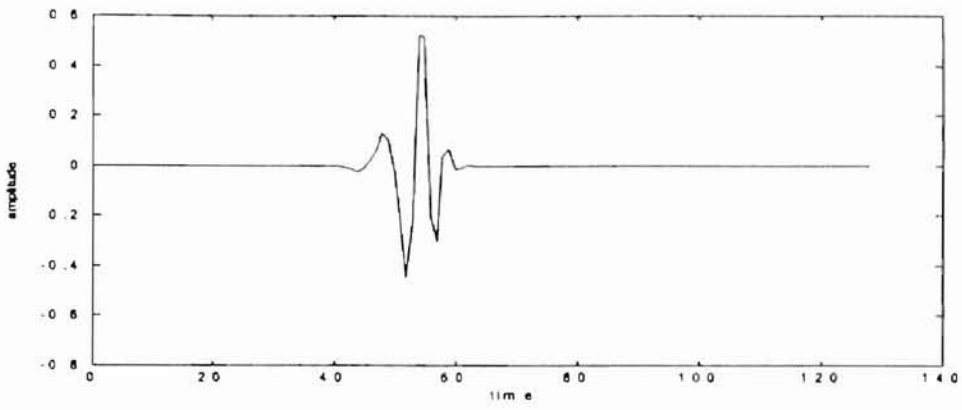

Figure 2-18: A signal localized in both time and frequency domain

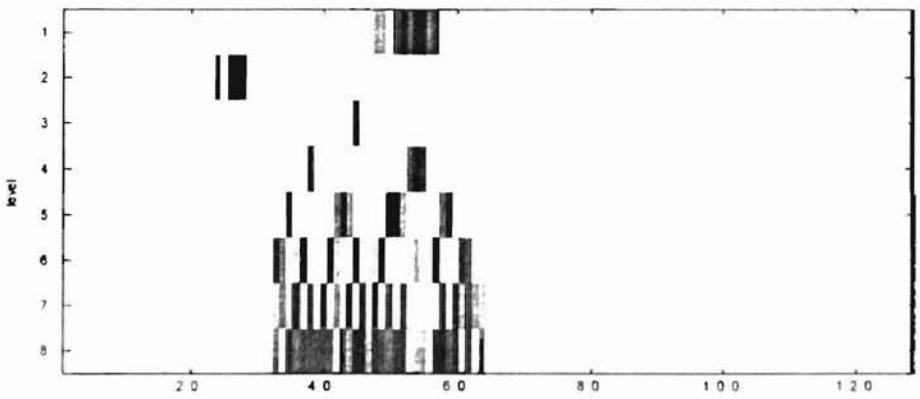

Figure 2-19: The WPD image representation of a time and frequency localized signal

It is clear now that WPD provides us flexibility that can adapt to the diverse time-frequency information in a signal. At levels near the top level time localized characteristics could be highly enhanced, while at levels near the bottom level frequency localized events are enhanced. Therefore, it is believed that the WPD provides the potential for dealing with signals exhibiting stationary and non-stationary characteristics.

CHAPTER III

CHOICE OF A FEATURE SET BASED ON WAVELET PACKET DECOMPOSITION

3.1 Overview

The wavelet packet transform is applied in classification problems based on time series vibration signatures. First, the vibration data is decomposed via the wavelet packet transform to extract the time-frequency dependent information. Features are then defined based upon the WPD coefficients. Second, simple statistical processing based on discriminant analysis is applied to identify a set of robust features that provides the most discrimination among the classes of vibration data. Then, a neural network classifier is trained based on this reduced feature set. With statistical-based feature selection criteria, a lot of feature components containing little discriminant information could be discarded, resulting in a feature subset with a reduced number of parameters. This will significantly ease the design of the classifier and enhance the generalization ability of the system. In following sections, we define the WPD based feature measurement used in this study. Then, we discuss some feature selection methods and present the ones applied in this study aiming to reduce the number of feature variables.

3.2 Feature Measures Based on WPD

One deficiency that wavelet bases inherently possessed is the lack of a translation invariant property. To illustrate this by example, consider two signals with a slight shift in time, as shown in Figure 3-1. When the two signals are decomposed via the wavelet packet transform, we can see appreciable differences between the two representations of the signals as shown in Figure 3-2 (a darker color corresponds to a larger WPD coefficient value). Therefore, direct assessment from all wavelet packet coefficients often turns out to be tedious or leads to inaccurate results.

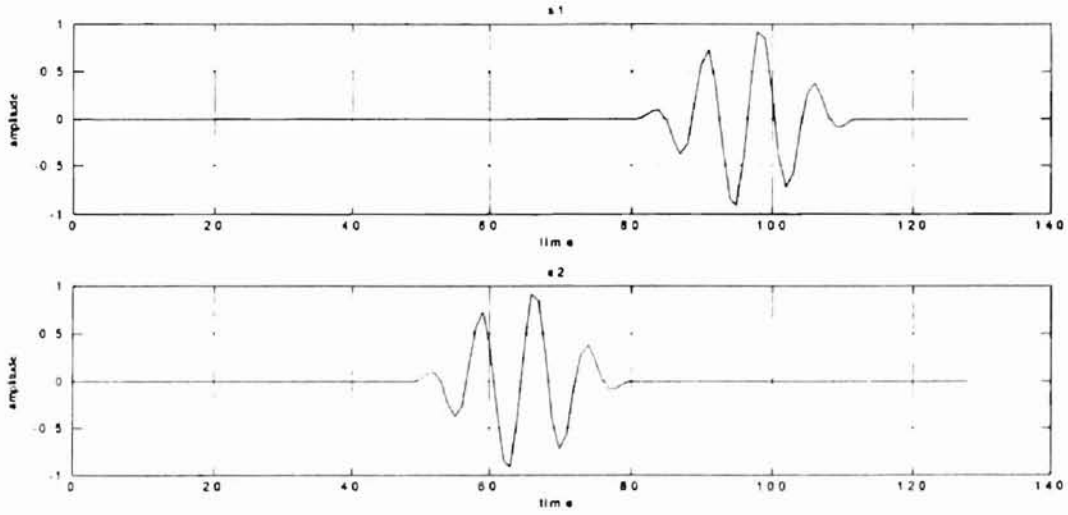

Figure 3-1: Signals with time shift

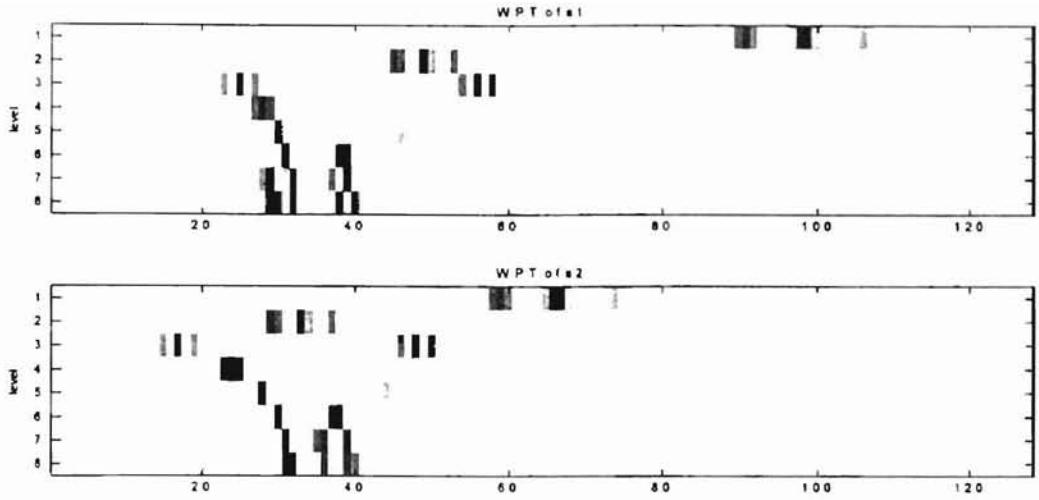

Figure 3-2: Wavelet packet decomposition of time shifted signals

Recall that each wavelet packet coefficient is given by:

$$w_{j,n,k} = < f, W_{j,n,k}(t) > = < f, 2^{j/2} W_n(2^j t - k) > \tag{3.1}$$

where j is a scaling parameter, k is a translation parameter and n is an oscillation parameter. Each $w_{j,n,k}$ coefficient measures a specific subband frequency content, controlled by the scale parameter j and the oscillation parameter n, of a signal around time instant $2^j t$.

We define the wavelet packet node energy as:

$$e_{j,n} = \sum_k w_{j,n,k}^2 \tag{3.2}$$

which measures the signal energy contained in some specific frequency band indexed by parameters j and n. In the sequel, we will call each (j,n) a wavelet packet node. Figure 3-3 displays the energy distribution that is calculated based on all coefficients in each wavelet packet node of the two signals given in Figure 3-1. We can see that node energy

values at level two, three or four show no clear difference between the two signals. This example reveals that the node energy representation provides us with a more robust signal feature for classification than using coefficients directly. In our strategy, each wavelet packet node energy value was defined as an individual feature component and was used as a robust rudimentary exploration of the specific signal features that provide useful information for classification purposes.

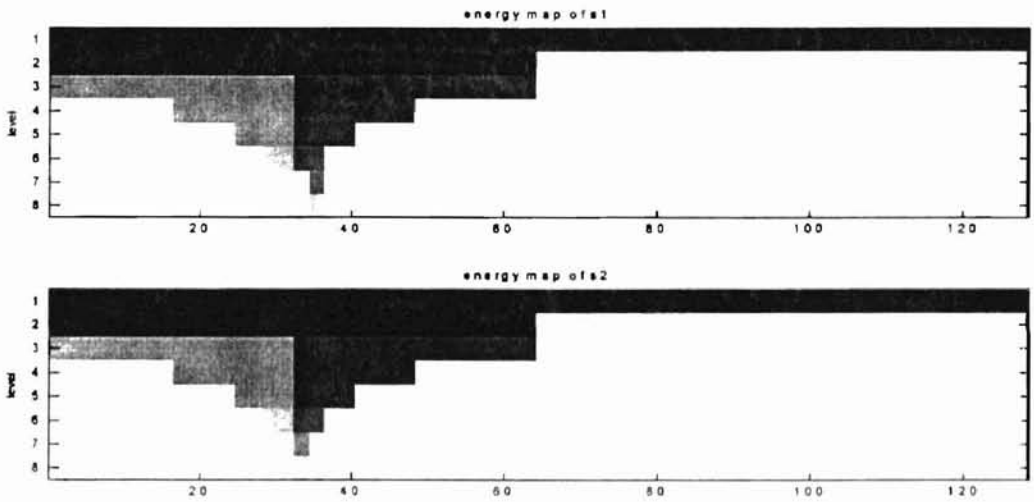

Figure 3-3. Wavelet packet node energy of time shifted signals

3.3 Dimension Reduction with Linear Transformation

One advantage of using wavelet packets transform to decompose a signal is that it allows us to examine different time-frequency resolution components in a signal. For example, by computing the full wavelet packet decomposition on a signal segment with $n = 2^J$ points for r resolution levels (where J, r denotes positive integers), the result is a

group of $2^1 + 2^2 + ... + 2^r = 2^{r+1} - 2$ sets of coefficients where each set corresponds to a wavelet packet node. If node energy as described before is used as a feature, we can obtain $2^{r+1} - 2$ feature components. However, direct manipulation on a whole set of node energies is prohibitive because the space normally has very high dimensionality and the existence of undesired components makes the classification unnecessarily difficult. In the training of a neural network classifier, it is desirable to use a lower dimensional vector as input to the neural network to ease the design of the classifier and enhance the generalization ability of the neural network classifier.

One popular technique in reducing the dimensionality is the Karhumen-Loéve (K-L) transform [19]. The K-L transform is optimal for signal representation in the sense that it provides the smallest mean square error for a given number of features. However, the features defined by the K-L transforms are not optimal for *class separability*. As an example, the data from two-class categories with a Gaussian distribution is shown in Figure 3-4. In the sense of K-L transform, the principal axis 1 with larger eigenvalue is a better vector than axis 2 to represent the vectors of this distribution. That is, the selection of axis 1 produces a smaller mean-square error of representation than the selection of axis 2 alone. However, as seen in the Figure 3-4, if the two distributions are mapped onto axis 1, the marginal density functions are heavily overlapped. On the other hand, if they are mapped onto axis 2, the marginal densities are well separated. Therefore, for classification purposes, axis 2 is a better feature than axis 1 alone, preserving more classification information.

As described previously, it is not the mean square error, as in the sense of K-L transform, but the classification accuracy that should be considered a primary criterion for

reducing the feature dimension. The ability to classify patterns relies on the implied assumption that different classes occupy distinct regions in the pattern space. Intuitively, the more distant the classes are from each other, the better the chance of successful recognition of class membership of patterns. One transformation associated with this assumption is based on within and between class scatter matrices that are used in linear discriminant analysis (LDA) of statistics [20]. The idea is to find a linear transformation that projects the samples onto a lower dimensional space in which the variability of samples within each class is as close as possible, and the dispersion of the class mean vectors about the mean vector is as separated as possible.

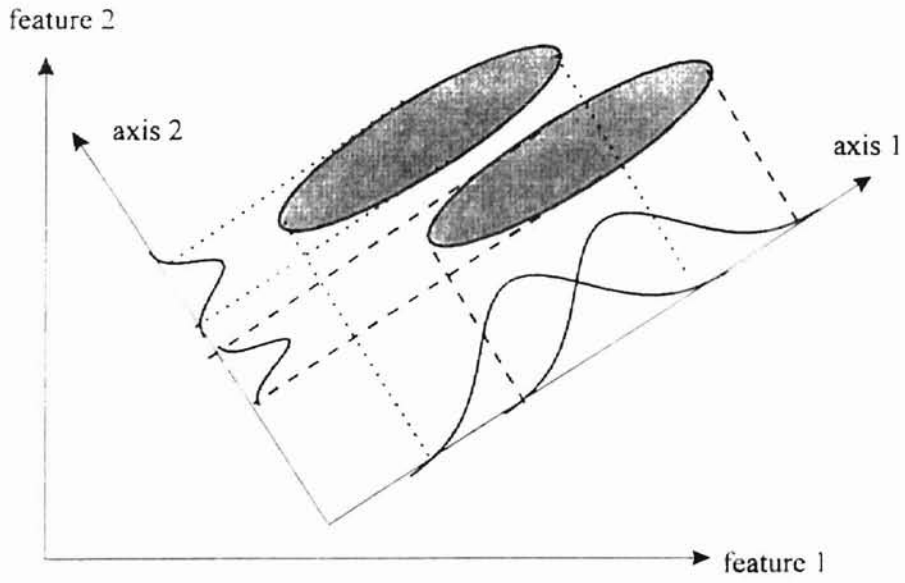

Figure 3-4: An example of feature extraction for classification

Specifically, consider an L-class problem. The variability of samples within each class is measured by the class sample covariance matrices:

$$S_c = (1/N_c)\sum_{i=1}^{N_c}(x_i^c - m_c)(x_i^c - m_c)^T,\ c = 1.2,\ldots, L \tag{3.3}$$

where x_i^c is a sample vector belong to class c, N_c is the number of samples belong to class labeled c and m_c is mean vector of class c:

$$m_c = (1/N_c)\sum_{i=1}^{N_c} x_i^c \tag{3.4}$$

In this way, the overall within-class variability could be estimated by the sample covariance matrix:

$$S_w = \sum_{c=1}^{L} p_c S_c \tag{3.5}$$

where p_c is the priori probability of class c. Similarly, the between-class covariance matrix measures the dispersion of the class mean vectors about the overall mean vectors:

$$S_b = \sum_{c=1}^{L} p_c (m_c - m)(m_c - m)^T \tag{3.6}$$

where m represents the expected vector of the mixture distribution and is given by

$$m = \sum_{c=1}^{L} p_c m_c \tag{3.7}$$

Now if $\bar{x} = A^T x$ denotes a linear transformation of the original variables, then the between- and within-class matrices in the transformed space are just $\overline{S_b} = A^T S_b A$ and $\overline{S_w} = A^T S_w A$. The goal is to find a subspace where the ratio of S_b and S_w are maximized. In this case it may be measured by the ratio of the determinant of the proceeding matrices (the determinant, being the product of the eigenvalues, is the product of the variance in the principal directions). The problem could thus be formulated as: find a transformation $\overline{A}$ such that:

$$\overline{A} = \arg \underset{A}{Max} \frac{|A^T S_b A|}{|A^T S_w A|} \tag{3.8}$$

The solution for Eq. (3.8) is given by the $\min(n, L-1)$ eigenvectors of $S_w^{-1} S_b$ [20]. Once the transformation map $\overline{A}$ is obtained, then the feature vector $\overline{A}^T x$ is computed for each sample, and finally it is assigned to the class which has the mean vector closest to this feature vector.

Although the vector found by LDA works well in most cases, several drawbacks might occur in practice. First, when we apply LDA to extract the discriminant feature vector, the mathematical procedure automatically combines the feature extractor and the classifier in a linear form. By restricting the form or criterion of the mapping, we implicitly assume an oversimplistic model of the pattern recognition system. Such a situation will arise if the classes are not linearly separable, and we restrict the feature extractor to a linear form.

Moreover, LDA involves the computation of the inverse of the covariance matrices, it may lead to numerical problems, especially when the matrices are estimated based on a limited data set. In our application on the classification of vibration signal data collected from multi-sensors, we might have thousands of time-frequency feature components while only hundreds of training samples are available. For example, given a 256 point signal, full decomposition of the signal to the 7th level and use of node energy as the feature component will result in a 254 dimensional feature vector. Combining all feature vectors from multiple sensors, say 8, will result in a 2032 dimension vector. However, only a few eigenvalues, such as 10, are dominant, so that

$$\lambda_1 + \lambda_2 + \ldots + \lambda_{2032} \cong \lambda_1 + \ldots + \lambda_{10} \tag{3.9}$$

This means that in a practical sense we are handling S_w with rank 10, even though the mathematical rank of S_w is still 2032, i.e. $\lambda_i \neq 0, \forall i$. In the calculation of S_w^{-1} the determinant $|S_w|$ is $\prod_{i=1}^{2032} \lambda_i$ and 2032-10=2022 λ_i are very close to zero. Suppose $\lambda_1 + ... + \lambda_{10} = 0.9$ out of $\sum_{i=1}^{2032} \lambda_i = 1$, then

$$\prod_{i=1}^{10} \lambda_i \times \prod_{j=11}^{2032} \lambda_j = \prod_{i=1}^{10} \lambda_i \times (0.1/2022)^{2022} \cong 0 \tag{3.10}$$

for the assumption $\lambda_{11} = \lambda_{12} = ... = \lambda_{2032} = 0.1/2022$.

This indeed leads to some computational difficulty in handling such a near-singular matrix. For this reason, we resort to employing the feature selection in feature measurement as described in the following section, which considers the numerical problems of calculating the inverse of covariance matrices as LDA does. Instead of trying to find a linear transformation to reduce the dimensionality, we evaluate the discriminant power of each individual feature component and discard those feature components containing little class separability information as measured by selected criterion. Then, neural networks is employed as a classifier to deal with nonlinearly separable case in the feature space.

3.4 Dimension Reduction with Feature Selection

The idea of feature selection in feature measurement space is to select the feature components that contain discriminant information and discard those feature components that provide little information useful for classification purposes [21]. Specifically, the feature component $\{f_k \mid k = 1,2,...,n\}$ is ranked:

$$J(f_1) \geq J(f_2) \geq \ldots \geq J(f_d) \geq \ldots \geq J(f_n) \quad (3.11)$$

where $J(.)$ is a criterion function for measuring the *discriminant power* of a specific feature component. The feature subset can be selected from the available features that have larger criterion function values.

To obtain a clearer picture of measuring the discriminant power of a feaure, it is essential to introduce a concept of probabilistic structure of classes. Consider the probability density function of class c1 and c2 given in Figure 3-5. For a specific feature variable x, if $p(x \mid c1)$ is zero for all x such that $p(x \mid c2) \neq 0$ as illustrated in Figure 3-5(a), then these two classes can be fully separable. On the other hand, when $p(x \mid c1) = p(x \mid c2)$ as in Figure 3-5(b), it is impossible to distinguish elements of class c1 from those belonging to c2. Intuitively, a criterion function for evaluating the discriminant power of a feature could be assessed by measuring the overlap between $p(x \mid c1)$ and $p(x \mid c2)$. A high overlap corresponds to a low discriminant power and vice versa.

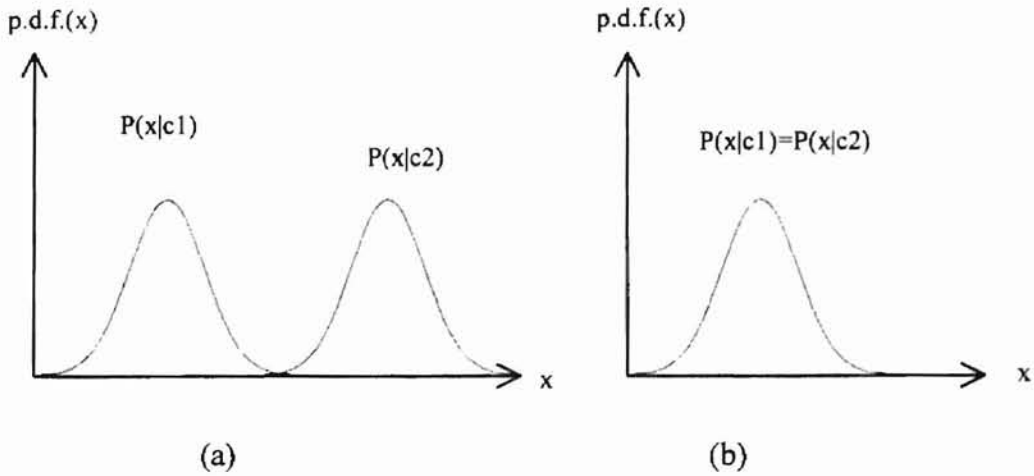

Figure 3-5: Probability density functions of (a) two well separated classes and (b) two completely overlapping classes

In general, a criterion function for measuring the overlap between classes has the following properties [21]:

(a) The measure is minimum when the conditional probability density function for class c1 and c2 are identical, i.e.

$$J(.) = 0, \text{ if } p(x \mid c1) = p(x \mid c2). \tag{3.12}$$

(b) The measure is non-negative.

(c) The measure attains a maximum when the classes are disjoint, i.e.

$$J(.) = \max, \text{ if } p(x \mid c1) = 0 \text{ when } p(x \mid c2) \neq 0, \forall x. \tag{3.13}$$

Although the above properties provide an intuitive justification of their suitability for feature selection, their relative potential can be assessed only if their relationship to the classification error is known. Nevertheless, these measures are closely related to the error probability. This relationship is a consequence of the fact that the measures gives a direct indication of the amount of the overlap of the class probability densities. Some criterions suggested for feature selection are listed below [21]:

Chernoff distance:

$$J(.) = -\ln \int p^{s}(x \mid c1) \cdot p^{1-s}(x \mid c2) dx, \tag{3.14}$$

where s is a parameter from the interval [0,1].

Matusita distance:

$$J(.) = \{ \int [p(x \mid c1)^{1/2} - p(x \mid c2)^{1/2}]^2 dx \}^{1/2} \tag{3.15}$$

In this study, however, we adopt a simple yet efficient criterion function known as Fisher's criterion [20]. In a two classes problem it is given by:

$$J_{f_k}(i,j)=\frac{\left|\mu_{i,f_k}-\mu_{j,f_k}\right|^2}{\sigma_{i,f_k}^{\ 2}+\sigma_{j,f_k}^{\ 2}} \tag{3.16}$$

where μ_{i,f_k}, μ_{j,f_k} are the mean values of the k-th feature, f_k, for class i and j, and $\sigma_{i,f_k}^{\ 2}$, $\sigma_{j,f_k}^{\ 2}$ are the variance of the k-th feature, f_k, for class i and j correspondingly. When there are more than two classes of data, the general approach is to take the summation of the pairwise combinations of $J_{f_k}(i,j)$:

$$J_{f_k}=\sum_{i=1}^{L-1}\sum_{j=i+1}^{L}J_{f_k}(i,j) \tag{3.17}$$

as an estimation of discriminant power for the specific feature f_k. Here L represents the number of classes in the problem. Eq. (3.17) provides us with a measure to evaluate the effectiveness of the "global" feature that is simultaneously suitable to differentiate all classes of signals. For a small number of classes, this approach may be sufficient. The more signal classes, the more ambiguous the equation (3.17) becomes. A large value of (3.17) may be due to a few significant terms with negligible majority (a favorable case) or to the accumulation of many terms with relatively small values (an unfavorable case). A feature that can effectively differentiate a pair of classes of signals, i.e. with a large discriminant measure as calculated by Eq. (3.16), might be averaged during the pairwise summation. To avoid such a problem, we propose two different approaches as described below.

Approach I:

Instead of trying to select features that are effective for the entire multi-class problem globally as measured by Eq.(3.17) , we select a feature subset based on Eq. (3.16) for each possible pair of classes. Then, we take the union of feature components selected from each pair of classes to form the final feature vector. Specfically, given an L-class problem with n feature components, the selection process is detailed in the following:

1. For each possible class pair $\{(i,j) \mid i = 1,2,...,L-1, j = i+1, i+2,...,L\}$, calculate the discriminant power measure for each feature component, f_k, i.e.:

$$J_{f_k}(i,j) = \frac{|\mu_{i,f_k} - \mu_{j,f_k}|^2}{\sigma_{i,f_k}^{\ 2} + \sigma_{j,f_k}^{\ 2}} \tag{3.18}$$

2. For each class pair, sort $J_{f_k}(i,j)$ such that:

$$J_{f1}(i,j) \geq J_{f2}(i,j) \geq ... \geq J_{fd}(i,j) \geq ... \geq J_{fn}(i,j) \tag{3.19}$$

Determine the feature subset $F_{i,j}$ for each class pair by selecting d feature components that have maximum $J_{f_k}(i,j)$ value:

$$F_{i,j} = \{f_k \mid k = 1,2,...,d\}, i = 1,2,...,L-1; j = i+1, i+2,...,L. \tag{3.20}$$

3. Form the final feature set by taking the union of each feature subset:

$$F_{final} = \{\bigcup_{i=1}^{L-1} \bigcup_{j=i+1}^{L} F_{i,j}\} \tag{3.21}$$

Approach II:

Another approach to avoid the influence of the pairwise summation process is similarly suggested by Watanabe [22]. Given an L-class signal classification problem,

we can consider the class k signals as the conceptual opposite of the class $\tilde{k}$ signals, which is the ensemble of data belonging to classes other than the $\tilde{k}$ class. Then, we apply the Fisher's criterion as was done in two-class problems to evaluate the discriminant power of each individual feature component.

1. For each class $k = 1,2,...,L$, we partition the data set to be class k signals and class $\tilde{k}$ signals. In this way, we can get L sets of data that can be used for selecting features.
2. For each of the L sets, use Fisher's criterion to evaluate the discriminat power for each feature component,

$$J_{f_k}(k,\tilde{k}) = \frac{|\mu_{k,f_k} - \mu_{\tilde{k},f_k}|^2}{\sigma_{k,f_k}^{\ 2} + \sigma_{\tilde{k},f_k}^{\ 2}},\ k = 1,2,...,L \tag{3.22}$$

3. For each of L sets, select d feature components that have larger criterion values for each set.
4. The final feature set is determined by taking the union of the feature components of L sets with d feature components selected form step 3.

Suitable feature components which offer favorable separation of classes are found as described; many classifiers could then be designed based on these features. The feedforward neural network is employed in the study because of its capability in dealing with nonlinearly separable distributions.

3.5 Using Neural Network as Classifier

Once suitable features have been extracted and selected from the vibration data as described, it is then necessary to determine the fault type based upon these features. Ideally, the features for normal and faulty conditions will occupy non-overlapping areas in the feature space. If not, then the classification algorithm will have to approximate a Bayes classifier [23].

Consider an L-class problem, the probability that a particular pattern, x, comes from class, c_i, $i = 1,2,...,L$, is denoted $p(c_i \mid x)$. If the pattern classifier decides that x came from c_j when it actually came from c_i, it incurs a loss, denoted $l(i \mid j)$. As pattern x may belong to any one of L classes under consideration, the average loss incurred in assigning x to class c_j is

$$r_j(x) = \sum_{k=1}^{L} l(k \mid j) p(c_k \mid x) \tag{3.23}$$

In general, the loss for a correct decision is zero, and it has the same nonzero value (say, 1) for any incorrect decision. i.e.

$$l(k \mid j) = \delta_{k,j} \tag{3.24}$$

where

$$\delta_{k,j} = \begin{cases} 1, k = j \\ 0, k \neq j \end{cases}. \tag{3.25}$$

Then the loss of assigning a pattern x to class c_j becomes

$$r_j(x) = \sum_{k \neq j} p(c_k \mid x). \tag{3.26}$$

The classifier has L possible classes to choose from for any given unknown pattern x. If it computes $r_j(x), j = 1,2,...,L$, for each pattern x, and assigns the pattern to the class with smallest loss, then total average loss with respect to all decisions will be minimum. The classifier that minimizes Eq. (3.26) is called the *Bayes Classifier*. Thus the Bayes classifier assigns an unknown pattern vector, x, to class c_i if:

$$r_i(x) < r_j(x) \text{ for } j = 1,2,...,L;\ j \neq i. \tag{3.27}$$

Substituting Eq. (3.26) into Eq.(3.27), the decision rule is then to choose label c_i if

$$\sum_{k \neq i} p(c_k \mid x) < \sum_{k \neq j} p(c_k \mid x),\ k = 1,2,...,L. \tag{3.28}$$

Note that each side of the Eq. (3.28) has all but one term in common. The decision rule then becomes to assign x to c_i if, for all $i \neq k$,

$$p(c_i \mid x) > p(c_k \mid x). \tag{3.29}$$

For the decision rule based on Eq. (3.29) to hold, the posteriori density functions $p(c_i \mid x); i = 1.2,...,L$ must be known; in practice it must be estimated from the available data set. To obtain the estimates of the posteriori density functions, neural networks are applied in the study for the following reasons. First, neural networks are universal approximators in the sense that they can theoretically approximate any continuous input-output mapping to any desired degree of accuracy. Hence, they can be used to approximate the *posteriori* function $p(c_i \mid x)$. Additionally, neural networks are inherently nonlinear in the activation function; they have the ability to capture the underlying non-linearity for the generation of incoming data.

CHAPTER IV

TEST RESULTS ON WESTLAND HELICOPTER GEAR BOX VIBRATION DATE SET

4.1 Data Description

In this chapter, the feasibility of the wavelet packet based feature classification technique was examined through numerical simulations on a real data set known as the Westland data set. The Westland data set [24] was chosen because it has been analyzed by a number of other researchers and because it is considered as a benchmark data set in the field. The vibration data used for simulation is archived at the Applied Research Laboratory at Penn State University, known as the Westland data set. In this data set, vibration data are recorded from an aft main power transmission of a U.S. Navy CH-46E helicopter. Vibration data are collected using eight accelerometers mounted at the known fault sensitive locations of the helicopter gearbox. The data are recorded for various seeded faults including the no defect case, listed in Table 4.1. Nine torque levels, ranging from 27 % up to 100 %, and various fault severity levels are applied. One tachometer is placed on the aft transmission in place of the rotor position motor. The tach signal is a 256 pulse-per-revolution signal with a once-per-revolution signal superimposed on it. Based on its position in the gearbox, one revolution describes a complete rotation of the

rotor position output, not that of the main shaft. The vibration data are sampled at 103,116.08 Hz rate. With the approximate 100 kHz sampling rate, there are between 897 and 904 samples within the period defined by the tachometer signal.

Table 4.1: Westland helicopter gearbox data description

Fault Type Number	Description
1	No Defect
2	Planetary Bearing Corrosion
3	Input Pinion Bearing Corrosion
4	Spiral Bevel Input Pinion Spalling
5	Helical Input Pinion Chipping
6	Helical Idler Gear Crack Propagation
7	Collector Gear Crack Propagation
8	Quill Shaft Crack Propagation

4.2 Signal Segmentation

For utilization of the Fast Wavelet Packet Transform algorithm, each 1024 time series data points of vibration signal is defined as a sample vector to be analyzed. The reason for using 1024 points is it covers one period defined by the tachometer period. It is reasonable to assume that fault symptoms can be fully described within the period. Figure 4-1 shows two signal segments and corresponding power spectrum for normal mode and fault 3. Looking at the spectrum of the vibration data segment, we observe a long flat

region toward the end of the frequency range; it is thus inferred that the bandwidth of the signal is much less than the sampling frequency. Based on this observation, the sample vector is first down sampled by 4 to yield a 256 point signal segment. This lowers the computational complexity without losing much information of the signal.

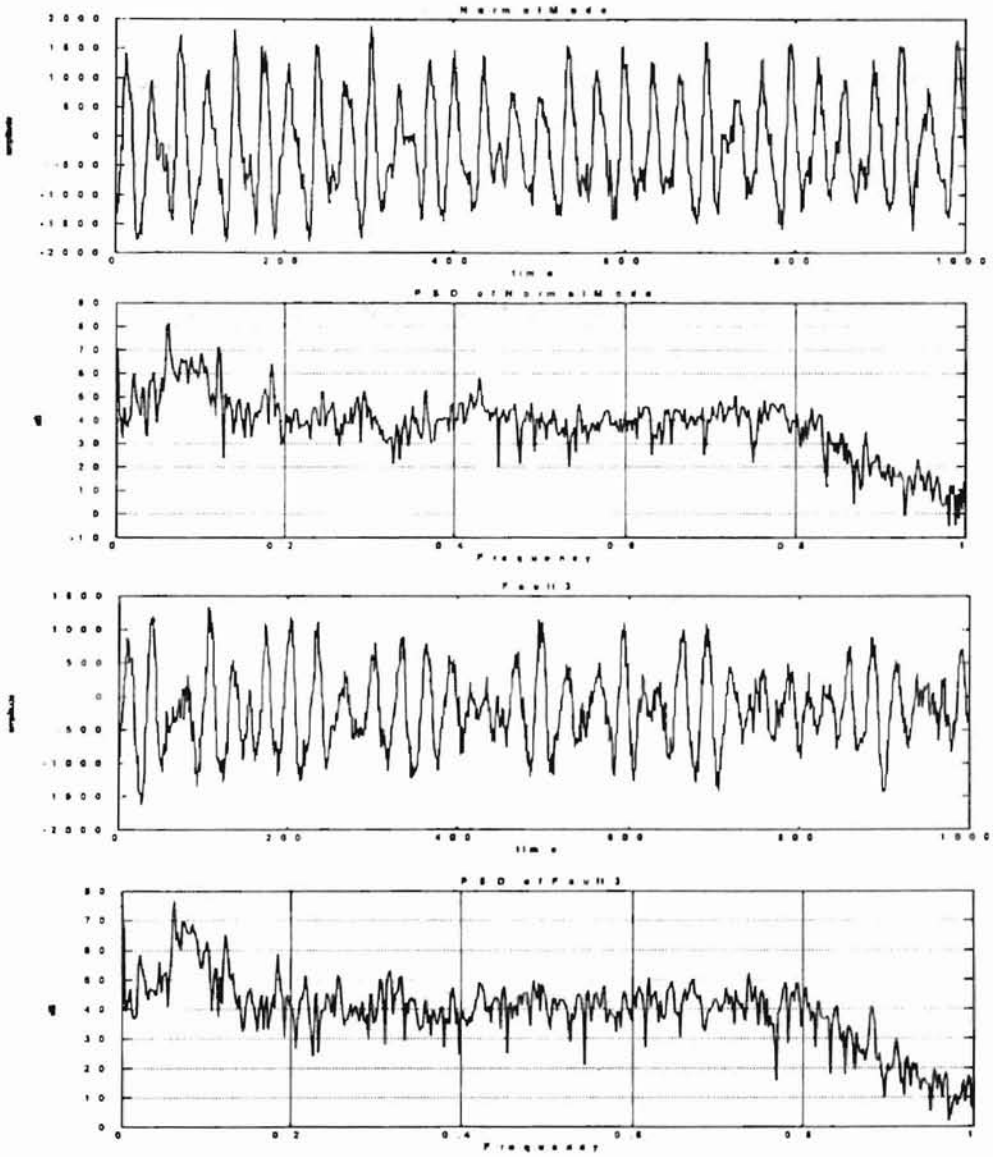

Figure 4-1: Typical vibration signals and corresponding PSD. The frequency axis is in units of π × radians.

Additionally, if a random signal has a nonzero mean, its power spectrum has an impulse at zero frequency. If the mean is relative large, this component will dominate the spectrum estimate, causing low-amplitude, low-frequency components to be obscured by the leakage. Therefore, in practice the mean is often estimated, and the resulting estimate is subtracted from the random signal before computing the power spectrum estimate. Although the sample mean is only an approximate estimate of the zero frequency components, subtracting it from the signal often leads to a better estimate at neighboring frequencies [25].

4.3 Generation of Training Data Set / Testing Data Set

There are total of 68 data sets available, which correspond to nine different torque levels and 8-class conditions. For each torque level, not all fault signals are available. Each file contains 412464 data points. The 412464 data points are segmented to 400 sample segments containing 1024 data points each. In this study, the first 50 samples collected represent the training data set, while the following 150 samples are used as testing data set. In this study, only the data set corresponding to torque level 100% is used for evaluation.

4.4 System Description

In the following simulations, each vibration signal segment is transformed in to a wavelet packet based energy vector as described in Section 3.4. The proposed two feature selection methods are then employed to identify a subset of feature components that will be used as the input to the neural network classifier. The steps are summarized below:

1. A seven-level wavelet packet decomposition is found for each vibration signal segment.
2. Energy based feature measures, as discussed in Section 3.2, are found for each wavelet packet decomposition of signal segment from step 1. This results in a 254 dimensional feature vector.
3. Identify a subset of feature components, as discussed in Section 3.4, to form input vector for neural network classifier.

4.5 Test Result Using One-Sensor Data

In the following simulations, we conducted tests on features extracted from both Fourier based features and wavelet packet based features for assessing the applicability of wavelet packet based analysis as a tool for vibration signatures. The Fourier based features are defined as the power spectrum of a 256-point signal segment, and the result is a 129 dimension vector where each component corresponds to one of the 129 uniform frequency band energies. The two feature selection processes, approach I and II described in Section 3.4, are applied on both wavelet packet based feature components and Fourier based feature components to select the best discriminant feature components. The obtained feature components are then used as input to train the neural network classifier. For each of the feature selection approaches, the eight highest discriminant (d=8) feature components (out of 254) are used to form the final feature vector. Table 4.2 provides the dimension of the final feature vector for the two approaches. In the following, the feature selection method *approach I*, as described in Section 3.4, is designated as *PWM* while the

approach II is designated as *KNK*. In general, the computation cost for PWM will be less than that of KNK.

The network architecture is D-D-8, where D is the dimension of the final feature vector. In the training process, the network is trained until the mean square error is below 0.01, or the maximum epochs (=10000) is reached. In practice the neural network will not produce a perfect decision, i.e. only one 1 in the output neuron while others are all 0's, and might produce values between zero and one. Hence, it was decided to use the maximum output value as the most likely fault condition. In all simulations, a clear winner can always be identified. The classification results are shown in Table 4.3 to Table 4.6. Note that the unit of error is % in all classification results. Tr. Err. is referring to the training error, while Test Err. is referring to the testing error.

Note that the performance of using different sensor data shows significant differences. For example, the testing errors of using data from sensor 5, 6, 8 are relatively higher than those of other sensor data for both the wavelet packet based and Fourier based approaches. It is thus inferred that some sensors are not sensitive to the detection of specific fault symptom. This suggests the need to use multiple sensor data to search the class specific features.

Table 4.2: Dimension of final feature vector using one sensor

	Wavelet packet feature							
	Sensor 1	Sensor 2	Sensor 3	Sensor 4	Sensor 5	Sensor 6	Sensor 7	Sensor 8
PWM	32	33	42	31	38	47	39	50
KNK	34	36	46	45	34	34	50	33
	Fourier Feature							
	Sensor 1	Sensor 2	Sensor 3	Sensor 4	Sensor 5	Sensor 6	Sensor 7	Sensor 8
PWM	31	37	31	31	43	51	37	54
KNK	28	30	37	35	31	35	37	35

Table 4.3: Classification results (Sensor 1 & 2)

		Sensor 1		Sensor 2	
		WPT	FT	WPT	FT
PWM	Tr. Err.	3.25	0.75	1	0
	Test Err.	21.92	4.25	4	2.17
KNK	Tr. Err.	2.75	1.25	1.00	0.25
	Test Err.	24.50	4.58	4.33	1.92

Table 4.4: Classification results (Sensor 3 & 4)

		Sensor 3		Sensor 4	
		WPT	FT	WPT	FT
PWM	Tr. Err.	0.75	0.25	2.25	1.25
	Test Err.	6.75	1.75	6.42	5.83
KNK	Tr. Err.	0.75	0.50	1.25	1.75
	Test Err.	7.25	1.42	8.00	5.08

Table 4.5: Classification results (Sensor 5 & 6)

		Sensor 5		Sensor 6	
		WPT	FT	WPT	FT
PWM	Tr. Err.	0.25	0.75	0.75	1.25
	Test Err.	50.33	50.33	62.92	57.92
KNK	Tr. Err.	1.25	1.25	4.00	3.25
	Test Err.	50.83	53.67	61.58	59.75

Table 4.6: Classification result using one sensor data (Sensor 7 & 8)

		Sensor 7		Sensor 8	
		WPT	FT	WPT	FT
PWM	Tr. Err.	1	1.5	0.5	0
	Test Err.	2.25	2.42	62	46.83
KNK	Tr. Err.	1.50	0.00	3.25	1.75
	Test Err.	5.08	3.08	61.08	45.17

4.6 Test Result Using Eight-Sensor Data

In the following tests, feature components from all eight sensors are all used to begin the feature selection process; i.e. the comparison of discriminant power is conducted on features coming from all eight sensors data. Table 4.7 provides the dimension of the final feature vector for the two approaches based on wavelet packet features and Fourier features respectively. All simulation settings, network architecture and MSE goal, are the same as previous tests. The classification results are displayed in

Table 4.8 and Table 4.9 corresponding to feature selection method PWM and KNK respectively.

In the tables which follow, FT refers to the Fourier based features while WPT refers to wavelet packet based features. The results show the performance is much improved when combining data from all sensors. It could be concluded that some fault symptom could only be detected by some sensors. If we use only one sensor, the crucial information for the specific fault symptom may not be detected and the overall classification performance may be lower. Additionally, it is observed that the performance of the Fourier based approach shows slightly better results than the wavelet packet based approach. It is concluded that features providing discriminant information may demonstrate narrow-band frequency characteristics in this data set. In such cases, the Fourier based approach is ideally the better candidate for extracting signal features.

Recall that there is a slight amount of frequency overlap among the wavelet basis functions, thus a particular frequency may be sensed by two different basis functions. This frequency leakage may lead to worse performance using wavelet packet based features. Nevertheless, the wavelet packet transform is still able to extract the essential discriminant features and achieve a satisfactory performance.

Table 4.7: Dimension of final feature vector

	Wavelet packet feature							
	d=1	d=2	d=3	d=4	d=5	d=6	d=7	d=8
PWM	7	14	26	35	42	50	58	62
KNK	8	16	24	32	39	47	55	63
	Fourier Feature							
	d=1	d=2	d=3	d=4	d=5	d=6	d=7	d=8
PWM	9	16	24	27	29	30	32	38
KNK	8	16	24	32	40	48	51	56

Table 4.8: Classification results (8-sensor data; PWM)

		d=1	d=2	d=3	d=4	d=5	d=6	d=7	d=8
WPT	Tr. Err.	0.5	0	0	0	0	0	0	0
	Test Err.	0.92	0.17	0	0.17	0.25	0	0.08	0.25
FT	Tr. Err.	0	0	0	0	0	0	0	0
	Test Err.	0	0	0	0.25	0	0	0	0

Table 4.9: Classification results (8-sensor data; KNK)

		d=1	d=2	d=3	d=4	d=5	d=6	d=7	d=8
WPT	Tr. Err.	0	0.25	0	0	0	0	0	0
	Test Err.	0.08	0	0	0	0	0.08	0.17	0.08
FT	Tr. Err.	2.75	0	0	0	0	0	0	0
	Test Err.	2.17	0.33	0	0.08	0	0	0	0

4.7 Test on Data Corrupted by Additive White Noise

A measured vibration signal can be considered to have the following components: the fault response caused by faulty equipment; vibration from normal machine components, vibration of neighboring machinery and measurement variation. In monitoring vibration signals, we considered the *noise* to consist of vibration from machine components (other than the faulty response), neighboring machinery and measurement noise. The presence of noise complicates the monitoring tasks in two forms: by masking the signal of interest and by increasing the vibration values beyond monitoring criteria, when in fact the component being monitored experiences no sign of malfunction. To test further the feasibility of the wavelet packet based feature extraction technique on the presence of noise, simulated data are artificially generated by adding different types of noise to the original vibration signals. The goal is to investigate the robustness of wavelet packet based features when the data are subjected to the presence of noise. In following simulations, the signals are first corrupted with artificially generated noise under different SNR, then the WPT and FT are applied on the corrupted signal to obtain the signal's time frequency feature. At last, the proposed feature selection method is used to identify discriminant feature components that will be used as input to train a neural network classifier. In this study, we use three types of noise to model the vibration signal other than the signal being monitored.

The first type of noise model used is white Gaussian noise, where no frequency is dominating as shown in Figure 4-2. This noise has an AR(0) model:

$$X_k = a_k \tag{4.1}$$

where a_k are normally distributed with mean zero and variance σ_a^2. In this study, we use the Matlab® function *randn*() to generate a_k.

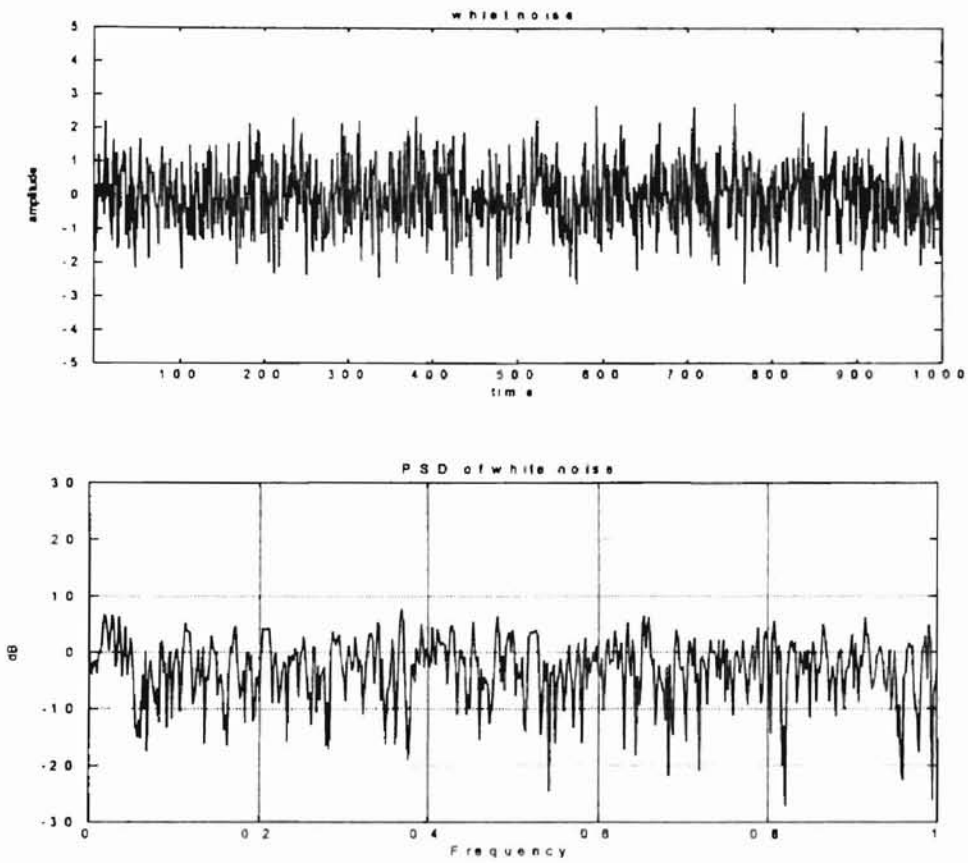

Figure 4-2: White Gaussian noise and its power spectrum. The frequency axis is in units of $\pi \times$ radians.

Tables 4.10 through Table 4.13 show the results under different SNR. The results reveal that the wavelet packet based approach demonstrates better results than the Fourier based approach. It was also observed that the difference of performance, between wavelet packet approach and Fourier based approach, is even higher when the noise power is increased.

Table 4.10: Classification results (white noise; SNR=0dB; PWM)

		d=1	d=2	d=3	d=4	d=5	d=6	d=7	d=8
WPT	Tr. Err.	0.25	0	0	0	0	0	0	0
	Test Err.	0.5	0.08	0.17	0.08	0.08	0.33	0.33	0.42
FT	Tr. Err.	14.25	1.25	1	0.25	0.5	0	0	0
	Test Err.	17.83	2.25	2.5	2.83	2.08	2.17	1.25	1.75

Table 4.11: Classification results (white noise; SNR=0dB; KNK)

		d=1	d=2	d=3	d=4	d=5	d=6	d=7	d=8
WPT	Tr. Err.	0	0.25	0.25	0	0	0	0	0
	Test Err.	0.25	0.5	1.08	0.08	0.17	0.17	0.25	0.42
FT	Tr. Err.	1.75	1	0.25	1	0.5	0	0	0.25
	Test Err.	6.33	6.08	3.42	2.08	4	2.25	2.17	1.5

Table 4.12: Classification results (white noise; SNR=-3dB; PWM)

		d=1	d=2	d=3	d=4	d=5	d=6	d=7	d=8
WPT	Tr. Err.	0.25	0	0.25	0.25	0	0.25	0	0
	Test Err.	0.67	0.08	0.92	0.83	0	0.17	0.08	0.08
FT	Tr. Err.	5	1	1.25	1.25	0.75	1	0.25	0
	Test Err.	7.25	5.42	5.92	6.92	4.83	4.83	6.67	5

Table 4.13: Classification results (white noise; SNR=-3dB; KNK)

		d=1	d=2	d=3	d=4	d=5	d=6	d=7	d=8
WPT	Tr. Err.	0.5	0	0	0	0.25	0	0.25	0.25
	Test Err.	1	0.17	0.17	0.33	0.75	0.08	1.17	1.33
FT	Tr. Err.	16.25	3.25	1.5	1.75	1	0.75	0.25	0
	Test Err.	19	10.58	5	6	4.5	5.08	4.83	4.17

4.8 Test on Data Corrupted by Additive Color Noise

The second type of noise used to corrupt original data is colored noise where a group of frequencies is dominant, as shown in Figure 4-3. Such a noise can be generally represented by an ARMA(n, n-1) model [26]:

$$\begin{aligned} X_k &= \phi_1 X_{k-1} + \phi_2 X_{k-2} + \ldots + \phi_n X_{k-n} \\ &\quad + a_k - \theta_1 a_{k-1} - \theta_2 a_{k-2} - \ldots - \theta_{n-1} a_{k-n+1} \end{aligned} \tag{4.2}$$

Coefficients ϕ_k and θ_k determine the center frequency and bandwidth of the noise. The a_k are normally distributed with zero mean and variance σ_a^2. In our tests, however, we generated such a noise by convolving a white noise sequence with a bandpass filter. We generated the colored noise such that the dominant frequencies lie between the digital frequency band 0 to 0.25π. This is the band where the original signal contains most of its energy, as can be seen from Figure 4-1. Table 4.14 through Table 4.17 show the classification results of conducted simulations corresponding to different SNR. In all simulations, better results are obtained via the wavelet packet based approach.

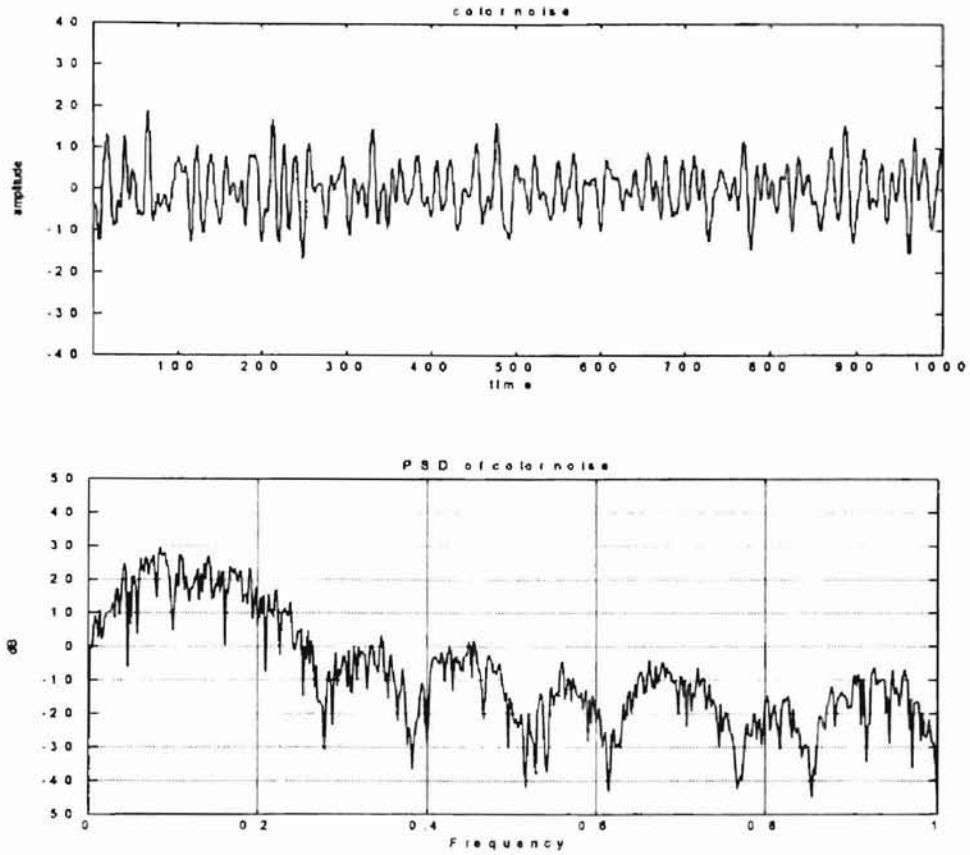

Figure 4-3: Color noise and its power spectrum. The frequency axis is in units of $\pi \times$ radians.

Table 4.14: Classification results (color noise, SNR=0dB; PWM)

		d=1	d=2	d=3	d=4	d=5	d=6	d=7	d=8
WPT	Tr. Err.	12.5	0	0.25	0	0	0	0	0
	Test Err.	13.08	0.5	0.25	0.08	0.5	0.25	0.67	1.08
FT	Tr. Err.	15.75	2.5	0.75	0.75	0.25	0.25	0	0.25
	Test Err.	18	5.92	9.5	6.83	4.75	5.5	4.33	3.75

Table 4.15: Classification results (color noise; SNR=0dB; KNK)

		d=1	d=2	d=3	d=4	d=5	d=6	d=7	d=8
WPT	Tr. Err.	0.5	0.75	0	0	0	0	0	0
	Test Err.	3	2.42	1.17	0.56	0.42	2.25	0.75	1
FT	Tr. Err.	2.25	1.5	0.75	0.75	0	0.75	0	0
	Test Err.	7.08	7.08	2.42	3.75	3.25	2.17	2.83	4.17

Table 4.16: Classification results (color noise; SNR=-3dB; PWM)

		d=1	d=2	d=3	d=4	d=5	d=6	d=7	d=8
WPT	Tr. Err.	12.5	0	0	0	0	0	0	0
	Test Err.	13.08	0.08	0.08	0.42	0.25	0.25	0.17	0.25
FT	Tr. Err.	27	4.5	2	1.25	1.25	0.5	0.25	0.5
	Test Err.	30.25	10.83	16.08	14.42	13.08	12.08	11.92	13.17

Table 4.17: Classification results (color noise; SNR=-3dB; KNK)

		d=1	d=2	d=3	d=4	d=5	d=6	d=7	d=8
WPT	Tr. Err.	0.25	0.25	0	0	0	0	0.25	0
	Test Err.	0.75	0.83	0.58	0.67	1.58	0.33	1.67	2.33
FT	Tr. Err.	7.25	3.25	2.25	1	0.5	1.75	0.25	1
	Test Err.	10	9.83	9.33	11.17	10	13.33	10.92	11.83

4.9 Test Result on Data Corrupted by Pink Noise

The third type of noise employed is pink noise, where power decreases as frequency increases, as depicted in Figure 4-4. It can be expressed by an AR(1) model:

$$X_k = \phi_1 X_{k-1} + a_k \tag{4.3}$$

where a_k are normally distributed with zero mean and variance σ_a^2. In the test, ϕ_1 is set to be 0.95, and resulting noise is displayed in Figure 4-4. The test results are shown in Table 4.18 through Table 4.21. Again it is confirmed that the wavelet packet based approach produces better results.

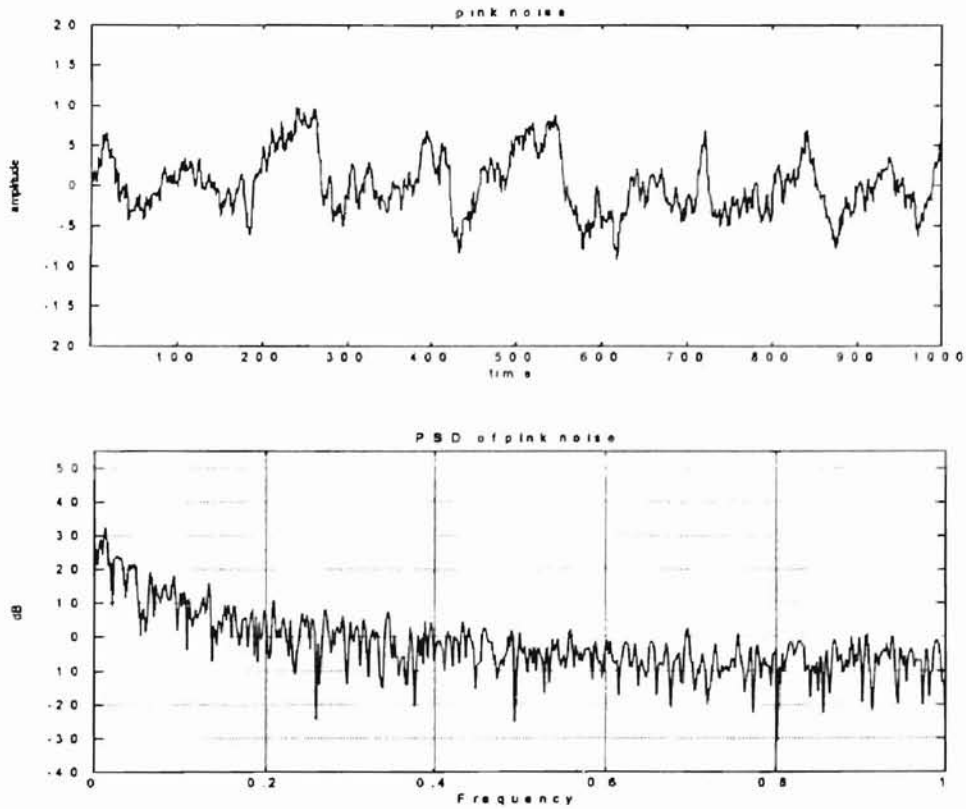

Figure 4-4: Pink noise and its power spectrum. The frequency axis is in units of $\pi \times$ radians.

Table 4.18: Classification results (pink noise; SNR=0dB; PWM)

		d=1	d=2	d=3	d=4	d=5	d=6	d=7	d=8
WPT	Tr. Err.	0.5	0	0	0	0	0	0	0
	Test Err.	0.5	0.08	0.67	0	1.33	0.08	0.33	0.08
FT	Tr. Err.	0.5	1.25	0	0.25	0	0	0	0
	Test Err.	1.33	2.58	2.25	1.5	2.5	0.92	0.42	0.5

Table 4.19: Classification results (pink noise; SNR=0dB; KNK)

		d=1	d=2	d=3	d=4	d=5	d=6	d=7	d=8
WPT	Tr. Err.	0	0.25	0.25	0	0.25	0	0	0
	Test Err.	0.83	0.92	0.17	0.42	0.25	0.58	0	0.17
FT	Tr. Err.	1.75	0.5	0.5	0.25	0	0.25	0.25	0
	Test Err.	3.58	4.58	1.92	2	1.91	0.67	1.42	0.83

Table 4.20: Classification results (pink noise; SNR=-3dB; PWM)

		d=1	d=2	d=3	d=4	d=5	d=6	d=7	d=8
WPT	Tr. Err.	0.5	0.25	0	0	0	0.25	0.25	0.25
	Test Err.	1.83	0.33	0.33	0.08	0.5	0.42	1.42	0.42
FT	Tr. Err.	2.5	1	0.75	0	0.5	0	0	0
	Test Err.	5.67	3.25	4.08	3.17	4.17	1.33	2.5	1.17

Table 4.21: Classification results (pink noise; SNR=-3dB; KNK)

		d=1	d=2	d=3	d=4	d=5	d=6	d=7	d=8
WPT	Tr. Err.	0.75	0	0	0	0	0.25	0.25	0.25
	Test Err.	1.25	0.17	1.25	0.5	0.17	0.75	0	0.17
FT	Tr. Err.	1.25	1.25	0.5	0.75	0.5	0	0	0
	Test Err.	7	5.25	3.5	4.17	3.08	3.42	3.17	2.83

4.10 Discussion on Test Results

In the sequel, we summarize the findings based on the conducted simulation results on the Westland data set.

1. By examining Table 4.3 and Table 4.4 where only one sensor is used for the searching of class specific feature components, it is clear that some sensors provide little class separability information in the sense of frequency analysis. This indeed confirmed our understanding that the faulted symptom is localized and can only be detected by neighboring sensors. It suggests that data collected from multiple sensors will provide better classification information and lead to better performance.
2. From the results of simulation on the *original data set*, it is observed that no improvement is made through the wavelet packet based approach on this data set, and in several cases it is even slightly worse than the Fourier based approach. As mentioned before, this could be due to the overlap of frequency content among wavelet packet basis functions.
3. Nevertheless, the wavelet packet based approach shows very promising results in a realistic environment for which the data are corrupted by noise.

CHAPTER V

CONCLUSION

5.1 Summary

This thesis has investigated the feasibility of applying the wavelet packet transform to the classification of vibration signals. Using the wavelet packet transform, a rich collection of time-frequency characteristics in a signal could be obtained and examined for classification purposes. In this study, we detailed our systematic feature selection process that exploits signal class differences in the wavelet packet node energy. This results in a reduced dimensional feature space compared to dimension of the original time domain signal. The wavelet packet based features, obtained by our method for vibration signals, yield nearly 100% correct classification when used as input to a neural network classifier.

In Chapter 2, we reviewed the Fourier based analysis on the extraction of frequency information from a signal and discussed the possible inherent drawbacks due to its fixed time-frequency resolution. The wavelet packet transform that overcomes the fixed time-frequency resolution was then presented. To alleviate the *time variant* characteristics of the wavelet packet transform coefficients, wavelet packet node energy was used as an essential time-frequency feature measure of the signal. Although the wavelet packet node energy provided us a multiresolution view of a signal, it

simultaneously introduced a higher dimension space compared to original time domain signal. To reduce the dimensionality, it was shown that LDA had some practical problems when the feature dimension was relatively higher compared to the number of collected samples. It involved calculation of the inverse of the covariance matrix. In such a case, two feature selection methodologies based on measures of the overlap of the conditional probability density function among different classes was proposed to avoid the possible numerical problems as presented in Chapter 3. In Chapter 4, the proposed wavelet packet based classification system, which combined a wavelet packet based feature extractor and a neural network classifier, was tested on a real data set known as the Westland data set. Numerically, it was observed that significant improvement can be achieved when using multiple sensor data. This validated our understanding that a faulted symptom is localized and can only be detected by the neighboring sensors. Both the Fourier based features and wavelet packet based features achieved excellent classification results on the original Westland data set when all eight sensor were utilized. Nonetheless, the improved time-frequency resolution of the wavelet packet transform are observed when we are confronted with signals corrupted by artificially synthesized noises. In the extended tests, the wavelet packet based approach showed very promising results compared to the Fourier based approach.

5.2 Suggestion for Future Work

Whereas satisfactory results are obtained from the study based on the Westland data set, there are some extensions of this research that are recommended for future studies.

1. Investigation of a more sophisticated feature selection criterion: In this study, a simplified criterion, Fisher's criterion, is used to measure the overlap of the conditional probability density function of a specific feature among different classes. The criterion is based upon the inherent assumption that the probability density distribution of a feature is Gaussian. In practice, the accuracy of the criterion may be degraded for non-Gaussian distribution.
2. One inherent problem with the two proposed feature selection methodologies, detailed in Chapter 3, is the lack of a criterion for determines d automatically. How to select the *best* d is a tough question and needs further research.

REFERENCES

1. G.G.Yen, "Health Monitoring of Vibration Signatures in Rotorcraft Wings," *Neural Processing Letters*, vol. 4, pp. 127-137, 1996.

2. R.J.Ferlez and D.C.Lang, "Gear-Tooth Fault Detection and Tracking Using the Wavelet Transform," *Prognosis of Residual Life Machinery and Structures, Proceedings of the 52nd Meeting of the Society for Machinery Failure Prevention Technology*, pp. 451-460, March, 1998.

3. Juei-Cheng Lo, et al, "Fault Prediction in Transmissions Using Wavelet Analysis," *Prognosis of Residual Life Machinery and Structures, Proceedings of the 52nd Meeting of the Society for Machinery Failure Prevention Technology*, pp. 441-450, March, 1998.

4. P.D.Samuel, et. al., "Fault Detection in the OH-58A Main Transmission Using the Wavelet Transform," *Prognosis of Residual Life Machinery and Structures, Proceedings of the 52nd Meeting of the Society for Machinery Failure Prevention Technology*, pp. 323-335, March, 1998.

5. B.Liu, et al, "Machinery Diagnosis Based on Wavelet Packets," Journal of Vibration and Control, vol. 3, pp. 5-17, 1997.

6. J.E.Lopez and K.Oliver, "Overview of Wavelet/Neural Network Fault Diagnostic Methods Applied to Rotating Machinery," *Technology Showcase Integrated Monitoring, Diagnostics and Failure Prevention, Proceeding of a Joint Conference*, pp.405-417, April 1996.

7. B.E.Parker, et al, "Helicopter Transmission Diagnostics using Vibration Signature Analysis," *Technology Showcase Integrated Monitoring, Diagnostics and Failure Prevention, Proceeding of a Joint Conference*, pp.419-430, April 1996.

8. M.A.Essawy, S.Diwakar and S.Zein-Sabatto, "Fault Diagnosis of Helicopter Gearboxes Using Neuro-Fuzzy Techniques," *Prognosis of Residual Life Machinery and Structures, Proceedings of the 52nd Meeting of the Society for Machinery Failure Prevention Technology*, pp. 293-303, March, 1998.

9. D.Paul, "Condition monitoring and defect diagnosis in manufacturing process using DDS and wavelets," Ph.d dissertation, Michigan Technological University, 1995.

10. O.Rioul and M.Vetterli, "Wavelets and Signal Processing," *IEEE Signal Processing Magnine*, pp. 14-38, Oct. 1991.

11. F.Hlawatsch and G.F.Boudreaux-Bartels, "Linear and Quadratic Time-Frequency Signal Representations," *IEEE Signal Processing Magnine*, pp. 21-67, Apr. 1992.

12. R.R.Coifman and M.V.Wickerhauser, "Entropy-Based Algorithms for Best Basis Selection," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 713-718, March 1992.

13. A.Papoulis, *The Fourier Integral and its Applications*, New York: McGraw-Hill, 1962.

14. I.Daubechies, "The Wavelet Transform, Time-Frequency Localization and Signal Analysis," *IEEE Transactions on Information Theory*, vol. 36, no. 5, pp. 961-1005, Sep 1990.

15. I.Daubechies, "Orthonormal Bases of Compactly Supported Wavelets," *Communications on Pure and Applied Mathematics*, vol. XLI, pp.909-996, 1988.

16. S.G.Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674-693, July, 1989.

17. A.N.Akansu and R.A.Haddad, *Multiresolution Signal Decomposition: Transforms, Subbands, Wavelets*, Academic Press, Inc. 1992.

18. M.V.Wickerhauser, *Adapted wavelet analysis from theory to software*, MA : Wellesley, 1994.

19. P.A.Devijver and J.Kittler, *Pattern Recognition - A Statistical Approach*, Prentice Hall, London, 1982.

20. K.Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, Inc. 1992.

21. J.Kittler, "Mathematical Methods of Feature Selection in Pattern Recognition," *International Journal on Man-Machine Studies*, vol. 7, pp.609-637, 1975.

22. S.Watanabe and T.Kaminuma, "Recent Developments of the Minimum Entropy Algorithms," *Proceedings of International Conference on Pattern Recognition, IEEE, 1998*, pp.536-540.

23. R.P.Lippmann, "Pattern classification using neural networks," *IEEE Communication Magazine*, pp. 47-64, Nov. 1989.

24. B.G.Cameron, "Final Report on CH-46 Aft Transmission Seeded Fault Testing," Westland Research Paper RP907, Westland Helicopter. 1 September 1993. Now accessible via WWW at http://wisdom.arl.psu.edu/Westland.

25. A.V.Oppenheim and R.W.Schafer, *Discrete-Time Signal Processing*, Prentice-Hall, Inc. 1989.

26. S.M.Pandit and S.M.Wu , Time Series and Systems Analysis with Applications, John Wiley, 1983. Reprinted by Krieger, 1993.

VITA

Kuo-Chung Lin

Candidate for the Degree of

Master of Science

Thesis: WAVELET PACKET FEATURE EXTRACTION FOR VIBRATION MONITORING

Major Field: Electrical Engineering

Biographical:

Education: Received Bachelor of Science degree in Electrical Engineering from Tamkang University, Taipei, Taiwan, R. O. C. in May 1990; completed the requirements for the Master of Science degree with a major in Electrical Engineering at Oklahoma State University in December 1998.

Experience: Employed by the Universal Scientific Industrial Co. Ltd. in Nantou, Taiwan as an assistant engineer, 1992-93; Employed by the Taiwan Provincial Health Department in Nantou, Taiwan as a computer technical consultant, 1993-94; Employed by the Nantou District Court in Nantou, Taiwan as computer technical consultant, 1994-96; Employed by the Electrical and Computer Engineering Department at Oklahoma State University in Stillwater, Oklahoma as a research assistant, 1997-98.