

Wavelets

Robert X. Gao • Ruqiang Yan

Wavelets

Theory and Applications for Manufacturing

Robert X. Gao
Department of Mechanical Engineering
University of Connecticut
Storrs, CT 06269 3139, USA
rgao@engr.uconn.edu

Ruqiang Yan
School of Instrument Science and
Engineering
Southeast University
Nanjing, Jiangsu
China, People's Republic
ruqiang@seu.edu.cn

ISBN 978 1 4419 1544 3 e ISBN 978 1 4419 1545 0
DOI 10.1007/978 1 4419 1545 0
Springer New York Dordrecht Heidelberg London

© Springer Science+Business Media, LLC 2011

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Since the publication of Alfred Haar's work on orthogonal function systems a hundred years ago, the world has witnessed a tremendous growth in the theory and practice of *wavelet*, even though a reported, systematic study of the subject field and its applications to engineering did not occur until the 1980s. Over the last 3 decades, a plethora of literature has been published, describing advancement in the wavelet theory and its successful applications in various fields of engineering: from image processing in biomedical engineering to signal processing in meteorology to bridge monitoring in civil engineering. The adaptive, multiresolution capability of the wavelet transform has also made it a powerful mathematical tool for the diagnostics of equipment operation conditions in manufacturing, such as tool breakage.

Past research on wavelets has been translated into a large volume of publications and significantly impacted the state-of-the-technology. These papers, together with a series of classic books, have taught generations of engineers the theory and applications of wavelets. Nevertheless, there exists a gap in the literature that is particularly dedicated to graduate students and practicing engineers in manufacturing who are interested in learning about and applying the theory of wavelet transform to solving problems related to equipment and process monitoring, diagnostics, and prognostics in manufacturing.

The book is intended to bridge such a gap by presenting a systematic yet easily accessible treatment of the mathematics of wavelet transform and illustrate, in concrete terms, how wavelet transform as a mathematical tool can be realized for applications in manufacturing. Contributing to the understating and adoption of wavelets by the manufacturing community is the primary motivation for this book, and the 12 chapters included herein provide an overview of some of the latest efforts in this vibrant field.

To establish a common ground for the treatment of signals, which is the focal point of this book, Chap. 1 starts by introducing a general classification scheme of signals typically countered in mechanical systems, from the point of view of their statistical behavior—deterministic and nondeterministic signals. Using a mass spring damper system as a physical embodiment, the analytical expression, waveform, and solution of deterministic signals are first illustrated. These are then

contrasted against the nondeterministic family of signals, and the concept of nonstationary, which provides the fundamental motivation for dedicating this book to wavelet transform in manufacturing, is introduced. Taking signals measured in two representative manufacturing processes as a realistic example, the link between manufacturing and signal processing, as well as the need for properly treating nonstationary signals, are established, motivating the dedication of the book to this subject matter.

Chapter 2 reviews several major events occurred in the field of signal processing since the invention of the Fourier transform in the nineteenth century, thereby recognizing the historical significance of spectral analysis. Such events have initiated and accompanied the conceptualization, formulation, and growth of the theory of wavelet transform. Based on the concept that signal transformation (for revealing the information content of the signal) can generally be represented by a convolution operation between the signal and a known *template* function, we sought to illustrate the common ground shared by the Fourier transform as well as its enhanced version (the short-time Fourier transform), which has a fixed length of the analysis window, and the wavelet transform, which features an analysis window of variable length.

The next three chapters, Chaps. 3–5, are devoted to introducing the fundamental mathematics involved in understanding what wavelet transform is and does, and how to apply it to decompose nonstationary signals as typically encountered in manufacturing. Aware of the existence of many excellent books on wavelets and at the same time, the recognized need by many graduate students and practicing engineers for a step-by-step treatment of some of the mathematical procedures involved to implement the wavelet transform, in terminologies familiar to engineers, we tried to take a balanced approach when writing these chapters. Specifically, we introduced the continuous version of the wavelet transform in Chap. 3, by first drawing the resemblance between a continuous, sinusoidal wave and a time-localized wavelet that is essentially a linear, integral transformation satisfying the *admissibility* condition. To provide the readers with a handy access to some of the most often encountered properties of the continuous wavelet transform (CWT) in one place, we included descriptions of concepts such as superposition, covariance under translation and dilation, and the Mayol principle, together with a mathematical *proof*, for each of these properties. By providing detailed proofs, we wish to encourage readers who might have initially felt intimidated by the wavelet mathematics to gain some confidence in approaching the topic from a practical yet mathematically rigorous perspective, instead of resorting to a strictly recipe type of operations. We then proceeded to give a step-for-step procedure for implementing the CWT, in two ways, such that readers can see, in concrete terms, where all the background information finally leads to, in terms of performing CWT on some representative signals.

Chapter 4 introduces the discrete version of the wavelet transform, or DWT. The chapter is motivated by the recognition that CWT, while enabling a 2D decomposition of signals in the time–frequency (via the scale) domain with high resolution, is computationally complex due to the generation of redundant data. In comparison, the DWT is computationally more efficient, thus it is better suited for image

compression and real-time applications. Using logarithmic discretization as an example, we first discussed how parameters are discretized to guarantee correct information retrieval as a result of the DWT process. Several derivation details are provided to illustrate the thought process. We then moved to the *dyadic* discretization method that allows for *orthogonal* wavelet basis to be constructed, based on the theory of multiresolution analysis (MRA). To satisfy readers who may be interested in knowing a bit more about the “why” and “how” related to MRA, we supplemented the explanation with several mathematical details, and illustrated why the process of DWT will lead to the generation of *detailed* and *approximate* information. In this context, we demonstrated that DWT, in essence, is about performing a series of low-pass and high-pass filtering operations, which can be implemented by following Mallat’s algorithm. Mirroring the structure of Chap. 3, we presented several commonly used wavelets for DWT and illustrated how they can be used for applications such as signal denoising, by means of soft and hard thresholding.

While enabling flexible time frequency resolution in signal decomposition, the relatively low resolution of DWT in analyzing the high-frequency region gives rise to the wavelet packet transform (WPT), which is the focus of Chap. 5. After a brief coverage of its definition and basic properties, two algorithms for implementing the WPT—the recursive algorithm developed by Mallat and a Fourier transform-based algorithm that leads to the harmonic wavelet packet transform—are introduced. We then illustrated how a signal’s time frequency composition of a vibration signal, which relates directly to the working state of manufacturing equipment, can be revealed by the WPT, and how WPT can be applied to removing Gaussian noise from a chirp signal. These applications exemplify how the enhanced resolution of WPT can provide an attractive tool for detecting and differentiating transient elements with high-frequency characteristics.

With the fundamentals of wavelet transform covered, Chaps. 6–8 describe several application scenarios where the effectiveness of wavelet transforms are demonstrated. The first application relates to signal *enveloping*, a technique commonly used for nondestructive testing and structural defect identification. Addressing the limitation of enveloping in requiring a priori knowledge for choosing the filtering band to extract a signal’s envelope, an adaptive, multiscale enveloping method (MuSEnS) that is rooted in the wavelet transform is described in Chap. 6, which effectively overcomes the limitation. Taking advantage of the Hilbert transform in extracting the envelop of an *analytic* signal and the fact that performing wavelet transform on a signal using a complex-valued base wavelet will result in an *analytic* signal, the chapter illustrates how a signal’s envelope can be readily calculated from the modulus of the corresponding wavelet coefficients. To illustrate the effectiveness of this technique in signal decomposition, two manufacturing-related applications—differentiation of ultrasonic pulses that are timely overlapped and spectrally adjacent for wireless pressure measurement in injection molding and bearing defect diagnosis in rotary machines—are demonstrated, using both experimentally measured signals and synthetic signals for quantitative evaluation.

While the localized signal decomposition capability of wavelet transform is particularly useful for transient events identification, the result of wavelet transform does not explicitly reveal distinct characteristic frequencies that are often times

indicative of defective modes of a machine, e.g., the ball passing frequency at the inner raceway of a rolling element, when a localized spalling is present. In such situations, the effectiveness of wavelet transform can be leveraged by the Fourier transform in identifying a signal's frequency components. This leads to the formulation of a *unified* time-scale-frequency analysis technique that adds spectral post-processing to the data set extracted by wavelet transform, for enhanced defect diagnosis. In Chap. 7, we demonstrate such an integrated method, in the context of a *generalized* signal transformation frame. An expression for both the Fourier transform and wavelet transform in the generalized frame is first presented, establishing the basis for crossdomain unification of the two transforms. Next, the viability of postspectral analysis of wavelet processed data is analytically justified, and the effectiveness of the technique in identifying the bearing defects under various operating conditions is demonstrated.

A question that naturally arises upon defect detection is the severity of the defect, which affects the proper scheduling of maintenance. To answer this question, we demonstrate in Chap. 8 how WPT can be applied in classifying machine defect severity, using vibration signals from rolling bearings as an example. We start the discussion by associating *features* (e.g., energy content or Kurtosis value) of a signal with the subfrequency bands of its decomposition, enabled by the WPT, and demonstrate how WPT can flexibly extract features from the subfrequency bands of the decomposed signal where the features are concentrated. The chapter further contains a discussion on how to process the features, once they are obtained, for classification purpose. Relevant techniques for selecting best-suited features using the Fisher linear discriminant analysis and principal component analysis, and classifying features to quantify defect severity levels are described. Two case studies presented toward the end of the chapter on ball and roller bearings confirm the validity of WPT for defect severity classification.

Chapter 9 continues the discussion on signal *classification*, with a focus on how it can be applied to differentiate different working conditions of a machine, for the purpose of diagnosis. The concept of *discriminant* features is first introduced, and a technique called the local discriminant bases (LDB) is described in detail. In a nutshell, the LDB algorithm determines an optimal set of wavelet packet nodes, each of which corresponding to a wavelet packet basis, to represent signals acquired under different machine states as different *classes*. Similar to the Shannon entropy feature introduced in Chap. 5 for signal compression, several features suited for diagnosis of rotating machines, e.g., relative entropy or correlation index, are identified in this chapter. We provided a step-by-step description of the LDB algorithm, for readers to see how the algorithm can be implemented. Using three synthetic signals with added white noise and vibration signals measured on a gearbox under different states of wear, we quantitatively demonstrated how the wavelet packet bases constructed using the LDB algorithm can more successfully differentiate and classify these signals than without using the LDB.

Given the abundance of the base wavelets in the published literature, it is natural to ask the question as to how to choose an optimal base wavelet for analyzing a particular type of signal. This is based on the understanding that (1) the choice of base wavelet made in the first place will affect the result obtained at the end, and

(2) each base wavelet may be developed for different purposes and emphasis; therefore, an educated approach to their selection is needed when solving a specific type of engineering problems. In this book, we have tried to address this issue of significant intellectual interest in two ways. First, in Chap. 10, we introduced a general strategy for base wavelet selection, using both *qualitative* measures (e.g., orthogonality and compact support) and *quantitative* measures (e.g., Shannon entropy and discrimination power). Subsequently, we presented several criteria for base wavelet selection, including the *energy-to-Shannon entropy ratio* and the *maximum information measure*. Using both real-valued and complex-valued base wavelets, we demonstrated how these criteria can be applied to selecting the best-suited base wavelet from a pool of candidates to decompose both a numerically formulated Gaussian-modulated sinusoidal test signal and a vibration signal measured on a defective ball bearing, thus confirming the effectiveness of these criteria.

Besides investigating how to choose an appropriate base wavelet from the existing library, another approach is to design a *customized* wavelet that is adapted to a specific type of application to maximize the degree of matching with the signal of interest, thus improving the effectiveness of feature extraction. Such a complementary technique is the focus of Chap. 11. After reviewing the fundamental issues involved in the wavelet design process and several customized wavelets, we described in detail the process of designing an *impulse* wavelet for bearing vibration analysis, based on the impulse response of the mechanical structure where the bearing is housed. The importance of satisfying the *dilation* equation to avoid information loss in the signal reconstruction is stressed, and the procedure of meeting this requirement is illustrated. Using the designed impulse wavelet, vibration signals from a defective bearing are analyzed, and the result is compared with that from using five standard wavelets available in the library, using the signal-to-noise ratio for the defect-characteristic frequency as the measure for comparison. The good performance of the impulse wavelet confirms the validity of the analytical procedure described in developing customized wavelets for enhanced signal analysis in a broad range of applications in engineering.

The last chapter of the book provides a brief survey of some new advancement reported in recent years that goes beyond the classical wavelet transform. These latest developments address some of the fundamental limitations inherent to the wavelet transform, e.g., when it is used to analyze signals of finite length and/or limited duration, or for capturing and defining image boundaries. We started the survey by introducing the second generation wavelet transform, or SGWT, which uses the so-called *lifting scheme* to replace the traditional mechanism of wavelet construction that uses translation and dilation. Major operation steps for realizing the SGWT, such as splitting, prediction, and updating, are described, and the effectiveness of the technique for separation and reconstruction of an intermittent linear chirp signal is demonstrated. Addressing the inherent limitations of classical wavelets (e.g., *isotropic*) and the specific challenges in image processing (e.g., in resolving image boundaries), we then introduced the ridgelet and curvelet transforms. The former was developed to address the need for analyzing anisotropic features in images, whereas the latter enables improved representation of curved

boundaries in images. For each transform, we have presented the definition and basic properties, and demonstrated a representative application in manufacturing.

As is true with any book, the writing reflects upon the authors' understanding of and knowledge about the subject matter. While we have strived to present to the readers a composition that is both rigorous in the mathematical treatment and relevant in the examples chosen to complement the theory, it is inherently difficult for a work of this size to be completely free of errors. We bear the responsibility for anything that is not correctly stated in the book and would greatly appreciate hearing from our readers about any mistakes they found such that we can correct them in future printings.

We thank the anonymous reviewers for their insightful and constructive comments that have both sharpened our thinking and provided clues to improving the pedagogic presentation. We are indebted to former graduate students at the Electromechanical Systems (EMS) Laboratory, whose intellectual contributions have made the book a reality. In particular, we thank Drs. Brian Holm-Hansen, Changting Wang, and Li Zhang, who dedicated a substantial part of their doctoral research to the study of wavelet for diagnosis in manufacturing equipment and processes, and Dr. Qingbo He, who spent a year as a postdoctoral research fellow at the EMS Laboratory, working on the characterization of physical activities, for their dedication to and enthusiasm in exploring the world of wavelets, which has laid the foundation for this book. We also thank the US National Science Foundation for funding a number of relevant projects, which has allowed us to systematically study this fascinating subject.

This book-writing project was initially planned to be completed in 1 year, but a number of events that took place in between have delayed the writing and made the time needed for ultimately completing the book nearly twice as long. We take this opportunity to express our sincere appreciation to the publisher for supporting this project. In particular, we thank Mr. Stephen Elliot, Senior Editor for Engineering, and Mr. Andrew Leigh, editorial assistant, for their earnest cooperation, editorial assistance, and above all, patience, which has created a relaxed environment for us to finish the writing while juggling many other deadline issues. Last but not the least, we sincerely thank our respective families for their understanding and carrying the load for us during the course of this project so that we could devote as much time as possible to the writing of the book. It is our hope that the book is worth their selfless support, and our readers will find in it something that is of value to their research.

Contents

1	Signals and Signal Processing in Manufacturing	1
1.1	Classification of Signals	1
1.1.1	Deterministic Signal	1
1.1.2	Nondeterministic Signal	3
1.2	Signals in Manufacturing	5
1.3	Role of Signal Processing for Manufacturing	11
1.4	References	13
2	From Fourier Transform to Wavelet Transform:	
	A Historical Perspective	17
2.1	Fourier Transform	18
2.2	Short-Time Fourier Transform	21
2.3	Wavelet Transform	26
2.4	References	31
3	Continuous Wavelet Transform	33
3.1	Properties of Continuous Wavelet Transform	35
3.1.1	Superposition Property	35
3.1.2	Covariant Under Translation	36
3.1.3	Covariant Under Dilation	36
3.1.4	Moyal Principle	37
3.2	Inverse Continuous Wavelet Transform	38
3.3	Implementation of Continuous Wavelet Transform	39
3.4	Some Commonly Used Wavelets	41
3.4.1	Mexican Hat Wavelets	41
3.4.2	Morlet Wavelet	41
3.4.3	Gaussian Wavelet	42
3.4.4	Frequency B-Spline Wavelet	43
3.4.5	Shannon Wavelet	43
3.4.6	Harmonic Wavelet	44

3.5	CWT of Representative Signals	45
3.5.1	CWT of Sinusoidal Function	45
3.5.2	CWT of Gaussian Pulse Function	46
3.5.3	CWT of Chirp Function.....	46
3.6	Summary	47
3.7	References.....	47
4	Discrete Wavelet Transform.....	49
4.1	Discretization of Scale and Translation Parameters.....	49
4.2	Multiresolution Analysis and Orthogonal Wavelet Transform.....	53
4.2.1	Multiresolution Analysis.....	53
4.2.2	Orthogonal Wavelet Transform.....	55
4.3	Dual-Scale Equation and Multiresolution Filters.....	56
4.4	The Mallat Algorithm.....	58
4.5	Commonly Used Base Wavelets.....	60
4.5.1	Haar Wavelet.....	61
4.5.2	Daubechies Wavelet	61
4.5.3	Coiflet Wavelet.....	62
4.5.4	Symlet Wavelet	63
4.5.5	Biorthogonal and Reverse Biorthogonal Wavelets.....	63
4.5.6	Meyer Wavelet.....	65
4.6	Application of Discrete Wavelet Transform.....	65
4.7	Summary	68
4.8	References.....	68
5	Wavelet Packet Transform	69
5.1	Theoretical Basis of Wavelet Packet	69
5.1.1	Definition.....	69
5.1.2	Wavelet Packet Property.....	72
5.2	Recursive Algorithm.....	73
5.3	FFT-Based Harmonic Wavelet Packet Transform	74
5.3.1	Harmonic Wavelet Transform	74
5.3.2	Harmonic Wavelet Packet Algorithm	75
5.4	Application of Wavelet Packet Transform	78
5.4.1	Time-Frequency Analysis.....	78
5.4.2	Wavelet Packet for Denoising	79
5.5	Summary	79
5.6	References.....	80
6	Wavelet-Based Multiscale Enveloping	83
6.1	Signal Enveloping Through Hilbert Transform	83
6.2	Multiscale Enveloping Using Complex-Valued Wavelet	86
6.3	Application of Multiscale Enveloping.....	87
6.3.1	Ultrasonic Pulse Differentiation for Pressure Measurement in Injection Molding.....	87
6.3.2	Bearing Defect Diagnosis in Rotary Machine.....	93

6.4	Summary	99
6.5	References.....	100
7	Wavelet Integrated with Fourier Transform:	
	A Unified Technique	103
7.1	Generalized Signal Transformation Frame	103
7.1.1	Fourier Transform in the Generalized Frame	106
7.1.2	Wavelet Transform in the Generalized Frame	107
7.2	Wavelet Transform with Spectral Postprocessing.....	109
7.2.1	Fourier Transform of the Measure Function	110
7.2.2	Fourier Transform of Wavelet-Extracted Data Set.....	112
7.3	Application to Bearing Defect Diagnosis.....	113
7.3.1	Effectiveness in Defect Feature Extraction.....	115
7.3.2	Selection of Decomposition Level.....	118
7.3.3	Effect of Bearing Operation Conditions.....	120
7.4	Summary	124
7.5	References.....	124
8	Wavelet Packet-Transform for Defect Severity Classification.....	125
8.1	Subband Feature Extraction	125
8.1.1	Energy Feature	126
8.1.2	Kurtosis	127
8.2	Key Feature Selection.....	128
8.2.1	Fisher Linear Discriminant Analysis	129
8.2.2	Principal Component Analysis.....	131
8.3	Neural-Network Classifier	134
8.4	Formulation of WPT-Based Defect Severity Classification.....	136
8.5	Case Studies.....	137
8.5.1	Case Study I: Roller Bearing Defect Severity Evaluation	137
8.5.2	Case Study II: Ball Bearing Defect Severity Evaluation	142
8.6	Summary	146
8.7	References.....	146
9	Local Discriminant Bases for Signal Classification	149
9.1	Dissimilarity Measures.....	149
9.1.1	Relative Entropy	150
9.1.2	Energy Difference.....	151
9.1.3	Correlation Index	151
9.1.4	Nonstationarity	152
9.2	Local Discriminant Bases.....	153
9.3	Case Study	155
9.4	Application to Gearbox Defect Classification.....	158
9.5	Summary	162
9.6	References.....	162

10	Selection of Base Wavelet	165
10.1	Overview of Base Wavelet Selection	165
10.1.1	Qualitative Measure	166
10.1.2	Quantitative Measure	168
10.2	Wavelet Selection Criteria.....	169
10.2.1	Energy and Shannon Entropy	170
10.2.2	Information Theoretic Measure	172
10.3	Numerical Study on Base Wavelet Selection	176
10.3.1	Evaluation Using Real-Valued Wavelets	176
10.3.2	Evaluation Using Complex-Valued Wavelets.....	179
10.4	Base Wavelet Selection for Bearing Vibration Signal.....	183
10.5	Summary.....	185
10.6	References	186
11	Designing Your Own Wavelet	189
11.1	Overview of Wavelet Design.....	189
11.2	Construction of an Impulse Wavelet	190
11.3	Impulse Wavelet Application	198
11.4	Summary.....	202
11.5	References	203
12	Beyond Wavelets	205
12.1	Second Generation Wavelet Transform	205
12.1.1	Theoretical Basis of SGWT	206
12.1.2	Illustration of SGWT in Signal Processing	208
12.2	Ridgelet Transform	210
12.2.1	Theoretical Basis of Ridgelet Transform.....	210
12.2.2	Application of the Ridgelet Transform.....	212
12.3	Curvelet Transform.....	214
12.3.1	Curvelet Transform.....	214
12.3.2	Application of the Curvelet Transform.....	217
12.4	Summary.....	218
12.5	References	219
	Index	221

Chapter 1

Signals and Signal Processing in Manufacturing

The term “signal” refers to a physical quantity that carries certain type of information and serves as a means for communication. As an example, the output of an accelerometer in the form of a voltage that varies with time is a signal that carries information about the vibration of the structure (e.g., a machine tool) on which the accelerometer is installed. Such a signal can serve as a means for communicating the operation status of the machine tool to the machine operator.

1.1 Classification of Signals

In general, any signal can be broadly classified as being either deterministic or nondeterministic (Bendat and Piersol 2000). Deterministic signals are those that can be defined explicitly by mathematical functions. An example is the vibration caused by imbalance in a rolling bearing, when the bearing’s gravitational center does not coincide with the rotational center. Nondeterministic signals, in comparison, are random in nature and are described in statistical terms. An example is the acoustic emission signals generated during a machining process. In real-world applications, whether a measured signal is deterministic or nondeterministic depends on its reproducibility. A signal that can be generated repeatedly with identical results is considered to be deterministic, otherwise it is nondeterministic.

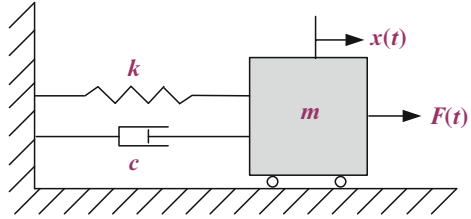
1.1.1 Deterministic Signal

There are two types of deterministic signals: periodic and transient. They are briefly explained and illustrated in the following.

1.1.1.1 Periodic Signal

A periodic signal is defined as a function that repeats itself exactly after a certain period of time, or cycle. Such a signal is mathematically expressed as

Fig. 1.1 A single degree of freedom (SDOF) mass spring damper system



$$x(t) = x(t + nT) \quad n \in \mathbb{Z} \quad (1.1)$$

In the above equation, \mathbb{Z} represents the integer set, n is an integer, and $T > 0$ represents the period. The simplest example of a periodic signal is the sinusoidal signal.

In practice, many physical systems can produce such a type of signal. A typical scenario is a single-degree-of-freedom (SDOF) mass-spring-damper system (Rao 2003). As illustrated in Fig. 1.1, the mass m is attached to the wall through a spring k and a damper c , and can vibrate in the horizontal direction. The motion (or displacement) of the mass-spring-damping system under input $F(t)$ is expressed as

$$m\ddot{x}(t) + c\dot{x}(t) + kx(t) = F(t) \quad (1.2)$$

where $x(t)$ is the displacement of the mass, $\dot{x}(t)$ the velocity of the mass, and $\ddot{x}(t)$ the acceleration of the mass.

Let us suppose that the system is under free vibration, with the external forcing input $F(t)$ being zero. Also assume that the damping coefficient $c = 0$. If the system is initially pulled away from the equilibrium position by a distance A_0 and released with the initial velocity equal to zero, so that

$$x(t = 0) = A_0 \quad \dot{x}(t = 0) = 0 \quad (1.3)$$

then the solution of (1.2) will generate a periodic signal with the period $T = 2\pi/\omega_n$. This will be a cosine function, as illustrated in Table 1.1a

A complex periodic signal can also be generated from the same system (Fig. 1.1) with $c = 0$, when the system is subject to a harmonic forcing input, $F(t) = F \cos(\omega t)$.

As illustrated in Table 1.1b, the complete response can be expressed as the sum of cosine waveforms of two different frequencies.

1.1.1.2 Transient Signal

A transient signal is defined as a function that lasts a short period of time. Such a signal can be generated by the system shown in Fig. 1.1, with the damping coefficient $c \neq 0$ and free vibration, as illustrated in Table 1.1c.

Table 1.1 Example of deterministic signals

Mathematical function	Waveform
(a) A simple periodic signal Condition $c = 0$ $F(t) = 0$ Solution $x(t) = A_0 \cos(\omega_n t)$	
(b) A complex periodic signal Condition $c = 0$ $F(t) = F \cos(\omega t)$ Solution $x(t) = A_1 \cos(\omega_n t) + A_2 \cos(\omega t)$	
(c) A transient signal Condition $c \neq 0$ $F(t) = 0$ Solution $x(t) = A_0 e^{-\zeta \omega_n t} \cos(\omega_d t)$	
(d) A mixed deterministic signal Condition $c \neq 0$ $F(t) = F \cos(\omega t)$ Solution $x(t) = A_0 e^{-\zeta \omega_n t} \cos(\omega_d t) + A_3 \cos(\omega t)$	

Note: $\omega_n = \sqrt{\frac{k}{m}}$, $\omega_d = \sqrt{1 - \zeta^2} \omega_n$, and $\zeta = \frac{c}{2m\omega_n} < 1$, $A_1 = A_0 \frac{F}{k - m\omega^2}$, $A_2 = \frac{F}{k - m\omega^2}$, $A_3 = \frac{F}{\sqrt{(k - m\omega^2)^2 + c^2 \omega^2}}$

Periodic and transient signals are often mixed together in real-world applications. Such a signal can be generated, for example, by the system shown in Fig. 1.1 with the damping coefficient $c \neq 0$, under a harmonic force, as illustrated in Table 1.1d.

1.1.2 Nondeterministic Signal

Nondeterministic signals, also called random signals, do not follow explicit mathematical expressions. They can be generally divided into two categories: stationary and nonstationary.

1.1.2.1 Stationary Signal

A signal $x(t)$ is considered *stationary* when none of its statistical properties change with time. Generally, wide-sense stationary (Bendat and Piersol 2000) is used to characterize the signal. This requires that it satisfies the following conditions on its mean function:

$$E\{x(t_1)\} = m_x(t_1) = m_x(t_1 + \tau) \quad \tau \in \mathbb{Z} \quad (1.4)$$

and the autocorrelation function:

$$E\{x(t_1), x(t_1 + \tau)\} = R_{xx}(t_1, t_1 + \tau) = R_{xx}(0, \tau) \quad \tau \in \mathbf{R} \quad (1.5)$$

In the above equations, the symbol τ is the real number, \mathbf{R} is defined as the real number set, and R_{xx} is the autocorrelation function of the signal $x(t)$. Equation (1.4) indicates that the mean function $m_x(t)$ must be time-invariant or remain unchanged as time goes by. As shown in (1.5), the autocorrelation function of the signal depends only on the time difference τ . The mean function and autocorrelation function of a signal can be obtained by time-averaging over a short time interval T as follows:

$$E\{x(t_1)\} = \frac{1}{T} \int_{t_1}^{t_1+T} x(t) dt \quad (1.6)$$

and

$$E\{x(t_1), x(t_1 + \tau)\} = \frac{1}{T} \int_{t_1}^{t_1+T} x(t)x(t + \tau) dt \quad (1.7)$$

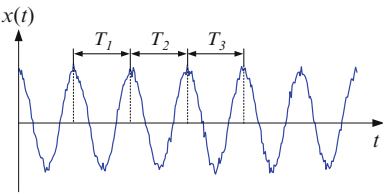
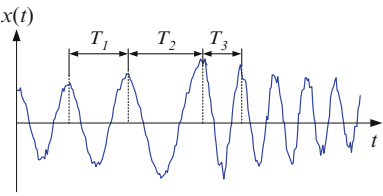
Table 1.2a illustrates an example of a stationary signal, which satisfies the two conditions expressed in (1.5) and (1.6).

1.1.2.2 Nonstationary Signal

A signal whose statistical properties change with time is called a nonstationary signal. As a result, a nonstationary signal does not satisfy the conditions specified in (1.4) and (1.5). Table 1.2b illustrates a nonstationary signal.

It should be noted that signal classification method as described above is not rigid and exclusive. No signals encountered in the real-world are exactly deterministic. Furthermore, there exist other means to classify a signal. For example, a signal can be considered as being either linear or nonlinear, as defined by the superposition principle. An SDOF mass-spring system is considered linear, if a linear relationship exists between the force input to the system and its corresponding displacement

Table 1.2 Example of nondeterministic signals

(a) Stationary signal	(b) Nonstationary signal
 <p>$T_1 = T_2 = T_3$</p>	 <p>$T_1 \neq T_2 \neq T_3$</p>

output. In real-world applications, a signal may contain some or several of the components described above.

1.2 Signals in Manufacturing

Signals are ubiquitously present in manufacturing machines and systems. For example, metal removal is essential to many manufacturing processes, as seen in turning, milling, and drilling (Schey 1999). During such a process, interactions between the cutting edge of the tool and the workpiece lead to removal of fragments of varying volumes, producing whereby time-varying or transient components in the vibration signals. Figure 1.2 illustrates the waveform of a vibration signal measured on a CNC milling machine center (shown in Fig. 1.3) when it is in production.

Another manufacturing process where transient signals may present is sheet metal stamping. The physical setup of a sheet metal stamping operation consists of three main components, namely, the die, the binder, and the punch (Suchy 2006), as shown in Fig. 1.4. During a stamping operation, the periphery of the sheet metal workpiece is held between the binder and die flange. As the punch moves down, the workpiece is pressed into the die, causing plastic deformation in the workpiece material. The flow of the workpiece material into the die is regulated by the binder force (Ahmetoglu et al. 1992; Koyama et al. 2004).

To characterize the stamping process, tonnage measurement has been conducted by placing accelerometers on the columns of the stamping machine. In Fig. 1.5, the output of an accelerometer is shown, in which four different phases of the forming operations are characterized: press idle, travel, contact, and free vibration. When the press is idle, the punch runs down until the binder touches the workpiece at point A. Then travel starts, and the signal amplitude increases as the stamping force increases, until it reaches point B where the punch touches the workpiece. After that, contact between the punch and workpiece is established, and metal forming starts. The signal quickly increases to its maximum at point C, as the punch pushes the sheet metal into the die. At point D, the forming process is completed, and the amplitude of the

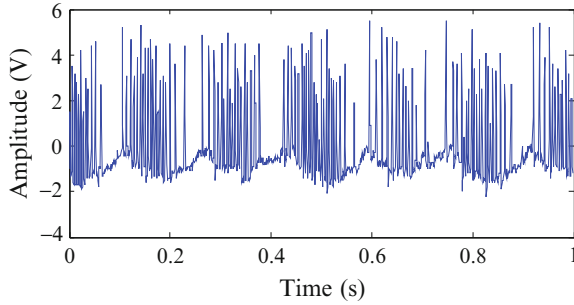


Fig. 1.2 Vibration signal measured during a milling process



Fig. 1.3 A CNC milling machine center (Haas Automation, Inc., <http://www.haascnc.com>)

vibration signal quickly drops to zero. After point E, vibration of the stamping machine diminishes with time, until the next stamping operation starts.

For nonmetallic material processing, injection molding is widely employed because of its capability in mass production of plastic parts. Figure 1.6 illustrates a typical injection molding machine. The injection molding process generally consists of four stages (Potsch and Michaeli 1995; Bryce 1996; Johannaber 2008): (1) plastication, where the raw material is melted in the barrel, (2) injection, during which the melted polymer is injected into the mold cavity, (3) packing, holding, and cooling, when additional polymer melt is forced into the cavity under high pressure to compensate for the volumetric shrinkage until the part is sufficiently solidified, and (4) ejection, where the mold opens and the part is ejected out of the mold by the push pins.

During each injection molding cycle, pressure within the mold cavity varies, as illustrated in Fig. 1.7. Such time-variation serves as a measure for identifying



Fig. 1.4 A typical stamping machine (BowStar Biz Management Ltd)

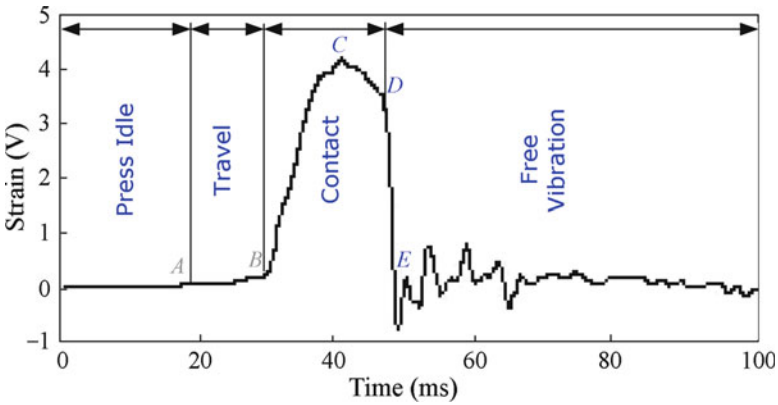


Fig. 1.5 A typical tonnage signal during stamping process



Fig. 1.6 A typical injection molding machine (Ferromatik Milacron)

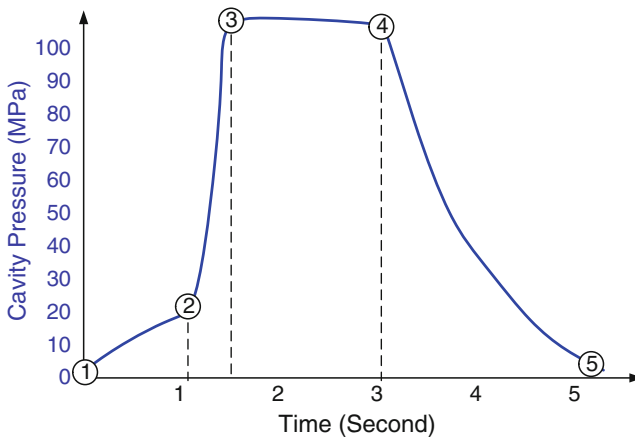


Fig. 1.7 Pressure signal measured during an injection molding process

and characterizing the various stages of the molding process. At point ① where the plasticized polymer enters the cavity, pressure starts to increase from zero in an approximately linear gradient relative to the duration of filling time. When the melt reaches the end of the cavity at point ②, the material is compacted to ensure reproduction of the mold cavity contour. Such a process is indicated by a fast pressure ramping rate as shown in the curve from ② to ③. During the holding phase from ③ to ④, a constant holding pressure is applied to the melt to compensate for the contraction of the polymer by injecting additional material into the cavity. As the molded part starts to cool down and solidify, viscosity of the material increases and the flow channel becomes constricted. As a result, pressure drop is seen from the sensor data as indicated by the section from ④ to ⑤.

The close association between signals and manufacturing, in addition to the various processes as illustrated above, is also seen in various components that have been employed in various machine equipment. One representative is the rolling bearings, which have been widely applied to providing loading support and rotational freedom in manufacturing, transportation, aerospace, and defense

(e.g., machine tools, trains, helicopters, power generator, etc.). Because of faulty installation, inappropriate lubrication, and other unpredictable adverse conditions during bearing operations, premature failure of bearings may occur, for example, in the form of surface spalling on the bearing raceways. As a result, impulsive signals will be generated every time when the rolling elements interact with the defects. These impulsive signals subsequently excite the machine system, leading to forced vibrations. Figure 1.8 illustrates a customized spindle system where bearings are installed. Vibration signals measured at two stages during a run-to-failure experiment on the spindle system are shown in Fig. 1.9.

Fig. 1.8 A customized spindle bearing test system

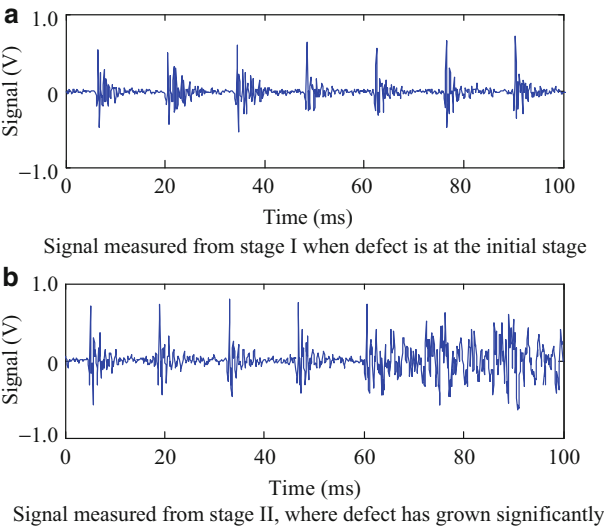
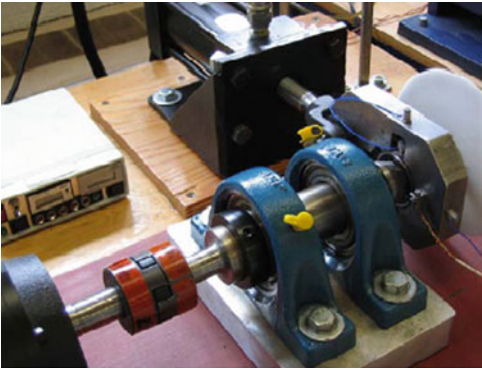


Fig. 1.9 The vibration signals from bearing run to failure test. (a) Signal measured from stage I when defect is at the initial stage and (b) signal measured from stage II, where defect has grown significantly

Gearbox, as illustrated in Fig. 1.10, has been employed in a wide range of machinery and control systems, because of its ability in transferring both power and motion with high efficiency. When a defect is developed in a gear, the vibration signal of the gearbox will contain amplitude and phase modulations that are periodic with respect to the rotation of the gear. Figure 1.11 shows an example of vibration signals measured on a gearbox under different running conditions.



Fig. 1.10 An automobile transmission gearbox (Topic Media PTY LTD)

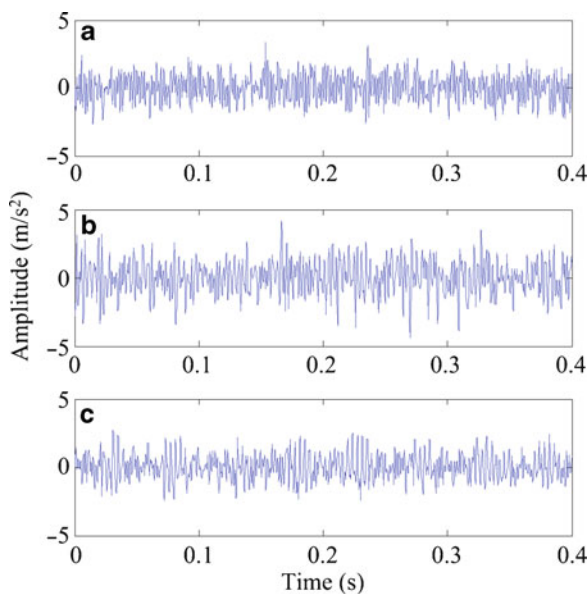


Fig. 1.11 Acceleration signals measured on a gearbox: (a) normal condition, (b) slight fault condition, and (c) severe fault condition

1.3 Role of Signal Processing for Manufacturing

Growing demand for high-quality and low-cost production has increased the need for condition monitoring, health diagnosis, and enhanced controls in manufacturing equipment and processes (Tönshoff et al. 1988; Byrne et al. 1996; Ganesan et al. 2004; Liang et al. 2004). Accordingly, sensor-based information acquisition and processing systems have gained increasing attention from the research community worldwide (Teti 1996; DimlaSnr 2000; Tseng and Chou 2002; Frankowiak et al. 2005). The goal of these efforts is to obtain information in real-time about the operation status of the machines and use the information for the following purposes:

1. Identification of machine faults at the incipient stage such that proper corrective measures can be taken before the faults have progressed to cause significant structural damage and costly downtime, thus enabling *adaptive* instead of fixed-time maintenance and production scheduling
2. More accurate control of the quality of products being manufactured, which is directly related to the working conditions of the machine

In addition to monitoring individual machines, data gathered from the sensors provide insight into the manufacturing process itself, and can be used to assist in high-level decision-making for production optimization.

Signals encountered in manufacturing machines typically consist of three major components:

1. A periodic component resulting from the cyclic interactions between the interfacing elements of the machine, such as vibrations caused by the interaction between the rolling elements and the raceway
2. A transient component caused by “one-time” events, such as the sudden breakage of a drilling bit or the inception of a crack inside a workpiece
3. Broadband background noise

Detection of the existence of these signals in real-time during the manufacturing process and extracting relevant information from the signals in a timely manner are of significant interest and importance, as they are precursors of potential machine defect and product quality deterioration that will negatively impact the manufacturing processes. On the other hand, detection of such signals can be challenging, as these signals are generally short in duration and weak in amplitude. Often times, they can be buried under strong background noises, making their detection difficult (Gu et al. 2002; Padovese 2004; Shi et al. 2004). Furthermore, the one-shot nature of these signals makes the assumption for stationary signals invalid, thus reducing the effectiveness of conventional signal processing techniques. For example, while Fourier transform has been extensively used in conjunction with filtering techniques, its effective utilization depends upon signals containing distinct characteristic frequency components of sufficient energy content, within a limited frequency band. If the feature components spread over a wide

spectrum, it would be difficult to use Fourier transform to differentiate them from disturbing or masking components, especially when the feature components are weak in magnitude. This has been shown in condition-monitoring studies of bearings with an incipient defect (Mori et al. 1996).

Time-frequency and time-scale techniques have been the subject of extensive research over the past decade for nonstationary signal analysis. Typical representatives include the short-time Fourier transform (STFT) and wavelet transform (Li and Ma 1997; Satish 1998). STFT was developed to address the limitation of the Fourier transform, which is rooted in its basis functions extending over an infinite period of time. As a result, Fourier transform is not well adapted to nonstationary transient signals with short durations. A solution to this problem is to perform a “time localized” Fourier transform within a sliding window, as in the case of STFT (Chui 1992). Popular choices for the window function include the Hamming, the Hann, and the Gaussian functions. When a Gaussian window is chosen, the STFT is called a Gabor transform (Gabor 1946). A one-sided Gaussian window has been used for detection of transient signals in a workpiece (Friedlander and Porat 1989). The disadvantage of the STFT is that its *time resolution* (the smallest separation in time of two signal components that can be discriminated) and *bandwidth* cannot be chosen to be simultaneously small, according to the uncertainty principle (Cohen 1989). The time-bandwidth product of the STFT must be greater than or equal to the inverse of 4π . The equal sign holds only when the window function is a Gaussian function. This means that the time-frequency resolution over the entire time-frequency plane is fixed, once the window function is chosen. As a result, a trade-off must be made between the time resolution and frequency resolution, when the STFT is applied to transient signal analysis.

To overcome the resolution limitation of the Gabor transform, the wavelet transform has been increasingly investigated for nonstationary signal analysis (Mallat 1989; Daubechies 1990, 1992; Rioul and Vetterli 1991). In contrast to the Gabor transform with fixed windows, the wavelet transform uses short windows at high frequencies and long windows at low frequencies (Rioul and Vetterli 1991). Such a nature leads to the wavelet transform being called the *constant relative-bandwidth* frequency analysis. Unlike the Fourier transform, which expresses a signal as the sum of a series of single-frequency sine and cosine functions, the wavelet transform decomposes a signal into a set of *basis* functions. These basis functions are obtained from a single base wavelet function by a two-step operation: *scaling* (through *dilation* and *contraction* of the base wavelet along the time axis, as will be explained in Chap. 2), and *time shift* (i.e., *translation* along the time axis). Essentially, the wavelet transform process measures the “similarity” between the signal being analyzed and the base wavelet. Through variations of the scales and time shifts of the base wavelet function, features hidden within the signal can be extracted, without requiring the signal to have a dominant frequency band.

Research on manufacturing machine and process monitoring and diagnosis using the wavelet transform has attracted increasing attention worldwide. For example, the adaptive capability has made wavelet transform a good analytical

tool for decomposing gearbox vibration signals. Studies have demonstrated its ability to detect incipient failures as well as differentiating different types of defects (Wang and McFadden 1993, 1995; Zheng et al. 2002). The discrete wavelet transform has been applied to analyzing spindle motor current for tool failure diagnosis in end-milling, under varying cutting conditions (Lee and Tarn 1999). Similar studies of wavelet transform for machine tool monitoring have been reported (Fu et al. 1998; Li et al. 2000). For detecting localized bearing defects and/or estimating the defect severity level, the advantage of wavelet transform has been extensively investigated (Wang and Gao 2003; Lou and Loparo 2004; Yan and Gao 2005; Chiementin et al. 2007; Wang et al. 2009), and the results have shown its superior performance over the conventional, Fourier-based approaches. Other applications of wavelet transform, including singularity detection (Sun and Tang 2002), denoising and extraction of weak signals (Altmann and Mathew 2001; Lin 2001), vibration signal compression (Tanaka et al. 1997; Staszewski 1998), and system and parameter identification (Robertson et al. 1998; Kim et al. 2001), have also been reported.

It can be concluded that the wavelet transform provides a powerful mathematical tool for the analysis, characterization, and classification of nonstationary signals typically seen in manufacturing. The adaptive, multiresolution capability of the wavelet transform makes it well suited for decomposing signals of varying time and frequency resolutions that are characteristic of the underlying defect mechanisms associated with a machine, a dynamical structure, or a manufacturing process. Such capability makes the wavelet transform an enabling tool for advancing the science base of signal processing in manufacturing. It is such significance and the associated potential impact that motivate this book, and it is the intention of the book to provide graduate students and practicing engineers with a systematic, comprehensive, yet easily accessible coverage of the fundamental theory and representative applications of wavelet transform in the broad and vibrant field of manufacturing research.

1.4 References

- Ahmetoglu MA, Altan T, Kinzel GL (1992) Improvement of part quality in stamping by controlling workpiece holder force and contact pressure. *J Mater Process Technol* 33:195–214
- Altmann J, Mathew J (2001) Multiple band pass autoregressive demodulation for rolling element bearing fault diagnostics. *Mech Syst Signal Process* 15:963–977
- Bendat JS, Piersol AG (2000) *Random data analysis and measurement procedures*, 3rd edn. Wiley, New York
- BowStar Biz Management Ltd, http://www.bowstar.hk.com/images/stamping_machine_1.jpg
- Bryce DM (1996) *Plastic injection molding: mold design and construction fundamentals*. Society of Manufacturing Engineers, Dearborn, MI
- Byrne G, Dornfeld D, Inasaki I, Ketteler G, König W, Teti R (1996) Tool condition monitoring (TCM) – the status of research and industrial application. *Ann CIRP* 44(2):541–567
- Chiementin X, Bolaers F, Dron J (2007) Early detection of fatigue damage on rolling element bearings using adapted wavelet. *J Vib Acoust* 129(4):495–506

- Chui CK (1992) An introduction to wavelets. Academic, New York
- Cohen L (1989) Time frequency distributions – a review. *Proc IEEE* 77(7):941–981
- Daubechies I (1990) The wavelet transform, time frequency localization and signal analysis. *IEEE Trans Inf Theory* 36(5):960–1005
- Daubechies I (1992) Ten lectures on wavelets. SIAM, Philadelphia, PA
- DimlaSnr DE (2000) Sensor signals for tool wear monitoring in metal cutting operations – a review of methods. *Int J Mach Tools Manuf* 40(8):1073–1098
- Ferromatik Milacron, http://www.ferromatik.com/de/information/presse/img/K_TEC_200_S_auf_weiss.jpg
- Frankowiak M, Grosvenor R, Prickett P (2005) A review of the evolution of microcontroller based machine and process monitoring. *Int J Mach Tools Manuf* 45(4–5):573–582
- Friedlander B, Porat B (1989) Detection of transient signals by the Gabor representation. *IEEE Trans Acoust Speech Signal Process* 37(2):169–180
- Fu J, Troy C, Phillips P (1998) Matching pursuit approach to small drill bit breakage prediction. *Int J Prod Res* 37(14):3247–3261
- Gabor D (1946) Theory of communication. *J Inst Electr Eng* 93:429–457
- Ganesan R, Das TK, Venkataraman V (2004) Wavelet based multiscale statistical process monitoring: a literature review. *IEEE Trans* 36:787–806
- Gu S, Ni J, Yuan J (2002) Non stationary signal analysis and transient machining process condition monitoring. *Int J Mach Tools Manuf* 42:41–51
- Johannaber F (2008) Injection molding machines: a user's guide. 4th edn. Hanser Gardner Publications, Cincinnati, OH
- Kim YY, Hong YC, Lee NY (2001) Frequency response function estimation via a robust wavelet de noising method. *J Sound Vib* 244:635–649
- Koyama H, Wagoner RH, Manabe K (2004) Workpiece holding force in panel stamping process using a database and FEM assisted intelligent press control system. *J Mater Process Technol* 152:190–196
- Lee BY, Tarng YS (1999) Application of the discrete wavelet transform to the monitoring of tool failure in end milling using the spindle motor current. *Int J Adv Manuf Technol* 15(4):238–243
- Li M, Ma J (1997) Wavelet decomposition of vibrations for detection of bearing localized defects. *NDTE Int* 30(3):143–149
- Li X, Tso S, Wang J (2000) Real time tool condition monitoring using wavelet transforms and fuzzy techniques. *IEEE Trans Syst Man Cybern C Appl Rev* 30(3):352–357
- Liang S, Hecker R, Landers R (2004) Machining process monitoring and control: the state of the art. *ASME J Manuf Sci Eng* 126(2):297–310
- Lin J (2001) Feature extraction of machine sound using wavelet and its application in fault diagnostics. *NDTE Int* 34:25–30
- Lou X, Loparo KA (2004) Bearing fault diagnosis based on wavelet transform and fuzzy inference. *Mech Syst Signal Process* 18(5):1077–1095
- Mallat S (1989) A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans Pattern Anal Mach Intell* 2(7):674–693
- Mori K, Kasashima N, Yoshioka T, Ueno Y (1996) Prediction of spalling on a ball bearing by applying the discrete wavelet transform to vibration signals. *Wear* 195(1–2):162–168
- Padovese LR (2004) Hybrid time frequency methods for non stationary mechanical signal analysis. *Mech Syst Signal Process* 18(5):1047–1064
- Potsch G, Michaeli W (1995) Injection molding: an introduction. Hanser Gardner Publications, Cincinnati, OH
- Rao SS (2003) Mechanical vibration. 4th edn. Prentice Hall, Old Tappan, NJ
- Rioul O, Vetterli M (1991) Wavelets and signal processing. *IEEE Signal Process Mag* 8(4):14–38
- Robertson AN, Park KC, Alvin KF (1998) Extraction of impulse response data via wavelet transform for structural system identification. *ASME J Vib Acoust* 120:252–260

- Satish L (1998) Short time Fourier and wavelet transform for fault detection in power transformers during impulse tests. *IEEE Proc Sci Meas Tech* 145(2):77 84
- Schey JA (1999) Introduction to manufacturing processes. 3rd edn, McGraw Hill Science/Engineering/Math, New York
- Shi DF, Tsung F, Unsworth PJ (2004) Adaptive time frequency decomposition for transient vibration monitoring of rotating machinery. *Mech Syst Signal Process* 18(1):127 141
- Staszewski WJ (1998) Wavelet based compression and feature selection for vibration analysis. *J Sound Vib* 211:735 760
- Suchy I (2006) Handbook of die design. 2nd edn, McGraw Hill, New York
- Sun Q, Tang Y (2002) Singularity analysis using continuous wavelet transform for bearing fault diagnosis. *Mech Syst Signal Process* 16:1025 1041
- Tanaka M, Sakawa M, Kato K, Abe M (1997) Application of wavelet transform to compression of mechanical vibration data. *Cybern Syst* 28(3):225 244
- Teti R (1996) A review of tool condition monitoring literature data base. *Ann CIRP* 44(2):659 666
- Tönshoff H, Wulfsberg J, Kals H, König W (1988) Developments and trends in monitoring and control of machining processes. *Ann CIRP* 37(2):611 622
- Topic Media PTY LTD. <http://www.zcars.com.au/images/ford powershift gearbox12.jpg>
- Tseng PC, Chou A (2002) The intelligent on line monitoring of end milling. *Int J Mach Tools Manuf* 42(1):89 97
- Wang C, Gao R (2003) Wavelet transform with spectral post processing for enhanced feature extraction. *IEEE Trans Instrum Meas* 52:1296 1301
- Wang WJ, McFadden PD (1993) Application of the wavelet transform to gearbox vibration analysis. *ASME Pet Div* 52:13 20
- Wang WJ, McFadden PD (1995) Application of orthogonal wavelets to early gear damage detection. *Mech Syst Signal Process* 9(5):497 507
- Wang C, Gao RX, Yan R (2009) Unified time scale frequency analysis for machine defect signature extraction: theoretical framework. *Mech Syst Signal Process* 23(1):226 235
- Yan R, Gao R (2005) An efficient approach to machine health diagnosis based on harmonic wavelet packet transform. *Robot Comput Integr Manuf* 21:291 301
- Zheng H, Li Z, Chen X (2002) Gear fault diagnosis based on continuous wavelet transform. *Mech Syst Signal Process* 16(2 3):447 457

Chapter 2

From Fourier Transform to Wavelet Transform: A Historical Perspective

To ensure safe and economical operation and product quality, manufacturing machines and processes are constantly monitored and evaluated for their working conditions, on the basis of signals collected by sensors, which are generally presented in the form of time series (e.g., time-dependent variation of vibration, pressure, temperature, etc.). To extract information from such signals and reveal the underlying dynamics that corresponds to the signals, proper signal processing technique is needed. Typically, the process of signal processing transforms a time-domain signal into another domain, with the purpose of extracting the characteristic information embedded within the time series that is otherwise not readily observable in its original form. Mathematically, this can be achieved by representing the time-domain signal as a series of coefficients, based on a comparison between the signal $x(t)$ and a set of known, template functions $\{\psi_n(t)\}_{n \in \mathbb{Z}}$ as (Chui 1992; Qian 2002)

$$c_n = \int_{-\infty}^{\infty} x(t) \psi_n^*(t) dt \quad (2.1)$$

where $(\cdot)^*$ stands for the complex conjugate of the function (\cdot) . The inner product between the two functions $x(t)$ and $\psi_n(t)$ is defined as

$$\langle x, \psi_n \rangle = \int x(t) \psi_n^*(t) dt \quad (2.2)$$

Then (2.1) can be expressed in the general form as

$$c_n = \langle x, \psi_n \rangle \quad (2.3)$$

The inner product in (2.3), in essence, describes an operation of comparing the “similarity” between the signal $x(t)$ and the template function $\{\psi_n(t)\}_{n \in \mathbb{Z}}$, that is, the degree of closeness between the two functions. The more similar $x(t)$ is to $\psi_n(t)$, the larger the inner product c_n will be. On the basis of this notion, this chapter presents a historical perspective on the evolution of the wavelet transform. This is realized by observing the similarities as well as differences between the wavelet

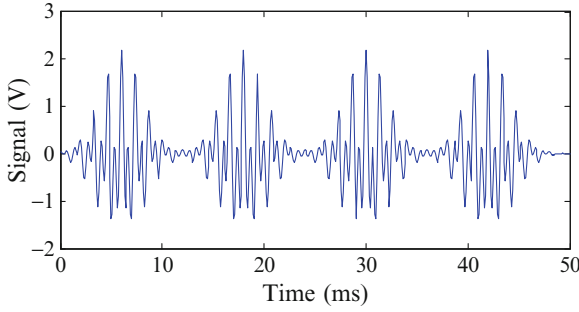


Fig. 2.1 A nonstationary signal $x(t)$

transform and other commonly used techniques, in terms of the choice of the template functions $\{\psi_n(t)\}_{n \in \mathbb{Z}}$. To illustrate the point, a nonstationary signal as shown in Fig. 2.1 is used as an example. The signal consists of four groups of impulsive signal trains, each containing two transient elements of different center frequencies at 1,500 and 650 Hz, respectively. The four groups are separated from one another by a 12-ms time interval. Within each group, the two transient elements are time-overlapped. The sampling frequency used to capture the signal is 10 kHz.

2.1 Fourier Transform

The Fourier transform is probably the most widely applied signal processing tool in science and engineering. It reveals the frequency composition of a time series $x(t)$ by transforming it from the time domain into the frequency domain. In 1807, the French mathematician Joseph Fourier (Fig. 2.2) found that any periodic signal can be presented by a weighted sum of a series of sine and cosine functions. However, because of the uncompromising objections from some of his contemporaries such as J. L. Lagrange (Herivel 1975), his paper on this finding never got published, until some 15 years later, when Fourier wrote his own book, *The Analytical Theory of Heat* (Fourier 1822). In that book, Fourier extended his finding to aperiodic signals, stating that an aperiodic signal can be represented by a weighted integral of a series of sine and cosine functions. Such an integral is termed the Fourier transform.

Using the notation of inner product, the Fourier transform of a signal $x(t)$ can be expressed as

$$X(f) = \langle x, e^{i2\pi ft} \rangle = \int_{-\infty}^{\infty} x(t) e^{-i2\pi ft} dt \quad (2.4)$$



*“An arbitrary function,
continuous or with
discontinuities, defined in a finite
interval by an arbitrarily
capricious graph can always be
expressed as a sum of sinusoids”
J.B.J. Fourier*

Fig. 2.2 Jean B. Joseph Fourier (1768–1830)

Assuming that the signal has finite energy,

$$\int_{-\infty}^{\infty} |x(t)|^2 dt < \infty$$

Accordingly, the inverse Fourier transform of the signal $x(t)$ can be expressed as

$$x(t) = \int_{-\infty}^{\infty} X(f) e^{i2\pi ft} df \quad (2.5)$$

Signals obtained experimentally through a data acquisition system are generally sampled at discrete time intervals ΔT , instead of continuously, within a total measurement time T . Such a signal, defined as x_k , can be transformed into the frequency domain by using the discrete Fourier transform (DFT), defined as

$$DFT(f_n) = \frac{1}{N} \sum_{k=0}^{N-1} x_k e^{i2\pi f_n k \Delta T} \quad (2.6)$$

where $N = T/\Delta T$ is the number of samples, and $f_n = n/T$, $n = 0, 1, 2, \dots, N-1$ are the discrete frequency components. The inverse DFT can then be expressed as

$$x_k = \frac{1}{\Delta T} \sum_{f_n=0}^{(N-1)/T} DFT(f_n) e^{i2\pi f_n k \Delta T} \quad (2.7)$$

Equations (2.4) and (2.6) indicate that the Fourier transform is essentially a convolution between the time series $x(t)$ or x_k and a series of sine and cosine functions that can be viewed as template functions. The operation measures the similarity between $x(t)$ or x_k and the template functions, and expresses the average frequency information during the entire period of the signal analyzed. In Fig. 2.3, such an operation is graphically illustrated.

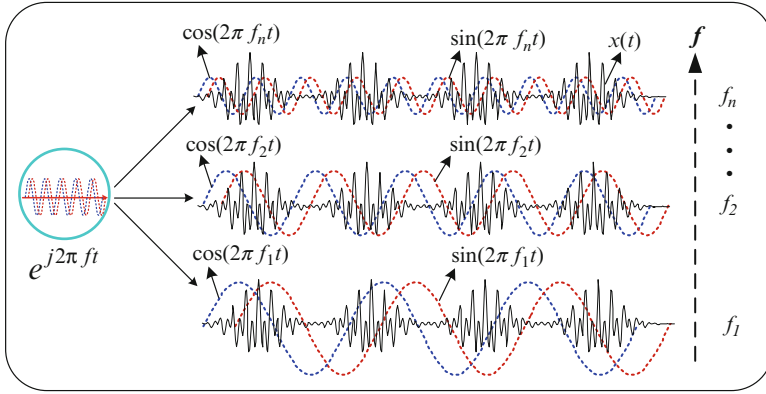


Fig. 2.3 Illustration of the Fourier transform of a continuous signal $x(t)$

To compute the DFT of a signal with N samples, multiplication of an $N \times N$ matrix that contains the primitive n th root of unity $e^{i2\pi/N}$ by the signal is needed. Such an operation takes a total of arithmetic operations on the order of N^2 to complete. The computational time increases quickly as the number of the samples increases. For example, a time series of $N = 256$ (i.e., 2^8) samples takes 65,536 operational steps to complete, whereas for $N = 4,096$ (i.e., 2^{12}), a total of 16,777,216 steps will be needed to compute its DFT. The high computational cost limited the widespread application of the DFT in its early stage, until a more efficient algorithm, called the Cooley Tukey algorithm, was introduced in 1965 (Cooley and Tukey 1965). This algorithm is also called the fast Fourier transform (FFT), and what it does is to recursively break down a DFT of a large data sample (i.e., a large N) into a series of smaller DFTs of smaller samples by dividing the transform with size N into two pieces of size $N/2$ at each step, and reduce the arithmetic operations to a total of $N \log(N)$. Comparing to the N^2 operations required for DFT, this represents a time reduction of up to 96%, when, for example, the data sample number N is 256.

In practice, the phenomena of *leakage* and *aliasing* can happen during the calculation of DFT (Körner 1988). Leakage is caused by the discontinuities involved when a signal is extended periodically for performing the DFT. Applying a window to the signal to force it to contain a full period can prevent leakage from happening. However, the window itself may contribute frequency information to the signal. Aliasing occurs when the Shannon's sampling theorem is violated, (Bracewell 1999) causing the actual frequency component to appear at different locations in the frequency spectrum. This can be solved by ensuring the sampling frequency to be at least twice as large as the maximum frequency component contained in the signal (Bracewell 1999). This requires, however, that the maximum frequency component is known a priori.

The Fourier transform of the signal shown in Fig. 2.1 is illustrated in Fig. 2.4. The figure shows two major frequency peaks at 650 and 1,500 Hz, respectively.

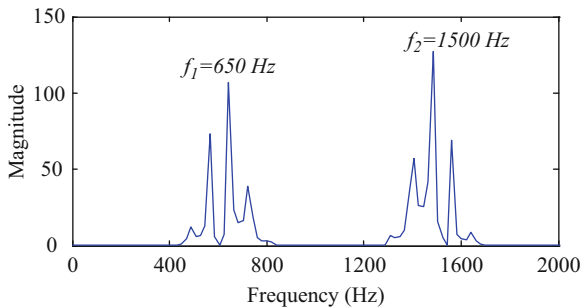


Fig. 2.4 Fourier transform results of the signal $x(t)$

However, it does not reveal how the signal's frequency contents vary with time; that is, the figure does not reveal if the two frequency components are continuously present throughout the time of observation or only at certain intervals, as is implicitly shown in the time-domain representation. Because the temporal structure of the signal is not revealed, the merit of the Fourier transform is limited; specifically, it is not suited for analyzing nonstationary signals. On the other hand, as signals encountered in manufacturing are generally nonstationary in nature (e.g., subtle, time-localized changes caused by structural defects are typically seen in vibration signals measured from rotary machines), a new signal processing technique that is able to handle the nonstationarity of a signal is needed.

2.2 Short-Time Fourier Transform

A straightforward solution to overcoming the limitations of the Fourier transform is to introduce an analysis window of certain length that glides through the signal along the time axis to perform a “time-localized” Fourier transform. Such a concept led to the short-time Fourier transform (STFT), introduced by Dennis Gabor (Fig. 2.5) in his paper titled “Theory of communication,” published in 1946 (Gabor 1946).

As shown in Fig. 2.6, the STFT employs a sliding window function $g(t)$ that is centered at time τ . For each specific τ , a time-localized Fourier transform is performed on the signal $x(t)$ within the window. Subsequently, the window is moved by τ along the time line, and another Fourier transform is performed. Through such consecutive operations, Fourier transform of the entire signal can be performed. The signal segment within the window function is assumed to be approximately stationary. As a result, the STFT decomposes a time domain signal into a 2D time-frequency representation, and variations of the frequency content of that signal within the window function are revealed, as illustrated in Fig. 2.6.

Fig. 2.5 Dennis Gabor
(1900–1979)

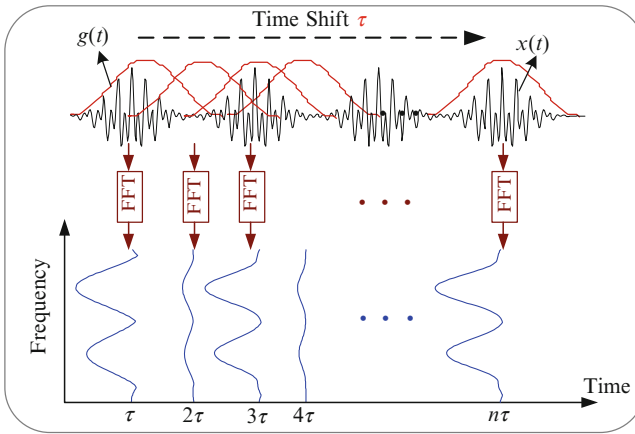
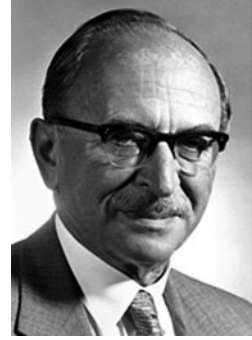


Fig. 2.6 Illustration of short time Fourier transform on the test signal $x(t)$

Using the inner product notation as before, the STFT can be expressed as

$$STFT(\tau, f) = \langle x, g_{\tau, f} \rangle = \int x(t) g_{\tau, f}^*(t) dt = \int x(t) g(t - \tau) e^{j2\pi f t} dt \quad (2.8)$$

Equation (2.8) can also be viewed as a measure of “similarity” between the signal $x(t)$ and the time-shifted and frequency-modulated window function $g(t)$. Over the past few decades, various types of window functions have been developed (Oppenheim et al. 1999), and each of them is specifically tailored toward a particular type of application. For example, the Gaussian window designed for analyzing transient signals, and the Hamming and Hann windows are applicable to narrowband, random signals, and the Kaiser-Bessel window is better suited for separating two signal components with frequencies very close to each other but with widely differing amplitudes. It should be noted that the choice of the window function

directly affects the time and frequency resolutions of the analysis result. While higher resolution in general provides better separation of the constituent components within a signal, the time and frequency resolutions of the STFT technique cannot be chosen arbitrarily at the same time, according to the uncertainty principle (Cohen 1989). Specifically, the product of the time and frequency resolutions is lower bounded by

$$\Delta\tau \cdot \Delta f \geq \frac{1}{4\pi} \quad (2.9)$$

where $\Delta\tau$ and Δf denote the time and frequency resolutions, respectively. Analytically, the time resolution $\Delta\tau$ is measured by the root-mean-square time width of the window function, defined as

$$\Delta\tau^2 = \frac{\int \tau^2 |g(\tau)|^2 d\tau}{\int |g(\tau)|^2 d\tau} \quad (2.10)$$

Similarly, the frequency resolution Δf is measured by the root-mean-square bandwidth of the window function, and is defined as (Rioul and Vetterli 1991)

$$\Delta f^2 = \frac{\int f^2 |G(f)|^2 df}{\int |G(f)|^2 df} \quad (2.11)$$

In (2.11), $G(f)$ is the Fourier transform of the window function $g(t)$. As an example, the Gaussian window function $g(t) = e^{-\alpha^2 t^2}$ (with α being a constant and τ controlling the window width) has the time and frequency resolutions of $\Delta\tau = \tau/(2\sqrt{\alpha})$ and $\Delta f = \sqrt{\alpha}/(\tau \cdot 2\pi)$, respectively. As a result, the time-frequency resolution provided by the Gaussian window when analyzing a signal $x(t)$ is $\Delta\tau \cdot \Delta f = 1/4\pi$. As the time and frequency resolutions of a window function are dependent on the parameter τ only, once the window function is chosen, the time and frequency resolutions over the entire time-frequency plane are fixed. Illustrated in Fig. 2.7 are two scenarios where the products of the time and frequency resolutions of the window function (i.e., the area defined by the product of $\Delta\tau \cdot \Delta f$) are the same, regardless of the actual window size (τ or $\tau/2$) employed.

The effect of the window size τ on the time and frequency resolutions is illustrated in Fig. 2.8, where STFT with the Gaussian window was performed on the signal shown in Fig. 2.1. Altogether three different window sizes (i.e., 1.6, 6.4, and 25.6 ms) were chosen. While the smallest window width of 1.6 ms has provided high time resolution in separating the four pulse trains contained in the signal, as illustrated in Fig. 2.8a, its frequency resolution was too low to differentiate the two time-overlapped transient elements within each group. As a result, the frequency elements 1,500 and 650 Hz are displayed as one lumped group on the time-frequency plane. In contrast, the largest window width

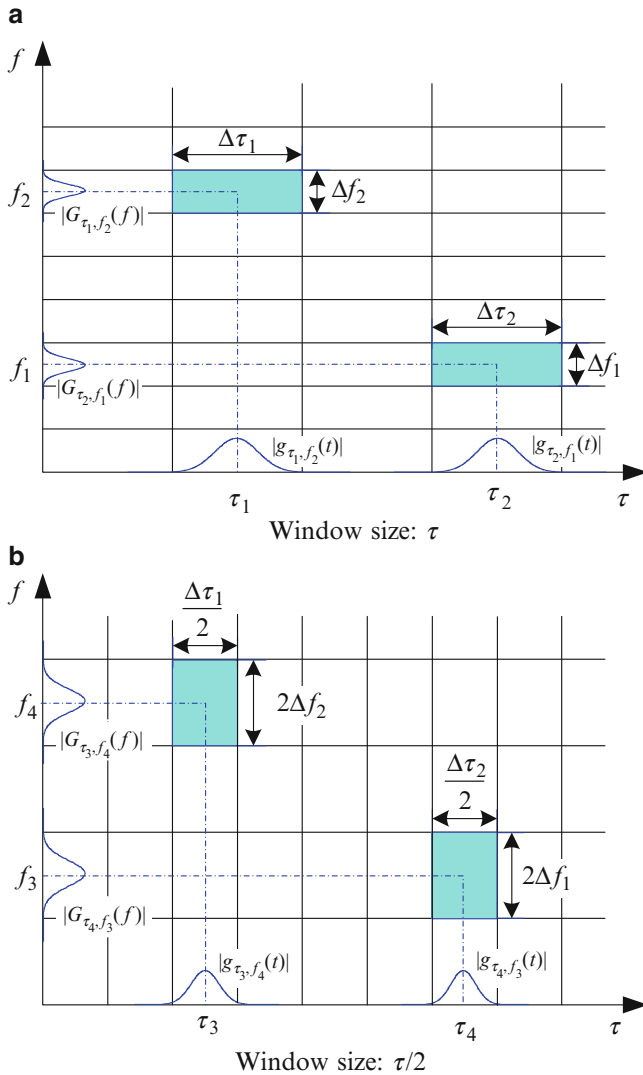
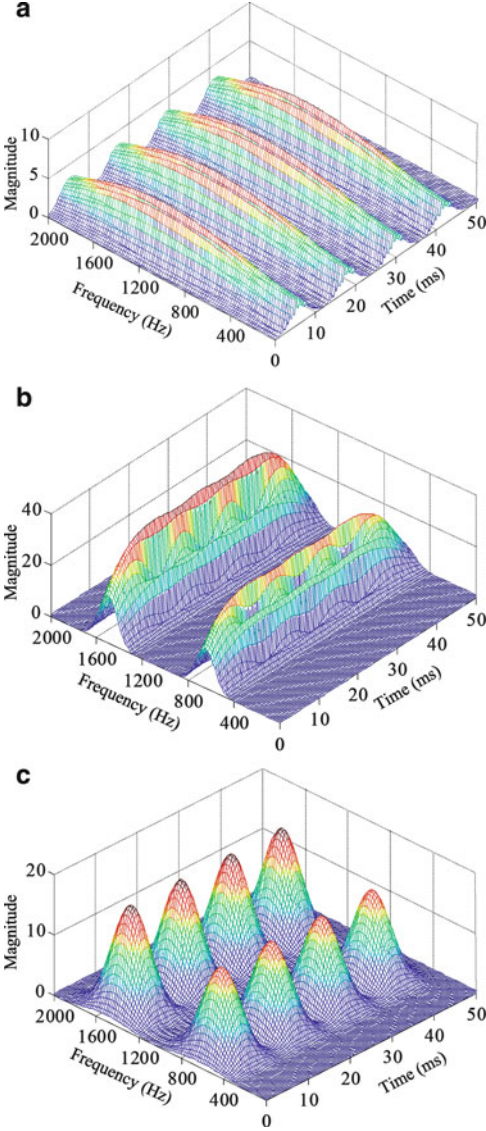


Fig. 2.7 Time frequency resolutions associated with the STFT technique. (a) Window size τ and (b) window size $\tau/2$

of 25.6 ms provided good frequency resolution to illustrate the two frequency components in Fig. 2.8b. However, the time-resolution was insufficient to differentiate the four pulse trains that are timely separated with a 12-ms interval. The best overall performance is given by the window width of 6.4 ms, shown in Fig. 2.8c, which allowed for all of the transients to be adequately differentiated on the time-frequency plane. Given that the specific frequency content of an

Fig. 2.8 Results of the STFT of the signal using three different window sizes. (a) Window size 1.6 ms, (b) window size 25.6 ms, and (c) window size 6.4 ms

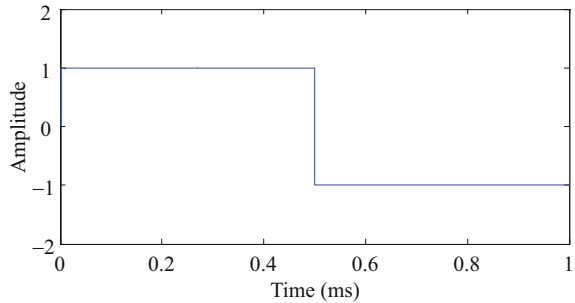


experimentally measured signal is generally not known a priori, selection of a suitable window size for effective signal decomposition using the STFT technique is not guaranteed. The inherent drawback of the STFT motivates researchers to look for other techniques that are better suited for processing nonstationary signals. One of such techniques, which is the focus of this book, is the wavelet transform.

Fig. 2.9 Alfred Haar
(1885–1933)



Fig. 2.10 The rectangular basis function



2.3 Wavelet Transform

From a historical point of view, the first reference to the wavelet goes back to the early twentieth century when Alfred Haar (Fig. 2.9) wrote his dissertation titled “On the theory of the orthogonal function systems” in 1909 to obtain his doctoral degree at the University of Göttingen. His research on orthogonal systems of functions led to the development of a set of rectangular basis functions (Haar 1910), as illustrated in Fig. 2.10. Later, an entire wavelet family, the Haar wavelet, was named on the basis of this set of functions, and it is also the simplest wavelet family developed till this date.

Essentially, Haar’s basis function consists of a short positive pulse followed by a short negative pulse, and it was used to illustrate a countable orthonormal system for the space of square-integrable functions on the real line (Haar 1910). Later, the Haar basis function was applied to compress images (DeVore et al. 1992).

Little advancement in the field of wavelets was reported after Haar’s work, until a physicist, Paul Levy (Fig. 2.11), investigated the Brownian motion in the 1930s. He discovered that the scale-varying function, that is, the Haar basis function, was better suited than the Fourier basis functions for studying subtle details in the Brownian motion. In addition, the Haar basis function can be scaled into different intervals, such as the interval $[0, 1]$ or the intervals $[0, 1/2]$ and $[1/2, 1]$, thereby providing higher precision when modeling a function than that provided by the Fourier basis function, as it can only have one interval $[-\infty, \infty]$.

Fig. 2.11 Paul Levy
(1886 1971)



Fig. 2.12 Jean Morlet
(1931 2007)



While several individuals, such as John Littlewood, Richard Paley (Littlewood and Paley 1931), Elias M. Stein (Jaffard et al. 2001), and Norman H. Ricker (Ricker 1953) have contributed, from the 1930s to the 1970s, to advancing the state of research in wavelets as it is called today, major advancement in the field was attributed to Jean Morlet (Fig. 2.12) who developed and implemented the technique of scaling and shifting of the analysis window functions in analyzing acoustic echoes while working for an oil company in the mid 1970s (Mackenzie 2001). By sending acoustic impulses into the ground and analyzing the received echoes, the existence of oil beneath the earth crust as well as the thickness of the oil layer can be identified. When Morlet first used the STFT to analyze these echoes, he found that keeping the width of the window function fixed did not work. As a solution to the problem, he experimented with keeping the frequency of the window function constant while changing the width of the window by stretching or squeezing the window function (Mackenzie 2001). The resulting waveforms of varying widths were called by Morlet the “Wavelet”, and this marked the beginning of the era of wavelet research. As a matter of fact, the approach that Morlet used was similar to what Haar did before, but the theoretical formation of the wavelet transform was first proposed only after Jean Morlet teamed up with Alex Grossmann to work out the idea that a signal could be transformed into the form of a wavelet and then transformed back into its original form without any information loss (Grossmann and Morlet 1984).

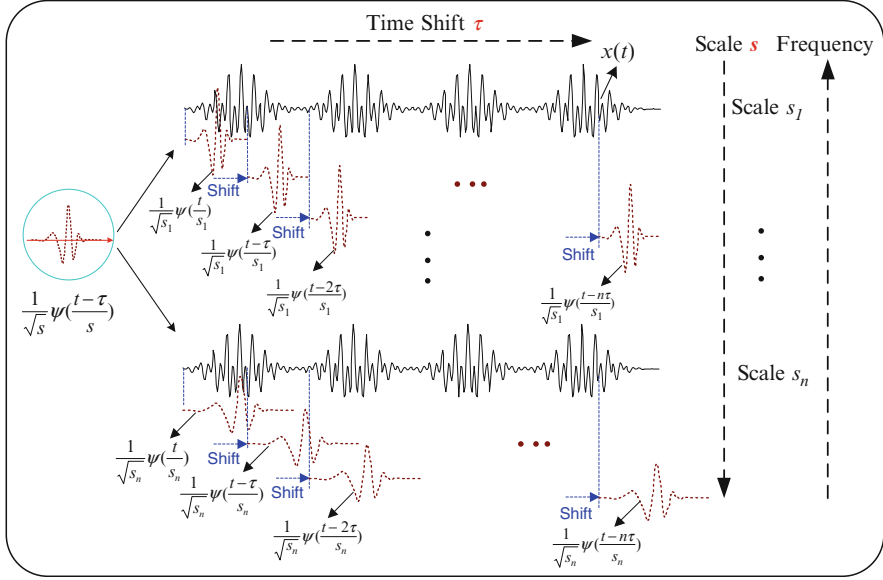


Fig. 2.13 Illustration of wavelet transform

In contrast to the STFT technique where the window size is *fixed*, the wavelet transform enables *variable* window sizes in analyzing different frequency components within a signal (Mallat 1998). This is realized by comparing the signal with a set of template functions obtained from the *scaling* (i.e., dilation and contraction) and *shift* (i.e., translation along the time axis) of a *base* wavelet $\psi(t)$ and looking for their similarities, as illustrated in Fig. 2.13.

Using again the notation of inner product, the wavelet transform of a signal $x(t)$ can be expressed as

$$wt(s, \tau) = \langle x, \psi_{s, \tau} \rangle = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} x(t) \psi^* \left(\frac{t - \tau}{s} \right) dt \quad (2.12)$$

where the symbol $s > 0$ represents the scaling parameter, which determines the time and frequency resolutions of the scaled base wavelet $\psi(t - \tau/s)$. The specific values of s are inversely proportional to the frequency. The symbol τ is the shifting parameter, which translates the scaled wavelet along the time axis. The symbol $\psi^*(\cdot)$ denotes the complex conjugation of the base wavelet $\psi(t)$. As an example, if the Morlet wavelet $\psi(t) = e^{i2\pi f_0 t} e^{-(\alpha t^2/\beta^2)}$ is chosen as the base wavelet, its scaled version will be expressed as

$$\psi \left(\frac{t - \tau}{s} \right) = e^{i2\pi f_0 \frac{t - \tau}{s}} e^{-\frac{\alpha (t - \tau)^2}{s^2 \beta^2}} \quad (2.13)$$

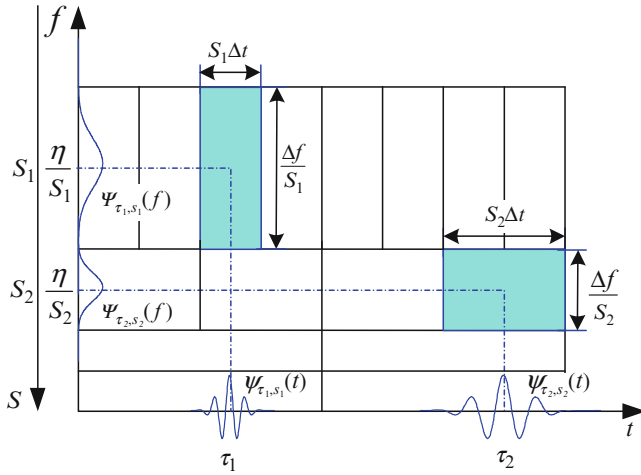


Fig. 2.14 Time and frequency resolutions of the wavelet transform ($s_2 = 2s_1$)

with the parameters f_0 , α , and β all being constants. The corresponding time and frequency resolutions of the Morlet wavelet will be calculated as $\Delta t = s\beta/2\sqrt{\alpha}$ and $\Delta f = \sqrt{\alpha}/(s \cdot 2\pi\beta)$, respectively. These expressions indicate that the time and frequency resolutions are directly and inversely proportional to the scaling parameter s , respectively. In Fig. 2.14, variations of the time and frequency resolutions of the Morlet wavelet at two locations on the time frequency (t f) plane, $(\tau_1, \eta/s_1)$ and $(\tau_2, \eta/s_2)$, are illustrated.

It is seen that changing the scale from s at the location $(\tau_1, \eta/s_1)$ to $s_2 = 2s_1$ at $(\tau_2, \eta/s_2)$ decreases the time resolution by half (as the width of the time window is doubled) while doubling the frequency resolution (because the width of the frequency window is reduced to half). Through variations of the scale s and time shifts (by τ) of the base wavelet function, the wavelet transform is capable of extracting the constituent components within a time series over its entire spectrum, by using small scales (corresponding to higher frequencies) for decomposing high frequency components and large scales (corresponding to lower frequencies) for low frequency components analysis. As an example, Fig. 2.15 illustrates the result of the wavelet transform performed on the signal shown in Fig. 2.1, using the Morlet base wavelet. It is evident that all the transient components are differentiated in the time scale domain.

Following up the impactful work of Morlet and Grossmann, numerous researchers have invested significant effort in further developing the theory of wavelet transform. Examples include Strömberg's early work on discrete wavelets in 1983 (Strömberg 1983), Grossmann, Morlet, and Paul's work on analyzing arbitrary signals in terms of scales and translations of a single base wavelet function (Grossmann et al. 1985, 1986), and Newman's work on Harmonic wavelet transform in 1993 (Newland 1993). Perhaps the most important step that has led to the prosperity of the wavelets was the invention of multiresolution analysis by

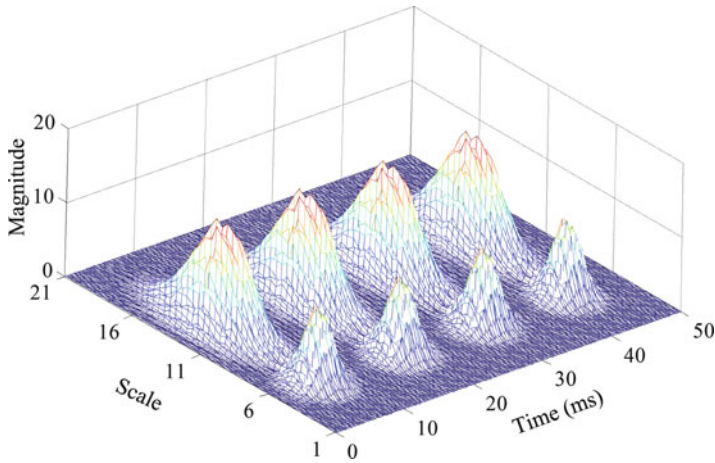


Fig. 2.15 Wavelet transform of the signal

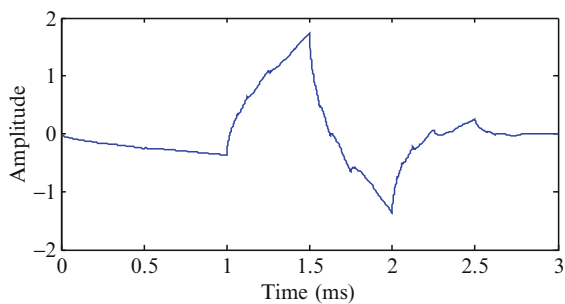
Fig. 2.16 Stephane Mallat



Stephane Mallat (Fig. 2.16) (Mallat 1989a, b, 1999) and Yves Meyer (Fig. 2.17) (Meyer 1989, 1993). Such an invention was introduced by a paper written by Meyer on orthonormal wavelets, entitled “Orthonormal wavelets” (Meyer 1989).

The key to multiresolution analysis is to design the scaling function of the wavelet such that it allowed other researchers to construct their own base wavelets in a mathematically grounded fashion. As an example, Ingrid Daubechies (Fig. 2.18) created her own family of wavelet, the Daubechies wavelets, around 1988 (Daubechies 1988, 1992), on the basis of the concept of multiresolution. Figure 2.19 illustrates one member of the Daubechies wavelet family: Daubechies 2 base wavelet. This type of wavelet is orthogonal and can be implemented using simple digital filtering techniques.

Since then, a proliferation of activities on wavelet transform and its applications in many fields has been seen. These include image processing, speech processing, as well as signal analysis in manufacturing which is the focus of this book.

Fig. 2.17 Yves Meyer**Fig. 2.18** Ingrid Daubechies**Fig. 2.19** Daubechies 2 base wavelet

2.4 References

- Bracewell, R (1999) The Fourier transform and its applications. 3rd edn. McGraw Hill, New York
Chui CK (1992) An introduction to wavelets. Academic, New York
Cohen L (1989) Time frequency distributions a review. Proc IEEE 77(7):941 981

- Cooley JW, Tukey JW (1965) An algorithm for the machine calculation of complex Fourier series. *Math Comput* 19:297–301
- Daubechies I (1988) Orthonormal bases of compactly supported wavelets. *Comm Pure Appl Math* 4:909–996
- Daubechies I (1992) Ten lectures on wavelets. SIAM, Philadelphia, PA
- DeVore RA, Jawerth B, Lucier BJ (1992) Image compression through wavelet transform coding. *IEEE Trans Inf Theory* 38(2):719–746
- Fourier J (1822) The analytical theory of heat. (trans: Freeman A). Cambridge University Press, London, p 1878
- Gabor D (1946) Theory of communication. *J IEEE* 93(3):429–457
- Grossmann A, Morlet J (1984) Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM J Math Anal* 15(4):723–736
- Grossmann A, Morlet J, Paul T (1985) Transforms associated to square integrable group representations. I. General results. *J Math Phys* 26:2473–2479
- Grossmann A, Morlet J, Paul T (1986) Transforms associated to square integrable group representations. II: examples. *Ann Inst Henri Poincaré* 45(3):293–309
- Haar A (1910) Zur theorie der orthogonalen funktionen systeme. *Math Ann* 69:331–371
- Herivel J (1975) Joseph Fourier. The man and the physicist. Clarendon Press, Oxford
- Jaffard S, Yves Meyer Y, Ryan RD (2001) Wavelets: tools for science & technology. Society for Industrial Mathematics, Philadelphia, PA
- Körner TW (1988) Fourier analysis. Cambridge University Press, London
- Littlewood JE, Paley REAC (1931) Theorems on Fourier series and power series. *J Lond Math Soc* 6:230–233
- Mackenzie D (2001) Wavelets: seeing the forest and the trees. National Academy of Sciences, Washington, DC
- Mallat SG (1989a) A theory of multiresolution signal decomposition: the wavelet representation. *IEEE Trans Pattern Anal Mach Intell* 11(7):674–693
- Mallat SG (1989b) Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$. *Trans Am Math Soc* 315:69–87
- Mallat SG (1998) A wavelet tour of signal processing. Academic, San Diego, CA
- Meyer Y (1989) Orthonormal wavelets. In: Combers JM, Grossmann A, Tachamitchian P (eds) Wavelets, time frequency methods and phase space, Springer Verlag, Berlin
- Meyer Y (1993) Wavelets, algorithms and applications. SIAM, Philadelphia, PA
- Newland DE (1993) Harmonic wavelet analysis. *Proc R Soc Lond A Math Phys Sci* 443(1917):203–225
- Oppenheim AV, Schafer RW, Buck JR (1999) Discrete time signal processing. Prentice Hall PTR, Englewood Cliffs, NJ
- Qian S (2002) Time frequency and wavelet transforms. Prentice Hall PTR, Upper Saddle River, NJ
- Ricker N (1953) The form and laws of propagation of seismic wavelets. *Geophysics* 18:10–40
- Rioul O, Vetterli M (1991) Wavelets and signal processing. *IEEE Signal Process Mag* 8(4):14–38
- Strömberg JO (1983) A modified Franklin system and higher order spline systems on \mathbb{R}^n as unconditional bases for Hardy space. Proceedings of Conference on Harmonic Analysis in Honor of Antoni Zygmund, vol 2, pp 475–494
- Jean B. Joseph Fourier, http://mathdl.maa.org/images/upload_library/1/Portraits/Fourier.bmp
- Dennis Gabor, http://nobelprize.org/nobel_prizes/physics/laureates/1971/gabor_autobio.html
- Alfred Haar, <http://www2.isye.gatech.edu/~brani/images/haar.html>
- Paul Levy, http://www.todayinsci.com/L/Levy_Paul/LevyPaulThm.jpg
- Jean Morlet, http://www.industrie-technologies.com/GlobalVisuels/Local/SL_Produit/Morlet.jpg
- Stephane Mallat, <http://www.cmap.polytechnique.fr/~mallat/Stephane.jpg>
- Yves Meyer, http://www.academie-sciences.fr/membres/M/Meyer_Yves.htm
- Ingrid Daubechies, <http://commons.princeton.edu/ciee/images/people/DaubechiesIngrid.jpg>

Chapter 3

Continuous Wavelet Transform

Wavelet transform is a mathematical tool that converts a signal into a different form. This conversion has the goal to reveal the characteristics or “features” hidden within the original signal and represent the original signal more succinctly. A base wavelet is needed in order to realize the wavelet transform. The wavelet is a small wave that has an oscillating wavelike characteristic and has its energy concentrated in time. Figure 3.1 illustrates a wave (sinusoidal) and a wavelet (Daubechies 4 wavelet) (Daubechies 1992).

The difference between a wave and a wavelet is that a wave is usually smooth and regular in shape, and can be everlasting, while in contrast, a wavelet may be irregular in shape, and normally lasts only for a limited period of time. A wave (e.g., sine and cosine) is typically used as a deterministic template in the Fourier transform for representing a signal that is time-invariant or stationary. In comparison, a wavelet can serve as both a deterministic and nondeterministic template for analyzing time-varying or nonstationary signals by decomposing the signal into a 2D, time-frequency domain.

Mathematically, a wavelet is a square integrable function $\psi(t)$ that satisfies the *admissibility condition* (Chui 1992; Meyer 1993; Mallat 1998):

$$\int_{-\infty}^{\infty} \frac{|\Psi(f)|^2}{(f)} df < \infty \quad (3.1)$$

In this equation, $\Psi(f)$ is the Fourier transform (i.e., frequency domain expression) of the wavelet function $\psi(t)$ (in the time domain). The admissibility condition implies that the Fourier transform of the function $\psi(t)$ vanishes at the zero frequency; that is,

$$|\Psi(f)|^2|_{f=0} = 0 \quad (3.2)$$

This means that the wavelet must have a band-pass like spectrum. A zero at the zero frequency also means that the average value of the wavelet $\psi(t)$ in the time domain is zero:

$$\int_{-\infty}^{\infty} \psi(t) dt = 0 \quad (3.3)$$

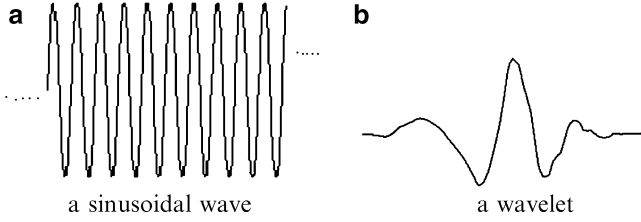


Fig. 3.1 Representation of a wave and a wavelet. (a) A sinusoidal wave and (b) a wavelet

Equation (3.3) indicates that the wavelet must be oscillatory in nature. Through the process of *dilation* (i.e., stretching or squeezing the wavelet function by $1/s$) and *translation*, (i.e., shift along time axis by τ), a family of scaled and translated wavelets can be obtained as

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-\tau}{s}\right), \quad s > 0, \tau \in \mathbb{R} \quad (3.4)$$

The purpose of having the factor $\frac{1}{\sqrt{s}}$ in (3.4) is to ensure that the energy of the wavelet family will remain the same under different scales. For example, by assuming that the energy of the wavelet function $\psi(t)$ is given by

$$\varepsilon = \int_{-\infty}^{\infty} |\psi(t)|^2 dt \quad (3.5)$$

the energy of the scaled and translated wavelets $\psi_{s,\tau}(t)$ can be calculated as

$$\varepsilon' = \int_{-\infty}^{\infty} \left| \frac{1}{\sqrt{s}} \psi\left(\frac{t-\tau}{s}\right) \right|^2 dt = \frac{1}{s} \int_{-\infty}^{\infty} \left| \psi\left(\frac{t}{s}\right) \right|^2 dt = \varepsilon \quad (3.6)$$

As a result, the energy of the original base wavelet $\psi(t)$ and the scaled and translated wavelets remains the same. The relationship between $\psi(t)$ and $\psi_{s,\tau}(t)$ is illustrated in Fig. 3.2, and the process through which a signal is decomposed by analyzing it with a family of scaled and translated wavelets such as $\psi_{s,\tau}(t)$ is called the wavelet transform.

Generally, the wavelet transform can be represented in continuous (i.e., continuous wavelet transform (CWT)) as well as in discrete forms (i.e., discrete wavelet transform). The CWT of a signal $x(t)$ is defined as (Rioul and Vetterli 1991)

$$wt(s, \tau) = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} x(t) \psi^*\left(\frac{t-\tau}{s}\right) dt \quad (3.7)$$

where $\psi^*(\cdot)$ is the complex conjugate of the scaled and shifted wavelet function $\psi(\cdot)$.

As shown in this definition, the CWT is an integral transformation. In this sense, it is similar to the Fourier transform in that integration operation will be performed in

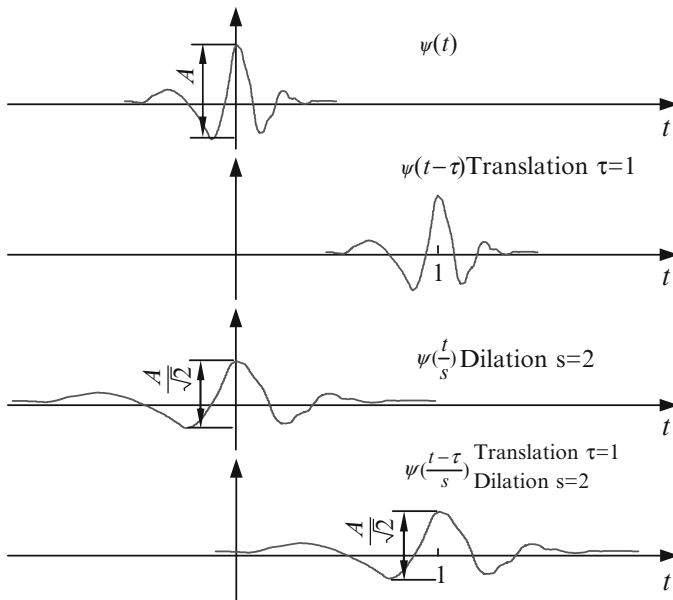


Fig. 3.2 Illustration of translation (by the time constant τ) and dilation (by the scaling factor s)

both transforms. On the other hand, as the wavelet contains two parameters (scale parameter s and translation parameter τ), transforming a signal with the wavelet basis means that such a signal will be projected into a 2D, time-scale plane, instead of the 1D frequency domain in the Fourier transform. Furthermore, because of the localization nature of the wavelet, the transformation will extract features from the signal in the time-scale plane that are not revealed in its original form, for example, what specific bearing defect-related spectral components existed at what time.

3.1 Properties of Continuous Wavelet Transform

Equation (3.7) indicates that the CWT is a linear transformation, characterized by the following properties.

3.1.1 Superposition Property

Suppose $x(t), y(t) \in L^2(R)$, and k_1 and k_2 are constants. If the CWT of $x(t)$ is $w_{t_x}(s, \tau)$ and the CWT of $y(t)$ is $w_{t_y}(s, \tau)$, then the CWT of $z(t) = k_1x(t) + k_2y(t)$ is given by

$$wt_z(s, \tau) = k_1 wt_x(s, \tau) + k_2 wt_y(s, \tau) \quad (3.8)$$

Proof: Let $wt_x(s, \tau) = \frac{1}{\sqrt{s}} \int x(t) \psi^*\left(\frac{t-\tau}{s}\right) dt$ and $wt_y(s, \tau) = \frac{1}{\sqrt{s}} \int y(t) \psi^*\left(\frac{t-\tau}{s}\right) dt$; then

$$\begin{aligned} wt_z(s, \tau) &= \frac{1}{\sqrt{s}} \int z(t) \psi^*\left(\frac{t-\tau}{s}\right) dt \\ &= \frac{1}{\sqrt{s}} \int [k_1 x(t) + k_2 y(t)] \psi^*\left(\frac{t-\tau}{s}\right) dt \\ &= k_1 \frac{1}{\sqrt{s}} \int x(t) \psi^*\left(\frac{t-\tau}{s}\right) dt + k_2 \frac{1}{\sqrt{s}} \int y(t) \psi^*\left(\frac{t-\tau}{s}\right) dt \\ &= k_1 wt_x(s, \tau) + k_2 wt_y(s, \tau) \end{aligned} \quad (3.9)$$

This proves the superposition property of the CWT.

3.1.2 Covariant Under Translation

Suppose the CWT of $x(t)$ is $wt_x(s, \tau)$; then the CWT of $x(t - t_0)$ is $wt_x(s, \tau - t_0)$. The proof of this property is shown below:

Proof: Let $x'(t) = x(t - t_0)$; then

$$wt_{x'}(s, \tau) = \frac{1}{\sqrt{s}} \int x(t - t_0) \psi^*\left(\frac{t-\tau}{s}\right) dt \quad (3.10)$$

Let $t' = t - t_0$; then

$$wt_{x'}(s, \tau) = \frac{1}{\sqrt{s}} \int x(t') \psi^*\left(\frac{t' + t_0 - \tau}{s}\right) dt' = wt_x(s, \tau - t_0) \quad (3.11)$$

This means that the wavelet coefficients of $x(t - t_0)$ can be obtained by translating the wavelet coefficients of $x(t)$ along the time axis with t_0 .

3.1.3 Covariant Under Dilation

Suppose the CWT of $x(t)$ is $wt_x(s, \tau)$; then the CWT of $x\left(\frac{t}{a}\right)$ is $\sqrt{a} wt_x\left(\frac{s}{a}, \frac{\tau}{a}\right)$

Proof: Let $x'(t) = x\left(\frac{t}{a}\right)$; then

$$wt_{x'}(s, \tau) = \frac{1}{\sqrt{s}} \int x'(t) \psi^*\left(\frac{t-\tau}{s}\right) dt = \frac{1}{\sqrt{s}} \int x\left(\frac{t}{a}\right) \psi^*\left(\frac{t-\tau}{s}\right) dt \quad (3.12)$$

Let $t' = \frac{t}{a}$; then (3.12) becomes

$$\begin{aligned} wt_{x'}(s, \tau) &= \frac{1}{\sqrt{s}} \int x(t') \psi^* \left(\frac{at' - \tau}{s} \right) d(at') \\ &= \frac{\sqrt{a}}{\sqrt{\frac{s}{a}}} \int x(t') \psi^* \left(\frac{t' - \frac{\tau}{a}}{\frac{s}{a}} \right) dt' = \sqrt{a} wt_x \left(\frac{s}{a}, \frac{\tau}{a} \right) \end{aligned} \quad (3.13)$$

Equation (3.13) indicates that, when a signal is dilated by a , its corresponding wavelet coefficients are also dilated by a along both the scale and time axes.

3.1.4 Moyal Principle

Suppose $x(t), y(t) \in L^2(R)$. If the CWT of $x(t)$ is $wt_x(s, \tau)$ and the CWT of $y(t)$ is $wt_y(s, \tau)$; that is,

$$wt_x(s, \tau) = \langle x(t), \psi_{s,\tau}(t) \rangle \quad (3.14a)$$

$$wt_y(s, \tau) = \langle y(t), \psi_{s,\tau}(t) \rangle \quad (3.14b)$$

then

$$\langle wt_x(s, \tau), wt_y(s, \tau) \rangle = C_\psi \langle x(t), y(t) \rangle \quad (3.15)$$

where $C_\psi = \int_0^\infty \frac{|\Psi(f)|^2}{f} df$. The proof of this property is as follows.

Proof According to the Parseval's theorem, the inner product of two functions in time domain can be equivalently given in the frequency domain as

$$\langle x(t), y(t) \rangle = \frac{1}{2\pi} \int X(f) Y^*(f) df \quad (3.16)$$

Consequently, we have

$$wt_x(s, \tau) = \langle x(t), \psi_{s,\tau}(t) \rangle = \frac{1}{2\pi} \int X(f) \Psi_{s,\tau}^*(f) df \quad (3.17a)$$

$$wt_y(s, \tau) = \langle y(t), \psi_{s,\tau}(t) \rangle = \frac{1}{2\pi} \int Y(f) \Psi_{s,\tau}^*(f) df \quad (3.17b)$$

From (3.4), we know that $\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}} \psi \left(\frac{t-\tau}{s} \right)$. Therefore,

$$\Psi_{s,\tau}(f) = \sqrt{s} \Psi(sf) e^{j2\pi f \tau} \quad (3.18a)$$

$$\Psi_{s,\tau}^*(f) = \sqrt{s}\Psi^*(sf)e^{j2\pi f\tau} \quad (3.18b)$$

By incorporating (3.18b) into (3.17) and utilizing the following integral relation,

$$\int e^{j2\pi(f-f')\tau} d\tau = 2\pi\delta(f-f') \quad (3.19)$$

the left side of (3.15) can be expanded as

$$\begin{aligned} \langle wt_x(s, \tau), wt_y(s, \tau) \rangle &= \frac{s}{2\pi} \iint \frac{ds}{s^2} X(f)Y^*(f)\Psi(sf)\Psi^*(sf)df \\ &= \frac{1}{2\pi} \int \left[\int \frac{|\Psi(sf)|^2}{s} ds \right] X(f)Y^*(f)df \end{aligned} \quad (3.20)$$

As $\int \frac{|\Psi(sf)|^2}{s} ds = \int \frac{|\Psi(sf)|^2}{sf} d(sf) = \int \frac{|\Psi(f')|^2}{f'} d(f') = C_\psi$, (3.20) can be expressed as

$$\langle wt_x(s, \tau), wt_y(s, \tau) \rangle = C_\psi \frac{1}{2\pi} \int X(f)Y^*(f)df = C_\psi \langle x(t), y(t) \rangle \quad (3.21)$$

This proves the existence of Moyal principle for CWT. It is noted that C_ψ is actually the *admissibility condition* of the wavelet. Only when this condition is satisfied can the Moyal principle exist. Furthermore, if $x(t) = y(t)$, then (3.15) becomes

$$\int_0^\infty \frac{ds}{s^2} \int_{-\infty}^\infty |wt_x(s, \tau)|^2 d\tau = C_\psi \int_{-\infty}^\infty |x(t)|^2 dt \quad (3.22)$$

This means that the integral of the square of wavelet coefficients is proportional to the energy of the signal.

3.2 Inverse Continuous Wavelet Transform

A transformation is considered to be meaningful in practice only when its corresponding inverse transformation exists. The same principle applies to the CWT. It can be shown that, as long as the wavelet satisfies the admission condition as defined in (3.1), the inverse CWT will exist. This means that a signal can be perfectly reconstructed from its corresponding wavelet coefficients, which can be written as

$$\begin{aligned}
x(t) &= \frac{1}{C_\psi} \int_0^\infty \frac{ds}{s^2} \int_{-\infty}^\infty w_{t_x}(s, \tau) \psi_{s, \tau}(t) d\tau \\
&= \frac{1}{C_\psi} \int_0^\infty \frac{ds}{s^2} \int_{-\infty}^\infty w_{t_x}(s, \tau) \frac{1}{\sqrt{s}} \psi\left(\frac{t - \tau}{s}\right) d\tau
\end{aligned} \tag{3.23}$$

where $C_\psi = \int_0^\infty \frac{|\Psi(f)|^2}{f} df < \infty$ is the admissibility condition of the wavelet $\psi(t)$.

The proof of (3.23) is shown below:

Proof Assume that $x_1(t) = x(t)$, and $x_2(t) = \delta(t - t')$. As $\langle x(t), \delta(t - t') \rangle = x(t')$,

$$C_\psi x(t') = C_\psi \langle x(t), \delta(t - t') \rangle \tag{3.24}$$

According to the Mayol principle shown in (3.15), (3.24) can be further written as

$$\begin{aligned}
C_\psi x(t') &= \langle w_{t_x}(s, \tau), w_{\delta(t - t')}(s, \tau) \rangle \\
&= \int_0^\infty \frac{ds}{s^2} \int_{-\infty}^\infty w_{t_x}(s, \tau) w_{\delta(t - t')}^*(s, \tau) d\tau \\
&= \int_0^\infty \frac{ds}{s^2} \int_{-\infty}^\infty w_{t_x}(s, \tau) \langle \psi_{s, \tau}(t), \delta(t - t') \rangle^* d\tau \\
&= \int_0^\infty \frac{ds}{s^2} \int_{-\infty}^\infty w_{t_x}(s, \tau) \langle \psi_{s, \tau}(t), \delta(t - t') \rangle d\tau \\
&= \int_0^\infty \frac{ds}{s^2} \int_{-\infty}^\infty w_{t_x}(s, \tau) \psi_{s, \tau}(t') d\tau \\
&= \frac{1}{\sqrt{s}} \int_0^\infty \frac{ds}{s^2} \int_{-\infty}^\infty w_{t_x}(s, \tau) \psi\left(\frac{t' - \tau}{s}\right) d\tau
\end{aligned} \tag{3.25}$$

This illustrates that the inverse CWT exists.

3.3 Implementation of Continuous Wavelet Transform

To implement the CWT, two approaches can be taken. The first approach is to obtain the wavelet coefficients directly from (3.7). The computation procedure is as follows:

1. The wavelet is placed at the beginning of the signal, and set $s = 1$ (the original, base wavelet).
2. The wavelet function at scale “1” is multiplied by the signal, integrated over all times, and then multiplied by $1/\sqrt{s}$.
3. Shift the wavelet to $t = \tau$, and get the transform value at $t = \tau$ and $s = 1$.
4. Repeat the procedure until the wavelet reaches the end of the signal.
5. Scale s is increased by a given value, and the above procedure is repeated for all s .
6. Each computation for a given s fills the single row of the time-scale plane.
7. Wavelet transform is obtained if all s are calculated.

The second approach to implementing the CWT is on the basis of the convolution theorem, which states that the Fourier transform of the convolution operation on two functions in the time domain is the product of the respective Fourier transforms of these two functions in the frequency domain (Bracewell 1999). The Fourier transform of (3.7) is expressed as

$$WT(s, f) = F\{wt(s, \tau)\} = \frac{1}{2\pi\sqrt{s}} \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x(t) \psi^*\left(\frac{t-\tau}{s}\right) dt \right) e^{j2\pi f\tau} d\tau \quad (3.26)$$

Applying the convolution theorem to (3.26) leads to

$$WT(s, f) = \sqrt{s} X(f) \Psi^*(sf) \quad (3.27)$$

where $X(f)$ denotes the Fourier transform of $x(t)$ and $\Psi^*(\cdot)$ denotes the Fourier transform of $\psi^*(\cdot)$. By taking the inverse Fourier transform, (3.27) is converted back into the time domain as

$$wt(s, t) = F^{-1}\{WT(s, f)\} = \sqrt{s} F^{-1}\{X(f) \Psi^*(sf)\} \quad (3.28)$$

where the symbol $F^{-1}[\cdot]$ denotes the operator of inverse Fourier transform. Therefore, the implementation of the CWT can be realized through a pair of Fourier and inverse Fourier transforms.

Figure 3.3 illustrates the procedure for implementing the CWT. After taking the Fourier transform of the signal $x(t)$ and the scaled base wavelet $\psi(s, t)$ to obtain their frequency information $X(f)$ and $\Psi(sf)$, respectively, the inner product between $X(f)$ and complex conjugate of $\Psi(sf)$ is calculated. Next, the CWT of the signal $x(t)$, denoted as $cwt(s, t)$, is obtained by taking the inverse Fourier transform on the inner product of $WT(s, f)$.

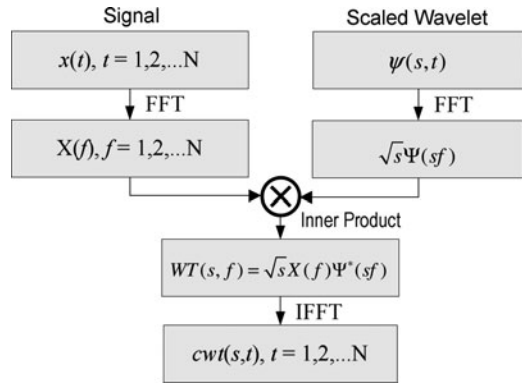


Fig. 3.3 Procedure for implementing the continuous wavelet transform

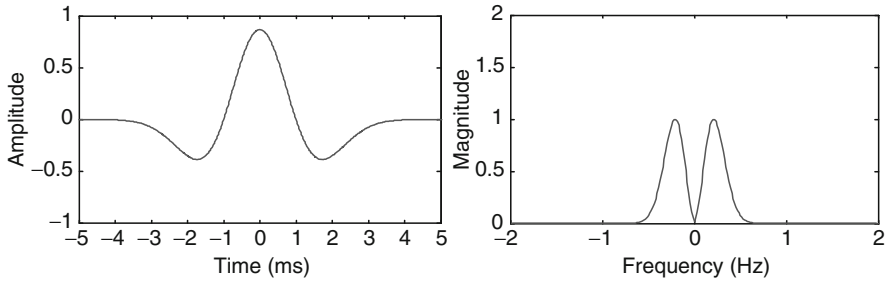


Fig. 3.4 Mexican hat wavelet (*left*) and its magnitude spectrum (*right*)

3.4 Some Commonly Used Wavelets

This section introduces several commonly used wavelets for performing the CWT.

3.4.1 Mexican Hat Wavelets

The Mexican hat wavelet is a normalized, second derivative of a Gaussian function, which is mathematically defined as (Mallat 1998)

$$\psi(t) = \frac{1}{\sqrt{2\pi}\sigma^3} \left(1 - \frac{\sigma^2}{t^2} \right) e^{-\frac{t^2}{2\sigma^2}} \quad (3.29)$$

Figure 3.4 illustrates the Mexican hat wavelet and its associated magnitude spectrum. The Mexican hat wavelet is often called the Ricker wavelet in geophysics, where it is frequently employed to model seismic data (Zhou and Adeli 2003; Erlebacher and Yuen 2004).

3.4.2 Morlet Wavelet

The Morlet wavelet is defined as (Grossmann and Morlet 1984; Teolis 1998)

$$\psi_M(t) = \frac{1}{\sqrt{\pi f_b}} e^{i2\pi f_c t} e^{-\frac{t^2}{f_b}} \quad (3.30)$$

where f_b is the bandwidth parameter and f_c denotes the wavelet center frequency. As an example, Fig. 3.5 illustrates the complex Morlet wavelet function and its corresponding magnitude spectrum when $f_b = 1$ Hz and $f_c = 1$ Hz. The Morlet wavelet has been widely used for identifying transient components embedded in a

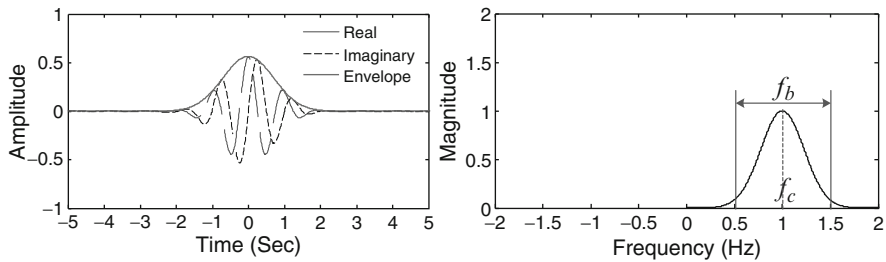


Fig. 3.5 Morlet wavelet (*left*) and its magnitude spectrum (*right*): $f_b = 1$ Hz and $f_c = 1$ Hz

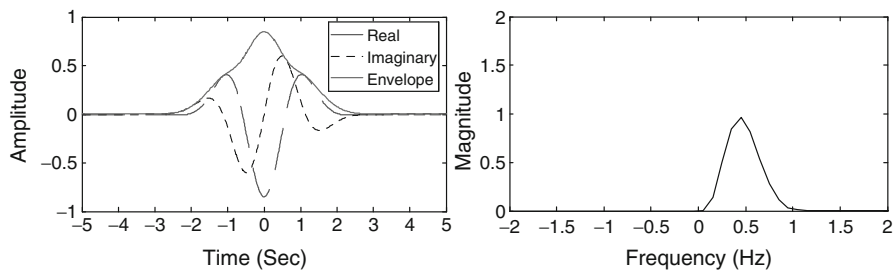


Fig. 3.6 Gaussian wavelet (*left*) and its magnitude spectrum (*right*): $N = 2$

signal, for example, bearing defect-induced vibration (Lin and Qu 2000; Nikolaou and Antoniadis 2002; Yan and Gao 2009).

3.4.3 Gaussian Wavelet

Mathematically, a Gaussian function is expressed as (Teolis 1998)

$$f(t) = e^{-j\omega t} e^{-t^2} \quad (3.31)$$

Taking the N th derivative of this function yields the Gaussian wavelet as

$$\psi_G(t) = c_N \frac{d^{(N)}f(t)}{dt^N}, \quad (3.32)$$

where N is an integer parameter (≥ 1) and denotes the order of the wavelet, and c_N is a constant introduced to ensure that $\|f^{(N)}(t)\|^2 = 1$. Figure 3.6 illustrates the Gaussian function with its magnitude spectrum for the case of $N = 2$. The Gaussian wavelet is often used for characterizing singularity that exists in a signal (Mallat and Hwang 1992; Sun and Tang 2002).

3.4.4 Frequency B-Spline Wavelet

A frequency B-spline wavelet is defined as (Teolis 1998)

$$\psi_B(t) = \sqrt{f_b} \left[\sin c \left(\frac{f_b t}{p} \right) \right]^p e^{j2\pi f_c t} \quad (3.33)$$

where f_b is the bandwidth parameter, f_c denotes the wavelet center frequency, and p is an integer parameter (≥ 2). The notation of $\sin c(\cdot)$ is a $\sin c$ function, which is defined as

$$\sin c(x) = \begin{cases} 1 & x = 0 \\ \frac{\sin x}{x} & \text{otherwise} \end{cases} \quad (3.34)$$

As an example, a B-spline wavelet for the case of $f_b = 1$ Hz, $f_c = 1$ Hz, and $p = 2$ together with its corresponding magnitude function is shown in Fig. 3.7. The application of the frequency B-spline wavelet has been seen in biomedical signal analysis (Moga et al. 2005; Fard et al. 2007).

3.4.5 Shannon Wavelet

The Shannon wavelet is a special case of the frequency B-spline wavelet for $p = 1$:

$$\psi_S(t) = \sqrt{f_b} \sin c(f_b t) e^{j2\pi f_c t} \quad (3.35)$$

where f_b is the bandwidth parameter and f_c denotes the wavelet center frequency. The notation of $\sin c(\cdot)$ is a $\sin c$ function and defined in (3.34). Figure 3.8 illustrates the Shannon wavelet for the case of $f_b = 1$ Hz and $f_c = 1$ Hz, with its corresponding magnitude spectrum. The Shannon wavelet has been shown for the analysis and synthesis of the $1/f$ processes (Shusterman and Feder 1998).

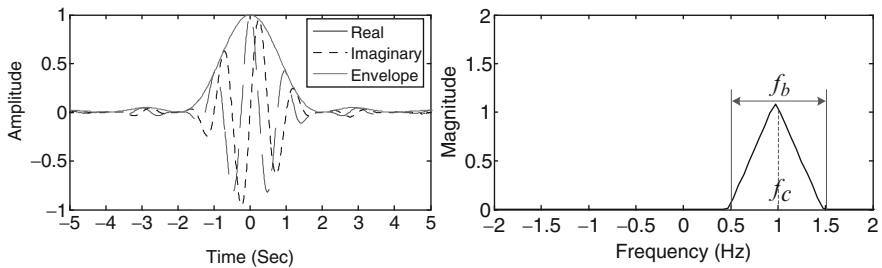


Fig. 3.7 Frequency B spline wavelet (left) and its corresponding magnitude spectrum (right): $p = 2$, $f_b = 1$ Hz, and $f_c = 1$ Hz

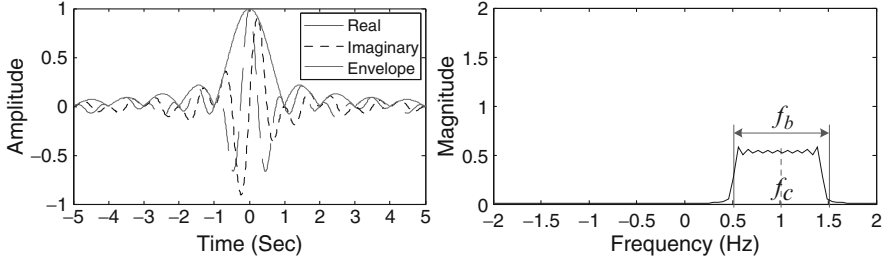


Fig. 3.8 Shannon wavelet (*left*) and its corresponding magnitude spectrum (*right*): $f_b = 1$ Hz and $f_c = 1$ Hz

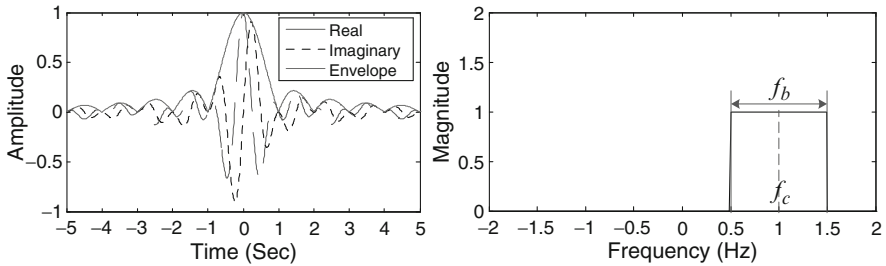


Fig. 3.9 Harmonic wavelet (*left*) and its magnitude spectrum (*right*): $m = 0.5$ Hz and $n = 1.5$ Hz

3.4.6 Harmonic Wavelet

The harmonic wavelet is defined in the frequency domain as (Newland 1994a, b; Yan and Gao 2005)

$$\Psi_{m,n}(f) = \begin{cases} \frac{1}{n-m} & m \leq f \leq n \\ 0 & \text{elsewhere} \end{cases} \quad (3.36)$$

where the symbols m and n are the scale parameters. These parameters are real but not necessarily integers. Furthermore, the bandwidth f_b and center frequency f_c are determined by the scale parameter as

$$f_b = n - m; \quad f_c = \frac{n + m}{2} \quad (3.37)$$

As an example, Fig. 3.9 shows the harmonic wavelet function and its corresponding magnitude spectrum for the case of $m = 0.5$ and $n = 1.5$. The harmonic wavelet was first designed by Newland for analyzing vibration signals (Newland 1993). Later, the application of harmonic wavelet has been extended to heart rate variability analysis (Bates et al. 1997) and image denoising (Iftekhharuddin 2002).

3.5 CWT of Representative Signals

Using the wavelets introduced in Sect. 3.4, the CWT is applied to several typical signals, as described below.

3.5.1 CWT of Sinusoidal Function

The first signal analyzed is a pure sinusoidal function. Figure 3.10a illustrates a 50 Hz sinusoidal signal, and Fig. 3.10b illustrates the CWT results of the signal. It is seen that the 50 Hz component is present all the time throughout the analysis duration.

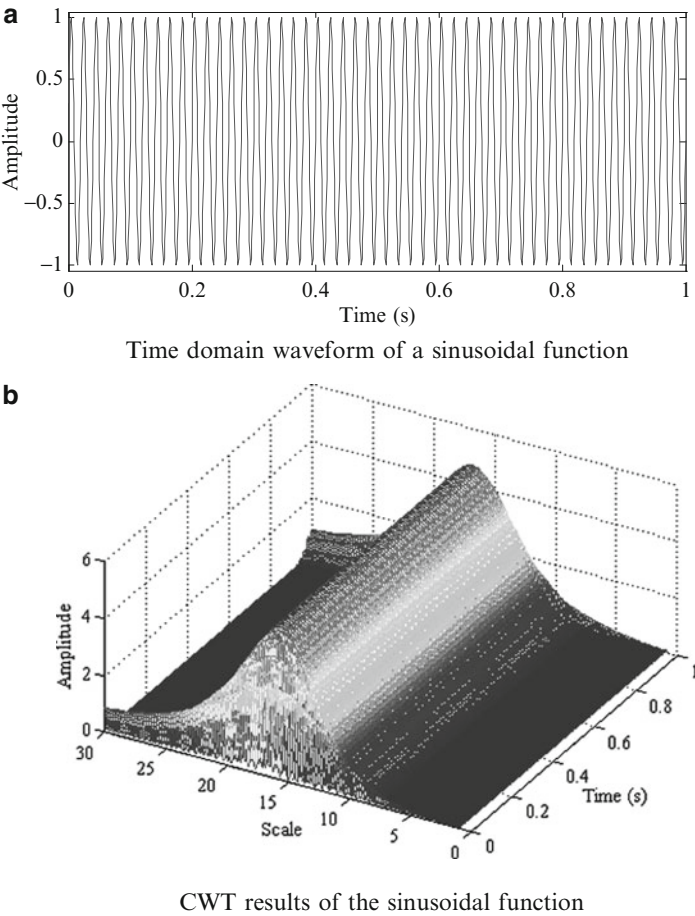


Fig. 3.10 A sinusoidal function (a) time domain waveform (b) CWT results

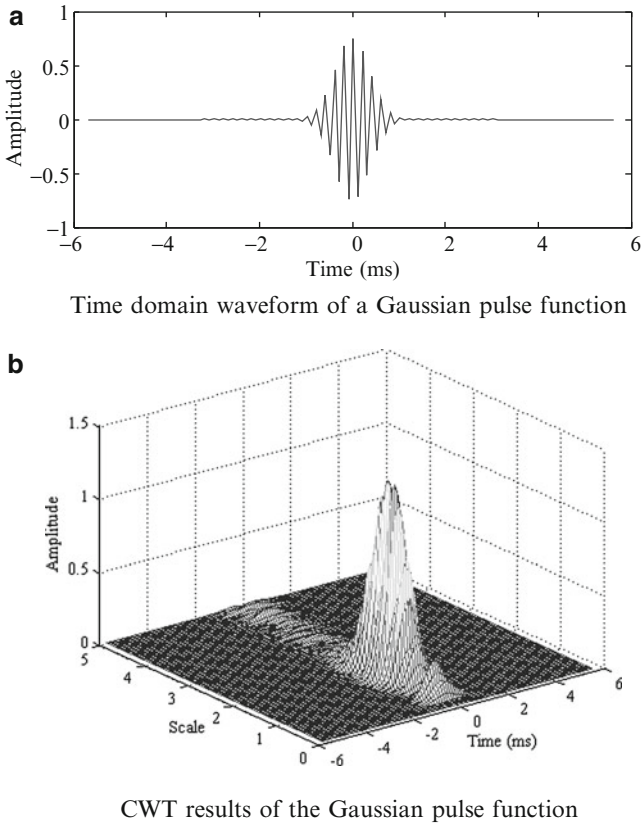


Fig. 3.11 A Gaussian pulse function (a) time domain waveform (b) CWT results

3.5.2 CWT of Gaussian Pulse Function

The second signal is a Gaussian pulse function. Figure 3.11a shows a Gaussian pulse signal with 10 kHz center frequency. Figure 3.11b illustrates the CWT results of the Gaussian pulse signal, which is identified in the time-scale domain at around 0 s.

3.5.3 CWT of Chirp Function

The last signal is a chirp function. Figure 3.12a shows an example of a chirp signal. It is a linear swept-frequency signal with the instantaneous frequency being 50 Hz at time zero. The instantaneous frequency 10 Hz is achieved after 1 s. Figure 3.12b illustrates the CWT results of the chirp signal, and the change of frequency along with time can be clearly seen.

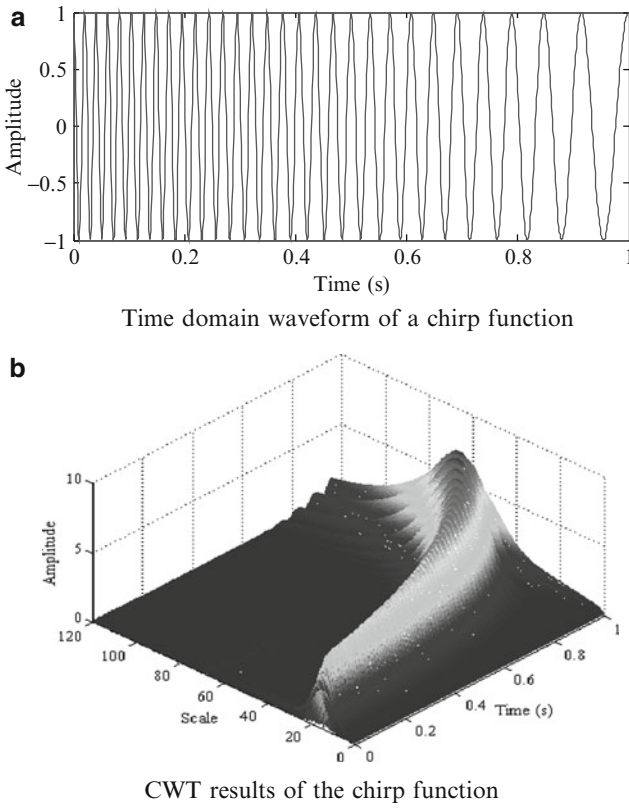


Fig. 3.12 A chirp function (a) time domain waveform (b) CWT results

3.6 Summary

This chapter begins with definition of a wavelet, where the *admissibility condition* that a wavelet should satisfy is emphasized. The CWT and its related properties are then introduced. Two approaches for implementing the CWT are discussed in Sect. 3.3, followed by the introduction of some commonly used wavelets in Sect. 3.4. Typical signals are analyzed using the CWT and the results are shown in Sect. 3.5.

3.7 References

Bates RA, Hilton MF, Godfrey KR, Chappell MJ (1997) Autonomic function assessment using analysis of heart rate variability. *Control Eng Pract* 5(12):1731–1737
Bracewell R (1999) *The Fourier transform and its applications*. 3rd edn. McGraw Hill, New York
Chui CK (1992) *An introduction to wavelets*. Academic, New York

- Daubechies I (1992) Ten lectures on wavelets. SIAM, Philadelphia, PA
- Erlebacher G, Yuen DA (2004) A wavelet toolkit for visualization and analysis of large data sets in earthquake research. *Pure Appl Geophys* 161(11–12):2215–2229
- Fard PJ, Moradi MH, Divide MR (2007) A novel approach in R peak detection using hybrid complex wavelet (HCW). *Int J Cardiol* 124(2):250–253
- Grossmann A, Morlet J (1984) Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM J Math Anal* 15(4):723–736
- Iftekharuddin KM (2002) Harmonic wavelet joint transform correlator: analysis, algorithm, and application to image denoising. *Opt Eng* 41(12):3307–3315
- Lin J, Qu L (2000) Feature extraction based on Morlet wavelet and its application for mechanical fault diagnosis. *J Sound Vib* 234(1):135–148
- Mallat SG (1998) A wavelet tour of signal processing. Academic, San Diego, CA
- Mallat SG, Hwang WL (1992) Singularity detection and processing with wavelets. *IEEE Trans Inf Theory* 38:617–643
- Meyer Y (1993) Wavelets, algorithms and applications. SIAM, Philadelphia, PA
- Moga M, Moga VD, Mihalas GhI (2005) Continuous wavelet transform in ECG analysis: a concept or clinical uses. *Connecting medical informatics and bio informatics*, pp 1143–1148, IOS Press
- Newland DE (1993) Random vibrations, spectral and wavelet analysis. Wiley, New York
- Newland DE (1994a) Wavelet analysis of vibration part I: theory, *ASME J Vib Acoust* 116(4):409–416
- Newland DE (1994b) Wavelet analysis of vibration part II: wavelet maps, *ASME J Vib Acoust* 116(4): 417–425
- Nikolaou NG, Antoniadis IA (2002) Demodulation of vibration signals generated by defects in rolling element bearings using complex shifted Morlet wavelets. *Mech Syst Signal Process* 16(4):677–694
- Rioul O, Vetterli M (1991) Wavelets and signal processing. *IEEE Signal Process Mag* 8(4):14–38
- Shusterman E, Feder M. (1998) *Analysis and synthesis of 1/f processes via Shannon wavelets*. *IEEE Trans Signal Process* 46(6):1698–1702
- Sun Q, Tang Y (2002) Singularity analysis using continuous wavelet transform for bearing fault diagnosis. *Mech Syst Signal Process* 16:1025–1041
- Teolis A. (1998) Computational signal processing with wavelets. Birkhäuser Boston, MA
- Yan R, Gao RX (2005) An efficient approach to machine health diagnosis based on harmonic wavelet packet transform. *Robot Comput Integr Manuf* 21:291–301
- Yan R, Gao R (2009) Multi scale enveloping spectrogram for vibration analysis in bearing defect diagnosis. *Tribol Int* 42(2): 293–302
- Zhou Z, Adeli H, (2003) Time frequency signal analysis of earthquake records using Mexican hat wavelets. *Comput Aided Civ Infrastruct Eng* 18(5):379–389

Chapter 4

Discrete Wavelet Transform

According to the definition of the continuous wavelet transform (CWT) given in (3.7), Chap. 3, the scale parameter s and translation parameter τ can be varied continuously. As a result, performing the CWT on a signal will lead to the generation of redundant information. Although the redundancy is useful in some applications, such as signal denoising and feature extraction where desired performance is achieved at the cost of increased computational time and memory size, other applications may need to emphasize reduced computational time and data size, for example, in image compression and numerical computation. Such requirements illustrate the need for reducing redundancy in the wavelet coefficients among different scales as much as possible, while at the same time, avoiding sacrificing the information contained in the original signal. This can be achieved by parameter discretization, as described in the following section.

4.1 Discretization of Scale and Translation Parameters

The approach to reducing redundancy is to use discrete values of scale and translation parameters. A natural way to implement this is to use a logarithmic discretization of the scale s and then link it to step size taken between the values of translation parameter τ . This type of discretization is expressed as

$$\begin{cases} s = s_0^j \\ \tau = k\tau_0 s_0^j \end{cases} \quad s_0 < 1, \tau_0 \neq 0, j \in \mathbb{Z}, k \in \mathbb{Z} \quad (4.1)$$

where the symbol \mathbb{Z} denotes an integer. The corresponding family of the base wavelet is then expressed as

$$\psi_{j,k}(t) = \frac{1}{\sqrt{s_0^j}} \psi\left(\frac{t - k\tau_0 s_0^j}{s_0^j}\right) \quad (4.2)$$

Generally, the values of $s_0 = 2$ and $\tau_0 = 1$ are adopted (Addison 2002). Consequently, (4.2) is expressed as

$$\psi_{j,k}(t) = \frac{1}{\sqrt{2^j}} \psi\left(\frac{t - k2^j}{2^j}\right) \quad (4.3)$$

As a result, the wavelet transform of a given signal $x(t)$ is obtained as

$$wt(j, k) = \langle x(t), \psi_{j,k}(t) \rangle = \frac{1}{\sqrt{2^j}} \int_{-\infty}^{\infty} x(t) \psi^*\left(\frac{t - k2^j}{2^j}\right) dt \quad (4.4)$$

where the symbol $\langle \cdot \rangle$ denotes inner product operation. Equation (4.4) poses the following two questions:

1. Can the results of the discretized wavelet transform represent the entire information content of the signal $x(t)$? In other words, can the wavelet coefficients obtained as a result of the wavelet transform be used to *perfectly* reconstruct the original signal $x(t)$?
2. Can any signal $x(t)$ be expressed as the summation of $\psi_{j,k}(t)$, in the form of the following equation?

$$x(t) = \sum_{j,k} C_{j,k} \psi_{j,k}(t) \quad (4.5)$$

In (4.5), $C_{j,k}$ represents the coefficient of the discrete wavelet transform (DWT), which corresponds to $wt(j, k)$ in (4.4). Finally, if the answer to question (2) is “yes,” then how can we calculate the coefficient $C_{j,k}$?

Assume that the answer to question (1) is “yes,” and we can select $\psi_{s,\tau}(t)$ and discretize s and τ properly. Then, there must exist a function $\tilde{\psi}_{j,k}(t)$, defined as the dual function of $\psi_{j,k}(t)$, that can be used for reconstructing the signal $x(t)$ described in question (1), as follows:

$$x(t) = \sum_{j,k} \langle x(t), \psi_{j,k}(t) \rangle \tilde{\psi}_{j,k}(t) \quad (4.6)$$

where the term $\tilde{\psi}_{j,k}(t)$ can be obtained by performing the scaling and translation operations on $\tilde{\psi}(t)$ as

$$\tilde{\psi}_{j,k}(t) = \frac{1}{\sqrt{2^j}} \tilde{\psi}\left(\frac{t - k2^j}{2^j}\right) \quad (4.7)$$

On the basis of the above assumption, if there exists another signal $y(t)$, we can obtain the inner product of the signals $x(t)$ and $y(t)$ as shown in (4.8). Note that the symbol $*$ indicates the complex conjugate operator:

$$\begin{aligned}
\langle y(t), x(t) \rangle &= \langle x(t), y(t) \rangle^* = \left\langle \sum_{j,k} \langle x(t), \psi_{j,k}(t) \rangle \tilde{\psi}_{j,k}(t), y(t) \right\rangle^* \\
&= \left(\sum_{j,k} \langle x(t), \psi_{j,k}(t) \rangle \langle \tilde{\psi}_{j,k}(t), y(t) \rangle \right)^* \\
&= \sum_{j,k} \langle y(t), \tilde{\psi}_{j,k}(t) \rangle \langle \psi_{j,k}(t), x(t) \rangle \\
&= \left\langle \sum_{j,k} \langle y(t), \tilde{\psi}_{j,k}(t) \rangle \psi_{j,k}(t), x(t) \right\rangle
\end{aligned} \tag{4.8}$$

Equation (4.8) implies that

$$y(t) = \sum_{j,k} \langle y(t), \tilde{\psi}_{j,k}(t) \rangle \psi_{j,k}(t) \tag{4.9}$$

which means that the answer to question (2) is also positive. It further implies that the coefficient $C_{j,k}$ can be calculated as

$$C_{j,k} = \langle y(t), \tilde{\psi}_{j,k}(t) \rangle \tag{4.10}$$

Therefore, once question (1) is answered, the answer to question (2) can be readily derived from it. The answer to the question (1) can be presented in mathematical terms as follows.

If a set of wavelet coefficients $\langle x(t), \psi_{j,k}(t) \rangle$ exists that describes complete information of the signal $x(t)$, then the following statements must hold:

1. When $x_1(t) = x_2(t)$, the inner product of $x_1(t)$ and the scaled and translated wavelet $\psi_{j,k}(t)$ can be expressed as

$$\langle x_1(t), \psi_{j,k}(t) \rangle = \langle x_2(t), \psi_{j,k}(t) \rangle \tag{4.11}$$

2. For $x(t) = 0$, we have

$$\langle x(t), \psi_{j,k}(t) \rangle = 0 \tag{4.12}$$

3. When $x_1(t)$ is very close to $x_2(t)$, the corresponding wavelet coefficients $\langle x_1(t), \psi_{j,k}(t) \rangle$ must be close to $\langle x_2(t), \psi_{j,k}(t) \rangle$. In other words, if $\|x_1(t) - x_2(t)\|$ is very small, then $\sum_{j,k} |\langle x_1(t), \psi_{j,k}(t) \rangle - \langle x_2(t), \psi_{j,k}(t) \rangle|^2$ must be very small, too.

Mathematically, this can be expressed as

$$\sum_{j,k} |\langle x_1(t), \psi_{j,k}(t) \rangle - \langle x_2(t), \psi_{j,k}(t) \rangle|^2 \leq B \|x_1(t) - x_2(t)\|^2, \quad B \in \mathbb{R}^+ \tag{4.13}$$

that is,

$$\sum_{j,k} |\langle x(t), \psi_{j,k}(t) \rangle|^2 \leq B \|x(t)\|^2 \quad (4.14)$$

In (4.13), the symbol R^+ denotes the set of positive real numbers, and B is a positive real number.

Furthermore, if we want to reconstruct $x(t)$ from the wavelet coefficient $\langle x(t), \psi_{j,k}(t) \rangle$, the following condition must hold:

When $\langle x_1(t), \psi_{j,k}(t) \rangle$ is very close to $\langle x_2(t), \psi_{j,k}(t) \rangle$, $x_1(t)$ must be very close to $x_2(t)$, too, which leads to

$$A \|x(t)\|^2 \leq \sum_{j,k} |\langle x(t), \psi_{j,k}(t) \rangle|^2, \quad A \in R^+ \quad (4.15)$$

where A is a positive real number.

Combining (4.15) with (4.14), we obtain the following equation:

$$A \|x(t)\|^2 \leq \sum_{j,k} |\langle x(t), \psi_{j,k}(t) \rangle|^2 \leq B \|x(t)\|^2, \quad A, B \in R^+ \quad (4.16)$$

This ensures that the DWT of a signal $x(t)$ can be obtained. Equation (4.16) is called a *wavelet frame* (Addison 2002). The values of the wavelet frame bounds, A and B , depending on both the scale parameter s and the translation parameter τ that are chosen for analysis and the base wavelet function used (Daubechies 1992). Particularly, if $A = B$, the wavelet frame is known as a *tight* frame. In such a case, the signal $x(t)$ can be reconstructed through the inverse discretized wavelet transform as

$$x(t) = \frac{1}{A} \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} wt(j, k) \psi_{j,k}(t) \quad (4.17)$$

If $A \neq B$, but the difference between A and B is not too large (Addison 2002), the signal $x(t)$ can still be reconstructed as

$$x'(t) = \frac{2}{A+B} \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} wt(j, k) \psi_{j,k}(t) \quad (4.18)$$

The difference between $x(t)$ and $x'(t)$ is determined by the values of A and B , and becomes small in practice when the ratio of B/A is approaching the value of one.

4.2 Multiresolution Analysis and Orthogonal Wavelet Transform

Of the various forms of wavelet discretization, the *dyadic* discretization with $s_0 = 2$ and $\tau_0 = 1$ has been widely used, as shown in (4.3). This is because it allows the selection of the base wavelet to be made in such a way that its corresponding family set $\psi_{j,k}(t)$ constitutes an *orthogonal* basis within the tight wavelet frame, characterized by $A = 1$. To construct a base wavelet having the characteristics of orthogonality, the multiresolution analysis (MRA) is presented here as the theoretical foundation.

4.2.1 Multiresolution Analysis

The concept of MRA was formed when Mallat was working on image processing in the 1980s (Mallat 1989a, b). At that time, the idea of studying images simultaneously at different scales had been popular for years already (Witkin 1983; Burt and Adelson 1983). This provided the background for using orthogonal wavelet bases as a tool to describe the information contained in the image, from coarse approximation to high-resolution approximation, and led to the formulation of MRA (Mallat 1989a, b). Theoretically, a MRA of the space $L^2(R)$ consists of a sequence of successive approximation subspaces $\{V_j, j \in Z\}$ that satisfies the following properties:

1. Monotonicity, that is, $\cdots \subset V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \subset \cdots$ (where the symbol \subset denotes a subset operator). This means that the subspace $\{V_j, j \in Z\}$ holds the successive inclusion relationship.
2. Completeness, that is, $\bigcap_{j \in Z} V_j = \{0\}$; $\bigcup_{j \in Z} V_j = L^2(R)$, where \cap denotes the intersect operator, and \cup denotes the union operator. This property indicates that all the subspaces together form a complete $L^2(R)$.
3. Dilation regularity, that is, $x(t) \in V_j \Leftrightarrow x(2^j t) \in V_0$, where \Leftrightarrow denotes “if and only if,” and \in denotes “is an element of.” The term $j \in Z$ indicates the multiresolution aspect of the subspaces $\{V_j, j \in Z\}$.
4. Translation invariance, that is, $x(t) \in V_0 \Rightarrow x(t - n) \in V_0$, for all $n \in Z$ (with \Rightarrow denotes “imply”).
5. Existence of orthogonal basis: there exists a function $\phi(t) \in V_0$, whose corresponding closed subspaces $\{\phi(t - n)\}_{n \in Z}$ form an orthogonal basis of the zero-scale space V_0 ; that is, $\int_R \phi(t - n)\phi(t - m) dt = \delta_{m,n}$.

The function $\phi(t)$ is the *scale* function, whose translated version $\phi_k(t) = \phi(t - k)$ satisfies the condition of $\langle \phi_k(t), \phi_{k'}(t) \rangle = \delta_{k,k'} (k, k' \in Z)$. The zero-scale space V_0 is composed of a set of closed subspaces, formed by $\phi_k(t)$ and is denoted as $V_0 = \overline{\text{span}}_k \{\phi(t - k)\}$.

Fig. 4.1 Inclusion relationship among closed subspaces $\{V_j, j \in \mathbb{Z}\}$

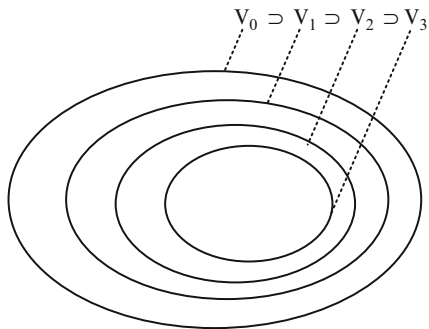
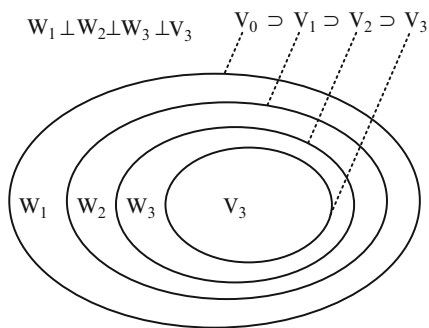


Fig. 4.2 Illustration of wavelet subspaces



From the above description, we know that all the closed subspaces $\{V_j, j \in \mathbb{Z}\}$ are formed from the same scale function $\phi(t)$ with different translation values, and the relationship among all the subspaces is illustrated in Fig. 4.1. It can be seen that the closed subspaces $\{V_j, j \in \mathbb{Z}\}$ hold the inclusion relationship, and they are not orthogonal. As a result, the scale function family $\phi_{j,k}(t) = 2^{-j/2}\phi(2^{-j}t - k)$ does not hold the orthogonal property; that is, $\{\phi_{j,k}(t)\}_{j \in \mathbb{Z}, k \in \mathbb{Z}}$ cannot be used as an orthogonal basis in $L^2(\mathbb{R})$ space.

To find the orthogonal bases in the $L^2(\mathbb{R})$ space, we can define W_j ($j \in \mathbb{Z}$) as the orthogonal complement of V_j in V_{j-1} , as illustrated in Fig. 4.2.

We can then write as follows:

$$V_{j-1} = V_j \oplus W_j \quad (4.19)$$

and

$$W_j \perp W_{j'}, \quad \text{for } j \neq j' \quad (4.20)$$

where the symbol \oplus denotes direct summation operator, and \perp denotes the orthogonal operator.

It follows that, for $j < J$, we can have the following relationship:

$$V_j = V_J \oplus \bigoplus_{k=0}^{J-j-1} W_{J-k} \quad (4.21)$$

where all the subspaces W_j ($j \in Z$) are orthogonal, and they form the $L^2(R)$ space as

$$L^2(R) = \bigoplus_{j \in Z} W_j \quad (4.22)$$

Furthermore, the W_j spaces inherit the scaling property from the V_j (Daubechies 1992); that is,

$$x(t) \in W_0 \Leftrightarrow x(2^{-j}t) \in W_j \quad (4.23)$$

Therefore, if $\{\psi_{0,k} \mid k \in Z\}$ is a set of orthogonal bases in W_0 space, then according to (4.23), for the scale $j \in Z$, $\{\psi_{j,k} = 2^{-j/2} \psi(2^{-j}t - k) \mid k \in Z\}$ is a collection of orthogonal bases in the W_j space. Accordingly, the entire collection of $\{\psi_{j,k} \mid j \in Z, k \in Z\}$ forms the sets of orthogonal bases in $L^2(R)$ space, and we call the function $\psi(t)$ the wavelet function, and the W_j space in (4.23) denotes the wavelet space in scale j .

4.2.2 Orthogonal Wavelet Transform

From the definition of the MRA, we know that

$$V_0 = V_1 \oplus W_1 = V_2 \oplus W_2 + W_1 = V_3 \oplus W_3 \oplus W_2 + W_1 = \dots \quad (4.24)$$

Therefore, for a given signal $x(t) \in V_0$, where V_0 is defined as zero-scale space, we can decompose it into two parts (the detailed information in W_1 and the approximate information in V_1). The approximate information in V_1 can then be further decomposed to get the next level of detailed information in W_2 and approximate information in V_2 , respectively. Such a decomposition process can be repeated until the designed scale j is reached. This, in a nutshell, is how a DWT is implemented.

Mathematically, we can define $x_a^j(t)$ as the approximate information at scale j after the signal $x(t)$ is projected onto the V_j space:

$$x_a^j(t) = \sum_k a_{j,k} \phi_k(2^{-j}t) = \sum_k a_{j,k} \phi_{j,k}(t), \quad k \in Z \quad (4.25)$$

where

$$a_{j,k} = \langle x(t), \phi_{j,k}(t) \rangle \quad (4.26)$$

are called the *approximate* coefficients.

Similarly, when the signal $x(t)$ is projected onto the W_j space, the detailed information at scale j is obtained as

$$x_d^j(t) = \sum_k d_{j,k} \psi_k(2^{-j}t) = \sum_k d_{j,k} \psi_{j,k}(t), \quad k \in \mathbb{Z} \quad (4.27)$$

where

$$d_{j,k} = \langle x(t), \psi_{j,k}(t) \rangle \quad (4.28)$$

are called *detailed* coefficients.

Consequently, when a given signal $x(t) \in L^2(\mathbb{R})$ is decomposed into the set of subspaces,

$$L^2(\mathbb{R}) = \sum_{j=-\infty}^J W_j \oplus V_J \quad (4.29)$$

with J being any predetermined scale, we will have

$$x(t) = \sum_{j=-\infty}^J \sum_{k=-\infty}^{\infty} d_{j,k} \psi_{j,k}(t) + \sum_{k=-\infty}^{\infty} a_{J,k} \phi_{J,k}(t) \quad (4.30)$$

If $J \rightarrow \infty$, (4.30) can be simplified as

$$x(t) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} d_{j,k} \psi_{j,k}(t) \quad (4.31)$$

Equation (4.31) is equivalent to (4.17) when $A = B = 1$. We know that in such cases the wavelet bases are orthogonal (Daubechies 1992). As a result, (4.30) and (4.31) express the inverse orthogonal wavelet transform, and (4.26) and (4.28) express the orthogonal wavelet transform. From the above description, we see that the idea of orthogonal wavelet transform and MRA has followed the same path; thus, MRA provides the theoretical basis for orthogonal wavelet transform.

4.3 Dual-Scale Equation and Multiresolution Filters

The inherent relationship between the scale function $\phi(t)$ and wavelet function $\psi(t)$ can be expressed in a dual-scale equation as

$$\phi(t) = \sum_n h(n) \phi_{-1,n}(t) = \sqrt{2} \sum_n h(n) \phi(2t - n) \quad (4.32)$$

$$\psi(t) = \sum_n g(n) \phi_{-1,n}(t) = \sqrt{2} \sum_n g(n) \phi(2t - n) \quad (4.33)$$

where

$$\begin{cases} h(n) = \langle \phi, \phi_{-1,n} \rangle \\ g(n) = \langle \psi, \phi_{-1,n} \rangle \end{cases} \quad (4.34)$$

It should be noted that the dual-scale relationship only exists between two successive scales j and $j - 1$; that is,

$$\phi_{j,0}(t) = \sum_n h(n) \phi_{j-1,n}(t) \quad (4.35)$$

$$\psi_{j,0}(t) = \sum_n g(n) \phi_{j-1,n}(t) \quad (4.36)$$

Furthermore, the coefficients $h(n)$ and $g(n)$ will not change with the scale j . This can be proved as follows:

Proof:

$$\begin{aligned} \langle \phi_{j,0}(t), \phi_{j-1,n}(t) \rangle &= \int_R [2^{-j/2} \phi(2^{-j}t)] [2^{-j/2} \phi^*(2^{-j+1}t - n)] dt \\ &= \sqrt{2} \int \phi(t') \phi^*(2t' - n) dt' \quad (\text{let } t' = 2^{-j}t) \\ &= \langle \phi(t), \phi_{-1,n}(t) \rangle = h(n) \end{aligned} \quad (4.37)$$

Similarly, we can prove that $\langle \psi_{j,0}(t), \phi_{j-1,n}(t) \rangle = g(n)$. This means that the coefficients $h(n)$ and $g(n)$ are determined by the scaling function $\phi(t)$ and wavelet function $\psi(t)$, respectively, and are not related to how we choose the scale j . Furthermore, if we perform an integral operation on both sides of (4.35), we obtain the following:

$$\int_R \phi_{j,0}(t) dt = \sum_n h(n) \int_R \phi_{j-1,n}(t) dt \quad (4.38)$$

As

$$\begin{aligned} \int_R \phi_{j-1,n}(t) dt &= 2^{-\frac{j-1}{2}} \int_R \phi(2^{-j+1}t - n) dt \\ &\stackrel{t'=2t}{=} \sqrt{2} \int_R 2^{-\frac{j}{2}} \phi(2^{-j}t' - n) \frac{1}{2} dt' \\ &= \frac{1}{\sqrt{2}} \int_R \phi_{j,n}(t) dt \\ &= \frac{1}{\sqrt{2}} \int_R \phi_{j,0}(t) dt \end{aligned} \quad (4.39)$$

substituting (4.39) in (4.38) leads to

$$\sum_n h(n) = \sqrt{2} \quad (4.40)$$

Similarly, we can perform an integral operation on both sides of (4.36) as

$$\int_R \psi_{j,0}(t) dt = \sum_n g(n) \int_R \phi_{j-1,n}(t) dt \quad (4.41)$$

Given that $\int_R \psi(t) dt = 0$, (4.41) can be simplified as

$$\sum_n g(n) = 0 \quad (4.42)$$

The coefficients $h(n)$ and $g(n)$ are called a pair of *low-pass* and *high-pass* wavelet filters, which are used to realize the DWT, on the basis of the Mallat algorithm, as described below.

4.4 The Mallat Algorithm

The dual-scale (4.32) can be rewritten as

$$\phi(t) = \sum_n h(n) \sqrt{2} \phi(2t - n) \quad (4.43)$$

Accordingly, the scaled and translation version of $\phi(t)$ can then be expressed as

$$\begin{aligned} \phi(2^{-j}t - k) &= \sum_n h(n) \sqrt{2} \phi(2(2^{-j}t - k) - n) \\ &= \sum_n h(n) \sqrt{2} \phi(2^{-j+1}t - 2k - n) \end{aligned} \quad (4.44)$$

Let $m = 2k + n$; then (4.44) can be rewritten as

$$\phi(2^{-j}t - k) = \sum_n h(m - 2k) \sqrt{2} \phi(2^{-j+1}t - m) \quad (4.45)$$

On the basis of the theory of MRA, we can define the following:

$$V_{j-1} = \overline{\text{span}_k \{2^{(j-1)/2} \phi(2^{-j+1}t - k)\}} \quad (4.46)$$

As a result, a given signal $x(t)$ in the V_{j-1} space can be expressed as

$$x(t) = \sum_k a_{j-1,k} 2^{(j+1)/2} \phi(2^{j+1}t - k) \quad (4.47)$$

If such a signal is projected (i.e., decomposed) onto the V_j and W_j spaces, the result can be expressed as

$$x(t) = \sum_k a_{j,k} 2^{j/2} \phi(2^j t - k) + \sum_k d_{j,k} 2^{j/2} \psi(2^j t - k) \quad (4.48)$$

where $a_{j,k}$ and $d_{j,k}$ are calculated as

$$a_{j,k} = \langle x(t), \phi_{j,k}(t) \rangle = \int_R x(t) 2^{j/2} \phi^*(2^j t - k) dt \quad (4.49)$$

$$d_{j,k} = \langle x(t), \psi_{j,k}(t) \rangle = \int_R x(t) 2^{j/2} \psi^*(2^j t - k) dt \quad (4.50)$$

Substituting (4.45) in (4.49) results in

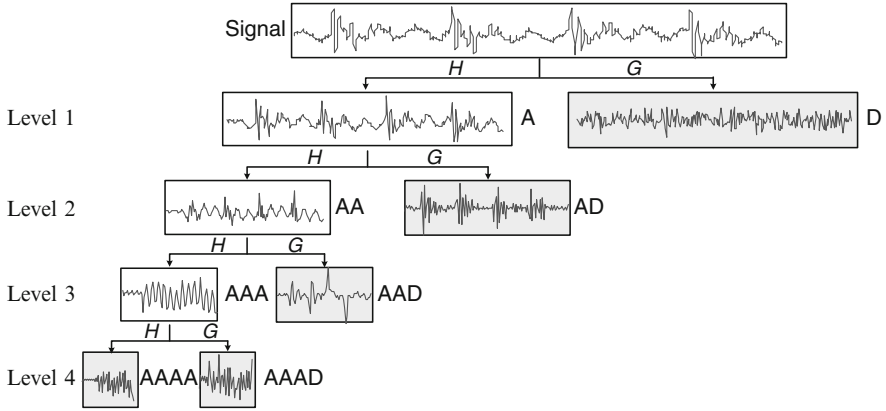
$$\begin{aligned} a_{j,k} &= \sum_m h(m - 2k) \int_R x(t) 2^{(j+1)/2} \phi^*(2^{j+1}t - m) dt \\ &= \sum_m h(m - 2k) \langle x(t), \phi_{j-1,m} \rangle \\ &= \sum_m h(m - 2k) a_{j-1,m} \end{aligned} \quad (4.51)$$

Similarly, (4.50) can be further rewritten as

$$\begin{aligned} d_{j,k} &= \sum_m g(m - 2k) \int_R x(t) 2^{(j+1)/2} \psi^*(2^{j+1}t - m) dt \\ &= \sum_m h(m - 2k) \langle x(t), \psi_{j-1,m} \rangle \\ &= \sum_m h(m - 2k) d_{j-1,m} \end{aligned} \quad (4.52)$$

This means that, through such a pair of filters, the signal $x(t)$ is decomposed into low- and high-frequency components, respectively, as (Mallat 1998)

$$\begin{cases} a_{j,k} = \sum_m h(m - 2k) a_{j-1,m} \\ d_{j,k} = \sum_m g(m - 2k) a_{j-1,m} \end{cases} \quad (4.53)$$



Note: H - Low pass filter; G - High pass filter; A - Approximate information; D - Detailed information

Fig. 4.3 Procedure of a four level signal decomposition using discrete wavelet transform. Note: H low pass filter, G high pass filter, A approximate information, D detailed information

In (4.53), $a_{j,k}$ is the *approximate* coefficient, which represents the low-frequency component of the signal, and $d_{j,k}$ is the *detailed* coefficient, which corresponds to the high-frequency component. The approximate coefficients at wavelet decomposition level j are obtained by convolving the approximate coefficients at the previous decomposition level $(j - 1)$ with the low-pass filter coefficients. Similarly, the detailed coefficients at wavelet decomposition level j are obtained by convolving the approximate coefficients at the previous decomposition level $(j - 1)$ with the high-pass filter coefficients. Such a process represents the idea of Mallat's algorithm to implement the DWT, and is schematically shown in Fig. 4.3.

From Fig. 4.3, we see that a signal is decomposed by a four-level DWT. After passing through the high-pass and low-pass filters on the first level (level 1), the output of the low-pass filter, denoted as the *approximate* coefficients of the level 1, is filtered again by the second-level filter banks. The process repeats itself, and at the end of the fourth level decomposition, the signal is decomposed into five feature groups: one group containing the lowest frequency components, denoted as the *approximate* information and labeled as $AAAA$, and four groups containing progressively higher frequency components, called the *detailed* information and labeled as $AAAD$, AAD , AD , and D . The levels 1–4 correspond to the wavelet scales $2^1 = 2$, $2^2 = 4$, $2^3 = 8$, and $2^4 = 16$, respectively.

4.5 Commonly Used Base Wavelets

This section introduces several commonly used orthogonal wavelets, which can be used as the basis for performing the DWT.

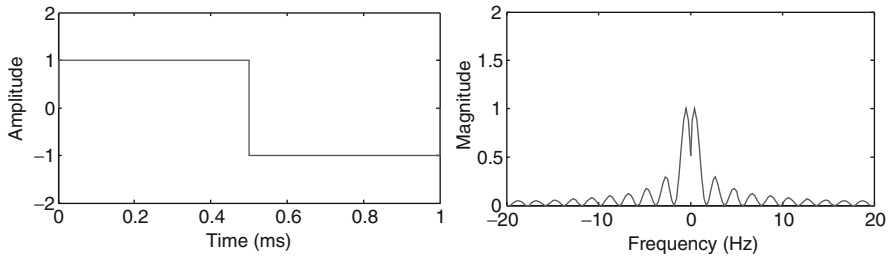


Fig. 4.4 Haar wavelet (*left*) and its magnitude spectrum (*right*)

4.5.1 Haar Wavelet

The Haar wavelet is mathematically defined as (Haar 1910)

$$\psi_{\text{Haar}}(t) = \begin{cases} 1 & 0 \leq t < \frac{1}{2} \\ -1 & \frac{1}{2} \leq t < 1 \\ 0 & \text{otherwise} \end{cases} \quad (4.54)$$

Its function and magnitude spectrum are illustrated in Fig. 4.4.

The Haar wavelet is orthogonal and symmetric in nature. The property of symmetry ensures that the Haar wavelet has linear phase characteristics, meaning that when a wavelet filtering operation is performed on a signal with this base wavelet, there will be no phase distortion in the filtered signal. Furthermore, it is the simplest base wavelet with the highest time resolution given by a compact support of one as shown in (4.54) (Daubechies 1992). However, the rectangular shape of the Haar wavelet determines its corresponding spectrum with slow decay characteristics, leading to a low frequency resolution. Examples of using the Haar wavelet for manufacturing related work include the stamping process monitoring (Zhou et al. 2006) and fault detection in dry etching process (Kim et al. 2010).

4.5.2 Daubechies Wavelet

The family of the Daubechies wavelets is orthogonal, however, asymmetric, which introduces a large phase distortion. This means that it cannot be used in applications where a signal's phase information needs to be kept. It is also a compactly supported base wavelet with a given support width of $2N - 1$, in which N is the order of the base wavelet (Daubechies 1992). In theory, N can be up to infinity. In real-world applications, the Daubechies wavelets with order up to 20 have been used. The Daubechies wavelets do not have explicit expression except for the one with $N = 1$, which is actually the Haar wavelet as discussed above. With an increase of the support width (i.e., an increase of the base wavelet order), the Daubechies

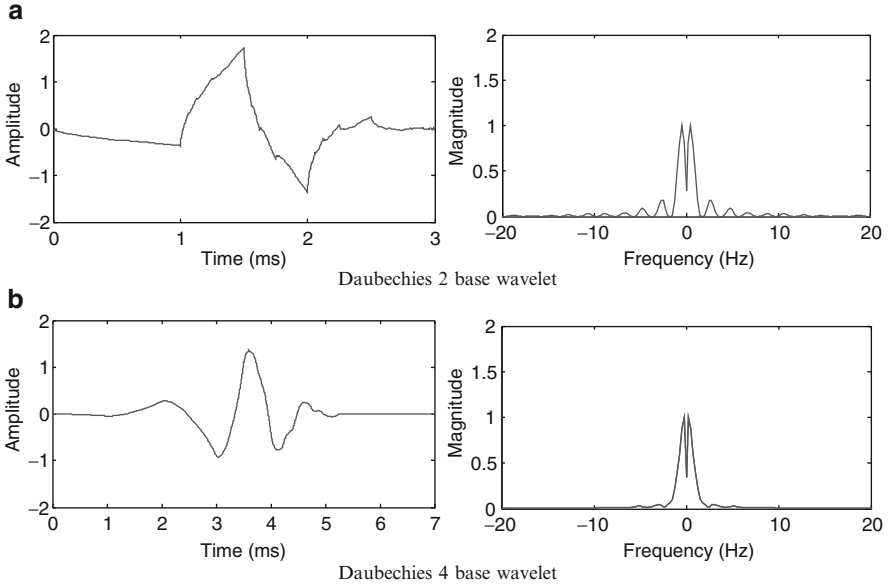


Fig. 4.5 Daubechies wavelet (*left*) and its magnitude spectrum (*right*). (a) Daubechies 2 base wavelet and (b) Daubechies 4 base wavelet

wavelet becomes increasingly smoother, leading to better frequency localization. Accordingly, the magnitude spectra for each of the Daubechies wavelets decay quickly, as illustrated in Fig. 4.5, where the Daubechies 2 base wavelet and Daubechies 4 base wavelet are used as examples.

The Daubechies wavelets have been widely investigated for fault diagnosis of bearings (Nikolaou and Antoniadis 2002; Lou and Loparo 2004) and automatic gears (Rafiee et al. 2010)

4.5.3 Coiflet Wavelet

The family of the Coiflet wavelets is orthogonal (Daubechies 1992), and near symmetric. This property of near symmetry leads to the near linear phase characteristics of the Coiflet wavelet. They are designed to yield the highest number of vanishing moments ($2N$) for both the base wavelet of the order N and the scaling function, for a given support width of $6N - 1$. Figure 4.6 illustrates the sample waveforms of the Coiflet wavelets, with their corresponding magnitude spectra at orders 2 and 4, respectively. The Coiflet wavelet has been used for fault diagnosis of rolling bearings (Sugumaran and Ramachandran 2009).

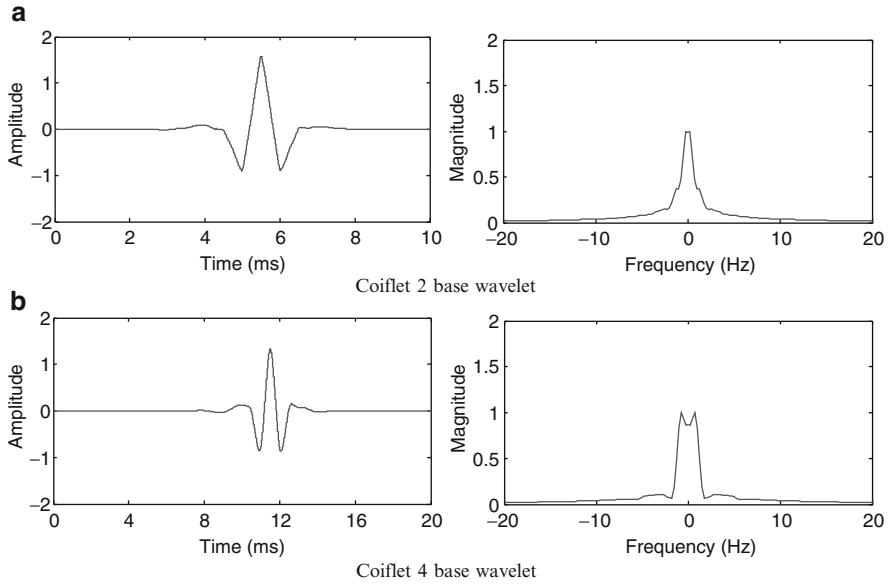


Fig. 4.6 Coiflet wavelet (*left*) and its magnitude spectrum (*right*). (a) Coiflet 2 base wavelet and (b) Coiflet 4 base wavelet

4.5.4 Symlet Wavelet

Symlet wavelets (Daubechies 1992) are orthogonal and near symmetric. This property ensures minimal phase distortion. A Symlet wavelet of order N has the number of vanishing moments N for a given support width of $2N - 1$. They are similar to the Daubechies wavelet, except for better symmetry. Waveforms with their corresponding magnitude spectra for the Symlet wavelet at orders 2 and 4 are illustrated in Fig. 4.7a, b, respectively. Examples of using the Symlet wavelet for signal decomposition in manufacturing-related problems include characterization of fabric texture (Shakher et al. 2004) and health monitoring of rolling bearings (Gao and Yan 2006).

4.5.5 Biorthogonal and Reverse Biorthogonal Wavelets

The family of biorthogonal and reverse biorthogonal wavelets (Daubechies 1992) is biorthogonal and symmetric. The property of symmetry ensures that they have linear phase characteristics. This type of base wavelet can be constructed by the spline method (Cohen et al. 1992). Figures 4.8 and 4.9 illustrate sample waveforms with their magnitude spectrum for several biorthogonal and reverse biorthogonal wavelets, respectively. In practice, this group of wavelets has been used for surface profile filtering in manufacturing process monitoring and diagnostics (Fu et al. 2003).

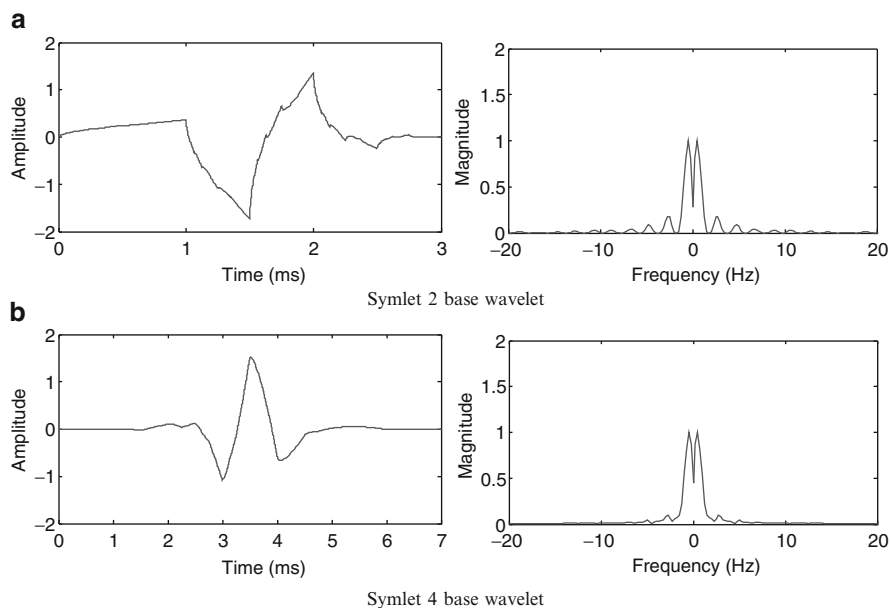


Fig. 4.7 Symlet wavelet (*left*) and its magnitude spectrum (*right*). (a) Symlet 2 base wavelet and (b) Symlet 4 base wavelet

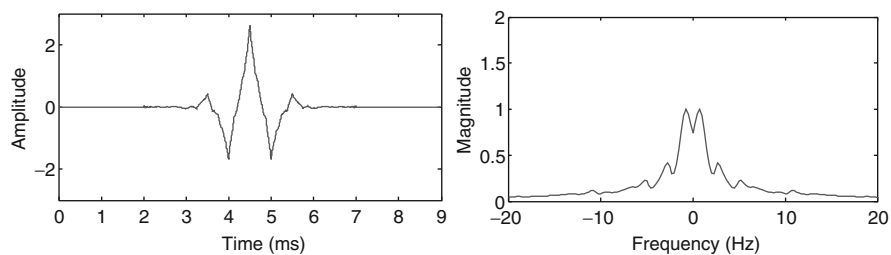


Fig. 4.8 Biorthogonal 2.4 wavelet (*left*) and its magnitude spectrum (*right*)

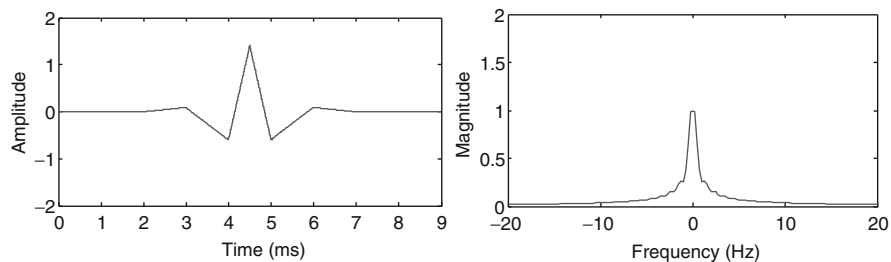


Fig. 4.9 Reverse biorthogonal 2.4 wavelet (*left*) and its magnitude spectrum (*right*)

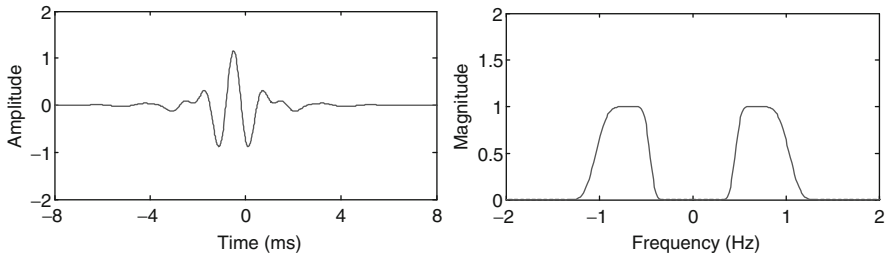


Fig. 4.10 Meyer wavelet (*left*) and its magnitude spectrum (*right*)

4.5.6 Meyer Wavelet

The Meyer wavelet is orthogonal and symmetric. However, it does not have a finite support. The Meyer wavelet has explicit expression and is defined in the frequency domain as follows:

$$\Psi_{Meyer}(f) = \begin{cases} \sqrt{2\pi} e^{i\pi f} \sin\left[\frac{\pi}{2} v(3|f| - 1)\right] & \frac{1}{3} \leq |f| \leq \frac{2}{3} \\ \sqrt{2\pi} e^{i\pi f} \cos\left[\frac{\pi}{2} v\left(\frac{3}{2}|f| - 1\right)\right] & \frac{2}{3} \leq |f| \leq \frac{4}{3} \\ 0 & |f| \notin \left(\frac{1}{3}, \frac{4}{3}\right) \end{cases} \quad (4.55)$$

where $v(\cdot)$ is an auxiliary function, expressed as

$$v(\alpha) = \alpha^4(35 - 84\alpha + 70\alpha^2 - 20\alpha^3), \quad \alpha \in \langle 0, 1 \rangle \quad (4.56)$$

The Meyer wavelet with its magnitude spectrum is illustrated in Fig. 4.10.

Typical applications of Meyer wavelet in manufacturing-related problems include signal denoising and bearing fault diagnosis (Abbasion et al. 2007).

4.6 Application of Discrete Wavelet Transform

One of the most popular applications of the DWT is to remove noise contained in a signal. This is based on the observation that a signal's energy is often distributed over a few wavelet coefficients with high magnitude, while energy of the noise is distributed across most of the wavelet coefficients with low magnitude. A thresholding scheme can therefore be devised to remove the noise. Mathematically, assume a signal with noise contamination expressed as

$$y(t) = x(t) + \sigma e(t) \quad (4.57)$$

where $x(t)$ is the signal, $e(t)$ is a Gaussian white noise $N(0,1)$, and σ represents the noise level. The objective of denoising is to suppress the noise $e(t)$ and to recover the signal $x(t)$. Generally, the denoising procedure consists of three steps:

1. *Signal decomposition*: Choosing a base wavelet and a decomposition level J , and then performing DWT up to level J on the signal.
2. *Detailed coefficients thresholding*: For each decomposition level from 1 to J , selecting a threshold and applying it to the detailed coefficients.
3. *Signal reconstruction*: Performing wavelet reconstruction to obtain denoised signal, based on the original approximate coefficients of level J and the modified detailed coefficients of levels 1 to J .

It should be noted that two thresholding approaches (hard thresholding and soft thresholding) can be used in the denoising process (Donoho 1995; Donoho and Johnstone 1995). Hard thresholding can be described as the process of setting the value of the detailed coefficient $d_{j,k}$ to zero, if its absolute value is lower than the threshold (denoted as thr). This is mathematically expressed as

$$\hat{d}_{j,k} = \begin{cases} d_{j,k} & |d_{j,k}| \geq thr \\ 0 & |d_{j,k}| < thr \end{cases} \quad (4.58)$$

Soft thresholding can be considered as an extension of the hard thresholding, as shown in Fig. 4.11. It sets those detailed coefficients to zero if their absolute values are lower than the threshold, and then shrinks the nonzero coefficients toward zero. Mathematically, this can be expressed as

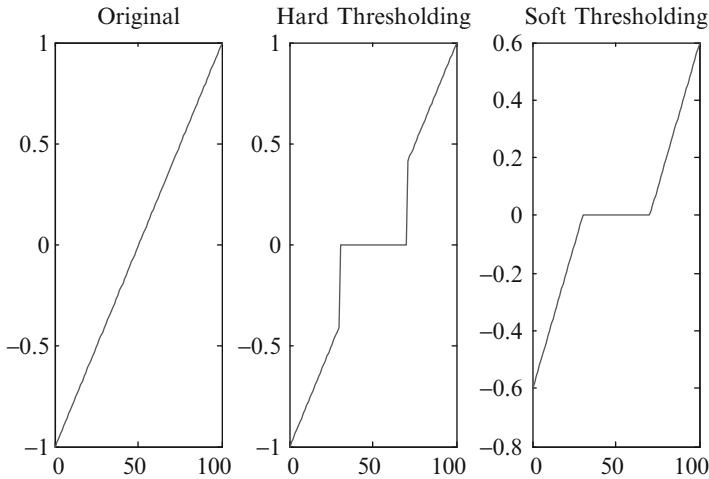


Fig. 4.11 Illustration of hard thresholding and soft thresholding

$$\hat{d}_{j,k} = \begin{cases} \text{sgn}(d_{j,k})(|d_{j,k}| - \text{thr}) & d_{j,k} \geq \text{thr} \\ 0 & d_{j,k} < \text{thr} \end{cases} \quad (4.59)$$

where

$$\text{sgn}(d_{j,k}) = \begin{cases} +1 & d_{j,k} \geq 0 \\ -1 & d_{j,k} < 0 \end{cases} \quad (4.60)$$

As an example, Fig. 4.12a shows a “blocks” test signal, and Fig. 4.12b shows that it is contaminated by a Gaussian white noise to make a signal-to-noise ratio of 4. The signal is decomposed up to level 3, with sym8 wavelet being the base wavelet. After performing soft thresholding to the detailed coefficients at each decomposition level, the signal is reconstructed as shown in Fig. 4.12c. As only a small number of large coefficients characterize the original “blocks” signal, this DWT-based denoising method performs well.

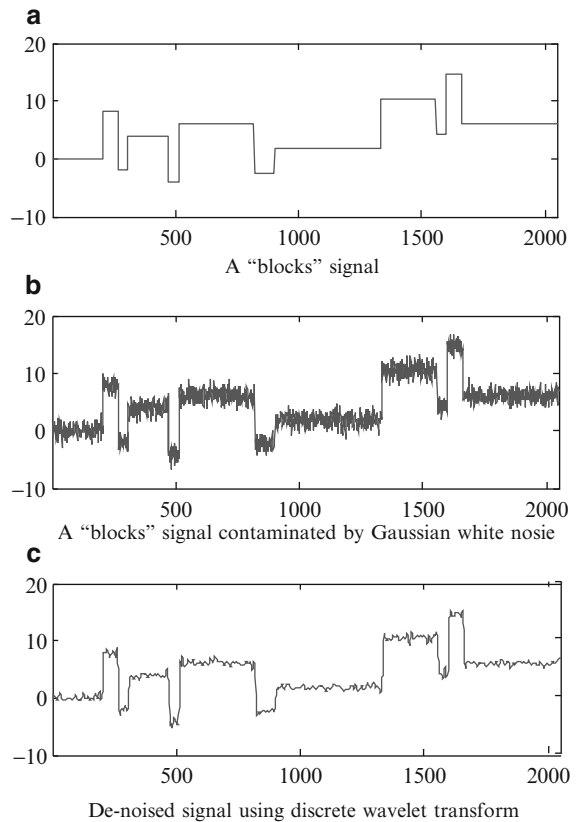


Fig. 4.12 Example of discrete wavelet transform for denoising. (a) A “blocks” signal, (b) a “blocks” signal contaminated by Gaussian white noise, and (c) denoised signal using discrete wavelet transform

4.7 Summary

This chapter begins with a description of the discretization of the scale and translation parameters. The MRA and orthogonal wavelet transform are then introduced in Sect. 4.2. After that, we describe in Sect. 4.3 the dual-scale equation and its associated wavelet filter pair. The Mallat algorithm for implementing the DWT is then discussed in Sect. 4.4, followed by the introduction of some commonly used wavelets in Sect. 4.5. Some typical applications of the DWT are shown in Sect. 4.6.

4.8 References

- Abbasian S, Rafsanjani A, Farshidianfar A, Irani N (2007) Rolling element bearings multi fault classification based on the wavelet denoising and support vector machine. *Mech Syst Signal Process* 21:2933–2945
- Addison N (2002) *The illustrated wavelet transform handbook*. Taylor & Francis, New York
- Burt P, Adelson E (1983) The Laplacian pyramid as a compact image code. *IEEE Trans Commun* 31:482–540
- Cohen A, Daubechies I, Feauveau, JC (1992) Biorthogonal bases of compactly supported wavelets. *Commun Pure Appl Math* 45:485–560
- Daubechies I (1992) *Ten lectures on wavelets*. SIAM, Philadelphia
- Donoho DL (1995) De noising by soft thresholding. *IEEE Trans Inform Theory*, 41(3): 613–627
- Donoho DL; Johnstone IM (1995) Adapting to unknown smoothness via wavelet shrinkage. *J Am Stat Assoc* 90(432):1200–1244
- Fu S, Muralikrishnan, Raja J (2003) Engineering surface analysis with different wavelet bases. *ASME J Manuf Sci Eng* 125(6):844–852
- Gao R, Yan R (2006) Non stationary signal processing for bearing health monitoring. *Int J Manuf Res* 1(1):18–40
- Haar A (1910) Zur theorie der orthogonalen funktionensysteme. *Math Annalen* 69:331–371
- Kim JS, Lee JH, Kim JH, Baek J, Kim SS (2010) Fault detection of cycle based signals using wavelet transform in FAB processes. *Int J Precision Eng Manuf* 11(2):237–246
- Lou X, Loparo KA (2004) Bearing fault diagnosis based on wavelet transform and fuzzy inference. *Mech Syst Signal Process* 18:1077–1095
- Mallat SG (1989a) A theory of multiresolution signal decomposition: the wavelet representation. *IEEE Trans Pattern Anal Machine Intell* 11(7) 674–693
- Mallat SG (1989b) Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$. *Trans Am Math Soc* 315:69–87
- Mallat SG (1998) *A wavelet tour of signal processing*. Academic, San Diego, CA
- Nikolaou NG, Antoniadis IA (2002) Rolling element bearing fault diagnosis using wavelet packets. *NDT&E Int* 35:197–205
- Rafiee J, Rafiee MA, Tse PW (2010) Application of mother wavelet functions for automatic gear and bearing fault diagnosis. *Expert Syst Appl* 37:4568–4579
- Shakher C, Ishtiaque SM, Singh SK, Zaidi HN (2004) Application of wavelet transform in characterization of fabric texture. *J Text Inst* 95(1–6):107–120
- Sugumaran V, Ramachandran KI (2009) Wavelet selection using decision tree for fault diagnosis of roller bearings. *Int J Appl Eng Res* 4(2):201–225
- Witkin A (1983) Scale space filtering. In: *Proceedings of international joint conference on artificial intelligence*, Karlsruhe, Germany, pp 1019–1023
- Zhou SY, Sun BC, Shi JJ (2006) An SPC monitoring system for cycle based waveform signals using haar transform. *IEEE Trans Automat Sci Eng* 3(1):60–72

Chapter 5

Wavelet Packet Transform

While discrete wavelet transform provides flexible time frequency resolution, it suffers from a relatively low resolution in the high-frequency region. This deficiency leads to difficulty in differentiating high-frequency transient components. The wavelet packet transform (WPT), in comparison, further decomposes the *detailed* information of the signal in the high-frequency region, thereby overcoming this limitation. Figure 5.1 schematically illustrates a WPT-based signal decomposition process, where a four-level WPT produces a total of 16 subbands, with each subband covering one-sixteenth of the signal frequency spectrum (Gao and Yan 2006). The enhanced signal decomposition capability makes WPT an attractive tool for detecting and differentiating transient elements with high-frequency characteristics.

In this chapter, we introduce the theoretical basis of a wavelet packet and algorithms to realize the WPT. Representative applications of the WPT are then introduced to illustrate this computational technique.

5.1 Theoretical Basis of Wavelet Packet

5.1.1 Definition

The wavelet packet is defined by the following equation (Wickerhauser 1991):

$$\begin{cases} u_{2n}^{(j)}(t) = \sqrt{2} \sum_k h(k) u_n^{(j)}(2t - k) \\ u_{2n+1}^{(j)}(t) = \sqrt{2} \sum_k g(k) u_n^{(j)}(2t - k). \end{cases} \quad \text{with } n = 0, 1, 2, \dots \text{ and } k = 0, 1, \dots, m \quad (5.1)$$

with $u_0^{(0)}(t)$ being the scaling function $\phi(t)$, that is, $u_0^{(0)}(t) = \phi(t)$, and $u_1^{(0)}(t)$ being the base wavelet function $\psi(t)$, that is, $u_1^{(0)}(t) = \psi(t)$ (Wickerhauser 1991). The superscript (j) in (5.1) denotes the j th level wavelet packet basis, and there will be 2^j wavelet packet bases at the j th level.

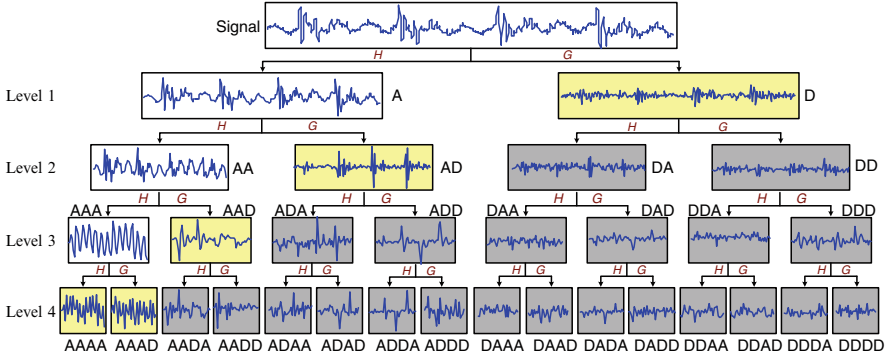


Fig. 5.1 Procedure for signal decomposition using wavelet packet transform. Note: *A* approximate information, *D* detailed information, *H* low pass filter, *G* high pass filter

To illustrate the derivation process of wavelet packet basis, the Haar wavelet (Haar 1910) is used here as an example. The coefficients $h(k)$ and $g(k)$ for Haar wavelet are defined as (Daubechies 1992)

$$\begin{cases} h(0) = h(1) = \frac{1}{\sqrt{2}}, & h(k) = 0 \text{ when } k = 2, 3, \dots, m \\ g(0) = g(1) = -\frac{1}{\sqrt{2}}, & g(k) = 0 \text{ when } k = 2, 3, \dots, m \end{cases} \quad (5.2)$$

From (5.1) and (5.2), the first level of the Haar wavelet packet basis, indicated by the superscript (1) is obtained as

$$\begin{cases} u_0^{(1)}(t) = u_0^{(0)}(2t) = \phi(2t) \\ u_1^{(1)}(t) = \sqrt{2} \frac{1}{\sqrt{2}} [u_0^{(1)}(2t) - u_0^{(1)}(2t-1)] = u_0^{(1)}(2t) - u_0^{(1)}(2t-1) \end{cases} \quad (5.3)$$

Similarly, the second and third levels of the Haar wavelet packet basis can be derived using (5.4) and (5.5), respectively:

$$\begin{cases} u_0^{(2)}(t) = \phi(4t) \\ u_1^{(2)}(t) = u_0^{(2)}(2t) - u_0^{(2)}(2t-1) \\ u_2^{(2)}(t) = u_1^{(2)}(2t) + u_1^{(2)}(2t-1) \\ u_3^{(2)}(t) = u_1^{(2)}(2t) - u_1^{(2)}(2t-1) \end{cases} \quad (5.4)$$

$$\begin{cases} u_0^{(3)}(t) = \phi(8t) \\ u_{2n}^{(3)}(t) = u_n^{(3)}(2t) + u_n^{(3)}(2t-1), & n = 1, 2, 3 \\ u_{2n+1}^{(3)}(t) = u_n^{(3)}(2t) - u_n^{(3)}(2t-1), & n = 0, 1, 2, 3 \end{cases} \quad (5.5)$$

Figure 5.2a c illustrates the waveforms of the Haar wavelet packet bases at levels 1 through 3 that are derived from the scaling function. Using the same approach, the Haar wavelet packet bases at all other levels can be obtained.

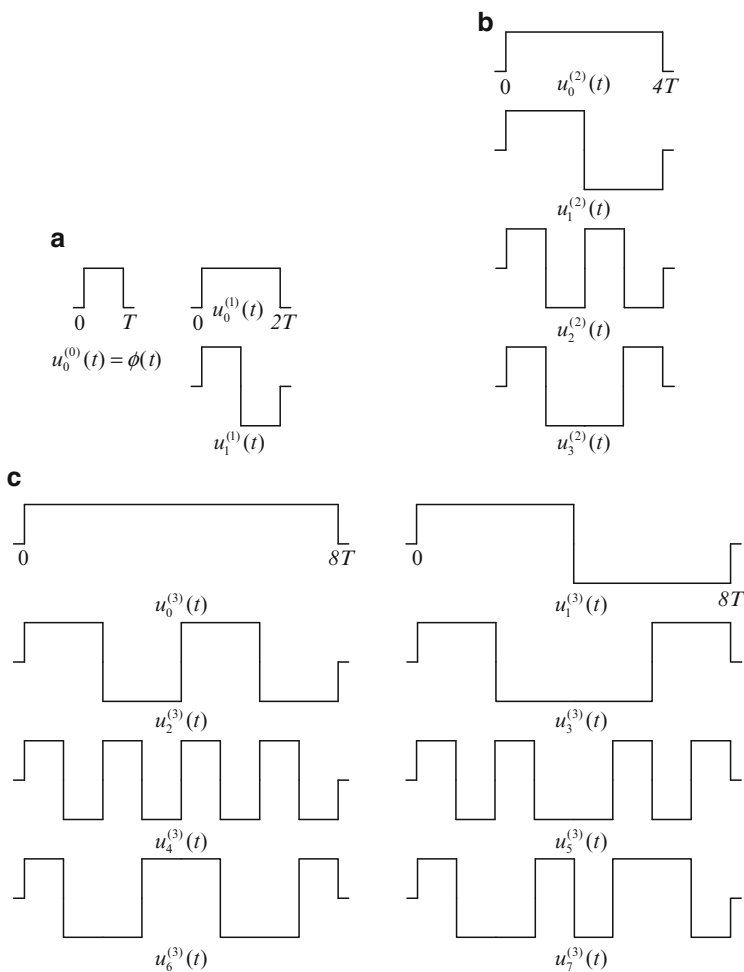


Fig. 5.2 Wavelet packet bases for the Haar wavelet: (a) level 1; (b) level 2; and (c) level 3

5.1.2 Wavelet Packet Property

Equation (5.1) indicates that the wavelet packet has the following properties (Wickerhauser 1991; Coifman et al. 1992).

5.1.2.1 Shift Orthogonality

If $\{u_n^{(j)}(t)\}_{n \in \mathbb{Z}}$ is the set of wavelet packet bases obtained from the scaling function $u_0^{(0)}(t) = \phi(t)$ of an orthogonal base wavelet, then these bases hold the property of shift orthogonality:

$$\langle u_n^{(j)}(t), u_n^{(j)}(t - k) \rangle = \delta_k, \quad k \in \mathbb{Z} \quad (5.6)$$

where $\langle \cdot \rangle$ denotes inner product operation. The symbol δ_k represents a Dirac function.

Proof: When $n = 0$, $u_0^{(j)}(t)$, and $u_1^{(j)}(t)$ are the scaled versions of $\phi(t)$ and $\psi(t)$, respectively. By definition of the scaling function and base wavelet function, they are orthogonal (Daubechies 1992).

When $n \neq 0$, as $u_2^{(j)}(t)$ and $u_3^{(j)}(t)$ are both a linear combination of $u_1^{(j)}(t)$ as seen in (5.3) and (5.4), and $u_1^{(j)}(t)$ is a scaled version of the wavelet function $\psi(t)$, which is orthogonal and normalized (Daubechies 1992), $u_2^{(j)}(t)$ and $u_3^{(j)}(t)$ are orthogonal.

As an example, if we have

$$\begin{cases} u_2^{(j)}(t) = \sqrt{2} \sum_{k'} h_{k'} u_1^{(j)}(2t - k') \\ u_2^{(j)}(t - k) = \sqrt{2} \sum_{k''} h_{k''} u_1^{(j)}(2t - 2k - k'') \end{cases} \quad (5.7)$$

where $k' = 0, 1, \dots, m$ and $k'' = 0, 1, \dots, m$.

Then

$$\langle u_2^{(j)}(t), u_2^{(j)}(t - k) \rangle = 2 \sum_{k'} \sum_{k''} h_{k'} h_{k''} \langle u_1^{(j)}(2t - k'), u_1^{(j)}(2t - 2k - k'') \rangle \quad (5.8)$$

The inner product on the right-hand side of (5.8) is equal to 1/2 when $k' = 2k + k''$; otherwise, it is equal to zero. Therefore,

$$\langle u_2^{(j)}(t), u_2^{(j)}(t - k) \rangle = \sum_{k'} h_{k'} h_{2k+k'} = \delta_k \quad (5.9)$$

Similarly, $u_4^{(j)}(t)$ and $u_5^{(j)}(t)$ are both linear combinations of $u_2^{(j)}(t)$, and they are also orthogonal. Using the same approach, wavelet packet bases of higher levels can be derived.

5.1.2.2 Orthogonal Relationship between $u_{2n}^{(j)}(t)$ and $u_{2n+1}^{(j)}(t)$

$$\langle u_{2n}^{(j)}(t), u_{2n+1}^{(j)}(t) \rangle = 0 \quad (5.10)$$

Proof From (5.1), we have

$$\begin{aligned} \langle u_{2n}^{(j)}(t), u_{2n+1}^{(j)}(t) \rangle &= 2 \int \sum_{k'} \sum_{k''} h_{k'} g_{k''} u_n^{(j)}(2t - 2k - k') u_n^{(j)}(2t - k'') dt \\ &= 2 \sum_{k'} \sum_{k''} h_{k'} g_{k''} \int u_n^{(j)}(2t - 2k - k') u_n^{(j)}(2t - k'') dt \end{aligned} \quad (5.11)$$

The result of integral part in (5.11) is equal to zero except when $k'' = 2k + k'$. Therefore,

$$\langle u_{2n}^{(j)}(t), u_{2n+1}^{(j)}(t) \rangle = \sum_{k''} h_{k'} g_{k''} = 0 \quad (5.12)$$

5.2 Recursive Algorithm

Once the wavelet packet basis is defined using (5.1), a recursive algorithm can be designed to implement WPT for signal decomposition. The result of the decomposition is given by (Mallat 1999):

$$\begin{cases} d_{j+1,2n} = \sum_m h(m - 2k) d_{j,n} \\ d_{j+1,2n+1} = \sum_m g(m - 2k) d_{j,n} \end{cases} \quad (5.13)$$

where $d_{j,n}$ denotes the wavelet coefficients at the j th level, n th subband, $d_{j+1,2n}$, and $d_{j+1,2n+1}$ denote the wavelet coefficients at the $(j+1)$ th level, $2n$ th, and $(2n+1)$ th subbands, respectively, and m is the number of the wavelet coefficients.

Theoretically, there are multiple ways ($>2^L$) to analyze a signal using an L -level decomposition (Mallat 1999). This makes it possible to optimize the signal decomposition process and improve the effectiveness. Various criteria, such as l_p ($p \leq 2$) norm, logarithmic entropy, and Shannon entropy, can be utilized as the cost function to facilitate the optimization process. A widely applied criterion for optimal WPT-based signal representation is the Shannon entropy (Coifman and Wickerhauser 1992). For wavelet coefficients at the n th subfrequency band within the level j , $d_{j,n} = \{d_{j,n} : n = 1, 2, \dots, 2^j\}$, the Shannon entropy is defined as

$$\text{Entropy}(d_{j,n}) = - \sum_i p_i \cdot \log(p_i) \quad (5.14)$$

where p_i is the probability distribution of the energy contained in the wavelet coefficients at the n th subfrequency band within the level j . The probability distribution function is defined as

$$p_i = |d_{j,n}(i)|^2 / \|d_{j,n}\|^2 \quad (5.15)$$

with $\sum_{i=1}^m p_i = 1$, and $p_i \cdot \log_2 p_i = 0$ if $p_i = 0$. The upper limit m represents the number of wavelet coefficients at the n th subfrequency band within the level j .

Equations (5.13) and (5.14) indicate that the entropy of the wavelet coefficients is bounded by

$$0 \leq E_{\text{entropy}}(d_{j,n}) \leq \log_2 m \quad (5.16)$$

From (5.16), we see that the Shannon entropy will have a large value if the energy content is spread out across the constituent wavelet coefficients within the subfrequency band. Conversely, it assumes a small value if the energy is concentrated on a few dominant components. As we want the signal information to be concentrated within as few coefficients as possible, the minimum Shannon entropy should be contained in the wavelet coefficients as a result of the signal decomposition. Mathematically, such a process involves comparing the entropy of the lower level (e.g., in the subbands *DAAA* and *DAAD*, Fig. 5.1) of the tree structure with the entropy of the higher source level (e.g., subband *DAA*), starting from the bottom of the decomposition (e.g., level 4). If the higher level has returned smaller entropy than the sum of the entropies from the lower level, then the higher level subfrequency band will be retained. Otherwise, it will be replaced by the two subfrequency bands at the lower level. Such a process is executed until it reaches the top level of the decomposition.

5.3 FFT-Based Harmonic Wavelet Packet Transform

Besides the recursive algorithm introduced in the previous section, another algorithm for WPT, based on the Fourier transform, has been shown to be effective when realizing the harmonic wavelet packet transform (HWPT) (Samuel et al. 2000; Yan and Gao 2005).

5.3.1 Harmonic Wavelet Transform

The mathematical expression of the harmonic wavelet is defined in Chap. 3 as

$$\Psi_{m,n}(f) = \begin{cases} \frac{1}{(n-m)} & m \leq f \leq n \\ 0 & \text{elsewhere} \end{cases} \quad (5.17)$$

Accordingly, its corresponding time domain expression is obtained by taking the inverse Fourier transform as (Yan and Gao 2005)

$$\psi_{m,n}(t) = \frac{e^{jn2\pi t} - e^{jm2\pi t}}{j2\pi(n-m)t} \quad (5.18)$$

If the harmonic wavelet is translated by a step $k/(m-n)$, in which k is the translation parameter, a generalized expression that is centered at $t = k/(n-m)$ with a bandwidth of $(n-m)$ can be obtained as (Newland 1994)

$$\psi_{m,n}\left(t - \frac{k}{n-m}\right) = \frac{\left(e^{jn2\pi\left(t - \frac{k}{n-m}\right)} - e^{jm2\pi\left(t - \frac{k}{n-m}\right)}\right)}{j2\pi(n-m)\left(t - \frac{k}{n-m}\right)} \quad (5.19)$$

On the basis of the generalized expression, the harmonic wavelet transform of a signal $x(t)$ can be performed as

$$hwt(m, n, k) = (n-m) \int_{-\infty}^{\infty} x(\tau) \psi_{m,n}^*\left(\tau - \frac{k}{n-m}\right) d\tau \quad (5.20)$$

where $hwt(m, n, k)$ is the harmonic wavelet coefficient.

By taking the Fourier transform of (5.20), an equivalent expression of the harmonic wavelet transform in the frequency domain can be expressed as

$$HWT(m, n, f) = X(f) \cdot \Psi^*((n-m)f) \quad (5.21)$$

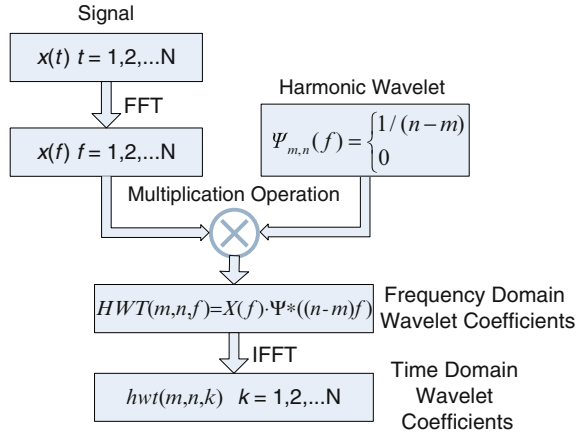
where $X(f)$ is the Fourier transform of the signal $x(t)$, and $\Psi^*((n-m)f)$ is the conjugate of $\Psi((n-m)f)$, which is the Fourier transform of the harmonic wavelet at the scale (m, n) . As the harmonic wavelet has compact frequency expression as shown in (5.17), the harmonic wavelet transform can be readily obtained through a pair of Fourier transform and inverse Fourier transform operations (Newland 1993).

As shown in Fig. 5.3, after taking the Fourier transform of a signal $x(t)$ to obtain its frequency domain expression $X(f)$, the inner product $HWT(m, n, f)$ of $X(f)$ and the conjugate of the harmonic wavelet $\Psi^*((n-m)f)$ at the scale (m, n) are calculated. Finally, the harmonic wavelet transform of the signal $x(t)$, denoted as $hwt(m, n, k)$, is obtained by taking the inverse Fourier transform of the inner product $HWT(m, n, f)$.

5.3.2 Harmonic Wavelet Packet Algorithm

The scale parameters m and n determine the bandwidth that the harmonic wavelet covers. Shown in Fig. 5.4a d are the real and imaginary parts of the generalized harmonic wavelet under two exemplary sets of scale parameters, $m = 0, n = 16$ and

Fig. 5.3 Algorithm for implementing the harmonic wavelet transform



$m = 16, n = 32$, while the translation parameter $k = 8$ remains the same. We can see that, through appropriate variation of these two scale parameters, the harmonic wavelet can be scaled to match the signal within different frequency regions associated with the same bandwidth of $(n - m)$ (16 in this example), as shown in Fig. 5.4e f. As a result, the HWPT operation is realized.

Similar to the WPT, the number of frequency subbands for the HWPT has to be s powers of two, in which s corresponds to the decomposition level of WPT. As a result, the signal can be decomposed into 2^s frequency subbands, with the bandwidth expressed in Hertz for each subband that is defined by

$$f_{\text{band}} = \frac{f_h}{2^s} \quad (5.22)$$

In (5.20), f_h is the highest frequency component of the signal to be analyzed. As the bandwidth of the harmonic wavelet is $(n - m)$, selection of the values for m and n of the HWPT has to satisfy the following condition:

$$m - n = f_{\text{band}} \quad (5.23)$$

Thus, the harmonic wavelet packet coefficients $hwpt(s, i, k)$ can be obtained as

$$hwpt(s, i, k) = hwt(m, n, k) \quad (5.24)$$

where s is the decomposition level, i is the index of the subband, and k is the index of the coefficient. In addition, the parameters m and n need to satisfy the following condition:

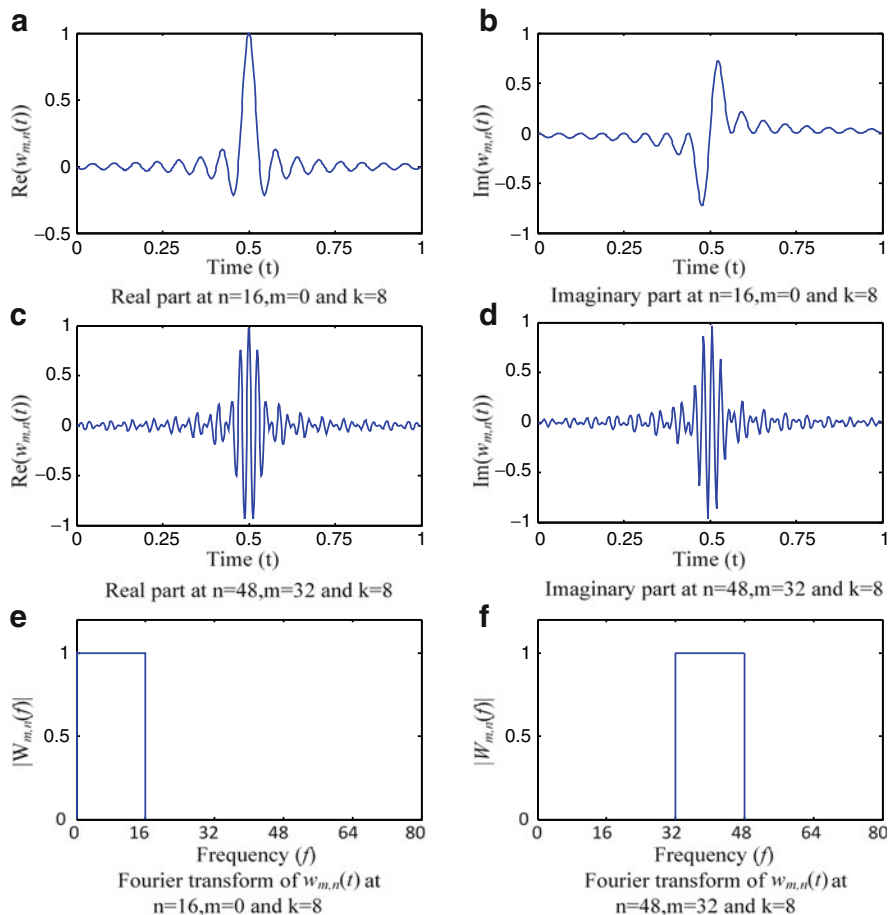


Fig. 5.4 Waveforms of the harmonic wavelet with their Fourier transforms under different scale parameters

$$\begin{cases} m = i \times f_{\text{band}} = i \times \frac{f_h}{2^s} \\ n = (i + 1) \times f_{\text{band}} = (i + 1) \times \frac{f_h}{2^s} \end{cases}, \quad i = 0, 1, \dots, 2^s - 1 \quad (5.25)$$

As a result, by selecting the appropriate parameter pairs (m, n) based on (5.25), the FFT-based HWPT algorithm can be realized through the computational process as illustrated in Fig. 5.3.

5.4 Application of Wavelet Packet Transform

Using the WPT, we can determine a signal’s time frequency composition, thereby having a good understanding of what is contained within the signal. Furthermore, the WPT can be applied to remove noise contained in the signal. In the following, we demonstrate two examples of these applications.

5.4.1 Time-Frequency Analysis

Figure 5.5 shows a vibration signal measured on a ball bearing during a run-to-failure test. Physically, when a localized defect is initiated in a rolling element bearing, for example, due to spalling on the surface of the bearing raceway, impact will be generated every time when a rolling element rolls over the defect. Such impacts subsequently excite the intrinsic modes of the bearing system, giving rise to transient vibrations at the mode-related resonant frequencies. As the defect size increases, different intrinsic modes of the bearing system will be excited, leading to frequency shifts of the impact-induced transient vibrations. Therefore, by evaluating the time frequency distributions of the vibration signal, degradation of the bearing’s health condition can be monitored.

Applying the WPT to the vibration data, we have seen in Fig. 5.6 that not only all the major transient elements are identified, but the corresponding frequency shifts are also clearly seen. The result also shows the increased number of frequency components after the 45-ms time point, reflecting the defect size propagation.

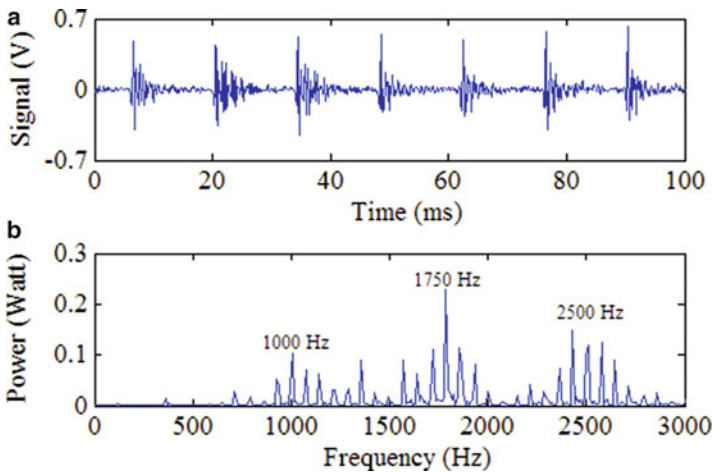


Fig. 5.5 Vibration signal from a ball bearing

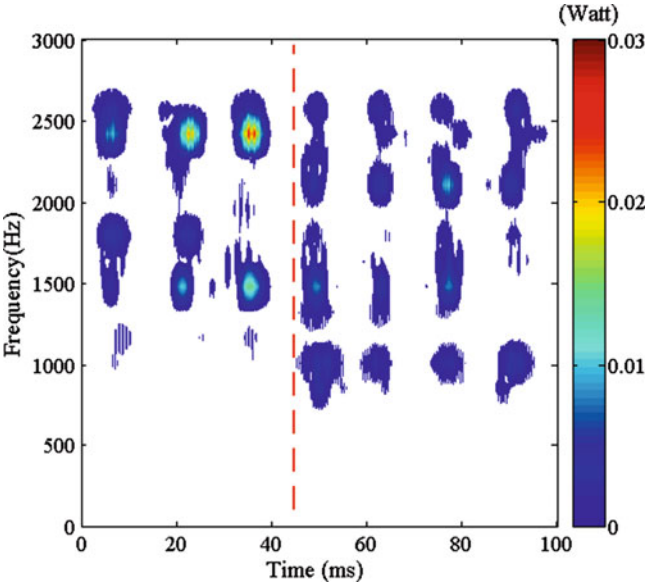


Fig. 5.6 Wavelet packet transform of the bearing vibration signal

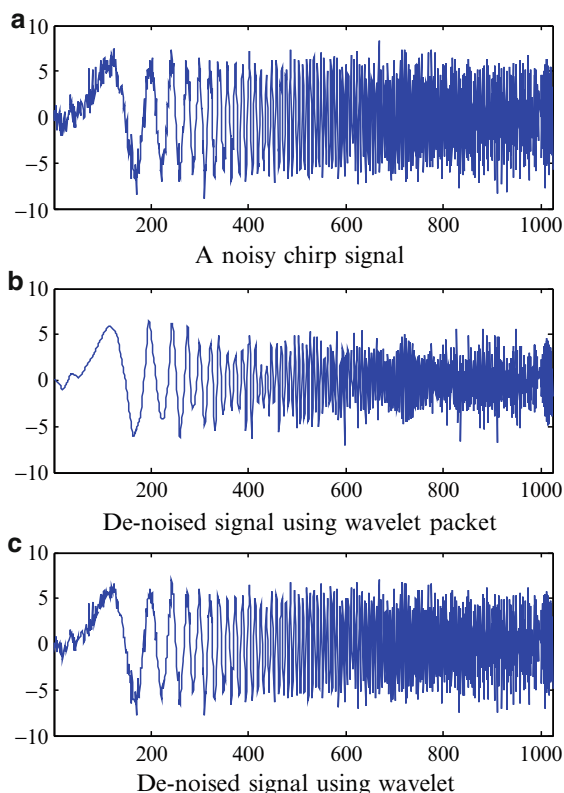
5.4.2 Wavelet Packet for Denoising

Figure 5.7a shows a noisy chirp signal, where Gaussian noise is added to the signal, leading to a signal-to-noise ratio of seven. The denoising idea illustrated here is in principle identical to that developed in the wavelet framework in Chap. 4. The only difference is that the WPT provides better flexibility because of a more complete analysis of the signal. In this example, the Stein’s unbiased estimate of risk (SURE) criterion threshold is used to construct the wavelet coefficients (Donoho 1995; Donoho and Johnstone 1995). For the purpose of comparison, the signal is processed using both the wavelet packets-based denoising and wavelet-based denoising techniques, and the results are shown in Fig. 5.7b, c, respectively. It can be seen that the performance of the wavelet packets-based denoising approach is better than that of the wavelet-based approach.

5.5 Summary

This chapter begins with the introduction of a theoretical basis of a wavelet packet, where the definition of the wavelet packet and its related properties are presented. Two approaches for implementing the WPT are then discussed. Applications of the WPT on time-frequency analysis and denoising are illustrated in Sect. 5.4.

Fig. 5.7 Example of wavelet packet for denoising. (a) A noisy chirp signal, (b) denoised signal using wavelet packet, and (c) denoised signal using wavelet



5.6 References

- Coifman RD et al (1992) Wavelet and signal processing. In: Ruskai (ed) Wavelet and their application. Jones and Bartlett Publishers, Boston, MA
- Coifman RR, Wickerhauser MV (1992) Entropy based algorithms for best basis selection. *IEEE Trans Inform Theory* 38(2):713–718
- Daubechies I (1992) Ten lectures on wavelets. SIAM, Philadelphia, PA
- Donoho DL (1995) De noising by soft thresholding. *IEEE Trans Inform Theory* 41(3):613–627
- Donoho DL, Johnstone IM (1995) Adapting to unknown smoothness via wavelet shrinkage. *J Am Stat Assoc* 90(432):1200–1244
- Gao R, Yan RQ (2006) Non stationary signal processing for bearing health monitoring. *Int J Manuf Res* 1(1):18–40
- Haar A (1910) Zur theorie der orthogonalen funktionensysteme. *Math Annalen* 69:331–371
- Mallat SG (1998) A wavelet tour of signal processing. Academic, San Diego, CA
- Newland DE (1993) Random vibrations, spectral and wavelet analysis. 3rd edn. Addison Wesley Longman, Boston, MA
- Newland DE (1994) Wavelet analysis of vibration part I: theory; part II: wavelet maps. *J Vib Acous* 116(4):409–425

- Samuel PD, Pines DJ, Lewicki DG (2000) A comparison of stationary and non stationary metrics for detecting faults in helicopter gearboxes. *J Am Helicopter Soc* 45:125–136
- Wickerhauser MV (1991) INRIA lectures on wavelet packet transform
- Yan R, Gao R (2005) An efficient approach to machine health evaluation based on harmonic wavelet packet transform. *Robot Comput Integrated Manuf* 21:291–301

Chapter 6

Wavelet-Based Multiscale Enveloping

The use of enveloping technique has been found in many engineering fields. For example, enveloping is employed for the detection of ultrasonic signals, as seen in nondestructive testing (McGonnagle 1966; Greguss 1980; Liang et al 2006). It also presents a complementary tool to spectral analysis in detecting structural defects in rolling bearings (e.g., surface spalling) and gearbox (e.g., broken teeth) (Tse et al 2001; Wang 2001). Generally, three steps are involved in envelope extraction, as illustrated in Fig. 6.1. First, the measured signal passes through a band-pass filter with its bandwidth covering the high-frequency components of interest. As a result, the rest of the frequency components outside of the passing band are rejected, leaving only bursts of the band-passed components in the signal, as shown in Fig. 6.1b. Next, the band-passed signal is rectified, and shown in Fig. 6.1c. Finally, the rectified signal passes through a low-pass filter that is designed to allow only the low-frequency envelope of the signal to pass through, as shown in Fig. 6.1d.

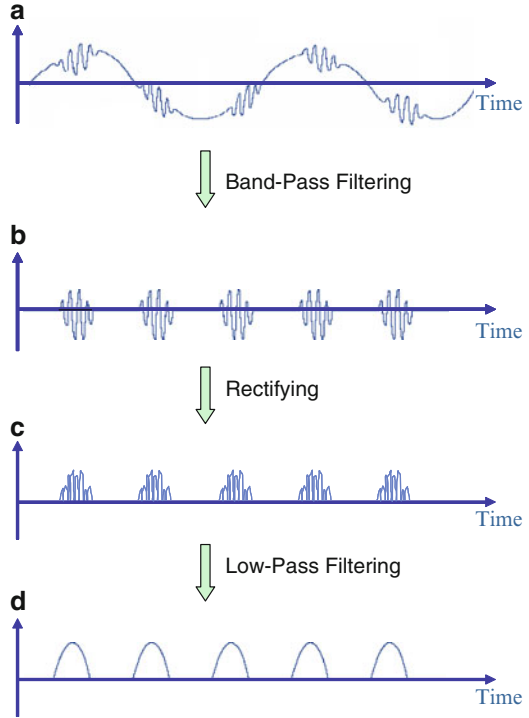
A limitation when applying this technique is that it requires a proper filtering band to be chosen to accurately extract the signal's envelope, for which a priori knowledge of the signal is desired. In this chapter, an adaptive, multiscale enveloping technique based on the wavelet transform is introduced, which overcomes the limitation of the conventional enveloping technique.

6.1 Signal Enveloping Through Hilbert Transform

The Hilbert transform has shown to present a good alternative to the conventional enveloping technique in extracting a signal's envelope (Hahn 1996). Mathematically, the Hilbert transform of a real-valued signal is defined as

$$\tilde{x}(t) = H[x(t)] = \int_{-\infty}^{\infty} \frac{x(\tau)}{\pi(t - \tau)} d\tau \quad (6.1)$$

Fig. 6.1 Procedure for traditional envelope extraction



where $H[\cdot]$ denotes the Hilbert transform operator. The symbol $\tilde{x}(t)$ represents the Hilbert transform result of a real-valued signal $x(t)$, and is the convolution of $x(t)$ and $(1/\pi t)$:

$$\tilde{x}(t) = x(t) \otimes \frac{1}{\pi t} \quad (6.2)$$

where the symbol \otimes denotes the “convolution” operation. According to the convolution theorem, the Fourier transform of the convolution of two signals is the product of the respective Fourier transforms of the two signals (Oppenheim et al. 1999). Accordingly, the Fourier transform of $\tilde{x}(t)$ can be expressed as

$$\tilde{X}(f) = X(f) \times F\left[\frac{1}{\pi t}\right] \quad (6.3)$$

where the symbol \times denotes the “product” operation, $X(f)$ is the Fourier transform of the signal $x(t)$, and $F[1/\pi t]$ denotes the Fourier transform of the term $1/\pi t$. Specifically, this is defined as

$$F\left[\frac{1}{\pi t}\right] = -j \operatorname{sgn} f = \begin{cases} -j & f > 0 \\ 0 & f = 0 \\ j & f < 0 \end{cases} \quad (6.4)$$

Combining (6.4) with (6.3) yields

$$\tilde{X}(f) = \begin{cases} -jX(f) & f > 0 \\ 0 & f = 0 \\ jX(f) & f < 0 \end{cases} \quad (6.5)$$

Through an inverse Fourier transform performed on (6.5), the Hilbert transform of the real-valued signal can be realized. Accordingly, a special type of complex-valued signal $z(t)$ can now be formulated as

$$z(t) = x(t) + j\tilde{x}(t) \quad (6.6)$$

with the real-valued signal $x(t)$ being its real part, and the Hilbert transform of the signal, $\tilde{x}(t)$, being the imaginary part. Because of the inherent linearity property of the Fourier transform, the corresponding expression of (6.6) in the frequency domain can then be given as

$$Z(f) = X(f) + j\tilde{X}(f) \quad (6.7)$$

Combining (6.7) with (6.5) yields

$$Z(f) = X(f) + j \begin{cases} -jX(f) & f > 0 \\ 0 & f = 0 \\ jX(f) & f < 0 \end{cases} = \begin{cases} 2X(f) & f > 0 \\ X(0) & f = 0 \\ 0 & f < 0 \end{cases}. \quad (6.8)$$

Equations (6.6) and (6.8) indicate that the complex-valued signal $z(t)$ is analytic in nature (Lawrence 1999). This means that it can also be expressed in terms of the complex polar coordinates as

$$z(t) = a(t) e^{j\theta(t)} \quad (6.9)$$

where

$$a(t) = \sqrt{x(t)^2 + \tilde{x}(t)^2} \quad (6.10)$$

$$\theta(t) = \tan^{-1} \left(\frac{\tilde{x}(t)}{x(t)} \right) \quad (6.11)$$

Equations (6.10) and (6.11) are called amplitude envelope function and instantaneous phase function of the signal $x(t)$, respectively. This indicates that performing the Hilbert transform on a real-valued signal $x(t)$, results in the formulation of a corresponding analytic signal $z(t)$, from which the envelope $a(t)$ of the signal can be extracted. Such a property of the Hilbert transform makes it well suited for enveloping, as described in the following section.

6.2 Multiscale Enveloping Using Complex-Valued Wavelet

Among various base wavelets commonly used for signal analysis (Lee and Tang 1999; Yen and Lin 2000; Yoshida et al. 2000; Prabhakar et al. 2002; Yan and Gao 2005a), the complex-valued wavelets have the property of being analytic in nature. Such wavelets are generally defined as

$$\psi(t) = \psi_R(t) + j\psi_I(t) = \psi_R(t) + jH[\psi_R(t)] \quad (6.12)$$

where $\psi_R(t)$ and $\psi_I(t)$ represent the real and the imaginary parts of the complex-valued wavelet, respectively, and $\psi_I(t)$ is the Hilbert transform of $\psi_R(t)$.

The wavelet transform $wt_c(s, \tau)$ of a signal $x(t)$ using complex-valued wavelet is expressed as

$$wt_c(s, \tau) = wt_R(s, \tau) + jwt_I(s, \tau) = wt_R(s, \tau) + jH[wt_R(s, \tau)] \quad (6.13)$$

where $wt_R(s, \tau)$ and $wt_I(s, \tau)$ are the real and imaginary parts of the transformation results, respectively. They are defined as

$$\begin{cases} wt_R(s, \tau) = |s|^{-1/2} \int_{-\infty}^{\infty} x(t) \psi_R^*\left(\frac{t-\tau}{s}\right) dt \\ wt_I(s, \tau) = H[wt_R(s, \tau)] = |s|^{-1/2} \int_{-\infty}^{\infty} x(t) H\left[\psi_R^*\left(\frac{t-\tau}{s}\right)\right] dt \end{cases} \quad (6.14)$$

Equations (6.13) and (6.14) indicate that the results of wavelet transform $wt_c(s, \tau)$ of a signal $x(t)$ using the complex-valued wavelet is also analytic. As a result, the signal's envelope at scale s , $env_{wt}(s, \tau)$, can be readily calculated from the modulus of the wavelet coefficients as

$$env_{wt}(s, \tau) = \|wt_c(s, \tau)\| = \sqrt{wt_R(s, \tau)^2 + H[wt_R(s, \tau)]^2} \quad (6.15)$$

As the wavelet transform itself can be considered as a series of band-pass filtering operations (implemented through the scaled parameter s) as described in Chap. 3, and the signal's envelope can be obtained by calculating the modulus of the wavelet coefficients when the complex-valued wavelet is used, a multiscale

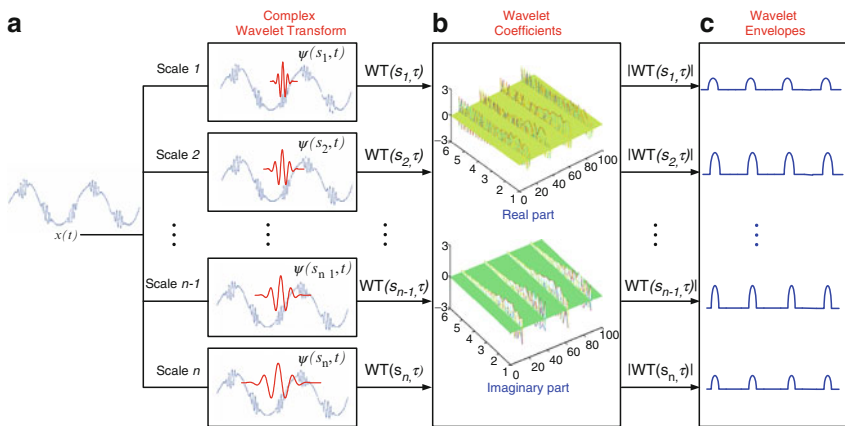


Fig. 6.2 Illustration of the multiscale enveloping algorithm

enveloping technique can be developed on the basis of the wavelet transform. Computationally, this technique first decomposes the signal (e.g., vibrations measured on a defective rolling bearing) into different wavelet scales by means of a complex-valued wavelet transform, as illustrated in Fig. 6.2a. A series of wavelet coefficients, which are expressed as real part and imaginary part, respectively, are then obtained (Fig. 6.2b). The envelope signal in each scale (Fig. 6.2c) is finally calculated from the modulus of the wavelet coefficients.

6.3 Application of Multiscale Enveloping

This section describes the application of the multiscale enveloping technique introduced above to two different mechanical systems.

6.3.1 Ultrasonic Pulse Differentiation for Pressure Measurement in Injection Molding

Online monitoring and control of pressure in the cavity of an injection machine has been shown to be critically important for improving product quality while maintaining low rejection rates in injection molding (Rawabdeh and Petersen 1999). The design of a self-powered wireless sensor has enabled the placement of multiple sensors within a mold to achieve comprehensive spatial coverage of the cavity pressure profile (Gao et al. 2001; Theurer et al. 2001). To overcome electromagnetic shielding caused by the steel mold that surrounds the sensors, ultrasonic wave has been explored as an alternative to electromagnetic wave for pressure data

transmission out of the mold (Zhang et al. 2004). Specifically, mold cavity pressure measured by a piezoceramic sensing element is digitized into a series of ultrasonic pulse trains, with each pulse train representing the crossing of a preset pressure threshold. The actual cavity pressure (denoted as ❶ in Fig. 6.3a) is reconstructed by multiplying the total number of the pulse trains (denoted as ❸ in Fig. 6.3a) with the known threshold value. Given a matrix arrangement of such wireless sensors within the mold cavity, spatial coverage of the cavity pressure profile can be obtained. An example of a sensor matrix consisting of six wireless sensors and a single receiver is illustrated in Fig. 6.3b.

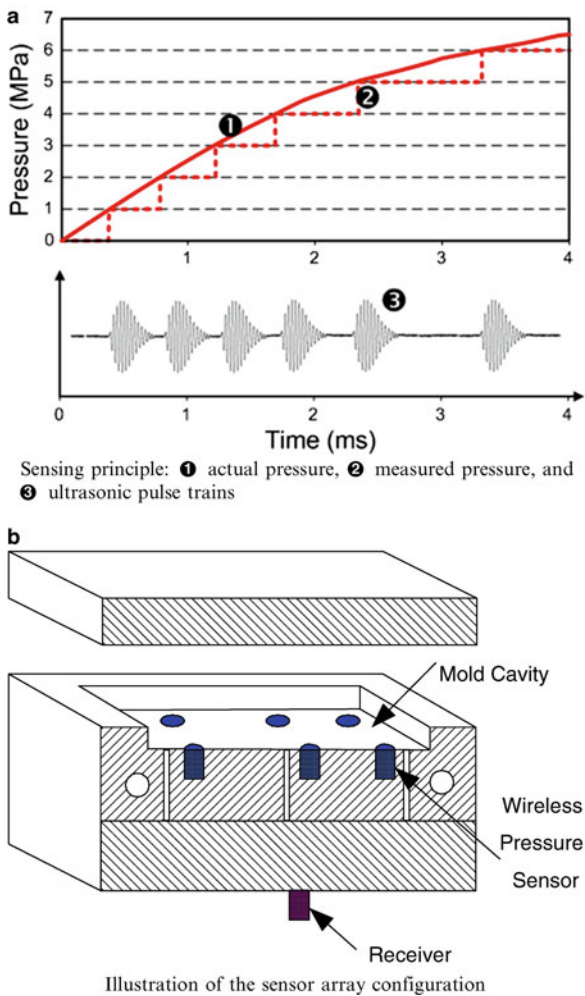


Fig. 6.3 Sensing principle and the sensor array arrangement in an injection mold. (a) Sensing principle: ❶ actual pressure, ❷ measured pressure, and ❸ ultrasonic pulse trains and (b) illustration of the sensor array configuration

The limitation of such a 1D enveloping technique is illustrated in Fig. 6.4, which illustrates a total of six ultrasonic pulse trains generated by six transmitters, with the center frequencies being 2,210, 2,480, 2,785, 3,140, 3,530, and 3,980 kHz, respectively. Each pulse train is related to the crossing of the melt pressure of a threshold level at a specific location along the cavity. The envelope of the signal is given in Fig. 6.4b. By thresholding the enveloped signal, the time of arrival of each pulse train can be determined. However, as the difference in frequency of the pulse trains cannot be accurately resolved, the spectrum of the multiple pulse trains appears in Fig. 6.4c as a lumped group, giving the appearance as if they were generated by a single transmitter.

Such a problem can be solved by using the multiscale enveloping technique introduced in this chapter, which decomposes the pulse trains into individual frequency subbands and extracts the respective envelope from the pulse trains in each subband. Multiplying the number of crossings by the envelope with each

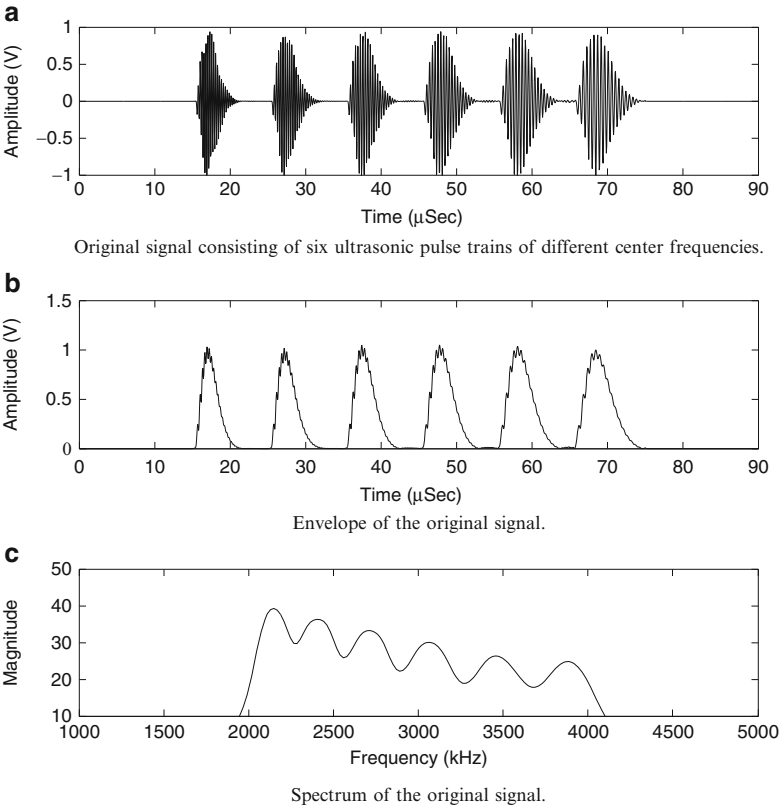


Fig. 6.4 Limitation of 1D enveloping technique. (a) Original signal consisting of six ultrasonic pulse trains of different center frequencies, (b) envelope of the original signal, and (c) spectrum of the original signal

respective threshold value, the cavity pressure profile can be reconstructed. This is illustrated in the following sections, through both simulation and experiments.

6.3.1.1 Simulation

The performance of the multiscale enveloping for pulse detection on a sensor matrix consisting of six spatially distributed ultrasonic transmitters is evaluated first by means of a computer simulation. The six spectrally adjacent ultrasonic pulse trains are centered at 2,210, 2,480, 2,785, 3,140, 3,530, and 3,980 kHz, respectively, labeled as ① through ⑥ in Fig. 6.5. The pulses are separated by an interval of 10 μs from one another, simulating the flow of polymer melt over the sensor matrix sequentially at a constant speed. As shown in Fig. 6.5, the six pulses could be detected and well separated into six levels (each level corresponds to a specific scale calculated on the basis of ultrasonic pulse center frequency) through a wavelet-based multiscale enveloping process.

In another simulation, the multiscale enveloping technique is applied to decomposing an ultrasonic signal consisting of two different types of ultrasound pulse trains:

1. spectrally identical (with the same center frequency of 3,980 kHz) and timely adjacent (5 μs apart), as labeled ① and ② in Fig. 6.6a
2. timely overlapped and spectrally adjacent (with center frequencies of 2,210 and 2,785 kHz, respectively) as labeled ③ and ④ in Fig. 6.6a.

As shown in Fig. 6.6b, pulses ① and ② are successfully differentiated both spectrally (at the same level 6, because of their identical center frequency) and

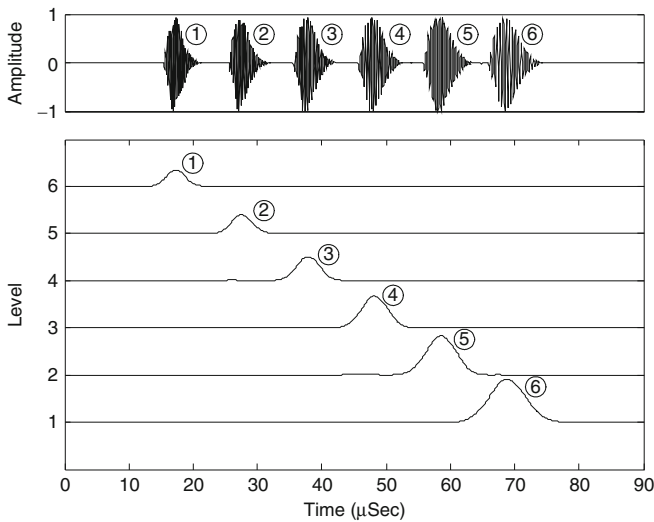


Fig. 6.5 Detection and differentiation of six spectrally adjacent pulse trains

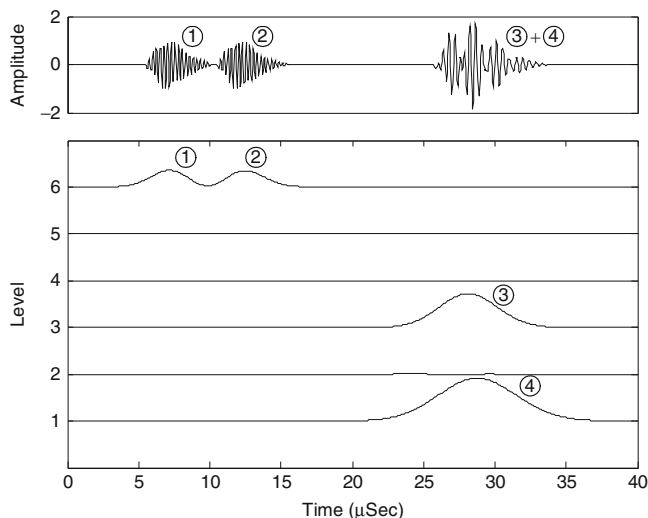


Fig. 6.6 Detection and differentiation of two pulse trains that are timely overlapped and spectrally adjacent and two pulse trains that are timely adjacent but spectrally identical

temporally (successive along the time axis, with 5 μ s separation). Similarly, the two pulses ③ and ④ are well separated into the first and third levels, reflecting on the different center frequencies that they contain.

6.3.1.2 Experimental Study

To experimentally verify the performance of the developed multiscale enveloping technique for ultrasonic pulse detection, three ultrasonic transmitters were designed and prototyped, with the center frequencies being 2,480, 2,785, and 3,140 kHz, respectively. An electrical pulser (model C-101-HV from PAC company) was used to electrically excite the transmitters. Ultrasonic pulses generated were then transmitted through a steel block of 6 cm thickness, which represents a realistic injection mold. The pulses were received by an ultrasonic receiver located on the opposite side of the steel block. The received ultrasonic pulses were measured and recorded using a digital oscilloscope (model TDS 3012B from Tektronix).

In the first experiment, a single transmitter (center frequency 3,140 kHz) was excited repetitively at 10 kHz. As a result, a series of pulses were generated with two adjacent pulses being timely separated by 100 μ s, as shown in Fig. 6.7a. For each train of pulses generated (by each excitation), the first arrived pulse with the highest amplitude, plus two reflections with decaying amplitudes were clearly observed. The received pulses were processed using the multiscale enveloping technique, and their corresponding envelopes were extracted. As shown in Fig. 6.7a, the first arrival and the first two reflections were clearly differentiated at level 4. As the reflections have much lower amplitude than the first arrival, they

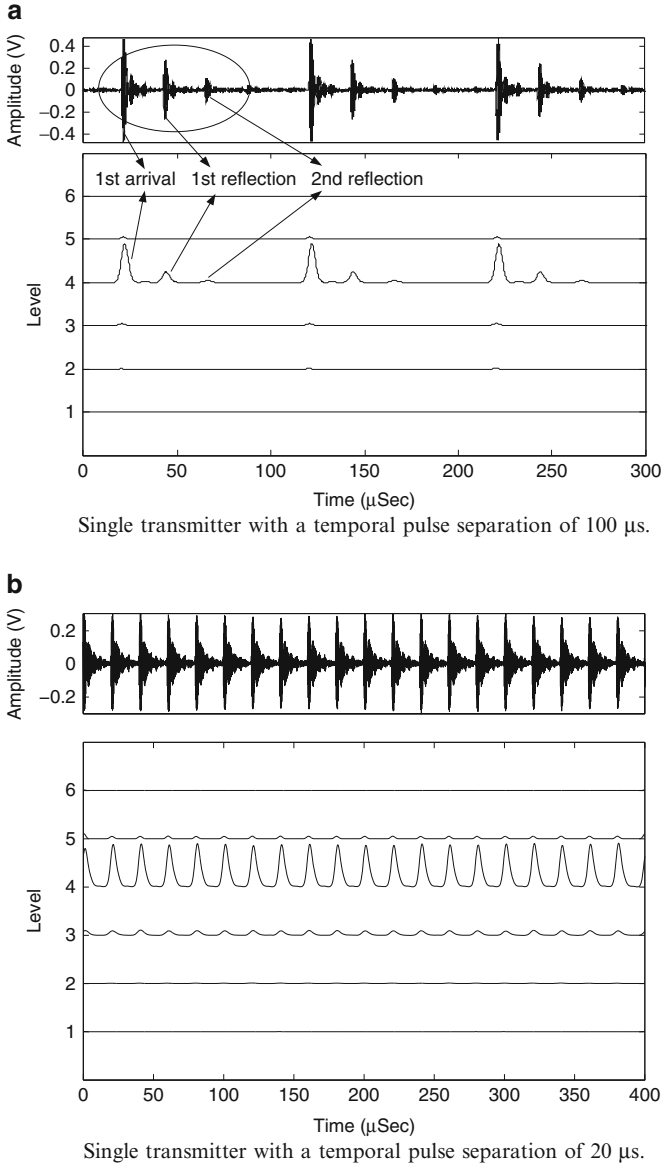


Fig. 6.7 Experimental detection of ultrasonic pulse trains having the same center frequency. (a) Single transmitter with a temporal pulse separation of 100 μs and (b) single transmitter with a temporal pulse separation of 20 μs

can be readily eliminated through thresholding from the extracted envelope. In the second experiment, the pulse repetition frequency was increased to 50 kHz, resulting in a temporal separation of 20 μs between the adjacent pulses. As shown in

Fig. 6.7b, the reflections were buried under the first arrivals; therefore, they did not affect the pulse detection.

To evaluate the pulse detector's ability in differentiating spectrally adjacent pulses in the frequency domain, the three transmitters (of center frequencies 2,480, 2,785, and 3,140 kHz) were placed side-by-side on one side of the steel block and excited simultaneously, with the excitation repetition frequency being 30 kHz (corresponding to 33 μ s pulse separation). The pulses received by the ultrasonic receiver are shown in the upper portion of Fig. 6.8a, where temporal overlap of the three transmitters cannot be differentiated in the time domain. Applying the multiscale enveloping technique, the envelopes of the three pulse trains were successfully extracted and differentiated in levels 2, 3, and 4, respectively, as shown in Fig. 6.8b.

To examine the robustness of the multiscale enveloping technique, repetition frequency of the excitation input to the transmitters was varied to be 30, 20, and 10 kHz for the three transmitters, resulting in a pulse separation of 33, 50, and 100 μ s, respectively. As shown in Fig. 6.8b, the pulse trains were again successfully detected and differentiated, with the corresponding envelopes separated into levels 2, 3, and 4, respectively.

6.3.2 *Bearing Defect Diagnosis in Rotary Machine*

A large number of applications in machine condition monitoring involve rotary machine components, for example, bearings, spindles, and gearboxes (Kiral and Karagülle 2003; Wu et al. 2004; Choy et al. 2005). To detect structural defects that may occur in these machine components, spectral analysis of the signal's envelope has been widely employed (McFadden and Smith 1984; Ho and Randall 2000). This is based on the consideration that structural impacts induced by a localized defect often excite one or more resonance modes of the structure and generate impulsive vibrations in a repetitive and periodic way. Frequencies related to such resonance modes are often located in higher frequency regions than those caused by machine-borne vibrations, and are characterized by an energy concentration within a relatively narrow band centered at one of the harmonics of the resonance frequency. By utilizing the effect of mechanical amplification provided by structural resonances, defect-induced vibration features can be separated from the background noise and interference for diagnosis purpose. However, as different resonance modes can be excited under varying machine operating conditions, consistent results are not guaranteed by simply applying the traditional enveloping spectral analysis. Research has found that complementing the wavelet-based multiscale enveloping with spectral analysis by means of the multiscale enveloping spectrogram (MuSenS) technique could significantly enhance the effectiveness of bearing defect diagnosis (Yan and Gao 2005b). Basically, the MuSenS starts with a signal's envelope extraction by using the developed wavelet-based multiscale enveloping technique; Fourier transform is then performed repetitively on the extracted

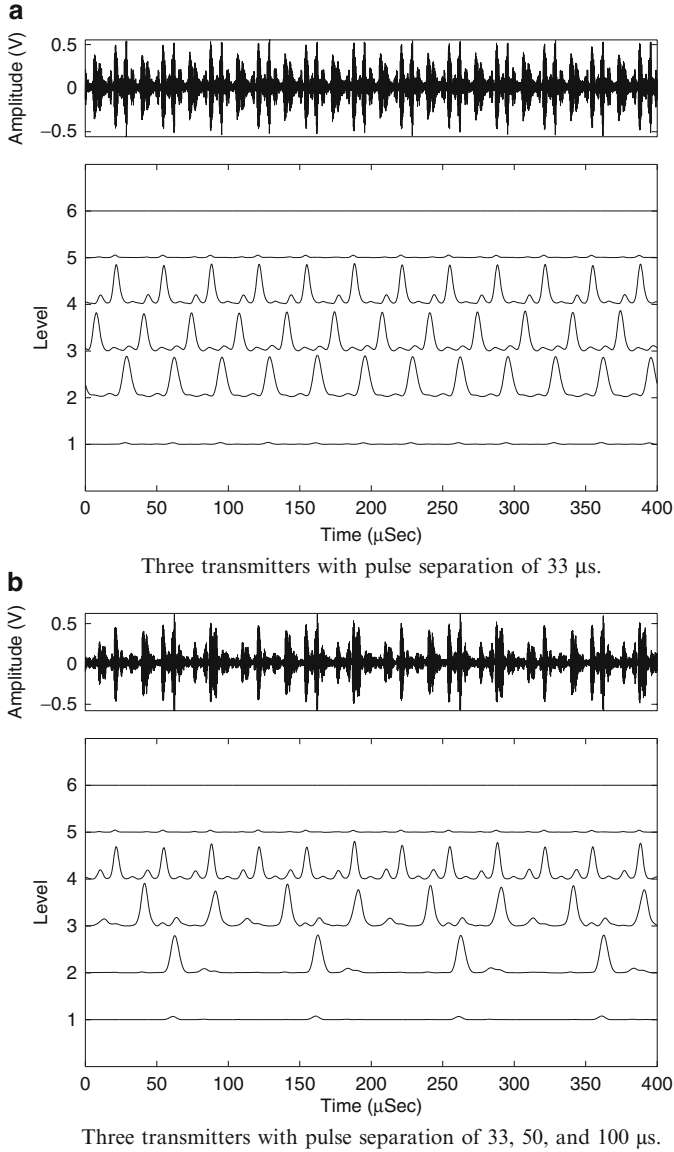


Fig. 6.8 Experimental detection and differentiation of temporally overlapped and spectrally adjacent ultrasonic pulse trains generated by three transmitters. (a) Three transmitters with pulse separation of 33 μs and (b) three transmitters with pulse separation of 33, 50, and 100 μs

envelope signal $env_{wt}(s, \tau)$ at each scale s , resulting in an “envelop spectrum” of the original signal at the various scales. Such envelop spectra can be expressed as

$$ENV_{wt}(s, f) = F[env_{wt}(s, \tau)] = \frac{1}{2\pi} \int_{-\infty}^{\infty} \|wt_c(s, \tau)\| e^{i2\pi f\tau} d\tau \quad (6.16)$$

where the envelope signal $env_{wt}(s, \tau)$ is obtained using (6.15), and calculated directly from the modulus of the wavelet coefficients $\|wt_c(s, \tau)\|$ of the original signal. The result of the $ENV_{wt}(s, f)$ operation is a 2D matrix, with each of its rows corresponding to the envelop spectrum of the vibration signal at a specified scale s , and each of its columns corresponding to a specific frequency component of the envelope spectrum across all the scales. By looking at the square of the magnitude of $ENV_{wt}(s, f)$, as seen in

$$E(s, f) = |ENV_{wt}(s, f)|^2 \quad (6.17)$$

which is termed as the energy spectrum, the final output $E(s, f)$ of the MuSEnS is obtained. The energy spectrum indicates how the energy content is distributed in the scale-frequency plane. For the purpose of visualization, such a result can be illustrated in a 3D scale-frequency-energy map, which can indicate the intensity and location of the defect-related frequency lines. The applications of the MuSEnS technique to bearing defect diagnosis are introduced in the next section.

6.3.2.1 Numerical Simulation Using the MuSEnS Algorithm

A synthetic signal that consists of different signal components for simulating vibration signal from the rolling element bearing is first constructed to quantitatively evaluate the MuSEnS technique. Generally, vibration signals from a bearing may include the following constituent components:

1. vibration caused by bearing imbalance with a characteristic frequency of f_u , equal to the bearing rotational speed, which occurs when the gravitational center of the bearing does not coincide with its rotational center
2. vibration caused by bearing misalignment at frequency f_m , equal to twice the shaft speed, which occurs when the two raceways of the bearing (inner and outer) fall out of the same plane, resulting in a raceway axis that is no longer parallel to the axis of the rotating shaft
3. vibration due to rolling elements periodically passing over a fixed reference position on the outer raceway, at the frequency f_{BPFO}
4. structure-borne vibration attributed by other components, which is broadband in nature, and can be modeled as white noise.

When a localized structural defect occurs on the surface of the bearing raceways (inner or outer), a series of impacts will be generated every time the rolling elements interact with the defects, subsequently exciting the bearing system. Such forced vibration is represented by high-frequency resonances that are amplitude modulated at the repetition frequency of the impacts.

For the numerical simulation, only defect-induced resonant vibration and structure-borne vibration are considered in the synthetic signal, as other vibration components can be filtered out through data preprocessing. The simulated resonant

vibration is obtained experimentally from the measured impulse response of a ball bearing (model 2214). This bearing has 17 rolling elements. When it rotates at 300 rpm, a total of eight impacts will be generated per bearing revolution, because of the ball defect interactions. This translates into an impact interval of 25 ms or a 40-Hz signal repetition frequency. Figure 6.9a illustrates such a series of impact-related vibrations. By adding white noise to these vibrations, a synthetic signal is then generated to simulate the actual bearing vibrations due to a localized outer raceway defect. The signal-to-noise ratio (SNR) of the synthetic system is set at -12 dB. The synthetic signal with its time and frequency domain waveform is shown in Fig. 6.9b, c. Because of the noise corruption, no apparent signal feature could be identified, except for the relatively dominant spectral components ranging from 2,500 to 3,500 Hz.

The synthetic signal is analyzed using the wavelet-based MuSenS technique, where the complex Morlet wavelet is used as the base wavelet for defect characteristic extraction. A series of equally spaced scales ranging from 1 to 6 (with an increment of s_1) were chosen to stretch the complex Morlet wavelet for extracting the defect-related feature embedded in the synthetic signal. The lower and upper limits of the scales correspond to the wavelet center frequency at 10,000 and

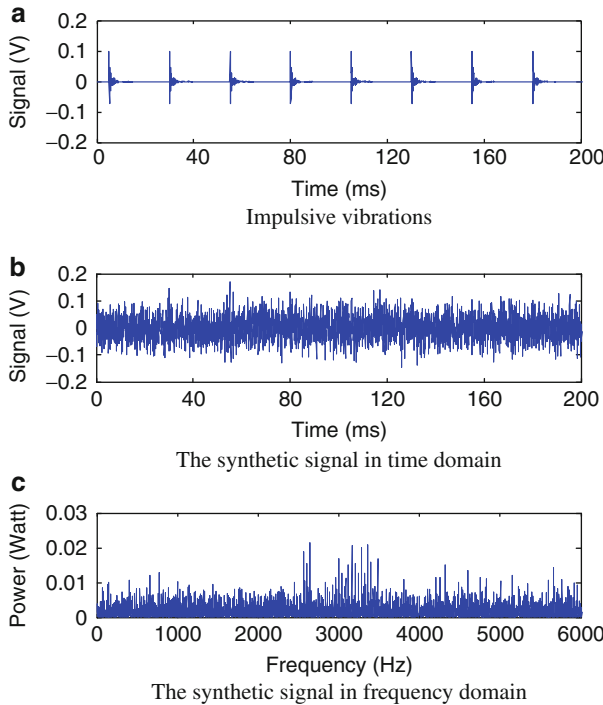


Fig. 6.9 The series of impulsive vibrations, the synthetic signal (signal to noise ratio (SNR) 12 dB), and its spectrum. (a) Impulsive vibrations, (b) the synthetic signal in time domain, and (c) the synthetic signal in frequency domain

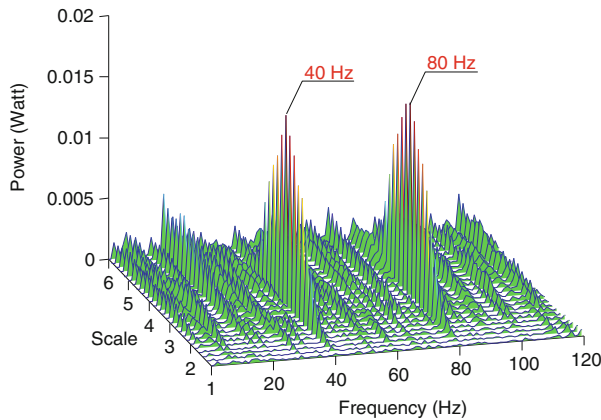


Fig. 6.10 Defect repetition frequency detection on the synthetic signal using multiscale enveloping spectrogram (MuSEnS) technique

1,667 Hz, respectively. This ensures that the defect-induced resonant vibration component can be fully covered by the wavelet transformation. To increase the possibility of matching the center frequency of a scaled wavelet with the frequency of the defect-induced resonant vibration, a small-scale interval is preferred. However, a small-scale interval leads to increased computational load, as more scales will be involved in the signal decomposition. A trade-off must therefore be made between the accuracy and computational time. On the basis of preliminary studies, an increment of $s_1 = 0.2$ was employed in this study. As the MuSEnS shown in Fig. 6.10, high-energy concentration can be identified at the 40-Hz frequency line, which corresponds to the defect-related repetition frequency. Also strongly represented in the spectrogram is the harmonics of the defect-related frequency at 80 Hz. This result demonstrates the effectiveness of the *MuSEnS* algorithm in identifying defect features hidden in bearing vibration signals.

6.3.2.2 Case Study

The first experimental case study of using *MuSEnS* algorithm to diagnose bearing defects is conducted on a roller bearing. A seeded defect in the form of 0.1 mm diameter hole is made in the outer raceway. The bearing is subject to a 3,665-N radial load, and the shaft rotational speed is 1,200 rpm (or a 20-Hz rotational frequency). On the basis of the geometrical parameters of the bearing and the rotational speed, a defect-related repetitive frequency of ($f_{\text{BPFO}} = 5.25f_{\text{rpm}}$) or 105 Hz can be analytically determined (Harris 1991). Figure 6.11 shows the bearing vibration signal acquired under the sampling frequency of 25 kHz. From its corresponding power spectrum, it is evident that frequency component related to bearing rotation is dominant in the frequency region of [0, 150] Hz. However,

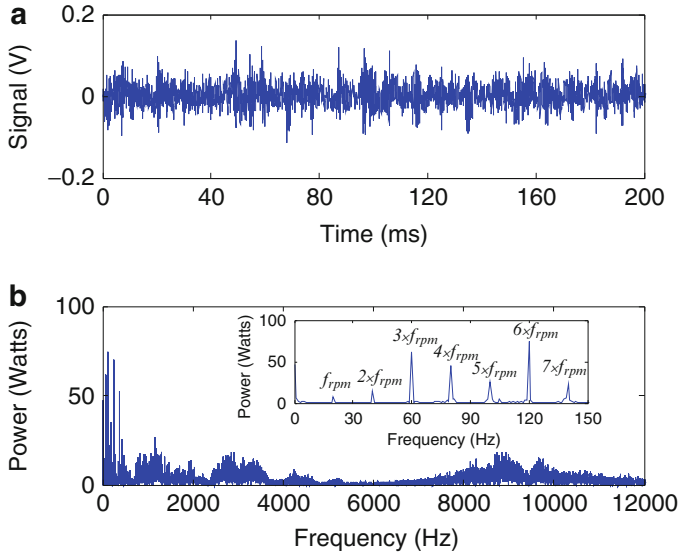


Fig. 6.11 Signal measured from a roller bearing and its power spectrum

defect-related frequency component of 105 Hz is submerged in the spectrum and therefore, cannot be identified.

The *MuSEnS* algorithm is then applied to decompose the bearing signal. The scales used are between 1 and 8, with an increment 0.2. These scales cover the frequency range of 1.56–12.5 kHz. The corresponding *MuSEnS* of the bearing vibration signal is shown in Fig. 6.12. Two major peaks are clearly shown at 20 and 105 Hz frequency lines, respectively. The 20-Hz component runs across the entire scale region, and is related to the bearing rotating speed. The 105-Hz component is identified at the scales of 1–2.4, and represents the repetitive frequency of the bearing due to the structural defect on the outer raceway. This demonstrates that the *MuSEnS* is able to clearly identify the existence of the structural defect, and pinpoint its location on the outer raceway for diagnosis purpose.

The second experimental case study of using the *MuSEnS* algorithm for diagnosis of bearing inner raceway defect is investigated on a ball bearing of model SKF 6220. The defect-related repetitive frequency is calculated to be ($f_{BPEI} = 5.9f_{rpm}$) or 59 Hz, based on the bearing geometry and rotational speed (600 rpm) (Harris 1991). A radial load of 10,000 N is applied to the bearing. As shown in Fig. 6.13, while frequency components related to the shaft speed and ball rotation are shown in the power spectrum, the defect-related repetitive frequency is not identified.

The *MuSEnS* algorithm is then applied to the same signal, with the decomposition scales chosen to be ~ 2 –10 at an increment of 0.2. The scales cover the frequency range from 500 to 2,500 Hz. As shown in Fig. 6.14, besides frequency components related to the shaft frequency and its harmonics, an appreciable peak can be seen at 59 Hz, which is the inner raceway defect-related repetitive frequency. This indicates that a structural defect exists on the inner raceway of the ball bearing. The peaks at 49 and 69 Hz frequency lines are attributed to the combined effect of

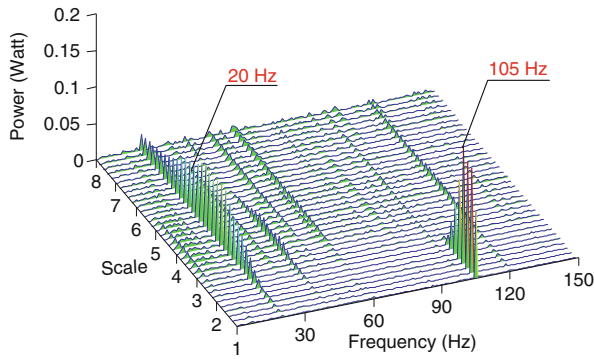


Fig. 6.12 MuSenS of vibration signals measured on a roller bearing with a structural defect on the outer raceway (speed: 1,200 rpm; radial load: 3,665 N)

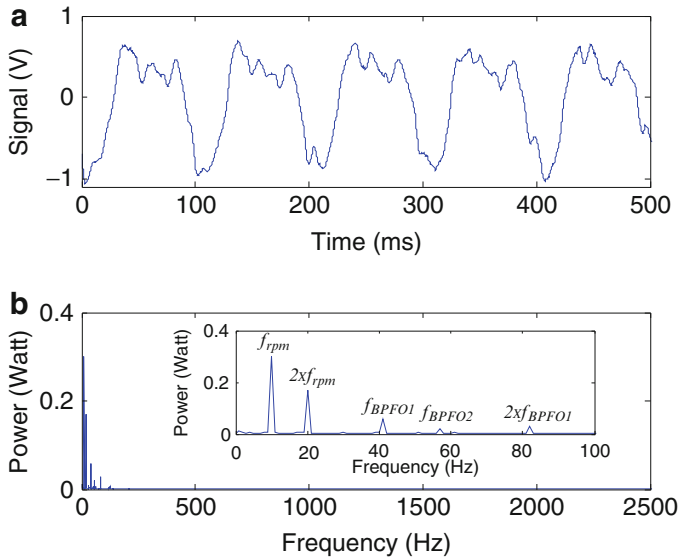


Fig. 6.13 Signal measured from a ball bearing and its power spectrum

bearing imbalance at 10 Hz frequency and the structural defect at 59 Hz frequency, as they can be calculated as 59 ± 10 Hz.

6.4 Summary

A wavelet-based multiscale enveloping technique is introduced in this chapter. This multidomain signal processing technique combines band-pass filtering (implemented through variation of the scale parameter s of the base wavelet) and enveloping

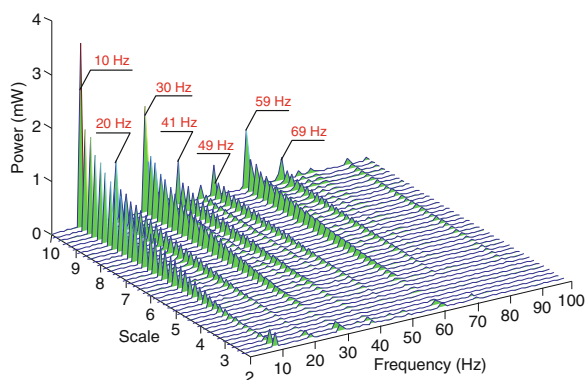


Fig. 6.14 MuSenS of vibration signals measured on a ball bearing with a structural defect on the inner raceway (speed: 600 rpm; radial load: 10,000 N)

(obtained through modulus of the wavelet coefficients) into a *single-step* operation. The effectiveness of the multiscale enveloping technique is demonstrated through studies in ultrasonic pulse differentiation for pressure measurement in injection molding and bearing defect diagnosis in rotary machines, both numerically and experimentally.

When the multiscale enveloping technique is used to identify the ultrasonic pulses generated from injection molding process, not only spectrally identical and timely adjacent but also timely overlapped and spectrally adjacent ultrasonic pulses could be detected and differentiated. This allows for comprehensive spatial coverage of the cavity pressure profile by placing multiple sensors with different working frequency ranges at different locations in the injection mold. When the wavelet-based enveloping is combined with spectral domain postprocessing, a new algorithm termed MuSenS was developed, which has been shown to be more accurate and illustrative in depicting critical features related to structural defects embedded in the bearings than the traditional enveloping spectral analysis. As many of the applications in manufacturing equipment and system monitoring involve rotary machine components (e.g., bearings, spindles, gearboxes, etc.), it is possible that the MuSenS technique may contribute to improving solutions to a wide variety of machine monitoring problems.

6.5 References

- Choy FK, Zhou J, Braun MJ, Wang L (2005) Vibration monitoring and damage quantification of faulty ball bearings. *ASME J Tribol* 127(3):776–783

- Gao R, Kazmer D, Theurer C, Zhang L (2001) Fundamental aspects for the design of a self energized sensor for injection molding process monitoring. In: Proceedings of NSF design and manufacturing research conference, Tempa, FL
- Greguss P (1980) Ultrasonic imaging: seeing by sound. The principles and widespread applications of image formation by sonic, ultrasonic, and other mechanical waves. Focal Press, New York
- Hahn SL (1996) Hilbert transform in signal processing. Artech House Inc., Norwood, MA
- Harris TA (1991) Rolling bearing analysis, 3rd edn. Wiley, New York
- Ho D, Randall RB (2000) Optimization of bearing diagnostic techniques using simulated and actual bearing fault signals. *Mech Syst Signal Process*, 14(5):763–788
- Kiral Z, Karagülle H (2003) Simulation and analysis of vibration signals generated by rolling element bearing with defects. *Tribol Int* 36:667–678
- Lawrence MS (1999) Computing the discrete time analytic signal via FFT. *IEEE Trans Signal Process* 47(9):2600–2603
- Lee BY, Tang YS (1999) Application of the discrete wavelet transform to the monitoring of tool failure in end milling using the spindle motor current. *Int J Adv Manuf Technol* 15(4):238–243
- Liang W, Que PW, Yang G (2006) Ultrasonic flaw detection during NDE of oil pipelines via a resonance filter. *Russ J Nondestruct Test* 42(6):398–403
- McFadden P, Smith J (1984) Vibration monitoring of rolling element bearings by the high frequency resonance technique – a review. *Tribol Int* 17(1):3–10
- McGonnagle WJ (1966) Nondestructive testing, 2nd edn. Gordon and Breach Science Publisher, New York
- Oppenheim AV, Schafer RW, Buck JR (1999) Discrete time signal processing, 2nd edn. Prentice Hall, Upper Saddle River, NJ
- Prabhakar S, Mohanty AR, Sekhar, AS (2002) Application of discrete wavelet transform for detection of ball bearing race faults. *Tribol Int* 35:793–800
- Rawabdeh IA, Petersen PF (1999) In line monitoring of injection molding operations: a literature review. *Injection Molding Technol* 3:47–53
- Theurer C, Zhang L, Gao R, Kazmer D (2001) Acoustic telemetry in injection molding. In: Proceedings of society of plastics engineers annual technical conference, process monitoring and control division, vol 51. Dallas, TX, pp 208–213
- Tse PT, Peng YH, Yam R (2001) Wavelet analysis and envelope detection for rolling element bearing fault diagnosis – their effectiveness and flexibilities. *ASME J Vib Acoust* 123(4):303–310
- Wang W (2001) Early detection of gear tooth cracking using the resonance demodulation technique. *Mech Syst Signal Process* 15(5):887–903
- Wu JD, Huang CW, Huang R (2004) An application of a recursive Kalman filtering algorithm in rotating machinery fault diagnosis. *NDT&E Int* 37:411–419
- Yan R, Gao R (2005a) An efficient approach to machine health evaluation based on harmonic wavelet packet transform. *Robot Comput Integrated Manuf* 21:291–301
- Yan R, Gao R (2005b) Multi scale enveloping spectrogram for bearing defect detection. *World tribology congress III*, Washington, DC, pp 855–856
- Yen G, Lin K (2000) Wavelet packet feature extraction for vibration monitoring. *IEEE Trans Ind Electron* 47(3):650–667
- Yoshida A, Ohue Y, Ishikawa H (2000) Diagnosis of tooth surface failure by wavelet transform of dynamic characteristics. *Tribol Int* 33:273–279
- Zhang L, Theurer C, Gao R, Kazmer D (2004) A self energized sensor for wireless injection mold cavity pressure measurement: design and evaluation. *ASME J Dyn Syst Meas Control* 126(2):309–318

Chapter 7

Wavelet Integrated with Fourier Transform: A Unified Technique

Fourier transform-based spectral analysis has been widely applied to processing signals, such as vibration and acoustic signals (Mori et al. 1996; Tandon and Choudhury 1999; Cavacece and Introini 2002), acquired from manufacturing systems. Because of noise contamination and signal interference, the constituent components of interest may be submerged in the signal and difficult to be revealed through a spectral analysis (Ho and Randall 2000). Furthermore, events occurred in the manufacturing system may be transient in nature, for example, the initiation and propagation of surface spalling in a ball bearing (Gao and Yan 2006; Orhan et al. 2006). As another example, the process of metal removal can be viewed as consisting of multiple, individual transient events in which a single chip of metal is removed (Ge et al. 2004; Obikawa and Shinozuka 2004; Byrne and O'Donnell 2007; Malekian et al. 2009). On the one hand, given the global analysis nature, it is difficult to apply the Fourier transform for localizing these transient events. On the other hand, the Fourier transform can identify a signal's frequency components, from which a specific event (e.g., localized bearing defects, which have distinct characteristic frequencies at inner raceway, outer raceway, or rolling element itself) can be detected. Leveraging the capability of wavelet transform in transient signal analysis, this chapter introduces a *unified* time scale frequency analysis technique through spectral postprocessing on the data set extracted by wavelet transforms to enhance the effectiveness of signal representation and identification.

7.1 Generalized Signal Transformation Frame

Fourier transform and wavelet transform were originated from different theoretical platforms, and each technique analyzes a signal from a different perspective. Specifically, the Fourier transform depicts the energy concentration of constituent frequency components within the signal, whereas the wavelet transform presents the similarity between the signal being analyzed and the base wavelet, in the time scale domain. To enable cross-domain unification of the two techniques for signal analysis, a common signal transformation platform needs to be established first, which is the focus of this chapter.

Let us first define a function $W_{1,0}(t)$ within a certain time interval, or *support*, expressed as $[0, L)$, where the symbol L represents the width of support. The function $W_{1,0}(t)$ is called the *base template* function for signal analysis. Next, we define $W_{s,u}(t)$, which is a derived version of $W_{1,0}(t)$. Comparing to $W_{1,0}(t)$, the magnitude of $W_{s,u}(t)$ is *scaled* by *scale* s , where $s \geq 0$ is an integer number, and its location along the time axis is *shifted* by *time* u , with $u \in R$, and R represents the set of all real numbers. The function $W_{s,u}(t)$ is called the *derived template* function, at scale s and time u , and is supported within the time interval $[u, u + sL]$. Generally, $W_{s,u}(t)$ can be expressed in terms of the *base template* function $W_{1,0}(t)$ as:

$$W_{s,u}(t) = \frac{1}{\sqrt{s}} W_{1,0}\left(\frac{t-u}{s}\right) \quad (7.1)$$

where $1/\sqrt{s}$ is a factor for purpose of normalization. Specifically, it ensures that the following relationship between the *derived template* function $W_{s,u}(t)$ and the *base template* function $W_{1,0}(t)$ is always satisfied:

$$\int_{-\infty}^{\infty} W_{s,u}^2(t) dt = \int_{-\infty}^{\infty} W_{1,0}^2(t) dt \equiv \|W_{1,0}(t)\|^2 \quad (7.2)$$

The physical significance of (7.2) is that all the derived template functions preserve the same amount of energy as the base template function does.

In a linear signal space, the set of all *derived template* functions $\{W_{s,u}(t): s \geq 0, u \in R\}$ forms a *continuous frame* Γ_c , spanned by scale s and time u . In accordance with the discrete data sampling process, where data points are taken at time instances $u = mkL$ from a derived template function with a scale factor of $s = k$, a discrete expression of the derived template function is obtained as:

$$W_{k,m} = \frac{1}{\sqrt{k}} W_{1,0}\left(\frac{t - mkL}{k}\right) \quad (7.3)$$

where $W_{k,m}(t)$ is a simplified expression for $W_{k,mkL}(t)$. In the above notation, k or $k^{-1} \in N$, $m \in Z$, and k and m represent a discrete version of the continuous parameters s and u . The notation $k^{-1} \in N$ corresponds to $s < 1$. The set $\{W_{k,m}(t): k \text{ or } k^{-1} \in N, m \in Z\}$, with N being the set of all nonnegative integers and Z being the set of all integers, forms a *discrete frame* Γ_d , spanned by the parameters k and m . The *continuous frame* Γ_c (or *discrete frame* Γ_d) provides a *generalized frame* for signal transformation, and is defined as *complete* in the linear signal space, if any signal function $x(t)$ can be expressed in it as (Kaiser 1994):

$$x(t) = \int_0^\infty \int_{-\infty}^\infty C(s,u) W_{s,u}(t) ds du \quad (7.4)$$

or, in the case of *discrete frame* Γ_d :

$$x(t) = \sum_{k=1}^{\infty} \sum_{m=-\infty}^{\infty} C(k, m) W_{k,m}(t) \quad (7.5)$$

In (7.4) and (7.5), the coefficient of the functions, $C(s, u)$ or $C(k, m)$, can be considered as a *measure* function, which expresses the extent to which the signal function $x(t)$ is correlated to the *derived template* function $\{W_{s,u}(t): s \geq 0, u \in R\}$ of *scale* s , at the specific *time* u .

Under the discrete frame Γ_d , the significance of the *measure* function expressed in (7.5) can be further illustrated, considering that there exists a complete *orthogonal* set $\{W_{k,m}(t)\}$ within such a frame. The *orthogonal* identity states that:

$$\int_{-\infty}^{\infty} W_{k_1,m_1}(t) W_{k_2,m_2}(t) dt = \begin{cases} \int_{-\infty}^{\infty} W_{k_1,m_1}^2(t) dt; & \text{for } k_2 = k_1, m_2 = m_1 \\ 0; & \text{otherwise} \end{cases} \quad (7.6)$$

where k_1 and $k_2 \in \{k\}$, and m_1 and $m_2 \in \{m\}$. Using the identity of (7.6), multiplying the two sides of (7.5) by $W_{k_1,m_1}(t)$, and integrating over the time interval $(-\infty, \infty)$, we have:

$$\begin{aligned} \int_{-\infty}^{\infty} x(t) W_{k_1,m_1}(t) dt &= \int_{-\infty}^{\infty} \sum_k \sum_m C(k, m) W_{k,m}(t) W_{k_1,m_1}(t) dt \\ &= C(k_1, m_1) \int_{-\infty}^{\infty} W_{k_1,m_1}^2(t) dt \end{aligned} \quad (7.7)$$

With $k = k_1$ and $m = m_1$, rearranging (7.7) yields,

$$C(k, m) = \frac{\int_{-\infty}^{\infty} x(t) W_{k,m}(t) dt}{\int_{-\infty}^{\infty} W_{k,m}^2(t) dt} = \frac{\int_{-\infty}^{\infty} x(t) W_{1,0}\left(\frac{t - mkL}{k}\right) dt}{\int_{-\infty}^{\infty} W_{1,0}^2(t) dt} \quad (7.8)$$

The equation indicates that the *measure* function $C(k, m)$ expresses the *correlation* (or *similarity*) between the signal function $x(t)$ and the *derived template* function $W_{k,m}(t)$ of scale k and at time mkL . This concept can be expanded to view a signal transformation operation, such as Fourier or wavelet transform, as a *correlation operation* between a signal and a *template* function, and the result expresses the measures of correlation between the two functions. When a *template* function derived from a *base template* function has a large value of $C(k, m)$, or correlation, with a signal feature at certain scale k and time mkL , the *template* function is said to have a good *match* with that corresponding feature. As a result, the feature will be effectively extracted by this particular *template* function. Signal components that are of little correlation to the *template* function will show small or no correlation measures, and thus be suppressed in the analysis. A signal may show different correlation measures with different *base template* functions. In Table 7.1, several

Table 7.1 Basic template functions expressed in the generalized transformation frame

	Frame $\{W_{k,m}(t)\}$	Properties
Fourier function	$W_{1,0}(t) = e^{-j2\pi t/L}, \quad t \in [0, L)$	Exponential function forms a complete orthogonal base
Haar function	$W_{1,0}(t) = \begin{cases} +1 & 0 \leq t < L/2 \\ 1 & L/2 \leq t < L \end{cases}, \quad t \in [0, L)$	Rectangular waveform forms a complete orthogonal base
Daubechies function	$W_{1,0}(t) = \psi_{1,0}^{(n)}(t), \quad t \in [0, L)^a$	Fractal shape forms a complete orthogonal base

^aThere are different Daubechies functions $\psi_{1,0}^{(n)}(t)$ depending on different order of n

basic template functions are expressed in the generalized signal transformation frame, and they are discussed in the following sections.

7.1.1 Fourier Transform in the Generalized Frame

A signal $x(t)$ of period T can be expressed through its Fourier transform as (Bracewell 1999):

$$x(t) = \sum_{n=0}^{\infty} c_n e^{j2\pi nt/T}, \quad -\infty < t < \infty, \quad n \in N \quad (7.9)$$

where c_n is the n th-order transformation coefficient. If a single-period complex exponential function is defined as the *base template* function:

$$W_{1,0}(t) = e^{j2\pi t/L}, \quad t \in [0, L) \quad (7.10)$$

then the corresponding *derived template* function can be expressed, per definition in (7.1), as:

$$W_{k,m}(t) = \frac{1}{\sqrt{k}} e^{j2\pi(\frac{t}{kL} - m)}, \quad t \in [mkL, mkL + kL) \quad (7.11)$$

Using the *derived template* function, the signal $x(t)$ shown in (7.9) can be expressed, in the generalized transformation frame, as:

$$\begin{aligned} x(t) &= \sum_k \sum_m C(k, m) W_{k,m}(t) = \sum_k \sum_m \frac{1}{\sqrt{k}} C(k, m) e^{j2\pi(\frac{t}{kL} - m)} \\ &= \sum_k \left(\sum_m \frac{1}{\sqrt{k}} C(k, m) \right) e^{j2\pi \frac{t}{kL}} = \sum_k C_k V_k(t) \end{aligned} \quad (7.12)$$

where the term

$$C_k = \sum_m \frac{1}{\sqrt{k}} C(k, m) \quad (7.13)$$

represents the sum of the normalized individual measure functions corresponding to the discrete scale k in the Fourier transform, and the term

$$V_k(t) = e^{j2\pi \frac{t}{kL}} \quad (7.14)$$

represents a periodic exponential function with a period of kL over the time interval $(-\infty, \infty)$, for a given scale k . Comparing (7.9) with (7.12), it can be seen that $k = 1/n$ when $L = T$, and $c_n = C_k$. The expression $k = 1/n$ indicates that a higher scale (k) corresponds to a lower frequency (n), and $c_n = C_k$ is defined only by the frequency n or scale k . From this discussion, the Fourier transform can be viewed in the generalized frame as a 1D function, defined by the scale k with the *orthogonal* base $\{V_k(t)\}$. There is no time information of the extracted signal features contained in this function. This explains why the Fourier transform does not provide time information of the extracted frequency components.

7.1.2 Wavelet Transform in the Generalized Frame

Wavelet transform decomposes a signal using finite time intervals (or *support*) at different scales, thus maintains the time location information of the signal features. Through variation of the scales of the template function used, nonstationary or *transient* features within a signal can be extracted more effectively than by using the exponential (sine and cosine) functions in the Fourier transform. The wavelet transform can be expressed as:

$$C(s, u) = \int_{-\infty}^{\infty} x(t) W_{s,u}(t) dt \quad (7.15)$$

where the term $C(s, u)$ represents the wavelet coefficients (or *measure* function, in the generalized signal transformation frame). The wavelet function $W_{s,u}(t)$ is defined by (7.1). To reduce computational load and avoid redundancy in signal representation, *discrete* instead of *continuous* wavelet transform is often employed for analyzing signals consisting of discrete data points acquired through a data acquisition process. Generally, a discrete wavelet transform discretizes a signal by using a scale of the power of 2 (Daubechies 1992; Kaiser 1994). At the scale $k = 2^n$ ($n \in \mathbb{N}$), the discrete wavelet function $W_{k,m}(t)$ ($m \in \mathbb{Z}$) has the support of $T_w = 2^n L$. Physically, the term $2^n L$ represents the time resolution, which increases linearly with the logarithm of the scale, enabling signal feature extractions at different resolutions. The frequency

resolution of wavelet transform is the inverse of the time resolution, or $1/T_w$. It increases as the scales become higher (i.e., when n increases), and is therefore well suited for analyzing slow-changing signals. At lower scales, the frequency resolution decreases, enabling analysis of fast-changing signals. Such is in contrast to the Fourier transform, which maintains a constant frequency resolution over the entire spectrum, and have a time resolution that is defined by the signal duration.

To illustrate the ability of wavelet transform in signal feature extraction, we analyze an frequency shifted keying (FSK) signal, which is commonly used for data modulation and wireless data transmission (Gibson 1999). An FSK signal is expressed as:

$$x(t) = \begin{cases} \text{square}(2\pi f_1 t) & \text{for message "1"} \\ \text{square}(2\pi f_2 t) & \text{for message "0"} \end{cases} \quad (7.16)$$

where $\text{square}(2\pi f_n t)$ represents a periodic square wave with a unit amplitude and frequency f_n . Figure 7.1 illustrates an example of such an FSK signal, where the frequency $f_1 = 30$ Hz is used to transmit the digit “1” and $f_2 = 125$ Hz is used to transmit the digit “0.” The message to be transmitted is [1 0 0 1 1 0 1 0 0 0]. Such a signal $x(t)$ is nonsinusoidal and nonstationary.

To analyze this signal, we choose the Haar wavelet as the base wavelet, since its square-shaped wave form best matches the shape of the FSK signal. Given that the wavelet has a support of $L = 1$ s (refer to Table 7.1), the measure function $C(s_1, m)$ can be calculated by using (7.8). For the message “1,” $C(s_1, m)$ is calculated to be $\sqrt{s_1}$, at scale $s_1 = 1/f_1$ and time $t = ms_1$. The measure $C(s_1, m)$ is zero for the message “0.” Similarly, at scale $s_2 = 1/f_2$, $C(s_2, m)$ is $\sqrt{s_2}$ at time $t = ms_2$ for the message “0,” and zero for a message “1.” The result of such a wavelet transform operation indicates that the Haar wavelet was able to locate the time instant of the message “1” or “0” at scales s_1 and s_2 , respectively, and expresses the FSK signal $x(t)$ in the generalized signal transform frame as a single-form expression:

$$x(t) = \sum_{m=-\infty}^{\infty} C(s_1, m) \psi_{s_1, m}^{(1)}(t) + C(s_2, m) \psi_{s_2, m}^{(1)}(t) \quad (7.17)$$

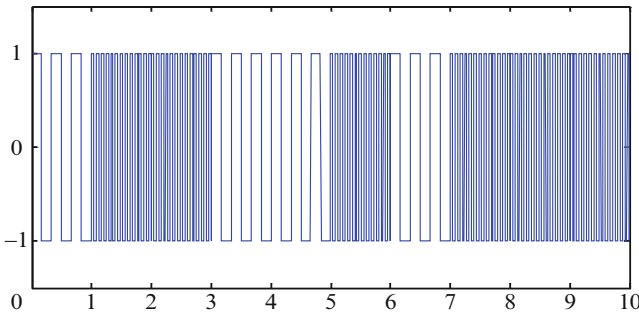


Fig. 7.1 A FSK signal $x(t)$

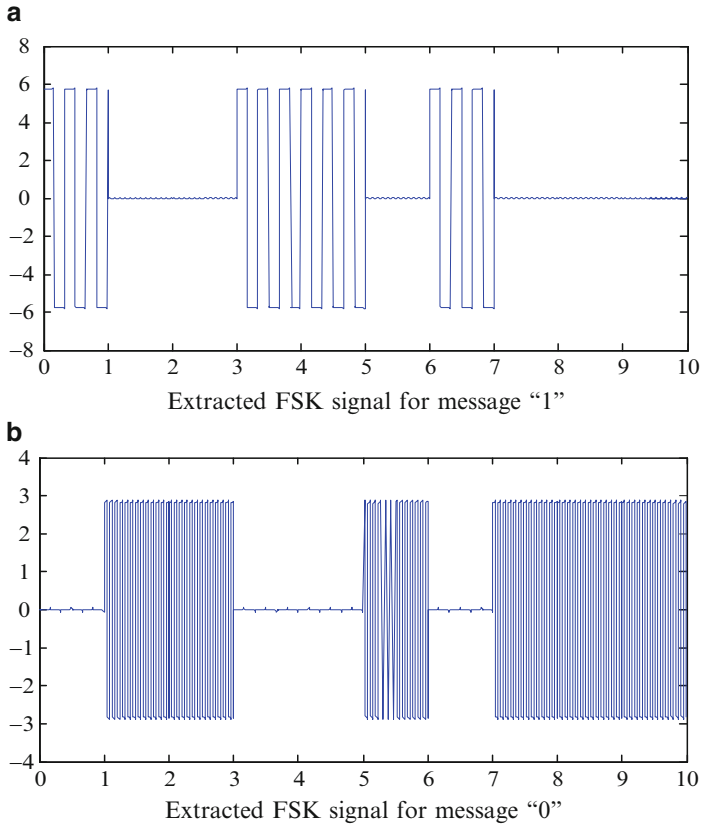


Fig. 7.2 Extracted FSK signals. (a) Extracted FSK signal for message “1” and (b) extracted FSK signal for message “0”

The results are shown in Fig. 7.2, where the messages “1” and “0” can be clearly separated into two different scales. In comparison, it is not feasible to use the Fourier transform to specify at which time which message (1 or 0) is transmitted. The sine and cosine template functions in the Fourier transform do not match the square waveform of the FSK signal, and consequently, the frequency component of the FSK signal will be spread out in a broad spectrum, especially when the message “0” and “1” are randomly transmitted.

7.2 Wavelet Transform with Spectral Postprocessing

As a time scale domain technique, the wavelet transform utilizes template functions of different time resolutions at different scales to extract “transient” features embedded in a signal. Such transient features can be generated by the interactions between the rolling elements in bearing and a localized defect (e.g., surface spalling) on the surface of the raceway. As the rolling elements periodically roll over

the localized defect, the “transient” feature will reoccur at a fundamental frequency f_0 , which is a function of the bearing rotational speed. Such a relationship will be reflected in a wavelet transform of the bearing vibration signal, in that the measure function $C(s_1, m)$ will retain the same fundamental frequency along the time axis, at one of its scales (s_1). As a result, the spectral feature of the transient signal is retained in the wavelet transform, although it is not explicitly expressed. Because of masking by noise and other signals with similar spectral characteristics that appear at the same scale, it can be difficult to rely on wavelet transform alone to identify such hidden patterns.

Such constraint of the wavelet transform can be compensated for by subsequently applying the Fourier transform to the measure function $C(s_1, m)$ resulting from the wavelet transform. Such a postspectral technique reveals the specific frequency location of the transient features, and presents a *unified* approach to transient signal processing. The following section explains how such a postspectral method is realized.

7.2.1 Fourier Transform of the Measure Function

In a complete linear signal space, the wavelet-extracted data set at *scale* s can be expressed as:

$$x_s(t) = \int_{u=-\infty}^{u=\infty} C_s(u) W_s(t-u) du = C_s(t) \otimes W_s(t) \quad (7.18)$$

where the symbol \otimes represents the *convolution* operation between the measured function $C_s(u)$ and the wavelet function $W_s(u)$. To perform Fourier transform on the data set, the Fourier transform of the measure function $C_s(u)$ is first derived. For this purpose, the wavelet transform defined in (7.15) at a fixed scale s is rewritten as:

$$C_s(u) = \int_{-\infty}^{\infty} x(t) W_s(t-u) dt \quad (7.19)$$

In (7.19), the terms $C_s(u)$ and $W_s(t-u)$ represent their respective counterparts in (7.15), $C(s, u)$ and $W_{s,u}(t)$, with a fixed scale s . Through a normalization operation, $\|W_{1,0}(t)\|^2$ in (7.2) is set as 1 for simplicity. With respect to time u , the Fourier transform of $C_s(u)$, denoted as $\tilde{C}_s(f)$, is derived as:

$$\tilde{C}_s(f) = \tilde{x}(f) \tilde{W}_{s,u}(f) \quad (7.20)$$

where the symbol $\tilde{x}(f)$ expresses the Fourier transform of the signal $x(t)$. The symbol $\tilde{W}_{s,u}(f)$ expresses the Fourier transform of the wavelet function $W_{s,u}(t)$, which is derived as:

$$\begin{aligned}
\tilde{W}_{s,u}(f) &= \int_{-\infty}^{\infty} W_{s,u}(t) e^{j2\pi ft} dt = \int_{-\infty}^{\infty} \frac{1}{\sqrt{s}} W_{1,0}\left(\frac{t-u}{s}\right) e^{j2\pi ft} dt \\
&= \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} W_{1,0}\left(\frac{t-u}{s}\right) e^{j2\pi ft} \left[s \cdot d\left(\frac{t-u}{s}\right)\right] \\
&= \sqrt{s} \int_{-\infty}^{\infty} W_{1,0}\left(\frac{t-u}{s}\right) e^{j2\pi fs\left(\frac{t-u}{s} + \frac{u}{s}\right)} d\left(\frac{t-u}{s}\right) \\
&= \sqrt{s} e^{j2\pi fu} \int_{-\infty}^{\infty} W_{1,0}\left(\frac{t-u}{s}\right) e^{j2\pi fs\left(\frac{t-u}{s}\right)} d\left(\frac{t-u}{s}\right) \\
&= \sqrt{s} e^{j2\pi fu} \tilde{W}_{1,0}(sf)
\end{aligned} \tag{7.21}$$

Combining (7.21) with (7.20) yields:

$$\tilde{C}_s(f) = \tilde{x}(f) \sqrt{s} e^{j2\pi fu} \tilde{W}_{1,0}(sf) \tag{7.22}$$

Let $\tilde{W}_{1,0}^*(sf) = e^{-j2\pi fu} \tilde{W}_{1,0}(sf)$, (7.22) can be further expressed as:

$$\tilde{C}_s(f) = \sqrt{s} \tilde{x}(f) \tilde{W}_{1,0}^*(sf) \tag{7.23}$$

where the superscript * denotes the conjugate operator.

Similar to (7.19), in the case of *discrete* wavelet transform, the discrete measure function $C_k(m)$ at a fixed scale k can be expressed as:

$$C_k(m) = \int_{-\infty}^{\infty} x(t) W_k(t - mkL) dt \tag{7.24}$$

The corresponding Fourier transform of $C_k(m)$ is expressed as:

$$\tilde{C}_k(f) = \tilde{x}(f) \tilde{W}_{k,m}(f) \tag{7.25}$$

In (7.25), $\tilde{W}_{k,m}(f)$ is derived as follows:

$$\begin{aligned}
\tilde{W}_{k,m}(f) &= \int_{-\infty}^{\infty} W_{k,m}(t) e^{j2\pi ft} dt = \int_{-\infty}^{\infty} \frac{1}{\sqrt{k}} W_{1,0}\left(\frac{t - mkL}{k}\right) e^{j2\pi ft} dt \\
&= \frac{1}{\sqrt{k}} \int_{-\infty}^{\infty} W_{1,0}\left(\frac{t - mkL}{k}\right) e^{j2\pi ft} \left[k \cdot d\left(\frac{t - mkL}{k}\right)\right] \\
&= \sqrt{k} \int_{-\infty}^{\infty} W_{1,0}\left(\frac{t}{k} - mL\right) e^{j2\pi f\left(\frac{t}{k} - mL + mL\right)} d\left(\frac{t}{k} - mL\right) \\
&= \sqrt{k} e^{j2\pi fmkL} \int_{-\infty}^{\infty} W_{1,0}\left(\frac{t}{k} - mL\right) e^{j2\pi fk\left(\frac{t}{k} - mL\right)} d\left(\frac{t}{k} - mL\right) \\
&= \sqrt{k} e^{j2\pi fmkL} \int_{-\infty}^{\infty} W_{1,0}\left(\frac{t}{k} - mL\right) e^{j2\pi fk\left(\frac{t}{k} - mL\right)} d\left(\frac{t}{k} - mL\right) \\
&= \sqrt{k} e^{j2\pi fmkL} \tilde{W}_{1,0}(kf)
\end{aligned} \tag{7.26}$$

As a result, (7.25) is given by:

$$\tilde{C}_k(f) = \tilde{x}(f)\tilde{W}_{k,m}(f) = \tilde{x}(f)\sqrt{k} e^{j2\pi fmkL} \tilde{W}_{1,0}(kf) \quad (7.27)$$

Let $\tilde{W}_{1,0}^*(kf) = \tilde{W}_{1,0}(kf)e^{-j2\pi fmkL}$, (7.27) can be further expressed as:

$$\tilde{C}_k(f) = \sqrt{k}\tilde{x}(f)\tilde{W}_{1,0}^*(kf) \quad (7.28)$$

Equations (7.23) and (7.28) illustrate that the Fourier transform of the measure function at scale s (for continuous transform) or k (for discrete transform) can be viewed as the original signal $x(t)$ passing through a data filter, which is a contracted (by a frequency factor of s or k) and amplified (by a factor of \sqrt{s} or \sqrt{k}) version of the filter represented by the base wavelet function. Such an operation establishes the link between measured function and data filtering, and is of significance in wavelet transform-based signal analysis.

7.2.2 Fourier Transform of Wavelet-Extracted Data Set

With the Fourier transform of the measure function obtained, the Fourier transform of the extracted (or reconstructed) data set from the continuous wavelet transform, denoted $x_s(t)$ as shown in (7.18), can be derived as:

$$\begin{aligned} \tilde{x}_s(f) &= \tilde{C}_s(f)\tilde{W}_s(f) \\ &= \sqrt{s}\tilde{x}(f)\tilde{W}_{1,0}^*(sf)\sqrt{s}\tilde{W}_{1,0}(sf) \\ &= s\tilde{x}(f)|\tilde{W}_{1,0}(sf)|^2 \end{aligned} \quad (7.29)$$

In case of a *discrete* wavelet transform, the Fourier transform of the data set $x_k(t)$ is obtained by setting $s = k$ and $u = mkL$ in (7.18), and its Fourier transform is expressed as:

$$\begin{aligned} \tilde{x}_k(f) &= \tilde{C}_k(f)\tilde{W}_k(f) \\ &= \sqrt{k}\tilde{x}(f)\tilde{W}_{1,0}^*(kf)\sqrt{k}\tilde{W}_{1,0}(kf) \\ &= k\tilde{x}(f)|\tilde{W}_{1,0}(kf)|^2 \end{aligned} \quad (7.30)$$

This indicates that the Fourier transform of the extracted data set $x_k(t)$ at scale k can be viewed as the Fourier transform of the original signal $x(t)$ passing through a low-pass filter and the filter being represented by the transfer function $|\tilde{W}_{1,0}(kf)|^2$. If the template function at scale k correlates well with the “transient” feature of the signal $x(t)$ in the time domain, then its Fourier transform will contain a strong “disturbance” component in its spectrum. As a result, the filter $|\tilde{W}_{1,0}(kf)|^2$ will extract the

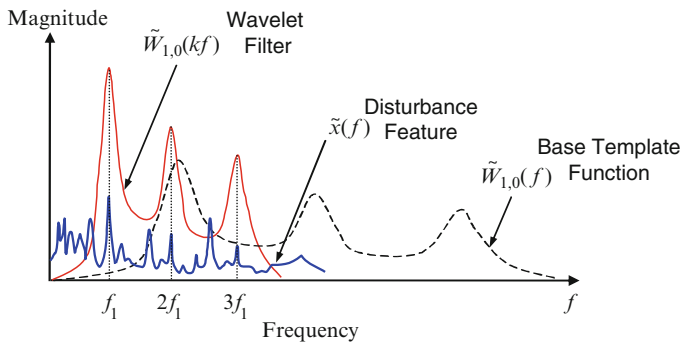


Fig. 7.3 Illustration of the filtering effect of wavelet transform

“disturbance” signal features from the original signal $x(t)$ at the scale k , as shown in (7.29) and (7.30). Because of a lower degree of correlation between this filter and other constituent components in the signal, other components will be attenuated at the scale k .

The filtering effect of postspectral processing of a wavelet transformed data series is illustrated in Fig. 7.3. The “transient” feature, represented by the solid thick line, is shown to have a fundamental frequency f_1 , characterized by a magnitude peak at f_1 and its harmonics ($2f_1$ and $3f_1$) in the frequency spectrum. When an appropriate base wavelet is selected (or designed) to decompose this signal (Holm-Hansen et al. 2004), a base template function will exist at a certain scale k where a high degree of correlation between the template function and the “transient” features can be identified. If the data set resulting from such a wavelet transform is subsequently processed by the Fourier transform, the result will be a data spectrum similar to that of the “transient” signal, with its major frequency components at f_1 , $2f_1$, and $3f_1$, respectively. Such a postspectral analysis can be viewed as filtering the data set in the frequency domain denoted by $\tilde{W}_{1,0}(kf)$.

7.3 Application to Bearing Defect Diagnosis

This section illustrates how the unified time scale frequency analysis technique described earlier can be applied to rolling bearing defect diagnosis. A custom-designed bearing test bed, as shown in Fig. 7.4, is set up to provide an experimental platform for evaluating the developed method. Axial and radial loads on the test bearing are applied through a hydraulic system, and the bearing rotation speed is varied by controlling the DC drive motor. Commercially available accelerometers (model 8624) are placed on the housing for vibration measurement with a data sampling frequency of 10 kHz. A deep-groove ball bearing of type 6220 with a seeded structural defect serves as the test bearing. The defect is implemented as a 0.25-mm diameter hole drilled on the inner raceway of the bearing, simulating the

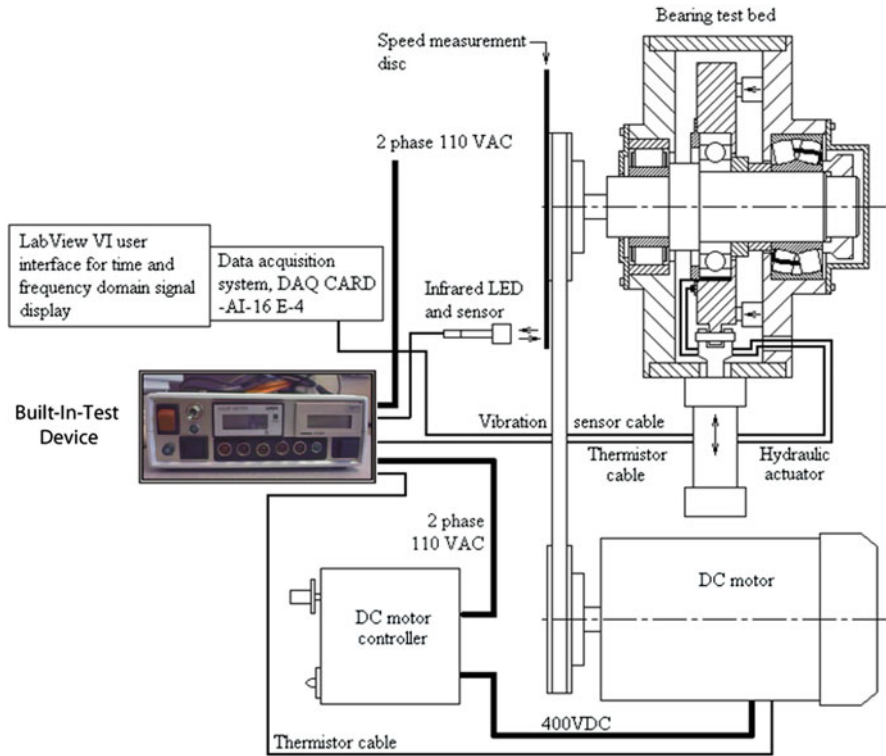


Fig. 7.4 Bearing test bed with hydraulic load application capability

condition of a surface spall. The relationship between the bearing rotational frequency f_r and the characteristic frequencies associated with defect-induced vibrations can be determined analytically as a function of the defect location, for example, on the inner raceway (f_{BPFI}), outer raceway (f_{BPFO}) or a rolling element (f_{BSF}) (Harris 1991). Specifically for the test bearing, these characteristic frequencies are calculated as $f_{BPFI} \approx 5.86f_r$, $f_{BPFO} \approx 4.1f_r$, $f_{BSF} \approx 5.3f_r$, respectively (SKF 1996). By identifying the existence of these characteristic frequencies and/or their combinations, the existence of bearing structural defects can be determined.

In the experimental evaluation, following aspects are studied:

1. The effectiveness of the unified time scale frequency analysis technique in extracting defect features (i.e., characteristic frequencies) from bearing vibration signals is compared to that of the Fourier transform and the discrete wavelet transform when it is applied alone.
2. The effectiveness of the new technique at different wavelet decomposition levels.
3. The effectiveness of the new technique under varying bearing operating conditions, such as the radial load, axial load, and shaft rotational speed.

7.3.1 Effectiveness in Defect Feature Extraction

To establish a basis for objective comparison, vibration signals are measured on both a defect-free (i.e., *healthy*) and a defective bearing of the same model (SKF 6220), under the same operation conditions: shaft speed $f_r = 600$ rpm (corresponding to 10 Hz rotational frequency), axial load of 7,038 N, and radial load of 17,468 N. Figures 7.5 and 7.6 shows the two signals in the time and frequency domains, respectively, while the related frequency resolution is approximately 0.3 Hz.

As shown in Fig. 7.6, both spectra indicate the existence of two dominant frequency components: (1) ball rotation (f_{BPFO1} , ball passing frequency on outer raceway, with the subscript “1” referring to the 6220 bearing) and (2) bearing misalignment (f_m). Ball rotation-related vibration has a peak value at the fundamental frequency $f_{\text{BPFO1}} = 41$ Hz, which is equal to $4.1 f_r$. Misalignment-related vibration has a characteristic frequency of $f_m = 20$ Hz ($= 2f_r$). In addition to these two major components, components related to bearing imbalance are also identified at the frequency $f_u = 10$ Hz (identical to f_r). However, visual comparison of the two spectra reveals no difference between them, as the characteristic frequency of $f_{\text{BPFI1}} = 58.6$ Hz, related to the inner raceway defect, is not recognized.

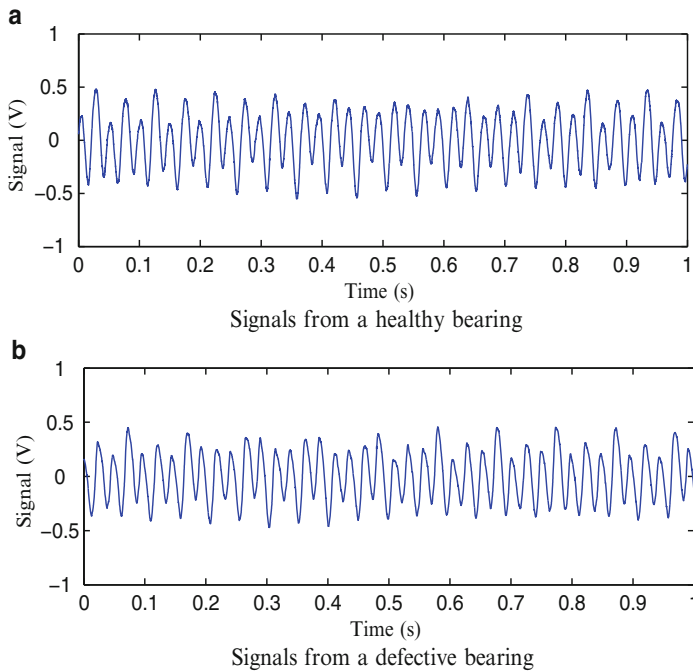


Fig. 7.5 Time domain signals from a healthy and a defective bearing. (a) Signals from a healthy bearing and (b) signals from a defective bearing

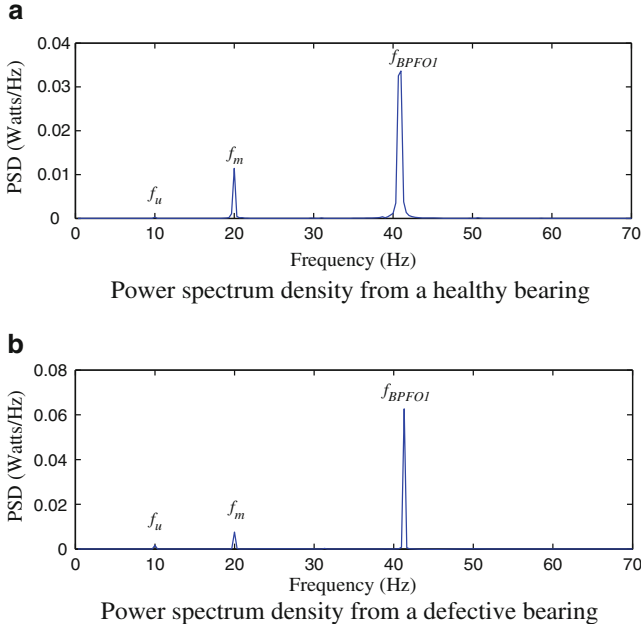


Fig. 7.6 Results of frequency domain analysis of the healthy and defective bearing. (a) Power spectrum density from a healthy bearing and (b) power spectrum density from a defective bearing

This illustrates that Fourier transform, when applied alone, may not be effective in detecting the existence of bearing structural defect.

The same signals are then analyzed using discrete wavelet transform, with the Daubechies 2 wavelet as the base wavelet. Figure 7.7 illustrates the wavelet coefficients of the vibration signals at the decomposition level 7, which has a corresponding frequency range of 39–78 Hz, thus covering the defect characteristic frequency of $f_{BPFI} = 58.6$ Hz. As seen in Fig. 7.7, the wavelet coefficients for the defective bearing have shown more dynamical variations than that of the healthy bearing. To quantify their difference, the root-mean-square (RMS) values of the two wavelet coefficients are calculated. It is found that the RMS value of the defective bearing (56 mV) is about 145% larger than that of the healthy bearing (22.8 mV). Although such an increase can be used as an indicator of structure defect in the bearing, it has the limitation that proper threshold needs to be set up a priori, to determine the quantitative extent that distinguishes a healthy bearing from a defective one. Another limitation of the wavelet transform is that the wavelet coefficients do not provide any indication on the specific location of the defect in the bearing, since it does not explicitly reveal the characteristic frequencies from the bearing.

Next, the bearing signal (as shown in Fig. 7.5) is analyzed using the unified time scale frequency method. For this purpose, the wavelet coefficients of the signal (shown in Fig. 7.7) is postprocessed using the Fourier transform.

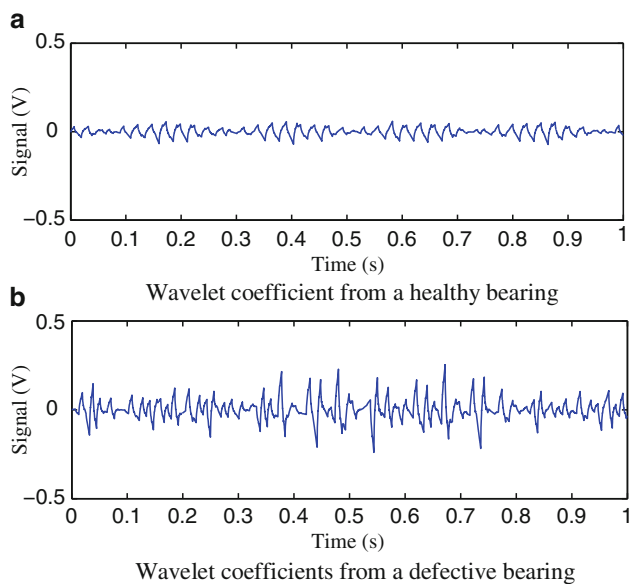


Fig. 7.7 Wavelet decomposition of bearing signals at decomposition level 7. (a) Wavelet coefficient from a healthy bearing and (b) wavelet coefficients from a defective bearing

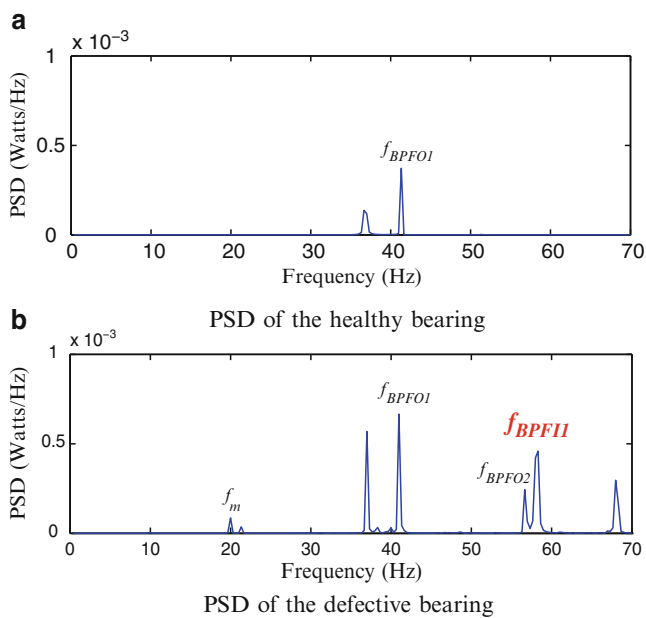


Fig. 7.8 Unified analysis for defect feature extraction at decomposition level 7. (a) PSD of the healthy bearing and (b) PSD of the defective bearing

Comparing the spectra of the healthy (Fig. 7.8a) and defective bearings (Fig. 7.8b), it is seen that the inner raceway defect can be clearly identified by its characteristic frequency at $f_{\text{BPF11}} = 58.6$ Hz. No distinctive peak is seen in the spectrum of the healthy bearing at this frequency. The spectrum further indicates several other major peaks at $f_m = 20$ Hz, $f_{\text{BPFO1}} = 41$ Hz, and $f_{\text{BPFO2}} = 56.5$ Hz. These are reflective of misalignment (at 20 Hz) of the defective bearing and rotational characteristic of other bearing. For example, $f_{\text{BPFO1}} = 41$ Hz is the ball passing frequency of the type 6220 bearing, and $f_{\text{BPFO2}} = 56.5$ Hz is found to be related to ball rotation of a different bearing (cylindrical bearing type 2322 with its vibration component indicated by the subscript “2”). This is based on the parameters of $Z = 14$, $D = 33.5$ mm, and $d_m = 175$ mm, and the characteristic frequency of the type 2322 bearing is calculated as $f_{\text{BPF12}} = 83.4$ Hz, $f_{\text{BPFO2}} = 56.5$ Hz, $f_{\text{BSF2}} = 50.1$ Hz. This bearing structurally supports the rotating shaft in the bearing test bed. Comparing with the Fourier analysis and the wavelet transform, the new, *unified* time scale frequency technique has shown to be more effective in extracting bearing defect features. In that it not only reveals the existence of a localized bearing defect, but also the defect characteristic frequency that is indicative of its specific location (e.g., inner raceway).

7.3.2 Selection of Decomposition Level

When evaluating the *unified* technique, a particular decomposition level (e.g., level 7) is chosen for the wavelet transform. The selection of an appropriate level is based on the signal sampling rate (or frequency) f_{sample} and the defect characteristic frequency f_{char} . The relationship is expressed as:

$$\frac{f_{\text{sample}}}{2^{L+1}} \leq f_{\text{Char}} \leq \frac{f_{\text{sample}}}{2^L} \quad (7.31)$$

where L denotes the wavelet decomposition level. As an example, when the sampling frequency is $f_{\text{sample}} = 10$ kHz, the frequency range associated with decomposition level $L = 7$ is calculated as 39–78 Hz. In Table 7.2, the frequency ranges covered by each of the decomposition levels under a 10 kHz sampling rate are shown. The essence of finding the best-suited decomposition level when wavelet transforming a dynamic signal is to ensure that its frequency range $[f_{\text{sample}}/2^{L+1}, f_{\text{sample}}/2^L]$ covers the characteristic frequency of structural defect in the bearing with the highest likelihood, if such a defect exists.

Table 7.3 lists the best-suited decomposition levels for analyzing bearing vibration signals specifically related to a localized inner raceway defect, under various bearing rotational (or shaft) speeds. Since at 600 rpm, the defect characteristic frequency $f_{\text{BPF11}} = 58.6$ Hz falls within the frequency range of 39–78 Hz, decomposition level 7 is chosen initially for data analysis.

Table 7.2 Frequency range associated with each decomposition level at a 10 kHz sampling rate

Decomposition level	Frequency range (Hz)	Decomposition level (<i>L</i>)	Frequency range (Hz)
1	2,500 5,000	5	156 312
2	1,250 2,500	6	78 156
3	625 1,250	7	39 78
4	312 625	8	19 39

Table 7.3 Best suited decomposition level for inner raceway defect frequency detection

Shaft speed (rpm)	f_{BPFI1} (Hz)	Decomposition level	Frequency range (Hz)
300	29.3	8	19 39
600	58.6	7	39 78
900	87.9	6	78 156
1,200	117.2	6	78 156
1,500	146.5	6	78 156

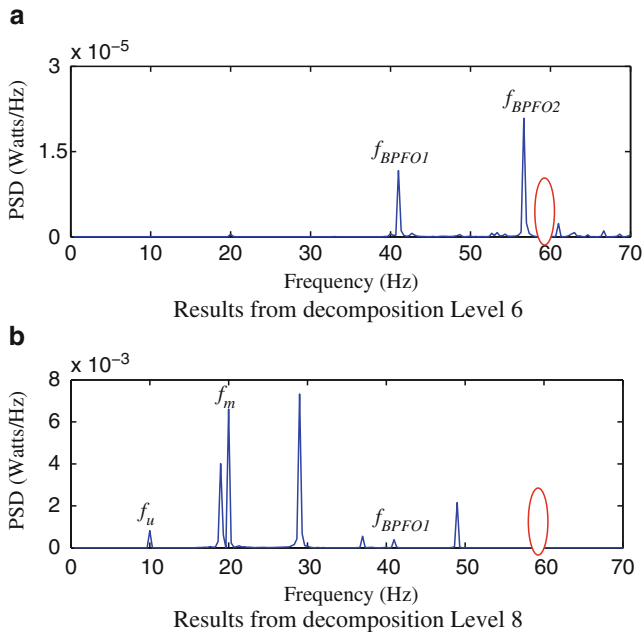


Fig. 7.9 Unified analysis using Daubechies 2 wavelet at levels 6 and 8. (a) Results from decomposition level 6 and (b) results from decomposition level 8

The importance of choosing proper decomposition level is illustrated in Fig. 7.9, where the results of decomposing defective bearing signal at levels 6 and 8 are shown. It is seen that none of the two levels (combined with the postspectral analysis) are able to identify the defect characteristic frequency at $f_{BPFI1} = 58.6$ Hz, due to the

mismatch between their respective frequency range (level 6 at 78–156 Hz and level 8 at 19–39 Hz) and the characteristic frequency of the inner raceway defect (at 58.6 Hz).

7.3.3 Effect of Bearing Operation Conditions

To investigate the effectiveness of the *unified* time scale frequency analysis method in defect feature extraction under varying bearing operating conditions, three groups of experiments are designed and conducted using a type 6220 ball bearing with a seeded defect.

7.3.3.1 Variation of Radial Load

The effect of radial loads on defect feature extraction is illustrated through the experimental results shown in Fig. 7.10, where four levels of radial loads are presented. It is noted that, as the radial load has progressively increased from 4,367 to 26,202 N, the peak value of the bearing defect frequency of f_{BPFI1} (= 58.6 Hz) has grown by 609.5%, as given in Table 7.4. Such an increase can be explained by the increased preload applied by the rolling elements of the bearing to the defect on the raceway. An increased preload enhances the impacts when the rolling elements roll over the defect, leading to increased amplitude of the defect feature.

7.3.3.2 Variation of Axial Load

Increase in the axial load has also shown to lead to an increase of the defect feature amplitude, as is evident when comparing the three different axial load conditions in Fig. 7.11. For example, when the axial load applied to the bearing increases from 0 to 4,192 N, the defect signal amplitude at defect characteristic frequency f_{BPFI1} has increased by 4.7%, from 3.18×10^{-3} to 3.33×10^{-3} W/Hz, as listed in Table 7.5. This is because the application of axial load on the bearing increases the extent of the bearing load zone distribution, resulting in an increased number of defect impacts within the load zone (Harris 1991). Such an increase enhances the overall energy content of the defect signal at its feature frequency, and is reflected by the increased defect feature amplitude.

7.3.3.3 Variation of Rotational Speed

The power spectral density graphs in Fig. 7.12 illustrate the effect of bearing rotational speed on the defect signal strength. As the speed increased from 300 to 1,200 rpm, the defect frequency amplitude increased by 71.2%, as listed in

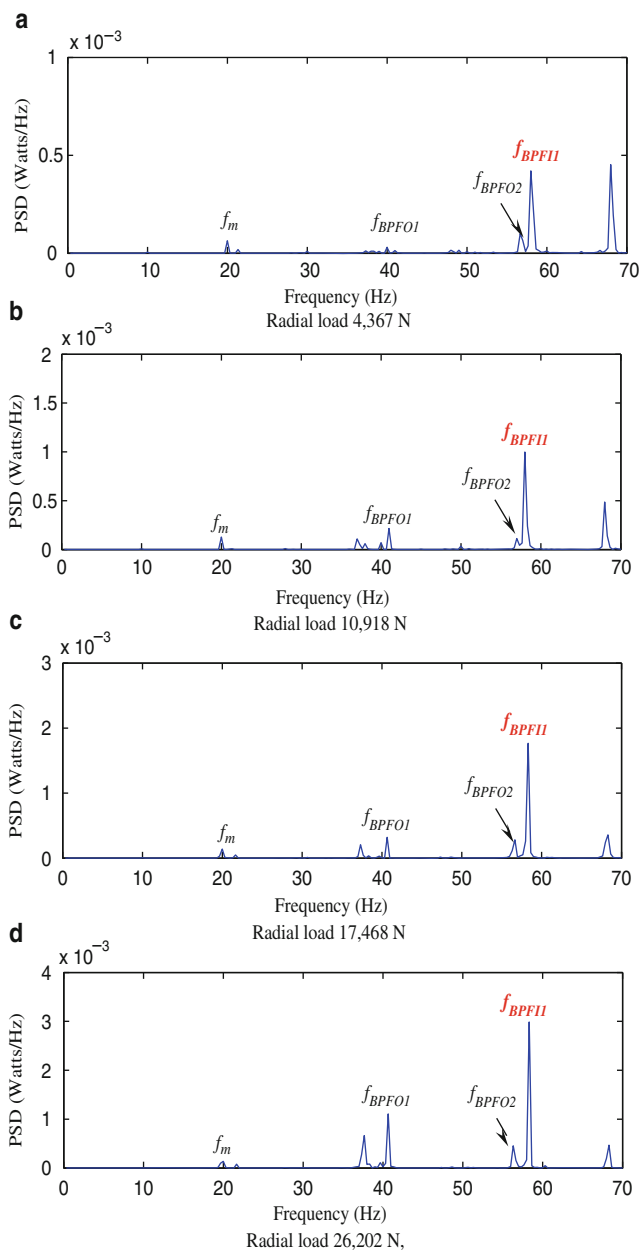


Fig. 7.10 Effect of the radial load on defect feature amplitude. (a) Radial load 4,367 N, (b) radial load 10,918 N, (c) radial load 17,468 N, and (d) radial load 26,202 N

Table 7.4 Effect of radial load on the defect signal amplitude (f_{BPF11})

Radial load (N)	PSD 10^{-3} (W/Hz)	Percentage of increase
4,367	0.42	
10,918	0.99	135.7
17,468	1.77	321.4
26,202	2.98	609.5

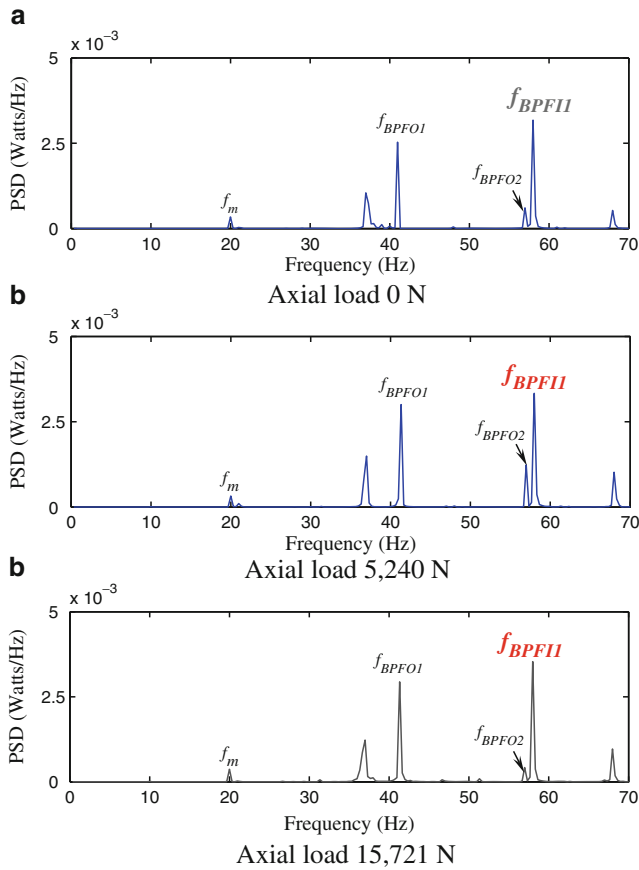


Fig. 7.11 Effect of axial load on defect feature amplitude. (a) Axial load 0 N, (b) axial load 5,240 N, and (c) axial load 15,721 N

Table 7.5 Effect of axial load on the defect feature amplitude (f_{BPF11})

Axial load	PSD 10^{-3} (W/Hz)	Percentage of increase
0	3.18	
5,240	3.33	4.7
15,721	3.54	11.3

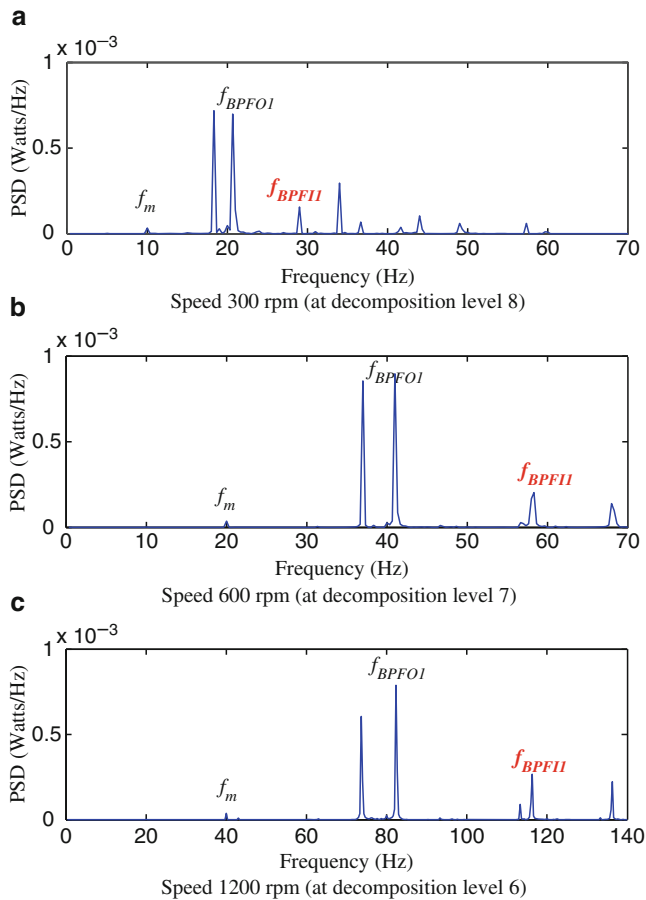


Fig. 7.12 Effect of bearing speed on defect amplitude. (a) Speed 300 rpm (at decomposition level 8), (b) speed 600 rpm (at decomposition level 7), and (c) Speed 1,200 rpm (at decomposition level 6)

Table 7.6 Effect of speed on the defect signal amplitude (f_{BPFII})		
Shaft speed (rpm)	PSD 10^{-4} (W/Hz)	Percentage of increase
300	1.56	
600	2.03	30.1
1,200	2.67	71.2

Table 7.6. The speed-related defect feature amplitude increase can be explained by the fact that with the increase of the speed, the number of impacts per bearing revolution increases proportionally, hence the total amount of defect impact energy also increases, leading to increased peak amplitude at the defect characteristic frequency f_{BPFII} .

7.4 Summary

This chapter introduces a *unified* time scale frequency analysis technique based on the combination of discrete wavelet transform with frequency domain postprocessing. The effectiveness of this technique in improving bearing defect diagnosis is then investigated. A localized defect of 0.25 mm in diameter at the inner raceway of a type 6220 bearing has been successfully detected, under various bearing operation conditions. It is shown that the Fourier-transform-based spectral analysis technique alone is not reliable to detect the transient components that are characteristic of localized bearing defect, whereas wavelet transform alone does not explicitly identify the specific location of the defect. Thus, the *unified* technique combines the advantages of both the time and frequency domain analyses and provides more information on the defect feature than each of the techniques employed individually. In addition to bearing defect diagnosis, the new technique provides a powerful tool for the detection, extraction, and identification of weak “defect” features submerged in vibration signals in a wide range of manufacturing equipment

7.5 References

- Bracewell, R (1999) The Fourier transform and its applications, 3rd edn. McGraw Hill, New York
- Byrne G, O'Donnell GE (2007) An integrated force sensor solution for process monitoring of drilling operations. *CIRP Ann Manuf Technol* 56(1):89–92
- Cavacece M, Introini A (2002) Analysis of damage of ball bearings of aeronautical transmissions by auto power spectrum and cross power spectrum. *ASME J Vib Acoust* 124(2):180–185
- Daubechies I (1992) Ten lectures on wavelets. SIAM, Philadelphia, PA
- Gao R, Yan R (2006) Non stationary signal processing for bearing health monitoring. *Int J Manuf Res* 1(1):18–40
- Ge M, Du, R, Zhang GC, Xu YS (2004) Fault diagnosis using support vector machine with an application in sheet metal stamping operations. *Mech Syst Signal Process* 18(1):143–159
- Gibson J (1999) Principle of digital and analog communication, 2nd edn. Prentice Hall, Inc, Upper Saddle River, NJ
- Harris TA (1991) Rolling bearing analysis, 3rd edn. Wiley, New York
- Ho D, Randall RB (2000) Optimization of bearing diagnostic techniques using simulated and actual bearing fault signals. *Mech Syst Signal Process*, 14(5):763–788
- Holm Hansen BT, Gao R, Zhang L (2004) Customized wavelet for bearing defect detection. *ASME J Dyn Syst Meas Control* 126(6):740–745
- Kaiser G (1994) A friendly guide to wavelets. Birkhäuser, Boston, MA
- Malekian M, Park SS, Jun M (2009) Tool wear monitoring of micro milling operations. *J Mater Process Technol* 209:4903–4914
- Mori K, Kasashima N, Yoshioka T, Ueno Y (1996) Prediction of spalling on a ball bearing by applying the discrete wavelet transform to vibration signals. *Wear* 195(1–2):162–168
- Obikawa T, Shinozuka J (2004) Monitoring of flank wear of coated tools in high speed machining with a neural network ART2. *Int J Mach Tools Manuf* 44:1311–1318
- Orhan S, Aktürk N, Çelik V (2006) Vibration monitoring for defect diagnosis of rolling element bearings as a predictive maintenance tool: comprehensive case studies. *NDTE Int* 39:293–298
- SKF Company (1996) SKF bearing maintenance handbook. SKF Company, Denmark
- Tandon T, Choudhury A (1999) A review of vibration and acoustic measurement methods for the defection of defects in rolling element bearings. *Tribol Int* 32:469–480

Chapter 8

Wavelet Packet-Transform for Defect Severity Classification

Once a defect is detected, the next question that comes up naturally is how severe the defect is. Since machine downtime is physically rooted in the progressive degradation of defects within the machine's components, accurate assessment of the severity of defect is critically important in terms of providing input to adjusting the maintenance schedule and minimizing machine downtime. This chapter describes how wavelet packet transform (WPT)-based techniques can classify machine defect severity, with specific application to rolling bearings.

8.1 Subband Feature Extraction

Because of the complex nature of machines and the intricacy of related parameters, it is generally difficult to assess the status of a machine directly from the measured time domain data. The general practice is to extract “features” to identify characteristics and patterns embedded in the data series that are indicative of status changes of the machine being monitored. The advent of wavelet transform has provided an effective tool for feature extraction from various time-varying signals, such as washing machines (Goumas et al. 2002), rolling bearings (Mori et al. 1996; Prabhakar et al. 2002), and machine tools (Lee and Tang 1999; Li et al. 2000a, 2000b). As an extension of the wavelet transform, WPT provides more flexible time frequency decomposition, especially in the high-frequency region, when compared with the wavelet transform. In particular, WPT allows for feature extraction (e.g., energy content or kurtosis value) from subfrequency bands of the decomposed signal where the features are concentrated, thereby directing the computation to where it is most needed. Prior efforts have studied different sets of wavelet packet vectors to represent bearing vibration under different defect conditions (Liu et al. 1997). Altmann and Mathew (2001) found that features extracted from wavelet packets that cover the multiple subfrequency bands yield a higher signal-to-noise (S/N) ratio than those from a conventional band-pass filter. For multistage gearbox vibration analysis, the Hilbert transform and WPT were combined to enable gear defect detection at the incipient stage (Fan and Zuo 2006).

Given a time domain signal $x(t)$, the WPT decomposes it into a number of subbands, as expressed by the resulting wavelet packet coefficients:

$$x(t) = \sum_{n=0}^{2^j-1} x_j^n(t) \quad (8.1)$$

In (8.1), the term $x_j^n(t)$ denotes the wavelet coefficients at the j level, n subband. From these coefficients, “features” will be extracted, at each subband, to provide information on the condition of the machine being monitored.

8.1.1 Energy Feature

The energy content of a signal provides a quantitative measure for the signal, which uniquely characterizes the signal. The amount of energy contained in a signal $x(t)$ is expressed as:

$$E_{x(t)} = \int |x(t)|^2 dt \quad (8.2)$$

The energy content of a signal can also be calculated from the coefficients of the signal’s transform. In the case of a WPT, the coefficients $x_j^n(t)$ quantify the amount of energy associated with each of the subbands. The total amount of energy contained in the signal is equal to the sum of the energy in each subband and expressed as:

$$E_{x(t)} = \sum_{n=0}^{2^j-1} \int |x_j^n(t)|^2 dt \quad (8.3)$$

Since the energy content of each subband of the signal is directly related to the severity of the defect, it presents an indicator or *feature* of the machine’s condition. From (8.3), the energy feature in each subband is defined as:

$$E_j^n = \int |x_j^n(t)|^2 dt \quad (8.4)$$

Similarly, when a signal is represented by discrete, sampled values $x(i)$ ($i = 1, 2, \dots, M$), the total energy feature in the subbands is calculated as:

$$E_j^n = \sum_{i=1}^M x_j^n(i)^2 \quad (8.5)$$

8.1.2 Kurtosis

Kurtosis is a dimensionless, statistical measure that characterizes the flatness of a signal's probability density function. An impulsive signal that is peaked has a larger kurtosis value than a signal that is flat and varies with time slowly, as illustrated in Fig. 8.1.

Mathematically, the kurtosis of a signal is defined as its fourth-order moment:

$$K_{x(t)} = \frac{E[(x(t) - \mu_{x(t)})^4]}{\sigma_{x(t)}^4} \quad (8.6)$$

where $\mu_{x(t)}$ and $\sigma_{x(t)}$ denotes the mean value and standard deviation of the signal $x(t)$, respectively. The symbol $E[\bullet]$ denotes the expectation operation. Table 8.1 lists the kurtosis values of several representative signals. It should be noted that the value of kurtosis does not depend on the amplitude of a signal.

For the wavelet packet coefficients in each subband, the corresponding kurtosis value is defined as:

$$K_j^n = \frac{E[(x_j^n(t) - \mu_{x_j^n(t)})^4]}{\sigma_{x_j^n(t)}^4} \quad (8.7)$$

where the symbols $\mu_{x_j^n(t)}$ and $\sigma_{x_j^n(t)}$ are the mean and standard deviation of the wavelet packet coefficients $x_j^n(t)$, respectively.

When the wavelet packet coefficients are sampled as $x_j^n(i)$, the kurtosis value is calculated as:

$$K_j^n = \frac{\sum_{i=1}^N [x_j^n(i) - \mu_{x_j^n(i)}]^4}{N\sigma_{x_j^n(i)}^4} \quad (8.8)$$

where the symbols $\mu_{x_j^n(i)}$ and $\sigma_{x_j^n(i)}$ are the mean and standard deviation of the wavelet packet coefficients $x_j^n(i)$, respectively.

Since the energy content of a signal provides a robust indicator of the signal, but is not sensitive in characterizing incipient defects, whereas the kurtosis

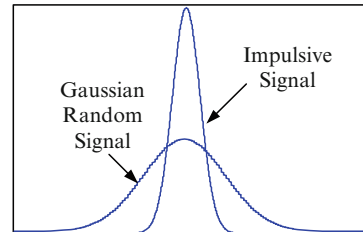


Fig. 8.1 Illustration of probability density functions of signals

Table 8.1 Kurtosis values of several typical signals

Signal	Kurtosis
Square signal	1.0
Sinusoidal signal	1.5
Gaussian signal	3.0
Pulse signal	>3.0

value has high sensitivity to incipient defects but has low stability (Yan and Gao 2004), these two features can be combined instead of being used alone to improve the signal characterization. Suppose we decompose a signal into j levels (e.g., $j = 4$), which generates 2^j or $2^4 = 16$ subbands. Given that the energy and kurtosis values are calculated from each subband, there will be 2×2^j or $2 \times 2^4 = 32$ features extracted from the signal. These features can be expressed in a feature vector as:

$$FV = [E_j^0, E_j^1, \dots, E_j^{2^{j-1}}, K_j^0, K_j^1, \dots, K_j^{2^{j-1}}]^T \quad (8.9)$$

For simplicity, (8.9) can be expressed as:

$$FV = \{f_l | l = 1, 2, \dots, p\}, \quad p = 2^{j+1} \quad (8.10)$$

where $f_1 = E_j^0, \dots, f_{2^j} = E_j^{2^{j-1}}, f_{2^j+1} = K_j^0, \dots, f_p = K_j^{2^{j-1}}$.

Determining which features shown in (8.10) are most effective for characterizing machine defect is generally not a simple, straightforward process because the usefulness of features may be affected by factors such as the specific location of the sensors, and consequently, the quality of the signal measured in terms of the S/N ratio or signal contamination. Furthermore, using more features may not necessarily improve the effectiveness of defect severity estimation, while increasing the computation cost (Malhi and Gao 2004). Since defect-induced signals are typically reflected in the variation of the characteristic frequencies (e.g., characteristic defect frequency shifts as the defect size grows), degradation of machine conditions is predominantly reflected in certain subfrequency bands, whereas other subfrequency bands contain information unrelated to the defect. This indicates that feature selection is needed for identifying significant features from the pool of WPT-based feature set.

8.2 Key Feature Selection

This section introduces two feature selection methods: Fisher linear discriminant (FLD) analysis and principal component analysis (PCA). The goal is to differentiate the signals (which represent different machine defect severity) by examining only

those subbands with distinct feature discrimination than others, thus improving the efficiency while not missing critical information related to the signal, to ensure reliability of the diagnostic operation.

8.2.1 Fisher Linear Discriminant Analysis

Distance measures, such as the Bhattacharyya distance, Kolmogorov distance, and FLD (Fukunaga 1990; Yen and Lin 2000; Duda et al. 2001), have been applied to components differentiation within a class pair. Generally, the greater the distance between two feature components within a class pair is, the easier it will be to separate them. Illustrated in Fig. 8.2 are two feature components representing the two signals from a healthy and a defective bearing, respectively. The features in the right-hand section of the figure are easier to be separated than those in the left-hand section because the probability distributions of the two components do not overlap, because of their relatively smaller standard distributions and larger distance between the mean values, compared with the two features in the left section. The result is that this pair of features has a higher discriminant power than the other pair of features.

The approach introduced here for efficient feature selection is to evaluate the discriminant power of each individual feature within a class pair. Features that have a low discriminant power are excluded from the data analysis process, as they contain little useful information. Such an approach can be realized by examining the rank order (Kittler 1975) of the feature vector $FV = \{f_l | l = 1, 2, \dots, p\}$, shown in (8.10), as:

$$J(f_1) \geq J(f_2) \geq \dots \geq J(f_d) \geq \dots \geq J(f_p) \quad (8.11)$$

where $J(\bullet)$ is a criterion function for evaluating the discriminant power of a specific feature. Here the utility of the FLD is introduced (Duda et al. 2000), where the criterion function for differentiating a class pair is given by:

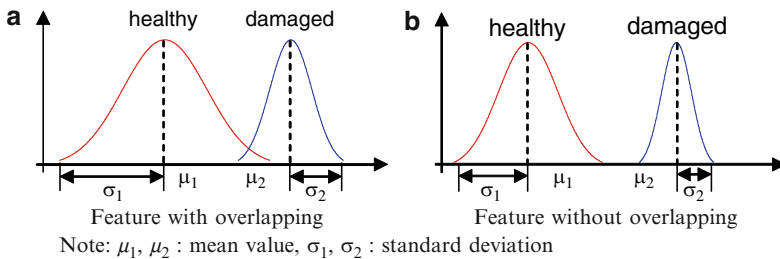


Fig. 8.2 Feature discrimination based on the distance between constituent components. (a) Feature with overlapping. (b) Feature without overlapping. Note: μ_1, μ_2 : mean value, σ_1, σ_2 : standard deviation

$$J_{f_l}(i, j) = \frac{|\mu_{i,f_l} - \mu_{j,f_l}|^2}{\sigma_{i,f_l}^2 + \sigma_{j,f_l}^2} \quad (8.12)$$

The symbols μ_{i,f_l} , μ_{j,f_l} and σ_{i,f_l}^2 , σ_{j,f_l}^2 represent the mean values and the variances of the l th feature, f_l , and for the classes i and j , respectively. Since typically more than two defect severity levels need to be evaluated in a machine defect diagnosis system, a k -class, p feature component problem with $k(k-1)/2$ class pairs is investigated for generality. The process for feature selection, based on the FLD analysis method, is illustrated in Fig. 8.3.

Features (i.e., subband energy and kurtosis values) are first extracted by means of the WPT from the signals measured on the machine (e.g., a milling machine, a spindle), under various operating conditions (e.g., speed and load). The mean values and variances of each individual feature f_l corresponding to each machine status are then calculated, for each set of operation conditions (e.g., 1,200 rpm, 3.6 kN radial load). For each possible class pair $\{(i, j) | i = 1, 2, \dots, k-1, j = i+1, i+2, \dots, k\}$ formed from two different machine states (e.g., health vs. light defect), the discriminant power measure $J_{f_l}(i, j)$ for each feature f_l , is calculated, using (8.12). Descending sorting $J_{f_l}(i, j)$ yields:

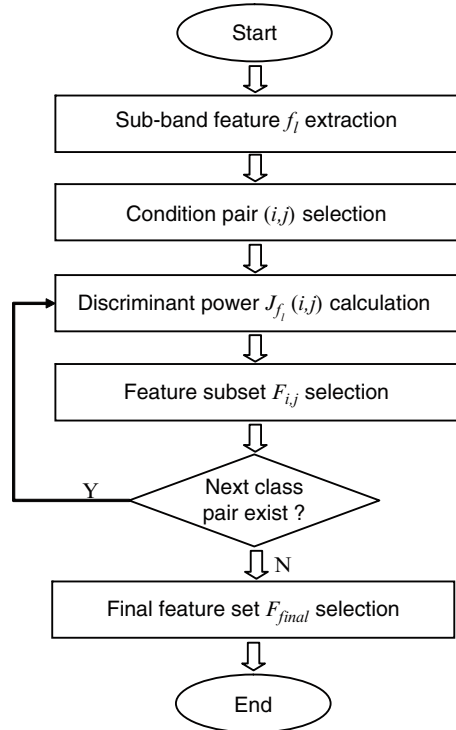


Fig. 8.3 Flowchart of the FLD feature selection process

$$J_{f_1}(i, j) \geq J_{f_2}(i, j) \geq \cdots \geq J_{f_d}(i, j) \geq \cdots \geq J_{f_p}(i, j) \quad (8.13)$$

The first group of d features that have the highest relative $J_{f_i}(i, j)$ values are chosen to form the feature subset $F_{i,j}$, for each class pair (i, j) :

$$F_{i,j} = \{ f_l | l = 1, 2, \dots, d \}, \quad i = 1, 2, \dots, k-1; j = i+1, i+2, \dots, k \quad (8.14)$$

The final feature set is obtained by taking the union of each feature subset across all the class pairs as:

$$F_{final} = \left\{ \bigcup_{i=1}^{L-1} \bigcup_{j=i+1}^L F_{i,j} \right\} \quad (8.15)$$

This feature set is subsequently selected for the machine defect severity classification.

8.2.2 Principal Component Analysis

PCA, as a multivariate statistical technique, has been intensively studied and utilized as an effective tool for process monitoring (Kano et al. 2001), structural damage identification (De Boe and Golival 2003), and machine health diagnosis (Baydar et al. 2001; He et al. 2008). This is due to its ability in dimension reduction and pattern classification. In general, the PCA technique seeks to determine a series of new variables, called the principal components, which indicates the maximal amount of variability in the data with a minimal loss of information (Jolliffe 1986), to best represent the data in a least square sense.

Suppose there are m feature vectors $FV_i (i = 1, 2, \dots, m)$ extracted from m signals, respectively. A $p \times m$ feature matrix X can then be formulated as:

$$X = [FV_1, FV_2, \dots, FV_m] \quad (8.16)$$

where the symbol FV denotes a p -dimensional feature vector as shown in (8.10). Depending on the decomposition level j of the WPT, the dimension of the feature vector is determined as $p = 2^{j+1}$. Correspondingly, a scatter matrix S is constructed from the feature matrix X as:

$$S = E[(X - \bar{X})(X - \bar{X})^T] \quad (8.17)$$

Where \bar{X} is the mean value of X , and $E[\bullet]$ is the statistical expectation operation (Duda et al. 2000). Performing singular value decomposition on the scatter matrix leads to:

$$S = A\Lambda A^T \quad (8.18)$$

where A is a $p \times p$ matrix whose columns are the orthonormal eigenvectors of the scatter matrix, and $A^T A = I_p$. The symbol Λ is a diagonal matrix whose diagonal elements $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ are the eigenvalues of the scatter matrix. Since the eigenvector in matrix A with the highest eigenvalue (i.e., λ_1 in the diagonal matrix Λ) is the first principle component of the p -dimensional feature vectors, it is better-suited than any other feature vectors as the representative feature that identifies the condition of the machine being monitored, for example, defect severity of a bearing. As a result, PCA ranks the order of eigenvectors by means of their respective eigenvalues, from the highest to the lowest. Such a ranking sequence reflects upon the order of significance of the corresponding components. By examining the accumulated variance (e.g., 90%) of the principle components, which is defined as:

$$\text{var} = \left(\frac{\sum_{i=1}^q \lambda_i}{\sum_{j=1}^p \lambda_j} \right) \times 100\% \quad (8.19)$$

a lower-dimensional feature vectors Y can be constructed as:

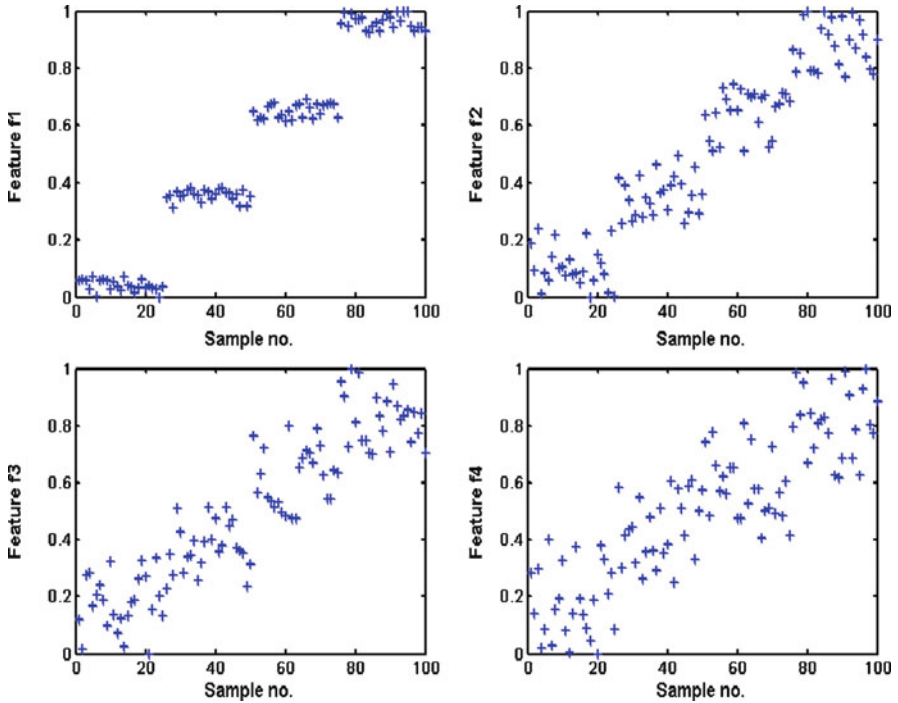


Fig. 8.4 Simulated data of (f_1, f_2, f_3, f_4) for developing a feature selection scheme

$$Y = A_{p \times q}^T X \quad (8.20)$$

where $q < p$, and $A_{p \times q}$ is the first q columns of A .

Given that the features transformed by the principal components are not directly connected to the physical nature of the defect, the eigenvectors in $A_{p \times q}$ for the transformed features are only used as the basis for choosing the most significant features from the original p -dimensional feature vectors. This is explained by means of a numerical simulation. As illustrated in Fig. 8.4, four normalized feature vectors, f_1, f_2, f_3 , and f_4 , are constructed with each of them forming clusters around four distinct levels of magnitudes. A total of 100 samples are considered for each of the four features, hence each feature is a 100-by-1 vector. The four features are simulated to have random variations from the same mean for each of the four clusters. This is similar in principle to the variation of a measured data feature for four different defect severities. Each of the four clusters for each feature contained 25 data points. The four features become less clearly differentiated from f_1 to f_4 , as overlap between the clusters increases.

It is evident that a suitable feature selection scheme should be able to rank f_1, f_2, f_3 , and f_4 in the same order as shown in Fig. 8.4. To derive the principal components for the simulated data set, the four normalized features are collected in a 4-by-100 matrix X :

$$X = [f_1, f_2, f_3, f_4]^T \quad (8.21)$$

The eigenvalues and the eigenvectors are calculated from the scatter matrix S . The matrix of eigenvectors can be represented as $A = [a_{ij}]$, where $i = 1$ to 4, and $j = 1$ to 4. The eigenvector a_4 consists of four components from the fourth column of the matrix A as $a_4 = [a_{1,4} \ a_{2,4} \ a_{3,4} \ a_{4,4}]$. Similar arrangement applies to a_1, a_2 , and a_3 (i.e., $a_1 = [a_{1,1} \ a_{2,1} \ a_{3,1} \ a_{4,1}]$, $a_2 = [a_{1,2} \ a_{2,2} \ a_{3,2} \ a_{4,2}]$, and $a_3 = [a_{1,3} \ a_{2,3} \ a_{3,3} \ a_{4,3}]$), respectively. The matrix A is a 4×4 square matrix because of the presence of the four features f_1 to f_4 . The eigenvector corresponding to the eigenvalue with the largest magnitude is chosen. As shown in Table 8.2, one of the four eigenvalues of the data set is much larger than the other three, indicating that most of the variance is concentrated in one direction. Table 8.3 lists the component magnitudes for the eigenvector corresponding to the largest eigenvalue. Since this corresponds to a_4 , the feature that is responsible for the maximum variance in the data is thus identified.

Subsequently, the magnitudes of the four components of e_4 are examined. As shown in Table 8.3, $|a_{1,4}| > |a_{2,4}| > |a_{3,4}| > |a_{4,4}|$. This result can be interpreted in terms of the directionality of the eigenvector (a_4) in the original feature space. If the unit vectors for the original feature space are represented as u_1, u_2, u_3 , and u_4 (where $u_1 = [1 \ 0 \ 0 \ 0]^T$, $u_2 = [0 \ 1 \ 0 \ 0]^T$, etc.), then a higher magnitude of $a_{i,4}$ denotes the similarity in direction of the eigenvector a_4 with u_i , when compared with the other unit vectors forming the basis for the original feature space. For the simulated data, the component $a_{1,4}$ has the largest magnitude, followed by $a_{2,4}$, $a_{3,4}$, and $a_{4,4}$.

Table 8.2 Eigenvalues for simulated data

λ_1	0.241
λ_2	0.775
λ_3	1.318
λ_4	32.023

Table 8.3 The fourth eigenvector component magnitudes for simulated data

Component	Magnitude
$a_{1,4}$	0.599
$a_{2,4}$	0.523
$a_{3,4}$	0.439
$a_{4,4}$	0.417

Thus, the feature represented along u_1 is the most sensitive, followed by those along u_2 , u_3 , and u_4 . As a result, the PCA-based scheme ranks the four features f_1 f_4 as desired and selects most representative features.

8.3 Neural-Network Classifier

Once a suitable feature set (e.g., 6) is chosen from the extracted features (e.g., 32), the machine defect severity levels can be evaluated by means of a status classifier. Neural network as a classifier has been applied to machine health diagnosis, for example, for classifying rotating machines with imbalance and rub faults (McCormick and Nandi 1997), bearing faults (Li et al. 2000), and tank reactor operation states (Maki and Loparo 1997). In general, a neural network consists of multiple layers of nodes or neurons, and each layer has a number of parallel nodes that are connected to all the nodes in the succeeding layer through different weights (Haykin 1994). Using a training algorithm, the weights are adjusted such that the neural network responds to the inputs with outputs corresponding to the severity of a structural defect at the output layer. Figure 8.5 illustrates the architecture of a feed-forward neural network. For the i th layer of links, the symbols $\mathbf{w}^{(i)}$, $\mathbf{b}^{(i)}$, $\mathbf{x}^{(i)}$, and $\mathbf{y}^{(i)}$ represent a vector of weights between the layers, node biases of a layer, inputs of nodes at one layer, and output at the output layer, respectively. At the output layer, a linear neuron is used to produce an output to indicate the machine defect severity level.

The weights of a multilayer feed-forward neural network are continuously updated, while it is trained with training data consisting of a set of machine defect feature input vectors (\mathbf{x}) and known output (d). This is realized by minimizing the error between the computed output of the network and the known output in the training process. Consider n pairs of input and output training data $\{(\mathbf{x}_p, d_p)\}$, $p = 1, 2, \dots, n$. For the p th pair data $\{\mathbf{x}_p, d_p\}$, the mean square error (MSE) of the network output y_p is expressed as:

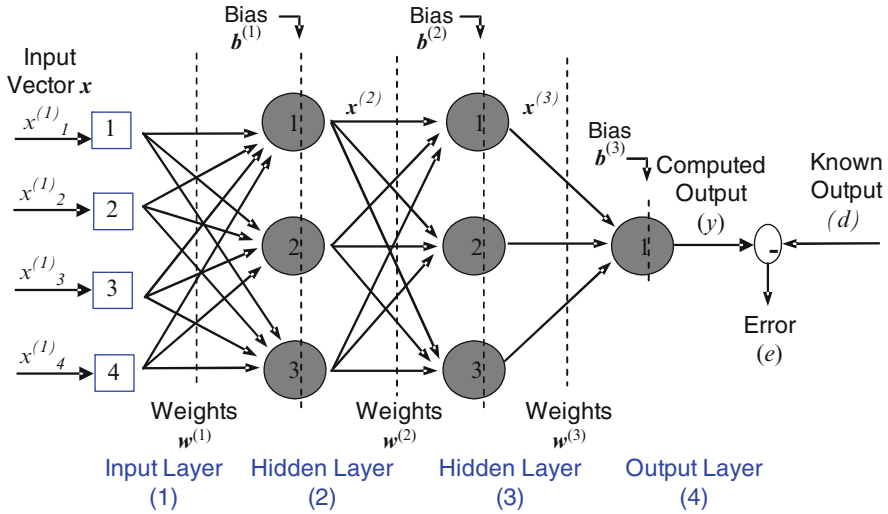


Fig. 8.5 Structure of a multilayer feed forward neural network

$$e_p = \sum_{m=1}^j (d_{pm} - y_{pm})^2 \quad (8.22)$$

where m is the number of nodes at the output layer. Assuming that each input vector corresponds to a single severity value, the value of m is 1. For the entire training data set, the total error Err , i.e., the learning error, is expressed as:

$$Err = \sum_{p=1}^n e_p = \sum_{p=1}^n \sum_{m=1}^j (d_{pm} - y_{pm})^2 \quad (8.23)$$

In the training process, the learning error is minimized through continuously updating the connection weights in its structure with certain learning rule. After training with the training data, the designed network with the resulting connection weights generalize the relationship between the input and output to correctly classify new input data. When input feature vectors associated with a defective measurement occur, for which the network is however not trained, the neural network will interpolate a defect severity by the location of the new input in the space spanned by the training data.

There are several gradient-based learning rules to minimize the learning error Err by changing the connection weight w of the multilayer feed-forward neural network. Different learning rules differ in how they use the gradients to update the weights w in training. Steepest decent with fixed learning rate is the traditional learning rule of the neural network, in which the weight $w^{(k)}$ between the k th layer and $(k + 1)$ th layer is tuned for each epoch, along the gradient direction by an amount:

in each subband, key feature selection process is performed to determine the most significant features from the feature set, which are subsequently used as input to a neural network-based classifier for defect severity classification. Figure 8.6 illustrates how the developed technique is realized. The left side of Fig. 8.6 depicts the training process of the hybrid technique in a manner of supervised learning (i.e., based on the available reference data, denoted as signal 1, \dots , n). In addition to providing inputs for constructing the neural-network classifier model, the results from the feature selection process are used to guide the feature vector selection in the evaluation process. The right side of Fig. 8.6 describes an evaluation process of the WPT-based algorithm. An input signal is passed through the process of signal decomposition, feature extraction, and selection. Eventually, the corresponding defect severity level is determined through the neural network classifier.

8.5 Case Studies

The application of the above described wavelet packet-based machine defect severity classification algorithm is described through two case studies.

8.5.1 Case Study I: Roller Bearing Defect Severity Evaluation

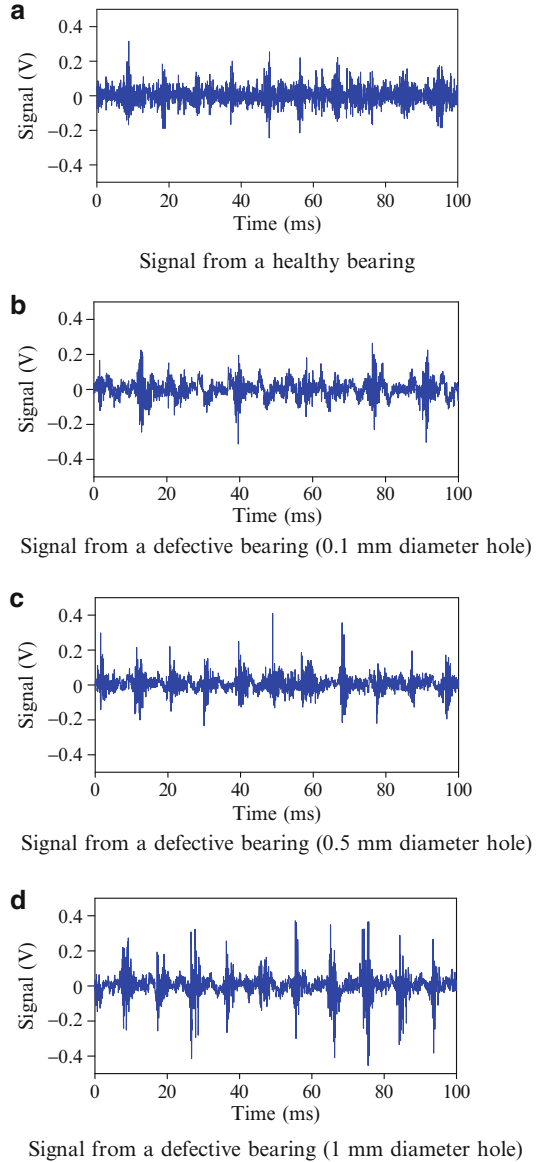
The first case study is to evaluate the defect severity level of a set of roller bearings (N205 ECP) with and without seeded defects. Specifically, vibration data were measured from both a new, “healthy” bearing that served as a reference base and three defective bearings containing localized defects of different sizes:

- (a) one 0.1 mm diameter hole in the outer raceway
- (b) one 0.5 mm diameter hole in the outer raceway
- (c) one 1 mm diameter hole in the outer raceway

In Fig. 8.7, segments of vibration signals measured from the healthy and defective bearings are shown.

To provide sufficient training and testing data sets to the neural-network classifier, a total of 240 vibration data sets were collected under a bearing rotating speed of 1,200 rpm and a radial load of 3,600 N. For each operation condition, 60 data sets were collected. Each data set was first decomposed by the WPT. Analysis has shown that features extracted from a four-level decomposition provided adequate information to differentiate the four defect conditions from each other (Gao and Yan 2007). On the basis of the information collected in the 16 subfrequency bands (since $2^4 = 16$), a feature vector was subsequently constructed, which contained 32 feature elements (i.e., 16 subband energy values and 16 subband kurtosis values).

Fig. 8.7 Vibration signals measured from roller bearings with different conditions. (a) Signal from a healthy bearing. (b) Signal from a defective bearing (0.1 mm diameter hole). (c) Signal from a defective bearing (0.5 mm diameter hole). (d) Signal from a defective bearing (1 mm diameter hole)



FLD analysis is then applied for feature selection. The means and variances of the feature element, f_i , are obtained for each of the four bearing conditions. Table 8.4 summarizes the discriminant power of the extracted features for different condition pairs, based on the Fisher discriminant criterion. The first three key features within each condition pair, for example, E_4^{12} , E_4^{13} , and E_4^{14} for the condition pair (healthy, light defect), are selected, and the final feature set is obtained through a union

Table 8.4 Discriminant power of the extracted features for various condition pairs in different subbands

Subband features	Healthy vs. light	Healthy vs. medium	Healthy vs. severe	Light vs. medium	Light vs. severe	Medium vs. severe
E_4^0	0.52	2.55	1.80	2.58	2.19	2.15
E_4^1	0.06	48.91	8.57	0.43	0.05	10.49
E_4^2	18.29	695.40	118.62	41.28	8.37	193.98
E_4^3	0.01	110.73	8.86	1.06	0.97	27.86
E_4^4	1,222.60	92.10	89.62	1,696.00	216.26	52.50
E_4^5	45.20	212.92	125.51	374.81	260.53	11.15
E_4^6	41.60	346.97	61.40	2.88	0.08	5.64
E_4^7	226.86	440.07	191.23	308.11	88.32	71.01
E_4^8	2.41	172.13	7.43	173.16	0.01	183.39
E_4^9	87.38	47.61	4.48	69.18	204.47	52.60
E_4^{10}	466.15	8.60	211.70	916.98	170.90	340.01
E_4^{11}	80.74	14.86	69.41	60.15	15.69	44.79
E_4^{12}	5,118.80	1,238.00	133.33	7,946.10	873.16	12.21
E_4^{13}	12,702.00	308.19	2,280.10	7,397.60	3,293.60	788.82
E_4^{14}	3,652.80	36.25	2,374.40	2,414.90	707.70	1,414.70
E_4^{15}	1,863.40	56.36	12,956.60	730.50	86.52	1,150.90
K_4^0	2.54	57.50	41.34	0.33	0.10	7.88
K_4^1	0.01	0.33	4.04	0.01	0.01	2.30
K_4^2	0.02	17.60	7.80	0.03	0.01	16.22
K_4^3	0.04	6.98	37.80	0.04	0.03	51.64
K_4^4	0.91	39.95	9.99	0.03	0.25	3.07
K_4^5	0.34	74.66	50.31	0.18	0.15	1.06
K_4^6	0.07	2.33	7.47	0.09	0.10	8.72
K_4^7	7.81	30.13	21.94	1.84	1.90	10.41
K_4^8	0.98	51.73	6.09	0.05	0.14	0.36
K_4^9	0.40	74.16	32.09	0.10	0.18	7.43
K_4^{10}	0.25	14.71	9.05	0.02	0.06	0.51
K_4^{11}	0.89	15.32	38.21	0.11	0.38	3.09
K_4^{12}	0.30	4.62	11.99	0.04	0.02	0.93
K_4^{13}	0.02	9.95	2.88	0.01	0.02	0.92
K_4^{14}	0.03	16.09	3.15	0.02	0.03	2.39
K_4^{15}	0.01	9.42	4.90	0.01	0.01	1.51

operation from all the six condition pairs as listed in Table 8.5, where the energy features E_4^2 , E_4^7 , E_4^{12} , E_4^{13} , E_4^{14} , and E_4^{15} are selected as the most representative features, because they possess higher discriminant power than others as listed in Table 8.4.

Next, the PCA technique was performed on the feature vectors. As seen in Fig. 8.8, the first five principal components represent over 90% variance, which preserves most of the information contained in the original feature set (Jolliffe 1986). This is

Table 8.5 Final feature set obtained through a union operation by Fisher linear discriminant analysis

Subband features	Healthy vs. light	Healthy vs. medium	Healthy vs. severe	Light vs. medium	Light vs. severe	Medium vs. severe	Final feature set
E_4^2		✓					✓
E_4^7		✓					✓
E_4^{12}	✓	✓		✓	✓		✓
E_4^{13}	✓		✓	✓	✓	✓	✓
E_4^{14}	✓		✓	✓	✓	✓	✓
E_4^{15}			✓			✓	✓

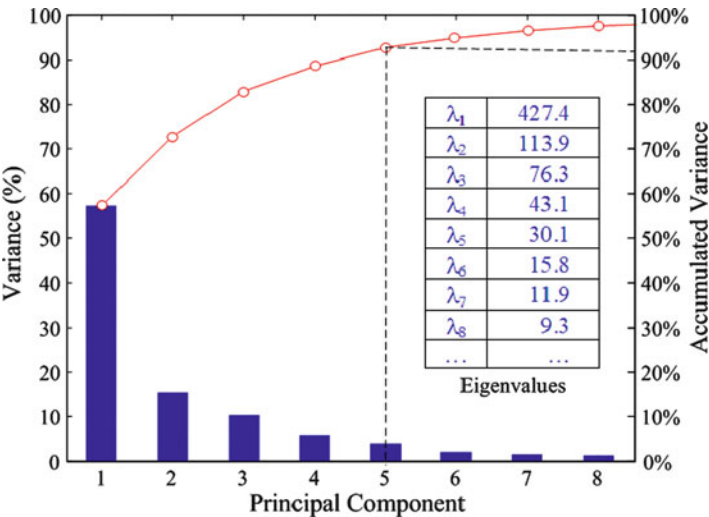


Fig. 8.8 Accumulated variance of principal components for the tested bearings

considered sufficient for constructing the corresponding subspace principle component matrix (based on their corresponding eigenvectors) for choosing the features from the original feature vector. Table 8.6 lists the eigenvectors that correspond to the first five principal components. By searching for those components with the largest magnitude in each eigenvector, the corresponding energy values E_4^0 and E_4^3 , kurtosis values K_4^1 , K_4^3 , and K_4^{15} are identified as the most representative features.

The selected feature set was input to a multiplayer perception (MLP) neural network for bearing defect severity classification. Since different ratios (e.g., 70 30 or 50 50) for the training and testing data were suggested for neural network-based classifier in the literature (Paya et al. 1997; Jack and Nandi 2001), but no single fixed ratio has been preferred, two thirds of the data sets corresponding to each condition were used for training the classifier, and the remaining one third for performance checking, from a total of 240 collected data sets. This was aimed at providing sufficient training data to ensure accuracy of the classifier. The classification rates are listed in Table 8.7. When the FLD-selected

Table 8.6 The first five eigenvectors of the extracted features for the roller bearing

Subband features	a_1	a_2	a_3	a_4	a_5
E_4^0	0.382	0.106	0.605	0.222	0.069
E_4^1	0.039	0.128	0.186	0.013	0.039
E_4^2	0.060	0.115	0.028	0.388	0.079
E_4^3	0.030	0.163	0.052	0.501	0.002
E_4^4	0.045	0.025	0.052	0.133	0.035
E_4^5	0.078	0.015	0.194	0.030	0.011
E_4^6	0.030	0.013	0.117	0.094	0.016
E_4^7	0.042	0.046	0.195	0.044	0.014
E_4^8	0.068	0.100	0.225	0.270	0.044
E_4^9	0.098	0.103	0.252	0.284	0.045
E_4^{10}	0.046	0.012	0.032	0.069	0.001
E_4^{11}	0.085	0.019	0.177	0.185	0.004
E_4^{12}	0.026	0.001	0.033	0.012	0.005
E_4^{13}	0.021	0.001	0.012	0.016	0.002
E_4^{14}	0.021	0.008	0.015	0.040	0.004
E_4^{15}	0.012	0.009	0.017	0.041	0.004
K_4^0	0.012	0.063	0.236	0.067	0.010
K_4^1	0.482	0.445	0.138	0.337	0.422
K_4^2	0.241	0.273	0.088	0.132	0.347
K_4^3	0.459	0.156	0.081	0.101	0.745
K_4^4	0.080	0.017	0.225	0.100	0.023
K_4^5	0.193	0.104	0.176	0.045	0.156
K_4^6	0.030	0.108	0.080	0.288	0.031
K_4^7	0.016	0.035	0.073	0.133	0.010
K_4^8	0.036	0.011	0.114	0.014	0.005
K_4^9	0.085	0.001	0.120	0.044	0.038
K_4^{10}	0.059	0.041	0.136	0.025	0.019
K_4^{11}	0.056	0.063	0.093	0.010	0.049
K_4^{12}	0.033	0.039	0.295	0.063	0.078
K_4^{13}	0.294	0.361	0.124	0.113	0.139
K_4^{14}	0.276	0.297	0.128	0.050	0.056
K_4^{15}	0.295	0.598	0.028	0.208	0.263

Table 8.7 Neural network classifier results of the roller bearing

Classification rate	WPT features with FLD (%)	WPT features with PCA (%)	WPT features only (%)	Raw data features (%)
No defect	100	95	90	85
0.1 mm hole	95	80	70	60
0.5 mm hole	100	95	95	95
1 mm hole	100	100	100	95
Overall	99	92	88	83

feature set was used as input to the MLP classifier, only 5% of measured data with the 0.1-mm hole in the bearing outer raceway is misclassified, out of the whole test data. This led to 98% overall classification success. When the PCA-selected feature set was used as the MLP input, a classification rate of 92% was identified, which is lower than the FLD-selected feature set. The rate, on the contrary, is still higher than the rates obtained using WPT features only as the input to MLP (88%) and using raw data features as the input (83%). This illustrates that the WPT-based feature extraction and selection method is effective in defect severity classification.

8.5.2 Case Study II: Ball Bearing Defect Severity Evaluation

For the second case study, a run-to-failure experiment was conducted on a deep groove ball bearing (type 1100KR) of 52 mm outer diameter, under a radial load of 5,498 N. The bearing contained a 0.27-mm wide groove across its outer raceway as an embedded defect, and was continually run under a rotational speed of 2,000 rpm. Upon reaching approximately 2.7 million revolutions, the defect has propagated throughout the entire raceway and rendered the bearing practically nonfunctional. This case study was designed to investigate the effect of continuous degradation of the defect, whereas case study I discussed above concerns with the effect of discrete defects.

Vibration signals were collected during the experiment at an interval of every 7 min. Figure 8.9 illustrates the trend of the vibration amplitude along the process of defect propagation. Three vibration signals are also shown in Fig. 8.10, and they are measured right after the bearing is physically examined at different test stages. For

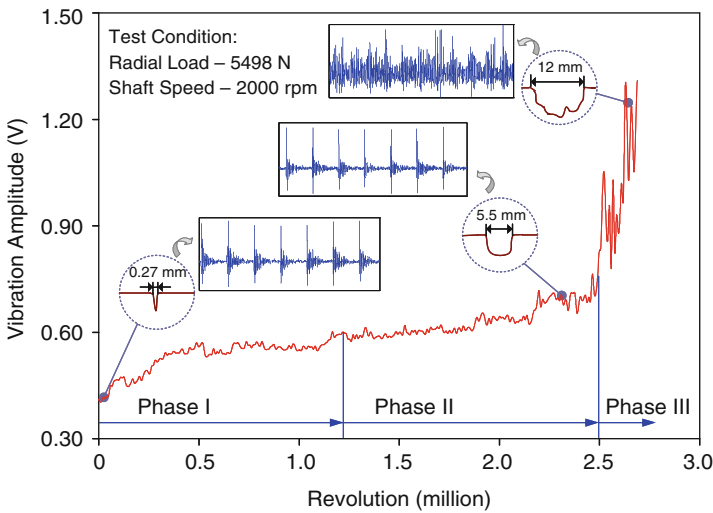


Fig. 8.9 Amplitude as a function of the ball bearing revolution

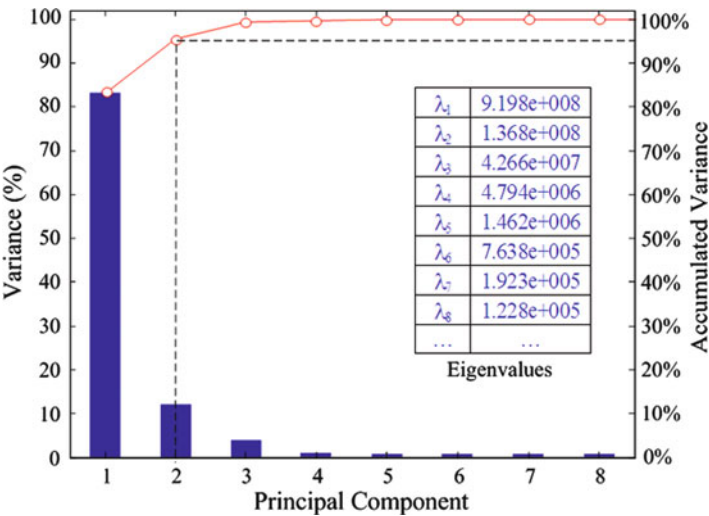


Fig. 8.10 Accumulated variance of principal components for the ball bearing 1100KR

purpose of defect severity evaluation, all the collected vibration data sets are divided into three sections with the threshold values of the amplitude being set at 0.6 and 0.9 V, respectively. As shown in Fig. 8.9, the three sections correspond to three different defect propagation phases during the run-to-failure test. It should be noted that, since no prior knowledge was available regarding the relationship between the vibration amplitude and the defect severity level, the choice of three phases investigated here is empirical.

The vibration signals were first decomposed into 16 subbands. The energy and kurtosis features were then calculated from the wavelet packet coefficients in each subband to formulate the feature vectors. FLD analysis was then used for feature selection. The means and variances of the feature element, f_i , were obtained for each of the four bearing initial conditions. Table 8.8 summarizes the discriminant power of the extracted features for different phase pairs, based on the Fisher discriminant criterion. The first three key features within each phase pair were selected and the final feature set was obtained through a union operation among different phase pairs (i.e., phase I, phase II; phase I, phase III; and phase II, phase III). As listed in Table 8.9, the kurtosis features K_4^0 , K_4^1 , K_4^2 , K_4^6 , and K_4^{10} are selected as the most representative features.

When the PCA was performed on the extracted subband feature vectors, the first two principal components shown in Fig. 8.10 represent over 90% variance, which was subsequently used as the reference features from the original feature set. Table 8.10 lists the eigenvectors corresponding to the first two principal components. The highest magnitude in the first eigenvector was found to be associated with the first component, and the highest magnitude in the second eigenvector was seen to be related to the third component. Accordingly, the

Table 8.8 Discriminant power of the extracted features for the ball bearing phase pair in various subbands

Subband features	Phase I vs. phase II	Phase I vs. phase III	Phase II vs. phase III
E_4^0	1.08×10^8	1.25×10^8	2.66×10^9
E_4^1	1.30×10^8	1.15×10^8	2.65×10^8
E_4^2	2.23×10^9	1.72×10^8	5.18×10^8
E_4^3	6.13×10^7	6.81×10^7	4.38×10^9
E_4^4	3.54×10^7	4.37×10^8	4.05×10^{10}
E_4^5	5.54×10^7	2.54×10^7	2.90×10^{10}
E_4^6	1.81×10^8	5.13×10^9	7.44×10^8
E_4^7	1.82×10^5	1.42×10^6	7.26×10^7
E_4^8	9.94×10^7	5.05×10^7	9.64×10^7
E_4^9	5.54×10^6	3.23×10^6	5.07×10^6
E_4^{10}	5.59×10^4	1.13×10^6	3.62×10^7
E_4^{11}	2.72×10^3	1.49×10^5	1.69×10^6
E_4^{12}	2.15×10^6	1.80×10^7	8.29×10^9
E_4^{13}	5.39×10^6	8.77×10^7	9.99×10^8
E_4^{14}	7.55×10^4	3.40×10^6	2.26×10^6
E_4^{15}	1.24×10^3	1.75×10^5	9.36×10^6
K_4^0	2.07×10^1	2.54E+01	1.75E+00
K_4^1	8.18×10^2	4.69×10^1	5.16×10^1
K_4^2	2.35×10^3	1.39E+00	3.14×10^1
K_4^3	9.96×10^5	4.10×10^2	4.11×10^2
K_4^4	7.98×10^4	1.93×10^5	1.34×10^5
K_4^5	3.72×10^4	4.19×10^3	4.61×10^3
K_4^6	1.22×10^3	9.56×10^1	2.37×10^1
K_4^7	6.30×10^6	4.94×10^2	9.81×10^3
K_4^8	3.26×10^4	2.18×10^6	2.76×10^6
K_4^9	2.85×10^5	9.04×10^7	9.47×10^7
K_4^{10}	3.05×10^3	1.05×10^6	8.54×10^7
K_4^{11}	1.57×10^4	6.45×10^7	7.02×10^7
K_4^{12}	1.13×10^3	3.02×10^5	2.23×10^5
K_4^{13}	5.42×10^4	3.45×10^6	2.04×10^7
K_4^{14}	2.64×10^4	5.86×10^6	3.50×10^6
K_4^{15}	7.76×10^4	6.13×10^5	9.54×10^6

Table 8.9 Final feature set obtained for the ball bearing through a union operation by FLD

Subband features	Phase I vs. phase II	Phase I vs. phase III	Phase II vs. phase III	Final feature set
K_4^0	✓	✓	✓	✓
K_4^1	✓		✓	✓
K_4^2		✓	✓	✓
K_4^6		✓		✓
K_4^{10}	✓			✓

Table 8.10 The first two eigenvectors calculated the extracted features for the ball bearing

Subband features	a_1	a_2
E_4^0	0.997	0.022
E_4^1	0.033	0.229
E_4^2	0.024	0.942
E_4^3	0.027	0.037
E_4^4	0.027	0.012
E_4^5	0.023	0.001
E_4^6	0.029	0.241
E_4^7	0.021	0.004
E_4^8	0.001	0.001
E_4^9	0.001	0.001
E_4^{10}	0.003	0.002
E_4^{11}	0.001	0.001
E_4^{12}	0.012	0.006
E_4^{13}	0.006	0.004
E_4^{14}	0.003	0.002
E_4^{15}	0.003	0.002
K_4^0	0.001	0.001
K_4^1	0.001	0.001
K_4^2	0.001	0.001
K_4^3	0.001	0.001
K_4^4	0.001	0.001
K_4^5	0.001	0.001
K_4^6	0.001	0.001
K_4^7	0.001	0.001
K_4^8	0.001	0.001
K_4^9	0.001	0.001
K_4^{10}	0.001	0.001
K_4^{11}	0.001	0.001
K_4^{12}	0.001	0.001
K_4^{13}	0.001	0.001
K_4^{14}	0.001	0.001
K_4^{15}	0.001	0.001

Table 8.11 Results of neural network classification rate of results of the ball bearing

Classification rate	WPT features with FLD (%)	WPT features with PCA (%)	WPT features only (%)	Raw data features (%)
Phase I	92	87	82	79
Phase II	91	84	81	77
Phase III	94	88	88	82
Overall	92	86	82	78

energy value at subbands 1 and 3 were (denoted as E_4^0 and E_4^2) identified as the most representative features.

Following the same procedure as described in case study I, 2/3 of the data sets corresponding to each defect propagation phase are used for training the MLP classifier, and the remaining 1/3 data points are used for performance checking. As shown in Table 8.11, when the features selected from the FLD approach were used as input to the MLP classifier, the classification rate for each phase is found to be 92%, 91%, and 94%, respectively. This led to the overall classification rate of 92%. In comparison, when features selected using PCA technique were used as input to the MLP, the classification rates were lower, 87%, 84%, and 88%, respectively. Furthermore, when feature set extracted from each subband and raw data was directly used as the MLP input, the classification rates dropped down to even lower values (e.g., 82% overall classification rate for WPT features only, and 78% overall classification rate for raw data features). This indicates again the effectiveness of the presented approach for defect severity classification.

8.6 Summary

This chapter introduces a wavelet packet-based signal processing approach for machine defect severity classification. After the subband energy and kurtosis features are extracted from realistic vibration signals using the wavelet-packet coefficients, the most representative features are chosen using the Fisher discriminant criterion and principal feature analysis, respectively. These features are used as inputs to the neural-network classifiers to evaluate the machine defect severity. The effectiveness of the approach has been experimentally verified through two case studies for rolling bearing defect severity classification. It is shown that the introduced approach provides a practical way for feature extraction and selection. In addition to bearing defect severity classification, this approach is applicable to classifying the working states of other machines and machine components, thus providing a useful tool for machine condition monitoring and diagnosis.

8.7 References

- Altmann J, Mathew J (2001) Multiple band pass autoregressive demodulation for rolling element bearing fault diagnosis. *Mech Syst Signal Process* 15:963–977
- Baydar N, Chen Q, Ball A, Kruger U (2001) Detection of incipient tooth defect in helical gears using multivariate statistics. *Mech Syst Signal Process* 15:303–321
- De Boe P, Golinvall JC (2003) Principal component analysis of a piezo sensor array for damage localization. *Int J Struct Health Monit* 2(2):137–144
- Duda R, Hart P, Stork D (2000) *Pattern classification*. Wiley Interscience, New York.
- Fan X, Zuo MJ (2006) Gearbox fault detection using Hilbert and wavelet packet transform. *Mech Syst Signal Process* 20:966–982

- Fukunaga K (1990) Introduction to statistical pattern recognition, 2nd edn. Academic, New York
- Gao R, Yan R (2007) Wavelet packet transform based hybrid signal processing for machine health monitoring and diagnosis. In: The 6th international workshop on structural health monitoring, Stanford, CA, pp 598–605
- Goumas SK, Zervakis ME, Stavrakakis GS (2002) Classification of washing machine vibration signals using discrete wavelet analysis for feature extraction. *IEEE Trans Instrum Meas* 51 (3):497–508
- Haykin, S (1994) Neural networks. Macmillan Publishing Company, New York
- He Q, Yan R, Kong F, Du R (2008) Machine condition monitoring using principle component representation. *Mech Syst Signal Process.* 23(2):446–466
- Jack LB, Nandi AK (2001) Support vector machines for detection and characterization of rolling element bearing faults. *Proc Inst Mech Eng* 215:1065–1074
- Jolliffe IT (1986) Principal component analysis. Springer Verlag New York Inc, New York
- Kano M, Hasebe S, Hashimoto I (2001) A new multivariate statistical process monitoring method using principal component analysis. *Comput Chem Eng* 25:1103–1113
- Kittler J (1975) Mathematical methods of feature selection in pattern recognition. *Int J Man Mach Stud* 7(5):609–637
- Lee BY, Tang YS (1999) Application of the discrete wavelet transform to the monitoring of tool failure in end milling using the spindle motor current. *Int J Adv Manuf Technol* 15(4):238–243
- Li B, Chow M, Tipsuwan Y, Hung JC (2000a) Neural network based motor rolling bearing fault diagnosis. *IEEE Trans Ind Electron* 47(5):1060–1069
- Li XL, Tso SK, Wang J (2000b) Real time tool condition monitoring using wavelet transforms and fuzzy techniques. *IEEE Trans Syst Man Cybern C Appl Rev* 30(3):352–357
- Liu B, Ling SF, Meng Q (1997) Machinery diagnosis based on wavelet packets. *J Vib Control* 3:5–17
- Maki Y, Loparo KA (1997) A neural network approach to fault detection and diagnosis in industrial processes. *IEEE Trans Control Syst Technol* 5(6):529–541
- Malhi A, Gao R. (2004) PCA based feature selection scheme for machine defect classification. *IEEE Trans Instrum Meas* 53(6):1517–1525
- McCormick AC, Nandi AK (1997) Classification of the rotating machine condition using artificial neural networks. *Proc Inst Mech Eng C* 211:439–450
- Mori K, Kasashima N, Yoshioka T, Ueno Y (1996) Prediction of spalling on a ball bearing by applying the discrete wavelet transform to vibration signals. *Wear* 195:162–168
- Paya BA, Esat II, Badi MNM (1997) Artificial neural network based fault diagnosis of rotating machinery using wavelet transforms as a preprocessor. *Mech Syst Signal Process* 11 (5):751–765
- Prabhakar S, Mohanty AR, Sekhar AS (2002) Application of discrete wavelet transform for detection of ball bearing race faults. *Tribol Int* 35(12):793–800
- Yan R, Gao R (2004) Harmonic wavelet packet transform for on line system health diagnosis. SPIE international symposium on sensors and smart structures technologies for civil, mechanical and aerospace systems, San Diego, CA, pp 512–522
- Yen G, Lin K (2000) Wavelet packet feature extraction for vibration monitoring. *IEEE Trans Ind Electron* 47(3):650–667

Chapter 9

Local Discriminant Bases for Signal Classification

The goal of analyzing signals from manufacturing machines is to extract relevant *features* from the waveforms to effectively characterize the working conditions of the machines (e.g., tool breakage and gear degradation). As we have shown in Chap. 5, the wavelet packet transform can lead to redundant signal decomposition within certain time frequency subspaces. When performing wavelet packet transform, the time frequency subspaces are collectively called the wavelet packet library. Each of the subspaces is denoted as a wavelet packet node. Such a way of signal decomposition provides the possibility of selecting a particularly suited set of wavelet packet nodes out of the wavelet packet library for a specific signal analysis task, such as data compression, regression, or classification (Saito 1994). However, the choice of wavelet packet nodes is dependent on the specific task. For example, the optimal wavelet packet transform technique introduced in Chap. 5 is geared toward signal compression (Coifman and Wicherhauser 1992), in which the wavelet packet nodes are selected based on minimizing an information cost function (e.g., Shannon entropy). This chapter introduces how to choose a good set of wavelet packet nodes from a wavelet packet library, for purpose of signal classification. Such a technique has shown to be effective for monitoring and diagnosis of rotating machines.

9.1 Dissimilarity Measures

To classify signals obtained from a machine under different working status, the *features* extracted from the signals should clearly differentiate different working status of the machine, where each status is considered as a distinct *class*. For example, signals measured on a new gearbox are denoted as one class, while signals measured on a gearbox with broken-teeth are denoted as another class. Such type of features is referred to as “discriminant” features of the signal. The main objective of signal classification by using the wavelet packet transform is to find an optimal set of wavelet packet nodes (each node representing a wavelet packet basis) that are capable of discriminating different *classes* as effectively as possible. This can be achieved by decomposing the signal of interest into different classes, using the local

discriminant bases (LDB) algorithm (Saito 1994; Saito and Coifman 1995). The optimal choice of LDBs depends on the nature of the signals and the dissimilarity measures used to distinguish classes. In general, the dissimilarity measure is aimed at evaluating the “statistical distances” of each wavelet packet node among different classes. Numerous dissimilarity measures have been developed (Basseville 1989; Saito 1994; Saito et al. 2002; Umapathy and Krishnan 2006; Umapathy et al. 2007), among which the following four measures have been typically associated with the application of the LDB algorithm.

9.1.1 Relative Entropy

Relative entropy is one of the first dissimilarity measures used for identifying the LDBs (Saito 1994). On the basis of the definition of relative entropy, this dissimilarity measure is defined as:

$$D_1(p^1, p^2) = \sum_{i=1}^n p_i^1 \log \frac{p_i^1}{p_i^2} \quad (9.1)$$

where $\sum_{i=1}^n p_i^1 = 1$ and $\sum_{i=1}^n p_i^2 = 1$. The symbols p^1 and p^2 denote nonnegative sequences, respectively. It is assumed that $\log 0 = -\infty$, $\log(x/0) = +\infty$ for $x > 0$, and $0 \times (\pm\infty) = 0$. The discriminant information $D_1(p^1, p^2)$ between these two sequences measures how differently p^1 and p^2 are distributed. From the definition, it is seen that the nonnegative sequences p^1 and p^2 can be considered as probability density function. Since the normalized energy of wavelet coefficients (i.e., representation of energy distribution) within each wavelet packet basis is actually an expression of the probability density function associated with that wavelet packet node, it can be used as replacement in (9.1). Consequently, the relative entropy can be used as a dissimilarity measure in the LDB algorithm. Furthermore, (9.1) indicates that the relative entropy measure D_1 is nonnegative and will be zero if the two sequences of p^1 and p^2 are the same. The more separate from each other the two sequences are, the higher the relative entropy measure D_1 will be. However, it should be noted that the relative entropy measure shown in (9.1) is only applicable to a two-class problem. For multiple-class problems (e.g., gearbox under four different conditions: such as (a) faultless, (b) slight-worn, (c) medium-worn, and (d) broken-teeth), the dissimilarity measure based on relative entropy is modified as:

$$D_1(\{p^m\}_{m=1}^L) = \sum_{a=1}^{L-1} \sum_{b=a+1}^L D_1(p^a, p^b) \quad (9.2)$$

where L is the number of classes. Equation (9.2) indicates that the dissimilarity measure of multiple-class problems is the summation of relative entropy for each pair of two-classes among all the classes.

9.1.2 Energy Difference

From the decomposition results of a signal's wavelet packet transform, the normalized energy associated with the wavelet packet node (j, k) is calculated as:

$$E_{j,k} = \frac{\sum_{l=1}^M |x_{j,k,l}|^2}{E_{x(t)}} \quad (9.3)$$

In (9.3), the symbols j and k represent the wavelet packet decomposition level and subfrequency band, respectively. These two symbols, collectively, represent a wavelet packet node (j, k) . The symbol $x_{j,k,l}$ denotes the l th wavelet packet coefficient within the node (j, k) , and the symbol M denotes the total number of coefficients within that node. $E_{x(t)}$ is the total energy contained in the signal.

The difference in the normalized energy associated with wavelet packet node (j, k) between the signals from two classes (denoted as class 1 and class 2) can be defined as a dissimilarity measure, given by:

$$D_2(E^1, E^2) = E_{j,k}^1 - E_{j,k}^2 \quad (9.4)$$

The symbols $E_{j,k}^1$ and $E_{j,k}^2$ represent the normalized energy associated with wavelet packet node (j, k) from class 1 and class 2 signals, respectively. Since each wavelet packet node corresponds to a time frequency subspace, the normalized energy computed at a node provides the energy distribution of the signal in a particular subfrequency band. The greater the difference at a particular node in the energy distribution of the two classes, the more significant the node for discriminating the classes will be. Similar to the definition expressed by (9.1), (9.4) represents the energy difference measure used for a two-class problem. For multiple-class problems, the dissimilarity measure based on the energy difference is expressed as:

$$D_2(\{E^m\}_{m=1}^L) = \sum_{a=1}^{L-1} \sum_{b=a+1}^L D_2(E^a, E^b) \quad (9.5)$$

where L is the number of classes. Equation (9.5) indicates the dissimilarity measure of multiple-class problems is the summation of energy difference for each pair of two-classes among all the classes.

9.1.3 Correlation Index

The dissimilarity measure can also be defined from the correlation between the wavelet packet node (j, k) from class 1 and class 2. This measure can be used to identify those nodes that can detect the difference in the temporal characteristics of

the signals between class 1 and class 2. The dissimilarity measure based on the correlation index, which is used in a two-class problem, is formulated as:

$$D_3(x^1, x^2) = \langle x_{j,k,l}^1, x_{j,k,l}^2 \rangle \quad (9.6)$$

where the symbols j , k , and l represent decomposition level, subfrequency band, and time position, respectively, and $x_{j,k,l}^1$ and $x_{j,k,l}^2$ are the coefficients of the corresponding wavelet packet node (j, k) of class 1 and class 2. An average low correlation index at a particular node indicates high dissimilarity between the classes. Similarly, for a multiple class problem, the dissimilarity measure based on correlation index is expressed as:

$$D_3(\{x^m\}_{m=1}^L) = \sum_{a=1}^{L-1} \sum_{b=a+1}^L D_3(x^a, x^b) \quad (9.7)$$

where L is the number of classes. Equation (9.7) indicates that the dissimilarity measure of multiple-class problems is the summation of correlation index for each pair of two-classes among all the classes.

9.1.4 Nonstationarity

Nonstationarity of the wavelet packet coefficients may also be used to measure the dissimilarity. It is computed as the set of variances along the segments of the wavelet packet coefficients at a given node (j, k) . The ratio of this variance between class 1 and class 2 indicates the amount of deviation in the nonstationarity between the two classes. Consequently, the dissimilarity measure based on nonstationarity, which is used in a two-class problem, can be defined as:

$$D_4(v^1, v^2) = \frac{\text{var}[v_{j,k}^1]}{\text{var}[v_{j,k}^2]} \quad (9.8)$$

where the symbols j and k represent the decomposition level and subfrequency band of the wavelet coefficients, respectively. The symbols v^1 and v^2 are variance vectors. Each of them contains L variances, obtained by equally segmenting the wavelet packet coefficients at node (j, k) for class 1 and class 2 signals, respectively. For example, given a signal in class 1 with 4,096 data points, there will be 1,024 wavelet packet coefficients at node $(2, 1)$, since $4,096/2^2 = 1,024$. If these wavelet packet coefficients are equally partitioned into eight segments (i.e., $L = 8$), then there will be eight elements in variance vector v^1 . Each of the elements is calculated from 128 wavelet packet coefficients ($1,024/8 = 128$). Variance vector v^2 can be obtained in the same way.

Similarly, for multiple class problems, the dissimilarity measure based on nonstationarity is expressed as:

$$D_4(\{v^m\}_{m=1}^L) = \sum_{a=1}^{L-1} \sum_{b=a+1}^L D_4(v^a, v^b) \quad (9.9)$$

where L is the number of classes. Equation (9.9) indicates that the dissimilarity measure of multiple class problems is the summation of nonstationarity for each pair of two-classes among all the classes.

9.2 Local Discriminant Bases

Utilizing one of the dissimilarity measures introduced earlier (e.g., relative entropy), the LDB algorithm can identify wavelet packet nodes that exhibit high discrimination, as indicated by a large statistical distance among the classes.

Let us assume that $\Omega_{0,0}$ denotes the wavelet packet node 0 of the parent tree (i.e., the signal itself). Then at each level, the wavelet packet node $\Omega_{j,k}$ is split into two mutually orthogonal subspaces (i.e., nodes $\Omega_{j+1,2k}$ and $\Omega_{j+1,2k+1}$), given by

$$\Omega_{j,k} = \Omega_{j+1,2k} \oplus \Omega_{j+1,2k+1} \quad (9.10)$$

where j indicates the level of the tree, and k represents the node index in level j , given by $k = 0, \dots, 2^j - 1$. This process is repeated until level J , giving rise to 2^J mutually orthogonal subspaces. The goal is to select a set of best subspaces that provide maximum dissimilarity information among different classes of the signals. This can be realized by a *pruning* approach, where the wavelet packet tree is pruned in such a way that, starting from the bottom decomposition level, a node is *split* if the cumulative discriminative measure of the *children* nodes is greater than that of the *parent* node. In other words, a node is split if the children nodes have better discriminative power than that of the parent node. Such a process is executed until it reaches the top level of the decomposition. As a result, the process will end with a subset of wavelet packet nodes that contribute to maximizing the statistical distance among different classes. As an example, Fig. 9.1 shows a wavelet packet tree for a two-level signal decomposition. The LDB algorithm first compares the discriminant

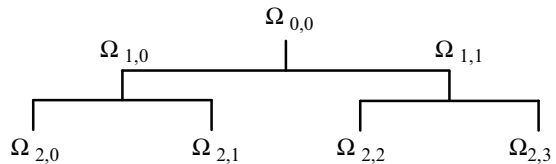


Fig. 9.1 Illustration of all nodes in two level wavelet packet decomposition

power associated with the coefficients of training signals in different classes at the $\Omega_{1,1}$ node with that of the $\Omega_{2,2}$ and $\Omega_{2,3}$ nodes, respectively. If the relative entropy of $\Omega_{1,1}$ is larger than that of $\Omega_{2,2}$ and $\Omega_{2,3}$, it keeps the bases belonging to node $\Omega_{1,1}$ and omits the other two nodes ($\Omega_{2,2}$ and $\Omega_{2,3}$). Otherwise it keeps the two nodes ($\Omega_{2,2}$ and $\Omega_{2,3}$) and disregards the basis of node $\Omega_{1,1}$. This process is applied to all the nodes in a sequential manner, up to the scale $j = 0$. As a result, a set of complete orthogonal wavelet packet bases having the highest discriminant power are obtained, which can be sorted out further for classification, according to a decreasing order.

Suppose that $A_{j,k}$ represents the desired local discriminant base restricted to the span of $B_{j,k}$, which is a set of wavelet packet coefficients at (j, k) node, and $\Delta_{j,k}$ is the array containing the discriminant measure of the same node, then the LDB algorithm for selecting the optimal wavelet packet base can be summarized as follows (Tafreshi et al. 2005):

LDB Algorithm Given a training dataset that consists of L class of signals $\{\{x_i^{(l)}\}_{i=1}^{N_l}\}_{l=1}^L$ with N_l being the total number of training signals in class l ,

Step 0: Choose a time frequency analysis method, such as the wavelet packet transform, to decompose the signals in the training dataset.

Step 1: Select a dissimilarity measure (e.g., relative entropy $D_1(\{p^m\}_{m=1}^L)$) to apply on the wavelet packet coefficients to the corresponding nodes (j, k) of the wavelet packet trees.

Step 2: Set $A_{j,k} = B_{j,k}$ where $B_{j,k}$ is the basis set spanning subspace of $\Omega_{j,k}$ node (j, k) , and then evaluate $\Delta_{j,k}$ for $k = 0, \dots, 2^j - 1$.

Step 3: Determine the best subspace $A_{j,k}$ for $j = J - 1, \dots, 0, k = 0, \dots, 2^j - 1$ by the following rule:

Set $\Delta_{j,k}$ as the dissimilarity measure, e.g., $\Delta_{j,k} = D_1(\{p^m\}_{m=1}^L)$

If $\Delta_{j,k} \geq \Delta_{j+1,2k} + \Delta_{j+1,2k+1}$, i.e., if the discriminant power of a parent node in wavelet packet tree is greater than those of children nodes,

Then

$$A_{j,k} = B_{j,k}$$

Else

$$A_{j,k} = A_{j+1,2k} \oplus A_{j+1,2k+1} \text{ and set } \Delta_{j,k} = \Delta_{j+1,2k} + \Delta_{j+1,2k+1}.$$

Step 4: Order sort the chosen basis functions by their power of discrimination in a decreasing order.

Step 5: Select the first $k(\leq l)$ highest discriminant base functions.

After step 3 is performed, a complete orthogonal basis is constructed. Orthogonality of the bases ensures that wavelet coefficients used as features during classification process are uncorrelated as much as possible. Subsequently, one can simply choose the first k highest discriminant bases in step 5 and use the corresponding coefficients as features in a classifier, or employ a statistical method, such as Fisher's criteria, to reduce the dimensionality of the problem first and then apply them into a classifier.

9.3 Case Study

To evaluate the effectiveness of the wavelet packet bases constructed using the LDB algorithm, three classes of signals are synthetically formed:

$$\begin{cases} x^{(1)}(t) = \text{Sine}(t) + n_1(t) & \text{for class 1} \\ x^{(2)}(t) = \text{Gauspuls}(t) + n_2(t) & \text{for class 2} \\ x^{(3)}(t) = \text{Tripuls}(t) + n_3(t) & \text{for class 3} \end{cases} \quad (9.11)$$

In (9.11), $\text{Sine}(t)$, $\text{Gauspuls}(t)$, and $\text{Tripuls}(t)$ represent the sinusoidal, Gaussian-modulated sinusoidal pulse, and triangle wave signals, respectively. The terms $n_1(t)$, $n_2(t)$, and $n_3(t)$ represent white noise. For each class, 100 training signals and 1,000 test signals were constructed, and the white noise was regenerated each time. Figure 9.2 shows one sample signal with 64 sampling points from each class. Each sample signal can be decomposed up to the six level (i.e., $2^6 = 64$), and the total number of nodes contained in the wavelet packet library for the signal is 127 (i.e., 1 for the 0 level, 2 for the first level, ..., 64 for the sixth level).

The LDB algorithm is first applied to the training signals to select a subset of wavelet packet nodes from a wavelet packet library that best discriminate the three classes. Figure 9.3 shows the selected wavelet packet nodes. It can be seen that the selected wavelet packet nodes (highlighted in black color) are distributed across different decomposition levels. Altogether, they form complete orthogonal bases. The first six selected LDB bases are shown in Fig. 9.4, with each containing 64 coefficients. It should be noted that these bases are sorted according to their discriminant power. A complete discriminant power for all the 64 LDB bases is

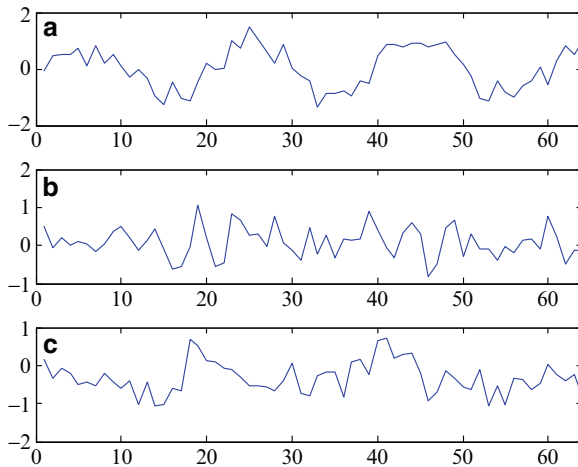


Fig. 9.2 Sample waveform from (a) class 1, (b) class 2, and (c) class 3

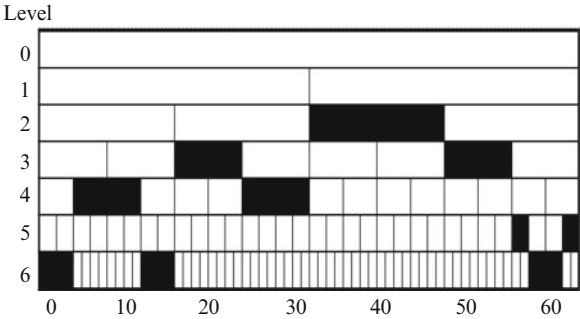


Fig. 9.3 The wavelet packet nodes selected by the LDB algorithm

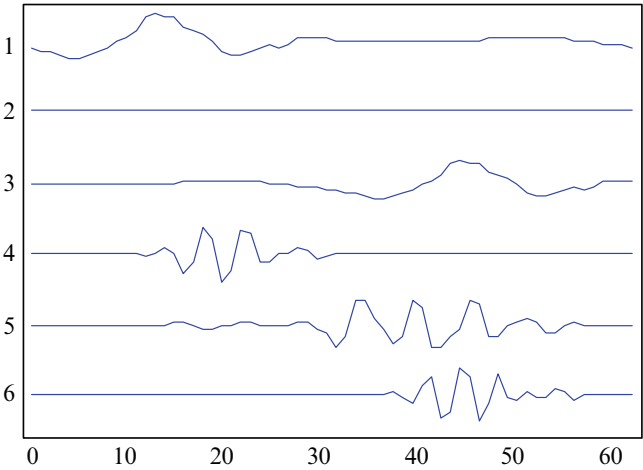


Fig. 9.4 The first six LDB bases selected from the signals

shown in Fig. 9.5 with a decreasing order. A rapid decrease of the discriminant power relative to the LDB bases is seen after the first few bases. Therefore, only the first few bases (e.g., the first six bases) with large discriminative power are considered for purpose of classification.

Wavelet coefficients constructed by projecting the signals onto the selected bases are then used to form *feature* variables for classification. For the training data set, two features (i.e., two wavelet coefficients) generated by the top two LDB bases have produced the clustering result as shown in Fig. 9.6.

To classify the three synthetic signals introduced above, these two features are used as input to a classifier. Various classifiers, such as linear discriminant analysis (LDA), neural network (NN), and support vector machine (SVM) can be considered for this purpose. In this example, the LDA classifier is selected due to its simplicity (Duda et al. 2000). Taking the two features as inputs to the LDA classifier, the

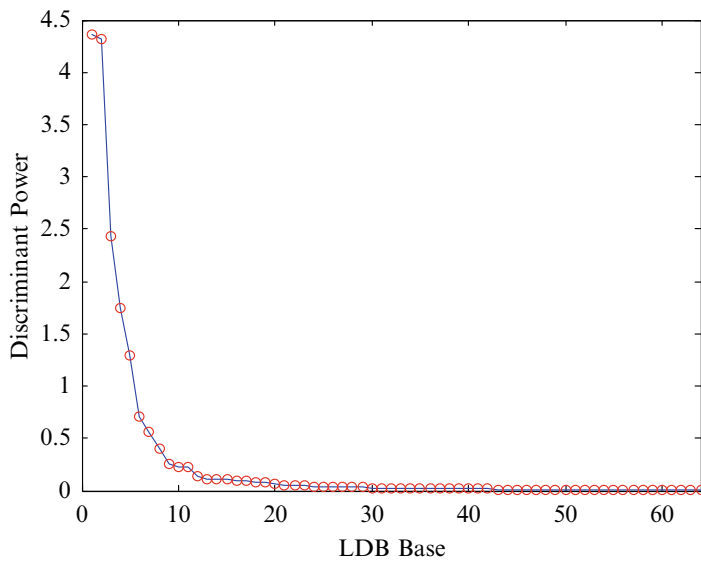


Fig. 9.5 Discriminant power of all 64 LDB bases

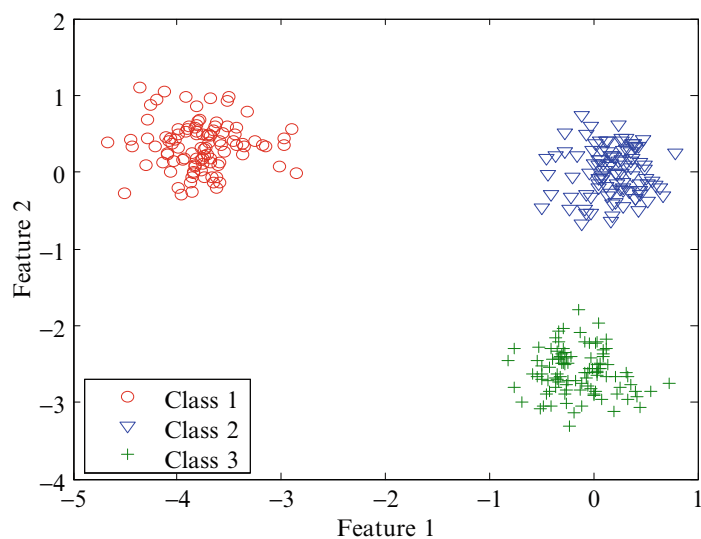


Fig. 9.6 Training signals represented by selected top two LDB features

training dataset are classified. The test signals are subsequently projected onto the top two LDB vectors to produce coefficients. Figure 9.7 indicates the scatter plot of the testing dataset by the two features. It is seen that, using the LDA classifier, all the testing data set are classified successfully.

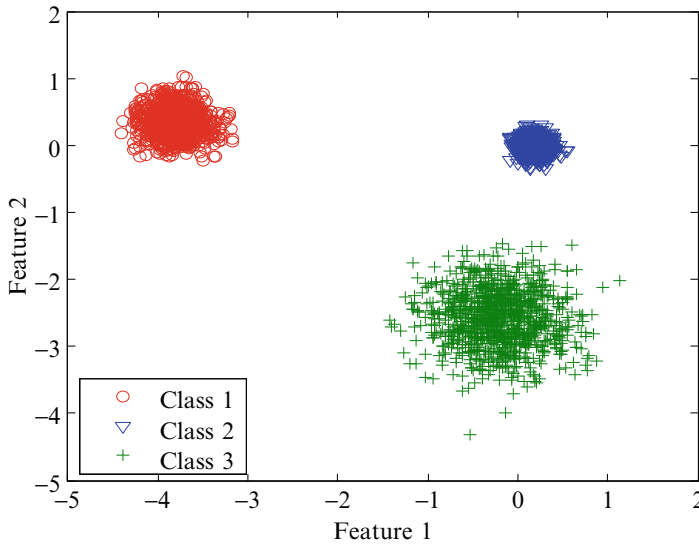


Fig. 9.7 Testing signals represented by selected top two LDB features

The above case study illustrates that the LDB-based wavelet packet base selection method is effective in producing features that enables effective discrimination of different classes.

9.4 Application to Gearbox Defect Classification

Application of the LDB algorithm to real-world classification problems has been reported in various areas, such as geophysical acoustic waveform classification (Saito and Coifman 1997), radar signal classification (Guglielmi 1997), automatic target recognition (Spooner 2001), ultrasonic echoes classification (Christian 2002), audio signal classification (Umapathy et al. 2007), biomedical signal analysis (Englehart et al. 2001; Umapathy and Krishnan 2006), and fault classification of mechanical systems (Tafreshi et al. 2005; Yen and Leong 2006). In this chapter, the LDB algorithm is applied to classifying the severity of defects in gearbox. Figure 9.8 illustrates the experimental set-up (Rafiee et al. 2007) where the vibration from a four-speed motorcycle gearbox is measured. An electrical motor drives the gearbox at a constant nominal rotational speed of 1,420 rpm. A tachometer measures the actual rotational speed to account for fluctuations caused by the load variations. Vibration signals are measured by a triaxial accelerometer mounted on the outer surface of the gearbox's housing, close to the input shaft of the gearbox. Four different working conditions of the test gear, including faultless, slight-worn, medium-worn, and with broken-teeth, are examined by analyzing the measured vibration data. The signals are sampled at 16,384 Hz. The sampled signals under the four different working conditions are shown in Fig. 9.9.

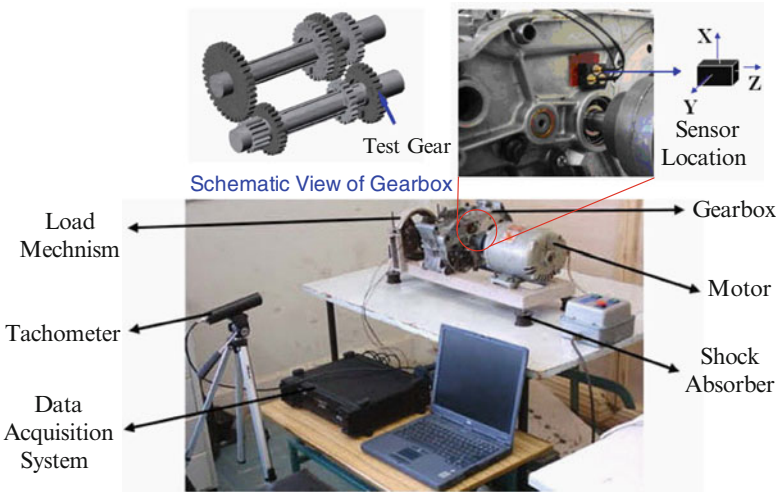


Fig. 9.8 Experimental setup to test a four speed motorcycle gearbox

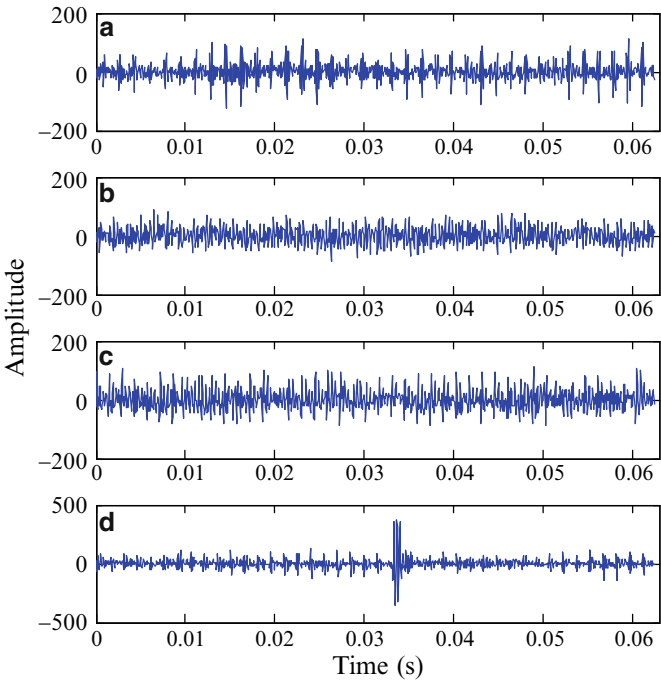


Fig. 9.9 Raw vibration signals of four gearbox conditions: (a) faultless, (b) slight worn, (c) medium worn, and (d) broken teeth

To classify the gearbox defect under different working conditions, 60 training signals and 80 testing signals, each containing 1,024 data points, are segmented from the raw signal corresponding to each defect condition. The LDB algorithm is then applied to the training data, where the relative entropy was chosen as the discriminant measure. Figure 9.10 shows the wavelet packet nodes selected from a four-level signal decomposition, and the first 6 LDB bases are shown in Fig. 9.11

To evaluate the effectiveness of the LDB algorithm, the performance of classification by using the testing data are compared between the LDB-selected nodes and all the 16 nodes at the 4th decomposition level. The energy values from the selected nodes are calculated and then used as inputs to a LDA classifier for characterizing the severity of the defect. Figures 9.12 and 9.13 illustrate the distribution of two energy features from node (3, 4) and node (4, 15) for the training and test data, respectively. The results of classification of gearbox defect severity are listed in Table 9.1. It is seen that, for the training data, although the misclassification rate for features with and without basis selection is the same, the LDB-selected features have resulted in a lower misclassification rate (1.56%) than those without (2.19%), for the testing data. This indicates the merit of the LDB in signal classification.

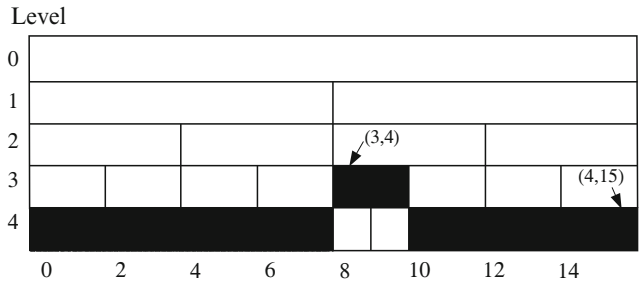


Fig. 9.10 Selected wavelet packet nodes for the gearbox data by the LDB algorithm

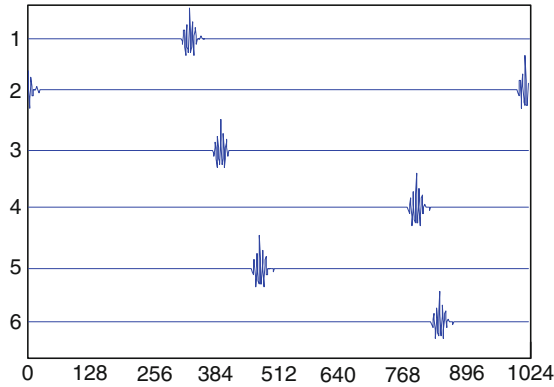


Fig. 9.11 The first six LDB bases selected from the gearbox vibration signals

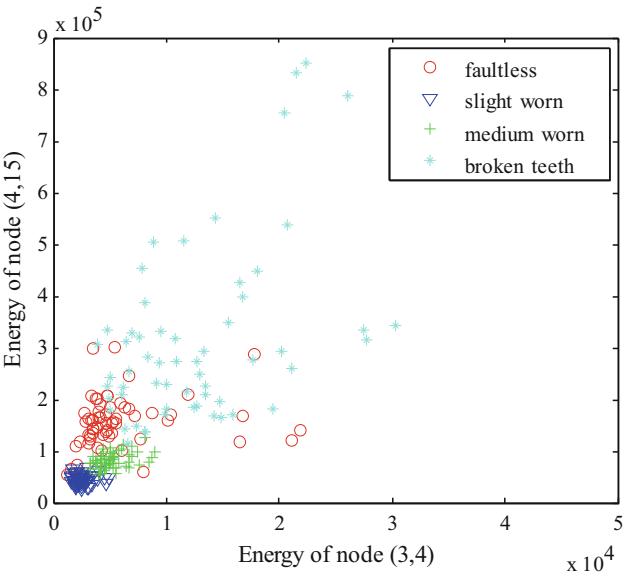


Fig. 9.12 Illustration of training samples by two features

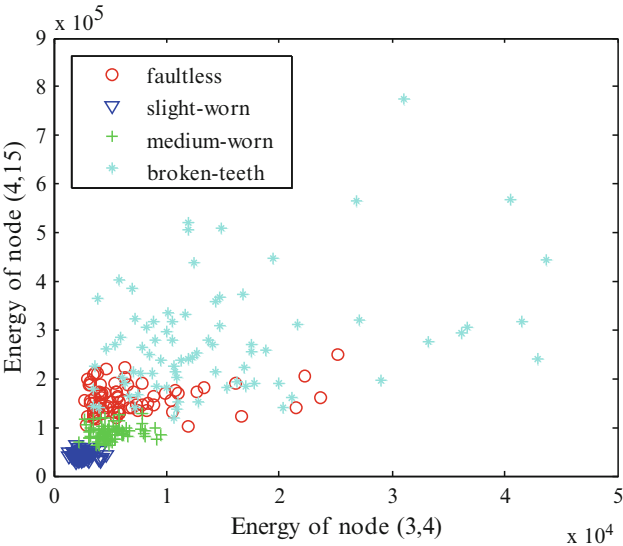


Fig. 9.13 Illustration of testing samples by two features

Table 9.1 Effect of wavelet packet node selection on gearbox defect severity classification

Features	Misclassification rate	
	Training signals (%)	Testing signals (%)
Wavelet packet nodes w/o LDB	0.83	2.19
Wavelet packet nodes w/LDB	0.83	1.56

9.5 Summary

The LDB provides an effective platform for the wavelet packet transform to decompose and classify signals. In this chapter, we have demonstrated that, using this approach, working conditions of a gearbox can be successfully classified by analyzing the measured vibration signals. Research on the theory of LDB has been continued in recent years. For example, the probability density of each class is estimated from the wavelet packet nodes to select the discriminant bases (Saito et al. 2002), and the features derived from this approach has been shown to be more sensitive to phase shifts than those from the original LDBs. Combing the LDB algorithm with signal-adapted filter banks (Strauss et al. 2003), a shape-adapted LDB approach has been developed for bio-signal processing. It can be expected that more powerful algorithms and computational tools are yet to come to better serve the need for signal classification in manufacturing.

9.6 References

- Basseville M (1989) Distance measures for signal processing and pattern recognition. *Signal Processing* 8(4):349–369
- Christian B (2002) Local discriminant bases and optimized wavelet to classify ultrasonic echoes: application to indoor mobile robotics. *Proc IEEE Sens* 1654–1659
- Coifman RR, Wicherhauser MV (1992) Entropy based algorithms for best basis selection. *IEEE Trans Inf Theory* 38(2):713–718
- Duda RO, Hart PE, Stork DG (2000) *Pattern classification*, 2nd edn. Wiley Interscience, New York
- Englehart K, Hudgins B, Parker PA (2001) A wavelet based continuous classification scheme for multifunction myoelectric control. *IEEE Trans Biomed Eng* 48(3):302–311
- Guglielmi RJM (1997) Wavelet feature definition and extraction for classification and image processing. Ph.D. Dissertation, Yale University
- Rafiee J, Arvani F, Harifi A, Sadeghi MH (2007) Intelligent condition monitoring of a gearbox using artificial neural network. *Mech Syst Signal Process* 21:1746–1754
- Saito N (1994) Local feature extraction and its applications using a library of bases, Ph.D. Dissertation, Yale University
- Saito N, Coifman RR (1995) Local discriminant bases and their applications. *J Math Imaging Vis* 5(4) 337–358
- Satio N, Coifman RR (1997) Extraction of geological information from acoustic well logging waveforms using time frequency wavelets. *Geophysics* 62(6):337–358

- Saito N, Coifman RR, Geshwind FB, Warner F (2002) Discriminant feature extraction using empirical probability density estimation and a local basis library. *Pattern Recognit* 35:2841–2852
- Spooner CM (2001) Application of local discriminant bases to HRR based ATR. In: *Proceeding of thirty fifth asilomar conference on signals, systems and computers*, pp 1067–1073
- Strauss DJ, Steidl G, Delb W (2003) Feature extraction by shape adapted local discriminant bases. *Signal Processing* 83:359–376
- Tafreshi R, Sassani F, Ahmadi H, Dumontb G (2005) Local discriminant bases in machine fault diagnosis using vibration signals. *Integr Comput Aided Eng*, 12:147–158
- Umapathy K, Krishnan S (2006) Modified local discriminant bases algorithm and its application in analysis of human knee joint vibration signals. *IEEE Trans Biomed Eng* 53(3):517–523
- Umapathy K, Krishnan S, Rao RK (2007) Audio signal feature extraction and classification using local discriminant bases. *IEEE Trans Audio Speech Lang Processing* 15(4):1236–1246
- Yen GG, Leong WF (2006) Fault classification on vibration data with wavelet based feature selection scheme. *ISA Trans* 45(2):141–151

Chapter 10

Selection of Base Wavelet

One of the advantages of wavelet transform for signal analysis is the abundance of the base wavelets developed over the past decades – there are a total of 13 wavelet families documented in the MATLAB library. From such abundance arises a natural question of how to choose a base wavelet that is best suited for analyzing a specific signal. The question is valid, since the choice in the first place may affect the result of wavelet transform at the end. As an example, Fig. 11.1 (top row, left) illustrates an impulsive signal and how it may appear as a time series (top row, right) in real-world applications. The three rows below illustrate three representative base wavelets and the results of using them to analyze the impulsive signal: (1) Daubechies wavelet (Daubechies 1992), (2) Morlet wavelet, and (3) Mexican hat wavelet. These base wavelets have been used for machine condition monitoring and health diagnosis studies, as reported in Shao and Nezu (2004), Li et al. (2000), and Abu-Mahfouz (2005). Comparing the wavelet transform results using these wavelets (shown in the right column, Fig. 10.1), it is apparent that only the Morlet wavelet is effective in extracting the impulsive component from the signal, as illustrated by the similarity in the waveform between the corresponding wavelet coefficient and the impulsive component. The Daubechies and Mexican-hat wavelets, in comparison, did not fully reveal the characteristics of impulsive component. Such an example motivates the study of base wavelet selection to achieve optimal result in feature extraction from a signal. In this chapter, we first present a general strategy for base wavelet selection, from both a qualitative and a quantitative aspect. Subsequently, we introduce several quantitative measures that can be used as guidelines for wavelet selection, to guarantee effective extraction of signal features.

10.1 Overview of Base Wavelet Selection

The topic of base wavelet selection has been addressed by researchers from different aspects. These prior approaches can be categorized as either qualitative or quantitative, and are reviewed in the following two sections.

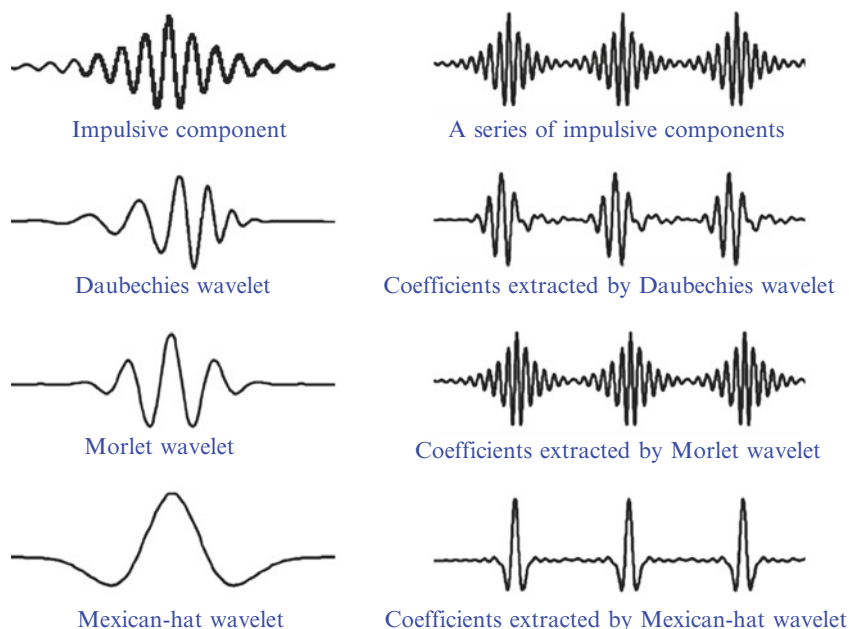


Fig. 10.1 Impulsive feature extraction using different base wavelets

10.1.1 Qualitative Measure

Base wavelets are characterized by a number of properties, such as *orthogonality*, *symmetry*, and *compact support*. Understanding these properties will be helpful for choosing a candidate base wavelet from the wavelet families for analyzing a specific signal. For example, the *orthogonality* property indicates that the inner product of the base wavelet is unity with itself, and zero with other scaled and shifted wavelets. As a result, using an orthogonal wavelet will result in efficient signal decomposition into nonoverlapping subfrequency bands. High computational efficiency can be achieved when orthogonal wavelets are used for implementing the discrete wavelet transform (DWT, see Chap. 4 for details) and wavelet packet transform (WPT, refer to Chap. 5). The symmetric property ensures that a base wavelet can serve as a linear phase filter. This is an important property in filtering operations, as the absence of it can lead to phase distortion. A *compact support* wavelet is one whose basis function is nonzero only within a finite interval. This allows the wavelet transform to efficiently represent signals that have localized features. The efficiency of such representation is important for data compression.

In recent years, the basic properties of wavelets have been extensively investigated to determine the suitability of a wavelet for specific applications. For example, based on the experiments conducted on a total of 23 Brodatz textures (Mojsilović

et al. 2000), it was concluded that the Biorthogonal wavelets with symmetry property enabled higher texture classification rate than the Daubechies wavelets, which is asymmetrical (e.g., 64.34% for Db3 vs. 82.17% for Bior3.3r). Similarly, the symmetric property of five wavelets (i.e., Haar, Db6, Coif4, Bior5.5, and Bior6.8) were reviewed (Fu et al. 2003), from which the Bior6.8 wavelet was chosen as the best-suited wavelet to separate the roughness, waviness, and geometrical form of an engineering surface into different frequency bands for both functional correlation and process diagnosis in manufacturing. In the area of biomedical engineering, the *regularity* and *symmetry* of base wavelets were considered as essential features for auditory-evoked potentials (AEP) signal analysis (Bradley and Wilson 2004). The morphology and latency of peaks, which characterize the AEP signal, were preserved when using a symmetric base wavelet, and the smooth peaks contained in the AEP signal were well matched when regularity of a base wavelet is greater than two. By taking into account the properties of *compact support*, *vanishing moment*, and *orthogonality*, the Coiflet 4 wavelet was selected to effectively separate burst and tonic components in the compound surface electromyogram (EMG) signals recorded from patients with dystonia (Wang et al. 2004). In addition to *orthogonality*, the property of complex or real basis was used to guide the choice of the base wavelet for electrocardiogram (ECG) signal analysis (Bhatia et al. 2006). The Morlet wavelet, Gaussian wavelet, Paul wavelet at order 4, and quadratic B-Spline wavelet were preselected as the candidates for ECG events detection and segmentation. In the area of image processing, the properties of *regularity*, *compact support*, *symmetry*, *orthogonality*, and *explicit expression* were used for recommending base wavelet for image sequence superresolution (Ahuja et al. 2005). It was concluded that the B-Spline family represents the most suitable base wavelet among the four candidates (i.e., Daubechies, Symlet, Coiflet, and B-Spline wavelets) for image sequence superresolution, as it is orthogonal, symmetric, and has the highest regularity, smallest support size, and explicit expression. In analyzing power system transients (Safavian et al. 2005), the Db4, Coiflet, and B-Spline wavelet were shown to be equally well-performing for the transient detection in a power system, as they share the same basic properties: finite support size and low vanishing moment.

Shape matching has been studied as an alternative approach to wavelet selection. For example, to measure the timing of multiunit bursts in surface EMGs from single trials (Flanders 2002), wavelets of different shapes, such as square, triangular, Gaussian and Mexican Hat, were investigated. The Db2 wavelet was chosen for its similarity to the shape of motor unit potentials hidden in the EMG signal. Also, base wavelets of different shapes were compared with ECG signals to determine their appropriateness for extracting a reference base from corrupted ECG, for magnetic resonance imaging (MRI) sequence triggering (Fokapu et al. 2005; Abi-Abdallah et al. 2006). To analyze impulses in vibration signals, researchers looked at the geometric shape of wavelets to determine the optimal choice (Yang and Ren 2004). It was found that components in a signal may be extracted effectively when a base wavelet with similar shape as the component is employed.

10.1.2 Quantitative Measure

The various approaches described in Sect. 10.1.1 illustrate the importance of choosing an appropriate base wavelet for effective signal processing. However, the basis properties of a wavelet only qualitatively determine its suitability for a particular application. As far as shape matching is concerned, it is generally difficult to accurately match the shape of a signal to that of a base wavelet through a visual comparison. These deficiencies motivate the study of quantitative measures for base wavelet selection.

The measures of *inequality* (Goel and Vidakovic 1995), which includes the Schur concave functions such as *Shannon entropy* and *Fishlow's measure* (Marshall and Olkin 1979), *Emelen's modified entropy measure* (Emlen 1973), and Schur convex functions (*Gini's coefficient* and *Schutz's coefficient* by Marshall and Olkin 1979) were proposed for wavelet selection in data compression and data denoising. A time-series, which was constructed by adding white noise into the sampled Db3 wavelet function, was used to evaluate each of the *inequality* measures. All of them, except for the *Fishlow's measure*, recognized the Db3 wavelet as the best base wavelet among a set of wavelets (Db1 Db20, Db30, Coif8, Coif12, and Coif18).

The *Shannon entropy* was also utilized to identify optimal base wavelet for velocity and temperature time series analysis in atmospheric surface layer (ASL) (Katul and Vidakovic 1996). The large scale eddy motion and small scale fluctuations in the ASL were successfully separated with the chosen Daubechies wavelet. In another study (Bedekar et al. 2005), the *Shannon entropy* was used to choose the Daubechies wavelet of order 3 from 23 preselected wavelets as the optimal wavelet for radio-frequency intravascular ultrasound (IVUS) data decomposition. It accurately decomposed 29 out of 30 IVUS data at all levels. The rest of the wavelets only decomposed less than 21 IVUS data.

In the field of biomedical engineering, study on horse gait classification has discussed an *uncertainty* model for wavelet selection (Arafat et al. 2003). The model combines the *fuzzy uncertainty* with the *probabilistic uncertainty* to provide a better measure, when compared with using either fuzzy or probabilistic uncertainty alone, for choosing an appropriate base wavelet to improve correct classification of different horse gait signals.

Study on assessing hypnotic state of anesthetized patients undergoing surgery (Bibian et al. 2001) has used the *discrimination power*, which is defined as the difference between the statistical features, such as *probability density function*, in the awake as well as in the anesthetized states, to select the appropriate base wavelet. It was found that among the Daubechies, Coiflet, Symlet, biorthogonal, and reverse biorthogonal wavelets, the Daubechies wavelet at order 8 provided the highest discrimination power, thus effectively estimated the hypnotic state. In another study on diagnosing cardiovascular ailments in patients (Singh and Tiwari 2006), experimental results have revealed the suitability of the Daubechies base wavelet at order 8 for the ECG signal denoising, as it

has the maximum *cross correlation* coefficient between the ECG signal and the chosen base wavelets, (Daubechies, Symlet, and Coiflet wavelets). The *cross correlation* measure has also been used in evaluating a base wavelet for detecting and locating the partial discharge (PD) occurred in operational transformers (Ma et al. 2002a, b; Yang et al. 2004). It was found that an optimal base wavelet would maximize the correlation coefficient between the signal of interest and the base wavelet, resulting in PD pulses being successfully separated from electrical noise.

For image denoising, two criteria, the *signal information extraction criterion* and the *distribution error criterion*, were proposed to select an optimal wavelet for improving the denoising performance (Zhang et al. 2005). The first criterion was implemented by calculating the *mutual information* of wavelet coefficients of the image without noise contamination and those of the image with noise contamination. The second criterion was the difference between the Gaussian and the actual distribution of the wavelet coefficients of the image without noise contamination. It was reasoned that the smaller the difference is the better the denoising performance will be, as the denoising performance is optimal only if the underlying signal distribution is Gaussian. Using these two criteria, it was found that the Bior1.3 wavelet provided the best performance among the eight wavelets (Bior1.1, Bior1.3, Bior2.2, Bior2.4, Bior3.3, Db2, Db3, and Db4) investigated when testing four benchmark images.

For automatic ultrasound nondestructive foreign body (FB) detection and classification in nonflat surface containers (Tsui and Basir 2006), the *relative entropy* was employed as a wavelet coefficient similarity measure to select the best base wavelet. The results have shown that the best base wavelet for FB shape classification is Bior3.1, while Haar (or Sym1 or reverse Bior1.1) and reverse Bior3.9 are the best for spherical and rectangular FB material classifications, respectively.

For analysis of impulses in vibration signals (Schukin et al. 2004), the *minimum total error* and *time-frequency resolution* were devised to evaluate different base wavelets on impulsive parameter identification of a single-degree-of-freedom system model. A comprehensive comparison among ten base wavelets (complex B-Spline, Gaussian, Shannon, etc.) indicated that the impulse wavelet is the most appropriate base wavelet for the analysis of impulses.

10.2 Wavelet Selection Criteria

The importance of the base wavelet has been addressed by various researchers, as summarized earlier. This section introduces several quantitative measures in evaluating the performance of base wavelets for the specific application domain of condition monitoring and health diagnosis in manufacturing.

10.2.1 Energy and Shannon Entropy

The energy content of a signal is a measure that uniquely characterizes the signal, thus can be used for base wavelet selection. The amount of energy contained in a signal $x(t)$ is expressed as:

$$E_{x(t)} = \int |x(t)|^2 dt \quad (10.1)$$

Similarly, when the signal is represented by discrete sample values $x(i) (i = 1, 2, \dots, N)$, the amount of energy is given by:

$$E_{x(i)} = \sum_{i=1}^N |x(i)|^2 \quad (10.2)$$

In (10.2), N is the length of the signal expressed by the number of data points, and $x(i)$ is the amplitude of the signal.

The energy content of a signal can also be calculated from its wavelet coefficients, and is expressed as:

$$E_{energy} = \iint |wt(s, \tau)|^2 ds d\tau \quad (10.3)$$

The corresponding sampled version is given by:

$$E_{energy} = \sum_s \sum_i |wt(s, i)|^2 \quad (10.4)$$

Equations (10.3) and (10.4) indicate that the energy associated with each particular scaling parameter s is expressed as:

$$E_{energy}(s) = \int |wt(s, \tau)|^2 d\tau \quad (10.5)$$

And the energy of the corresponding sampled version is described as:

$$E_{energy}(s) = \sum_{i=1}^N |wt(s, i)|^2 \quad (10.6)$$

where N is the number of wavelet coefficients and $wt(s, i)$ represents the wavelet coefficients.

If a major frequency component corresponding to a particular scale s exists in the signal, then the wavelet coefficients at that scale will have relatively high

magnitudes at the time when this major frequency component occurs. As a result, the energy related to such frequency component will be extracted from the signal when applying the wavelet transform to the signal. For purpose of condition monitoring and health diagnosis, the higher the energy content extracted from the defect-induced transient vibrations is the more effective the wavelet transform of the signal will be. Therefore, the energy content can serve as a criterion for selecting the base wavelet. This is formulated in the following criterion.

1. *Maximum energy criterion:* The base wavelet that extracts the largest amount of energy from the signal being analyzed represents the most appropriate wavelet for extracting features from defect-induced transient vibrations.

Given that for the same amount of energy within a subfrequency band, the specific condition of the signal may be significantly different (e.g., only several frequency components with high magnitude and others with negligible magnitude vs. a widespread spectrum), the spectral distribution (or concentration) of the energy needs to be considered also to ensure effective feature extraction. The energy distribution of the wavelet coefficients is quantitatively described by the Shannon entropy (Cover and Thomas 1991):

$$E_{entropy}(s) = - \sum_{i=1}^N p_i \cdot \log_2 p_i \quad (10.7)$$

where p_i is the energy probability distribution of the wavelet coefficients, defined as:

$$p_i = \frac{|wt(s, i)|^2}{E_{energy}(s)} \quad (10.8)$$

with $\sum_{i=1}^N p_i = 1$, and $p_i \cdot \log_2 p_i = 0$ if $p_i = 0$.

Equations (10.7) and (10.8) indicate that the entropy of the wavelet coefficients is bounded by:

$$0 \leq E_{entropy}(s) \leq \log_2 N \quad (10.9)$$

in which $E_{entropy}(s)$ will be equal to (1) zero, if all other wavelet coefficients are equal to zero except for one wavelet coefficient, and (2) $\log_2 N$, if the probability of energy distribution for all the wavelet coefficients are the same (i.e., $1/N$). This leads to the conclusion that the lower the entropy value is, the higher the energy concentration will be. Therefore, an appropriate base wavelet should yield large magnitude at a few wavelet coefficients and negligible magnitude at others when the signal is decomposed into various scales, leading to the minimum Shannon entropy. The corresponding Shannon entropy-based wavelet selection criterion can thus be designed as:

2. *Minimum Shannon entropy criterion:* The base wavelet that minimizes the entropy of the wavelet coefficients represents the most appropriate wavelet for defect-induced transient vibration analysis.

Combining the strengths of the two criteria described earlier, we note that an appropriate base wavelet should extract the maximum amount of energy from the signal being analyzed, while minimizing the Shannon entropy of the corresponding wavelet coefficients. This lead to the *energy-to-Shannon entropy* ratio, which is defined as:

$$R(s) = \frac{E_{\text{energy}}(s)}{E_{\text{entropy}}(s)} \quad (10.10)$$

In (10.10), the energy $E_{\text{energy}}(s)$ and the entropy $E_{\text{entropy}}(s)$ are calculated from (10.6) and (10.7), respectively. By maximizing the *energy-to-Shannon entropy* ratio $R(s)$, an appropriate base wavelet can be selected from a set of candidate base wavelets. This leads to the following criterion for wavelet selection:

3. *Energy-to-Shannon entropy ratio measure*: The base wavelet that has produced the maximum energy-to-Shannon entropy ratio should be chosen as the most appropriate wavelet for defect-induced transient vibration signal analysis.

10.2.2 Information Theoretic Measure

The *energy* and *Shannon entropy*-related criteria are solely based on the content of the wavelet coefficients themselves. Since the coefficients of a signal's wavelet transformation are inherently related to the signal, information theoretic measures, which describes the relationship between a pair of data sequence, can be explored for best suited base wavelet selection. These are introduced in the following sections.

10.2.2.1 Joint Entropy

The joint entropy $H(X, Y)$ between two data sequences X and Y is defined to measure information associated with them as a whole (Cover and Thomas 1991). This is expressed as

$$H(X, Y) = - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y) \quad (10.11)$$

where $p(x, y)$ is the joint probability distribution of the two data sequences.

10.2.2.2 Conditional Entropy

With probability distribution of the data sequence X known, the amount of information contained in the other data sequence Y can be measured by the condition entropy $H(Y|X)$ as (Cover and Thomas 1991):

$$\begin{aligned}
H(Y|X) &= - \sum_{x \in X} p(x) H(Y|X = x) \\
&= - \sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log p(y|x)
\end{aligned} \tag{10.12}$$

In (10.12), $p(x)$ is the probability distribution of the data sequence X , and $p(y|x)$ denotes the conditional probability distribution of the data sequence Y when the data sequence X is known. The conditional probability distribution $p(y|x)$ is expressed as (Mendenhall and Sincich 1995):

$$p(y|x) = \frac{p(x, y)}{p(x)} \tag{10.13}$$

with $p(x, y)$ being the joint probability distribution of the two data sequence X and Y . As a result, (10.12) can be further expressed as:

$$\begin{aligned}
H(Y|X) &= - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)} \\
&= - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y) + \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x) \\
&= H(X, Y) + \sum_{x \in X} p(x) \log p(x) = H(X, Y) - H(X)
\end{aligned} \tag{10.14}$$

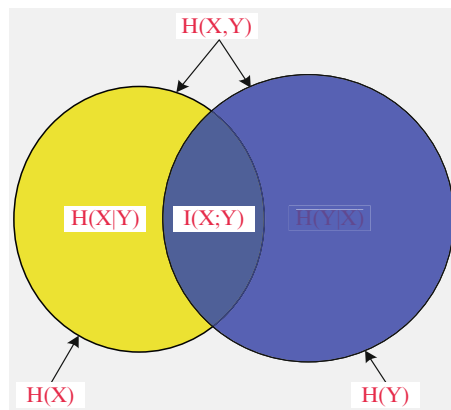
Equation (10.14) indicates that, given the data sequence X , the condition entropy of data sequence Y can be calculated by the joint entropy between the two data sequences, minus the entropy of the data sequence X .

10.2.2.3 Mutual Information

The mutual information $I(X; Y)$ measures the amount of information that data sequence X contains about data sequence Y , which is defined as (Cover and Thomas 1991):

$$\begin{aligned}
I(X; Y) &= \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\
&= \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y) - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log [p(x)p(y)] \\
&= -H(X, Y) - \sum_{x \in X} p(x) \log p(x) - \sum_{y \in Y} p(y) \log p(y) \\
&= -H(X, Y) + H(X) + H(Y)
\end{aligned} \tag{10.15}$$

Fig. 10.2 Relationships among entropies and mutual information



Equation (10.15) indicates that the mutual information is the sum of the entropies $H(X)$ and $H(Y)$, minus the joint entropy $H(X, Y)$.

The relationships among the joint entropy, condition entropy, and mutual information are illustrated in a Venn diagram (Cover and Thomas 1991), as shown in Fig. 10.2. It is noted that the mutual information $I(X; Y)$ is represented by the intersection of the two data sequences. The greater the mutual information is, the more similar the two data sequences will be. The condition entropy $H(X|Y)$ or $H(Y|X)$ expresses the information that is particular to each corresponding data sequence itself, while the joint entropy $H(X, Y)$ includes all information of the two data sequences.

The above-described relationship can be applied to base wavelet selection by taking the signal to be analyzed and its corresponding wavelet coefficients as two data sequences X and Y , respectively. Since defect-induced transient features are represented by the wavelet coefficients, a high value of mutual information between the vibration signal and wavelet coefficients can be expected when an appropriate wavelet is chosen. It should be noted that, when a vibration signal is obtained, the information $H(X)$ is fixed. Similarly, the information $H(Y)$ of the wavelet coefficients is fixed once a base wavelet is chosen. On the basis of the relationship described earlier, low values of both the joint entropy and condition entropy are desired for choosing an appropriate wavelet for characterizing defect-induced transient vibrations.

Following are several criteria for base wavelet selection, based on the information theoretic measures.

4. *Minimum joint entropy criterion:* The base wavelet that minimizes the joint entropy between the signal and the wavelet coefficients represents the most appropriate wavelet for defect-induced transient feature extraction.
5. *Minimum condition entropy criterion:* The base wavelet that minimizes the condition entropy between the signal and the wavelet coefficients represents the most appropriate wavelet for defect-induced transient feature extraction.

6. *Maximum mutual information criterion*: The base wavelet that maximizes the mutual information between the signal and the wavelet coefficients represents the most appropriate wavelet for defect-induced transient feature extraction.

10.2.2.4 Relative Entropy

In contrast to the mutual information, which measures shared information between two data sequences, the relative entropy (also known as *Kullback Leibler distance* or the *divergence*) is a measure of the distance between probability distributions of data sequences X and Y (Cover and Thomas 1991). The relative entropy is defined as

$$D(X||Y) = \sum_{x \in X} p(x) \log \frac{p(x)}{p(y)} \quad (10.16)$$

with $p(x) \log \frac{p(x)}{p(y)} = 0$ if $p(x) = 0$, and $p(x) \log \frac{p(x)}{p(y)} = \infty$ if $p(y) = 0$.

Equation (10.16) states that the relative entropy value is always nonnegative, and it is zero if and only if both probability distributions are equivalent [i.e., $p(x) = p(y)$]. The smaller the relative entropy is, the more similar the distributions of the two data sequences will be. For applications in machine condition monitoring and health diagnosis, it is expected that an appropriately chosen base wavelet will be able to extract features related to defect-induced transient vibrations completely. Consequently, a small relative entropy value between the signal (i.e., data sequence X) and its corresponding wavelet coefficients (i.e., data sequence Y) is desired. The following criterion reflects this consideration:

7. *Minimum relative entropy criterion*: The base wavelet that minimizes the relative entropy between the signal and the wavelet coefficients represents the most appropriate wavelet for defect-induced transient feature extraction.

Synthesizing the above criteria, an appropriate wavelet should minimize the joint entropy, condition entropy, and relative entropy while maximizing the mutual information. Such consideration is captured in the following information measure:

$$Info(s) = \frac{I(X; Y)}{H(X, Y) \times H(Y|X) \times D(X||Y)} \quad (10.17)$$

In (10.17), the joint entropy $H(X, Y)$, condition entropy $H(Y|X)$, relative entropy $D(X||Y)$, and mutual information $I(X; Y)$ are calculated using (10.11), (10.14), (10.16), and (10.15), respectively. Maximizing the information measure $Info(s)$ leads to the following comprehensive criterion:

8. *Maximum information criterion*: The base wavelet that has produced the maximum information value should be chosen to be the most appropriate wavelet for defect-induced transient feature extraction.

10.3 Numerical Study on Base Wavelet Selection

To quantitatively evaluate the base wavelet selection criteria described earlier, a Gaussian-modulated sinusoidal test signal is numerically simulated. Mathematically such a signal can be expressed as

$$x(t) = e^{-\beta(t-t_0)^2} \sin[2\pi f(t-t_0)] \quad (10.18)$$

The symbol β denotes the attenuation factor, and t_0 is the time delay of the signal. This type of signal has been widely used for simulating transient vibrations involved in mechanical systems (Ho and Randall 2000; Schukin et al. 2004; Yang and Ren 2004). Figure 10.3 illustrates the test signal, in which the center frequency is 48 Hz, and the sampling frequency is 1,024 Hz. In the following, the criteria presented in the above sections are evaluated for choosing best suited base wavelet from both real-valued and complex-valued wavelets.

10.3.1 Evaluation Using Real-Valued Wavelets

The performance of real-valued wavelets on processing the test signal is evaluated first, for which the DWT is performed to decompose the test signal. The decomposition level L of the wavelet transform is determined by the sampling frequency f_q and

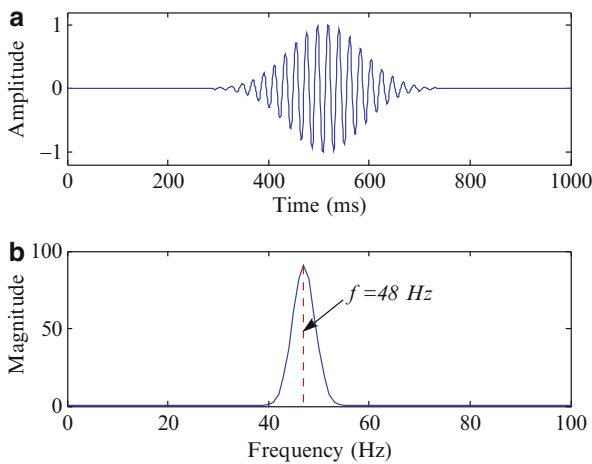


Fig. 10.3 Test signal: Gaussian modulated sinusoidal signal

frequency component to be identified in the signal, as expressed in the following equation:

$$\frac{f_q}{2^{L+1}} \leq f_{char} \leq \frac{f_q}{2^L}$$

(10.19)

In (10.19), f_q is the sampling frequency, and f_{char} is related to the characteristic frequency component of the signal (e.g., $f_{char} = 48$ Hz for the test signal). In Table 10.1, the respective frequency ranges covered by each of the decomposition levels under the sampling rate of 1,024 Hz are shown. Since the center frequency (48 Hz) of the test signal falls within the frequency range of 32–64 Hz, which is covered by the decomposition level 4 (corresponding to scale $s = 2^4 = 16$), this level is chosen for the evaluation of each wavelet.

Thirty candidate base wavelets were preselected from seven wavelet families. The energy extracted from the Gaussian-modulated sinusoidal signal by these wavelets is listed in Table 10.2. It is shown that the *Meyer* wavelet has extracted the highest amount of energy, thus is considered the most appropriate base wavelet for analyzing the Gaussian-modulated sinusoidal signal with the given parameters. It is also found that the amount of energy that is extracted from the signal increases with increasing order of the base wavelet, for each wavelet family. This is because base wavelets of higher order within a wavelet family possess higher degree of regularity. As a result, they are better suited for extracting energy from the Gaussian-modulated sinusoidal test signal than their lower-ordered counterparts in the same wavelet family.

The Shannon entropy of the extracted Gaussian-modulated sinusoidal signal is then calculated, as listed in Table 10.3. On the basis of the minimum Shannon entropy, the *Symlet 3* wavelet is considered as the most appropriate base wavelet.

Table 10.1 Frequency range for each decomposition level under a 1,024 Hz sampling rate

Decomposition level (L)	Frequency range (Hz)	Decomposition level (L)	Frequency range (Hz)
1	256–512	4	32–64
2	128–256	5	16–32
3	64–128	6	8–16

Table 10.2 Energy extracted from the test signal: real valued wavelets

Base wavelet	Energy (J)	Base wavelet	Energy (J)	Base wavelet	Energy (J)
Haar	33.855	Coif4	60.662	Bior2.6	53.645
Db2	45.546	Coif5	61.856	Bior4.4	52.310
Db4	54.433	Sym2	45.546	Bior5.5	54.614
Db6	58.167	Sym3	51.143	Bior6.8	58.415
Db8	60.207	Sym4	54.433	rBio1.3	45.326
Db10	61.471	Sym6	58.167	rBio2.4	55.138
DB20	63.687	Sym8	60.217	rBio2.6	55.546
Coif1	46.065	Meyr	64.146	rBio4.4	59.235
Coif2	55.038	Bior1.3	53.481	rBio5.5	61.123
Coif3	58.692	Bior2.4	49.198	rBio6.8	60.464

This conclusion is not consistent with the Meyer wavelet selected by the maximum energy criterion. To resolve such conflict, the energy-to-Shannon entropy ratio is calculated and the results are listed in Table 10.4. From the *maximum energy-to-Shannon entropy ratio* criterion, the *Meyer* wavelet possesses the highest values, thus is considered the most appropriate wavelet to analyze the Gaussian-modulated sinusoidal signal.

Various criteria based on information theoretic measures have also been studied to evaluate the performance of each of the real-valued candidate wavelets, as listed in Table 10.5 10.8. It is noted that all of the four measures (i.e., joint entropy, condition entropy, mutual information, and relative entropy) point to the *Meyer* wavelet as the most suited wavelet when analyzing the Gaussian-modulated sinusoidal signal. This is because it maximizes the maximum mutual information, while minimizing the joint entropy, condition entropy, and relative entropy. The comprehensive criterion “*maximum information*” that integrates the effect of these four measures, as illustrated in (10.17), has also shown that the Meyer wavelet is the most appropriate wavelet. This is verified in Table 10.9, in which a maximum information value is obtained when the Meyer wavelet is chosen as the base wavelet.

Table 10.3 Shannon entropy of the extracted signal: real valued wavelets

Base wavelet	Shannon entropy	Base wavelet	Shannon entropy	Base wavelet	Shannon entropy
Haar	3.667	Coif4	2.945	Bior2.6	5.959
Db2	3.137	Coif5	2.985	Bior4.4	5.673
Db4	3.475	Sym2	3.137	Bior5.5	6.042
Db6	3.171	Sym3	2.800	Bior6.8	4.069
Db8	3.491	Sym4	3.011	rBio1.3	4.579
Db10	3.121	Sym6	3.598	rBio2.4	4.665
DB20	3.653	Sym8	3.613	rBio2.6	4.949
Coif1	2.856	Meyr	2.959	rBio4.4	4.664
Coif2	3.609	Bior1.3	6.197	rBio5.5	5.034
Coif3	3.617	Bior2.4	6.214	rBio6.8	4.069

Table 10.4 Energy to Shannon entropy ratio of the extracted signal: real valued wavelets

Base wavelet	Energy to Shannon entropy ratio	Base wavelet	Energy to Shannon entropy ratio	Base wavelet	Energy to Shannon entropy ratio
Haar	9.229	Coif4	20.594	Bior2.6	9.002
Db2	14.512	Coif5	20.719	Bior4.4	9.220
Db4	15.662	Sym2	14.512	Bior5.5	9.047
Db6	18.341	Sym3	18.265	Bior6.8	14.356
Db8	17.246	Sym4	18.079	rBio1.3	9.896
Db10	19.693	Sym6	16.162	rBio2.4	11.817
DB20	17.428	Sym8	16.662	rBio2.6	11.221
Coif1	16.124	Meyr	21.678	rBio4.4	12.699
Coif2	15.244	Bior1.3	8.629	rBio5.5	12.141
Coif3	16.224	Bior2.4	7.915	rBio6.8	14.858

Table 10.5 Joint entropy of the extracted signal: real valued wavelets

Base wavelet	Joint entropy	Base wavelet	Joint entropy	Base wavelet	Joint entropy
Haar	4.086	Coif4	3.261	Bior2.6	3.414
Db2	3.409	Coif5	3.246	Bior4.4	3.338
Db4	3.394	Sym2	3.398	Bior5.5	3.415
Db6	3.495	Sym3	3.291	Bior6.8	3.379
Db8	3.358	Sym4	3.250	rBio1.3	3.603
Db10	3.256	Sym6	3.441	rBio2.4	3.329
DB20	3.086	Sym8	3.336	rBio2.6	3.619
Coif1	3.082	Meyr	2.957	rBio4.4	3.341
Coif2	3.614	Bior1.3	3.682	rBio5.5	3.348
Coif3	3.366	Bior2.4	3.313	rBio6.8	3.453

Table 10.6 Condition entropy of the extracted signal: real valued wavelets

Base wavelet	Condition entropy	Base wavelet	Condition entropy	Base wavelet	Condition entropy
Haar	1.539	Coif4	0.715	Bior2.6	0.792
Db2	0.851	Coif5	0.699	Bior4.4	0.868
Db4	0.847	Sym2	0.851	Bior5.5	0.833
Db6	0.948	Sym3	0.745	Bior6.8	1.057
Db8	0.812	Sym4	0.704	rBio1.3	0.783
Db10	0.710	Sym6	0.895	rBio2.4	1.073
DB20	0.539	Sym8	0.790	rBio2.6	0.795
Coif1	0.536	Meyr	0.411	rBio4.4	0.802
Coif2	1.067	Bior1.3	1.136	rBio5.5	0.907
Coif3	0.819	Bior2.4	0.767	rBio6.8	0.792

Table 10.7 Mutual information of the extracted signal: real valued wavelets

Base wavelet	Mutual information	Base wavelet	Mutual information	Base wavelet	Mutual information
Haar	1.243	Coif4	1.261	Bior2.6	1.045
Db2	0.69	Coif5	1.317	Bior4.4	1.101
Db4	1.074	Sym2	0.869	Bior5.5	1.165
Db6	1.151	Sym3	0.990	Bior6.8	1.559
Db8	1.174	Sym4	1.085	rBio1.3	0.969
Db10	1.291	Sym6	1.110	rBio2.4	0.873
DB20	1.502	Sym8	1.241	rBio2.6	1.011
Coif1	0.760	Meyr	1.721	rBio4.4	0.913
Coif2	1.078	Bior1.3	0.511	rBio5.5	0.979
Coif3	1.148	Bior2.4	0.997	rBio6.8	1.435

10.3.2 Evaluation Using Complex-Valued Wavelets

The criteria for choosing an appropriate complex-valued wavelet can be evaluated by applying the continuous wavelet transform to the test signal. The scale whose

Table 10.8 Relative entropy of the extracted signal: real valued wavelets

Base wavelet	Relative entropy	Base wavelet	Relative entropy	Base wavelet	Relative entropy
Haar	0.4851	Coif4	0.163	Bior2.6	1.155
Db2	1.09	Coif5	0.111	Bior4.4	0.564
Db4	0.506	Sym2	1.091	Bior5.5	0.423
Db6	0.290	Sym3	0.764	Bior6.8	0.371
Db8	0.194	Sym4	0.507	rBio1.3	0.240
Db10	0.125	Sym6	0.281	rBio2.4	0.182
DB20	0.022	Sym8	0.194	rBio2.6	1.146
Coif1	1.149	Meyr	0.002	rBio4.4	0.985
Coif2	0.435	Bior1.3	1.155	rBio5.5	0.515
Coif3	0.266	Bior2.4	0.564	rBio6.8	0.817

Table 10.9 Information value of the extracted signal: real valued wavelets

Base wavelet	Information value	Base wavelet	Information value	Base wavelet	Information value
Haar	0.296	Coif4	3.226	Bior2.6	0.773
Db2	0.232	Coif5	5.155	Bior4.4	1.054
Db4	0.679	Sym2	0.232	Bior5.5	1.550
Db6	1.139	Sym3	0.471	Bior6.8	2.915
Db8	2.146	Sym4	0.858	rBio1.3	0.187
Db10	4.348	Sym6	1.221	rBio2.4	0.292
DB20	40	Sym8	2.347	rBio2.6	0.461
Coif1	0.339	Meyr	1,000	rBio4.4	0.371
Coif2	0.594	Bior1.3	0.082	rBio5.5	0.537
Coif3	1.495	Bior2.4	0.633	rBio6.8	1.534

Table 10.10 Energy extracted from the test signal: complex valued wavelets

Base wavelet	Energy (J)
Morlet wavelet	96.243
Gaussian wavelet	58.942
B Spline wavelet	57.257
Shannon wavelet	14.789
Harmonic wavelet	15.835

corresponding center frequency is equal to that of the frequency component of interest (e.g., 48 Hz in the test signal) is chosen for the wavelet transform. In general, the scale of the wavelet, s , and the corresponding center frequency of the scaled wavelet, f_{s-c} , are related by (Abry 1997):

$$s = \frac{f_q f_{b-c}}{f_{s-c}} \quad (10.20)$$

where f_q is the sampling rate, f_{b_c} is the center frequency of the base wavelet, and f_{s_c} is the center frequency of the scaled wavelet.

Table 10.10 lists the energy values extracted from the test signal, whereas Table 10.11 lists the corresponding Shannon entropy of the extracted signal. It is seen that the maximum energy criterion selected the *complex Morlet* wavelet as the most suited wavelet among the five complex-valued wavelets, while the *minimum Shannon entropy* criterion identified the *Complex Gaussian* wavelet. Such a conflict is resolved by the integrated *energy-to-Shannon entropy* ratio criterion. In Table 10.12, it is shown that the *Complex Morlet* wavelet led to the maximum energy-to-Shannon entropy ratio. As a result, the *Complex Morlet* wavelet is considered the most suited base wavelet.

The information theoretic measures can also be used to evaluate performance for each of the candidate wavelets, as listed in Table 10.13 10.16. It is noted that the *Complex Morlet* wavelet is again identified as the most suited wavelet by using the following three criteria: *minimum joint entropy*, the *minimum condition entropy*,

Table 10.11 Shannon entropy of the extracted signal: complex valued wavelets

Base wavelet	Shannon entropy
Morlet wavelet	7.322
Gaussian wavelet	7.290
B Spline wavelet	7.365
Shannon wavelet	7.690
Harmonic wavelet	7.453

Table 10.12 Energy to Shannon entropy ratio of the extracted signal: complex valued wavelets

Base wavelet	Energy to Shannon entropy ratio
Morlet wavelet	13.143
Gaussian wavelet	8.085
B Spline wavelet	7.772
Shannon wavelet	1.923
Harmonic wavelet	2.125

Table 10.13 Joint entropy of the extracted signal: complex valued wavelets

Base wavelet	Joint entropy
Morlet wavelet	3.121
Gaussian wavelet	3.280
B Spline wavelet	3.248
Shannon wavelet	4.167
Harmonic wavelet	3.639

Table 10.14 Condition entropy of the extracted signal: complex valued wavelets

Base wavelet	Condition entropy
Morlet wavelet	0.575
Gaussian wavelet	0.734
B Spline wavelet	0.702
Shannon wavelet	1.614
Harmonic wavelet	1.093

Table 10.15 Mutual information of the extracted signal: complex valued wavelets

Base wavelet	Mutual information
Morlet wavelet	1.665
Gaussian wavelet	1.808
B Spline wavelet	1.705
Shannon wavelet	1.296
Harmonic wavelet	1.551

Table 10.16 Relative entropy of the extracted signal: complex valued wavelets

Base wavelet	Relative entropy
Morlet wavelet	0.009
Gaussian wavelet	0.060
B Spline wavelet	0.011
Shannon wavelet	0.214
Harmonic wavelet	0.027

Table 10.17 Information value of the extracted signal: complex valued wavelets

Base wavelet	Information value
Morlet wavelet	111.111
Gaussian wavelet	12.346
B Spline wavelet	66.667
Shannon wavelet	0.864
Harmonic wavelet	13.889

and the *minimum relative entropy*. However, when the *maximum mutual information* criterion is applied, the *Complex Gaussian* wavelet is identified as the winner. We apply once again the comprehensive criterion “maximum information,” and the conflict is successfully resolved. As shown in Table 10.17, a maximum information value is obtained when the complex Morlet wavelet is chosen as the base wavelet.

The reason why the *Complex Morlet* wavelet is the most suited base wavelet for analyzing the Gaussian-modulated sinusoidal signal can be explained from a physical point of view by comparing their corresponding analytical expressions, as shown in (10.18) and (10.21) below:

$$\psi_M(t) = \frac{1}{\sqrt{\pi f_b}} e^{j2\pi f_c t} e^{-\frac{t^2}{f_b}} \quad (10.21)$$

Tuning the bandwidth f_b and center frequency f_c of the Complex Morlet wavelet, the scaled Complex Morlet wavelet can be expressed as:

$$\psi(t) = \sqrt{\frac{120}{\pi}} e^{j2\pi 48t} e^{-120t^2} \quad (10.22)$$

Equation (10.22) illustrates a perfect match of the scaled Complex Morlet wavelet to the Gaussian-modulated sinusoidal signal given in (10.18). As a result, its

wavelet coefficients best represent the test signal, which is why this wavelet has extracted the maximum amount of energy from the test signal.

In summary, using the Gaussian-modulated sinusoidal test signal, we have demonstrated how to systematically choose a base wavelet from a number of candidates by using the various quantitative measures. The two comprehensive criteria, i.e., *energy-to-Shannon entropy* measure and *maximum information* measure, have shown to be effective in choosing the most suited base wavelet for decomposing vibration signals for machine condition monitoring and health diagnosis.

10.4 Base Wavelet Selection for Bearing Vibration Signal

We now demonstrate how the two comprehensive wavelet selection criteria, *maximum energy-to-Shannon entropy* ratio and *maximum information* measure, have been applied to selecting base wavelet for bearing vibration signals. Figure 10.4a illustrates the waveform of a vibration signal measured from a ball bearing that contains a localized defect on its outer raceway. The sampling rate is 10,000 Hz. The spectrum in Fig. 10.4b indicates a major peak frequency component at 1,840 Hz. This component is used as the reference base for determining the decomposition

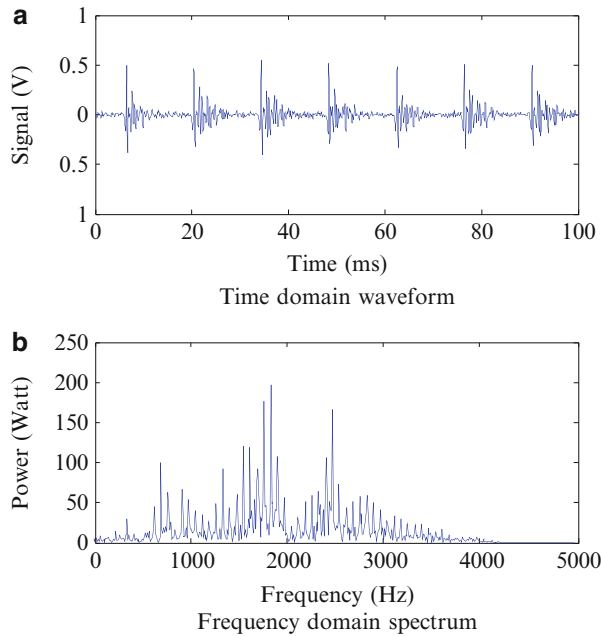


Fig. 10.4 Bearing vibration signal and its corresponding spectrum. (a) Time domain waveform and (b) frequency domain spectrum

level (for DWT) as well as for the scale selection (when performing CWT). The two criteria have been applied to evaluating real-valued and complex-valued wavelets, respectively.

The real-valued wavelets are evaluated first. The decomposition level of the DWT is chosen to be 2, which corresponds to scale 4 ($s = 2^2$). This scale covers the frequency range from 1,250 to 2,500 Hz, within which the major peak frequency component is located (at 1,840 Hz). After calculation of the energy and Shannon entropy values of the bearing vibration signal by each of the candidate wavelets, the *energy-to-Shannon entropy* ratio is calculated for each wavelet, and the results are listed in Table 10.18. On the basis of the *maximum energy-to-Shannon entropy* ratio criterion, the *reverse Biorthogonal* wavelet 5.5 (denoted as *rBio5.5*) was considered as the most appropriate wavelet for analyzing the bearing vibration signal.

Various similarity measures, including joint entropy, condition entropy, relative entropy, and mutual information, have also been calculated to evaluate the candidate base wavelets. By integrating these similarity measures into the *maximum*

Table 10.18 Energy to Shannon entropy ratio of the extracted bearing vibration signal: real valued wavelets

Base wavelet	Energy to Shannon entropy ratio	Base wavelet	Energy to Shannon entropy ratio	Base wavelet	Energy to Shannon entropy ratio
Haar	56.279	Coif4	75.980	Bior2.6	69.647
Db2	80.793	Coif5	76.473	Bior4.4	69.864
Db4	104.750	Sym2	80.120	Bior5.5	91.454
Db6	71.343	Sym3	73.969	Bior6.8	77.721
Db8	74.153	Sym4	59.229	rBio1.3	43.843
Db10	93.488	Sym6	77.946	rBio2.4	69.435
DB20	85.949	Sym8	68.515	rBio2.6	70.795
Coif1	66.550	Meyr	77.757	rBio4.4	76.204
Coif2	72.738	Bior1.3	39.720	rBio5.5	109.920
Coif3	75.050	Bior2.4	63.477	rBio6.8	78.777

Table 10.19 Information value of the extracted bearing vibration signal: real valued wavelets

Base wavelet	Information value	Base wavelet	Information value	Base wavelet	Information value
Haar	0.106	Coif4	0.143	Bior2.6	0.180
Db2	0.223	Coif5	0.142	Bior4.4	0.181
Db4	0.143	Sym2	0.223	Bior5.5	0.219
Db6	0.127	Sym3	0.180	Bior6.8	0.167
Db8	0.116	Sym4	0.108	rBio1.3	0.105
Db10	0.136	Sym6	0.171	rBio2.4	0.162
DB20	0.129	Sym8	0.122	rBio2.6	0.169
Coif1	0.173	Meyr	0.108	rBio4.4	0.140
Coif2	0.169	Bior1.3	0.101	rBio5.5	0.242
Coif3	0.124	Bior2.4	0.179	rBio6.8	0.158

Table 10.20 Energy to Shannon entropy ratio of the extracted bearing vibration signal: complex valued wavelets

Base wavelet	Energy to Shannon entropy ratio
Morlet wavelet	60.765
Gaussian wavelet	56.044
B Spline wavelet	35.051
Shannon wavelet	12.476
Harmonic wavelet	14.504

Table 10.21 Information value of the extracted bearing vibration signal: complex valued wavelets

Base wavelet	Information value
Morlet wavelet	0.189
Gaussian wavelet	0.068
B Spline wavelet	0.105
Shannon wavelet	0.017
Harmonic wavelet	0.091

information criterion, it is found that the *reverse Biorthogonal* wavelet 5.5 is the most suited wavelet for analyzing the bearing vibration signal. Details of the results are listed in Table 10.19.

Continuous wavelet transform is also applied to analyze the bearing signals, using the five commonly seen complex-valued wavelets. The energy and Shannon entropy values of the bearing vibration signal extracted by each wavelet are first calculated, and their corresponding *energy-to-Shannon* entropy ratios are then determined. As listed in Table 10.20, the *Complex Morlet* wavelet indicates the maximum *energy-to-Shannon entropy* ratio, thus is considered the most appropriate base wavelet for bearing signal analysis.

In addition, the value of the information measure of the bearing vibration signal extracted by each candidate wavelet is calculated, and the result is shown in Table 10.21. Based on the *maximum information* criterion, the *Complex Morlet* wavelet is again identified as the most suited wavelet, since it demonstrates the highest information value compared to other four candidate wavelets.

10.5 Summary

Using a number of quantitative measures, we presented a systematic approach in selecting a base wavelet that is best suited for analyzing nonstationary signals, typically seen in manufacturing. These measures are examined from two difference aspects: (1) their corresponding wavelet coefficients (i.e., the energy and Shannon entropy measures) and (2) the relationship between the signal being analyzed and the

coefficients of the base wavelet used for the analysis (i.e., joint entropy, condition entropy, mutual information, and relative entropy). Based on these measures, two comprehensive base wavelet selection criteria (i.e., the *maximum energy-to-Shannon entropy ratio* and the *maximum information measure*) are identified as the quantitative measure for determining the best suited wavelet. Both numerical study and experimental data analysis have shown that these two criteria provide quantitative guidance to base wavelet selection for effective signal analysis.

10.6 References

- Abi Abdallah D, Chauvet E, Bouchet Fakri L, Bataillard A, Briguet A, Fokapu O (2006) Reference signal extraction from corrupted ECG using wavelet decomposition for MRI sequence triggering: application to small animals. *BioMed Eng Online* 5(11):1–12
- Abry P (1997) Wavelet and turbulence – multi resolutions, algorithms of decomposition, invariance of scale and signals of pressure. Diderot Editeur, Paris
- Abu Mahfouz I (2005) Drill flank wear estimation using supervised vector quantization neural networks. *Neural Comput Appl* 14(3):167–175
- Ahuja N, Lertrattanapanich S, Bose NK (2005) Properties determining choice of mother wavelet. *IEEE Proc Vis Image Signal Process* 152:659–664
- Arafat S, Skubic M, Keegan K (2003) Combined uncertainty model for best wavelet selection. In: *The IEEE international conference on fuzzy systems*, pp 1195–1199
- Bedekar D, Nair A, Vince DG (2005) Choosing the optimal mother wavelet for decomposition of radio frequency intravascular ultrasound data for characterization of atherosclerotic plaque lesions. *Proc SPIE* 5750:490–502
- Bhatia P, Boudy J, Andreao RV (2006) Wavelet transformation and pre selection of mother wavelet for ECG signal processing. In: *Proceedings of the 24th IASTED international multi conference: biomedical engineering*, Innsbruck, Austria, 15–17 February, pp 390–395
- Bibian S, Zikov T, Dumont GA, Ries CR, Puil E, Ahmadi H, Huzmenzan M, Macleod BA (2001) Estimation of an anesthetic depth using wavelet analysis of electroencephalogram. In: *23rd International conference of the IEEE engineering in medicine and biology society*, Istanbul, Turkey, October
- Bradley AP, Wilson WJ (2004) On wavelet analysis of auditory evoked potentials. *Clin Neurophysiol* 115:1114–1128
- Cover TM, Thomas JA (1991) *Elements of information theory*. Wiley, New York
- Daubechies I (1992) *Ten lectures on wavelets*. SIAM, Philadelphia, PA.
- Emlen JM (1973) *Ecology: an evolutionary approach*. Addison Wesley, Reading, MA
- Flanders M (2002) Choosing a wavelet for single trial EMG. *J Neurosci Methods* 116:165–177
- Fokapu O, Abi Abdallah D, Briguet A (2005) Extracting a reference signal for cardiac MRI gating: experimental study for wavelet functions choice. *Proceedings of 12th international workshop on systems, signals & image processing*, Chalkida, Greece, pp 419–423
- Fu S, Muralikrishnan B, Raja J (2003) Engineering surface analysis with different wavelet bases. *ASME J Manuf Sci Eng* 125:844–852
- Goel P, Vidakovic B (1995) Wavelet transformations as diversity enhancers. *Proc SPIE Int Soc Opt Eng*, 2569:845–857.
- Ho D, Randall RB (2000) Optimization of bearing diagnostic techniques using simulated and actual bearing fault signals. *Mech Syst Signal Process* 14(5):763–788
- Katul G, Vidakovic B (1996) The partitioning of attached and detached eddy motion in the atmospheric surface layer using Lorentz wavelet filtering. *Bound Layer Meteorol* 77(2):153–172

- Li X, Tso SK, Wang J (2000) Real time tool condition monitoring using wavelet transforms and fuzzy techniques. *IEEE Trans Syst Man Cybern C Appl Rev* 30(3):352–357
- Ma X, Zhou C, Kemp IJ (2002a) Automated wavelet selection and thresholding for PD detection. *IEEE Electr Insul Mag* 18(2):37–45
- Ma X, Zhou C, Kemp IJ (2002b) Interpretation of wavelet analysis and its application in partial discharge detection. *IEEE Trans Dielectr Electr Insul* 9(3):446–457
- Marshall AW, Olkin I (1979) *Inequalities: theory of majorization and its application*. Academic, New York
- Mendenhall W, Sincich TL (1995) *Statistics for engineering and the sciences*, 4th edn. Prentice Hall, Englewood Cliffs, NJ
- Mojsilović A, Popović MV, Rackov DM (2000) On the selection of an optimal wavelet basis for texture characterization. *IEEE Trans Image Process* 9(12):2043–2050
- Safavian LS, Kinsner W, Turanli H (2005) A quantitative comparison of different mother wavelets for characterizing transients in power systems. In: *Canadian Conference on Electrical and Computer Engineering*, Saskatoon, Canada, May, pp 1453–1456
- Schukin EL, Zamaraev RU, Schukin LI (2004) The optimization of wavelet transform for the impulsive analysis in vibration signals. *Mech Syst Signal Process* 18(6):1315–1333
- Shao Y, Nezu K (2004) Extracting symptoms of bearing faults in the wavelet domain. *Proc Inst Mech Eng I J Syst Control Eng* 218(1):39–51
- Singh BN, Tiwari AK (2006) Optimal selection of wavelet basis function applied to ECG signal denoising. *Digit Signal Process* 16:275–287
- Tsui PPC, Basir OA (2006) Wavelet basis selection and feature extraction for shift invariant ultrasound foreign body classification. *Ultrasonics* 45:1–14
- Wang S, Liu X, Yianni J, Aziz TZ, Stein JF (2004) Extracting burst and tonic components from surface electromyograms in dystonia using adaptive wavelet shrinkage. *J Neurosci Methods* 139:174–184
- Yang L, Judd MD, Bennoch CJ (2004) Denoising UHF signal for PD detection in transformers based on wavelet technique. *IEEE conference on electrical insulation and dielectric phenomena*, Boulder, CO, October 17–20, pp 166–169
- Yang WX, Ren XM (2004) Detecting impulses in mechanical signals by wavelets. *EURASIP J Appl Signal Process* 8:1156–1162
- Zhang L, Bao P, Wu X (2005) Multiscale LMMSE based image denoising with optimal wavelet selection. *IEEE Trans Circuits Syst Video Technol* 15(4):469–481

Chapter 11

Designing Your Own Wavelet

To achieve effective signal signature extraction, Chap. 10 introduced several quantitative measures for selecting appropriate base wavelets from a pool of available wavelet families, such as Daubechies, Myer, and Morlet wavelets. This chapter introduces a complimentary technique focusing on wavelet customization. The goal is to design a wavelet that is specifically adapted to the signal of interest. Because such a customized wavelet would have a higher degree of matching with the signal than other wavelets, the effectiveness of signature extraction will improve.

11.1 Overview of Wavelet Design

Researchers have studied various techniques for designing base wavelets. In the late 1980s to early 1990s, Daubechies' work has led to the publication of orthonormal (Daubechies 1988) and biorthonormal (Cohen et al. 1992) base wavelets with compact support. These wavelets are independent of the signal to be analyzed. Tewfik et al. (1992) have developed cost functions for finding the optimal orthonormal wavelet basis to represent a specified signal within a finite number of scales. Their work has been extended by assuming band limited signals and finding the optimal M-band wavelet basis within a finite number of scales (Gopinath et al. 1994), for representing a desired signal. During the same period, Aldroubi and Unser (1993) proposed a method to match a wavelet basis to a desired signal by either projecting the desired signal onto an existing wavelet basis, or transforming the wavelet basis under certain conditions such that the error norm between the desired signal and the new wavelet basis is minimum. Recently, Chapa and Rao (2000) have developed two sets of equations for designing a wavelet directly from a signal of interest. The first set of equations derives expressions for continuously matched wavelet spectrum amplitudes, whereas the second set provides a direct discrete algorithm for calculating approximations to the optimal complex wavelet spectrum. By formulating wavelet design as a constrained optimization problem and then solving it by converting the optimization problem into an iterative line-search problem through a first-order parameterization of the perfect reconstruction

constraint, a signal-adapted, biorthogonal filter banks of finite length was constructed by Lu and Antoniou (2001). Later, Shark and Yu (2003) proposed a genetic algorithm-based design method to construct orthonormal wavelet filter banks with an optimal shift-invariant property. On the basis of a generalized Mexican-hat function, the authors also designed a new class of continuous wavelets for arbitrary transient signals (Shark and Yu 2006), where signal matching is achieved by minimizing the spectral difference between the reference signal and the generalized Mexican-hat wavelet. Gupta et al. (2005a) have proposed to construct wavelets that are matched to a given signal in the statistical sense. The main idea is to first estimate a high-pass wavelet filter from the statistics of the signal, and then obtain a FIR/IIR biorthogonal perfect reconstruction filter bank. This leads to the construction of a statistically matched wavelet. The authors have also designed both biorthogonal and semiorthogonal wavelet from a signal by maximizing projection of the signal onto successive scaling subspaces while minimizing energy of the signal in the wavelet subspace (Gupta et al. 2005b). Using the same idea, Guido et al. (2006) designed a spikelet wavelet that has shown improved performance on pattern recognition of signals corresponding to neural action potentials of H1, a motion sensitive neuron in the fly's visual system. These prior efforts motivate our study of application-specific base wavelets for improved signature extraction in signals related to manufacturing.

11.2 Construction of an Impulse Wavelet

Considering that base wavelets available in the literature (e.g., provided by MATLAB) are developed primarily from a mathematical point of view without reference to a specific physical system although in real-world applications, signals to be analyzed are generally produced by physical systems it would be interesting, from an intellectual pursuit point of view, to study how to construct a customized base wavelet from the physical phenomena being analyzed. Naturally, such construction process will have to satisfy the mathematical requirement for designing a base wavelet. With this in mind, we introduce an impulse wavelet designed for analyzing vibration signals measured from rolling bearings, which are widely used in manufacturing machines.

Generally, a base wavelet must satisfy the conditions as described in Chaps. 3 and 4 to ensure that a signal's wavelet transformation does not result in loss of information so that the signal can be properly reconstructed from the corresponding wavelet coefficients. Mathematically, such a reconstruction exists if a scaling function $\phi(t)$, which satisfies the following *dilation* equation (Burrus et al. 1998; Cui et al. 1994), can be found as:

$$\frac{1}{\sqrt{2}}\phi\left(\frac{t}{2}\right) = \sum_n h_n \phi(t - n) \quad (11.1)$$

In (11.1), h_n is a set of scaling coefficients applied to $\phi(t - n)$. Equation (11.1) indicates that the dilated version of $\phi(t)$ can be written as the sum of translated versions that are scaled by the coefficients h_n . Furthermore, it indicates that a scaling function at one scale can be constructed from a number of scaling functions at a previous scale. In general, the construction of a base wavelet starts from the scaling function that satisfies (11.1). Such scaling function is then used to derive the base wavelet.

Assuming an impulsive input is applied to a rolling bearing, a corresponding output signal can be defined by the convolution integral in the continuous form as (Inman 1996; Lutes and Sarkani 1997):

$$x(t) = \int_0^t R(\tau)h(t - \tau)d\tau \quad (11.2)$$

where $R(\tau)$ denotes the impulsive input, and $x(t)$ denotes the output signal.

In the discrete form, the impulsive input $R(\tau)$ is sampled at $R(n)$, and the output signal can be obtained as:

$$x(t) = \sum_n R(n)h(t - n) \quad (11.3)$$

In (11.2) and (11.3), the symbol $h(\bullet)$ represents the impulse response of the rolling bearing. Considering the discrete form of convolution expression, the similarity between (11.3) and (11.1) becomes apparent: the output $x(t)$ in (11.3) can be viewed as the sum of translated versions of the impulse response $h(\bullet)$ that are scaled by the input $R(n)$. If the impulse response satisfies (11.1), then it can be used to form a scaling function that contains relevant information on the underlying dynamics of the bearing being monitored. Subsequently, the scaling function can be used to construct a base wavelet for analyzing vibration signals measured from the bearing. Because of the nature of such a derivation, it is expected that the base wavelet presents a more direct and meaningful decomposition of the bearing signal than the standard wavelets commonly found in the literature.

To construct the base wavelet, several impulse responses of a ball bearing have been taken through hammer strikes, as shown in Fig. 11.1. The corresponding spectra are shown in Fig. 11.2. It is seen that the frequency components below 1,500 Hz are consistent in terms of their magnitudes, but those above 1,500 Hz have varied. The magnitudes of these high frequency components are smaller than those of the lower frequency components.

Because of their relatively small magnitude and random behavior, the high frequency components were treated as noise and removed from the signal with a low pass filter. The cutoff frequency for the filter was chosen to be 1,500 Hz, as the spectral components of each impulse were stable below this frequency. As seen in Fig. 11.3 where the original and filtered signals are shown, the filter is effective in removing noise from the impulse response and retaining the frequency constant of the original signal.

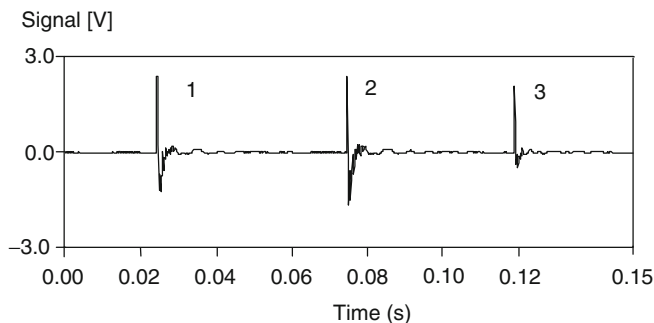


Fig. 11.1 Waveform of three consecutive impulse responses from a ball bearing

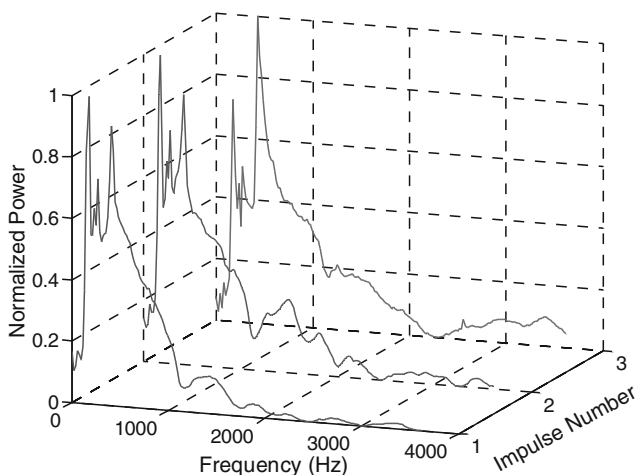


Fig. 11.2 Spectra of three consecutive impulse responses from a ball bearing

In order for the filtered impulse response shown in Fig. 11.3 to be used as a scaling function, it must satisfy the dilation equation, for which the length of the support interval of the signal must be at least one. A function with a support interval of less than one would have a gap between $h_n\phi(t - n)$ and $h_{n+1}\phi(t - n - 1)$, within which nothing contributes to the sum on the right hand side of (11.1). Consequently, such a function would not satisfy the dilation equation. To address this issue, the impulse response is first dilated such that its support is greater than one. After dilation, the coefficients h_n are determined from a recursive relationship that is derived from the dilation equation.

As an example of the procedure of satisfying the dilation equation, a standard Daubechies scaling function $\phi(t)$ (Fig. 11.4) is illustrated below. It should be noted first that an explicit expression for the Daubechies scaling function does not exist (Daubechies 1992).

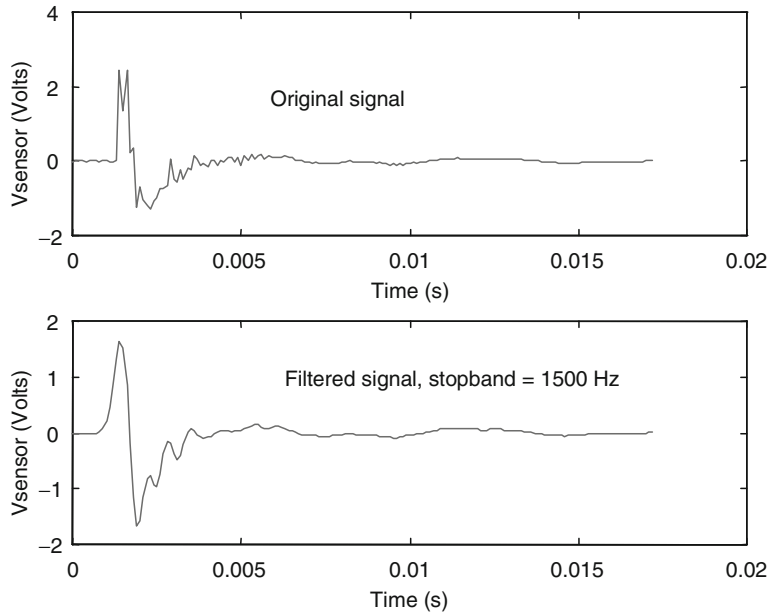


Fig. 11.3 Impulse response of a ball bearing

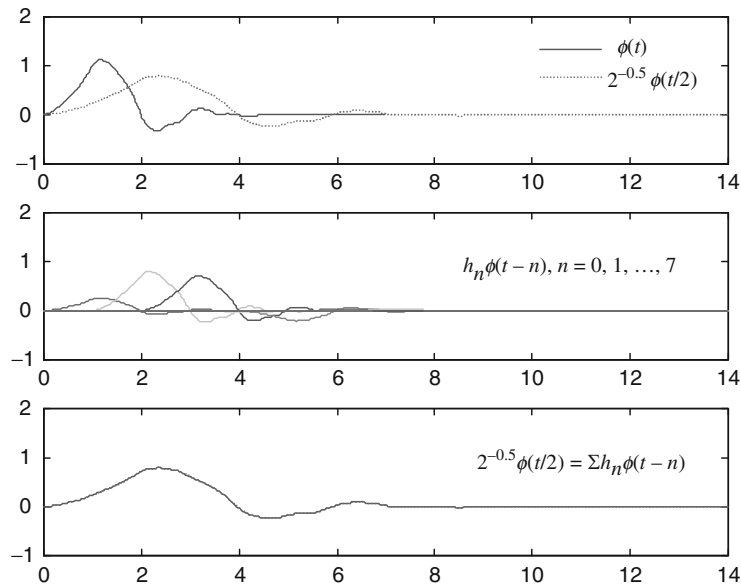


Fig. 11.4 The Daubechies scaling function

After the sequence of coefficients h_n is obtained, they are used to form an FIR filter denoted by H^* . The FIR filters H^* and G^* are quadrature mirror filters if, for a signal $x(t)$:

$$||H^*x(t)|| + ||G^*x(t)|| = ||x(t)|| \quad (11.4)$$

Together, H^* and G^* form a pair of reconstruction filters for the wavelet decomposition of a signal. This process, called *deconstruction*, is implemented via the adjoints of H^* and G^* , which are denoted by H and G , respectively (Kaiser 1994). For the Daubechies scaling function $\phi(t)$ shown in Fig. 11.4a, the filter coefficients are as follows: $h_n = \{0.2304, 0.7148, 0.6309, -0.0280, -0.1870, 0.0308, 0.0329, -0.0106\}$, $n = 0, 1, \dots, 7$ (Misiti et al. 1997). In Fig. 11.4b, the scaled and translated version of $\phi(t)$ (i.e., $h_n\phi(t-n)$ for $n = 0, 1, 2, \dots, 7$) is shown. Since $\phi(t)$ is a valid scaling function and h_n are valid filter coefficients, the dilation equation is satisfied, as shown in Fig. 11.4c.

The above procedure is repeated for the impulse response as shown below in Fig. 11.5. It should be noted that the impulse response $\phi(t)$ here is a function of the bearing dynamics, not an exact solution to the dilation equation. However, a set of filter coefficients h_n can be determined such that the impulse response approximately satisfies the dilation equation.

The filter coefficients can be calculated such that the dilation equation is satisfied exactly at integer values of t . The solution is recursive: each h_n , for $n > 0$, can be

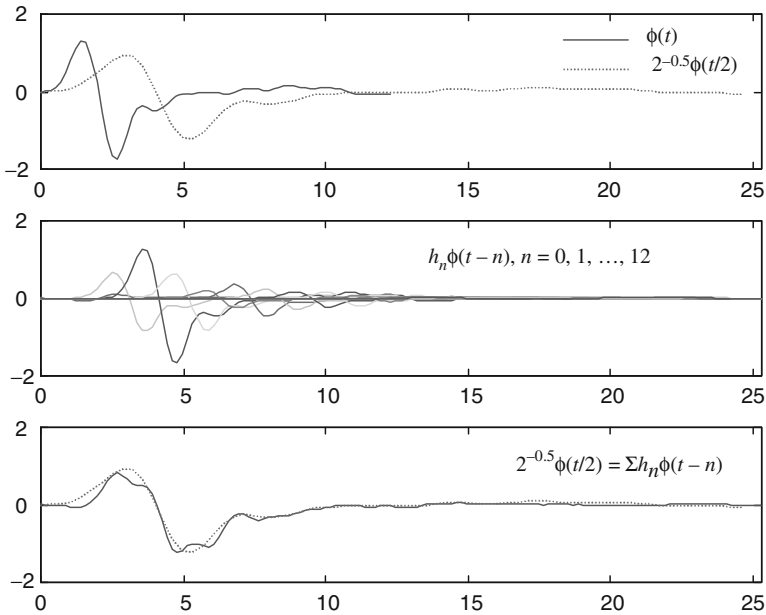


Fig. 11.5 The impulse scaling function obtained from the ball bearing

explicitly determined as a function of $\phi(t)$ and $h_0, h_1, h_2, \dots, h_{n-1}$. The first coefficient, h_0 , is simply a function of $\phi(t)$. These solutions are obtained by evaluating the dilation equation at integer values of t . For $t = 1$, (11.1) gives $\phi(1/2)/\sqrt{2} = h_0\phi(1)$. The terms $h_1\phi(0)$, $h_2\phi(-1)$, etc., do not appear because $\phi(t) = 0$ for $t \leq 0$. (Recall that ϕ is compactly supported.) Thus, h_0 is determined by:

$$h_0 = \frac{2^{-1/2}\phi(1/2)}{\phi(1)} \quad (11.5)$$

Similarly, for $t = 2$, (11.1) gives:

$$\frac{1}{\sqrt{2}}\phi(1) = h_0\phi(2) + h_1\phi(1) \quad (11.6)$$

For $t = 3$:

$$\frac{1}{\sqrt{2}}\phi\left(\frac{3}{2}\right) = h_0\phi(3) + h_1\phi(2) + h_2\phi(1) \quad (11.7)$$

For $t = N + 1$:

$$\frac{1}{\sqrt{2}}\phi\left(\frac{N+1}{2}\right) = h_0\phi(N+1) + h_1\phi(N) + h_2\phi(N-1) + \dots + h_N\phi(1) \quad (11.8)$$

Equations (11.5)–(11.8) determine a recursive definition for each filter coefficient h_n . With the first coefficient h_0 given by (11.5), the remaining coefficients are given by:

$$h_n = \frac{2^{-1/2}\phi((n+1)/2) - \sum_{k=0}^{n-1} h_k\phi(n+1-k)}{\phi(1)}, \quad \text{for } n \geq 1 \quad (11.9)$$

Since the scaling function $\phi(t)$ is given by the impulse response of the bearing, each of the filter coefficients h_n can be readily determined from (11.5) and (11.9). Furthermore, note that the dilation equation can be written as:

$$\frac{1}{\sqrt{2}}\phi\left(\frac{t}{2}\right) = h_0\phi(t) + \sum_{j=1}^N h_j\phi(t-j) \quad (11.10)$$

Since h_n is given by (11.9), the dilation equation can be rewritten as:

$$\frac{1}{\sqrt{2}}\phi\left(\frac{t}{2}\right) = \frac{\sum_{j=1}^N \left[2^{-1/2}\phi((j+1)/2) - \sum_{k=0}^{j-1} h_k\phi(j+1-k) \right] \phi(t-j)}{\phi(1)} + h_0\phi(t) \quad (11.11)$$

Collecting terms yields the following form for the dilation equation:

$$\begin{aligned} \frac{1}{\sqrt{2}} \phi\left(\frac{t}{2}\right) &= \frac{\sum_{j=1}^N [2^{-1/2} \phi((j+1)/2) \phi(t-j)]}{\phi(1)} \\ &\quad - \frac{\sum_{j=1}^N \sum_{k=0}^{j-1} h_k \phi(j+1-k) \phi(t-j)}{\phi(1)} + h_0 \phi(t) \end{aligned} \quad (11.12)$$

Note that only the second term on the right hand side of (11.12) contains filter coefficients h_n , which are determined by (11.5) and (11.9). Equation (11.12) serves to illustrate the interesting relationship that the dilation equation imposes between h_n and $\phi(t)$. Particularly, the expression given by (11.12) shows that the dilated version of the scaling function is related not only to $\phi(t-n)$ scaled by the filter coefficient h_n , but also to $\phi(t-n)$, scaled by $\phi(t)$, which is evaluated at integer values of t . The recursive relationship given by (11.9) gives h_n such that the dilation equation is satisfied at $x = \{0, 1, 2, \dots\}$. At other points, the sum on the right hand side of (11.1) might differ from the left hand side. In practical treatment of an impulse scaling function such as shown in Fig. 11.5a, (11.5) and (11.9) are first used to obtain an initial set of filter coefficients. These coefficients are then optimized by minimizing the following error function:

$$E_{\text{rms}} = \sqrt{\frac{1}{T} \int_0^T \left(\frac{1}{\sqrt{2}} \phi\left(\frac{t}{2}\right) - \sum_n h_n \phi(t-n) \right)^2 dt} \quad (11.13)$$

The error E_{rms} is a scalar valued function of the vector of filter coefficients h_n , and the optimization is accomplished by finding the vector which minimizes E_{rms} . Since E_{rms} is a measure of how well the dilation equation is satisfied, the vector h_n minimizing E_{rms} is the best set of filter coefficients that can be obtained from $\phi(t)$. Using this technique, the filter coefficients are determined to be: $h_n = \{-0.0529, 0.4897, 0.9601, 0.4848, 0.1467, 0.2653, 0.1723, 0.1295, 0.1208, 0.0495, -0.0182, -0.0255, 0.0131\}$, for $n = 0, 1, \dots, 12$. The translated and scaled versions of $\phi(t)$ corresponding to these h_n (i.e., $h_n \phi(t-n)$) are plotted in Fig. 12.5b. As indicated by Fig. 12.5c, the impulse response is an approximate solution to the dilation equation ($E_{\text{rms}} = 0.0984$). The low pass filter coefficients derived from this scaling function $\phi(t)$ can then be used to determine the corresponding wavelet $\psi(t)$ (Young 1993; Mallat 1998). The coefficients for the high pass reconstruction filter G^* are determined from (11.4). The wavelet is evaluated by upsampling G^* , convolving it with H^* , and then iteratively repeating this procedure:

$$H_{n+1}^* = \uparrow G^* * H_n^* \quad (11.14)$$

where \uparrow is a dyadic up-sampling operator. Thus, after N iterations, $\psi(t) \cong H_{N+1}^*$. Figure 11.6 shows the result of four iterations of (11.14), which produced a

customized wavelet, based on the impulse response of the rolling bearing structure. The set of FIR filters based on $\psi(t)$ and synthesis based on $\phi(t)$ are given in Table 11.1, where the filters have been normalized to have a norm of $1/\sqrt{2}$.

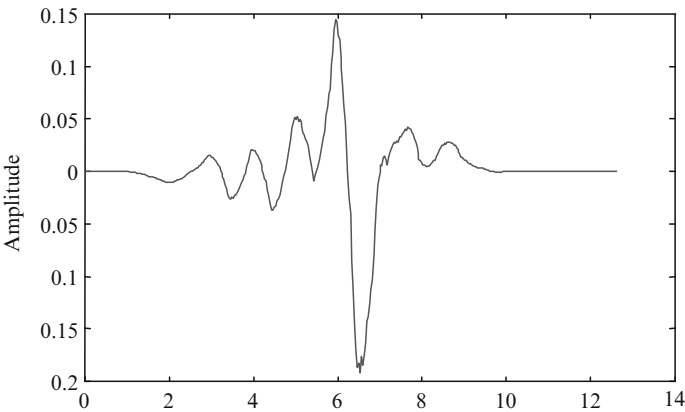


Fig. 11.6 The wavelet derived from the impulse response

Table 11.1 Normalized filter coefficients

Deconstruction		Reconstruction	
Low pass	High pass	Low pass	High pass
0	0.0274	0.0274	0
0.0068	0.2532	0.2532	0.0068
0.0132	0.4964	0.4964	0.0132
0.0094	0.2507	0.2507	0.0094
0.0256	0.0759	0.0759	0.0256
0.0625	0.1372	0.1372	0.0625
0.0670	0.0891	0.0891	0.0670
0.0891	0.0670	0.0670	0.0891
0.1372	0.0625	0.0625	0.1372
0.0759	0.0256	0.0256	0.0759
0.2507	0.0094	0.0094	0.2507
0.4964	0.0132	0.0132	0.4964
0.2532	0.0068	0.0068	0.2532
0.0274	0	0	0.0274

11.3 Impulse Wavelet Application

As an application example, the impulse wavelet is used to diagnose bearing defect. Figure 11.7a shows a vibration signal measured on a SKF 6220 ball bearing with a 0.25-mm hole on its inner raceway. This signal is sampled at 10 kHz, and the rotating speed of the bearing is 600 rpm (i.e., corresponding to 10 Hz rotating frequency). Based on geometric dimensions of the bearing and the rotating speed (Harris 1991), the defect characteristic frequency on the inner raceway of such bearing is $f_{BPFI1} = 58.6$ Hz. As illustrated in Fig. 11.7b, such a defect-related frequency component is not seen in its power spectral density (PSD) resulted from the Fourier transform.

Utilizing the wavelet integrated with Fourier transform technique, which is described in Chap. 7, the same vibration signal is first analyzed by the wavelet transform. The impulse wavelet, developed from the impulse response of the rolling bearing as described above, is used as the base wavelet. Fourier transform is then performed on the wavelet coefficients obtained from the wavelet transform to expose explicitly the related frequency components. Figure 11.8 illustrates the resulting wavelet coefficients and their corresponding PSD. It is seen that the defect-related frequency component f_{BPFI1} at 58.6 Hz is clearly shown in the spectrum, thus verifying the existence of a localized inner raceway defect.

To demonstrate the signature extraction capability of the designed impulse wavelet for bearing defect diagnosis, a comparison study is carried out, where

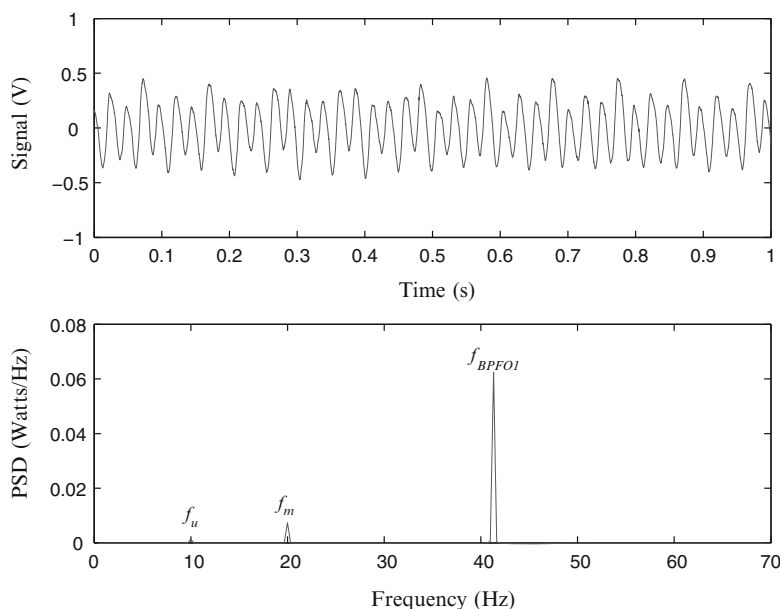


Fig. 11.7 Vibration signal and its PSD from a defective bearing

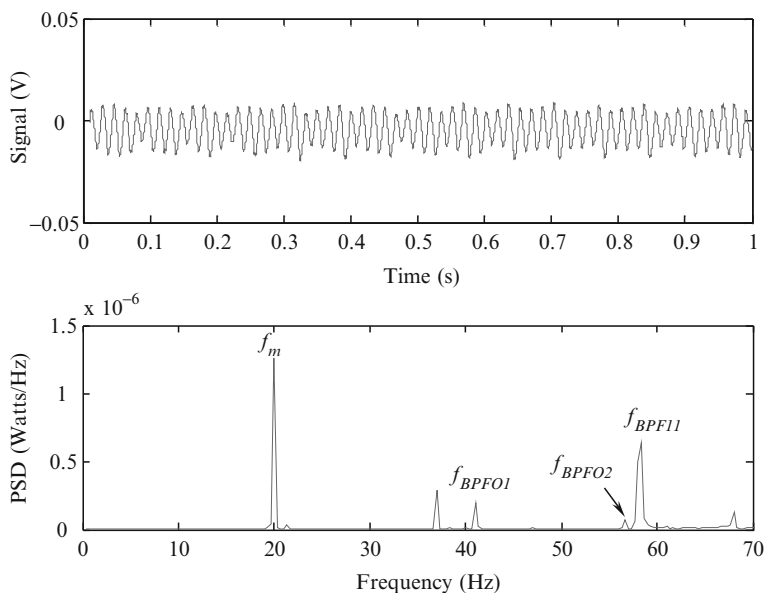


Fig. 11.8 Wavelet integrated Fourier spectrum using the customized impulse wavelet

five standard base wavelets from the literature: Daubechies 2 and 4 wavelets, Coiflets 1, Symlets 3, and Biorthogonal 2.2 (Daubechies 1992; Lou and Loparo 2004; Zhang et al. 2005) are used to analyze the vibration signal. The upper parts of Figs. 11.9–11.13 are intermediate results (i.e., wavelet coefficients) of the integrated wavelet-Fourier transform analysis, and the lower parts of these figures are their corresponding PSDs. It is shown that all the five standard base wavelets can identify the defect-related frequency component, and the results are shown in the lower parts of Figs. 11.9–11.13.

In the spectra of Figs. 11.9–11.13, there exists a frequency component f_{BPFO2} at 56.5 Hz, which has a distinct magnitude. Such a frequency component is identified as from the ball rotation of another bearing in the support structure (Yan et al. 2009). To enable a quantitative performance comparison of the developed impulse wavelet and other five standard base wavelets, a signal-to-noise ratio measure is introduced, which is the amplitude ratio between the defect frequency f_{BPF11} and the adjacent frequency f_{BPFO2} . As listed in Table 11.2, the impulse wavelet has shown the highest signal-to-noise ratio in detecting the defect-characteristic frequency of $f_{BPF11} = 58.6$ Hz. This result can be attributed to the nature of this customized wavelet, which is derived from the actual impulse response of the bearing structure. The direct link to the dynamics of the bearing and thus inherent better match to the bearing signature than a standard wavelet has made it more effective in exposing the constituent features for defect identification.

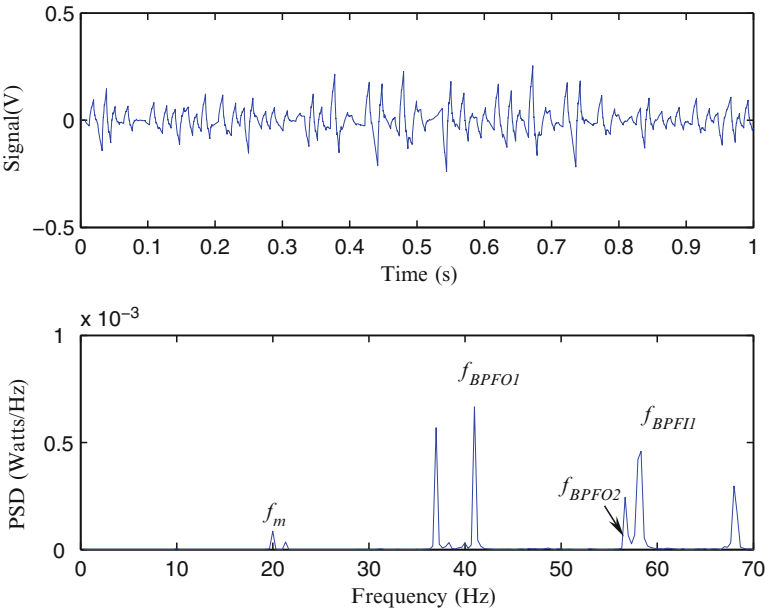


Fig. 11.9 Wavelet integrated Fourier spectrum results using Daubechies 2 (Db2) wavelet

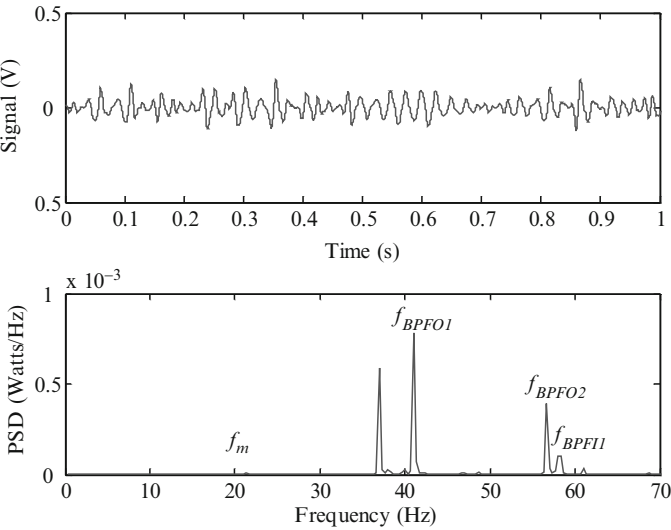


Fig. 11.10 Wavelet integrated Fourier spectrum results using Daubechies 4 (Db4) wavelet

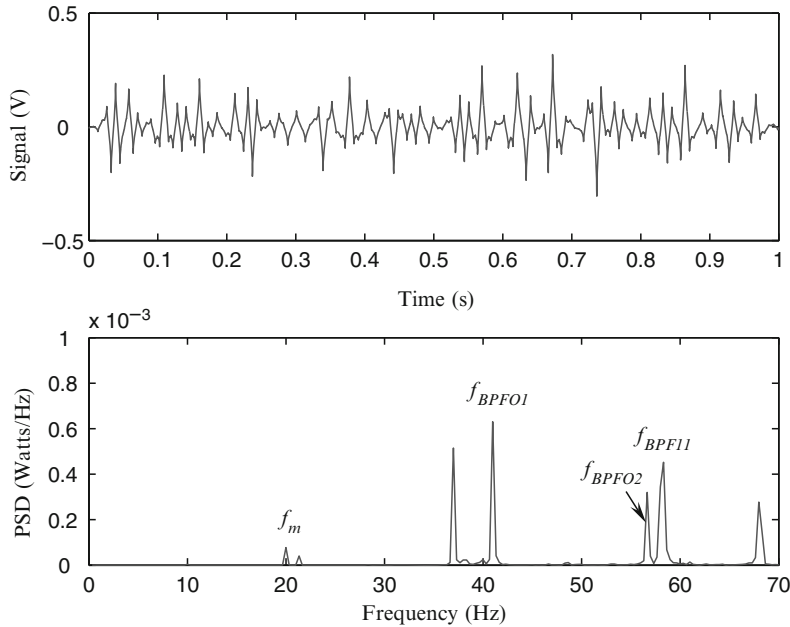


Fig. 11.11 Wavelet integrated Fourier spectrum results using Coiflets 1 (Coif1) wavelet

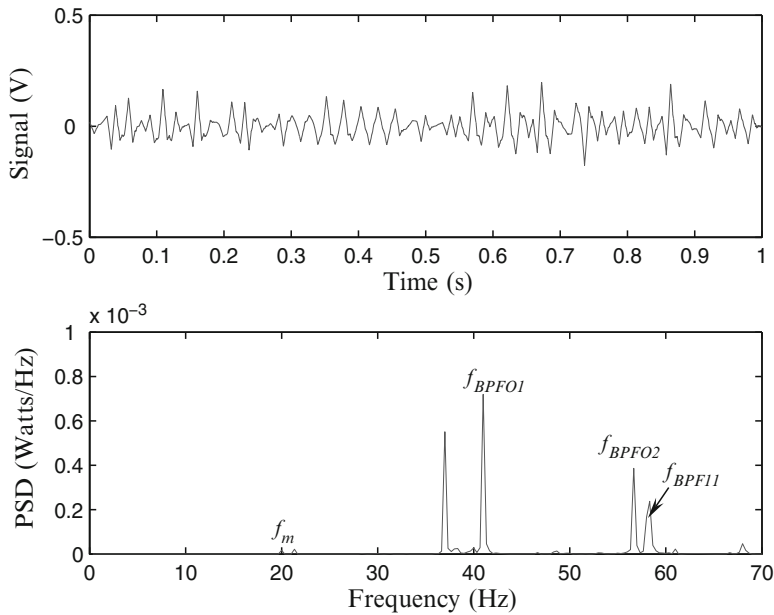


Fig. 11.12 Wavelet integrated Fourier spectrum results using Symlets 3 (Sym3) wavelet

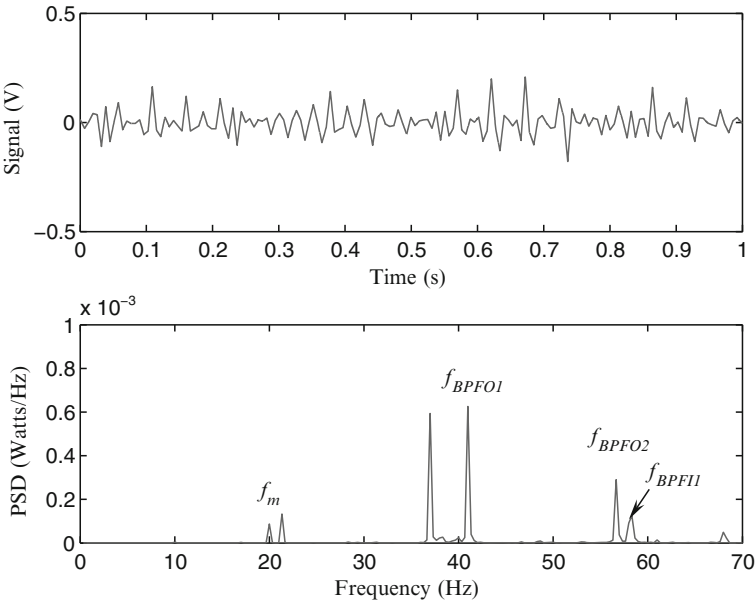


Fig. 11.13 Wavelet integrated Fourier spectrum results using Biorthogonal 2.2 (Bior2.2) wavelet

Table 11.2 Comparison of signal to noise ratios for different base wavelets

Base wavelet	f_{BPFI1}/f_{BPFO2}
Impulse	9.55
Db2	1.88
Db4	0.27
Coif1	1.41
Sym3	0.62
Bior2.2	0.43

11.4 Summary

This chapter introduces the procedure to design a wavelet based on the dynamics of the physical system being analyzed. Using the impulse response of a rolling bearing system, an impulse wavelet has been constructed for defect-induced signature extraction. Experimental study has verified the effectiveness of the impulse wavelet in identifying bearing localized defect of the bearing in its inner raceway, as illustrated in the comparative study involving five standard wavelets from the literature. Although the impulse wavelet development is based on a specific type of bearing, the analytical procedure described in this chapter should be applicable to the analysis of other types of mechanical systems.

11.5 References

- Aldroubi A, Unser M (1993) Families of multiresolution and wavelet spaces with optimal properties. *Numer Funct Anal Optim* 14:417–446
- Burrus CS, Gopinath R, Guo H (1998) *Introduction to wavelets and wavelet transforms: a primer*. Prentice Hall, Englewood Cliffs, NJ
- Chapa JO, Rao RM (2000) Algorithm for designing wavelets to match a specified signal. *IEEE Trans Signal Process* 48(12):3395–3406
- Cohen A, Daubechies I, Feauveau JC (1992) Biorthogonal bases of compactly supported wavelets. *Commun Pure Appl Math* 45:485–560
- Cui CK, Montefusco L, Puccio L (1994) *Wavelet: theory, algorithms, and applications*. Academic, New York
- Daubechies I (1988) Orthonormal bases of compactly supported wavelets. *Commun Pure Appl Math* 41:909–996
- Daubechies I (1992) *Ten lectures on wavelets*. Society of Industrial and Applied Mathematics, Pennsylvania, PA
- Gopinath RA, Odegard JE, Burrus CS (1994) Optimal wavelet representation of signals and wavelet sampling theorem. *IEEE Trans Circuits Syst II Analog Digital Signal Process* 41:262–277
- Guido RC, Slaets JFW, Koberle R, Almeida LOB, Pereira JC (2006) A new technique to construct a wavelet transform matching a specified signal with applications to digital, real time, spike, and overlap pattern recognition. *Digit Signal Process* 16:22–44
- Gupta A, Joshi SD, Prasad S (2005a) A new approach for estimation of statistically matched wavelet. *IEEE Trans Signal Process* 53(5):1778–1793
- Gupta A, Joshi SD, Prasad S (2005b) A new method of estimating wavelet with desired features from a given signal. *Signal Processing* 85:147–161
- Harris T (1991) *Rolling bearing analysis*. Wiley, New York
- Inman D (1996) *Engineering vibration*. Prentice Hall, Englewood Cliffs, NJ.
- Kaiser G (1994) *A Friendly Guide to Wavelets*. Birkhauser, Boston, MA
- Lou X, Loparo KA (2004) Bearing fault diagnosis based on wavelet transform and fuzzy inference. *Mech Syst Signal Process* 18:1077–1095.
- Lu WS, Antoniou A (2001) Design of signal adapted biorthogonal filter banks. *IEEE Trans Circuits Syst I Fundam Theory Appl* 48:90–102.
- Lutes L, Sarkani S (1997) *Stochastic analysis of structural and mechanical vibrations*. Prentice Hall, Englewood Cliffs, NJ
- Mallat S (1998) *A wavelet tour of signal processing*. Academic, Boston, MA
- Misiti M, Misiti Y, Oppenheim G, Poggi J (1997) *Wavelet toolbox for use with Matlab*. The Math Works, Inc., Natick, MA
- Shark L, Yu C (2003) Design of optimal shift invariant orthonormal wavelet filter banks via genetic algorithm. *Signal Processing* 83:2579–2591
- Shark L, Yu C (2006) Design of matched wavelets based on generalized Mexican hat function. *Signal Processing* 86:1451–1469
- Tewfik AH, Sinha D, Jorgensen P (1992) On the optimal choice of a wavelet for signal representation. *IEEE Trans Inf Theory* 38:747–765
- Yan R, Gao R, Wang C (2009) Experimental evaluation of a unified time scale frequency technique for bearing defect feature extraction. *ASME J Vib Acoust* 131:041012 1–12
- Young R (1993) *Wavelet theory and its applications*. Kluwer Academic Publishers, Boston, MA
- Zhang S, Mathew J, Ma L, Sun Y (2005) Best basis based intelligent machine fault diagnosis. *Mech Syst Signal Process* 19:357–370

Chapter 12

Beyond Wavelets

In previous chapters, we have introduced the theoretical foundation and practical applications related to the wavelet transform. The ability of wavelet transform in adaptive time-scale representation and decomposition of a signal into different subfrequency band presents an efficient signal analysis method without introducing calculation burden (Sweldens 1998). Consequently, it has become a prevailing tool for nonstationary signal processing (e.g., transient pattern identification and location). Given, however, the great variety of signals that appear in real-world applications, there remains plenty of room for continued advancement in the theory of the classical wavelet transform. For example, one of the limitations of the wavelet transform is to modify the base wavelet function to better analyze signals of finite length or duration, instead of infinite or periodic signals (Sweldens 1997). In addition, it has limitations in precisely capturing and defining the geometry of image edges. In this chapter, we introduce several new developments in signal and image processing that address these limitations and extend beyond the scope of the classical wavelet transform method (Jiang et al. 2006, 2008; Li et al. 2008; Zhou et al. 2010).

12.1 Second Generation Wavelet Transform

Second generation wavelet transform (SGWT), as an advanced mathematical tool for time-scale representation of signals, has been developed to overcome deficiencies of the classical wavelet transform. Specifically, the mechanism of constructing a base wavelet from the translation and dilation of a fixed function has been replaced by the so-called *lifting scheme* (Sweldens 1996, 1998). The resulting wavelet transform has the following properties (Uytterhoeven et al. 1997):

1. It is a generic method that is faster to calculate and easier to implement than the classical wavelet transform.
2. It can transform signals with a finite length without extension of the signal to infinite duration.

3. It can be applied to irregular signal samplings and extended for the determination of weighting functions.
4. Its inverse transform shares the same complexity as the forward transform.

12.1.1 Theoretical Basis of SGWT

The architecture of the lifting scheme can be illustrated in Fig. 12.1. The forward procedure of the lifting scheme, similar to its counterpart in the classical discrete wavelet transform, is to obtain both the approximation and detail of the original signal. It mainly incorporates three critical operational steps: (1) splitting, (2) prediction, and (3) updating. When starting the lifting scheme process, the signal $x(i)$ is first split into two subsets, the odd sample x_{odd} and the even sample x_{even} , by means of a sample sequence. For example, given a signal $x(i)$, where $i = 1, 2, 3, \dots, 2n$ (n is a natural number), it will be split as:

$$\begin{cases} x_{\text{odd}} = \{x(2i - 1)\} \\ x_{\text{even}} = \{x(2i)\} \end{cases}, \quad i = 1, 2, 3, \dots, n \quad (12.1)$$

When the splitting procedure of the signal $x(i)$ is completed, the odd and even subsamples are obtained and the signal is subsampled by a factor of 2.

Following the splitting operation is the prediction operation, which predicts the odd data sample with the even data sample as:

$$\overline{x_{\text{odd}}} = P(x_{\text{even}}) \quad (12.2)$$

In (12.2), P is the prediction operator that is independent of the signal. The difference between the predicted result and the odd sample is considered as the detail of the original signal, d , described as:

$$d = x_{\text{odd}} - \overline{x_{\text{odd}}} = x_{\text{odd}} - P(x_{\text{even}}) \quad (12.3)$$

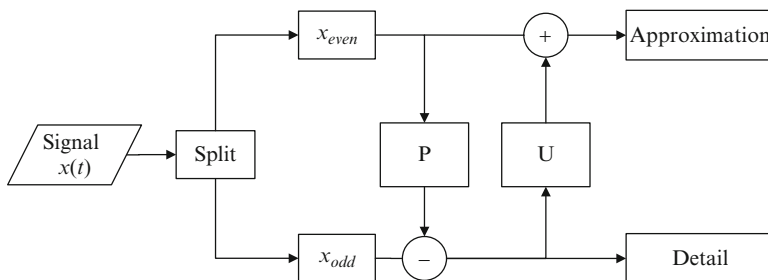


Fig. 12.1 Forward transform procedure of lifting scheme

Given the x_{even} and the detail, the approximation can be calculated with the updating operator U as:

$$a = x_{\text{even}} + U(d) \quad (12.4)$$

Similar to the prediction operation, the updating operation is also independent of the signal to be analyzed. The functions of prediction and updating operators are similar to that of a pair of $h(n)$ and $g(n)$ filters in the classical wavelet transform, and they can be derived from the scaling function $\phi(t)$ and wavelet function $\psi(t)$ by iteration algorithm (Claypoole 1999; Claypoole et al. 2003). It should be noted that the prediction and updating operators can be optimized using different algorithms, such as the Claypoole's optimization algorithm (Claypoole 1999; Claypoole et al. 2003).

Based on the forward procedure described above, the signal is decomposed into two parts: approximation and detail. This process can be iterated by taking the approximation as the input signal to continue the decomposition. Furthermore, by iterated decomposition of the detail and the approximation together, wavelet packet transform can be realized with the lifting scheme.

The decomposition is invertible, and the signal reconstruction procedure is illustrated in Fig. 12.2.

As the forward procedure realizes decomposition of the original signal, the reverse procedure realizes signal reconstruction. Similar to the forward procedure, the reverse procedure involves both the prediction operator and the updating operator. This means that the following relationship exists:

$$\begin{cases} x_{\text{odd}} = d + P(x_{\text{even}}) \\ x_{\text{even}} = a - U(d) \end{cases} \quad (12.5)$$

The signal can then be reconstructed by merging x_{even} and x_{odd} .

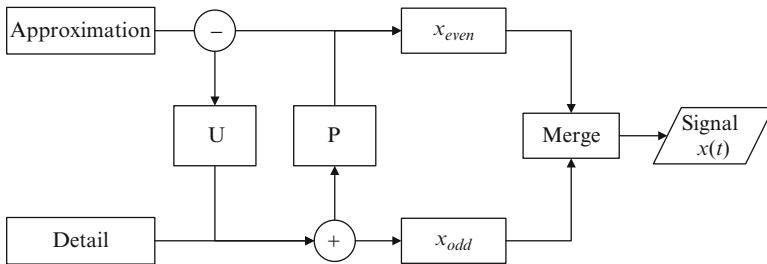


Fig. 12.2 Reverse transform procedure of lifting scheme

12.1.2 Illustration of SGWT in Signal Processing

Several examples are illustrated here on the application of the SGWT algorithm. The first example involves a signal that consists of two frequency components, sampled at 100 Hz as:

$$x(t) = \sin(2\pi \cdot 11t) + \sin(2\pi \cdot 41t) \quad (12.6)$$

The signal can be separated by one-level SGWT. Performing SGWT on this signal, where the db8 wavelet function is used as the starting point to derive the prediction and updating operators and to get the approximation part $a1$ and detailed part $d1$. The result is shown in Fig. 12.3. The accuracy of the decomposition result is evaluated through calculation of the error. Specifically, the absolute values of subtracting approximation coefficients in $a1$ and detailed coefficients in $d1$ at each sampling point from the original signal are summed up, as illustrated below:

$$\text{error} = \sum_{i=1}^N [x(i) - (a1(i) + d1(i))] \quad (12.7)$$

Performing the above calculation, the resulted error is only 1.26×10^{-12} , which verifies the accuracy of the decomposition.

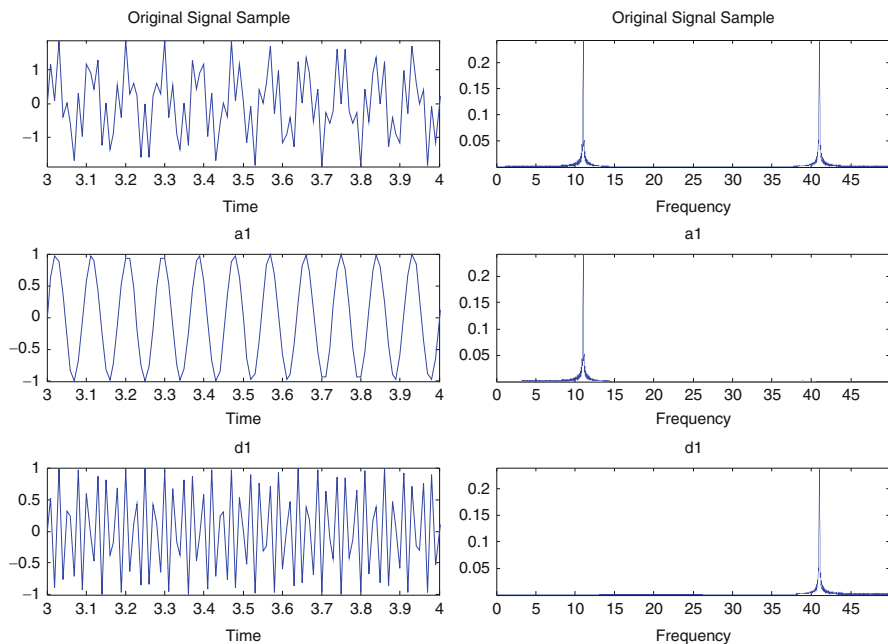


Fig. 12.3 Separation and reconstruction of two sine waves

The second example involves an intermittent linear chirp signal as expressed in (12.7):

$$x(t) = \begin{cases} \sin\left[2\pi\left(\frac{t+20}{3}\right)t\right] & t \in [1, 4] \cup [6, 9] \\ 0 & \text{else} \end{cases} \quad (12.8)$$

The signal is decomposed using the SGWT, and the results are shown in Fig. 12.4.

Adopting the same concept as that in the first example, the error of this transform is calculated as 4.09×10^{-13} , which again verifies the accuracy of the decomposition result using the SGWT.

In the area of manufacturing, surface topography has been considered as one of the factors that affect the functional performance of components. The features of a surface, such as the roughness and waviness, have direct impact on the wear rate of the component. To identify these surface features, the SGWT has been used for surface analysis (Jiang et al. 2001a, b, 2008). As an example, the bearing surface of a worn metallic femoral head is shown in Fig. 12.5a, where two different types of scratches (a regular scratch that is related to manufacturing process and a random scratch that is generated during the service time) exist (Jiang et al. 2001b). With the application of the SGWT to processing the bearing surface, the waviness feature can be clearly seen in Fig. 12.5b.

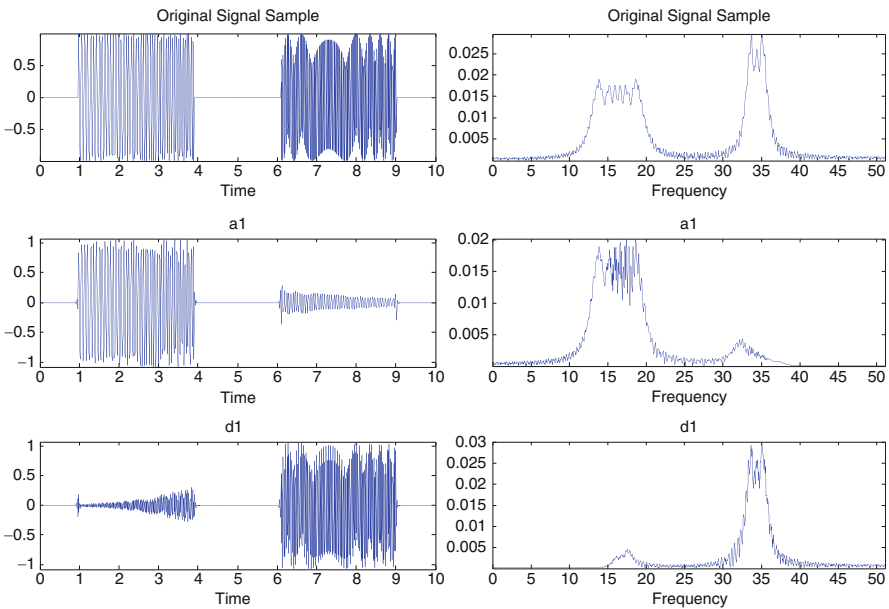


Fig. 12.4 Separation and reconstruction of intermittent linear chirp

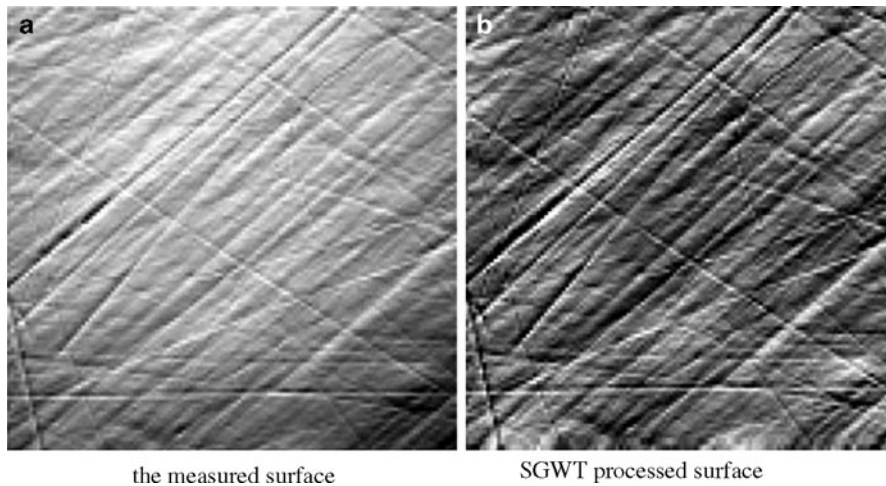


Fig. 12.5 Bearing surface of a new metallic femoral head (Jiang et al. 2001b). (a) The measured surface and (b) SGWT processed surface

12.2 Ridgelet Transform

Classical wavelet transform has been found insufficient in addressing certain problems in image processing. Prominent among the limitations is the fact that wavelets are essentially *isotropic* (i.e., its characteristics is uniform in all directions) in nature and are therefore inadequate for analyzing anisotropic features in images (Starck et al. 2006). Such constraint on the applicability of wavelets to image processing has led to research in improved methods of representation and analysis. One such method is called ridgelet analysis, which was developed by researchers at the Stanford University in 1998 (Candes 1998; Candes and Donoho 1999). The analysis is based on ridge functions that were known since the late 1970s (Logan and Shepp 1975).

12.2.1 Theoretical Basis of Ridgelet Transform

Ridgelets and the associated ridgelet analysis present a multiscale representation of mathematical functions through the superposition of *ridge* functions. The ridge functions are expressed as $r(a_1x_1 + a_2x_2 + \cdots + a_nx_n)$ (Candes and Donoho 1999). They are a set of functions with n variables, and are constant along the hyperplanes $a_1x_1 + a_2x_2 + \cdots + a_nx_n = c$. A graphical representation of the ridgelet functions is given in Fig. 12.6.

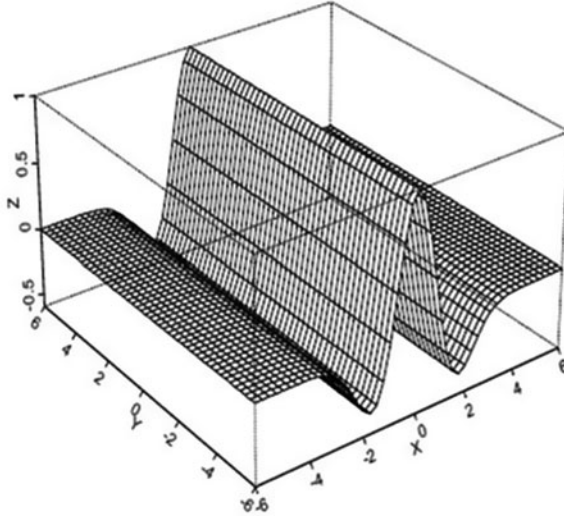


Fig. 12.6 A ridgelet with ridge functions marked by lines parallel to y axis (Starck et al. 2003)

Furthermore, a ridge function can also be expressed as a multivariate function ($f : \mathbb{R}^n \rightarrow \mathbb{R}$) of a set of real numbers as

$$f(x_1, x_2, \dots, x_n) = g(a_1x_1 + \dots + a_nx_n) = g(ax) \quad (12.9)$$

where $g : \mathbb{R} \rightarrow \mathbb{R}$ is a function of a set of real numbers, and $a = (a_1, a_2, \dots, a_n) \in \mathbb{R}$ is a vector representing the direction. The multivariate function f with n variables can be further approximately represented by a superposition of m ($m < n$) ridge functions (Candes 1998; Candes and Donoho 1999; Starck et al. 2006) as

$$f(x_1, x_2, \dots, x_n) \approx \sum_{i=1}^m c_i \sigma(a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n) \quad (12.10)$$

where c_i denotes the coefficients and m denotes the number of ridge functions.

Ridgelet transform is associated with the ridge functions, and the concept of ridgelet transform is similar to that of the Fourier transform in that it is associated with the periodic sine and cosine functions, as mathematically expressed below.

Consider a smooth, univariate function ψ , such that $\psi : \mathbb{R} \rightarrow \mathbb{R}$, with a vanishing mean, $\int \psi(t) dt = 0$. Given this function, we can further define a bivariate function $\psi_{a,b,\theta} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ as (Candes and Donoho 1999):

$$\psi_{a,b,\theta}(x) = a^{-1/2} \psi\left(\frac{x_1 \cos \theta + x_2 \sin \theta - b}{a}\right) \quad (12.11)$$

In (12.11), $x = (x_1, x_2) \in \mathbb{R}^2$, $a > 0$ is the dilation parameter, $b \in \mathbb{R}$ represents the translation parameter, and $\theta \in [0, 2\pi)$ represents the direction parameter.

Equation (12.11) represents a ridgelet, whereas the dilation and translation parameters given above perform the function of scaling and translating the ridgelet, similar to the dilation (by the scaling factor s) and translation (by the time constant τ) operations in a wavelet. The function $\psi_{a,b,\theta}(x)$ is constant along the lines (i.e., ridges) $x_1 \cos \theta + x_2 \sin \theta = \text{constant}$. Transverse to these ridges, it is a wavelet function. Accordingly, the continuous ridgelet transform of any integrable bivariate function can be expressed as (Candes 1998; Candes and Donoho 1999; Starck et al. 2006).

$$R_f(a, b, \theta) = \int \psi_{a,b,\theta}(x) f(x) dx \quad (12.12)$$

It is interesting to note that $\psi_{a,b,\theta}$ is defined on the \mathbb{R}^2 space and the associated transform is therefore 2D. The inverse of (12.12) used in reconstruction is given as (Candes and Donoho 1999)

$$f(x) = \int_0^{2\pi} \int_{-\infty}^{\infty} \int_0^{\infty} \frac{4\pi}{a^3} R_f(a, b, \theta) \psi_{a,b,\theta}(x) da db d\theta \quad (12.13)$$

12.2.2 Application of the Ridgelet Transform

The 2D nature of the ridgelet transform makes it very well suited for analyzing and processing images. Prominent ridgelet applications include denoising, edge detection, and classification of tissues from images of internal human organs. As an example, Fig. 12.7 shows two images of a supernova before and after denoising using ridgelets, respectively (Starck et al. 2003).

It can be seen that the original x-ray image (the left side of Fig. 12.7) is blurred by the noise, while the image becomes clear after the noise has been filtered out using the ridgelet transform (the right side of Fig. 12.7).

Another example of the application of the ridgelet transform is to characterize surface topography (Ma et al. 2005). This is an important issue, as it has impacts on the mechanical and physical properties of the system. Figure 12.8 shows the results of extracting deep scratches from a honed surface, which were seen in an automotive engine cylinder (Ma et al. 2005). Their distribution of such scratches on the surface and their amplitudes directly affect the gas or air flow and pressure balance in an engine. In Fig. 12.8b, deep scratches are reconstructed by means of a ridgelet transform.

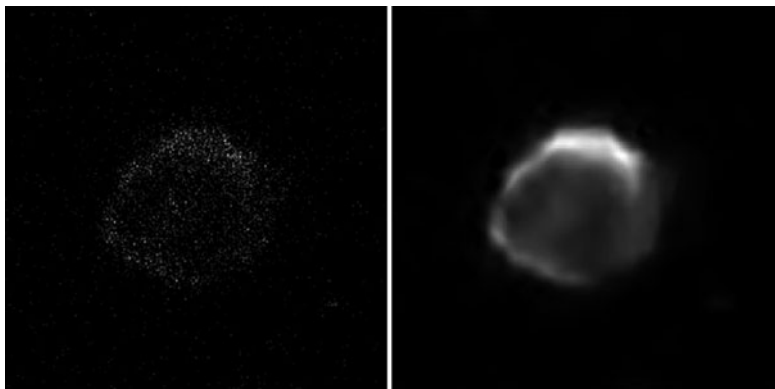


Fig. 12.7 Image denoising using the ridgelet transform. Reproduced from Starck et al. (2003)

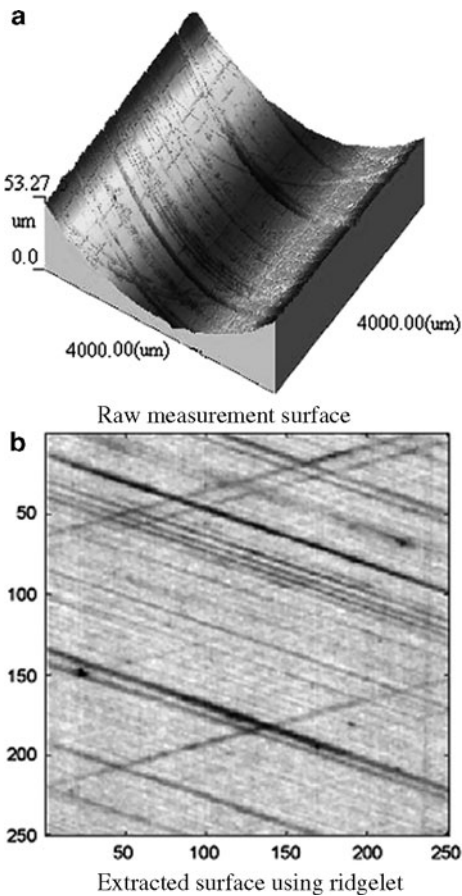


Fig. 12.8 Extraction of linear scratches from a honed surface of the automotive engine cylinder (Ma et al. 2005). (a) Raw measurement surface and (b) extracted surface using ridgelet

12.3 Curvelet Transform

Ridgelet transform is a relatively new system of representation and analysis, which has been shown to be effective in resolving edges (Candes and Donoho 1999; Do and Vetterli 2003; Dettori and Semler 2007). However, ridgelets are limited to resolving only straight edges; curved edges cannot be represented with as few coefficients as required for a straight edge (Do and Vetterli 2003; Starck et al. 2003). The inadequacy of wavelets and ridgelets in resolving edges has been the primary driving force behind the search for better representation and analysis. Curvelet analysis, introduced in the year 2000 (Candes and Donoho 2000), holds potential in addressing the shortcomings of wavelets and ridgelets. A brief introduction to the fundamentals of curvelets is given in this section.

12.3.1 Curvelet Transform

The curvelet transform is defined as the inner product of the function f to be analyzed and a family of curvelets $\gamma_{ab\theta}$ (Candes and Donoho 2000, 2005a, b):

$$\Gamma_f(a, b, \theta) = \langle f, \gamma_{ab\theta} \rangle \quad (12.14)$$

where $a > 0$ is the scale parameter, b is the translation parameter, and $\theta \in [0, 2\pi)$ is the orientation parameter. The symbol Γ_f represents the curvelet transform. The family of curvelets is explained by starting with two smooth, nonnegative, real valued windowing functions called the *radial* window $W(r)$ and the *angular* window $V(t)$, respectively (Candes and Donoho 2005a, b). The two windowing functions are subject to the following two admission conditions:

$$\int_0^\infty \frac{1}{a} W(ar)^2 da = 1, \quad \forall r > 0 \quad (12.15)$$

$$\int_{-1}^1 V(t)^2 dt = 1 \quad (12.16)$$

In (12.5), $r \in (1/2, 2)$ is the radial coordinate and in (12.6) $t \in [-1, 1]$ denotes the time variable. According to the definition in (12.14), at a given scale a , a family of curvelets can be generated by translation and rotation of a basic element γ_{a00} as shown in (12.17) (Candes and Donoho 2005a, b):

$$\gamma_{ab\theta} = \gamma_{a00}(R_\theta(x - b)) \quad (12.17)$$

where R_θ is a 2×2 rotation matrix that is related to planar rotation by an angle θ .

The basic element itself is expressed mathematically as:

$$\gamma_{a00}(r, \omega) = W(ar) V\left(\frac{\omega}{\sqrt{a}}\right) a^{3/4} \quad (12.18)$$

where r and ω are polar coordinates defined in the frequency domain.

Generally, the discrete curvelet transform is often used to process the function f , which also starts with two window functions: the *radial* window $W(r)$ and the *angular* window $V(t)$ (Candes et al. 2006). The transform is subject to the conditions expressed in (12.19) (12.21) as:

$$\sum_{j=-\infty}^{\infty} W(2^j r)^2 = 1 \quad (12.19)$$

$$\sum_{l=-\infty}^{\infty} V(t - l)^2 = 1 \quad (12.20)$$

$$U_j(r, \theta) = 2^{-3j/4} W(2^{-j} r) V\left(\frac{2^{j/2} \theta}{2\pi}\right) \quad (12.21)$$

In (12.21), the window function U_j is derived from the radial window $W(r)$ and the angular window $V(t)$ and expressed in the Fourier domain. The symbols $r \in (3/4, 3/2)$ and θ denote polar coordinates, and $t \in (-1/2, 1/2)$ is the time variable.

Based on these notations, a family of curvelets at a fixed scale of 2^j is defined as:

$$\varphi_{j,l,k}(x) = \varphi_j(R_{\theta_l}(x - x_k^{(j,l)})) \quad (12.22)$$

where $x_k^{(j,l)} = R_{\theta_l}^{-1}(k_1 2^{-j}, k_2 2^{-j/2})$ represents the position information, and $R_{\theta} = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix}$ represents the rotation information in terms of θ radians, respectively.

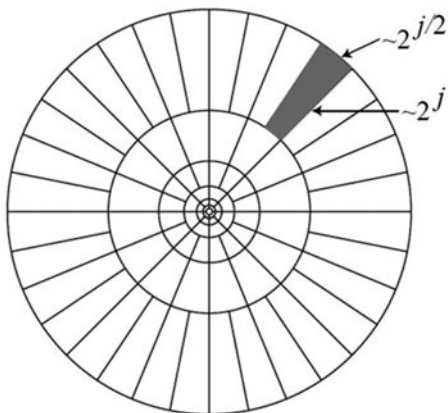
Accordingly, the inner product between a curvelet $\varphi_{j,l,k}$ and a function f results in the curvelet coefficients as:

$$c(j, l, k) = \langle f, \varphi_{j,l,k} \rangle = \int_{\mathbb{R}^2} f(x) \varphi_{j,l,k}(x) dx \quad (12.23)$$

where $c(j, l, k)$ are the curvelet coefficients.

The physical interpretation of (12.23) in the Fourier domain at a scale of 2^j can be illustrated in Fig. 12.9. The concentric circles represent the family of curvelets $\varphi_{j,l,k}(x)$ and the shaded portion represents one curvelet from this family.

Fig. 12.9 A family of curvelets in polar coordinates, with the shaded area representing support for a single curvelet (Candes et al. 2006)



While the curvelets described above are expressed in the polar coordinates, Cartesian coordinates are desirable to implement the curvelet transform (Donoho and Duncan 2000; Candes et al. 2006). Consequently, the window functions expressed in (12.19)–(12.21) are expressed in the Cartesian coordinates as

$$W_j(\omega) = \sqrt{\Phi_{j+1}^2(\omega) - \Phi_j^2(\omega)}, \quad j > 0 \quad (12.24)$$

$$V_j(\omega) = V\left(\frac{2^{\lfloor j/2 \rfloor} \omega_2}{\omega_1}\right) \quad (12.25)$$

$$U_j(\omega) = W_j(\omega) V_j(\omega) \quad (12.26)$$

where $\Phi(\omega_1, \omega_2) = \phi(2^{-j} \omega_1) \phi(2^{-j} \omega_2)$, $2^j \leq \omega_1 \leq 2^{j+1}$, $-2^{-j/2} \leq (\omega_2/\omega_1) \leq 2^{-j/2}$, and $0 \leq \phi \leq 1$, $\phi = \begin{cases} 1, & [-1/2, 1/2] \\ 0, & [-2, 2] \end{cases}$.

Currently, there are two prevalent methods of calculating the curvelet coefficients. The first method, called the digital curvelet transform via unequipped fast Fourier transform (Candes et al. 2006), involves the following four steps:

1. Calculate the 2D FT of the function of interest ($f[t_1, t_2]$), to obtain its Fourier samples, $\hat{f}[n_1, n_2]$.
2. Resample $\hat{f}[n_1, n_2]$ for each scale/angle pair (j, l) to obtain sampled values $\hat{f}[n_1, n_2 - n_1 \tan \theta_l]$.
3. Multiply the resampled \hat{f} with the window function U_j , to obtain $\tilde{f}_{j,l}[n_1, n_2] = \hat{f}[n_1, n_2 - n_1 \tan \theta_l] U_j[n_1, n_2]$, where $\tan \theta_l = l 2^{-\lfloor j/2 \rfloor}$.
4. Calculate the inverse of the 2D FT of every $\tilde{f}_{j,l}$ to obtain the discrete curvelet coefficients, $c(j, l, k)$.

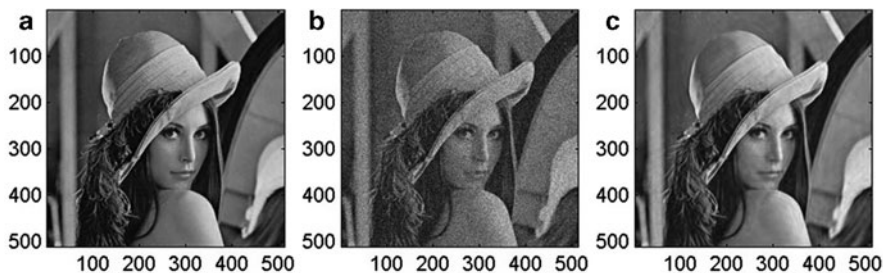
The second method is called digital curvelet transform via wrapping (Candes et al. 2006), and involves the following computational steps:

1. Calculation of 2D FT of the function of interest ($f[t_1, t_2]$), to obtain its Fourier samples, $\hat{f}[n_1, n_2]$.
2. Calculation of the product $U_{j,l}[n_1, n_2] \hat{f}[n_1, n_2]$.
3. Wrapping the product $U_{j,l}[n_1, n_2] \hat{f}[n_1, n_2]$ around the origin to obtain $\tilde{f}_{j,l}[n_1, n_2] = W(U_{j,l} \hat{f})[n_1, n_2]$.
4. Calculation of inverse 2D FFT of every $\tilde{f}_{j,l}$ to obtain discrete curvelet coefficients, $c(j, l, k)$.

12.3.2 Application of the Curvelet Transform

Because of its multiscale nature, most of the curvelet applications are related to image processing. Specifically, the curvelet transform has been applied to image compression, contrast enhancement, feature extraction from noisy images, pattern detection, noise filtration, edge detection, etc. As an example, a denoising operation on a test image “Lena” is shown in Fig. 12.10. This image, which is 512×512 pixels in size (Fig. 12.10a), was contaminated with random noise (peak signal-to-noise ratio, PSNR: 22.1), as shown in Fig. 12.10b. Curvelet transform is then applied for image denoising via thresholding of its curvelet coefficients. The result is a filtered image as seen in Fig. 12.10c, which has an improved PSNR value of 31.1.

In manufacturing, the curvelet has been used for surface characterization. Fig 12.11a shows a surface of a worn metallic femoral head. When additive white noise is added into it, the surface has a signal-to-noise ratio (SNR) of 56.43 dB. When the wavelet transform is used to remove the noise, the SNR of the surface has increased to 58.72 dB. Finally, the performance of the denoising operation was further improved when the curvelet transform is applied to process the surface, where the SNR has increased to 61.62 dB.



12.10 (a) Original image, (b) image contaminated with noise, and (c) image after denoising

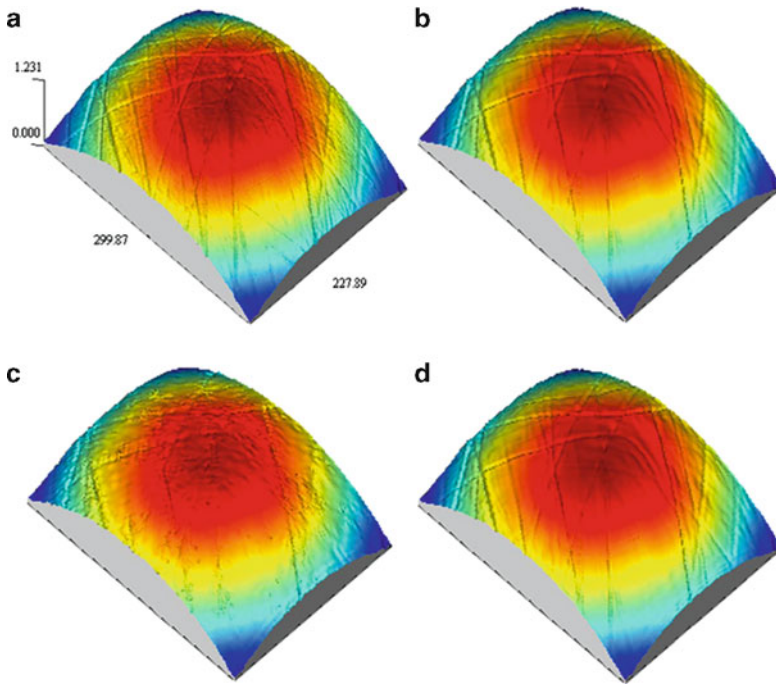


Fig. 12.11 Denoising of a microscalar surface (Ma 2007). (a) Original surface (scale unit: μm), (b) noisy surface (signal to noise ratio, SNR = 56.43 dB), (c) denoised surface using wavelet (SNR = 58.72 dB), and (d) denoised surface using curvelet (SNR = 61.62 dB)

12.4 Summary

This chapter briefly presents development in signal processing that goes beyond the classical wavelet transform. The SGWT based on the lifting scheme is first introduced to allow for the design of the base wavelet functions to better fit the signal to be analyzed. The ridgelet and curvelet transforms are then introduced from the point of view of overcoming the limitations in detecting edges (straight and curved) of images when applying the classical wavelet transform. These techniques, together with further advancement reported in the literature, such as multiwavelet transform (Cotronei et al. 1998), dual-tree wavelet transform (Selesnick et al. 2005), and contourlet transform (Do and Vetterli 2005), promise to continually push the envelope of signal and image processing to better serve the needs for a wide range of engineering problems.

12.5 References

- Candes EJ (1998) Ridgelets: theory and applications. Ph.D. Dissertation, Stanford University
- Candes EJ, Donoho DL (1999) Ridgelets: a key to higher dimensional intermittency. *Philos Trans R Soc Math, Phys Eng Sci* 357:2495–2509
- Candes EJ, Donoho DL (2000) Curvelets – a surprisingly effective nonadaptive representation for objects with edges. In: Rabut C, Cohen A, Schumaker LL (eds) *Curves and surfaces*. Vanderbilt University Press, Nashville, TN
- Candes EJ, Donoho DL (2005a) Continuous curvelet transform: 1. resolution of the wavefront set. *Appl Comput Harmon Anal* 19:162–197
- Candes EJ, Donoho DL (2005b) Continuous curvelet transform: 2. discretization and frames. *Appl Comput Harmon Anal* 19:198–222
- Candes EJ, Demanet L, Donoho DL, Ying L (2006) Fast discrete curvelet transforms. *SIAM Multiscale Model Simul* 5:861–899
- Claypoole R (1999) Adaptive wavelet transforms via lifting. Thesis: computer engineering, Rice University
- Claypoole R, Davis G, Sweldens W (2003) Nonlinear wavelet transform for image coding via lifting. *IEEE Trans Image Process* 12(12):1449–1459
- Cotronei M, Montefusco LB, Puccio L (1998) Multiwavelet analysis and signal processing. *IEEE Trans Circuits Syst II Analog Digital Signal Process* 45(8): 970–987
- Dettori L, Semler L (2007) A comparison of wavelet, ridgelet and curvelet based texture classification algorithms in computed tomography. *Comput Biol Med* 37:486–498
- Do MN, Vetterli M (2003) The finite ridgelet transform for image representation. *IEEE Trans Image Process* 12:16–28
- Do MN, Vetterli M (2005) The contourlet transform: an efficient directional multiresolution image representation. *IEEE Trans Image Process* 14(12):2091–2106
- Donoho DL, Duncan MR (2000) Digital curvelet transform: strategy, implementation and experiments. *Proc SPIE* 4056:12–29
- Jiang HK, Wang ZS, He ZJ (2006) Wavelet design for extracting weak fault feature based on lifting scheme. *Front Mech Eng China* 1(2):199–203
- Jiang X, Blunt L, Stout KJ (2001a) Application of the lifting wavelet to rough surfaces. *J Int Soc Precision Eng Nanotechnol* 25:83–89
- Jiang X, Blunt L, Stout KJ (2001b) Lifting wavelet for three dimensional surface analysis. *Int J Mach Tools Manuf* 41:2163–2169
- Jiang X, Scott P, Whitehouse D (2008) Wavelets and their application in surface metrology. *CIRP Ann Manuf Technol* 57:555–558
- Li Z, He ZJ, Zi YY, Jiang HK (2008) Rotating machinery fault diagnosis using signal adapted lifting scheme. *Mech Syst Signal Process* 22(3):542–556
- Logan BF, Shepp LA (1975) Optimal reconstruction of a function from its projections. *Duke Math J* 42:645–659
- Ma J, Jiang X, Scott P (2005) Complex ridgelets for shift invariant characterization of surface topography with line singularities. *Phys Lett A* 344:423–431
- Ma J (2007) Curvelets for surface characterization. *Appl Phys Lett* 90: 054109 1–3
- Selesnick IW, Baraniuk RG, Kingsbury NG (2005) The dual tree complex wavelet transform. *IEEE Signal Process Mag* 22(6): 123–151
- Starck JL, Donoho DL, Candes EJ (2003) Astronomical image representation by the curvelet transform. *Astron Astrophys* 398:785–800
- Starck JL, Moudden Y, Abrial P, Nguyen M (2006) Wavelets, ridgelets and curvelets on the sphere. *Astron Astrophys* 446:1191–1204
- Sweldens W (1996) The lifting scheme: a custom design construction of biorthogonal wavelets. *Appl Comput Harmon Anal* 3:186–200

- Sweldens W (1997) Second generation wavelets: theory and application. http://www.ima.umn.edu/industrial/97_98/sweldens/fourth.html. Accessed 30 June 2009
- Sweldens W (1998) The lifting scheme: a construction of second generation wavelets. *SIAM J Math Anal* 29(2):511–546
- Uytterhoeven G, Dirk R, Adhemar B (1997) Wavelet transforms using the lifting scheme. Department of Computer Science, Katholieke Universiteit Leuven, Belgium
- Zhou R, Bao W, Li N, Huang X, Yu DR (2010) Mechanical equipment fault diagnosis based on redundant second generation wavelet packet transform. *Digit Signal Process* 20(1):276–288

Index

A

Admissibility condition, 33, 38, 47
Aliasing, 20
Angular window, 214, 215
Approximate coefficient, 55, 60, 67
Axial load, 114, 115, 120, 122

B

Band pass filter, 83, 84, 86, 89, 125
Base template function, 104 106, 113
Base wavelet, 12, 28 31, 33, 34, 39, 40, 49, 52, 53,
56, 60 64, 66, 67, 69, 72, 79, 83 100, 103,
108, 112, 113, 116, 165 186, 189 191,
197 199, 202, 205, 218
Basis function, 12, 26, 154, 166
Bearing, 1, 8, 9, 13, 35, 42, 62, 63, 65, 78,
79, 83, 87, 93 99, 103, 109, 110,
113 125, 129, 132, 134 146, 183 186,
190 195, 197 199, 202, 209, 210
Bearing defect diagnosis, 93 99, 100,
113 124, 198
Bearing vibration signal, 79, 97, 98, 110, 114,
183 186
Biorthogonal wavelet, 63 65, 167, 168, 184

C

Center frequency, 41, 43, 44, 46, 90 92, 96, 97,
176, 177, 180 182
Children node, 153, 154
Chirp function, 46 47
Class, 129 131, 149 155, 157, 158, 162, 190
Classification, 1 5, 13, 125 146, 149 162,
167 169, 212
Coiflet wavelet, 62, 63
Compact support, 66, 166, 167, 189
Completeness, 53

Complex conjugate, 17, 34, 40, 50
Complex valued signal, 85
Complex valued wavelet, 86 87, 176, 179 185
Computation procedure, 39
Conditional entropy, 172 173
Continuous frame, 104
Continuous wavelet transform, 33 47, 49,
107, 112, 179, 184, 185
Convolution theorem, 40, 84
Correlation index, 151 152
Covariant, 36 37
Curvelet analysis, 214
Curvelet transform, 214 218
Customized wavelet, 189, 197, 199

D

Daubechies function, 106
Daubechies, I., 30, 31
Daubechies scaling function, 192 194
Daubechies wavelet, 30, 61 63, 165 168
Decomposition level, 60, 66, 67, 76, 116 119,
123, 131, 151 153, 155, 160, 176,
177, 183
Deconstruction, 194, 197
Defect feature extraction, 115 118, 120
Defect severity, 13, 125 146, 160, 162
Denoising, 13, 44, 49, 65 67, 79, 80, 168,
169, 212, 213, 217, 218
Derived template function, 104 106
Detailed coefficient, 56, 66, 67, 208
Deterministic signal, 1 5
Diagnosis, 11 13, 62, 65, 93 100, 113 124,
130, 131, 134, 146, 149, 165, 167, 169,
171, 175, 183, 198
Dilation, 12, 28, 34 37, 205, 212
equation, 190, 192, 194 196
regularity, 53

Discrete Fourier transform, 19
 Discrete frame, 104, 105
 Discrete wavelet transform, 13, 34, 49 69, 107,
 111, 112, 114, 116, 124, 166, 206
 Discriminant power, 129, 130, 138, 139, 143,
 144, 154 157
 Discrimination, 129, 153, 154, 158, 168
 Dissimilarity measure, 149 154
 Dual scale equation, 56 58, 68

E

Eigen value, 132 134
 Energy, 11, 19, 33, 34, 38, 65, 74, 93, 95,
 97, 103, 104, 120, 123, 125 128,
 130, 137, 139, 140, 143, 146, 150,
 151, 160, 161, 170 172, 177, 178,
 180 185, 190
 Energy difference, 151
 Energy to Shannon entropy ratio, 172, 178,
 181, 183 185
 Enveloping, 83 100
 Error function, 196
 Expectation operation, 127, 131

F

Feature, 11, 12, 33, 35, 49, 60, 93, 96, 97,
 100, 105, 107 110, 112 118,
 120 146, 149, 154, 156, 158, 160 162,
 165 168, 171, 174, 175, 199, 209,
 210, 217
 Feature selection, 128 134, 136 138, 143
 Fisher linear discriminant, 128 131, 140
 Fourier function, 106
 Fourier transform, 11, 12, 17 31, 33 35, 40,
 74, 75, 77, 84, 85, 93, 103 124, 198,
 199, 211, 216
 Frequency B Spline wavelet, 43
 Frequency shifted keying (FSK) signal,
 108, 109

G

Gabor transform, 12
 Gaussian modulated sinusoidal signal,
 176 178, 182
 Gaussian pulse function, 46
 Gaussian wavelet, 42, 167, 180 182, 185
 Gearbox, 10, 13, 83, 93, 100, 125, 149,
 150, 158 162
 Gearbox defect classification, 158 161
 Generalized frame, 104, 106 109

H

Haar, A., 26
 Haar function, 106
 Haar wavelet, 26, 61, 70, 71, 108
 Hard thresholding, 66
 Harmonic wavelet, 44, 180 182, 185
 packet transform, 74 77
 transform, 29, 74 76
 Hilbert transform, 83 86, 125

I

Impulse response, 96, 191 199, 202
 Impulse wavelet, 169, 190 202
 Inclusion relationship, 53, 54
 Inequality, 168
 Injection molding, 6, 8, 87 90, 100
 Inner product, 17, 18, 22, 28, 37, 40, 50, 51,
 72, 75, 166, 214, 215
 Inverse continuous wavelet transform, 38 39
 Inverse discretized wavelet transform, 52
 Isotropic, 210

J

Joint entropy, 172 175, 178, 179, 181,
 184, 186

K

Kurtosis, 125, 127 128, 130, 137, 140,
 143, 146

L

Leakage, 20
 Levy, P., 26, 27
 Lifting scheme, 205 207, 218
 Linear discriminant analysis (LDA), 156,
 157, 160
 Local discriminant base, 149 162
 Low pass filter, 60, 70, 83, 84

M

Mallat algorithm, 58 60, 68
 Mallat, S., 30
 Manufacturing, 1 13, 17, 21, 30, 61, 63, 65,
 100, 103, 124, 149, 162, 167, 169,
 185, 190, 209, 217
 Maximum information, 175, 178, 182 183, 185
 Measure function, 105, 107, 108, 110 112
 Message, 108, 109

Mexican hat wavelet, 41, 165, 167, 190
 Meyer wavelet, 65, 177, 178
 Meyer, Y., 30, 31, 33
 Milling, 5, 6, 130
 Modulus, 86, 87, 95, 100
 Monotonicity, 53
 Morlet, J., 27
 Morlet wavelet, 28, 29, 41 42, 96, 165, 167,
 180 182, 185, 189
 Moyal principle, 37 38
 Multiresolution analysis (MRA), 29, 30,
 53 56, 58, 68
 Multi scale, 83 100, 210, 217
 Multi scale enveloping, 83 100
 Multiscale enveloping spectrogram
 (MuSEnS), 93, 95 100
 Mutual information, 169, 173 175, 178, 179,
 182, 184, 186

N

Neural network classifier, 134 137, 140,
 141, 145, 146, 156
 Non deterministic signal, 1, 3 5
 Non stationarity, 21, 152 153
 Non stationary signal, 21

O

Orthogonal basis, 53, 54, 154
 Orthogonality, 53, 72, 154, 166, 167
 Orthogonal wavelet, 30, 53 56, 60, 68,
 154, 166
 Orthonormal eigenvectors, 132
 Orthonormal system, 26

P

Parent node, 153, 154
 Parseval's theorem, 37
 Periodic signal, 1 3, 18, 205
 Prediction, 206 208
 Pressure measurement, 87 93, 100
 Principal component analysis (PCA),
 128, 131 134, 139, 141 143,
 145, 146
 Probability density function, 127, 150, 168
 Pruning approach, 153
 Pulse differentiation, 87 93, 100

Q

Quadrature mirror filters, 194

R

Radial load, 97 100, 113 115, 120 122,
 130, 137, 142
 Radial window, 214, 215
 Real valued signal, 83 86
 Real valued wavelet, 176 180, 184
 Reconstruction, 66, 189, 190, 194, 196, 197,
 207 209, 212
 Rectifying, 84
 Recursive algorithm, 73 74
 Reflection, 91, 93
 Regularity, 53, 167, 177
 Relative entropy, 150, 153, 154, 160, 169,
 175, 178, 180, 182, 184, 186
 Ridge analysis, 210
 Ridge function, 210, 211
 Ridgelet transform, 210 214
 Rotary machine, 21, 93 99
 Rotational speed, 95, 97, 98, 110, 114, 120,
 123, 142, 158

S

Sampling theorem, 20
 Scale function, 53, 54, 56
 Scaling coefficient, 191
 Scatter matrix, 131 133
 Second generation wavelet transform
 (SGWT), 205 210, 218
 Shannon entropy, 73, 74, 149, 168, 170 172,
 177, 178, 181, 183 185
 Shannon wavelet, 43 44, 180 182, 185
 Shift orthogonality, 72
 Short time Fourier transform (STFT), 12,
 21 28
 Signal transformation, 103 109
 Sinusoidal function, 45
 Soft thresholding, 66, 67
 Spalling, 9, 78, 83, 103, 109
 Spectral post processing, 103, 109 113
 Spindle, 9, 13, 93, 100, 130
 Splitting, 206
 Stamping, 5 7, 61
 Stationary signal, 4, 5, 11, 12
 Sub band, 125 130, 136, 137, 139 141,
 143 146
 Superposition property, 35 36
 Surface, 9, 63, 78, 83, 95, 103, 109, 114,
 158, 167 169, 209, 210, 212, 213,
 217, 218
 Symlet wavelet, 63, 64
 Symmetry, 61 63, 166, 167
 Synthetic signal, 95 97, 156

T

Template function, 17 19, 28, 104 107, 109, 112, 113
 Time frequency analysis, 78 79, 154
 Time frequency resolution, 12, 23, 24, 69, 169
 Time scale frequency analysis, 103, 113, 114, 120, 124
 Transient signal, 2 3, 5, 12, 22, 103, 110, 113, 190
 Translation, 12, 28, 29, 34 36, 49 52, 54, 58, 68, 75, 76, 205, 212, 214
 Translation invariance, 53
 Transmitter, 89, 90 94

U

Ultrasonic pulse, 87 93, 100
 Ultrasound pulse train, 88 90, 92, 94
 Uncertainty principle, 12, 23
 Unified technique, 103 124
 Updating, 135, 136, 206 208
 Upsampling, 196

W

Wave, 33, 34, 87, 108, 155
 Waveform, 2, 3, 5, 27, 45 47, 62, 63, 71, 77, 96, 106, 109, 149, 155, 158, 165, 183, 192
 Wavelet
 design, 189 190
 frame, 52, 53
 function, 12, 29, 33, 34, 39, 41, 44, 52, 55 57, 69, 72, 107, 110, 112, 168, 205, 207, 208, 212, 218
 packet coefficient, 76, 126, 127, 143, 146, 151, 152, 154
 packet transform, 69 80, 125 146, 149, 151, 154, 162, 166, 207
 transform, 12, 17 31, 33 47, 49 69, 74 76, 83, 86, 87, 97, 103, 105, 107 114, 116, 118, 124, 125, 165, 166, 171, 172, 176, 179, 180, 184, 185, 190, 198, 205 210, 217, 218
 Window function, 12, 21 23, 27, 215, 216
 Window size, 23 25, 28