Wavelet-Based Techniques for Classification of Power Quality Disturbances

by

Tuan Anh Hoang, B.E. (Hons.), IEEE Member
School of Engineering

Submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

University of Tasmania

Electrical Engineering

February 2003

Statement of Originality

This thesis contains no material which has been accepted for the award of any other degree or diploma in any tertiary institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person, except when due reference is made in the text.

Tuan Anh Hoang

Statement of Authority of Access

This thesis may be made available for loan and limited copying in accordance with the *Copyright Act 1968.*

Tuan Anh Hoang

To my dearly loved parents Hoang Minh Trien & Vu Thi Lien;

To my lovely wife Nguyen Hong Thanh Thao;

To my brother, sisters and my daughter Thu.

ABSTRACT

The quality of power supply has become an important issue for electricity utilities and their customers. In recent years there has been a rising incidence of damage attributed to the power quality supplied to the customers of electric utilities. Meanwhile, there has been a rapid increase in the already widespread use of electronic equipment and modern power electronic devices. These trends have both decreased the quality of power on the electric grid and increased the equipment's sensitivity to power quality disturbances.

In order to improve the quality of the power supply, identifying the type and source of troublesome disturbances is an essential task. Existing automatic disturbance classification methods have replaced the traditional visual inspection of the disturbance waveforms. However, they are not reliable because those methods rely on the classification capability of large neural networks operating on inputs derived by simply pre-processing the disturbance signals with discrete wavelet transforms [134,135,136,137,138]. Long and redundant feature vectors both take a long time to train the network and result in a reduced classification rate. In this thesis, we aim to develop an efficiency method that automatically classifies power quality disturbances by using wavelet transform techniques to generate short and nonredundant feature vector.

Because of the wide range of power quality disturbances and their characteristic waveforms, ranging from very simple stationary and deterministic harmonics to highly transient and stochastic waveforms, different and appropriate analysis techniques are needed to achieve the overall classification objective. It is well known that the traditional Fourier analysis is ideal for analysing steady state signal. Although it is very powerful, Fourier analysis does not have the temporal resolution needed to cope with sharp changes and discontinuities in signals.

Recent years have witnessed a proliferation in the applications of wavelet transforms to signal analysis in a wide variety of fields, from geo-physics to telecommunications to bio-medical engineering. This has occurred because wavelet analysis provides dual localisations in both the time and the frequency domains. Moreover, wavelet analysis allows the flexibility of choosing a wavelet that suits a particular application. Especially by using the simple and flexible lifting scheme, we can construct a time-variant or space-variant wavelet – known as second-generation wavelet. The second-generation wavelet analysis makes optimal use of the correlation between neighbouring signal samples and between neighbouring frequency components to construct '*local*' wavelets, which adapt to the local characteristics of the signal.

Common types of wavelet schemes are the orthonormal or biorthonormal wavelet transforms that are typically used in compression and coding applications. This is due to the fact that those schemes can be implemented with fast algorithms and they are non-redundant representations of a signal. Unfortunately, they suffer the limitation of not being translation invariant; a totally different set of transformed coefficients is obtained when the same signal is shifted. This is the major concern in pattern recognition applications.

There exist a number of wavelet schemes that have the shift invariance property in their multiresolution representations. In this thesis, the local maxima and the matching pursuit techniques are presented as the two most appropriate techniques for power quality solutions. This is because the two techniques can efficiently decompose a signal and have the ability to precisely measure power quality disturbance characteristics so that they represent the disturbances by a compact, time-invariance feature vector.

The final task of classification is the selection of an appropriate classifier for use with the feature vector. There are two main approaches of pattern recognition: one is parametric and the other is non-parametric [129]. Parametric approaches can be either deterministic or statistical. The statistical parametric approach requires a good assumption about the statistical distribution of the data. On the other hand, the non-parametric approach, known as the neural network approach, does not require any statistical assumption about the data. In our statistical approach, we use a two-layer network structure with locally tuned nodes in the hidden layer, known as Radial Basis Function (RBF) network [106,120,121]. The network has only a local learning capability and a limited learning inference from the training data, but trains quickly as the training of the two layers is decoupled.

In an RBF network, the crucial concern is the selection of cluster centres and their widths. However, current techniques give suboptimum positions of cluster centres and their widths, thus limiting the classification rate. To improve the performance of an RBF network, we propose to modify the structure of the RBF network by introducing the weight matrix to the input layer (in contrast to the direct connection of the input to the hidden layer of a conventional RBF) so that the training space in the RBF network is adaptively separated by the resultant decision boundaries and class regions. During training iterations, cluster centres, their widths and the input layer weights are optimally determined together and concurrently adjusted to maximise the discriminant between classes, thus minimising the classification error. In this way the network has the ability to deal with complicated problems, which have a high degree of interference in the training data, and achieves a higher classification rate over the current classifiers using RBF.

For the classification of different types of disturbances that may be present on a power supply, in this thesis we show that our automatic classification techniques achieve superior recognition rates over current techniques. This improvement is done in two steps. The first improvement is the extraction of disturbance features using appropriate signal processing tools from which we obtain an efficiency and translation invariant feature vector. The second improvement is the designing of an appropriate classifier which maximises the inter-class discriminant function.

ACKNOWLEDGMENTS

Firstly, I would like to express my deepest gratitude to my supervisor Professor D. Thong Nguyen for his valuable guidance, encouragement and support. His valuable advice and his experience have greatly benefited me academically as well as personally. His willingness and motivation have given me an assured mind and confidence in my research work. I would also like to express my appreciation for the research scholarship that he has helped me obtain from his ARC grant for the Power Quality studies project.

I am also grateful to my co-supervisor Associate Professor Michael Negnevitsky for his guidance and contribution during my research. His valuable advice and support have immensely advantaged me with study resources at the early stage of my research.

I gratefully acknowledge all the staff of the School of Engineering at the University of Tasmania, especially Mr. Peter Watt and Dr. Allan Beasley for their kindness and assistance. Many thanks go to the technical staff of the School for their help and support for my research facilities. I am also grateful to all my fellow Ph.D students for their companionship, particularly, Thanh Kieu and Martin Ringrose.

I would like to express my deep sense of gratitude to my parents, who have always provided great opportunities, encouragement and love to their children. I would like also to express my love and heartfelt appreciation to my wife for her love and patience. My love is also dedicated to my daughter, who is so wonderful and who brings us lots of happiness while we are far away from our homeland.

Finally, I like to extend my appreciation and thanks to the financial support from the School of Engineering at the University of Tasmania together with the local Energy Industry, in particular, Hydro-Electric Corporation Tasmania, Comalco Aluminium and Pasminco Metal. Thanks also given to all friends who have made my life in Tasmania enjoyable and memorable.

Tuan Anh Hoang

Hobart, 20th February 2003.
School of Engineering
University of Tasmania, TAS 7001, AUSTRALIA

PREFACE

Traditional Fourier transform has been used as a powerful tool for signal analysis and decomposition. Unfortunately, due to the lack of time localisation, Fourier analysis cannot deal with transient or non-stationary signals.

Unlike Fourier transform, wavelet transform has the advantage of optimal trade-off between time resolution and frequency resolution. It can thus provide both time and frequency localisation. Wavelet analysis is therefore well suited to the analysis of non-stationary signals.

Wavelet transform has been recognised as a suitable tool to use in power quality monitoring and disturbance recognition or classification since its first publication in 1996 [130]. In 1999, at the time this thesis was initiated, there are some research works found in literature in this field. These works however simply re-process the disturbance signal using wavelet transform, while retaining a large, redundant and translation variant feature vectors. This requires large neural network classifiers, making these methods inefficient and requiring a long training time.

The scope of this research is to find an efficient wavelet based technique for analysing and extracting the features of power quality disturbances, which can be used for a stability classification for power quality disturbances.

Thesis Organisation

This thesis is organised into nine chapters:

Chapter 1 provides an overview and introduction to power quality problems. It gives a short background on power system quality and discusses the need for automatic disturbance classification in power quality monitoring and analysis.

Chapter 2 presents the definitions, the causes and the effects of different types of power quality disturbances in the power supply. A method for generating the transient disturbances is introduced.

Chapter 3 covers two important types of local time-frequency decomposition, which are the short-time Fourier transform, and the wavelet transform. Various wavelet analysis schemes and their properties are also discussed in this chapter.

Chapter 4 presents the translation invariant wavelet transform modulus maxima representation. It studies the Lipschitz exponents of a signal that provide a measurement on the local regularities of the signal. For oscillation singularities, the general modulus maximum of each modulus maxima chain gives an estimation of the local frequencies of the signal. The Lipschitz exponents and the general modulus maxima are used to characterise the transient disturbances. This chapter presents some numerical demonstrations for transient disturbances using wavelet transform modulus maxima.

Chapter 5 presents a popular translation invariant representation called matching pursuit. It discusses fast implementation of the matching pursuit and introduces the orthogonal matching pursuit. Numerical demonstrations for the transient disturbances using matching pursuit are also presented.

Chapter 6 covers different pattern recognition approaches. It examines in detail the two commonly used approaches, parametric and non-parametric. A suitable pattern recognition approach for our problem is also discussed in this chapter.

Chapter 7 provides a modification of the radial basis function network by adding the feature weights or the input layer weight into the network. Two new training techniques, which are the knowledge base technique and the generalised technique, are proposed to train this network. Comparisons are made between the conventional radial basis function network, the backpropagation network and the modified radial basis function network trained by one of the two training techniques.

Chapter 8 presents an automatic classification method for classifying 10 power quality disturbance types. Comparisons for the four transient disturbance types are made between the classification method developed by the authors and other current classification techniques.

Chapter 9 gives a summary of the thesis contribution, major results obtained and suggestions for further research.

Supporting Publications

1. D.T. Nguyen and T.A. Hoang, "Detection of Disturbances on Electricity Supply Using Wavelets," *Proceedings of Australian Universities Power Engineering Conference and IEAust Electric Energy Conference AUPEC/EECON'99*, Darwin, Australia, pp. 231-235, 26-29 Sep. 1999.

2. T.A. Hoang and D.T. Nguyen, "Appropriate Processing for the Classification of Power Quality Disturbances," *Proceedings of Australian Universities Power Engineering Conference AUPEC 2000*, Brisbane, Australia, pp. 196-201, 24-27 Sep. 2000.

3. D.T. Nguyen and T.A. Hoang, "Analysis of Power Transient Disturbances using Wavelet Transform Modulus Maxima Technique," *Proceedings of Australian Universities Power Engineering Conference AUPEC 2000*, Brisbane, Australia, pp. 190-195, 24-27 Sep. 2000.

4. T.A. Hoang and D.T. Nguyen, "Wavelet-based Neural Network for the Classification of Power Quality Disturbances", *Proceedings of Artificial Intelligence in Science and Technology Conference AISAT 2000*, Tasmania, Australia, pp. 273-278, 17-20 Dec. 2000.

5. T.A. Hoang and D.T. Nguyen, "Classification of Power Quality Disturbances using Wavelets", under review by *IEEE Transaction on Power Delivery*, Oct. 2001.

6. M. Negnevitsky, M.J. Ringrose, J. Huang and T.A. Hoang, "Neuro-Fuzzy Classifier for Recognition of Power Quality Disturbances, Quality and Reliability of Supply", *Proceedings of 6th International Transmission and Distribution Conference and Exhibition*, Brisbane, Australia, 11-14 Nov. 2001.

7. T.A. Hoang and D.T. Nguyen, "Wavelet-based Classification of Power Quality Disturbances using Radial Basis Function Networks," *Proceedings 6th IASTED*

International Multi-Conference on Power and Energy Systems, Marina Del Rey, USA, pp. 282-287, 13-15 May 2002.

8. T.A. Hoang and D.T. Nguyen, "Matching Pursuit for the Recognition of Power Quality Disturbances," *Proceedings 33rd Annual IEEE Power Electronic Specialists Conference PESC'02*, Cairns, Australia, pp. 1791-1796, 23-27 Jun. 2002.

9. T.A. Hoang and D.T. Nguyen, "Training Radial Basis Function Networks for Wavelet-Based Classification of Power Quality Disturbances," *Proceedings 4th IASTED International Conference on Signal and Image Processing SIP 2002*, Hawaii, USA, Paper # 359-112, 12-14 Aug. 2002.

10. T.A. Hoang and D.T. Nguyen, "Improving the Training of Radial Basis Function Network for Classification of Power Quality Disturbances," *IEE Electronic Letters*, vol. 38, no. 17, pp. 976-977, Aug. 2002.

11. T.A. Hoang and D.T. Nguyen, "Optimal learning for patterns classification in RBF networks," *IEE Electronic Letters*, vol. 38, no. 20, pp. 1188 –1190, Sep. 2002.

12. D.T. Nguyen and T.A. Hoang, "Classification of Power Quality Disturbances Using Radial Basis Function Networks," *Proceedings of International Power Quality Conference IPQC 2002*, Singapore, 21-25 Oct. 2002.

13. T.A. Hoang and D.T. Nguyen, "A Comprehensive Training for Wavelet-based RBF Classifier for Power Quality Disturbances," to appear in *Proceedings of IEEE TENCON'02*, Beijing, China, 28-31 Oct. 2002.

CONTENTS

GLOSSARY

PARAMETER

$\phi(t)$	Scaling function
$\psi(t)$	Wavelet function
$\Psi(\omega)$	Fourier transform of wavelet function
η	Gaussian noise
ξ, ω	Frequency parameter
2^j	Resolution at scale, j
a_j	Approximation for discrete 1-D wavelets, at j^{th} scale
d_j	Detail for discrete 1-D wavelets, at j^{th} scale
$E(t, \omega)$	Time-frequency energy
$f(t)$	Continuous signal in spatial domain
$g(n)$	Discrete time domain highpass analysis filter
$\tilde{g}(n)$	Discrete time domain highpass synthesis filter
$h(n)$	Discrete time domain lowpass analysis filter
$\tilde{h}(n)$	Discrete time domain lowpass synthesis filter
$\bar{h}(n)$	Mirror image of $h(n)$
i	Square root of -1
I_d	Identity operator
j	j^{th} scale in the discrete wavelet transform, $j \in Z$
$\mathbf{L}^2(R^n)$	Space of measureble, square integrable n-dimensional functions
R^n	Euclidean space of n-dimentions
s	Scale in continuous wavelet transform
s_j	Scale in discrete wavelet transform, at j^{th} scale
t, τ	Signal parameter (temporal or spatial) in 1-D
V_j	A vector space that is the set of all approximations of a signal in $\mathbf{L}^2(R)$
Z	Set of integers in 1-D
W_j	Orthogonal complement to V_j

ABBREVIATIONS

AC	Alternating Current
CWT	Continuous Wavelet Transform
DC	Direct Current
DWT	Discrete Wavelet Transform
FIR	Finite Impulse Response (filters)
FT	Fourier Transform
HF	High-Frequency Capacitor Switching Disturbance
IT	Impulse Transient Disturbance
NT	Notch Disturbance
LF	Low-Frequency Capacitor Switching Disturbance
pu	per unit
RMS	Root Mean Square
STFT	Short Time Fourier Transform
WT	Wavelet Transform

MATHEMATICAL SYMBOLS

∂	Partial derivative
$\forall$	For all
$\exists$	There exists
Σ	The sum of
Π	The product of
$\lvert\ \rvert$	Absolute value of
$\lVert\ \rVert$	Magnitude of
$<\ >$	Inner product of
$x^*(t)$	Complex conjugate of $x(t)$
∞	Infinity
π	pi
μ	Mean value
σ	Standard deviation

$\lim_{j \to \infty}$	The limit as j tends to infinity
$\circ$	Convolution
$A \oplus B$	Append the vector space, A, to the vector, B
$\otimes$	Tensor product
$\subset$	Subset of
$\cup$	Union of
$\cap$	Intersection of
$\perp$	Orthogonal to

List of Figures

List of Tables

Chapter 1

AN OVERVIEW & INTRODUCTION TO POWER QUALITY PROBLEMS

1.1 WHAT IS POWER QUALITY?

The term ‘power quality’ applies to a wide variety of electromagnetic phenomena on power systems. Although it is difficult to define precisely, power quality (PQ) fundamentally describes the consistency of the voltage and current waveforms on a power system or within an electrical power customer facility. It is perhaps more relevant to discuss the commonly used term a ‘*PQ problem*’. A PQ problem can be categorised as a disturbance caused by a piece of equipment, a combination of pieces of equipment, or a network configuration that impacts on the voltage and/or current waveforms in such a way that adversely affects the operation of other equipment on the system. Optimally, the voltage waveform at all points on a power system should be a perfect sinusoid at nominal voltage and frequency. This is impossible to achieve in practice due to inconsistent loads and the dynamic nature of power system operation, so limits are defined for acceptable deviations away from the optimum. Different electrical equipment or systems have different abilities to cope with the deviations of the power supply. Nowadays much electronic equipment is very sensitive to changes on the power supply and can easily mis-operate if the voltage exceeds these limits.

PQ problems have been evident for many years, but it is only in the last decade that they have become a major concern to electric power utilities and customers. This is mainly due to the increasing use of microprocessor-based electronics, power electronic devices and power factor correction capacitor banks. The increasing use of electronic and power electronic devices presents a two-fold problem with regard to PQ. These devices are typically more sensitive to PQ disturbances and also produce more PQ problems than the electromechanical power system devices used in the past. Capacitor banks reduce reactive power flow, decreasing losses, but can also trigger sensitivities to a range of PQ disturbances, especially harmonics. They can also produce transient disturbances when switching of capacitor banks is implemented.

PQ problems represent a huge cost to business in lost productivity and equipment damage. It is not uncommon for a momentary utility breaker operation to result in a loss of $10,000 to an average sized industry by shutting down its production line, which will require many hours to restart. Large-scale industries can lose millions of dollars per hour in lost production if one machine or process in a production line fails due to a PQ problem. Utilities and customers are spending increasing amounts of money monitoring, studying and improving PQ in order to minimise the economic impact of these problems. Customers are becoming more aware of PQ issues and the problems poor PQ can cause. Of major concern in the study of PQ, is locating the source of PQ disturbance on a power system. Generally, the source must be located before a solution can be formulated.

1.2 CLASSIFICATION OF POWER QUALITY DISTURBANCES

The current practice is to perform manual studies into possible causes of the problem [1,130], which is at best, a highly inefficient and costly task. In many cases it is a relatively simple task for a human to categorise a disturbance based on a visual inspection of a recorded waveform. However, as is the case with many visual tasks that humans find simple, automatic computer classification has proved to be much more difficult. Although the need for automatic determination of disturbance source location has been apparent for some time, there are only a small number of solutions towards this goal that can be found in the literature. It is obvious that because each power system is different, the system response to a PQ disturbance will be different, making it a very complex task to develop a generalised PQ source locator. Also power system parameters change constantly with the load and generation schemes.

At the University of Tasmania in Australia, a collaborative effort between its School of Engineering and the local energy industry has been pursued to develop an accurate and robust PQ monitoring and disturbance recognition system. Our technique, in this collaborative project, uses wavelet analysis for the detection and feature extraction of power quality disturbances, and neural pattern classification for the identification of their origin.

There are a number of methods for detecting particular types of power system disturbances, in particular for fault detection [2,3,4] and transformer inrush [5], but these are limited in their application. Using an individual artificial neural network, the authors of [6] attempt to automatically classify PQ disturbances in the time domain. This work has been further refined by pre-processing the signals using the discrete wavelet transform (DWT) in a time-scale domain [134,135,137]. However these techniques use the entire set of DWT coefficients, which is very large and is lack of time shift-invariance. Therefore, it takes a long time to train and does not guarantee a convergence of the neural network. Moreover, the methods cannot produce a good classification rate, as the DWT coefficient at any point in the time-scale domain is completely unpredictable and results in a substantial overlap in range between different types of disturbance. Various wavelet kernels have been investigated in [7,132].

This thesis proposes a number of new wavelet-based techniques for automatically identifying the type and source of PQ disturbances, and shows that the wavelet-based techniques that we propose produce compact and time-invariant feature vectors. These two characteristics render efficiency and stability to the training and convergence of classifiers of PQ disturbances.

1.3 THESIS CONTRIBUTIONS

A general block diagram for PQ disturbance classification is shown in Figure 1. The first step in classification is to understand the sources and characteristics of the signal that we are dealing with. So in Chapter 2, we take a close look at all technical aspects of power quality and disturbances. In that chapter, the definition of each type of disturbance and its characteristics are presented. Depending on the natural characteristics and the phenomena of disturbances, they are classified into two main groups: transient and steady states. Since different disturbance types may have different specific characteristics, we intend to use different signal processing tools to analyse different types.

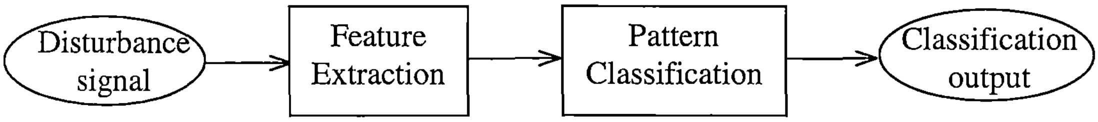


Figure 1.1: A general block diagram for PQ disturbance classification.

For the next step of classification, different signal processing techniques are used to extract signal features for further processing. The signal is analysed in the time and frequency domains, so that it can be reduced to a number of components, the so-called *time-frequency atoms*, that are localised either in time, in frequency or in both.

The number of atoms and the efficiency of the method are greatly dependent on the correlation between the signal and the analysis kernel. In this thesis, the author proposes some new feature extracting techniques that use appropriate signal processing tools such as Fourier transform (FT), *wavelet transform modulus maxima* (WTMM) and the *matching pursuit* for power quality disturbances classification.

In Chapter 3 an introductory analysis of several time-frequency techniques is presented. The properties of signals are revealed by transforms that decompose signals into elementary functions that are well concentrated in time and frequency. We present two important types of local time-frequency decomposition, which are the short-time Fourier transform (STFT) and wavelet transform (WT). The localised nature of time-frequency decomposition is important. However it is limited by the *Heisenberg uncertainty principle.*

The issue of translation invariance is significant in pattern classification. A DWT or wavelet frame is not adequate to deal with this issue. Chapter 4 presents the WTMM technique that maintains the translation invariance by only sampling the scale parameter and leaving the time parameter in a continuous manner. Moreover, singularities and irregular structures often carry essential information in a signal, and they can be detected by following the WTMM coefficients at fine scales. So at the end, the signal features that we obtain from the WTMM are much more compact than the entire set of WT coefficients and they are also time-invariant.

In STFTs, the window of the time-frequency atoms has a constant size. Therefore, STFTs are not well adapted to signal structures that are much smaller or much larger in window size. On the other hand, a wavelet transform is built by relating the frequency to the scale (window size). The resulting family of waveforms are dilated and translated versions of a single mother wavelet. This has a limitation on the estimation of frequencies that are well localized in the Fourier transform domain, especially at high frequencies. In general, adaptive signal decomposition involves the expansion of a signal over a set of waveforms, which are selected appropriately from a large and redundant dictionary. Chapter 5 presents a general algorithm called *matching pursuit* that performs such an adaptive decomposition.

The most important step in a classification task is the selection of feature vector and the designing of a correspondent distance metric for pattern matching. Chapter 6 discusses the two main approaches of pattern recognition: one is parametric and the other is non-parametric [129]. Parametric approaches can be either deterministic or statistical. The statistical parametric approach requires a good assumption of the underlying distribution of the data. On the other hand, the non-parametric approach, known as the neural network approach, does not require any statistical assumption of the data. In PQ classification problems, most researchers [134,135,137] simply pre-process the signal but still retain a large and redundant amount of data, then rely on the capacity of a neural network for the classification. These methods are therefore inefficient and may not produce a high classification rate. In this thesis, we propose methods that extensively study the signal characteristics via a number of signal processing tools. Then a small number of signal features that have clear statistical distribution between different types of disturbance are selected. With this small number of selected signal features, a small classifier network is required by either using the parametric or non-parametric approach.

There are a variety of Artificial Neural Networks (ANNs) with different structures. Each network has its own advantages and disadvantages. Multilayer perceptrons (MLPs) including the backpropagation neural networks, are typical of *globally generalising* networks which have the capability of robust learning inference and generalisation from the training data. These networks, however, are very slow in learning and suffer from the possibility of being trapped in local maxima of the chosen optimisation cost function. Existence of optimisation techniques such as genetic algorithm [8], learning automata [9], and proper initialisation of connection weights [10] are capable to achieve a global minimum, they require extensive computation and also in many applications there is inadequate prior knowledge about the training examples to allow a good estimate for the initialised setting. In our statistical approach, we use a two-layer network structure with locally tuned nodes in the hidden layer, known as Radial Basis Function (RBF) network [120]. The network, therefore, has only a local learning capability and a limited learning inference from the training data, but is much faster for training because the training of the two layers is decoupled. When using an MLP network the training is iteratively coupled together. Furthermore, the locally tuned or self-organised ability of the hidden layer in an RBF network is equivalent to a very efficient initialisation of the connections, thus giving the network the ability to avoid local minima.

In the classification of power quality disturbances, the prior knowledge of the classes does in fact allow us to effectively initialise the RBF network. Disturbance types have their attributes widely spread over well-known time and frequency ranges. Although these ranges overlap, their cluster centres are well separated. This type of training data should favour the use of networks with self-organising capability in the first hidden layer since they are being free from local maxima and are much faster to train. This is our motivation for the use of an RBF for classification of transient power quality disturbances.

In Chapter 7, we propose two new techniques to improve the performance of an RBF network. In many applications, some features of a data pattern are more important or more discriminating than other features, e.g. the formant frequencies of a voiced sound, the dominant components in a principal component analysis. The pattern matching gives more weight to these components in the feature vector. Therefore, to increase the discriminant between classes, a feature weight vector is introduced to the distance measurement. We then carry out a comprehensive *knowledge-base* training algorithm for the RBF classifier so that at its convergence the network gives both the optimal feature weight vector as well as the cluster centres and their scaling width. Moreover, in most cases the importance level of a given feature is different in different classes and the general form of the feature weight should be a matrix. Also in some cases it is difficult to construct a knowledge base precisely. Therefore, we take a further step to modify the structure of the RBF network by introducing a weight matrix to the input layer in contrast to the direct connection of the input to the hidden layer of a conventional RBF. We then train this weight matrix as a single layer perceptron together with the *clustering* training. This still retains the speed advantage of an RBF network over an MLP, while archiving a higher classification rate.

Chapter 8 presents the proposed techniques for the classification of power quality disturbances, and the classification results of these techniques. Depending on the signal characteristics, appropriate signal processing tools are proposed to extract different discriminating features from the disturbance signal, hence enhancing the classification results. We also present some comparative results between our techniques and current techniques.

The last chapter, Chapter 9, provides a conclusion to the work contributed by this thesis, as well as suggestions for future studies.

Chapter 2

POWER QUALITY DISTURBANCES

2.1 INTRODUCTION

In recent years, power quality has become an important concern for utility, facility and consulting engineers [10,65,135,134]. End user equipment is more sensitive to disturbances that arise both on the supplying power system and within customer facilities. Also, this equipment is more interconnected in the network and industrial process, therefore the causes of disturbances on the system are much more severe. A recent E SOURCE survey claimed more than 50% of larger users are significantly affected by power quality on their company's overall performance. In high-tech industry, a single power outage can easily cause the company losses of US$1 million or more, while the average outage that facilities experience is 3.5 outages per year.

It is important to understand the phenomena of power quality variation that causes problems with sensitive loads. Categories of these variations must be developed with a consistent set of definitions so that measuring equipment and analysis tools can be designed to handle different types of disturbance. Also the first and the most important step in identifying the source of a disturbance is to correlate the disturbance waveform with the possible cause, i.e. recognising the category of the disturbance (e.g. load switching, capacitor switching, lightning, remote fault condition, ect.). This requires a full understanding of the characteristics of each disturbance category. Once the category for the cause has been determined, the identification becomes much more straightforward.

This chapter describes the characteristics of different types of PQ disturbances and presents a method for generating PQ disturbances.

2.2 GENERAL CLASSES OF PQ PROBLEMS

Power quality disturbances usually result in voltage or current waveform being deviated from the normal level. However it is essential to maintain the voltage waveform within a certain limit so that it can retain the quality of power supply. Moreover, the power supply system can only control the quality of the voltage. It has no control over the currents that particular loads may draw. Therefore, the limits defining normal operations in power systems are generally given in terms of voltage [11,13].

Several international standards have been proposed for PQ problems, such as the Institute of Electrical and Electronic Engineers Standards Coordinating Committee 22 (IEEE SC22) or the International Electrotechnical Commission (IEC) standard or the American National Standards Institute (ANSI).

Basically, PQ disturbance can be divided into seven main categories in terms of their deviation of the voltage waveform. They are:

1. Transients
2. RMS voltage variation
3. Voltage imbalance

4. Waveform distortion
5. Voltage fluctuation
6. Power frequency variation

As in [11], based on the source, duration and severity of disturbances on the power system, disturbances in these main categories can be further sub-grouped as shown in Table 2.1. In this table, the three most important disturbance attributes, which are spectral content, duration, and voltage magnitude, are presented for each type.

2.3 SOURCES AND DEFINITIONS

2.3.1 Transients

According to the IEC standard, a transient is defined as: pertaining to or designating a phenomenon or a quantity which varies between two consecutive steady states during a short time compared with the time-scale of interest.

The term *transients* has long been used in the analysis of power system variations to denote an event that is undesirable but momentary in nature. In power systems, when we encounter the word transient we probably think of a damped oscillatory transient due to a RLC network. A transient can also result from a lightning strike for which a surge arrester is used for protection. Transients can be classified into two categories, *impulsive* and *oscillatory*. These terms reflect the waveshape of a current or voltage transient.

Impulsive transient:

An impulse transient is a sudden non-power frequency change in the steady state condition of voltage or current that is unidirectional in polarity. To characterise impulsive transients, three parameters are normally used, that is their amplitude and the rise and decay times. We can also reveal impulse transients by their spectral content. Lightning is the most common cause of impulse transients. Figure 2.1 shows an example of an impulsive transient caused by lightning.

Categories	Typical spectral content	Typical duration	Typical voltage magnitude
1. Transients			
1.1 Impulsive	5ns - 0.1ms rise	50ns - 10ms	
1.2 Oscillatory	400Hz - 5MHz	5μs - 50ms	0 - 8pu
2. RMS voltage variation			
2.1 Interruption		0.5cycles - 1min	< 0.1pu
2.2 Sag (dip)		0.5cycles - 1min	0.1 - 0.9pu
2.3 Swell		0.5cycles - 1min	1.1 - 1.2pu
2.4 Interruption, sustained		> 1min	0.0 u
2.5 Under-voltages		> 1min	0.8 - 0.9pu
2.6 Over-voltages		> 1min	1.1 - 1.2pu
3. Voltage unbalance		Steady state	0.5 - 2%
4. Waveform distortion			
4.1 dc offset		Steady state	0 - 0.1%
4.2 Harmonics	0 - 100th harmonics	Steady state	0 - 20%
4.3 Interharmonics	0 - 6kHz	Steady state	0 - 2%
4.4 Notching		Steady state	
4.5 Noise	Broadband	Steady state	0 - 1%
5. Voltage fluctuation	< 25Hz	Intermittent	0.1 - 7%
6. Power frequency variations		< 10s	

Table 2.1: Categories and Characteristics of Power System Electromagnetic Phenomena

Due to the sudden change of impulse transient, it can be responded to by many circuit components, and produce oscillation transients which have significantly different characteristics when viewed from different parts of the power system [11].

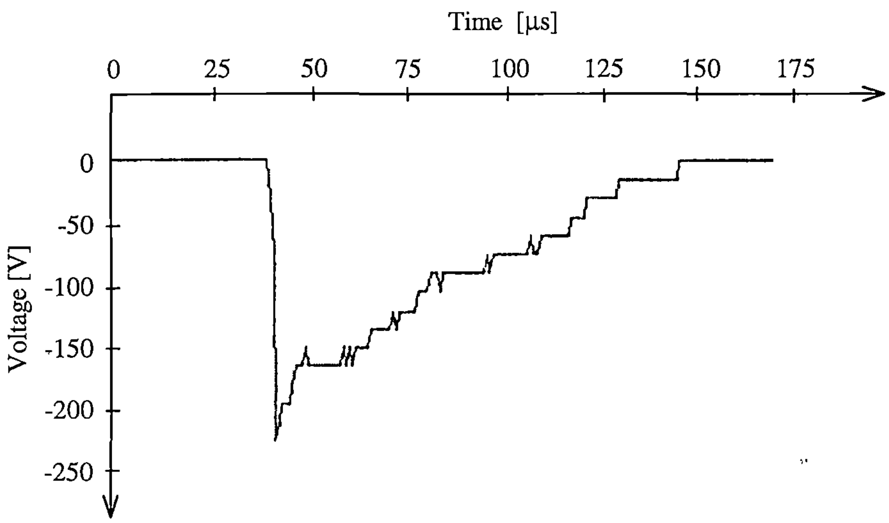


Figure 2.1: An impulsive transient caused by lightning

A common problem caused by impulse transients is divides damage. Often the over-voltage in impulse transients can be high enough to damage inadequately protected equipment. Impulses can cause electronic equipment to malfunction or be temporary offline.

Oscillatory transient:

An oscillatory transient is a sudden, non-power frequency change in the steady state condition of voltage or current that includes both positive or negative polarity value. An oscillatory transient normally takes the form of a damped sinusoid. It is described by its spectral content (predominate frequency), duration and magnitude.

This category of disturbance is frequently come across in utility subtransmission and distribution systems. Capacitor switching mainly causes oscillatory transient

[12,13]. Also many other types of events can cause this type of disturbance, namely transformer switching operation, ferro-resonant, as well as the response of circuits on the power system to an impulsive transient. An example of low frequency capacitor switching is shown in Figure 2.2.

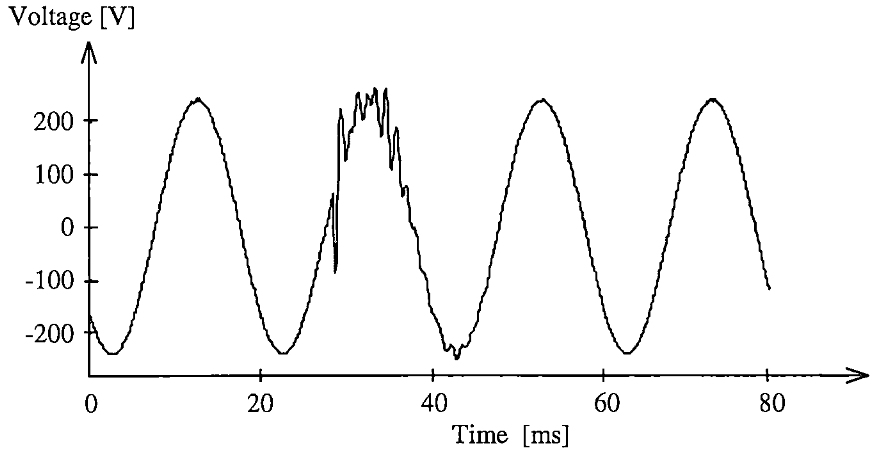


Figure 2.2: Oscillatory transient caused by a capacitor switching

Major causes of oscillation transients are variable-speed drive trip-outs and electronic equipment malfunctions. Large magnitude oscillation transients can also cause damage to unprotected equipment and as its large potential energy is contained, arrestors require a high quality to survive.

2.3.2 RMS Voltage Variations

A RMS voltage variation is a variation of the RMS value of the voltage from nominal voltage for a time greater than one-half a cycle of the power frequency. It can be a reason of fault conditions, irregular loose connections in power wiring, or switching of large loads that require high starting current, as well as variation in load on the system.

Depending on the system condition and the location of the fault, it can cause a short duration (less than one minute) of either voltage drop (*sag*), or voltage rise (*swell*), or a complete loss of voltage (*interruption*). Long duration variations (greater than one minute) such as *over-voltages* or *under-voltages* are generally not due to system faults, but are caused by system switching operations and load variations [10,11].

Sags:

The term voltage sag has been used for many years in the power quality community. It describes a short duration voltage decrease whose RMS value varies between 0.1 and 0.9 pu for durations from 0.5 cycles to one minute.

Sags are commonly associated with system faults. However heavy loads switching can also cause voltage sags. They are described by their RMS value and duration. The voltage drop in sags can cause sensitive electronic equipment to drop out or to malfunction. An example of voltage sag that is associated with a single-line-to-ground (SLG) fault is shown in Figure 2.3.

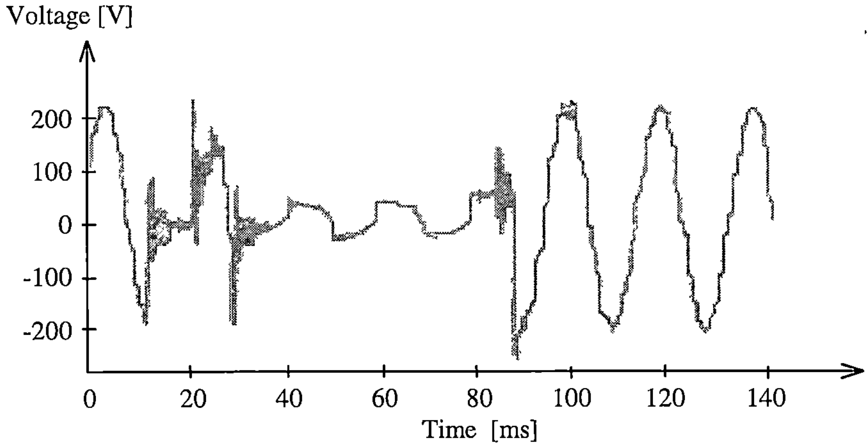


Figure 2.3: A voltage sag caused by a SLG fault (from [11]).

Swells:

A voltage swell is a temporary increase in the RMS value of the voltage of between 1.1 to 1.8 pu (typical between 1.1 to 1.2 pu), at the power frequency, for durations from 0.5 cycles to one minute.

As in sags, the main causes of swells are system faults, but they not as common as sags. During a SLG fault, the voltage on the un-faulted phases can temporary rise and create swell. The over-voltage in swells can damage electronic equipment if severe enough, or cause sensitive equipment to drop out or malfunction.

Interruption:

Interruptions are a type of short duration variation, where there is a complete loss of voltage (< 0.1 pu) on one or more phase conductors for a time period from 0.5 cycles to one minute.

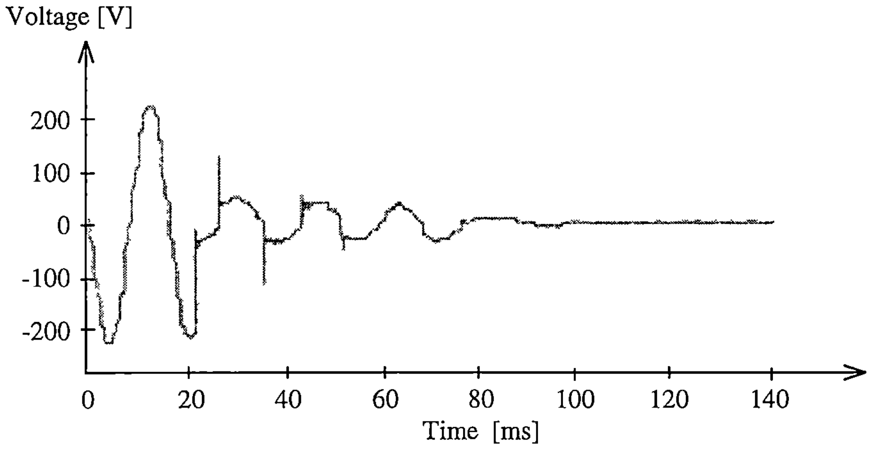


Figure 2.4: A momentary interruption caused by a fault and subsequent recloser operations (from [11]).

An interruption can be the result of power system faults or equipment failures. It is characterised by only the duration since it is associated with a total lost of voltage. The duration of an interruption is normally determined by the time taken from the instant protective equipment recloses to the time that the fault is cleared. Long interruptions are likely to cause most electronic devices to shut down [11]. Figure 2.4 shows an interruption caused by fault, followed by several protective reclosers operating until the fault is finally cleared.

Under-voltage:

Under-voltage is used to describe a specific type of long duration variation, which has a voltage value of between 0.1 and 0.9 pu (typical between 0.8 to 0.9 pu) and lasts for more than one minute. It is typical a result of loads switching on, or a capacitors switching off. Overloaded circuits, or incorrect tap settings on transformers, can also cause under-voltage.

Under-voltage is normally associated with cause other than system fault, and can be controlled by voltage regulation equipment.

Over-voltage:

Over-voltage is used to describe a specific type of long duration variation, which has a voltage value of greater than 1.1 pu and lasts for more than one minute. It is mainly caused by load reductions, capacitor switching on, or incorrect tap setting on transformers.

Under and over-voltage may cause equipment to drop out, malfunction or be damaged. As with all equipment, the abnormal voltage levels produce excess heating, shortening their lifespan and reducing their efficiency.

Sustained interruptions:

Sustained interruptions are a type of long duration variation, which has a complete loss of voltage on one or more phase conductors for a time greater than one minute.

The distinction between an interruption and sustained interruption is that the latter are longer than one minute and are often permanent. They normally cause mechanical devices to shut down, in some cases with costly consequences. This type of PQ problem requires human intervention to repair before power is restored.

2.3.3 Voltage Imbalance

Voltage imbalance (unbalance) is defined as the difference in the voltages amplitude of three phases in a three-phase system. It can be measured using symmetrical components. The ratio of either the negative or zero-sequence component to the positive sequence component can be used to specify the percent unbalance [11].

Single-phase loads on a three-phase circuit can cause the voltage to unbalance by less than two percent. Server voltage imbalance can be caused by single-phasing conditions.

2.3.4 Waveform Distortion

Waveform distortion is a steady state deviation from the perfect sine wave of power frequency. Basically, the characteristics of distortion can be reviewed from the spectral content. Waveform distortion contains five primary categories: DC offset, harmonics, inter-harmonics, notching and noise.

DC offset:

DC offset is defined by the presence of a DC component in voltage or a current in an AC power system. DC current is normally due to the effect of half-wave rectification. It causes biasing in transformer cores which results in excessive heating and loss of transformer life.

Harmonics:

Harmonic distortions are the presence of sinusoidal voltages or currents whose frequencies are integer multiples of the fundamental power frequency (power system operates at fundamental frequency, i.e. 50 or 60Hz) in the power signal. It is caused by non-linear loads on the power system. Figure 2.5 shows a harmonic distortion of a voltage waveform.

To specify the harmonic distortion levels, the complete harmonic spectrum is reviewed with magnitudes and phase angles of each harmonic component. The total harmonic distortion is also commonly used to determine the effective value of the harmonics distortion level.

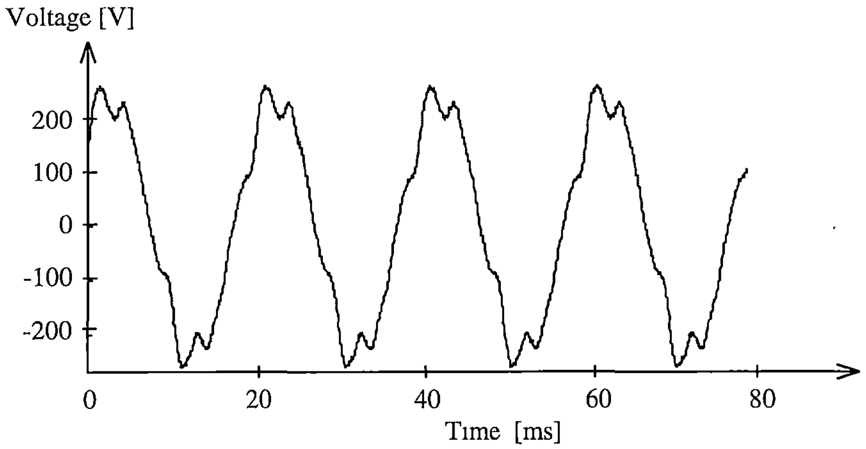


Figure 2.5: A harmonic voltage waveform

The harmonic currents can cause excess heating in supply transformers and capacitor banks, interference with nearby telecommunication lines, and harmonic voltage distortions [12]. Harmonic voltage distortions cause electronic equipment to overheat, malfunction or drop out. The equipment lifespan is shortened due to the additional heating and the sub-optimal efficiency condition.

Inter-harmonics:

Inter-harmonics are periodic waveform distortions of voltage or current whose frequency components are not at integer multiples of the fundamental power signal. They can appear as a discrete frequency or as a wide band spectrum.

The main sources of inter-harmonics are induction motors, static frequency converter, cycloconverter and arcing devices as well as DC transmission links across separate power systems.

Notching:

Notching is a periodic voltage distortion caused by commutation between different phases in power electronic devices. An example of notching that caused is by a three-phase converter is shown in Figure 2.6. During the commutation between two phases, the voltage is pulled as close as zero depending on the system impedances.

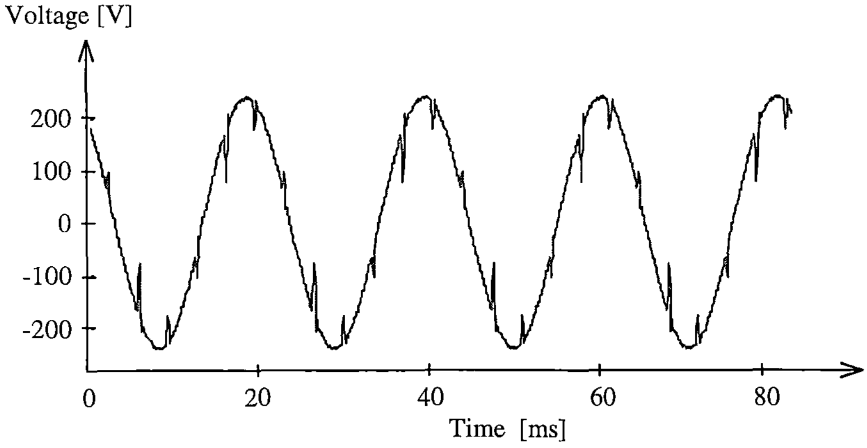


Figure 2.6: Notching caused by a three-phase converter

Notching can be characterised by harmonic spectrum as it occurred in a periodic manner. However, notching can produce very high frequency components that may not be measured by most harmonic measurement devices [11].

The notches can sometime produce the voltage waveform sufficiently close to zero and cause errors in instruments and systems that rely on zero crossing.

Noise:

Noise is defined as a broadband distortion signal with spectral content less than 200 kHz. Basically, any unwanted distortion of the power signal that is not classified as harmonics or transients distortion is referred as noise.

Many power electronic devices are sources of noises. They are control circuits, arcing equipment, load with solid-state rectifiers and switching power supplies. Electronics devices such as microprocessors and programmable controllers are most affected by noise.

2.3.5 Voltage Fluctuation

Voltage fluctuations are distortion of voltage envelopes in a continuous manner. The variations of voltage envelope are in the range of 0.9 to 1.1 pu and produce the phenomenon known as *flicker*. Flickers are the effects of the voltage variations on lamps as perceived by the human eye (i.e. fluctuation frequencies < 25 Hz). Figure 2.7 shows an example of voltage flicker.

As the terms voltage fluctuation and voltage flicker are interchangable in most standards, we will use the two terms with no distinction. One of the most common causes of voltage fluctuation on power system is arc furnace.

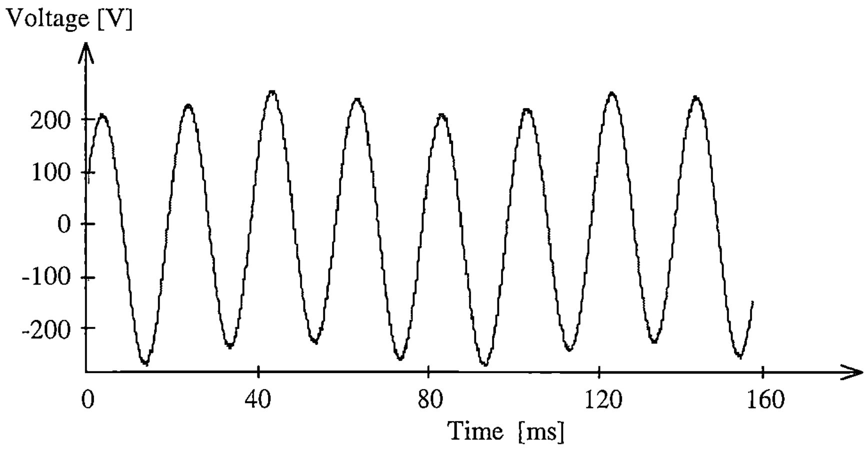


Figure 2.7: Voltage flicker caused by arc furnace operation

2.3.6 Power Frequency Variations

Power frequency variations are a deviation of the power system fundamental frequency from its nominal value (e.g. 50 or 60Hz). Since it is directly related to the rotation speed of the generator supplying the system, there are very small variations in frequency if a change exists in the dynamic balance between load and generation.

The change of frequency in power system is limited by the large net inertia of the system generators and generation control systems. However, significant frequency changes can be found if there are faults on the bulk power transmission system, or a large group of load is disconnected, or a large source of generation goes offline [11].

2.4 POWER QUALITY STUDY

Studies show that up to 80 percent of most small business' PQ problems are caused by disturbances created inside a facility or business (see Figure 2.8). For example, in a large power using building, fans, air conditioning equipment and other large applications cycle on and off. They can cause power dips, surges and transients that affect other equipment in the building. Lightning is another major source of disturbance that accounts for around 15 percent of power disturbances.

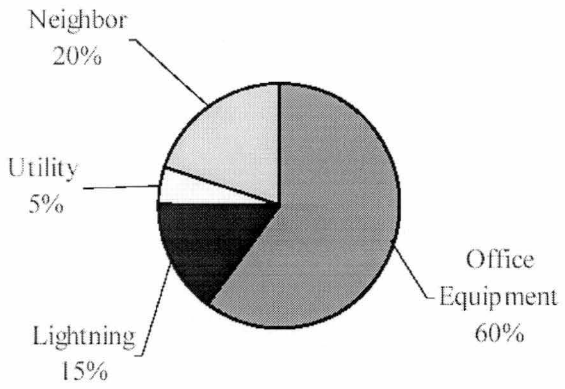


Figure 2.8: Source of power quality disturbances (Florida power study 1993).

To improve system stability and reliability, we need an efficient and prompt detection, classification and characterisation of PQ disturbance events and their sources, so that further identification of the location of these events can be made for maintenance and control of the system. Another aspect of PQ study in the recognition and characterisation of disturbance events is to coordinate these events with equipment performance. It is desired that the response of the sensitive equipment during the each event be explained and correlated to specific features of the event, so that the equipment operating characteristics can be turned for improved ride-though ability or immunity of the equipment to specific events [14].

To alleviate some of the problems that power line disturbances create, and to improve the performance of sensitive equipment, there are a number of protective devices available. Each of these devices has a different problem-solving function and is listed as below (1100-1992 IEEE Standard).

(i) ***Transient Voltage Surge Suppressors:***

This is an electronic device that attenuates noise or high-speed voltage transients caused from equipment switching, lightning strikes or faults. They are installed between the power source and sensitive equipment and are most effective if installed within close proximity to the piece of equipment being protected. Transient Voltage Surge Suppressors cut noise and voltage transients only and do not regulate voltage to limit surges and sags.

The two major types of transient voltage surge suppressors are filters and transient diverters. Filters serve as a block to high frequency current which is often noise, while letting the low frequency power current to pass through unaffected. Transient diverters offer a very low impedance path to ground whenever the voltage across the device exceeds a certain value, and thus reduces the voltage that could otherwise be presented to the sensitive equipment.

(ii) ***Shielded Isolation Transformers:***

This is a transformer where the primary winding is isolated from the secondary winding by an electrostatic shield. The shield reduces the passage of common-mode (line-to-ground) noise or transients, but is limited in rejecting normal-mode (line-to-line) noise, and does not regulate voltage nor protect against sags and surges. Those transformers with multiple shields are most effective and provide good protection.

(iii) ***Voltage Regulators:***

This device is designed to control the incoming voltage in order to sustain a constant output voltage. This is performed to protect sensitive equipment against over-voltage or under-voltage. Voltage regulators can give steady long-term voltage levels for varying inputs but do not protect against spikes nor attenuate noise.

(iv) ***Power Line Filters:***

This device is used to reduce voltage waveform distortion affecting sensitive equipment. In most cases, filters are used to screen out high-frequency noise, i.e. harmonic filters including line reactors, tuning reactors and capacitors.

(v) ***Power Line Conditioners:***

This piece of equipment combines the functions of a voltage transient suppressor, isolation transformer and a voltage regulator into one operational unit. A Power Line Conditioner will provide protection from all except the most severe or lengthy disturbance. These devices offer characteristics of two or more protection devices.

(vi) ***Standby Power Source:***

Under normal operation, the Standby Power Source provides for transient voltage and a degree of noise suppression. In the event of a brownout or blackout, the Standby Power Source rapidly transfers to an inverter to supply power to the load.

(vii) ***Motor Generator Sets:***

Motor Generator Sets consist of an AC motor coupled to a generator. The utility line energizes the motor that drives the generator. Since Motor Generators isolate the incoming power source from the load, they provide protection against noise and transients on the incoming power supply. Motor Generators protect against transients, but not against blackouts. However, if the generator is equipped with a heavy flywheel, it may ride through some momentary outages. This ride-through time is determined by the load and actual design of the unit. The Motor Generator is a relatively expensive device.

(viii) ***Uninterruptible Power Supply:***

For small equipment, this device provides full protection from all power disturbances. The Uninterruptible Power Supply provides an alternate power source in the event of utility power interruptions or failure and supplies power from a few minutes to several hours. The Uninterruptible Power Supply comes in two different modes: on-line and standby. A standard on-line or true Uninterruptible Power Supply consists of a rectifier or charger, battery and inverter. The rectifier or charger converts the incoming AC power to DC. The DC power charges the battery, which then runs the inverter section. The inverter then reshapes the DC power to AC, and sends it to the critical load. When the utility AC power is interrupted or fails, the batteries will supply DC voltage to the DC buss without a switching interruption which is then converted to AC voltage through the inverter to the load. The standby Uninterruptible Power Supply has a transfer switch that is programmed to select either the normal utility supply or transfer to battery/inverter should there be a power interruption that falls outside the operating limits.

2.5 MODELLING FOR POWER QUALITY

There are many different methods for analysing and simulating PQ disturbances, and the type and complexity of the model depends on the class of disturbance being analysed. Due to the fact that disturbance on the power system encompasses a wide range of variation and characteristics of disturbances on the power system, two main analysing techniques are normally used for different disturbance types. The first one is the time domain analysis, which is commonly used to analyse transient disturbances (e.g. transients and notches). The second one is the steady state analysis that is used for most other disturbance which are much less complex.

In this thesis, we emphasise using the wavelet-based technique to classify transient disturbances. For steady state disturbance types, we introduce a simple steady state analysing technique that uses Fourier transform for the classification. This technique is demonstrated later in Chapter 8.

The four common transient disturbance types that we include in our classification are impulsive transient (IT), high frequency capacitive switching transient (HF), low frequency capacitive switching transient (LF), and aperiodic notch (NT). The attributes of these types of disturbances are well documented in [11] and are presented above. However, in real-world applications, the limitation of recording equipment such as insufficient sampling frequency and limited storage space, prevents complete information about disturbances to be recorded or even misses out the disturbances altogether. Therefore, for the four classes of transient we select, only the three most important attributes are considered in our work – amplitude range, frequency range including dominant frequencies, and duration.

Since measured data of power quality disturbances is notoriously scarce because of its commercial implication, in this thesis data is obtained from a limited number of measured disturbances and simulation results based on the characteristics of disturbances (see Table 2.1). These disturbances are used for training and testing the performance of our recognition techniques throughout this thesis.

A disturbance can travel to the monitoring point through different paths in the network and may be the subject to different frequency drifts. We then simulate an oscillatory transient that is due to impulse or due to capacitive switching as a sum of P uncorrelated dominant frequencies plus noise [91,140], i.e.

$$d(t) = \sum_{i=1}^{P} A_i \sin(\omega_i t + \phi_i).e^{-\gamma_i t} + \eta(t) \qquad (2.1)$$

in which the variable $\omega_i = N(\omega_0, \sigma_\omega^2)$, i.e. normally distributed with mean ω_0 and standard deviation σ_ω, while ω_0, A_i, and ϕ_i are uniformly distributed over their respective ranges (Table 2.2). The values of the damping factors γ_i are chosen so that they produce the desired overall length of the simulated oscillatory transient disturbance.

A transitory notch is simulated differently from the other three types by adjusting its rise and decay times and the additive oscillatory and noise components. Transient disturbances caused by narrow notches and impulses often appear very similar in their waveforms, providing a real challenge to our classification problem.

At a sampling frequency of 12.8kHz, the simulation parameters we have used to generate the disturbance *d*(*t*) in (2.1) are given in Table 2.2 below. It is important to note that the ranges of the actual frequency ω_t of the three classes of oscillatory transient do overlap due to its assumed normal distribution.

Class of oscillatory transient	Frequency range for ω_0 [Hz]	Frequency deviation σ_ω [Hz]	Range for factor γ_t	Typical Magnitude [pu]	Typical Duration [ms]
IT	3000-6000	300	0.35-0.65	0.1-5	< 1
HF	1200-3000	150	0.025-0.05	0.1-2	2-5
LF	500-1200	100	0.01-0.025	0.1-2	5-20
NT	3000-6000	300	0.2-0.4	0.1-0.5	0.4-2.5

Table 2.2: Simulation parameters for the generation of disturbance

2.6 CONCLUSION

Power quality has become an important concern for customers since more and more sensitive equipment is connected to the power supply nowadays. As the power travels through transmission lines and energises electric equipment, the various pieces of equipment it energises can change the quality of the power, making it less suitable for other equipment or applications. These changes in power quality are especially common in large industrial and commercial complexes, which include increases and decreases in voltage, momentary power outages, and noise on the electrical system. If the power quality is too poor, is can even cause equipment to malfunction or to breakdown.

Time and steady state models are used to analyse transient disturbances and steady state disturbances. In particular, wavelet analysis is suitable for analysing transient disturbances. In the next three chapters, we present different wavelet-based techniques to analyse and extract the transient disturbances features, which are further used for the classification of these transient disturbance types.

Chapter 3

TIME-FREQUENCY ANALYSIS TECHNIQUES

3.1 INTRODUCTION

Traditional Fourier techniques such as Fourier Series and Fourier Transform have been used as powerful tools for signal analysis and decomposition. However, due to the lack of time localisation of Fourier techniques, they are not really suited to the analysis of non-stationary signals. To overcome this weakness, in 1946 Dennis Gabor formulated a fundamental approach for signal decomposition in terms of elementary signals [16]. His approach has since become a paradigm for the spectral analysis techniques associated with time-frequency or time-scale methods such as Short-time Fourier transforms (STFT), Wigner transforms and Wavelet transforms (WT).

Wavelet theory was developed as a unifying framework in the 1980s, although similar ideals and construction took place as early as the beginning of the century [17,18]. The idea of decomposing the signal into various time-scale resolutions has in fact recently emerged independently in many different fields of mathematics, quantum physics, engineering, economic analysis and seismic geology. In the mid-1980s, a group of geophysicists, theoretical physicists and mathematicians, namely Morlet, Grossmann, Meyer and Lemarie built a strong mathematical foundation around the subject and named their work '*Wavelet*'. A couple of years later, Daubechies, Cohen and Mallat added their contribution to the theory of wavelets which established connections to discrete signal processing results [19,20]. The wavelet transform is defined as "a tool that cups up data or functions or operators into different frequency components, and then studies each component with a resolution matched to its scale" [21]. It has advantages over the Fourier techniques in analysing non-stationary signals which contain discontinuities and sharp transitions.

This chapter introduces some most popular time-frequency analysis techniques, among which we will concentrate on the wavelet transform techniques as signal processing tools used throughout this thesis. The next section presents the relationship between time and frequency resolutions that states the localised nature of time-frequency decomposition. The third section is concerned with the STFT technique, which has been very popular in the last few decades in dealing with non-stationary signals. Its main weakness is the fixed window length (i.e. time resolution). Wavelet transforms are presented in the fourth and fifth sections. They cover the continuous wavelet transform (CWT), the discrete wavelet transform (DWT), its two particular bases, which are the orthogonal basis and biorthogonal basis, and finally the second-generation WT.

3.2 TIME AND FREQUENCY RESOLUTIONS

For many applications in signal processing and harmonic analysis, signals are decomposed into a family of functions that are well localized both in time and frequency. Such functions are called *time-frequency atoms*. The decomposition properties depend on the choice of time-frequency atoms. To extract information from complex signals, it is often necessary to adapt the time-frequency decomposition to particular signal structures.

In $\mathbf{L}^2$(R), consider a general family of time-frequency atoms of norm 1 $\{\phi_\gamma\}_{\gamma\in\Gamma}$, where γ may be a multi-indexed parameter. The corresponding linear time-frequency transform of a signal f in $\mathbf{L}^2$(R) is defined by the inner product of f and ϕ_γ

$$Tf(\gamma) = < f, \phi_\gamma > = \int_{-\infty}^{+\infty} f(t)\phi_\gamma^*(t)dt \tag{3.1}$$

where * denotes complex conjugate. The inner product in (3.1) shows how much correlation there is between the signal f and the atoms ϕ_γ.

In the time-frequency plane (t,ω), the slice of information provided by $< f, \phi_\gamma >$ is represented by a region whose location and width depend on the time-frequency spread of ϕ_γ. In the time domain, ϕ_γ centres at t_c and spreads around t_c with a variance $\sigma_t^2(\gamma)$ that

$$t_c = \int_{-\infty}^{+\infty} t\left|\phi_\gamma(t)\right|^2 dt \tag{3.2}$$

$$\sigma_t^2(\gamma) = \int_{-\infty}^{+\infty} (t - t_c)^2 \left|\phi_\gamma(t)\right|^2 dt \tag{3.3}$$

According to Parseval and Planchevel formulars, inner producs and norms in $\mathbf{L}^2$(R) are conserved by the Fourier transform up to a factor of 2π.

$$Tf(\gamma) = \int_{-\infty}^{+\infty} f(t)\phi_\gamma^*(t)dt = \frac{1}{2\pi}\int_{-\infty}^{+\infty} F(\omega)\Phi_\gamma^*(\omega)d\omega \tag{3.4}$$

and
$$\frac{1}{2\pi}\int_{-\infty}^{+\infty}\left|\Phi_\gamma(\omega)\right|^2 d\omega = \int_{-\infty}^{+\infty}\left|\phi_\gamma(t)\right|^2 dt$$
$$= \left\|\phi_\gamma\right\|^2$$

Hence, Φ_γ is centred on frequency ω_c and spreads around ω_c with a variance $\sigma_\omega^2(\gamma)$ where

$$\omega_c = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \omega\left|\Phi_\gamma(\omega)\right|^2 d\omega \tag{3.5}$$

$$\sigma_\omega^2 = \frac{1}{2\pi}\int_{-\infty}^{+\infty}(\omega-\omega_c)^2\left|\Phi_\gamma(\omega)\right|^2 d\omega \tag{3.6}$$

If $\phi_\gamma(t)$ approaches zero when t is outside a neighbourhood of t_c and $\Phi_\gamma(\omega)$ is very small for ω far from ω_c, then the 'Heisenberg box' provided by $< f, \phi_\gamma >$ is localised and centred at (t_c, ω_c) in the time-frequency plane. Figure 3.1 illustrates the Heisenberg box whose width along the time axis is $\sigma_t(\gamma)$ and along the frequency axis is $\sigma_\omega(\gamma)$.

By choosing $\sigma_t(\gamma)$ and $\sigma_\omega(\gamma)$ appropriately, the entire time-frequency plane can be covered completely. However, the resolutions in time and frequency cannot be arbitrarily small according to the well-known *Heisenberg uncertainty principle* [27]. It states that the Time-Bandwidth product of a signal cannot be less than a certain minimum value and that the time-frequency resolution of ϕ_γ (area of the Heisenberg box) is at least 1/2, i.e.

$$\sigma_t\,\sigma_\omega \geq 1/2 \tag{3.7}$$

The uncertainty principle states the trade-off involved in achieving good time resolution or good frequency resolution. It limits the joint resolution of ϕ_γ in time and in frequency. Gaussian functions are therefore often used since they meet the bound with equality [22,23].

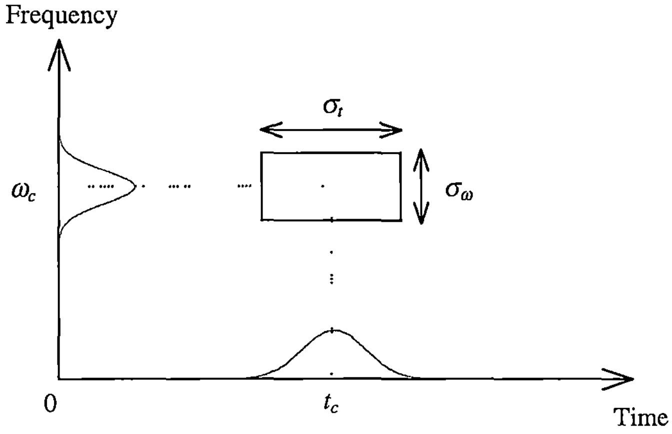


Figure 3.1: Heisenberg box representing the resolution of an atom ϕ_γ in the time-frequency plane.

3.3 SHORT-TIME FOURIER TRANSFORMS (STFT)

In 1946, Gabor [16] introduced windowed Fourier transforms or short-time Fourier transforms. These transforms are modified versions of the traditional Fourier transform. The idea is to introduce a 'local frequency' parameter so that the 'local' Fourier transform looks at the signal through a window over which the signal can be approximately stationary. Using a real and symmetric window, a short-time Fourier atom is constructed by translating this window with a time τ and modulating it with a frequency ω.

$$g_{\omega,\tau}(t) = g(t - \tau)e^{-j\omega t} \tag{3.8}$$

The window is normalised so that $\|g_{\omega,\tau}\| = 1$ for any $(\tau,\omega) \in \mathrm{R}^2$. Mathematical expression of the STFT can be represented as an inner product of the analysed signal $f(t)$ and the shifted-modulated version of the window $g(t)$ [23,24].

$$\mathrm{STFT}f(\omega,\tau) = \langle f(t), g_{\omega,\tau}(t)\rangle = \int_{-\infty}^{+\infty} f(t)g^*(t-\tau)e^{-j\omega t}dt \tag{3.9}$$

The window's translation in time by τ corresponds to a shift of the tile in the time-frequency plane by τ in time axis, while modulating it with $e^{-j\omega t}$ corresponds to a shift of the tile by ω in the frequency axis. The STFT maps the signal into a two-dimensional function on the time-frequency plane.

However, the time-frequency resolution of the existence windows is limited by the uncertainty principle. There is a trade-off between time resolution and frequency resolution of a certain type of window. Short windows certainly enhance the time resolution but at the expense of frequency resolution, while long windows have poor time resolution but improve the frequency resolution.

In STFT, since the time-frequency atoms are constructed by translating and modulating the same window, their time-frequency resolution remains the same across the time-frequency plane that is illustrated as in Figure 3.2. Also since the window is real and symmetric, an atom $g_{\omega,\tau}(t) = g(t - \tau)e^{-j\omega t}$ is centred at (τ,ω_c) in the time-frequency plane. The spreads of time and frequency around this centre are independent of τ and ω_c, and are given by

$$\sigma_t^2(\gamma)=\int_{-\infty}^{+\infty}(t-\tau)^2\left|g_{\omega,\tau}(t)\right|^2dt=\int_{-\infty}^{+\infty}t\left|g(t)\right|^2dt \tag{3.10}$$

$$\sigma_\omega^2=\frac{1}{2\pi}\int_{-\infty}^{+\infty}(\omega-\omega_c)^2\left|G_{\omega,\tau}(\omega)\right|^2d\omega=\frac{1}{2\pi}\int_{-\infty}^{+\infty}\omega^2\left|G(\omega)\right|^2d\omega \tag{3.11}$$

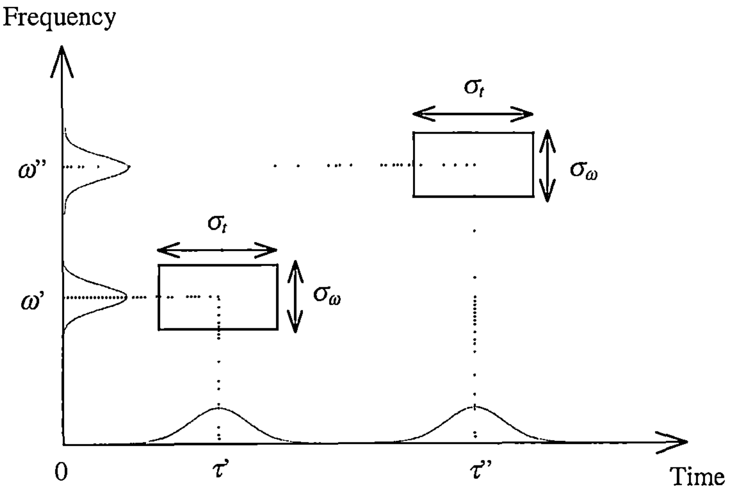


Figure 3.2: Heisenberg box of two STFT atoms $g_{\tau',\omega'}(t)$ and $g_{\tau'',\omega''}(t)$

There is no admissibility constraint on the window used in STFT since it is sufficient for the window to have finite energy. The signal $f(t)$ can be recovered from its STFT by a double integral as

$$f(t)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}STFTf(\omega,\tau)g_{\omega,\tau}(t)d\omega d\tau \tag{3.12}$$

Formula (3.12) can also be written as

$$f(t)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}<f,g_{\omega,\tau}>g_{\omega,\tau}(t)d\omega d\tau \tag{3.13}$$

The decomposition in STFT first looks like that in an orthonormal space in $\mathbf{L}^2$(R). However it does not, as the space of $\{g_{\omega,\tau}\}_{\omega,\tau\in R^2}$ is not orthonormal but is very redundant. It is possible to reduce the redundancy of the transform space by discretising the STFT on a rectangular grid $(n\tau_0,m\omega_0)_{m,n\in N}$. The choice of ω_0 is limited by the bandwidth (BW) of the lowpass window function $g(t)$, and τ_0 is chosen to be smaller than 2π/BW, in order to have adequate sampling [25]. Such discretisation guarantees an adequate covering of the entire time-frequency plane, and also guarantees the recovery of the complete signal $f(t)$ from its STFT.

In STFT, once a window has been chosen, the time and frequency resolution given by (3.10) and (3.11) is fixed over the entire time-frequency plane. Hence, if a signal contains short bursts as well as long quasi-stationary components, then we cannot analyse the signal both with good time resolution for the bursts and good frequency resolution for the quasi-stationary components. This is the main weakness of the STFT.

3.4 WAVELET TRANSFORM

To overcome the resolution limitation in the STFT, recent years have witnessed a proliferation of applications of wavelet transforms to signal analysis in a wide variety of fields from geo-physics to telecommunications to bio-medical engineering. The wavelet transform has the advantage of an optimal trade-off between time resolution and frequency resolution in the time-frequency plane, thus providing an effective multi-resolution analysis.

To analyse a non-stationary signal, we wish to achieve good time resolutions for high frequency bursts, and good frequency resolutions for low frequency components. As the joint of time and frequency resolution of analysis filters is limited by the uncertainty principle, good time resolutions for high frequency bursts can only be achieved with short windows requiring the filters to have large bandwidth. While good frequency resolutions for low frequency components requires small bandwidth filters that have long windows, it is natural to impose that the frequency resolution $\Delta\omega$ of a filter is proportional to the centre frequency of that filter ω, or

$$\frac{\Delta\omega}{\omega} = c \tag{3.14}$$

where c is a constant. The analysis is viewed as a filter bank, in which the decomposing bandpass filters have a constant relative bandwidth, called 'constant-Q' analysis [23,25,26]. The frequency responses of the constant-Q analysis are spread in a logarithmic scale, which is in contrast to the regularly spaced over the frequency axis of the STFT. Figure 3.3 shows the tiling of the time-frequency planes for the STFT and for the WT, which has the constant-Q frequency responses analysis. Figure 3.4 shows three basis functions and their corresponding time-frequency resolution windows of the STFT and the WT.

3.4.1 Continuous Wavelet Transform (CWT)

Continuous Wavelet Transforms follows exactly the same ideas of a filter bank while adding a simplification that all impulse responses of the filters in the filter bank are scaled (i.e. stretched or compressed) versions of the same prototype $\psi(t)$, i.e.

$$\psi_s(t) = \frac{1}{\sqrt{|s|}}\psi\left(\frac{t}{s}\right), \quad s \in R\backslash\{0\} \tag{3.15}$$

where s is a scale factor, and $1/\sqrt{|s|}$ is used for energy normalisation so that $\|\psi_s(t)\|^2 = \|\psi(t)\|^2$. For convenience, the norm $\|\psi(t)\|^2$ is normalised to one. The CWT of a signal $f(t)$, with respect to a mother wavelet $\psi(t)$, at a scale s and location τ is then defined as [21,25]

$$CWTf(s,\tau) = \frac{1}{\sqrt{|s|}}\int_{-\infty}^{+\infty} f(t)\psi^*\left(\frac{t-\tau}{s}\right)dt, \quad s \in R\setminus\{0\},\ \tau \in R \tag{3.16}$$

Expression in (3.16) can be rewritten as an inner product

$$CWTf(s,\tau) = < f(t), \psi_{s,\tau}(t) > \tag{3.17}$$

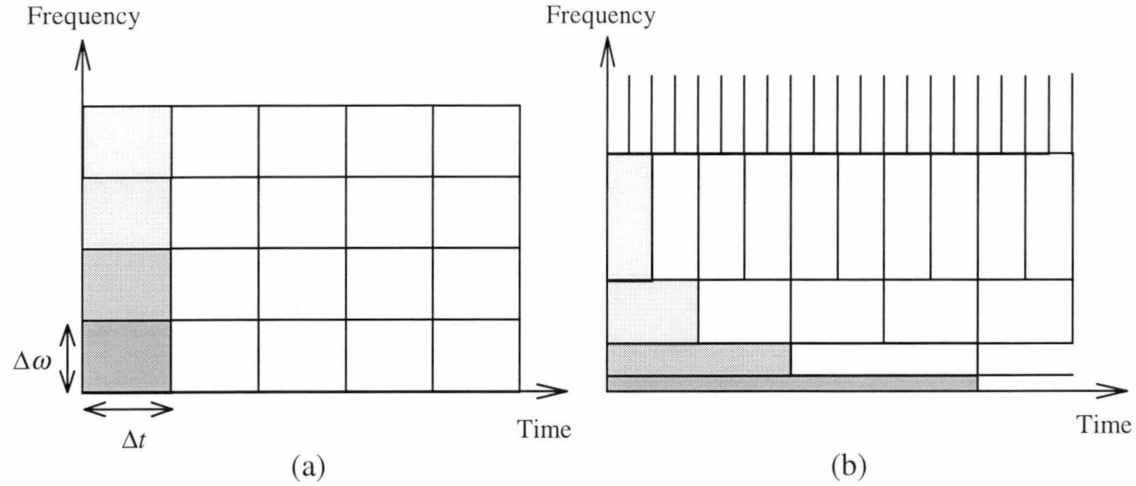


Figure 3.3: The tiling of the time-frequency plane for the (a) STFT and the (b) WT.

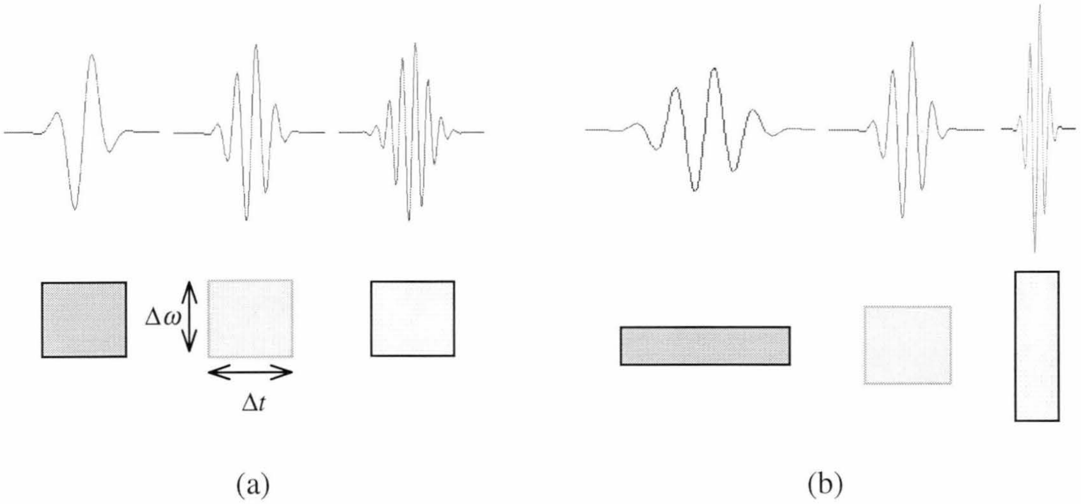


Figure 3.4: Three basis functions and their corresponding time-frequency resolution windows of the (a) STFT and (b) WT.

where $\psi_{s,\tau}(t)$ is the scaled (or dilated) and shifted (or translated) version of the mother wavelet $\psi(t)$, that is

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{|s|}}\psi\left(\frac{t-\tau}{s}\right) \tag{3.18}$$

In order for the family $\{\psi_{s,\tau}(t)\}_{s\in R^+,\tau\in R}$ to cover the entire time-frequency plane and the transform to be a constant-Q frequency response analysis, the mother wavelet $\psi(t)$ has to be a bandpass function. Therefore, the scaling operation on the wavelet will only shift and spread its spectrum in a logarithmic scale along the frequency axis. This implies the wavelet $\psi(t)$ in $\mathbf{L}^2$(R) has a zero mean,

$$\int_{-\infty}^{+\infty}\psi(t)dt = 0 \tag{3.19}$$

In fact, the condition in (3.19) is the only requirement for the transform in (3.16) to be invertible [25]. However this condition is not sufficient for the transform to be complete and to maintain an energy conservation. An additional condition is required - that the Fourier transform $\Psi(\omega)$ of the wavelet is continuously differentiable [27]. In particular, the wavelet $\psi(t)$ must satisfy the '*admissibility condition*' defined by [21,26,27]

$$C_\psi = \int_{-\infty}^{+\infty}\frac{|\Psi(\omega)|^2}{|\omega|}d\omega < \infty \tag{3.20}$$

(3.20) implies $\Psi(0) = 0$, i.e. $\psi(t)$ is a bandpass function.

Then the function $f(t)$ can be recovered from its CWT via the '*Resolution of Identity*' equation defined by [21,24,26], for any function $f(t)$ and $g(t)$ in $\mathbf{L}^2(R)$, as

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} CWTf(s,\tau)CWTg^*(s,\tau)\frac{dsd\tau}{s^2} = C_\psi < f,g > \tag{3.21}$$

The reconstruction formula of $f(t)$ is obtained from (3.21) as

$$f(t) = \frac{1}{C_\psi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} CWTf(s,\tau)\psi_{s,\tau}(t)\frac{dsd\tau}{s^2} \tag{3.22}$$

The transform maintains an energy conservation so that

$$\|f(t)\|^2 = \int_{-\infty}^{+\infty} |f(t)|^2 dt = \frac{1}{C_\psi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |CWTf(s,\tau)|^2 \frac{dsd\tau}{s^2} \tag{3.23}$$

Like the STFT, CWT is a redundant representation as the scaling and translating parameters vary in a continuous manner. The redundancy is characterised by a *reproducing kernel equation* such that a CWT$f(s',\tau')$ can be written as a continuous linear combination of the set of weighted {CWT$f(s,\tau)$}

$$CWTf(s',\tau') = \frac{1}{C_\psi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} K(s',\tau',s,\tau) CWTf(s,\tau) \frac{d\tau ds}{s^2} \tag{3.24}$$

where

$$K(s',\tau',s,\tau) = < \psi_{s,\tau}, \psi_{s',\tau'} > \tag{3.25}$$

is the reproducing kernel, which measures the correlation of two wavelets $\psi_{s,\tau}(t)$ and $\psi_{s',\tau'}(t)$. So any function in time-frequency plane is the CWT of a signal $f(t)$ in $\mathbf{L}^2(\mathrm{R})$ if and only if it satisfies the reproducing kernel equation (3.24) [20,21].

The time-frequency resolution of a CWT depends on the time-frequency spread of the wavelet atoms $\psi_{s,\tau}(t)$. For simplicity, the centre of $\psi(t)$ is assumed to be at zero, which implies that $\psi_{s,\tau}(t)$ will be centred at $t = \tau$. The spread in time of $\psi_{s,\tau}(t)$ is

$$\int_{-\infty}^{+\infty} (t-\tau)^2 \left| \frac{1}{\sqrt{|s|}} \psi\left(\frac{t-\tau}{s}\right) \right|^2 dt = s^2 \sigma_t^{\ 2} \tag{3.26}$$

where $\sigma_t^{\ 2} = \int_{-\infty}^{+\infty} t^2 |\psi(t)|^2 dt$ is the spread in time of $\psi(t)$.

To analyse a real signal, it is necessary to use an 'analytic wavelet' whose spectrum has no negative frequencies [21,27], then CWT$f(s,\tau)$ = 0 if s < 0. The centre frequency ω_c of $\Psi(\omega)$ is

$$\omega_c = \frac{1}{2\pi} \int_0^{+\infty} \omega |\Psi(\omega)|^2 d\omega \tag{3.27}$$

And the spectrum of $\psi_{s,\tau}(t)$ is a dilation of $\Psi(\omega)$ by 1/s

$$\Psi_{s,\tau}(\omega) = \sqrt{|s|}\Psi(s\omega)e^{-j\omega\tau} \tag{3.28}$$

The centre frequency of $\Psi_{s,\tau}(\omega)$ is therefore ω_c/s. The frequency spread around this centre is

$$\frac{1}{2\pi}\int_0^{+\infty}\left(\omega-\frac{\omega_c}{s}\right)^2\left|\Psi_{s,\tau}(\omega)\right|^2 d\omega = \frac{\sigma_\omega{}^2}{s^2} \tag{3.29}$$

where $\sigma_\omega{}^2 = \frac{1}{2\pi}\int_0^{+\infty}(\omega-\omega_c)^2\left|\Psi(\omega)\right|^2 d\omega$ is the spread in frequency of $\Psi(\omega)$

Hence the energy spread of the wavelet atoms $\psi_{s,\tau}(t)$ corresponds to a Heisenberg box which is centred at $(\tau,\omega_c/s)$ and has the size of $s\sigma_t$ along the time axis and σ_ω/s along the frequency axis. When the wavelet is scaled and translated, its time-frequency resolution will change depending on s, but the Time-Bandwidth product remains unchanged and equal to $\sigma_t\sigma_\omega$. Figure 3.5 shows the time-frequency resolution of two wavelet atoms in the time-frequency plane.

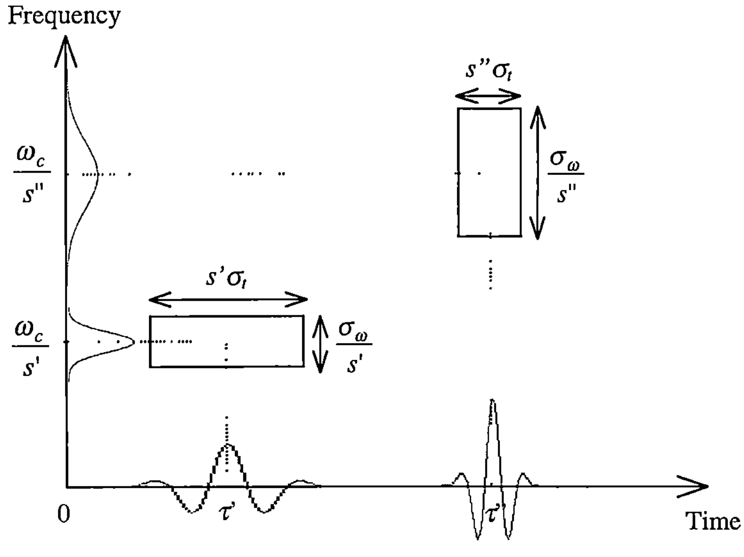


Figure 3.5: Heisenberg box of two wavelet atoms.

3.4.2 Discrete Wavelet Transform (DWT) and Frame

In the CWT, consider the family of wavelet $\{\psi_{s,\tau}(t)\}$ that is given in (3.18)

$$\psi_{s,\tau}(t) = |s|^{-1/2}\psi\left(\frac{t-\tau}{s}\right) \qquad \text{where } s \in R\backslash\{0\},\ \tau \in R$$

Over positive frequencies, the wavelet's spectrum $\Psi(\omega)$ has an effective support that is proportional to $1/s$. Therefore, to obtain full coverage of the time-frequency plane, the discretisation of the scale parameter s is chosen to be an exponential sequence $\{s_0^j\}_{j\in Z}$ with a sufficiently small dilation step $s_0 > 1$. For $j \neq 0$, the wavelet basis function is rescaled and the translation step size at scale j cannot be chosen independently of j. Since the effective support in time of the wavelet at scale j is proportional to s_0^j, the translation step size at this scale is chosen to be equal to s_0^j times the step size τ_0 at scale $j = 0$. Therefore, the discretisation of the wavelet family is [27]

$$\psi_{j,n}(t) = \frac{1}{\sqrt{s_0{}^j}}\psi\left(\frac{t - n\tau_0 s_0{}^j}{s_0{}^j}\right) \qquad j,n \in Z,\ s_0 > 1,\ \tau_0 > 0 \tag{3.30}$$

So in the discretisation scheme, at small scales the wavelets have narrow support width (i.e. wide frequency bandwidths) and are translated by a small step size, while at larger scales the wavelets will have wider support width (i.e. narrower frequency bandwidths) and are translated by a larger step size. The dilation s_0 and translation τ_0 steps are chosen so that the entire time-frequency plane is covered (Figure 3.3 b). In particular, when $s_0 = 2$ and $\tau_0 = 1$, we have the Dyadic DWT.

The discrete wavelet transform of a function $f(t)$ in $\mathbf{L}^2(R)$ is defined as

$$\begin{aligned} Wf(j,n) &= \frac{1}{\sqrt{s_0{}^j}}\int_{-\infty}^{+\infty} f(t)\psi^*\left(\frac{t - n\tau_0 s_0{}^j}{s_0{}^j}\right)dt \\ &= < f, \psi_{j,n} > \end{aligned} \tag{3.31}$$

Two important questions now can be asked about the completeness and the stability of the DWT:

1. Do the discrete wavelet coefficients $< f,\psi_{j,n} >$ completely characterise the function *f*(*t*) ?

2. Can we reconstruct *f*(*t*) in a numerically stable way from its discrete coefficients $< f,\psi_{j,n} >$?

In the case of the CWT, the *Resolution of Identity* equation (3.21) can give an immediate answer to both questions. In the DWT case, there is no analogy of the resolution of identity but the theory of *frames* provides the answers to these questions.

In 1952, Duffin and Schaeffer [28] introduced the theory of frames in the context of the non-harmonic Fourier series. In 1985, Grossmann [29] pointed out the connection between frames and numerically stable reconstruction from DWT. If the discrete wavelet family $\{\psi_{j,n}(t)\}$ constitutes a frame in $\mathbf{L}^2(R)$, then any function *f*(*t*) in $\mathbf{L}^2(R)$ can be characterised from its inner products $\{ < f,\psi_{j,n} > \}$.

The set of wavelets $\{\psi_{j,n}(t)\}$ is a frame of $\mathbf{L}^2(R)$ if there exist two constants $A > 0$ and $B > 0$ such that for any $f \in \mathbf{L}^2(R)$ [20,21,26]

$$A\|f\|^2 \le \sum_j \sum_n |< f,\psi_{j,n} >|^2 \le B\|f\|^2 \qquad (3.32)$$

A frame defines a complete and stable signal representation. In addition, it specifies a *continuity condition* in DWT. The continuity condition implies that if the distance between any two functions f_1 and f_2 in $\mathbf{L}^2(R)$, $\|f_1 - f_2\|$, is small then their distance in the transform domain $\|< f_1,\psi_{j,n} > - < f_2,\psi_{j,n} >\|$ is also small.

Under the stability condition (3.32), the DWT operator $W =< f,\psi_{j,n} >$ is the linear operator that maps a function f in $\mathbf{L}^2(R)$ onto $l^2(Z^2)$ which contains sets of any sequence $c = \{c_{j,n}\}_{j,n \in Z}$ that

$$\|c\|^2 = \sum_j \sum_n \|c_{j,n}\|^2 < \infty \qquad (3.33)$$

It follows from the stability condition (3.32) that $\|Wf\|^2 \le B\|f\|^2$, which implies that W is a one-to-one bounded linear operator. The adjoint operator W^* of W is defined by

$$W^*c = \sum_j \sum_n c_{j,n}\psi_{j,n} \tag{3.34}$$

Since, we have

$$\begin{aligned} < Wf, Wf > &= \sum_j \sum_n |< f, \psi_{j,n} >|^2 \\ &= \sum_j \sum_n << f, \psi_{j,n} > \psi_{j,n}, f > \\ &= < W^*(Wf), f > \end{aligned} \tag{3.35}$$

The lower bound of the stability condition becomes

$$< W^*(Wf), f > \ \ge A\|f\|^2$$

$$W^*W \ge AI_d \tag{3.36}$$

where I_d is the identity operator. In terms of W, the stability condition (3.32) can therefore be rewritten as

$$AI_d \le W^*W \le BI_d \tag{3.37}$$

This implies that W^*W is invertible and its inverse $(W^*W)^{-1}$ is bounded by

$$B^{-1}I_d \le (W^*W)^{-1} \le A^{-1}I_d \tag{3.38}$$

Hence, any function $f(t)$ in $\mathbf{L}^2(R)$ can be reconstructed from its DWT values in (3.31) by using the formula

$$\begin{aligned} f(t) &= (W^*W)^{-1}(W^*W)f \\ &= \sum_j \sum_n < f, \psi_{j,n} > (W^*W)^{-1}\psi_{j,n} \end{aligned} \tag{3.39}$$

By setting

$$\tilde{\psi}_{j,n} = (W^*W)^{-1}\psi_{j,n} \tag{3.40}$$

the reconstruction formula (3.39) can be written as

$$\begin{cases} < f, g >= \sum_j \sum_n < f, \psi_{j,n} >< \tilde{\psi}_{j,n}, g > \\ f(t) = \sum_j \sum_n < f, \psi_{j,n} > \tilde{\psi}_{j,n} = \sum_j \sum_n < f, \tilde{\psi}_{j,n} > \psi_{j,n} \end{cases} \tag{3.41}$$

for any function $f(t)$ and $g(t)$ in $\mathbf{L}^2(R)$. The family $\{\tilde{\psi}_{j,n}(t)\}$ is called the *dual frame* of $\{\psi_{j,n}(t)\}$. Therefore, given a wavelet frame $\{\psi_{j,n}(t)\}_{j,n\in Z}$, we only need to compute $\tilde{\psi}_{j,n} = (W^*W)^{-1}\psi_{j,n}$ in order to reconstruct a function in $\mathbf{L}^2(R)$ from its DWT values.

If the bounds A and B are close to each other, then (3.38) shows that $(W^*W)^{-1}$ is close to $\frac{2}{A+B}I_d$ so that $\tilde{\psi}_{j,n}(t)$ is close to $\frac{2}{A+B}\psi_{j,n}(t)$. In fact

$$f(t) = \frac{2}{A+B}\sum_j \sum_n < f, \psi_{j,n} > \psi_{j,n} + Rf \tag{3.42}$$

where $R = I_d - \frac{2}{A+B}W^*W$ defines a residue operation. Formula (3.37) gives

$$-\frac{B-A}{B+A}I_d \leq R \leq \frac{B-A}{B+A}I_d \tag{3.43}$$

and the norm of R is given by

$$\|R\| \leq \frac{B-A}{B+A} \tag{3.44}$$

So if A is close to B, $\|R\|$ becomes small. Then the term Rf in (3.42) can be dropped out and we obtain an accurate reconstruction of the function $f(t)$. The closer A is to B, the more accurate the reconstruction is. When the frame bounds are equal, that is $A = B$ then the frame is called a *tight frame*. In this case, the wavelets behave the same as an orthonormal basis, although they may not even be linearly independent [19].

When the wavelets $\{\psi_{j,n}(t)\}_{j,n\in Z}$ are linearly independent and span $\mathbf{L}^2(R)$, this family of wavelets is said to be a Riesz basis in $\mathbf{L}^2(R)$ [21,26,27], which is characterised by

$$f(t)=\sum_j\sum_n d_{j,n}\psi_{j,n}(t) \tag{3.45}$$

with

$$A\|f\|^2\le\sum_j\sum_n|d_{j,n}|^2\le B\|f\|^2 \tag{3.46}$$

Since $\{\psi_{j,n}(t)\}_{j,n\in Z}$ are linearly independent, one can derive from the reconstruction formula (3.41) that the dual basis functions $\{\tilde{\psi}_{j,n}(t)\}_{j,n\in Z}$ are also linearly independent in $\mathbf{L}^2(R)$. By replacing $f(t) = \psi_{j',n'}(t)$ in (3.41), we obtain

$$\psi_{j',n'}(t)=\sum_j\sum_n<\psi_{j',n'},\tilde{\psi}_{j,n}>\psi_{j,n} \tag{3.47}$$

and the linear independence implies that

$$<\psi_{j',n'},\tilde{\psi}_{j,n}>=\delta_{j'-j,n'-n} \tag{3.48}$$

where the delta function $\delta_{j'-j,n'-n}=1$ if $j' = j$ and $n' = n$, otherwise this function is zero. The WT representation is therefore non-redundant if the wavelets form a Riesz basis in $\mathbf{L}^2(R)$.

3.4.3 Orthogonal Wavelet Bases

In a tight frame and the frame bounds are equal to one (i.e. $A = B = 1$), the set of wavelets $\{\psi_{j,n}(t)\}$ forms an orthonormal basis of $\mathbf{L}^2(R)$. The operation $(W^*W)^{-1}$ in (3.38) becomes the identity operator I_d, and the analysis wavelets are

$$\tilde{\psi}_{j,n}(t)=(W^*W)^{-1}\psi_{j,n}(t)=\psi_{j,n}(t) \tag{3.49}$$

which implies that the analysis wavelets and synthesis wavelet are the same in an orthonomal basis, and $f(t)$ can be reconstructed by

$$f(t)=\sum_j\sum_n<f,\psi_{j,n}>\psi_{j,n}(t) \tag{3.50}$$

As a frame is complete in $\mathbf{L}^2(R)$, therefore $\{\psi_{j,n}(t)\}$ span $\mathbf{L}^2(R)$. The orthonormality of $\{\psi_{j,n}(t)\}$ means that

$$<\psi_{j',n'},\psi_{j,n}>=\delta_{j'-j,n'-n} \tag{3.51}$$

Hence, orthonormal WT has a non-redundant kernel that can represent the signal efficiently in the time-frequency plane. It has been used widely in image coding and compression fields since a large number of the transform coefficients can be discarded without affecting the reconstruction quality [30].

In the discrete time case, two methods were developed for coding purposes that require critical sampling of a minimum number of samples. Those methods are namely *multiresolution analysis* [31] and *filter banks* [32,36].

Multiresolution Analysis

Mallat [20,33] showed that the decomposition of $f(t)$ using an orthonormal wavelet basis $\{\psi_{j,n}(t)\}_{j,n\in Z}$ can indeed be interpreted as the difference between two successive resolutions. The ideal of multiresolution analysis is to compute the approximation of signals in $\mathbf{L}^2(R)$ at various resolutions with orthogonal projections on different lowpass subspaces $V_j \subset \mathbf{L}^2(R)$. Authors of [33,34] showed that a sequence $\{V_j\}_{j\in Z}$ of closed subspaces of $\mathbf{L}^2(R)$ is a multiresolution analysis if the following properties are satisfied:

1. $\forall j\in Z,\quad V_{j+1}\subset V_j$ (3.52)
2. $\lim_{j\to-\infty} V_j = \text{Closure}\left(\bigcup_{j=-\infty}^{+\infty} V_j\right)=\mathbf{L}^2(R)$ (3.53)
3. $\lim_{j\to+\infty} V_j = \bigcap_{j=-\infty}^{+\infty} V_j = \{0\}$ (3.54)
4. $\forall j\in Z,\ f(t)\in V_j \Leftrightarrow f(2^j t)\in V_0$ (3.55)
5. $\forall j,n\in Z^2,\ f(t)\in V_j \Leftrightarrow f(t-2^j n)\in V_j$ (3.56)
6. There exists $\phi(t) \in V_0$ such that $\{\phi(t-n)\}_{n\in Z}$ is an orthonormal basis of V_0

A multiresolution analysis is entirely characterised by a scaling function $\phi(t)$ that generates an orthogonal basis of the spaces V_j. The multiresolution properties 1 and 4 imply that $2^{-1/2}\phi(t/2) \in V_1 \subset V_0$, while the property 6 states that $\{\phi(t - n)\}_{n \in Z}$ is an orthonormal basis of V_0, then the former basis can be decomposed in terms of the latter basis as

$$\frac{1}{\sqrt{2}}\phi\left(\frac{t}{2}\right) = \sum_n h(n)\phi(t-n) \tag{3.57}$$

where $$h(n) = \left\langle \frac{1}{\sqrt{2}}\phi\left(\frac{t}{2}\right), \phi(t-n) \right\rangle$$

This is the well known 'dilation by 2' equation in multi-resolution analysis. As the norm $\|\phi(t)\|$ is normalised to 1, $\sum_n |\, h(n) \,|^2 = 1$. By taking the Fourier transform of (3.57), we have

$$\Phi(2\omega) = \frac{1}{\sqrt{2}} H(\omega)\Phi(\omega) \tag{3.58}$$

where $$H(\omega) = \sum_n h(n)e^{-j\omega n}$$

From (3.58), for any $j \geq 0$ we have

$$\Phi(2^{-j+1}\omega) = \frac{1}{\sqrt{2}} H(2^{-j}\omega)\Phi(2^{-j}\omega) \tag{3.59}$$

Therefore, $\Phi(\omega)$ can be expressed directly as a product of dilation of $H(\omega)$

$$\Phi(\omega) = \prod_{j=1}^{J} \frac{H(2^{-j}\omega)}{\sqrt{2}} \Phi(2^{-J}\omega) \tag{3.60}$$

If $\Phi(\omega)$ is continuous at $\omega = 0$, then $\lim\limits_{J \to +\infty} \Phi(2^{-J}\omega) = \Phi(0)$ and (3.60) can be rewritten

$$\Phi(\omega) = \prod_{j=1}^{+\infty} \frac{H(2^{-J}\omega)}{\sqrt{2}} \Phi(0) \tag{3.61}$$

To guarantee the existence of an orthonormal basis $\{\phi(t - n)\}_{n\in Z}$ for V_0 and an outcome of the dyadic scale multiresolution analysis (3.61), [33,34] showed that the infinite product $H(\omega)$ must satisfy

$$|H(\omega)|^2 + |H(\omega+\pi)|^2 = 2 \tag{3.62}$$

and

$$|H(0)| = \sqrt{2} \tag{3.63}$$

Since the subspace V_j is included in V_{j-1}, the orthogonal bandpass complement of V_j in V_{j-1} is denoted by W_j that

$$V_{j-1} = V_j \oplus W_j, \quad V_m \perp W_m \tag{3.64}$$

It follows that $\{W_j\}_{j\in Z}$ are orthogonal spaces which sum to $\mathbf{L}^2(R)$

$$\underset{j\in Z}{\oplus} W_j = \mathbf{L}^2(R) \tag{3.65}$$

and all W_j are scaled versions of W_0

$$f \in W_j \Leftrightarrow f(2^j t) \in W_0 \tag{3.66}$$

Because of the multiresolution properties 1 to 4, one can show that there also exists $\psi(t)$ such that $\{\psi(t - n)\}_{n\in Z}$ constitutes an orthonormal basis for W_0 [20,34]. Since $2^{-1/2}\psi(t/2) \in W_1 \subset V_0$, it can thus be decomposed in $\{\phi(t - n)\}_{n\in Z}$

$$\frac{1}{\sqrt{2}}\psi\left(\frac{t}{2}\right) = \sum_n g(n)\phi(t-n) \tag{3.67}$$

where $g(n) = \left\langle \frac{1}{\sqrt{2}}\psi\left(\frac{t}{2}\right), \phi(t-n) \right\rangle$ and $\sum_n | g(n) |^2 = 1$

Taking the Fourier transform of (3.67) yields

$$\Psi(2\omega) = \frac{1}{\sqrt{2}} G(\omega)\Phi(\omega) \tag{3.68}$$

where $G(\omega) = \sum_n g(n)e^{-j\omega n}$

Since $\{\psi(t - n)\}_{n\in Z}$ constitutes an orthonormal basis for W_0, then (3.66) implies that the family $\{\psi_{j,n}\}_{n\in Z}$ is an orthonormal basis of W_j, where the function $\psi_{j,n}(t)$ is defined by

$$\psi_{j,n}(t)=\frac{1}{\sqrt{2^J}}\psi\left(\frac{t-2^J n}{2^j}\right) \tag{3.69}$$

To guarantee that the family $\{\psi_{j,n}\}_{n\in Z}$ is an orthonormal basis of W_j and W_j be the orthogonal bandpass complement of V_j in V_{j-1}, the Fourier transform $G(\omega)$ of the bandpass $\psi(t)$ must satisfy [33,34]

$$|G(\omega)|^2+|G(\omega+\pi)|^2=2 \tag{3.70}$$

and

$$G(\omega)H^*(\omega)+G(\omega+\pi)H^*(\omega+\pi)=0 \tag{3.71}$$

Let us choose

$$G(\omega)=p(\omega)H^*(\omega+\pi) \tag{3.72}$$

$p(\omega)$ is a 2π periodic function and since $H^*(\omega)$ $H^*(\omega+\pi)$ cannot equal to zero for all ω, (3.71) reduces to

$$p(\omega)+p(\omega+\pi)=0 \tag{3.73}$$

One can note that, $p(\omega)$ is not uniquely established by the multiresolution analysis. A function that satisfies (3.73) is $p(\omega) = e^{-j\omega}$ and

$$G(\omega)=e^{-j\omega}H^*(\omega+\pi) \tag{3.74}$$

By taking the inverse Fourier transform of (3.74) yields

$$g(n)=(-1)^{1-n}h(1-n) \tag{3.75}$$

Given that $\{W_j\}_{j\in Z}$ are orthogonal spaces which sum to $\mathbf{L}^2(R)$, therefore any $f(t)$ in $\mathbf{L}^2(R)$ can be decomposed as

$$f(t)=\sum_j\sum_n d_{j,n}\psi_{j,n}(t) \tag{3.76}$$

where the coefficients $d_{j,n} = <f(t), \psi_{j,n}(t)>$. Since the family $\{\psi_{j,n}(t)\}_{j,n\in Z}$ are bandpass functions, $d_{j,n}$ are therefore referred to the *detail* signals. The transform retains the energy conservation such that

$$\begin{aligned}\|f\|^2 &= \left\langle \sum_j \sum_n d_{j,n}\psi_{j,n}(t), f \right\rangle \\ &= \sum_j \sum_n d_{j,n}{}^2\end{aligned} \tag{3.77}$$

Mallat's Pyramid Algorithms and Filter Banks

To compute the orthogonal wavelet coefficients of a signal measured at a finite resolution, Mallat [35,20] developed a fast algorithm called the pyramid algorithm, which decomposes successively each approximation of a function $f(t)$ in V_j into a coarser approximation of $f(t)$ in V_{j+1} plus a detail signal of $f(t)$ (the wavelet coefficients) in W_{j+1}. Although derived from the multiresolution analysis, the pyramid algorithms turn out to be equivalent to conjugate mirror filters used in discrete multirate filter banks [36,37,32].

As elements of $V_j = V_{j+1} \oplus W_{j+1}$, the orthogonal projection of $f(t)$ on V_j can therefore be decomposed as the sum of orthogonal projections on V_{j+1} and W_{j+1}:

$$P_jf(t) = P_{j+1}f(t) + Q_{j+1}f(t) \tag{3.78}$$

$P_jf(t)$ and $Q_jf(t)$ denote the orthogonal projection of function $f(t)$ onto subspaces V_j and W_j respectively and are given by

$$P_j f(t) = \sum_n a_j(n)\phi_{j,n}(t) \tag{3.79}$$

$$Q_j f(t) = \sum_n d_j(n)\psi_{j,n}(t) \tag{3.80}$$

where $a_j(n) =< f, \phi_{j,n} >$ and $d_j(n) =< f, \psi_{j,n} >$. The sequences a_j and d_j are respectively called the *smoothed version* and the *detailed version* of $f(t)$ at the scale j.

Using (3.57), any $\phi_{j+1,m}(t) \in V_{j+1} \subset V_j$ can be decomposed in the orthonormal basis $\{\phi_{j,n}(t)\}_{n\in Z}$ of V_j as

$$\begin{aligned}\phi_{j+1,m}(t) &= \sum_n < \phi_{j+1,m}, \phi_{j,n} > \phi_{j,n}(t) \\ &= \sum_n h(n-2m)\phi_{j,n}(t)\end{aligned} \qquad (3.81)$$

Also by using (3.67), any $\psi_{j+1,m}(t) \in V_{j+1} \subset V_j$ can be decomposed in this basis as

$$\begin{aligned}\psi_{j+1,m}(t) &= \sum_n < \psi_{j+1,m}, \phi_{j,n} > \phi_{j,n}(t) \\ &= \sum_n g(n-2m)\phi_{j,n}(t)\end{aligned} \qquad (3.82)$$

Taking the inner product with $f(t)$ on both sides of (3.81) and (3.82) gives the decomposition of the lowpass coefficient $a_{j+1}(m)$ and detail coefficient $d_{j+1}(m)$

$$\begin{aligned}a_{j+1}(m) &= \sum_n h(n-2m)a_j(n) \\ d_{j+1}(m) &= \sum_n g(n-2m)a_j(n)\end{aligned} \qquad (3.83)$$

Let $\bar{h}(n) = h(-n)$ and $\bar{g}(n) = g(-n)$ be the mirror image of $h(n)$ and $g(n)$ respectively, then a_{j+1} and d_{j+1} are computed by taking every second sample of the convolution of a_j with $\bar{h}$ and $\bar{g}$ respectively

$$\begin{aligned}a_{j+1}(m) &= a_j \circ \bar{h}(2m) \\ d_{j+1}(m) &= a_j \circ \bar{g}(2m)\end{aligned} \qquad (3.84)$$

The pyramid algorithm is illustrated by Figure 3.6. The lowpass filter $\bar{h}$ removes the higher frequencies of the sequence a_j to produce the courser approximation a_{j+1}, while the highpass filter $\bar{g}$ collects the remaining highest frequencies to produce the detail signal d_{j+1}.

In practice, we will start from a fine enough subspace V_L so that $f(t)$ can be approximated by its orthogonal projection onto V_L, and the pyramid will stop after a finite number of levels, level J. This means that the information

$a_L = \{< f, \phi_{L,n} >\}_{n \in Z}$ can be rewritten as the wavelet coefficients d_{L+1}, d_{L+2}, d_{L+3}, ..., d_J and a final coarse approximation a_J.

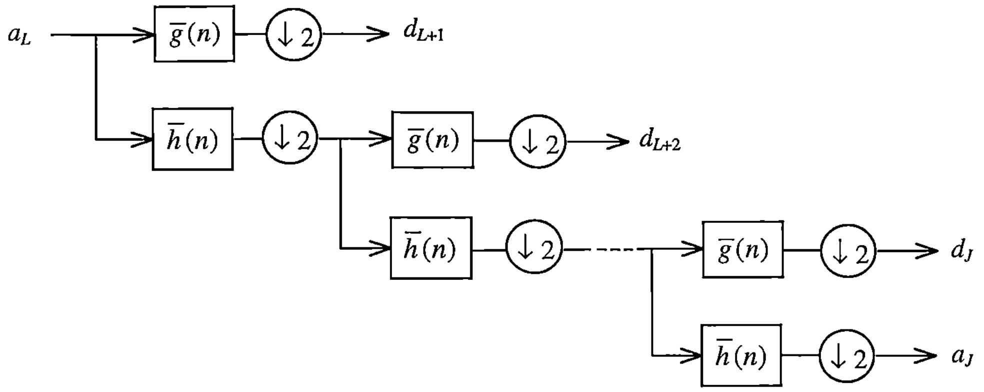


Figure 3.6: Orthogonal WT using the pyramid algorithm

The initialisation of the pyramid algorithm is to approximate $f(t)$ at a subspace V_L such that $a_L = \{< f(t), 2^{-L/2}\phi(2^{-L}t - n) >\}_{n \in Z}$. If V_L is fine enough to present the resolution of the signal $f(t)$, then sampling $f(t)$ will be sufficient. This is because $\phi(t)$ is a lowpass function, whose integral equals to 1. If L is sufficiently small (towards minus infinitive) then $\phi_{L,n}(t)$ will be sufficiently short lived such that

$$< f(t), 2^{-L/2}\phi(2^{-L}t - n) > \approx 2^{-L/2} f(2^L n) \qquad (3.85)$$

Hence, if $f(t)$ is regular, then there will be a resolution at which it can be closely approximated by its inner product with the lowpass $\phi_{L,n}(t)$. From this we obtain the initial approximation for the pyramid algorithm.

In the reconstruction direction, a signal can be reconstructed from its wavelet coefficients d_{L+1}, d_{L+2}, d_{L+3}, ..., d_J and a final coarse approximation a_J. At a scale j ($L < j \leq J$), since W_{j+1} is the orthogonal complement of V_{j+1} in V_j, the union of the two bases $\{\psi_{j+1,n}(t)\}_{n \in Z}$ and $\{\phi_{j+1,n}(t)\}_{n \in Z}$ is therefore an orthonormal basis of V_j. Then any function $\phi_{j,m}(t)$ in V_j can be decomposed in this basis as

$$\phi_{j,m}(t) = \sum_n < \phi_{j,m}(t), \phi_{j+1,n}(t) > \phi_{j+1,n}(t) + \sum_n < \phi_{j,m}(t), \psi_{j+1,n}(t) > \psi_{j+1,n}(t)$$

$$= \sum_n h(m-2n)\phi_{j+1,n}(t) + \sum_n g(m-2n)\psi_{j+1,n}(t) \qquad (3.86)$$

Taking the inner product with $f(t)$ on both sides of (3.86) gives

$$a_j(m) = \sum_n h(m-2n)a_{j+1}(n) + \sum_n g(m-2n)d_{j+1}(n) \qquad (3.87)$$

Therefore, the reconstruction is an interpolation that inserts a zero between each sample of the sequences a_{j+1} and d_{j+1} (i.e. upsampling by 2), then they are convolved with the filters $h(n)$ and $g(t)$ respectively. The two convolved outputs are added to give the finer approximation a_j. Starting from the coarsest resolution $j = J$ (i.e. d_J and a_J), the iterative pyramid reconstruction algorithm reconstructs the original approximation a_L as shown in Figure 3.7.

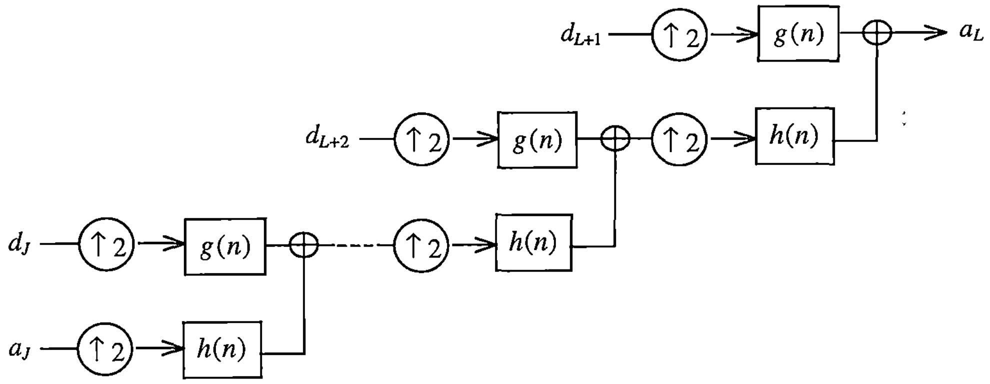


Figure 3.7: Inverse orthonormal WT using the pyramid algorithm

By substituting a_{j+1} and d_{j+1} from (3.83) into (3.87), we then have the following relationship between $g(n)$ and $h(n)$

$$\sum_n [h(m-2n)h(l-2n) + g(m-2n)g(l-2n)] = \delta_{m-l} \qquad (3.88)$$

Taking the Fourier transform on both sides of (3.88) with respect to $(m-l)$, we then have

$$\left|H(\omega)\right|^2+\left|G(\omega)\right|^2+(-1)^l\left[H(\omega)H^*(\omega+\pi)+G(\omega)G^*(\omega+\pi)\right]=2 \quad (3.89)$$

Equation (3.89) can only hold for all l if

$$\begin{aligned} &H(\omega)H^*(\omega+\pi)+G(\omega)G^*(\omega+\pi)=0 \\ &\left|H(\omega)\right|^2+\left|G(\omega)\right|^2=2 \end{aligned} \quad (3.90)$$

In fact, the pyramid analysis is equivalent to the filter bank algorithm (subband coding scheme) [33,38,39]. The classical multirate filter banks were first developed in 1975 for speech coding purposes [36]. A two-channel multirate filter bank analyses an input sequence $a_0(n)$ by convolving it with a lowpass filter $\bar{h}$ and a highpass filter $\bar{g}$ $(\bar{h}(n)=h(-n)$ and $\bar{g}(n)=g(-n))$ then by subsampling by 2 so that the outputs are

$$\begin{aligned} &a_1(n)=a_0*\bar{h}(2n) \\ &d_1(n)=a_0*\bar{g}(2n) \end{aligned} \quad (3.91)$$

At the reconstruction, a reconstructed signal $\tilde{a}_0(n)$ is obtained by first upsampling by 2 (by inserting a zero between consecutive samples) the lowpass sequence $a_1(n)$ and highpass sequence $d_1(n)$. These sequences are then filtered with a dual lowpass filter $\tilde{h}$ and a dual highpass filter $\tilde{g}$ respectively and finally are summed up to give $\tilde{a}_0(n)$ as shown in Figure 3.8

$$\tilde{a}_0(m)=\sum_n \tilde{h}(m-2n)a_1(n)+\sum_n \tilde{g}(m-2n)d_1(n) \quad (3.92)$$

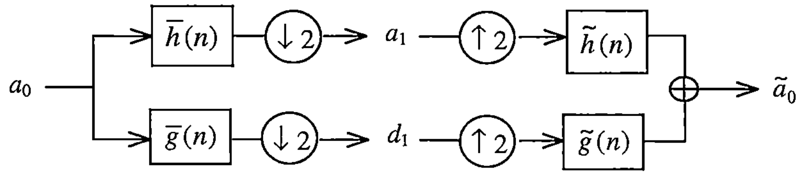


Figure 3.8: A two-channel multirate filter bank

As we can see from above, the decomposition and reconstruction processes of a two-channel multirate filter bank are equivalent to that of a pyramid algorithm. In the following, we study the necessary and sufficient condition on the filter h, g, $\tilde{h}$ and $\tilde{g}$ to guarantee a perfect reconstruction in a filter bank.

From (3.91), we have the relationship between the Fourier transforms of two subsampled sequences a_1 and d_1 and the Fourier transform of a_0 as

$$\begin{aligned} A_1(2\omega) &= \frac{1}{2}\left[A_0(\omega)H^*(\omega) + A_0(\omega+\pi)H^*(\omega+\pi)\right] \\ D_1(2\omega) &= \frac{1}{2}\left[A_0(\omega)G^*(\omega) + A_0(\omega+\pi)G^*(\omega+\pi)\right] \end{aligned} \tag{3.93}$$

And the reconstruction formula (3.92) gives

$$\tilde{A}_0(\omega) = A_1(2\omega)\tilde{H}(\omega) + D_1(2\omega)\tilde{G}(\omega) \tag{3.94}$$

Hence

$$\begin{aligned} \tilde{A}_0(\omega) = &\frac{1}{2}\left[H^*(\omega)\tilde{H}(\omega) + G^*(\omega)\tilde{G}(\omega)\right]A_0(\omega) + \\ &\frac{1}{2}\left[H^*(\omega+\pi)\tilde{H}(\omega) + G^*(\omega+\pi)\tilde{G}(\omega)\right]A_0(\omega+\pi) \end{aligned} \tag{3.95}$$

To guarantee $\tilde{a}_0 = a_0$ for all a_0, the aliasing term $A_0(\omega + \pi)$ must be cancelled out and $A_0(\omega)$ remains a unit gain, which provides the necessary and sufficient conditions of the four filters to guarantee a perfect reconstruction [40,41].

$$\begin{aligned} &H^*(\omega+\pi)\tilde{H}(\omega) + G^*(\omega+\pi)\tilde{G}(\omega) = 0 \\ &H^*(\omega)\tilde{H}(\omega) + G^*(\omega)\tilde{G}(\omega) = 2 \end{aligned} \tag{3.96}$$

In an orthogonal basis, the reconstruction filters are the same as the decomposition filters (i.e. $\tilde{h} = h$ and $\tilde{g} = g$), then the conditions in (3.96) become the pyramid's condition in (3.90). This implies that the pyramid analysis is a special case of the filter bank algorithms.

Depending on the application, the design of wavelets may require some wavelet characteristics. Some of these characteristics are spatial compactness, regularity or smoothness, symmetry or anti-symmetry. However, some characteristics are mutually exclusive in the design of orthogonal wavelets [21,42]. For example, orthogonal wavelets cannot be symmetric or anti-symmetric (with the exception of the Haar wavelet) or these wavelets cannot have both compactness and smoothness. Also the transition from passband to stopband of orthogonal wavelets is not sharp, so the design problem is still open [42]. In the next section we describe the biorthogonal wavelet bases, which have more flexibility in the design of wavelets, but still guarantee a perfect reconstruction. This is because the biorthogonal wavelet bases only require the family of wavelets to form a Riesz basis in $\mathbf{L}^2(R)$.

3.4.4 Biorthogonal Wavelet Bases

Biorthogonal wavelet bases are related to multiresolution analysis. In this thesis, for decomposition, a *primal* scaling function $\phi(t)$ and a *primal* wavelet function $\psi(t)$ are used, while a *dual* scaling function $\tilde{\phi}(t)$ and a *dual* wavelet function $\tilde{\psi}(t)$ are involved for reconstruction. The requirement that the family $\{\phi(t - n)\}_{n\in Z}$ forms an orthonormal basis for V_0 in the orthogonal wavelet case is now reduced to the requirement that this family forms a Riesz basis of the space V_0. This means the family $\{\phi(t - n)\}_{n\in Z}$ or the family $\{\psi(t - n)\}_{n\in Z}$ forms a tight frame in $\mathbf{L}^2(R)$. From (3.48), biorthogonal wavelet bases mean that

$$<\psi_{j',n'},\tilde{\psi}_{j,n}>=\delta_{j'-j,n'-n} \tag{3.97}$$

and any $f(t)$ in $\mathbf{L}^2(R)$ has two possible decompositions

$$\begin{aligned} f(t) &= \sum_j \sum_n < f,\psi_{j,n} > \tilde{\psi}_{j,n} \\ &= \sum_j \sum_n < f,\tilde{\psi}_{j,n} > \psi_{j,n} \end{aligned} \tag{3.98}$$

Since the family $\{\phi(t - n)\}_{n\in Z}$ forms a Riesz basis of the space V_0, the dual scaling function $\{\tilde{\phi}(t-n)\}_{n\in Z}$ also forms a Riesz basis in its space $\tilde{V}_0$. Let V_j and $\tilde{V}_j$ be the subspaces defined by

$$\begin{aligned} f(t)\in V_j \quad &\Leftrightarrow \quad f(2^j t)\in V_0 \\ f(t)\in \tilde{V}_j \quad &\Leftrightarrow \quad f(2^j t)\in \tilde{V}_0 \end{aligned} \tag{3.99}$$

For any $j \in Z$, one can verify that $\{\phi_{j,n}(t\}_{n\in Z}$ and $\{\tilde{\phi}_{j,n}(t)\}_{n\in Z}$ are Riesz bases of V_j and $\tilde{V}_j$. And we have

$$<\phi_{j',n'},\tilde{\phi}_{j,n}>=\delta_{j'-j,n'-n} \qquad \forall (j',j,n',n)\in Z^4 \tag{3.100}$$

Their respective bandpass complementary spaces W_j and $\tilde{W}_j$ are

$$V_j \oplus W_j = V_{j-1} \text{ and } \tilde{V}_j \oplus \tilde{W}_j = \tilde{V}_{j-1} \tag{3.101}$$

In biorthogonal, W_j is not orthogonal to V_j but is to $\tilde{V}_j$ whereas $\tilde{W}_j$ is not orthogonal to $\tilde{V}_j$ but is to V_j. Hence these two multiresolution hierarchies in biorthogonal work like two 'zipped' spaces that allow perfect reconstruction.

The construction of the basis functions $\phi(t)$ and $\psi(t)$ and their dual basis functions $\tilde{\phi}(t)$ and $\tilde{\psi}(t)$ involve the construction of the filters ($h(n)$, $g(n)$) and $(\tilde{h}(n),\tilde{g}(n))$, whose Fourier transforms satisfy

$$\begin{aligned} \Phi(2\omega) &= \frac{1}{\sqrt{2}}H(\omega)\Phi(\omega) \quad \text{and} \quad \Psi(2\omega) = \frac{1}{\sqrt{2}}G(\omega)\Psi(\omega) \\ \tilde{\Phi}(2\omega) &= \frac{1}{\sqrt{2}}\tilde{H}(\omega)\tilde{\Phi}(\omega) \quad \text{and} \quad \tilde{\Psi}(2\omega) = \frac{1}{\sqrt{2}}\tilde{G}(\omega)\tilde{\Psi}(\omega) \end{aligned} \tag{3.102}$$

In the time domain, these two-scale difference equations become

$$\begin{aligned} \phi(t) &= \sqrt{2}\sum_n h(n)\phi(2t-n) \quad \text{and} \quad \psi(t) = \sqrt{2}\sum_n h(n)\psi(2t-n) \\ \tilde{\phi}(t) &= \sqrt{2}\sum_n \tilde{h}(n)\tilde{\phi}(2t-n) \quad \text{and} \quad \tilde{\psi}(t) = \sqrt{2}\sum_n \tilde{h}(n)\tilde{\psi}(2t-n) \end{aligned} \tag{3.103}$$

Note that the wavelets $\{\psi_{j,n}(t)\}$ are not orthogonal to one another. A similar statement is made for wavelets in the dual space $\{\tilde{\psi}_{j,n}(t)\}$. But rather, wavelets in the same space are related by the two-scale difference equations given in (3.103). The filter $h(n)$ and $\tilde{h}(n)$ also satisfy the biorthogonal condition as (3.100) implies

$$\sum_m h(m-2n)\tilde{h}(m) = \delta_n \tag{3.104}$$

and their Fourier transforms satisfy

$$H^*(\omega)\tilde{H}(\omega) + H^*(\omega+\pi)\tilde{H}(\omega+\pi) = 2 \tag{3.105}$$

Similar to the filter bank case, for a perfect reconstruction in the biorthogonal basis the four filters $h(n)$, $\tilde{h}(n)$, $g(n)$ and $\tilde{g}(n)$ must satisfy

$$\begin{aligned} &H^*(\omega+\pi)\tilde{H}(\omega) + G^*(\omega+\pi)\tilde{G}(\omega) = 0 \\ &H^*(\omega)\tilde{H}(\omega) + G^*(\omega)\tilde{G}(\omega) = 2 \end{aligned} \tag{3.106}$$

From (3.105) and (3.106), one can obtain a relationship between $H(\omega)$, $\tilde{H}(\omega)$, $G(\omega)$ and $\tilde{G}(\omega)$ as

$$\begin{aligned} &G(\omega) = P_1(\omega)\tilde{H}^*(\omega+\pi) \\ &\tilde{G}(\omega) = P_2(\omega)H^*(\omega+\pi) \\ &P_1^*(\omega)P_2(\omega) = 1 \\ &P_2(\omega) + P_2(\omega+\pi) = 0 \end{aligned} \tag{3.107}$$

For $P_1(\omega) = P_2(\omega) = e^{-j\omega}$, the conditions in (3.107) give a set of four filters $h(n)$, $\tilde{h}(n)$, $g(n)$ and $\tilde{g}(n)$ that are related by

$$\begin{aligned} &g(n) = (-1)^{n+1}\tilde{h}(1-n) \\ &\tilde{g}(n) = (-1)^{n+1}h(1-n) \end{aligned} \tag{3.108}$$

3.5 SECOND GENERATION WAVELETS

Existing *first-generation wavelets* are all translates and dilates of one or more *constant* basic shapes called mother wavelets. Translation and dilation in the time domain are *linear* algebraic operations in the Fourier domain. Polynomial factorisation in the frequency domain is required in the construction of wavelets [43]. The Fourier transform, therefore, plays a crucial role in the design of these basically linear time-invariant (LTI) wavelets. These wavelets remain invariant over the entire signal or image to be analysed or at least over the time or spatial duration of the analysis frame. First-generation wavelets are therefore not suitable for applications in bounded domains such as finite-length signals and isolated objects in images (e.g. texts and subtitles) as they introduce ringing artifacts at the boundaries. They are also not appropriate in non-Euclidean spaces such as curves and curved surfaces (e.g. for face recognition), and irregular sampling grids, in which LTI Fourier-based (frequency domain) techniques are not available.

Current research on *second-generation wavelets* is concentrated on the two original motivations for the lifting scheme [44,45]: to design adaptive (time-varying) perfect reconstruction filter banks or second-generation wavelets [44], and to factorise existing first-generation wavelets for faster transforms [45]. The lifting scheme in Figure 3.9 provides a simple and flexible alternative technique for the construction of time-variant or space-variant wavelets, entirely in the time or spatial domain and adapted to the local characteristics at every sample of the signal. In other words, a second-generation 'wavelet' changes its shape from sample to sample in the signal. Lifting is also an effective technique to factorise existing wavelets into simple basic building blocks for faster computation of the corresponding wavelet transform [45]. The basic concept of the lifting scheme is to start with a very simple or trivial wavelet and design a single operator S to '*lift*' the simple wavelet to a more sophisticated wavelet, satisfying certain required properties such as smoothness and vanishing moments.

Let $H(\omega)$ be the lowpass filter for the scaling function $\phi(t)$, and $G(\omega)$ be the bandpass filter for the mother wavelet $\psi(t)$. $\tilde{H}(\omega)$ and $\tilde{G}(\omega)$ are dual filters of $H(\omega)$ and $G(\omega)$ respectively. The scheme starts with a simple analysis filter pair $(H(\omega), G(\omega))$ and the corresponding simple synthesis pair $(\tilde{H}(\omega), \tilde{G}(\omega))$, then uses an operator S to design a more sophisticated set of filters $(H^{new}(\omega), G^{new}(\omega), \tilde{H}^{new}(\omega), \tilde{G}^{new}(\omega))$.

For perfect reconstruction without aliasing [4] of the FIR filters, we must have

$$\begin{aligned} &H^*(\omega+\pi)\tilde{H}(\omega)+G^*(\omega+\pi)\tilde{G}(\omega)=0 \\ &H^*(\omega)\tilde{H}(\omega)+G^*(\omega)\tilde{G}(\omega)=2 \end{aligned} \tag{3.109}$$

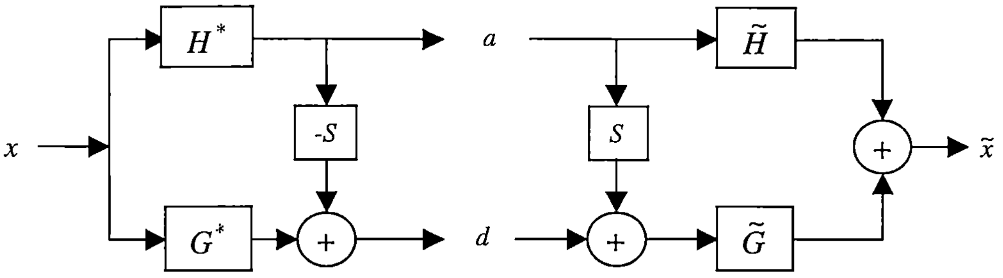


Figure 3.9: Diagram of a lifting scheme

The essence of the lifting scheme is in the lifting theorem [44] which states that if we take an initial set of biorthogonal filters $\{H(\omega), G(\omega), \tilde{H}(\omega), \tilde{G}(\omega))$, then a new set of biorthogonal filters $(H^{new}(\omega), G^{new}(\omega), \tilde{H}^{new}(\omega), \tilde{G}^{new}(\omega))$ can be found as

$$\begin{aligned} &H^{new}(\omega)=H(\omega) \\ &G^{new}(\omega)=G(\omega)-S.H(\omega) \\ &\tilde{H}^{new}(\omega)=H(\omega)+S^*.\tilde{G}(\omega) \\ &\tilde{G}^{new}(\omega)=\tilde{G}(\omega) \end{aligned} \tag{3.110}$$

where S is an operator to be designed. We can see from the set of equations in (3.110), that there are a number of advantages of the lifting scheme. Second-generation wavelets can be designed directly in the time or spatial domain without having to deal with complex Fourier analyses using a single operator S. Furthermore, once S is fixed, the lifting scheme assures that the new filters are biorthogonal. We can also observe that any required properties or conditions (e.g. to give a particular waveshape and smoothness) on the new wavelet $\psi(t)$ through its generating filter $G(\omega)$, can directly translate into required properties or conditions on S.

In [44], a 'Lazy wavelet' is introduced as a simple initial candidate, even simpler than the Haar wavelet, to start the lifting scheme. The Lazy wavelet does nothing except subsampling the data into two groups (or phases), a group of *even-indexed* samples going to the upper branch of Figure 3.10 and a group of *odd-indexed* samples to the lower branch. In this figure, the Lazy wavelet acts as a two-band *polyphase* transform splitting the input data into two phases. Odd samples x_o are lifted with the help of even samples x_e using an S operator which, in this case, consists of two steps: a *predictor P* and an *update* filter *U*.

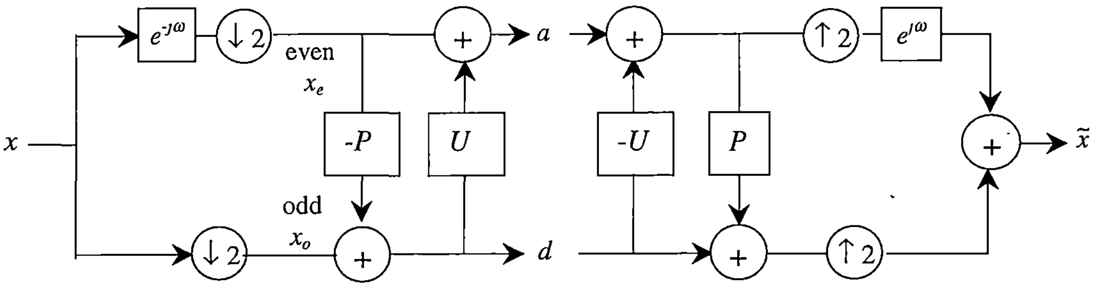


Figure 3.10: Lifting scheme starting with the Lazy wavelet

The first lifting step makes optimal use of the correlation between neighbouring signal samples to predict odd samples from the even samples. The prediction result is subtracted from the odd samples yielding the bandpass or detail signal *d*, i.e.

$$d = x_o - P(x_e) \tag{3.111}$$

The sub-sampling action resulting from the polyphase splitting usually cannot guarantee adequate spectral separation and may produce aliasing in the two-polyphase sets x_e and x_o. The second step in a lifting scheme is *to update* x_e by replacing it with an aliasing-free smoother set *a*, ready for the next lower resolution lifting stage. In the original lifting scheme [44], the update operator *U*(.) is a linear combination of the detail signal *d*, i.e.

$$a = x_e + U(d) \tag{3.112}$$

If $P(.)$ is an accurate predictor, then d will be a very sparse set. The lifting scheme replaces x_o by d, thus achieving data compression. The essence of the lifting scheme is that all the steps involved are invertible regardless of the choice and the nature of $P(.)$ and $U(.)$, eg. linear or non-linear, time invariant or variant. This is because both $U(.)$ and $P(.)$ operations are always invertible. The lifting can be expressed using invertible matrices as follows:

$$\begin{bmatrix} a \\ d \end{bmatrix} = \begin{bmatrix} 1 & U(.) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -P(.) & 1 \end{bmatrix} \begin{bmatrix} x_e \\ x_o \end{bmatrix}$$

$$= \begin{bmatrix} 1-UP & U \\ -P & 1 \end{bmatrix} \begin{bmatrix} x_e \\ x_o \end{bmatrix} \qquad (3.113)$$

If we cascade the lifting stage in Figure 3.10 J times, each stage using the smoothed signal a from the previous stage as input, we have a J-resolution decomposition for the signal x, i.e.

$$x = (d_1, d_2, d_3, \ldots\ldots. d_J, a_J) \qquad (3.114)$$

Polyphase representation is a convenient tool to express the special structure of the modulation matrix. In half-band filtering, the decimation of 2 creates even and odd phases and it is convenient to express all filters in polyphase, e.g.

$$H(\omega) = \sum_n h(2n)e^{-j\omega 2n} + \sum_n h(2n+1)e^{-j\omega(2n+1)}$$

$$= H_e(2\omega) + e^{-j\omega} H_o(2\omega) \qquad (3.115)$$

where

$$H_e(\omega) = \sum_n h(2n)e^{-j\omega n} \quad \text{and} \quad H_o(\omega) = \sum_n h(2n+1)e^{-j\omega n} \qquad (3.116)$$

Therefore

$$H_e(2\omega) = \tfrac{1}{2}\left[H(\omega) + H(\omega+\pi)\right]$$

$$H_o(2\omega) = \tfrac{1}{2}e^{j\omega}\left[H(\omega) - H(\omega+\pi)\right] \qquad (3.117)$$

and similarly for $G_e(2\omega)$ and $G_o(2\omega)$. We define the *polyphase matrix* of the half-band filtering as:

$$P(\omega) = \begin{bmatrix} H_e(2\omega) & H_o(2\omega) \\ G_e(2\omega) & G_o(2\omega) \end{bmatrix} \qquad (3.118)$$

For designing two operators $P(.)$ and $U(.)$, we recall that every wavelet or filter bank designed via the lifting scheme automatically guarantees perfect reconstruction and biorthogonality conditions. The prediction and update combination to give lowpass and bandpass outputs can also be expressed as the polyphase matrix [44]

$$\begin{aligned} \mathrm{P}(\omega) &= \begin{bmatrix} H_e(2\omega) & H_o(2\omega) \\ G_e(2\omega) & G_o(2\omega) \end{bmatrix} \\ &= \begin{bmatrix} 1-UP & U \\ -P & 1 \end{bmatrix} \end{aligned} \qquad (3.119)$$

The design of $P(.)$ and $U(.)$ is beyond the scope of this thesis, and can be found elsewhere [44,45].

3.6 CONCLUSION

It is obvious that the main advantage of wavelet transform is the freedom to choose the shape of the mother wavelet $\psi(t)$ and to vary the scale s to suit the local characteristics of the signal at $t = \tau$. The shape of the analysis wavelet is chosen to be as similar as possible to the local shape of the signal so as to maximise their inner product. In first-generation wavelets, the effectiveness of using wavelet transform mainly depends on the appropriateness of the chosen wavelet, while, in second-generation wavelets, the wavelets adapt themselves to maximise their correlation with the local signal structures.

In an infinite dimensional space, a signal can be perfectly reconstructed from its orthogonal projections on a family of orthogonal wavelet bases. However, if we loosen the orthogonality requirement, we must still entail a partial energy equivalence in the transform to guarantee the stability of the basis and the wavelet base is linearly independent. This implies that the basis is biorthogonal or the Riesz basis. The advantage of biorthogonal wavelet bases over orthogonal wavelet bases is that they provide more flexibility in the design of wavelets (e.g. allowing symmetry, smoothing), but still guarantee a perfect reconstruction.

The wavelet decompositions using orthogonal or biorthogonal bases are efficient techniques to decompose discrete signals. They are, however, not suitable for patterns recognition since they suffer from the lack of translation invariance. In Chapter 4 and Chapter 5, we present some translation-invariant techniques that are appropriate for extracting the features of PQ disturbances so that we produce efficient and shift-invariant feature vectors for further use in the classification process.

Chapter 4

WAVELET TRANSFORM MODULUS MAXIMA TECHNIQUES

4.1 INTRODUCTION

In the previous chapter, we have seen that the discrete wavelet transforms using orthogonal or biorthogonal bases are very efficient for the decomposition of discrete signals. The existence of the fast decomposition technique and the efficiency make it suitable for computer implementations. However, the discrete wavelet transforms suffer from the lack of translation invariance. When a function is translated, due to the critical down sampling, the transform coefficients associated with a mother wavelet for this function are not translated but are completely modified. This is the major inconvenience of the discrete wavelet transform in pattern recognition applications.

Several methods have been proposed so that they retain the shift-invariance property in their multiresolution representations. Some methods such as auto-correlation shell in [46] or methods that are used in [47,48] obtain the translation invariance by entailing high oversampling rates, in which no down sampling with the changing scales is allowed. The authors in [49,50] propose to modify the wavelet transform and wavelet packet decompositions leading to orthonormal best-basis representations, which characterise signals with lower costs and translation invariance. Approximation methods such as zero crossing [51,52,53] or local maxima [54,55] that critically sample signals at their inflection points, and the local structures of the analysed signal are revised by the evolution across scales of these wavelet transform coefficients at the inflection points. These methods, in particular the local maxima method, can approximate the signal with a very small number of coefficients. Some other methods such as the basis pursuit [56] or the matching pursuit [57,75] are computationally expensive.

In this chapter, we present the WT local maxima technique and show that the technique can efficiently represent a signal with its shift-invariant coefficients. Also, the technique has the ability to precisely measure transient power quality disturbance characteristics.

4.2 THE ISSUE OF TRANSLATION INVARIANCE

Although they provide a non-redundant multiresolution representation of a signal, and can be computed very efficiently using the pyramid algorithm, the orthogonal and the biorthogonal WT suffer the major problem of not being translation invariant [58,59,60].

To demonstrate this point, let us consider an example of four-level orthogonal WT decomposition using a Dauberchy-4 (D4) wavelet. Figure 4.1 shows the input signal that is chosen to be one of the wavelet basis functions in the third subband. Then its WT, which is plotted in

Figure 4.2, has a single impulse in the third subband, while other wavelet coefficients in the transform are zero. Now if the signal is shifted by one sample to the right, we obtain a completely different set of wavelet coefficients in the transform. They spread broadly across the subband shown in Figure 4.3. This behaviour is clearly not suitable for pattern recognition applications.

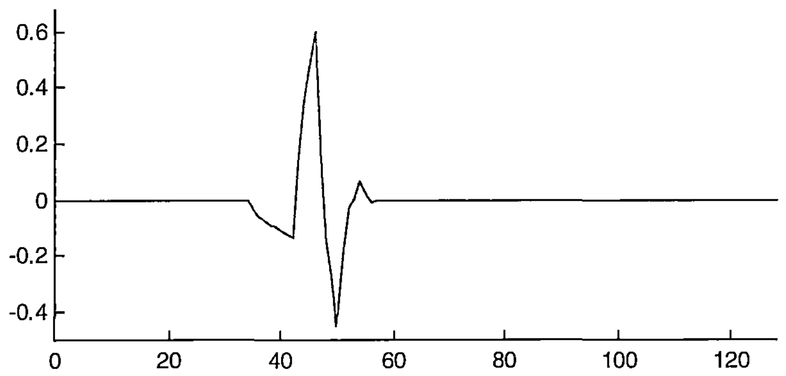


Figure 4.1: A D4 wavelet basis functions in the third subband.

Consider the orthogonal or biorthogonal WT schemes shown in Figure 3.6. The signal is first filtered by a lowpass filter $\bar{h}(n)$ and a highpass filter $\bar{g}(n)$, thus dividing the signal spectrum into two parts. The lowpass signal and the highpass signal are down sampled by 2. It is well known that sharp cutoff filters require very high order and are highly sensitive to quantisation and often cause instability problems in IIR filters. Therefore, in order to cover completely the signal frequency band, the responses of $\bar{h}(n)$ and $\bar{g}(n)$ should not be bandlimited but overlapped. This is the reason why the orthogonal and biorthogonal WT are not translation invariant [61].

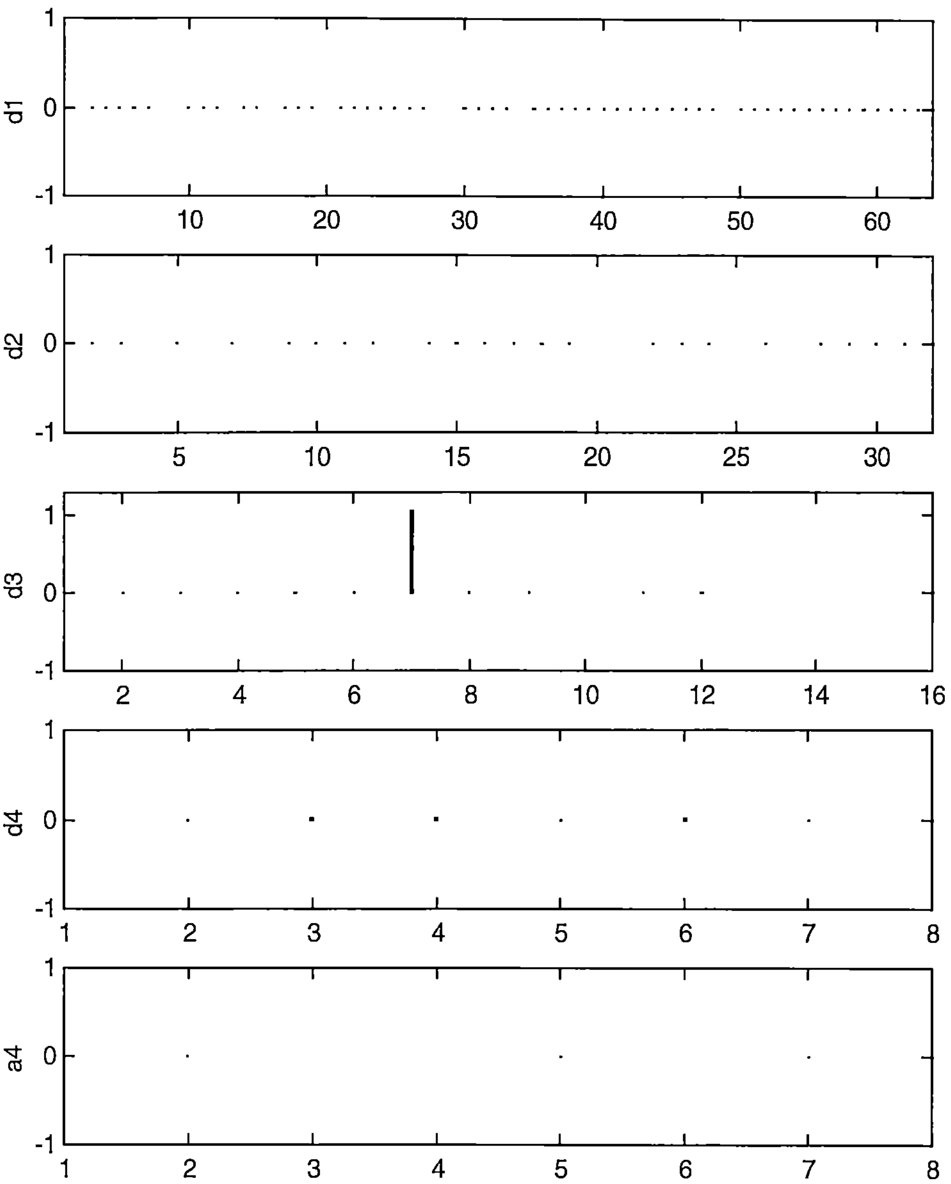


Figure 4.2: The four-level wavelet coefficients of the signal in Figure 4.1.

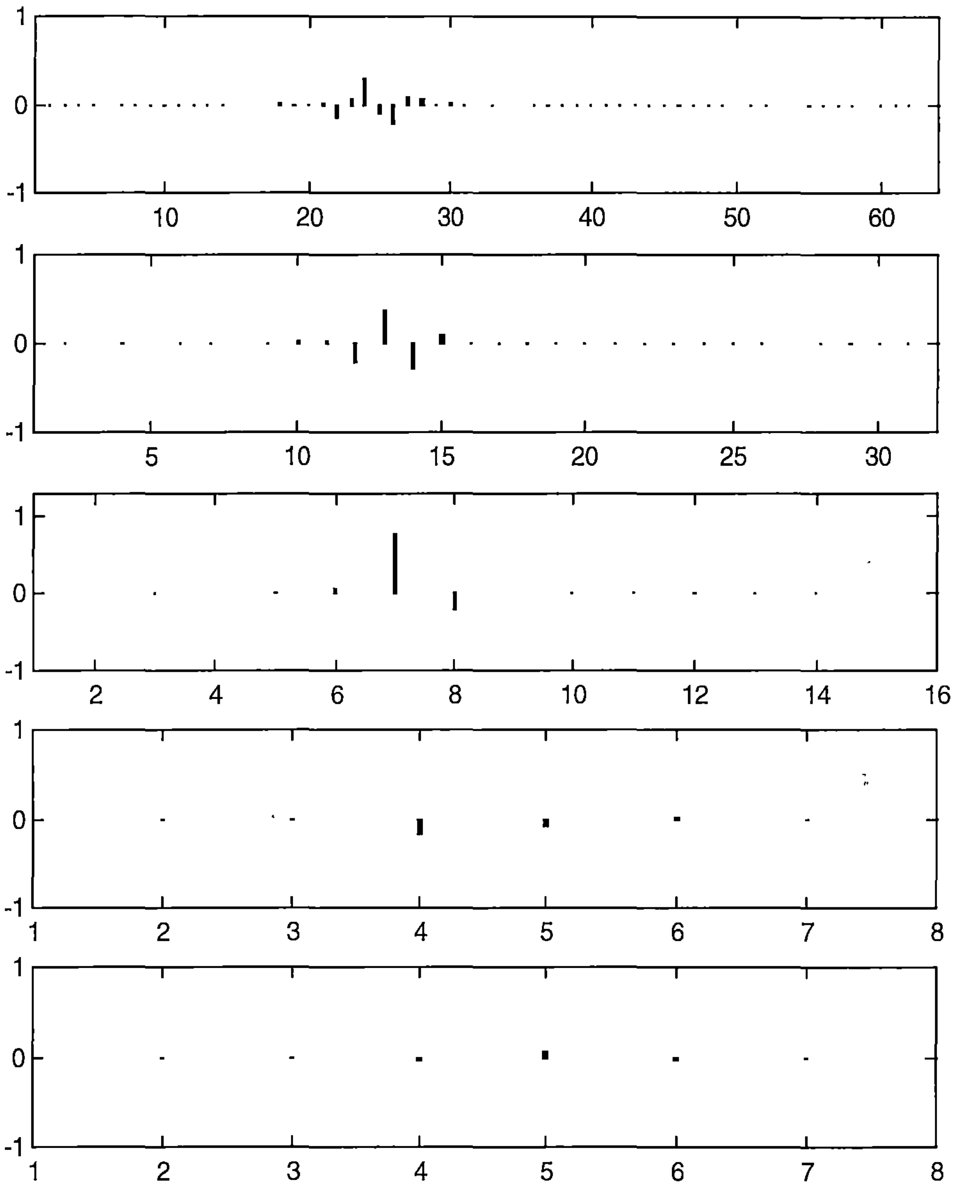
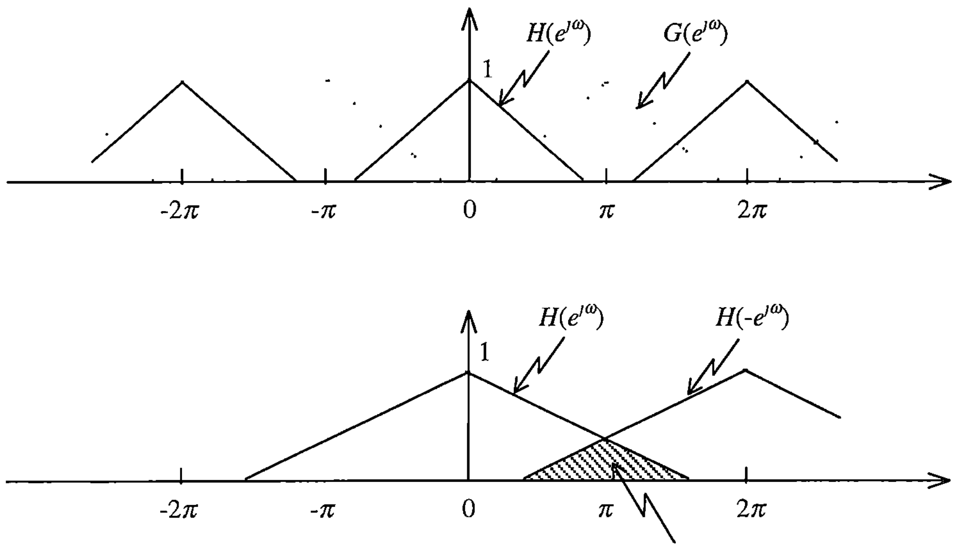


Figure 4.3: The four-level wavelet coefficients of the signal in Figure 4.1 shifted by one sample.

To demonstrate this point, let us consider the lowpass filter $\overline{h}(n)$ that is not bandlimited to [-π/2, π/2] and the highpass filter $\overline{g}(n)$ that is not bandlimited to [π/2, 3π/2]. The subsampling by 2 on the lowpass and highpass signals corresponds to a stretching by 2 of their digital spectrums (i.e. ωT radians): $\overline{H}(e^{j\omega/2})$ and $\overline{G}(e^{j\omega/2})$. In fact, the sub-sampling by 2 produces two terms. The first term is $\overline{H}(e^{j\omega/2})$ or $\overline{G}(e^{j\omega/2})$, and the second term is its frequency shifted version by 2π, i.e. $\overline{H}(-e^{j\omega/2})$ or $\overline{G}(-e^{j\omega/2})$. Since the two filters are not properly bandlimited, those two terms of each filter introduce aliasing and prevent the signal from being reconstructed exactly from them. This is shown in Figure 4.4 below. Note that perfect reconstruction of orthogonal or biorthogonal WT can only be obtained by choosing the appropriate reconstruction filters so that the aliasing is cancelled out during reconstruction.



Region of overlap,
prevents recovery of $H(e^{j\omega})$.
Similar for the recovery of $G(e^{j\omega})$

Figure 4.4: The subsampling by 2 cause aliasing and prevents recovery of the original signals

One can see that the DWT schemes above can only achieve the translation invariance by keeping all coefficients without down sampling the signals and obtaining an overcomplete representation. In this way, the transforms retain all the shift versions of the signal. This representation is, however, not efficient and very redundant.

4.3 CHARACTERISATION OF LOCAL REGULARITY WITH WAVELET TRANSFORM

Singular structures in a signal such as edges and discontinuities contain a lot of information, which are usually used to characterise the signal. In order to characterise singular structures, it is necessary to specifically measure the local regularity of the signal. Following in this section, we study the *Lipschitz exponents* of the signal that provide uniform regularity and measurements over time intervals, and also at any point of the signal. A remarkable property of the wavelet transform is its ability to characterise the local regularity so that the local Lipschitz exponents can be measured from the decay of the wavelet transform magnitude at fine scales [54,62].

4.3.1 Lipschitz Definition

Let n be a positive integer and $n \leq \alpha \leq n + 1$. A function $f(t)$ is said to be pointwise Lipschitz α at τ, if there exists a constant $K > 0$ and a polynomial p_τ of degree n such that [54,55]

$$|f(t) - p_\tau(t)| \leq K|t-\tau|^\alpha \ , \forall t \in R \tag{4.1}$$

- A function $f(t)$ is uniformly Lipschitz α over a time interval $[a,b]$ if it satisfies (4.1) for all $\tau \in [a,b]$, with a constant K that is independent of τ.

- The Lipschitz regularity of $f(t)$ at $\tau \in [a,b]$ is the superior bound of all values α such that $f(t)$ is Lipschitz α.

From (4.1), one can clarify that the polynomial $p_\tau(t)$ is uniquely defined at each point τ. In fact, if $f(t)$ is n times continuously differentiable over $[a,b]$, $p_\tau(t)$ can be shown to be the first $n + 1$ terms of the Taylor Series expansion of f(t) at τ, i.e.

$$p_\tau(t) = \sum_{k=0}^{n} \frac{f^{(k)}(\tau)}{k!}(t-\tau)^k \tag{4.2}$$

At a point τ, the larger the α the more regular the signal is at that point. A classical tool for measuring the Lipschitz regularity of a function $f(t)$ is to look at the asymptotic decay of its Fourier transform $F(\omega)$ [54]. A bounded function $f(t)$ is uniformly Lipschitz α over R if it satisfies

$$\int_{-\infty}^{+\infty} |F(\omega)|(1+|\omega|^\alpha)d\alpha < \infty \tag{4.3}$$

The FT gives a global regularity condition over the whole real line but we cannot determine whether the function is locally more regular at a particular point τ. This is because the FT basis functions do not provide any localisation in the time domain. In contrast, the wavelet transforms are well localised in time, and that they provide a measurement of the Lipschitz regularity over any interval and at any point in the signal.

4.3.2 Characterisation of Local Regularity with WT

To measure the local regularity of a signal, the wavelet vanishing moments are crucial. If the wavelet has n vanishing moments then the wavelet transform can be interpreted as a multiscale differential operator of order n [27].

The approximation error of the Taylor series $p_\tau(t)$ in (4.2) to the signal in a neighbourhood of τ, $\varepsilon_\tau(t) = f(t) - p_\tau(t)$, satisfies the Lipschitz property (4.1) that

$$|\varepsilon_\tau(t)| \leq K|t-\tau|^\alpha \tag{4.4}$$

The purpose is to estimate the Lipschitz exponent α using wavelet transform while ignoring the polynomial $p_\tau(t)$. The wavelet must have $n > \alpha$ vanishing moments, so that it is orthogonal to polynomials of degree less than n [27].

$$\int_{-\infty}^{+\infty} t^k \psi(t)dt = 0 \quad \text{for } 0 < k < n \tag{4.5}$$

As $\alpha < n$, the polynomial $p_\tau(t)$ has a degree at most $(n - 1)$. This implies that its WT with the change of variable $t' = (t - u)/s$ is

$$Wp_\tau(s,u) = \int_{-\infty}^{+\infty} p_\tau(t)\frac{1}{\sqrt{s}}\psi\left(\frac{t-u}{s}\right)dt = 0 \tag{4.6}$$

Then the WT of $f(t)$ in the neighbour of τ becomes the WT of the approximation error, that is

$$Wf(s,u) = W\varepsilon_\tau(s,u) \tag{4.7}$$

If the wavelet has a compact support, the wavelet transform $Wf(s,\tau)$ depends upon the value of $f(t)$ in the neighbourhood of τ, of size proportional to the scale s. When s is small, it provides a localised characteristic of $f(t)$. Suppose the wavelet $\psi(t)$ has n vanishing moments and it is therefore n times continuously differentiable. Then from (4.4) and (4.7), it is shown in [54,55] that if $f(t)$ is uniform Lipschitz $\alpha < n$ over $[a,b]$, then there exists a constant $A > 0$ such that

$$|Wf(s,\tau)| \le As^{\alpha+1/2} \tag{4.8}$$

Conversely, if $f(t)$ is bounded and $|Wf(s,\tau)|$ satisfy (4.8) for an $\alpha < n$, then $f(t)$ is uniformly Lipschitz α on $[a+\varepsilon, b\text{-}\varepsilon]$, for any $\varepsilon > 0$.

Equation (4.8) provides a condition on the asymptotic decay of $|Wf(s,\tau)|$ when s goes to zero. It also shows that $|Wf(s,\tau)|$ decays like $s^{\alpha+1/2}$ over intervals where $f(t)$ is uniformly Lipschitz α. If $f(t)$ is uniformly Lipschitz $\alpha > n$, the decay of $|Wf(s,\tau)|$ at fine scales gives no information about the Lipschitz regularity α since this decay can only be a maximum of $s^{n+1/2}$.

In practice, measuring the decay of the absolute value of the CWT in a whole time-scale plane (s,τ) is not necessary. In fact, this decay can be controlled from its local maxima values [54]. The following section represents the WTMM technique.

4.4 WT MODULUS MAXIMA

Often in transient signals or images, points of sharp variations are among the most important features for analysing their properties. The concept of multiscale edge detection has been used to detect the contours of small structures as well as a larger object in image [54,55,63,64]. Most multiscale edge detectors smooth the signal at various scales and detect sharp variation points from their first derivatives. The extrema from the first derivative correspond to the inflection points of the smoothed signal.

The WT at a point (s_0,t_0) is a *modulus maximum* if $|Wf(s_0,t)|$ is a strict local maximum at t_0, i.e.

$$\frac{\partial Wf(s_0,t)}{\partial t}=0 \quad \text{at } t=t_0 \tag{4.9}$$

Clearly, a translation in a signal results in the same translation of its WTMM representation.

4.4.1 Detection of Singularities

Consider a wavelet that is the n^{th} order derivative of a smoothing function. When $n =$ 1, i.e. the wavelet is the first order derivative of a smoothing function, the local extrema of the CWT characterise sharp variations or edges in a signal. Equivalently, when $n = 2$ the wavelet is the second derivative of a smoothing function, and the zero-crossings of the CWT express the same information [54,55].

We call a smoothing function any real function $\theta(t)$ such that $\theta(t) = O(1/(1+t^2))$ and whose integral is equal to 1. Gaussian function is an important smoothing function and is often used in many applications. Suppose that $\theta(t)$ is twice differentiable and its first and second order derivatives are, respectively,

$$\psi^{(1)}(t) = \frac{d\theta(t)}{dt} \quad \text{and} \quad \psi^{(2)}(t) = \frac{d^2\theta(t)}{dt^2} \tag{4.10}$$

Such function $\psi^{(1)}(t)$ and $\psi^{(2)}(t)$ can be considered as wavelets since they satisfy the admissibility condition in (3.20). The WT of $f(t)$ with respect to each of these wavelets can be written as convolutions and are defined by

$$\begin{aligned} W^{(1)}f(s,t) &= f * \psi_s{}^{(1)}(t) \\ W^{(2)}f(s,t) &= f * \psi_s{}^{(2)}(t) \end{aligned} \tag{4.11}$$

Let the scale version of the smoothing function be $\theta_s(t) = \frac{1}{\sqrt{s}}\theta\left(\frac{t}{s}\right)$, (4.11) can be rewritten as

$$\begin{aligned} W^{(1)}f(s,t) &= f * \left(s\frac{d\theta_s}{dt} \right)(t) \\ &= s\frac{d}{dt}(f * \theta_s)(t) \end{aligned} \tag{4.12}$$

and

$$\begin{aligned} W^{(2)}f(s,t) &= f * \left(s^2\frac{d^2\theta_s}{dt^2} \right)(t) \\ &= s^2\frac{d^2}{dt^2}(f * \theta_s)(t) \end{aligned} \tag{4.13}$$

The WT $W^{(1)}f(s,t)$ and $W^{(2)}f(s,t)$ are therefore, respectively, the first and second derivative of the signal $f(t)$ smoothed by $\theta_s(t)$. Hence the local extrema of $W^{(1)}f(s,t)$ correspond to the zero-crossing of $W^{(2)}f(s,t)$ and to the inflection points of $f*\theta_s(t)$. For small scales s, the smoothing of $f(t)$ by $\theta_s(t)$ is negligible and the edge detection provides the locations of all locally sharp variations of $f(t)$. While at large scales s, the convolution of $f(t)$ with $\theta_s(t)$ removes small fluctuations and only detects larger amplitude variations of $f(t)$.

In the WT extrema representation, the magnitude of extrema and their corresponding locations are recorded. For an exact equivalence, the WT zero-crossings representation requires the positions of the zero-crossings as well as the integral values of the function between two zero-crossings. In fact, these integral values are needed in order to stabilise the zero-crossings representation [51]. As pointed out in [55], the WT extrema approach has several important advantages over the WT zero-crossing approach. The local maxima of the absolute value of $W^{(1)}f(s,t)$ are sharp variation points of $f*\theta_s(t)$, whereas the zero-crossing of $W^{(2)}f(s,t)$ are either sharp variation points or low variation points of $f*\theta_s(t)$, which can be shown as in Figure 4.5. As we are only interested in sharp variations in the signal, we thus only retain the values of the local extrema of $W^{(1)}f(s,t)$ and their corresponding locations. We call this representation the WTMM representation.

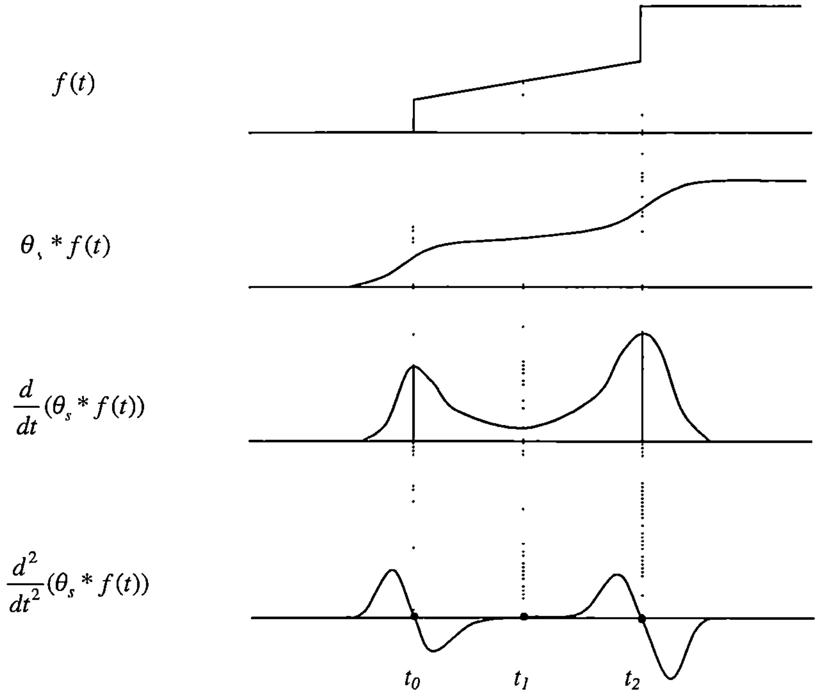

$f(t)$
$\theta_s * f(t)$
$\frac{d}{dt}(\theta_s * f(t))$
$\frac{d^2}{dt^2}(\theta_s * f(t))$
t_0
t_1
t_2

Figure 4.5: WT modulus maxima $W^{(1)}f(s,t)$ and WT zero-crossings $W^{(2)}f(s,t)$ of a two-steps signal $f(t)$.

If the wavelet used is the first derivative of a smoothing function, it has only one vanishing moment. A wavelet with more vanishing moments has the advantage of being able to measure the Lipschitz regularity up to a higher order, but it also increases the number of maxima lines. The number of maxima at a given scale often increases linearly with the number of moments of the wavelet. In order to minimize the amount of computation, we want to have the minimum number of maxima necessary to detect the interesting irregular behaviour of the signal. This means that we must choose a wavelet with as few vanishing moments as possible, but with enough moments to detect the Lipschitz exponents of highest order that we are interested in [54]. In power signals, transients and high frequency disturbances have negative or small positive values for their Lipschitz exponents [65,66]. Therefore, to minimize the amount of computation, we use a wavelet with only one vanishing moment. A very popular and efficient wavelet in this category is a *Quadratic Spline*, which is the first derivative of a *Cubic Spline* function, which is shown in Figure 4.6. Then the local maxima of the wavelet transform $|Wf(s,t)|$ give the locations of the sharp variation points in the signal $f(t)$.

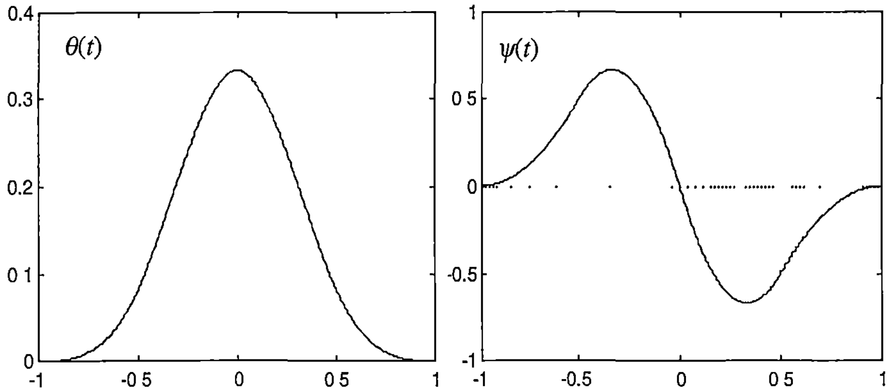


Figure 4.6: Graph of the smoothing cubic spline function $\theta(t)$ and the corresponding quadratic spline wavelet $\psi(t)$.

4.4.2 Completeness of the WTMM

WTMM technique is a 'loose' representation of signals. We might wonder how much information is carried by the position of the local maxima of $|Wf(s,t)|$ and the value of $Wf(s,t)$ at the corresponding location. Different methods for reconstruction of a signal from its WTMM have been proposed in [55,67,68]. To obtain an efficient numerical implementation, modulus maxima are detected only along a dyadic sequence of scales (i.e. $s = 2^j$, $j \in Z$). Reconstruction methods recover an approximation of the original signal with a signal to noise ratio of the order of 40dB [55,70]. This indicates that the WTMM provide a 'nearly complete' characterisation of signals, and that small errors, which mostly concentrate at high frequency, remain to be identified mathematically [55]. Following in this section, some properties of a dyadic WT, as well as an iterative reconstruction method from the WTMM, are presented.

The dyadic WT of a function $f(t)$ with respect to $\psi(t)$ at scale 2^j is defined as

$$Wf(2^j,t) = f * \psi_j(t) \tag{4.14}$$

or in the Fourier domain as

$$WF(2^j,\omega) = \sqrt{2^j}\,F(\omega)\Psi(2^j\omega) \tag{4.15}$$

To ensure that the signal $f(t)$ can be reconstructed from its wavelet transform, and the reconstruction is stable, there must exist two positive constants A and B such that [21,55]

$$A \le \sum_j \left|\Psi(2^j\omega)\right|^2 \le B \tag{4.16}$$

From (4.16), we can obtain a semi-discrete version of the frame equation of (3.32), in which the constants A and B are interpreted as the lower and upper frame bounds respectively.

$$A\|f\|^2 \le \sum_j \left|Wf(2^j,t)\right|^2 \le B\|f\|^2 \tag{4.17}$$

The reconstruction wavelet is a dual wavelet $\tilde{\psi}(t)$ whose Fourier transform satisfies

$$\sum_{j} \Psi(2^{j}\omega)\tilde{\Psi}(2^{j}\omega) = 1 \tag{4.18}$$

Since the frame (4.17) is redundant, this dual wavelet is not uniquely specified. An example of dual wavelet that satisfies (4.18) is given in [26]

$$\tilde{\Psi}(\omega) = \frac{\Psi^{*}(\omega)}{\sum_{j} |\Psi(2^{j}\omega)|^{2}} \tag{4.19}$$

The function $f(t)$ is recovered from its dyadic WT with

$$f(t) = \sum_{j} Wf(2^{j},\cdot) * \tilde{\psi}_{j}(t) \tag{4.20}$$

Similar to the CWT, the dyadic WT $(Wf(2^{j},t))_{j\in Z}$ is over complete. This means that any sequence of functions $(g_{j}(t))_{j\in Z}$ is not a priori the dyadic WT of a function $f(t) \in$ $\mathbf{L}^{2}(R)$. The space V of all dyadic WT of functions in $\mathbf{L}^{2}(R)$ is in fact be a closed subspace of $l^{2}(\mathbf{L}^{2}(R))$. Any sequence $(g_{j}(t))_{j\in Z}$ that is the dyadic WT of a function in $\mathbf{L}^{2}(R)$ must satisfy the semi-discrete reproducing kernel equation

$$g_{j}(t) = \sum_{l} g_{l} * K_{l,j}(t)\,, \quad \forall j \in Z \tag{4.21}$$

where the semi-discrete reproducing kernel $K_{l,j}(t) = \tilde{\psi}_{l} * \psi_{j}(t)$. (4.21) can be rewritten in term of the dyadic WT operator W and the inverse dyadic WT operator W^{-1} as

$$(g_{j}(t))_{j\in Z} = W(W^{-1}(g_{l}(t))_{l\in Z}) \tag{4.22}$$

In practice the input signal is measured at a finite resolution. The WT is therefore cannot be computed at an arbitrary fine scale. By introducing a smoothing function $\phi(t)$ whose Fourier transform is an aggregation of the wavelets at scales 2^{j} larger than 1,

$$|\Phi(\omega)|^{2} = \sum_{j=1}^{\infty} |\Psi(2^{j}\omega)|^{2} \tag{4.23}$$

The smoothing operator at scale 2^J is defined by

$$Sf(2^J,t) = f * \phi_J(t) \tag{4.24}$$

The function $Sf(2^l,t)$ at any given scale 2^l can be reconstructed from the dyadic WT of $f(t)$ at scales larger than 2^l, $(Wf(2^j,t))_{l<j<+\infty}$. Conversely, the information of the dyadic WT at scales larger than 2^l can be computed from $Sf(2^l,t)$ [55]. The wavelet decomposition is also limited to a finite larger scale 2^J. Then the decomposition of the signal $f(t)$ by the dyadic WT between scale 1 and 2^J is the set of functions $\left\{(Wf(2^j,t))_{1\leq j\leq J}, Sf(2^J,t)\right\}$. The signal can then be reconstructed perfectly from its dyadic WT [55,69].

Since we can obtain an exact reconstruction of a signal $f(t)$ from its dyadic WT, the reconstruction of the signal $f(t)$ from its WTMM is equivalent to the reconstruction of $(Wf(2^j,t))_{1\leq j\leq J}$ from the positions of the local maxima of $(|Wf(2^j,t)|)_{1\leq j\leq J}$ and the value of $Wf(2^j,t)$ at these locations. At a scale 2^j, there is an infinite number of functions $g_j(t)$ which have the same local maxima as $Wf(2^j,t)$. However, any such sequence of functions $(g_j(t))_{j\in Z}$ is not necessarily a dyadic WT of a function in $\boldsymbol{L}^2(R)$. In fact the dyadic WT must satisfy the semi-discrete reproducing kernel equations (4.21).

A common method for reconstructing the signal from WTMM is the projection-based method [55,70]. Let Γ be the set of all sequence of functions $(g_j(t))_{j\in Z}$ that have the same local modulus maxima of $(Wf(2^j,t))_{j\in Z}$ at all scale 2^j. Then $(Wf(2^j,t))_{j\in Z}$ $\in \Gamma$, and the local maxima representation is complete only when the intersection of Γ and the space V of all dyadic WT sequences of functions in $\mathbf{L}^2(R)$ reduces to one element, i.e.

$$\Gamma \cap V = \left\{(Wf(2^j,t))_{j\in Z}\right\} \tag{4.25}$$

There is no mathematical proof of this statement. However, by performing numerical simulations on a large number of signals, Mallat [54,55] showed that it is always possible to reconstruct a signal from its WTMM.

As V is the dyadic wavelet space, the projection operator P_V onto V is the reproducing kernel equation

$$P_V = WW^{-1} \tag{4.26}$$

The projection operator P_V projects orthogonally an element not on V onto an element in V closest to it. Defining a projection operator P_Γ that transforms any sequence $(g_j(t))_{j\in Z}$ not on Γ into a sequence $(h_j(t))_{j\in Z} \in \Gamma$ closest to it with respect to the Sobolev norm [55], so that the norm

$$\left|(\varepsilon_j(t))_{j\in Z}\right|^2 = \sum_j \left(\left\|\varepsilon_j(t)\right\|^2 + 2^{2j} \left\| \frac{d\varepsilon_j(t)}{dt} \right\|^2 \right) \quad \text{is minimised} \tag{4.27}$$

where $(\varepsilon_j(t))_{j\in Z} = (h_j(t) - g_j(t))_{j\in Z}$ is the difference. Beside the minimisation of the Sobolev norm in (4.27), the authors in [67,70] propose some new convex constraints and improve the constraints used in [55], to be imposed onto the reconstruction so as to obtain a better estimate of the true solution. The purpose of using those constrains is to prevent the oscillations that they may have from the minimisation of the Sobolev norm in (4.27), thus avoiding spurious local maxima.

The composite operator P is defined by

$$P = P_\Gamma P_V \tag{4.28}$$

Then any element at the intersection of Γ and V is a fixed point of P. By iterating the operator P, we can compute such a fixed point. The space Γ is not convex but is close to being convex [54,55], and since Γ is an affined space and V is a Hilbert space, it is also shown in [55] that the alternate projections onto the set Γ and the space V converge strongly to the solution. The alternate projections algorithm is illustrated in Figure 4.7.

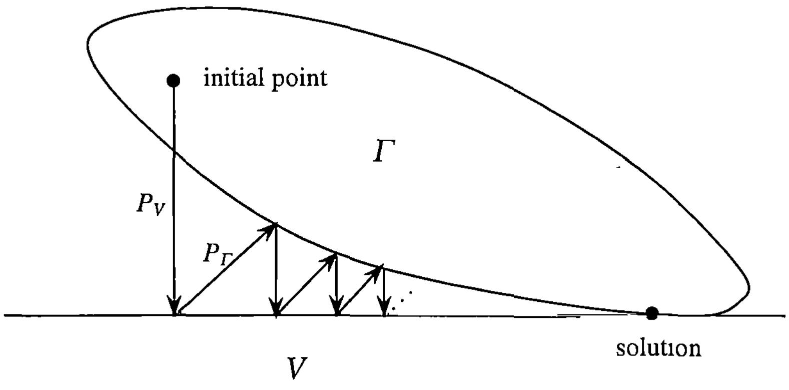


Figure 4.7: The reconstruction of the dyadic WT by alternately projecting operators P_Γ and P_V.

4.5 SINGULARITIES MEASUREMENT USING THE WTMM

This section explains how to measure Lipschitz exponents by using the WTMM. In general, if we want to estimate the Lipschitz exponents up to an order n, then we use a compactly supported wavelet that has n vanishing moments and is n times continuous differentiable.

We can show that for any interval, if the WT has no modulus maxima at fine scales, then the signal is not singular within that interval. We define a *modulus maxima chain* of $Wf(s,t)$ originated from $t = t_0$, being the path followed by the evolution across scales of the modulus maximum of $Wf(s,t)$ due to an inflection point at t_0. However, not all singularities of a signal $f(t)$ can be characterised by following the maxima chains to fine scales. This is the case when the signal $f(t)$ has fast oscillations. Then singularity at a point on $f(t)$ can be influenced by these fast oscillations within its neighbourhood [54]. In the following, we study the characterisation of singularities when locally the signal has no oscillations, and when it contains fast oscillations.

4.5.1 Non-Oscillating Singularities

From the mathematical roots in the characterisation of Sobolev spaces in 1930s, the study of the pointwise Lipschitz exponential has been a delicate topic for wavelet transform. Assume that the wavelet $\psi(t)$ has a compact support equal to $[-C,C]$. Then the *cone of influence* of a point τ in the scale-space plane is the set of points (s,t) that is included in the support of $\psi_{s,\tau} = s^{-1/2}\psi((t-\tau)/s)$, i.e.

$$|t-\tau| \leq Cs \tag{4.29}$$

Suppose that the wavelet has n vanishing moments and is n times continuous differentiable. If the signal $f(t)$ has an isolated singularity at point t_0 with a Lipschitz exponent $\alpha < n$, then the WTMM of that singularity for all scales s less than some scale $s_0 > 0$ belong to the cone of influence of t_0 satisfying (4.8), i.e.

$$|Wf(s,t)|_{\max} \leq As^{\alpha+1/2} \tag{4.30}$$

where A is a positive constant. Equation (4.30) is equivalent to

$$\log(|Wf(s,t)|_{\max}) \leq \log(A) + (\alpha + 1/2)\log(s) \tag{4.31}$$

Hence when the WTMM belongs to the cone of influence of the isolated singularity at t_0, (4.31) proves that the Lipschitz exponent α at t_0 can be measured from the maximum slope of straight lines remaining above $\log(|Wf(s,t)|_{\max})$.

4.5.2 Oscillating Singularities

If the signal $f(t)$ is oscillating quickly in the neighbourhood of an inflection point t_0, the whole singularity behaviour at t_0 is dominated by these oscillations. In this case, we cannot characterise correctly the Lipschitz regularity of the signal $f(t)$ at point t_0 from its wavelet transform within a cone which is strictly smaller then the cone of influence of t_0. In fact, below the cone of influence (i.e. at scales larger than s_0), the singularity behaviour and size of the oscillations of the signal $f(t)$ can be estimated [54].

Along each modulus maxima chain, there is a *general modulus maximum* corresponding to the scale at which the analysis wavelet and the signal are in 'resonance'. Thus the general modulus point is the strongest maxima point in the chain. It provides information on the extent of oscillation and the local frequency of the signal $f(t)$. Since all modulus maxima chains of all inflection points have their roots at the finest scale, the detection of modulus maxima chains starts from the modulus maxima of the first scale. At different scales, modulus maxima of $Wf(s,t)$ of the same inflection point t_0 lie within the cone of influence of t_0. However, as the signal has fast oscillation, cones of influence of different inflection points can overlap. In practice when the signal is in discrete form, the WT decomposition starts from the first scale $s = 1$ and the scale step Δs cannot be set to an arbitrarily small value. Then in order to detect or to follow correctly a modulus maxima chain that originate from an inflection point, the modulus maxima that belong to this chain must satisfy the following conditions [71]

- Starting from the first scale, a chain has to be continuous across scales. This means for any two adjacent scales, positions of the modulus maxima belonging to this chain are not different by more than 1 sample.

- $Wf(s,t)$ has the same sign across scales.

- Since the quadratic spline wavelet is the first derivative of a smoothing function, it can only measure the Lipschitz exponent α within a range [-1,1]. Therefore, from (4.31) the modulus maxima value of $Wf(s,t)$ at a scale $s+\Delta s$ ($|Wf_{s+\Delta s}|$) and that at its previous scale s ($|Wf_s|$) are related as

$$-1+1/2 \le \frac{\log|Wf_{s+\Delta s}| - \log|Wf_s|}{\log(s+\Delta s) - \log s} \le 1+1/2$$

or

$$\frac{s}{2(s+\Delta s)} \le \frac{|Wf_{s+\Delta s}|}{|Wf_s|} \le \frac{3(s+\Delta s)}{2s} \tag{4.32}$$

The general modulus maxima (GMM) of $Wf(s,t)$ at a point $(s_{\max}, t_{\max})$ is a strict local maximum of $|Wf(s,t)|$ within the two-dimensional neighbourhood in the scale-space plane (s,t). If a wavelet is equal to a derivative of a Gaussian, there will be only one GMM of $Wf(s,t)$ along each modulus maxima chain [54,55]. This general modulus maximum represents the best match between the wavelet at the scale $s_{\max}$ and translation $t_{\max}$, and part of the signal in the neighbourhood of t_0. Then the scale $s_{\max}$ is related to the local frequency of the signal by

$$s_{\max} = \frac{\omega_m}{\omega_0} \tag{4.33}$$

where ω_m is the frequency that $|\Psi(\omega)|$ reaches its maximum i.e. the spectral peak of $\Psi(\omega)$, and ω_0 is the estimated signal frequency in the neighbourhood of t_0. We can thus estimate the frequency ω_0 from the value $s_{\max}$ at which the general modulus maximum occurs.

For the quadratic spline wavelet, its FT and the FT of its corresponding lowpass function are given by [55]

$$\Psi(\omega) = i\omega\left(\frac{\sin(\omega/4)}{\omega/4}\right)^4 \text{ and } \Phi(\omega) = \left(\frac{\sin(\omega/2)}{\omega/2}\right)^3 \tag{4.34}$$

which are shown in Figure 4.8. From (4.34), we obtain the maximum of the wavelet frequency spectrum in terms of the sampling frequency ω_s as

$$\omega_m = \frac{3.37\omega_s}{2\pi} \tag{4.35}$$

The GMM position $(s_{\max},t_{\max})$ gives a local frequency approximation of the signal $f(t)$ in the neighbourhood of t_0. Instead of only keeping general points that give the estimation of local frequencies at t_0, instantaneous frequencies at any location τ in the time axis can be estimated by finding some scales $s_{\max}$ such that $|Wf(s_{\max},\tau)|$ are maximum across the scale axis. The technique of finding these instantaneous frequencies is known as the '*wavelet ridges*' technique [72,73]. However, since a signal is completely characterised by its WTMM, we are only interested in the information that is given from the GMM positions. In this way, the technique provides a very compact time-invariant set of features that allows us to analyse the transient characteristics of signals.

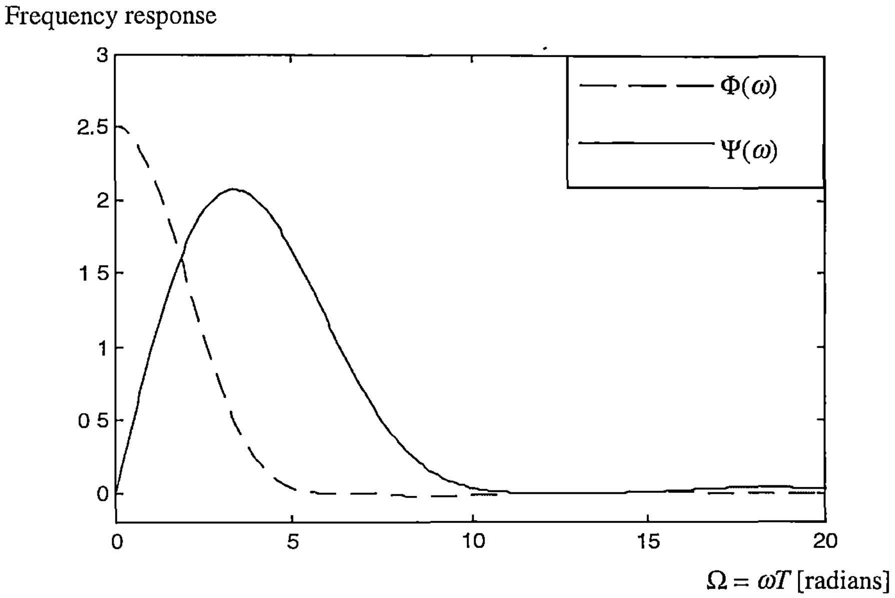


Figure 4.8: Frequency spectrums of the quadratic spline wavelet and its corresponding lowpass function.

4.6 NUMERICAL DEMONSTRATIONS

As mentioned in Chapter 2, transient power quality disturbances such as impulse transient, high frequency capacitor switching, and low frequency capacitor switching contain fast oscillations. Hence, their singularities and the information on their oscillation frequencies can be measured from their WTMM as well as from their GMM [66,71,139].

Noise in power systems has a typical voltage magnitude of less than 1% of the power signal. In the following examples, we use the maximum noise level of 1% voltage magnitude. Then, if a disturbance has a voltage magnitude of 0.1pu, the noise presents in the disturbance with a relative voltage magnitude of 10%. Using the quadratic spline wavelet for the calculation of the WT, in Figure 4.9 we show the

WTMM and the GMM representation of an impulse transient disturbance. Figure 4.10 is for a low frequency capacitor switching disturbance, Figure 4.11 is for a high frequency capacitor switching disturbance, and Figure 4.12 is for a single notch disturbance. Shown in figure (a), each disturbance signal has a duration of 20ms sampled at 12.8KHz (i.e. 256 samples). The wavelet transforms $Wf(s,n)$ are calculated for 30 scales $s = s_0^j$, where $s_0 = 1.1$ and $j = 0$ to 29.

The location of sharp transients in the signals can be seen clearly at the small scales of the wavelet transform $Wf(s,n)$ in each figure (b). The position of modulus maxima of $Wf(s,n)$ are located with a threshold of 1% of the maximum modulus maxima on each scale. Modulus maxima chains are then detected and shown in each figure (c). Finally the position of GMM on each chain is located and shown in each figure (d).

The modulus maxima at small scales of a noisy low frequency disturbance are predominantly due to noise. This is because the wavelet coefficients at small scales are less sensitive to low frequency components while more sensitive to fast changing components. But as the scale increases, modulus maxima that are due to those low frequency components of the disturbance get larger and those due to noise get smaller and fewer. Therefore, at scale $s_{\max}$ a small threshold would eliminate completely the modulus maxima due to noise.

On the other hand, for disturbances containing very high frequencies (e.g. impulse transients), their modulus maxima have the highest values at small scales and are reduced at higher scales. This is due to their negative Lipschitz exponential (which is similar to noise). However, in the time domain the magnitude of noise is much smaller than the disturbances. Hence the GMM of noise are also much smaller than those produced by high frequency disturbances. Therefore, we remove small GMM as they are produced either by noise or by disturbance components with small amplitudes.

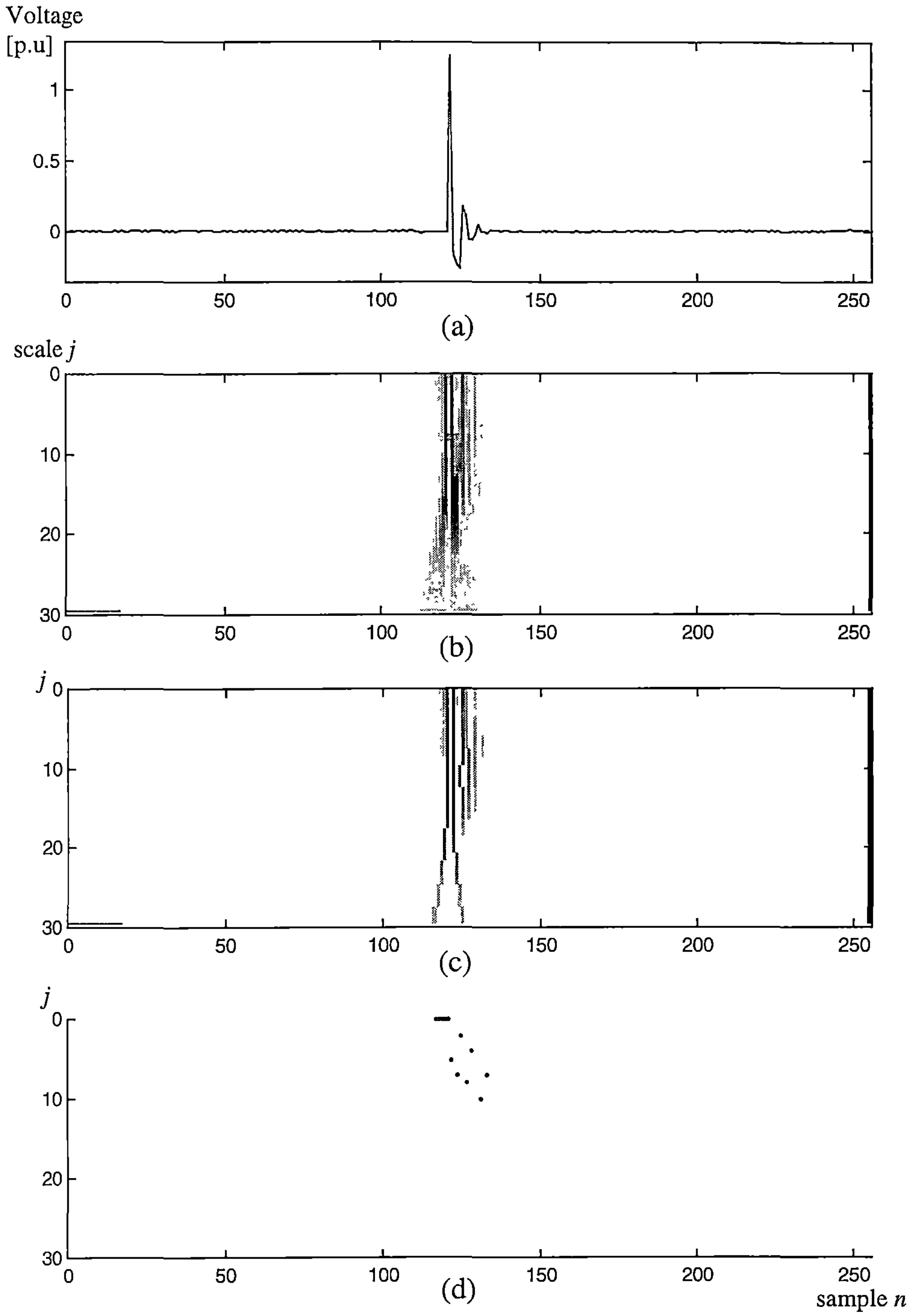


Figure 4.9: (a) 256 samples of a impulse transient disturbance *f*(*n*). (b) The wavelet transforms $Wf(s,n)$, $s = s_0^j$, $s_0 = 1.1$, $j = 0{:}29$. (c) Modulus maxima chains. (d) General modulus maxima representation of *f*(*n*).

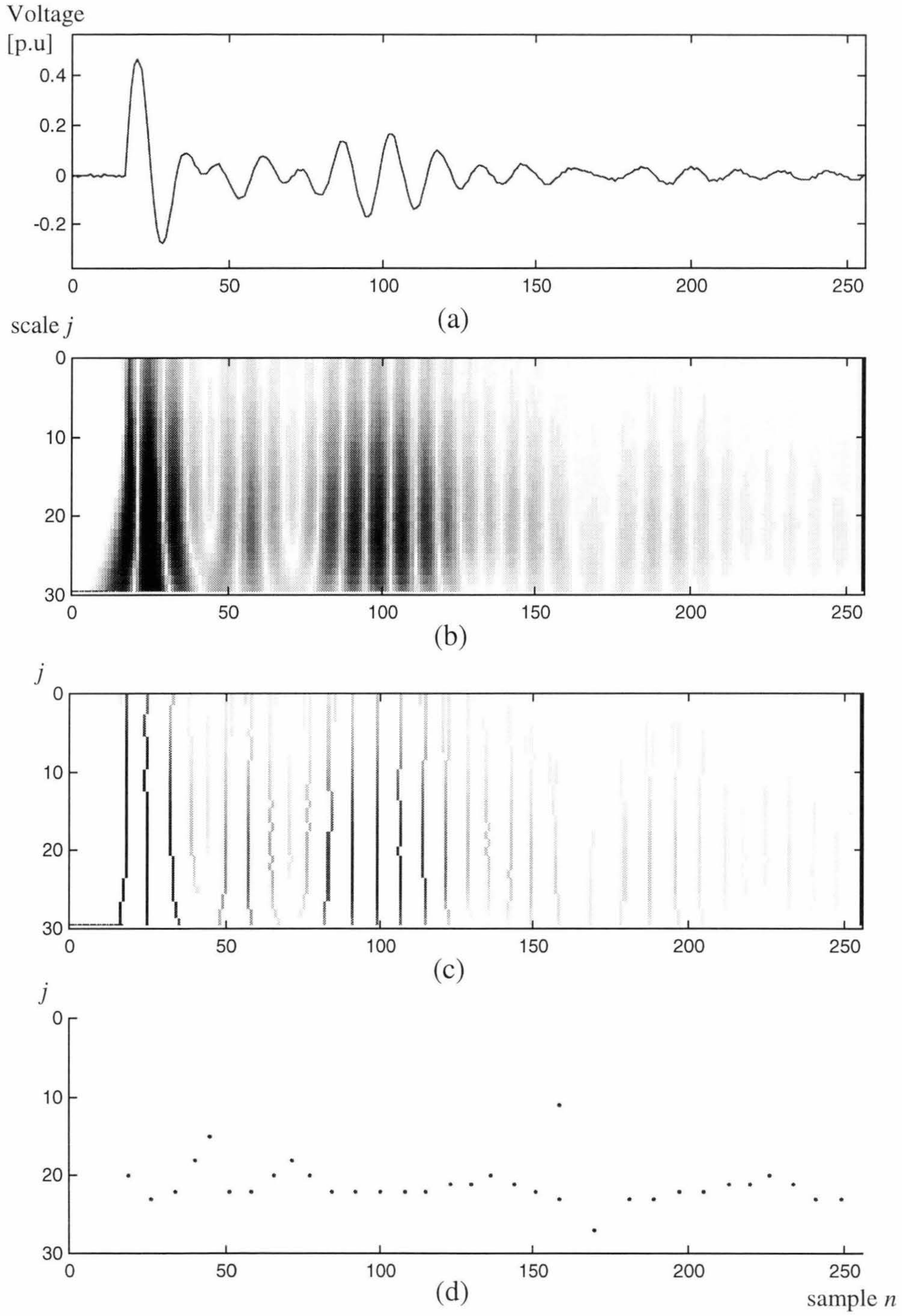


Figure 4.10: (a) 256 samples of a low frequency capacitor switching disturbance $f(n)$. (b) The wavelet transforms $Wf(s,n)$, $s = s_0^j$, $s_0 = 1.1$, $j = 0{:}29$. (c) Modulus maxima chains. (d) General modulus maxima representation of $f(n)$.

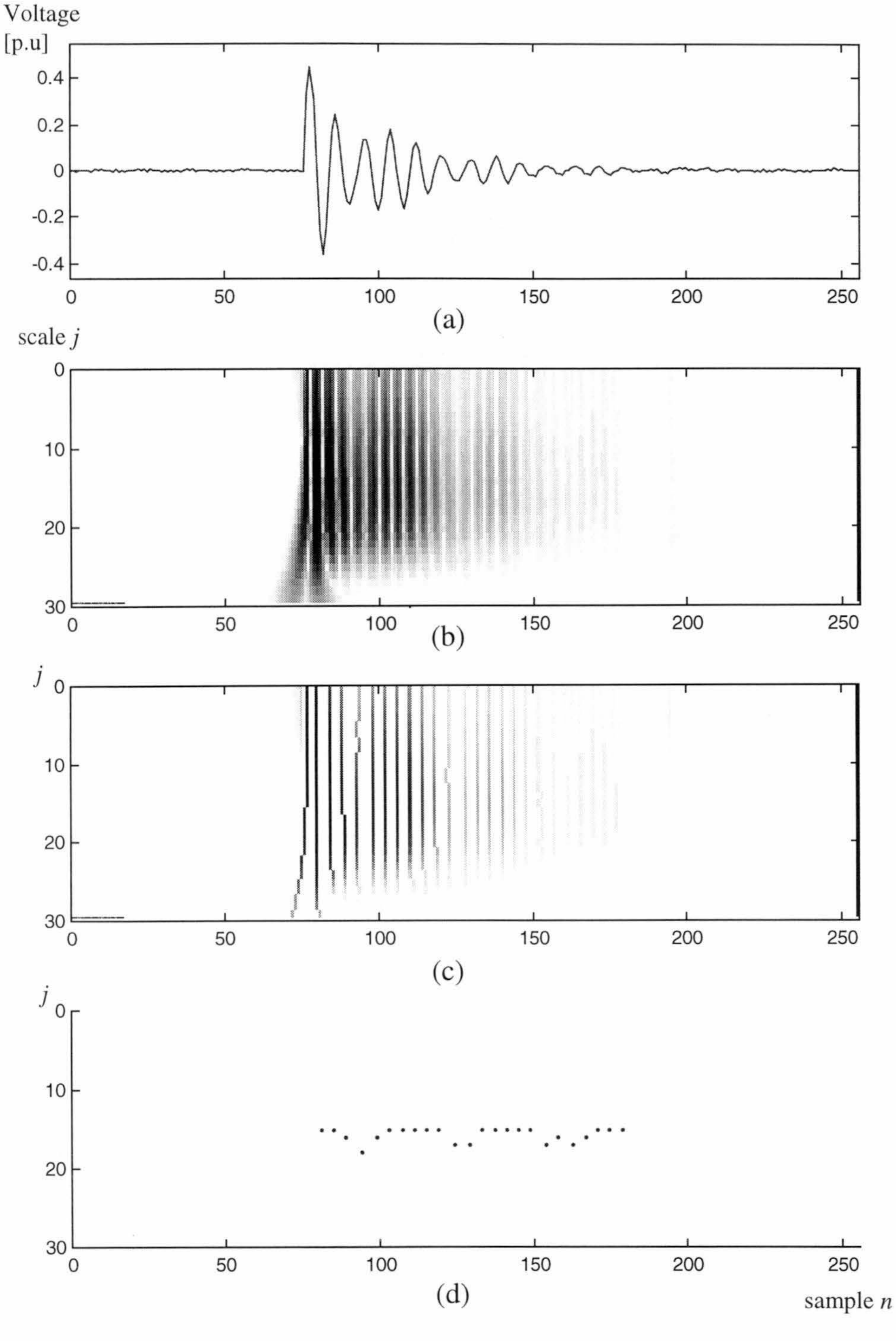


Figure 4.11: (a) 256 samples of a high frequency capacitor switching disturbance $f(n)$. (b) The wavelet transforms $Wf(s,n)$, $s = s_0^j$, $s_0 = 1.1, j = 0{:}29$. (c) Modulus maxima chains. (d) General modulus maxima representation of $f(n)$.

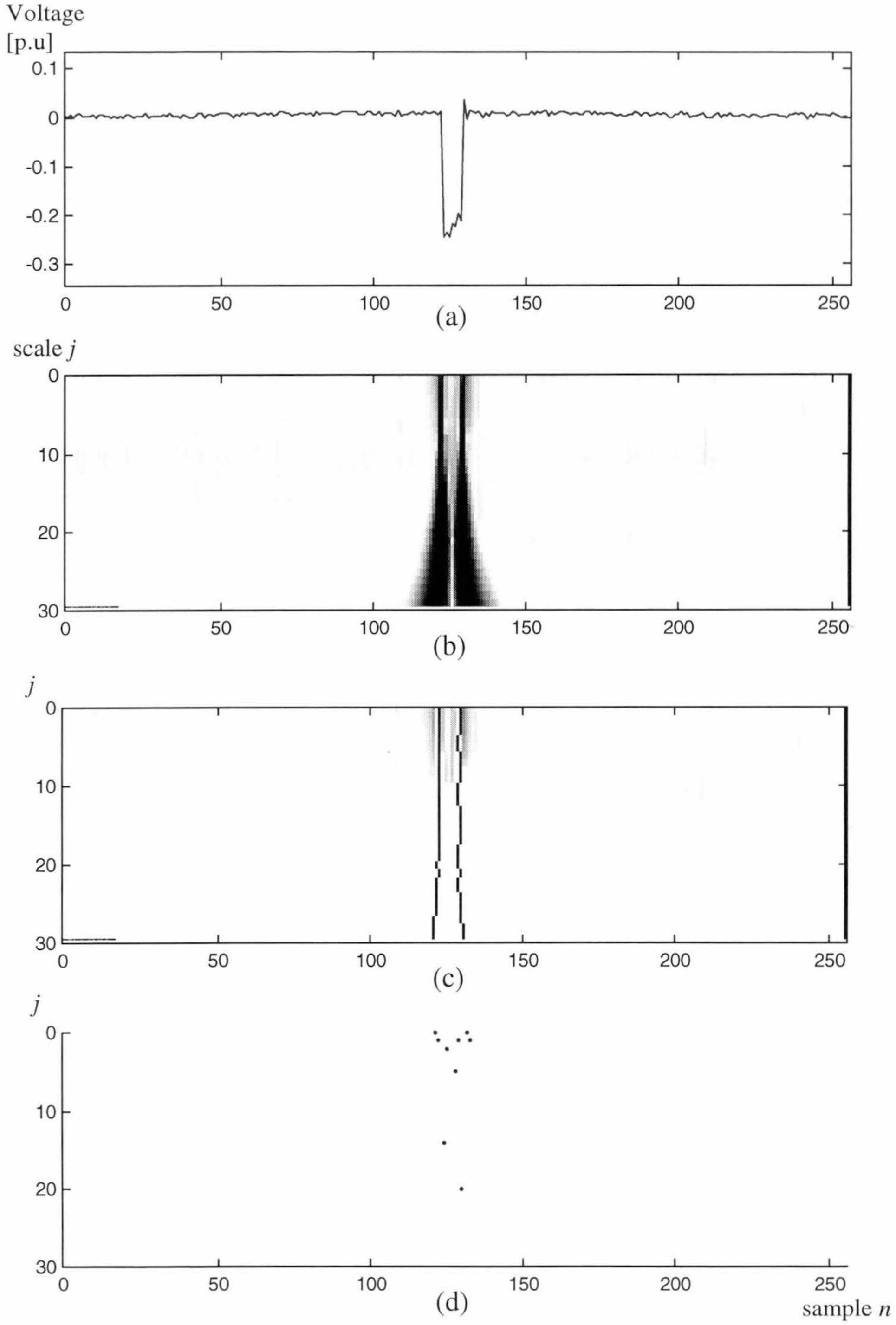


Figure 4.12: (a) 256 samples of a single notch disturbance *f*(*n*). (b) The wavelet transform *Wf*(*s*,*n*), $s = s_0^j$, $s_0 = 1.1$, $j = 0{:}29$. (c) Modulus maxima chains. (d) General modulus maxima representation of *f*(*n*).

If the sampling frequency is limited, impulse transient disturbances, including single impulses that contain very fast oscillations, are usually characterised by negative Lipschitz exponents at small scales. As a result, their GMM appear at small scales. On the other hand, capacitor switching disturbances have slower oscillations which are characterised by positive Lipschitz exponents at small scales. Hence, their GMM appear at larger scales. When we come to notch disturbances, the characteristics of this disturbance type are not fast oscillations but are discontinuous, in which the discontinuities at the starting and ending of notches are normally characterised by small values of Lipschitz exponents ($\alpha \approx 0$). This makes their GMM positions unstable and sensitive to noise. Their classification rate is thus not generally high compared to the other disturbance types when using a conventional RBF classifier. This will be illustrated later in Chapter 8 of this thesis.

Chapter 5

MATCHING PURSUIT TECHNIQUE

5.1 INTRODUCTION

Linear expansion in a single basis, whether it is a Fourier, wavelet, or any other basics is not flexible enough. A Fourier basis provides a poor representation of functions well localized in time, while a wavelet basis is not well adapted to represent functions whose Fourier transform has a narrow high frequency support. It is important to have a flexible decomposition for representing signal components whose localizations in time and frequency vary widely. The signal is decomposed into waveforms whose time-frequency properties are adapted to its local structures.

In this chapter, we present a popular translation invariant algorithm called matching pursuit [74,75,76,77,84] that provides an adaptive decomposition. In fact the method obtains the translation invariant property by decomposing any signal into a linear expansion of waveforms that belong to a redundant and shift-invariant dictionary of functions. These waveforms are selected in order to best match the signal structure at each iteration. Depending on applications, the matching pursuit technique allows a flexible choice of functions and the size of the dictionary. For example, specific dictionaries are constructed for inverse electro-magnetic problems [78], face recognition [79], data compression [80], analyzing of sleep electroencephalogram (EEG) [81,82] and recognition of power quality disturbances [83].

Unlike an orthogonal expansion, original matching pursuit [75] is non-linear. To improve the matching pursuit technique so that the projection converges with a finite number of iterations, an orthogonal matching pursuit technique that uses a Gram-Schmidt algorithm is proposed in [77,84]. The orthogonal matching pursuit, however, requires a significant computation cost for the Gram-Schmidt orthogonalization.

5.2 TIME-FREQUENCY ATOMIC DECOMPOSITION

In many applications in signal processing and harmonic analysis, signals are decomposed over a family of functions that are well localized both in time and frequency. Such functions are called time-frequency atoms. The decomposition properties depend upon the choice of time-frequency atoms. To extract information from complex signals, it is often necessary to adapt the time-frequency decomposition to particular signal structures.

Our signal space is $\mathbf{L}^2(R)$, a general family of time-frequency atoms can be generated by scaling, translating and modulating a single window function $g(t) \in \mathbf{L}^2(R)$. Suppose that $g(t)$ is real and centered at 0. For convenience, the norm of $g(t)$ is set to 1. For any scale $s > 0$, frequency modulation ξ and translation τ, we denote $\gamma = (s,\tau,\xi)$ and define

$$g_\gamma(t) = \frac{1}{\sqrt{s}} g\left(\frac{t-\tau}{s}\right) e^{\imath\xi t} \tag{5.1}$$

where the factor $1/\sqrt{s}$ normalizes the norm of $g_\gamma(t)$ to 1. The index γ is an element of the set $\Gamma = R^{+}\times R^{2}$. The function $g_\gamma(t)$ is centered at location τ and its energy is concentrated in the neighborhood of τ, whose size is proportional to s. Equation (5.1) yields the Fourier transform of $g(t)$,

$$\hat{g}_\gamma(\omega) = \sqrt{s}\hat{g}(s(\omega - \xi))e^{-\imath(\omega-\xi)\tau} \tag{5.2}$$

Since $|\hat{g}(\omega)|$ is even, $|\hat{g}_\gamma(\omega)|$ is centered at the frequency $\omega = \xi$. Its energy is concentrated in the neighborhood of ξ, whose size is proportional to $1/s$. The dictionary of time-frequency atoms, $D = (g_\gamma(t))_{\gamma\in\Gamma}$ is a very redundant set of functions in $\mathbf{L}^2(R)$ that includes windowed Fourier frames and wavelet frames [21]. To decompose efficiently a signal $f(t)$ over D, the time-frequency atoms $(g_{\gamma_m}(t))_{m\in N}$ with $\gamma_m = (s_m,\tau_m,\xi_m)$ in D are chosen such that they are best adapted to expand the function $f(t)$. Then the signal $f(t)$ can be written

$$f(t) = \sum_{m=-\infty}^{+\infty} a_m g_{\gamma_m}(t) \tag{5.3}$$

where a_m are expansion coefficients, which show how much correlation between the function $f(t)$ and the time-frequency atoms $g_{\gamma_m}(t)$.

In a windowed Fourier transform, the time-frequency atoms $g_{\gamma_m}(t)$ have a constant scale $s_m = s_0$ for all atoms $g_{\gamma_m}(t)$ and thus they are mainly localized over an interval whose size is proportional to s_0. Therefore, a windowed Fourier transform is not well adapted to signal structures that are much smaller or much larger than s_0. On the other hand, a wavelet transform is built by relating the frequency ξ_m to the scale s_m with $\xi_m = \xi_0 / s_m$, where ξ_0 is a constant. The resulting family of waveforms are dilated and translated of a single mother wavelet with complex phases. This has a limitation on the estimation of frequencies that are contained in the signal whose Fourier transform is well localized, especially at high frequencies.

In general, adaptive signal decomposition involves the expansion of a function over a set of waveforms, which are selected appropriately among a large and redundant dictionary. The next section describes a general algorithm called *matching pursuit* that performs such adaptive decomposition.

5.3 MATCHING PURSUITS

A dictionary is defined as a family $D = (g_\gamma(t))_{\gamma\in\Gamma}$ of vectors in a Hilbert space H, such that $\|g_\gamma(t)\| = 1$. Let V be the closed linear span of the dictionary, so the dictionary is completed if and only if $V = H$.

A signal $f(t)$ in H is computed as a linear expansion over a set of vectors selected from D which best matches the inner structures of f. A matching pursuit is a greedy algorithm which successively approximates $f(t)$ with projection onto elements of D. Let $g_{\gamma_0}(t) \in D$, the signal $f(t)$ can be decomposed into [75]

$$f(t) = < f, g_{\gamma_0} > g_{\gamma_0}(t) + Rf(t) \tag{5.4}$$

where $Rf(t)$ is the residual vector after approximating $f(t)$ in the direction of $g_{\gamma_0}(t)$. Since $Rf(t)$ is orthogonal to $g_{\gamma_0}(t)$, we have

$$\|f\|^2 = |< f, g_{\gamma_0} >|^2 + \|Rf\|^2 \tag{5.5}$$

To minimize the residue $\|Rf\|$, $g_{\gamma_0}(t) \in D$ must be chosen so that $|< f, g_{\gamma_0} >|$ is maximum. In some cases, it is computationally more efficient to find a vector $g_{\gamma_0}(t)$ that is almost optimal

$$|< f, g_{\gamma_0} >| \geq \alpha \sup_{\gamma\in\Gamma} |< f, g_\gamma >| \tag{5.6}$$

where $\alpha \in (0,1]$ is an optimality factor. The pursuit iterates this procedure by subdecomposing the residue. At step $m \geq 0$, the m^{th} order residue $R^m f$ is computed by projecting it onto a vector $g_{\gamma_m}(t) \in D$ that matches $R^m f$ almost at best, as it was done for $f(t)$ in the first step, i.e.

$$R^m f = < R^m f, g_{\gamma_m} > g_{\gamma_m} + R^{m+1} f \tag{5.7}$$

and

$$|< R^m f, g_{\gamma_m} >| \geq \alpha \sup_{\gamma\in\Gamma} |< R^m f, g_\gamma >| \tag{5.8}$$

which defined the $m+1^{th}$ order residue $R^{m+1}f$. The orthogonality of the residue $R^{m+1}f$ and $g_{\gamma_m}(t)$ implies

$$\left\|R^m f\right\|^2 = \left|< R^m f, g_{\gamma_m} >\right|^2 + \left\|R^{m+1} f\right\|^2 \tag{5.9}$$

By summing (5.7) from 0 to M, the decomposition of f over D up to $(M+1)^{\text{th}}$ order residue is:

$$f = \sum_{m=0}^{M} < R^m f, g_{\gamma_m} > g_{\gamma_m} + R^{M+1} f \tag{5.10}$$

Similarly, by summing (5.9) from 0 to M, we obtain

$$\left\|f\right\|^2 = \sum_{m=0}^{M} \left|< R^m f, g_{\gamma_m} >\right|^2 + \left\|R^{M+1} f\right\|^2 \tag{5.11}$$

Therefore, the signal $f(t)$ is decomposed into a set of dictionary elements $\{g_{\gamma_m}(t)\}_{0 \le m \le M}$ that are chosen to best match its residues. The matching pursuit decomposition in (5.10) is non-linear. However it maintains an energy conservation as though it were a linear, orthogonal decomposition [75]. When m tends to infinity, [27,75] proves that the residue $\|R^m f\|$ converges exponentially to 0, and that there exists $\lambda > 0$ such that for all $m \ge 0$

$$\left\|R^m f\right\| \le 2^{-\lambda m} \left\|f\right\| \tag{5.12}$$

As a consequence, $f(t)$ can be decomposed into

$$f = \sum_{m=0}^{+\infty} < R^m f, g_{\gamma_m} > g_{\gamma_m} \tag{5.13}$$

and

$$\left\|f\right\|^2 = \sum_{m=0}^{+\infty} \left|< R^m f, g_{\gamma_m} >\right|^2 \tag{5.14}$$

The double sequence $\left(< R^m f, g_{\gamma_m} >, \gamma_m\right)_{m \in N}$ is called a *structure book.* It specifies the expansion coefficients and the index of each chosen vector within the dictionary that are used to characterize the signal $f(t)$. In most cases, a signal of size N can obtain a sufficiently precise approximation with far fewer than N number of iterations [75].

5.3.1 Fast Implementations of Matching Pursuits

In an N dimensional space, suppose the dictionary D may have an infinite number of elements and it is complete (i.e $V = H$). When the dictionary is very redundant, we can project the signal $f(n)$ into a sub-dictionary $D_\alpha \subset D$, which could have many fewer elements $\left(g_\gamma\right)_{\gamma\in\Gamma_\alpha}$ than in D. Suppose that Γ_α is a finite index set included in Γ such that for any $f(n) \in H$

$$\sup_{\gamma\in\Gamma_\alpha}\left|< f, g_\gamma >\right| \geq \alpha \sup_{\gamma\in\Gamma}\left|< f, g_\gamma >\right| \tag{5.15}$$

Depending on the value of α and the redundancy of the dictionary D, the set Γ_α can be much smaller than Γ.

Instead of projecting the residue to all the dictionary elements at each iteration, a fast implementation of matching pursuit is to compute the $m+1^{th}$ order residue from the previous m^{th} order residue with a simple updating formula derived from (5.7) as:

$$< R^{m+1}f, g_\gamma >=< R^m f, g_\gamma > - < R^m f, g_{\gamma_m} >< g_{\gamma_m}, g_\gamma > \tag{5.16}$$

The algorithm is initialized by computing the inner products $\left(< R^0 f, g_\gamma >\right)_{\gamma\in\Gamma_\alpha}$. The updating formula in (5.16) allows us to find the inner products for the next stage by computing only $< g_{\gamma_m}, g_\gamma >$. To reduce the computational load, it is necessary to construct dictionaries with vectors having a sparse interaction. This means that each $g_\gamma(t) \in D$ has non-zero inner products with only a small fraction of all other dictionary vectors.

At stage m of the pursuit, we suppose that the inner products $\left(< R^m f, g_\gamma >\right)_{\gamma\in\Gamma_\alpha}$ for $m \geq 0$ have already been computed. We search in D_α for an element $g_{\tilde{\gamma}_m}$ such that

$$\left|< R^m f, g_{\tilde{\gamma}_m} >\right| = \sup_{\gamma\in\Gamma_\alpha}\left|R^m f, g_\gamma\right| \tag{5.17}$$

Since the search is in the sub-space Γ_α, then in Γ we are able to find in the neighborhood of $\tilde{\gamma}_m$ an index γ_m so that its dictionary element g_{γ_m} matches f even better than $g_{\tilde{\gamma}_m}$. The search is performed with a local search to find γ_m in the neighborhood of $\tilde{\gamma}_m$ where $\left|< R^m f, g_\gamma >\right|$ reaches a local maximum. This can be done in the time-frequency dictionaries where a sub-dictionary can sufficiently indicate a time-frequency region where almost best match is located.

$$\left|< R^m f, g_{\gamma_m} >\right| \geq \left|< R^m f, g_{\tilde{\gamma}_m} >\right|$$
$$\geq \sup_{\gamma \in \Gamma_\alpha} \left| R^m f, g_\gamma \right| \tag{5.18}$$

The number of iterations of $f(n)$ over D depends upon the desired precision ε. We decompose $f(n)$ over D up to a stage M such that

$$\left\| R^{M+1} f \right\| = \left\| f - \sum_{m=0}^{M} < R^m f, g_{\gamma_m} > g_{\gamma_m} \right\| \leq \varepsilon \left\| f \right\| \tag{5.19}$$

From (5.11), it proves that (5.19) is equivalent to

$$\left\| f \right\|^2 - \sum_{m=0}^{M} \left|< R^m f, g_{\gamma_m} >\right|^2 \leq \varepsilon \left\| f \right\|^2 \tag{5.20}$$

Note that we can obtain the translation invariance in matching pursuits by using translation invariance dictionaries. A dictionary D is translation invariant if for any element $g_\gamma(n)$ belongs to D, then its shifted versions $g_\gamma(n\text{-}p)$, $-n \leq p < N - n$, also belong to D. Then the matching pursuit decomposition of a signal $f(n)$ over D

$$f(n) = \sum_{m=0}^{M} < R^m f, g_{\gamma_m} > g_{\gamma_m}(n) + R^{M+1} f(n) \tag{5.21}$$

and the decomposition of the shifted signal $f_p(n) = f(n\text{-}p)$ selects a translation by p of the same vectors g_{γ_m} with the same decomposition coefficients [85], i.e.

$$f_p(n) = \sum_{m=0}^{M} < R^m f, g_{\gamma_m} > g_{\gamma_m}(n - p) + R^{M+1} f_p(n) \tag{5.22}$$

Hence, the signal patterns can be characterized as being invariant with a translation in the signal. However, translation invariant dictionaries are necessarily large and redundant, which often requires expensive computations.

5.3.2 Dictionaries of Time-Frequency Atoms

For dictionaries of time-frequency atoms that are derived from (5.1), a matching pursuit yields an adaptive time-frequency transform. Since dictionaries of time-frequency atoms are complete in $\mathbf{L}^2(R)$, in the space of transform, any signal $f(t) \in \mathbf{L}^2(R)$ is completely decomposed into a sum of complex time-frequency atoms $\{g_{\gamma_m}\}$ that best match its residues.

$$f(t) = \sum_{m=0}^{+\infty} < R^m f, g_{\gamma_m} > g_{\gamma_m}(t) \tag{5.23}$$

where $\gamma_m = (s_m, \tau_m, \xi_m)$ and

$$g_{\gamma_m}(t) = \frac{1}{\sqrt{s_m}} g\left(\frac{t - \tau_m}{s_m} \right) e^{i\xi_m t} \tag{5.24}$$

From the atomic decomposition of a function $f(t)$ in (5.23), the time-frequency energy distribution of $f(t)$ can be derived by adding the Wigner-Ville distribution of all selected atoms [75,87]. The Wigner-Ville distribution of $f(t)$, $Pf(t,\omega)$, is the cross Wigner-Ville distribution of $f(t)$ and itself, $P[f,f](t,\omega)$ and is defined by [26,86]

$$P[f,f](t,\omega) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} f\left(t + \frac{\tau}{2} \right) f^*\left(t - \frac{\tau}{2} \right) e^{-i\omega\tau} d\tau \tag{5.25}$$

From (5.23), the Wigner-Ville distribution of $f(t)$ is derived as

$$\begin{aligned} Pf(t,\omega) &= \sum_{m=0}^{+\infty} \left| < R^m f, g_{\gamma_m} > \right|^2 W g_{\gamma_m}(t,\omega) \\ &+ \sum_{m=0}^{+\infty} \sum_{\substack{l=0 \\ l \neq m}}^{+\infty} < R^m f, g_{\gamma_m} > < R^l f, g_{\gamma_l} >^* P\left[g_{\gamma_m}, g_{\gamma_l} \right](t,\omega) \end{aligned} \tag{5.26}$$

Since the double sum is the cross term of the Wigner-Ville distribution, which contains the terms that one usually tries to remove in order to obtain a clear picture of the energy distribution, we then only keep the first sum and define

$$Ef(t,\omega) = \sum_{m=0}^{+\infty} \left| < R^m f, g_{\gamma_m} > \right|^2 Pg_{\gamma_m}(t,\omega) \tag{5.27}$$

or

$$Ef(t,\omega) = \sum_{m=0}^{+\infty} \left| < R^n f, g_{\gamma_m} > \right|^2 Pg\left(\frac{t-\tau_m}{s_m}, s_m(\omega - \xi_m) \right) \tag{5.28}$$

Since the energy of a signal remains the same through Wigner-Ville distribution

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} Pg(t,\omega)\,dt\,d\omega = \left\| g(t) \right\|^2 = 1 \tag{5.29}$$

then the energy conversion equation (5.14) implies

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} Ef(t,\omega)\,dt\,d\omega = \left\| f(t) \right\|^2 \tag{5.30}$$

Therefore, we can interpret $Ef(t,\omega)$ as an energy density of $f(t)$ in the time-frequency plane. It does not contain the cross terms as in Wigner and Cohen distributions. In the case $g(t)$ is a Gaussian window of unity norm

$$g(t) = 2^{1/4} \exp(-\pi t^2), \tag{5.31}$$

then its Wigner-Ville distribution is

$$Pg(t,\omega) = \frac{1}{\pi} \exp\left(-2\pi t^2 - \frac{\omega^2}{2\pi} \right) \tag{5.32}$$

So the time-frequency energy distribution $Ef(t,\omega)$ in (5.28) becomes

$$Ef(t,\omega) = \frac{1}{\pi} \sum_{m=0}^{+\infty} \left| < R^m f, g_{\gamma_m} > \right|^2 \exp\left(-\frac{2\pi}{s_m^{\;2}}(t-\tau_m)^2 - \frac{s_m^{\;2}}{2\pi}(\omega - \xi_m)^2 \right) \tag{5.33}$$

Through this expression, $Ef(t,\omega)$ remains positive in the case of Gaussian windows. In the time-frequency plane, the time-frequency energy distribution $Ef(t,\omega)$ is presented as a sum of two dimensional Gaussian blobs centred at (τ_m, ξ_m).

5.3.3 Discrete Matching Pursuit with Gabor Dictionaries

Since Gaussian windows have optimal time-frequency energy distribution as regard to the uncertainty principle, by scaling, translating and modulating a Gaussian window, a Garbor dictionary that is translation invariant with time and frequency is constructed in [75,87]. Using the Gaussian window in (5.31), a discrete window of period N at each scale 2^j is designed by sampling and periodizing, giving

$$g_j(n) = K_j \sum_{p=-\infty}^{+\infty} g\left(\frac{n-pN}{2^j}\right) \tag{5.34}$$

where the constant K_j is used to normalize the norm $\|g_j(n)\|$. At this scale 2^j, the discrete window $g_j(n)$ is then modulated and translated in time to obtain a discrete Gabor atom $g_\gamma(n)$, $\gamma = (2^j, p, k)$, that is

$$g_\gamma(n) = g_j(n-p)\exp\left(\frac{i2\pi kn}{N}\right) \tag{5.35}$$

By letting Γ be the discrete set of index $\gamma = (2^j,p,k)$ for $j \in [0, \log_2 N]$ and $(p,k) \in [0, N-1]^2$. The resulting discrete Gabor dictionary $D = \{g_\gamma\}_{\gamma \in \Gamma}$ is time and frequency translation invariant modulo N [27,85]. For real signals, the matching pursuit decomposes a signal in this dictionary by projecting $R^m f$ on the plane generated by $(g_{\gamma+}, g_{\gamma-})$, with $\gamma^{\pm} = (2^j, p, \pm k)$. Since the residue $R^m f$ of a real signal is real, we can verify that the projection is equivalent to projecting $R^m f$ on a real vector

$$g_{(\gamma,\phi)}(n) = K_{(\gamma,\phi)} g_j(n-p)\cos\left(\frac{2\pi kn}{N} + \phi\right) \tag{5.36}$$

where the constant $K_{(\gamma,\phi)}$ sets the norm $\|g_{(\gamma,\phi)}\|$ to 1, and the phase $\phi \in [0, 2\pi)$ that was hidden in the complex numbers, now appears in the real atoms so as to maximize the inner product with $R^m f$. In fact real atoms are related to complex atoms by

$$g_{(\gamma,\phi)}(n) = \frac{K_{(\gamma,\phi)}}{2}\left[e^{i\phi} g_{\gamma+}(n) + e^{-i\phi} g_{\gamma-}(n)\right] \tag{5.37}$$

From (5.37), we can derive the normalization constant

$$K_{(\gamma,\phi)} = \frac{\sqrt{2}}{\sqrt{1+\mathrm{Re}(e^{j2\phi} < g_{\gamma+}, g_{\gamma-} >)}} \tag{5.38}$$

For any real residue $R^m f$, we have

$$|< R^m f, g_{(\gamma,\phi)} >| = K_{(\gamma,\phi)} \, |\mathrm{Re}(e^{-i\phi} < R^m f, g_\gamma >)| \tag{5.39}$$

To maximize the inner product with $R^m f$, the phase ϕ of $g_{(\gamma,\phi)}$ is chosen to be equal to the complex phase of $< R^m f, g_\gamma >$ so that

$$|\mathrm{Re}(e^{-i\phi} < R^m f, g_\gamma >)| = |< R^m f, g_\gamma >| \tag{5.40}$$

The matching pursuit decomposes any real signal $f(n)$ into

$$f(n) = \sum_{m=0}^{+\infty} < R^m f, g_{(\gamma_m,\phi_m)} > g_{(\gamma_m,\phi_m)}(n) \tag{5.41}$$

The indexes (γ_m,ϕ_m) in (5.41) are chosen to best match the residue of $f(n)$. The time-frequency energy distribution of the signal $Ef(n,k)$ is presented via the matching decomposition (5.41) by summing the Wigner-Ville distribution $Pg_{\gamma_m}(n,k)$ of the complex atoms g_{γ_m} as in (5.33) [75,87],

$$Ef(n,k) = \frac{1}{\pi} \sum_{m=0}^{+\infty} \left|< R^m f, g_{(\gamma_m,\phi_m)} >\right|^2 Pg_{\gamma_m}(n,k) \tag{5.42}$$

where

$$Pg_{\gamma_m}(n,k) = \exp\left(-\frac{2\pi}{(2^{J_m})^2}(n-\tau_m)^2 - \frac{(2^{J_m})^2}{N}(k-k_m)^2 \right) \tag{5.43}$$

The fast implementation of matching pursuit in a Gabor dictionary is performed with a sub-dictionary $D_\alpha = \{g_\alpha\}_{\alpha \in \Gamma_\alpha}$. At each scale 2^j, the time-frequency plane is subsampled at time $p(a2^j)$ and at frequency $k(Na2^{-j})$, where $p \in [0, Na^{-1}2^{-j})$, $k \in [0, a^{-1}2^j)$, and the sampling factor $a \leq 1$ is small enough to detect the high energy regions of the signal. At each iteration, once the best match atom $g_{\tilde{\gamma}_m}$ is found in D_α, then in D the matching is improved by searching in the neighborhood of $g_{\tilde{\gamma}_m}$ to find an atom g_{γ_m} that locally maximizes the correlation with the signal residue $R^m f$.

Since the selection of vector g_{γ_m} by the matching pursuit algorithm is not a priori for orthogonalizing to all previously selected vectors $\{g_{\gamma_p}\}_{0\le p<m}$, then when subtracting the projection of $R^m f$ over g_{γ_m}, this introduces new components that are in the directions of $\{g_{\gamma_p}\}_{0\le p<m}$. Therefore, the algorithm can require an infinite number of iterations to converge. One can avoid this by projecting the residues onto an orthogonal basis.

5.4 ORTHOGONAL MATCHING PURSUIT

To improve the convergence rate of the projection for matching pursuit and hence guarantee the projection to converge within a finite number of iterations, an orthogonal matching pursuit technique is proposed in [84,77]. The technique uses a Gram-Schmidt algorithm that computes an orthogonal family $\{u_p\}_{0\le p<m}$ from the previously selected vectors $\{g_{\gamma_p}\}_{0\le p<m}$.

Initializing with $u_0 = g_{\gamma_0}$, at step $m \ge 0$, an orthogonal matching pursuit first selects the vector g_{γ_m} that satisfies (5.8). The technique then orthogonalizes g_{γ_m} with respect to $\{g_{\gamma_p}\}_{0\le p<m}$ by using the Gram-Schmidt algorithm to give

$$u_m = g_{\gamma_m} - \sum_{p=0}^{m-1} \frac{< g_{\gamma_m}, u_p >}{\|u_p\|^2} u_p \tag{5.44}$$

The family $\{u_p\}_{0\le p\le m}$ is an orthogonal basis of V_{m+1}. The orthogonal matching pursuit projects the residue $R^m f$ over u_m instead of g_{γ_m} that defines the residue $R^{m+1}f$

$$R^m f = \frac{< R^m f, u_m >}{\|u_m\|^2} u_m + R^{m+1} f \tag{5.45}$$

One can show that $R^m f$ is orthogonal to the vectors $\{g_{\gamma_p}\}_{0\le p<m}$, and that equation (5.44) implies

$$< R^m f, u_m > = < R^m f, g_{\gamma_m} > \tag{5.46}$$

Hence, (5.45) can be rewritten as a residue updating equation

$$R^{m+1}f = R^m f - \frac{< R^m f, g_{\gamma_m} >}{\|u_m\|^2} u_m \tag{5.47}$$

This residue updating equation is similar to that of a matching pursuit in (5.7), but instead of subtracting the projection of $R^m f$ over g_{γ_m} in the direction of g_{γ_m}, we subtract it in a direction orthogonal to all previously selected vectors $\{g_{\gamma_p}\}_{0\le p<m}$. As $R^{m+1}f$ is orthogonal to u_m, we have

$$\|R^{m+1}f\|^2 = \|R^m f\|^2 - \frac{|< R^m f, g_{\gamma_m} >|^2}{\|u_m\|^2} \tag{5.48}$$

In an N dimensional space, the orthogonal matching pursuit converges with a finite number $M \le N$ of iterations so that $R^M f = 0$. Summing (5.47) for $0 \le m < M$ yields

$$f = \sum_{m=0}^{M-1} \frac{< R^m f, g_{\gamma_m} >}{\|u_m\|^2} u_m \tag{5.49}$$

and an energy conservation

$$\|f\|^2 = \sum_{m=0}^{M-1} \frac{|< R^m f, g_{\gamma_m} >|^2}{\|u_m\|^2} \tag{5.50}$$

The objective is to expand f over the original dictionary vectors $\{g_{\gamma_m}\}_{0\le m<M}$, and this requires a change of basis. From the Gram-Schmidt relations (5.44), we can decompose u_m over the family $\{g_{\gamma_p}\}_{0\le p<M}$

$$u_m = \sum_{p=0}^{m} b_{p,m} g_{\gamma_p} \tag{5.51}$$

The coefficients $b_{p,m}$ can be calculated one by one through the iterations of the orthogonal matching pursuit. By inserting expression (5.51) into (5.49), we obtain an expansion of f in $\{g_{\gamma_p}\}_{0\le p<M}$

$$f = \sum_{m=0}^{M-1} \frac{< R^m f, g_{\gamma_m} >}{\|u_m\|^2} \sum_{p=0}^{m} b_{p,m} g_{\gamma_p} \tag{5.52}$$

Rearranging the terms of this double summation to give

$$f = \sum_{p=0}^{M-1} a_p g_{\gamma_p} \tag{5.53}$$

where

$$a_p = \sum_{m=p}^{M-1} b_{p,m} \frac{< R^m f, g_{\gamma_m} >}{\|u_m\|^2}$$

For the first few iterations, the matching pursuit often selects vectors that are almost orthogonal. Then orthogonal and non-orthogonal pursuits select nearly the same set of vectors, so the Gram-Schmidt orthogonalization is not needed. When the number of iterations increases and gets close to *N*, the residue's norms of the orthogonal matching pursuit decrease faster than that of the non-orthogonal matching pursuit. For large *M*, the orthogonal matching pursuit has the convergent advantage at the cost of requiring significant computations for the Gram-Schmidt orthogonalization. The non-orthogonal matching pursuit is thus more often used for large signals [27].

5.5 NUMERICAL DEMONSTRATIONS

Since our primary purpose is to study the characteristics of power quality disturbance signals via signal transformations, we should not consider all selected atoms from the matching pursuit, but examine only the first few selected atoms that contain the majority signal energy. Moreover, since the first few iterations of the matching pursuit often select atoms that are almost orthogonal, the Gram-Schmidt orthogonalization is not needed. This provides a minimal number of calculations, and speeds up the classification process.

We perform the matching pursuit of the Gabor dictionary in Section 5.3.3, which decomposes a disturbance waveform *f*(*n*) into a sum of Gabor atoms that are selected to best match its residues. Figure 5.1 shows the matching pursuit decomposition for an impulse transient disturbance waveform. While Figure 5.2 is for a low frequency capacitor switching signal, Figure 5.3 is for a high frequency capacitor switching signal, and Figure 5.4 is for a single notch signal. In each figure, figure (a) is the disturbance signal $f(n)$ that has a duration of 20ms and is sampled at 12.8KHz (i.e. 256 samples). The time-frequency energy distribution of the disturbance signal $Ef(n,k)$ obtained from the matching pursuit decomposition is shown in (b). Figure (c) presents the energy conversion $\|R^mf\|^2/\|f\|^2$ versus the number of iterations m of the matching pursuit.

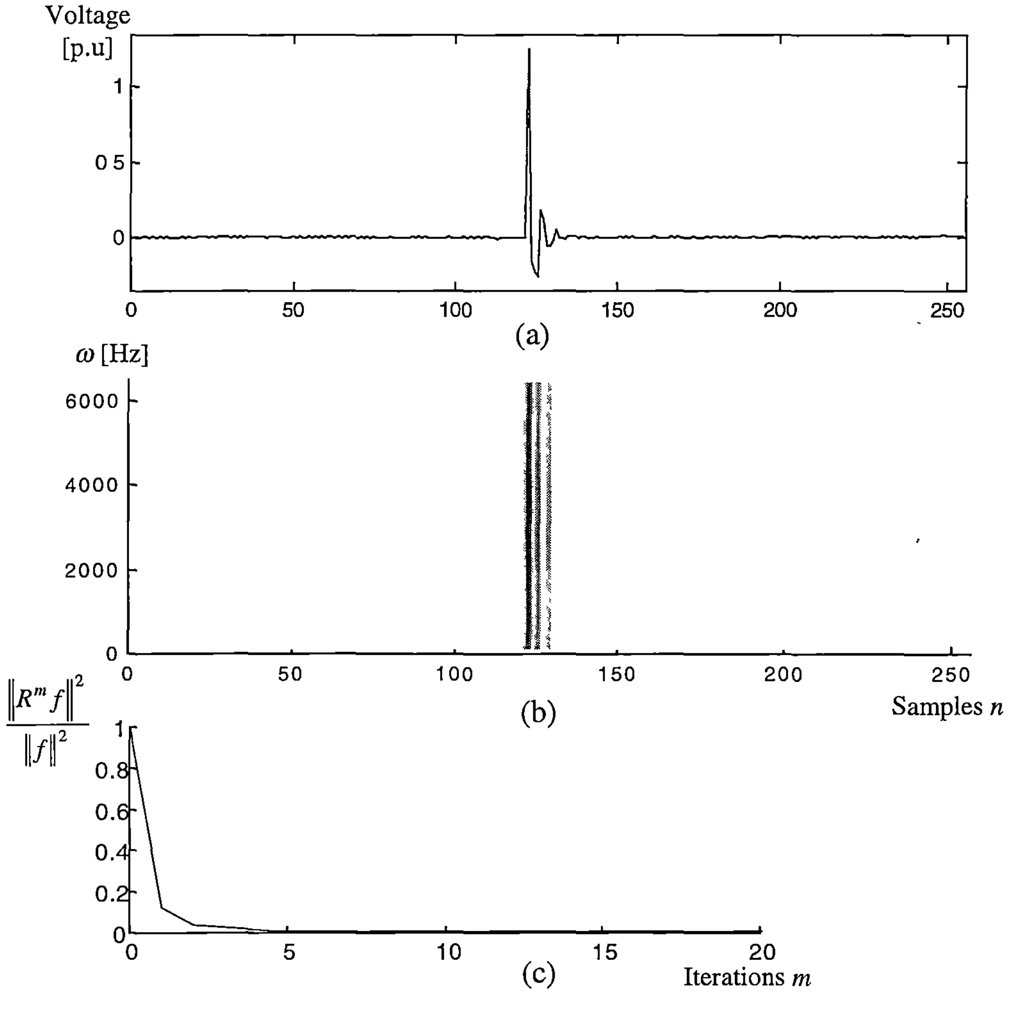


Figure 5.1: (a) 256 samples of an impulse transient disturbance $f(n)$; (b) Time-frequency energy distribution of the disturbance signal $Ef(n,k)$ obtained from the matching pursuit with Gabor dictionary; (c) Energy conversion $\|R^mf\|^2/\|f\|^2$ versus the number of iterations m of the matching pursuit.

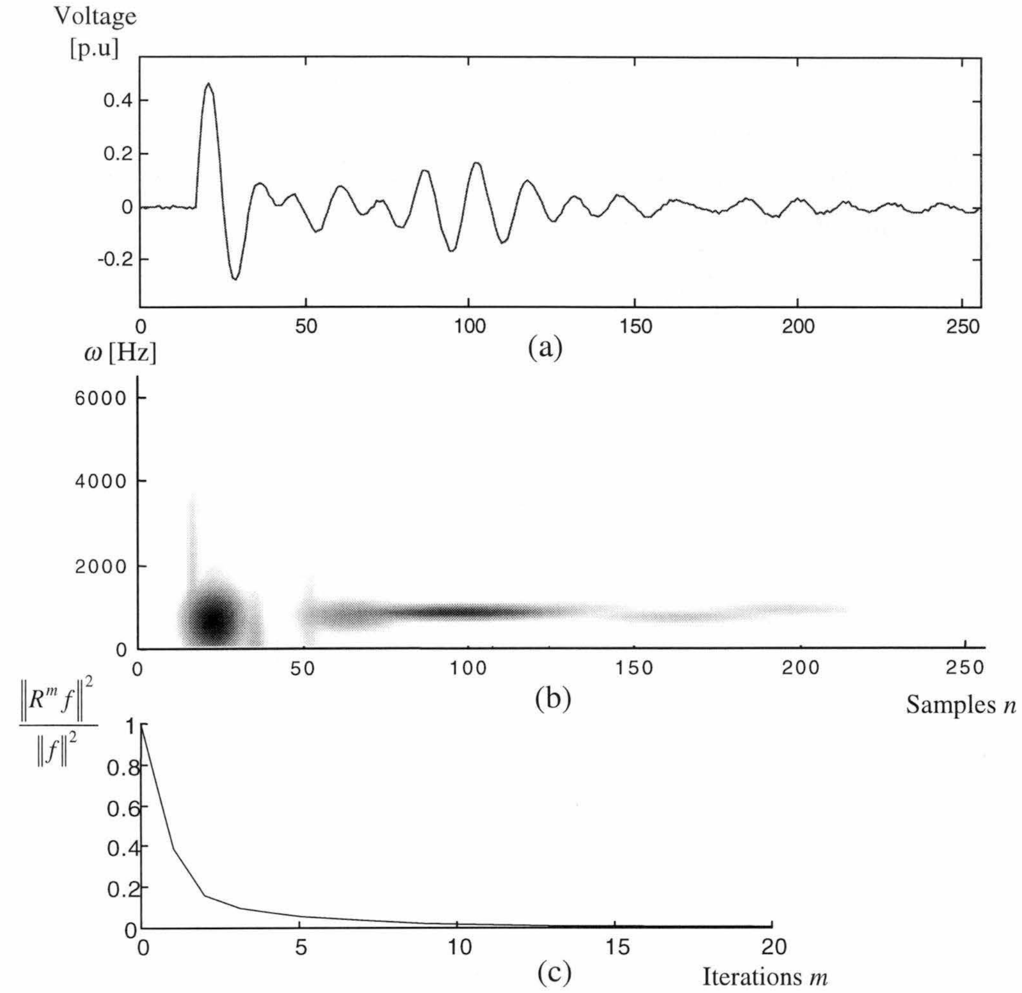


Figure 5.2: (a) 256 samples of a low frequency capacitor switching disturbance $f(n)$; (b) Time-frequency energy distribution of the disturbance signal $Ef(n,k)$ obtained from the matching pursuit with Gabor dictionary; (c) The energy conversion $\|R^m f\|^2/\|f\|^2$ versus the number of iterations m of the matching pursuit.

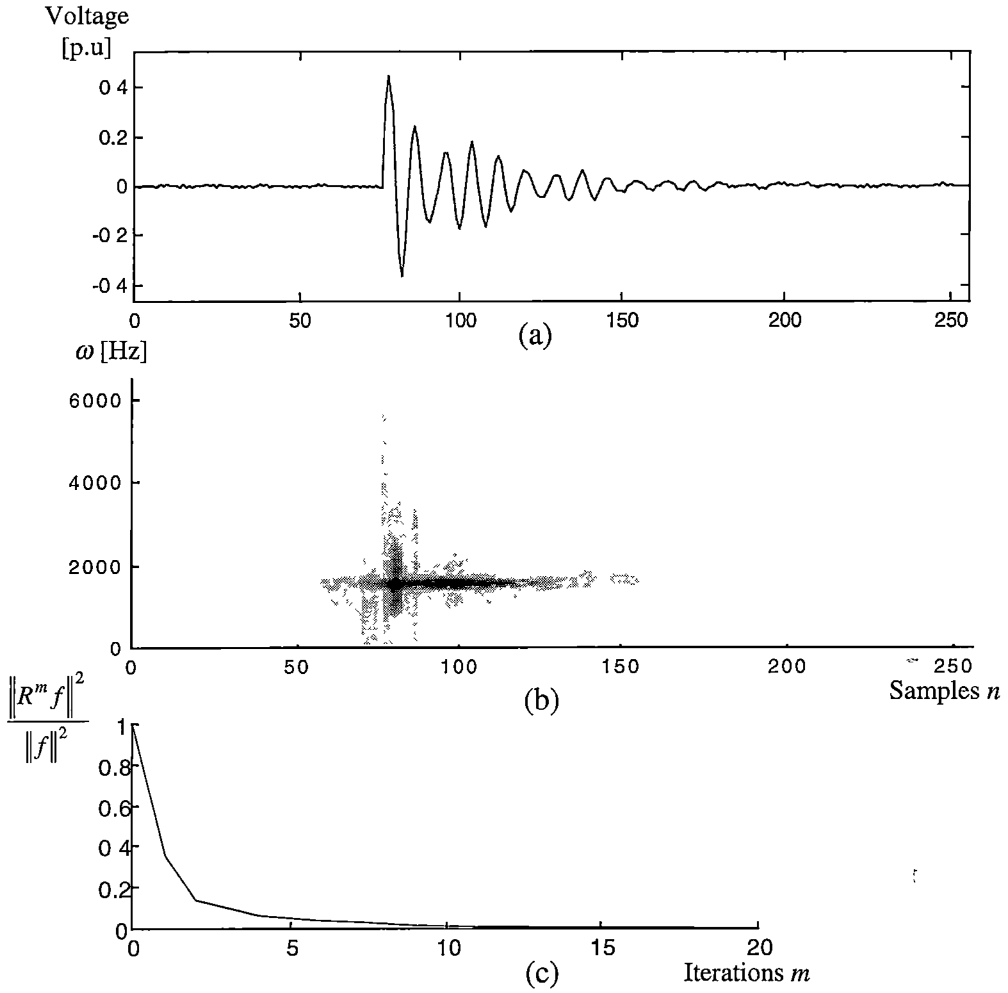


Figure 5.3: (a) 256 samples of a high frequency capacitor switching disturbance $f(n)$; (b) Time-frequency energy distribution of the disturbance signal $Ef(n,k)$ obtained from the matching pursuit with Gabor dictionary; (c) The energy conversion $\|R^m f\|^2/\|f\|^2$ versus the number of iterations m of the matching pursuit.

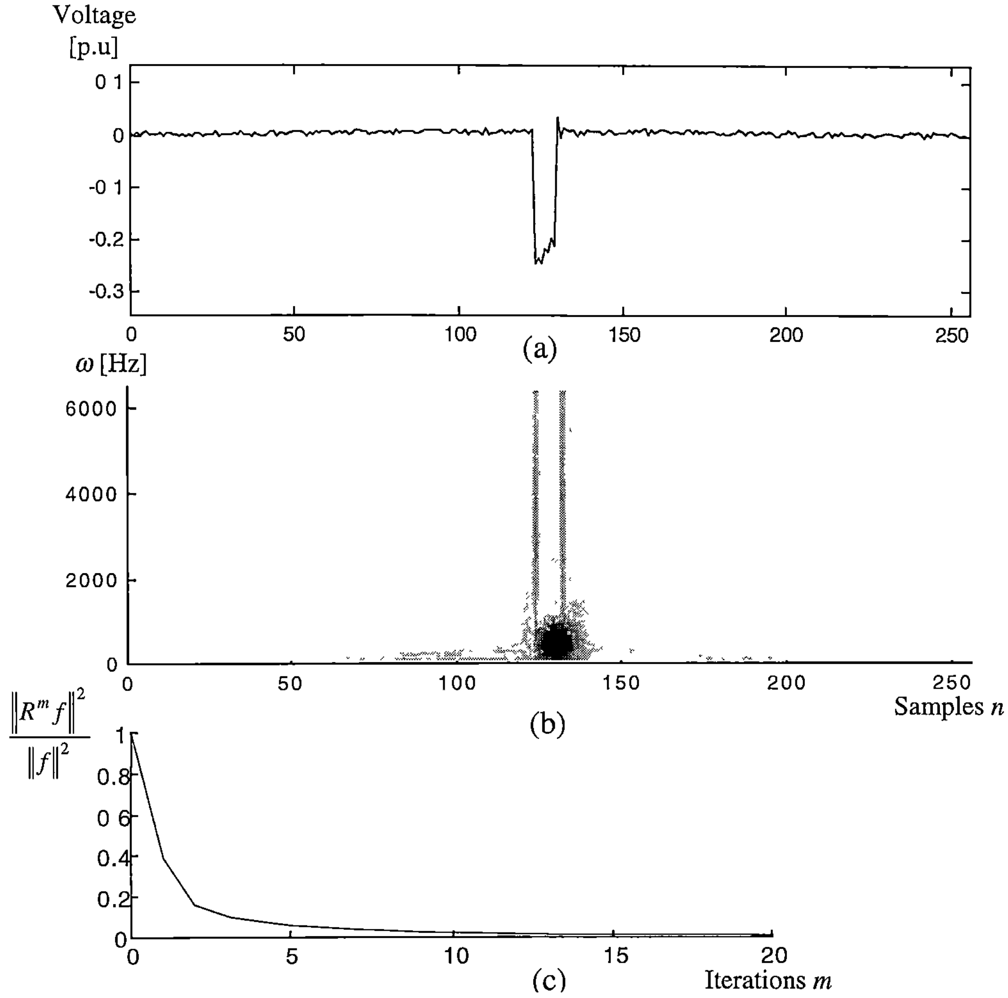


Figure 5.4: (a) 256 samples of a single notch disturbance *f*(*n*); (b) Time-frequency energy distribution of the disturbance signal *Ef*(*n*,*k*) obtained from the matching pursuit with Gabor dictionary; (c) The energy conversion $\|R^m f\|^2/\|f\|^2$ versus the number of iterations *m* of the matching pursuit.

These examples show clearly that the convergence rate of the matching pursuit decays toward zero very quickly, and that we obtain a sufficiently precise signal approximation by a very small number of iterations. The convergence rate can be different depending on the structure and length of the signal. However in the above examples, disturbance signals are approximated to more than 95% of their energy after only five iterations. The technique thus provides a very efficient decomposition, and also is a translation-invariant representation.

Since the matching pursuit is a greedy algorithm, the extraction of features (e.g. modulation frequencies, window sizes, ...) from the selected vectors must be done carefully. There are a limited number of selected vectors from the matching decomposition that closely match the local structures of the signal at different locations, while many other vectors are selected as they best match the signal residues, whose structures can be completely different from signal. This is because the subtraction between the signal and the selected vectors can create new components that do not have in the signal. This is the case when the signal has been subtracted many times. Figure 5.3 shows clearly the appearance of such new components. They appear at the position where the disturbance signal changes from a sharp transition to regular oscillations.

From the discussion above we should only use vectors that match closely to the signal structure. This agrees with our objective stated at the beginning of this section. We thus examine only the first five selected vectors that contain most of the signal energy [83]. The discarded energy is considered as insignificant and is produced from small signal components and noise.

Chapter 6

PATTERN RECOGNITION APPROACHES

6.1 INTRODUCTION TO PATTERN RECOGNITION

Automatic recognition, description and classification of a variety of objects and patterns are important problems in engineering and scientific disciplines such as computer vision, marketing, biology, psychology, medicine, artificial intelligence and remote sensing. The demand for automatic pattern recognition systems is rapidly growing due to the performance requirements of speed, accuracy and cost on huge databases. Pattern recognition systems have been designed using the following approaches: (i) template matching, (ii) statistical methods, (iii) syntactic methods and (iv) neural networks.

Among the various frameworks of pattern recognition that have been formulated and developed in the past fifty years, the statistical approach is the most intensively studied and used in practice [88]. Recently, neural network techniques and methods using statistical learning theory have gained increasing attention as they have the capability of robust learning inference and generalisation from the training data. The main difference between neural networks and other approaches is that neural networks can learn the relationships of a complex non-linear input-output, and use sequential training procedures. In spite of what seem to be differences, neural network models are similar or have a close relationship to statistical pattern recognition [89,90].

The concepts of statistical decision theory are to determine the decision boundaries between classes. Depending on the information available about the class condition densities, different techniques can be used to design a classifier. If all information of the class distribution is specified, then a clear deterministic Bayes decision rule can be "drawn" out for the designing of the classifier. However, in practice the information of the class distribution is normally not known and must be learned from the available training set. In the case that the class condition densities are known (e.g. Gaussian distribution), but some information on the densities (e.g. the mean, covariance matrix) are unknown, we then have a parametric decision problem. The Bayesian approach in this situation is used to estimate the unknown parameters in the density functions. In the case that we do not know the class condition densities, we then have a nonparametric problem in which the class distribution must be estimated or the decision boundaries are directly constructed from the training data. The multilayer perceptron is in fact viewed as the nonparametric method that constructs the decision boundaries from the training data. Radial basis function network is a special class of the neural network that contains the radial basis function - a form of statistical parametric approach. Unlike the multilayer perceptron, the radial basis network is required to estimate the class distribution (i.e. the mean, covariance matrix) from the training data.

In this chapter, we present two important categories of pattern recognition. The first one is a classical and well-known statistical approach called Bayesian. The second is the artificial neural network approach. In particular, radial basis function network that is imported from statistical leaning theory, is popular due to its fast training, and we will show that this type of classifier is suitable in classifying power quality disturbances [91,92].

6.2 BAYESIAN PARAMETER ESTIMATION APPROACH

6.2.1 General Consideration

In an N-dimensional space, let $\boldsymbol{x}(k) = (x_1(k),x_2(k),\ldots x_N(k))^T$, $k \in [1,K]$ be the k^{th} pattern and the available classes that $\boldsymbol{x}(k)$ might be a member are c_h, $h = 1,2, \ldots, H$. The statistics of the overall phenomenon can be described in terms of the following probabilities:

$P(c_h)$ ≡ the *a priori* probability that a pattern belongs to class c_h,

$p(\boldsymbol{x}(k))$ ≡ the probability that a pattern is $\boldsymbol{x}(k)$,

$p(\boldsymbol{x}(k)|c_h)$ ≡ the class conditional probability that a pattern is $\boldsymbol{x}(k)$, given that it belongs to class c_h,

$P(c_h|\boldsymbol{x}(k))$ ≡ the *a posteriori* conditional probability that the pattern's class membership is c_h, given that the pattern is $\boldsymbol{x}(k)$,

$P(c_h,\boldsymbol{x}(k))$ ≡ the joint probability that the pattern is $\boldsymbol{x}(k)$ and that its class membership is c_h

The normalisation conditions of the patterns and the classes are

$$\sum_{k=1}^{K} p(\boldsymbol{x}(k)) = 1, \quad \text{and} \quad \sum_{h=1}^{H} P(c_h) = 1 \tag{6.1}$$

The joint probability can be expressed by two different ways,

$$\begin{aligned} P(c_h,\boldsymbol{x}(k)) &= p(\boldsymbol{x}(k) \mid c_h)P(c_h) \\ &= P(c_h \mid \boldsymbol{x}(k))p(\boldsymbol{x}(k)) \end{aligned} \tag{6.2}$$

Then the a priori estimate of the probability of a certain class can be converted to the a posteriori probability

$$P(c_h \mid x(k)) = \frac{p(x(k) \mid c_h)P(c_h)}{p(x(k))} \tag{6.3}$$

This expression is known as the Bayes' relation, which is used to estimate values of the a posteriori, or *measurement condition*, probability $P(c_h|x(k))$ if those statistics are not known directly, but the class conditional probabilities and a priori probabilities are known. In principle and in practice, a decision rule can be make based on the values of the a posteriori probability such that a pattern $x(k)$ is decided to belong to class $c = c_c$ if and only if [93,94]

$$P(c_c \mid x(k)) > P(c_h \mid x(k)), \ \forall h \neq c \tag{6.4}$$

This means that the probability of the pattern $x(k)$ belonging to the class c_c is highest.

6.2.2 Discriminant Functions

Gaussian models are popular because of their mathematical tractability and because the Gaussian distribution is the natural result of a combination of a large number of samples in the practical world (Central-Limit theorem). Consider a multidimensional Gaussian probability density distribution function of a N-dimension random variable x with mean μ [94,101],

$$p(x) = \frac{1}{(2\pi)^{N/2}(\det.\Sigma)^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right] \tag{6.5}$$

where Σ is the covariance matrix of x and $(\det.\Sigma)$ is the determinant of Σ. Then each class of Gaussian variable is specified by its mean (i.e. the centre) and its covariance (i.e. the spread or receptive field).

Define a *discriminant function* for the h^{th} class c_h, as the probability of finding that class given a feature vector $x(k)$, i.e.

$$d_h(k) = P(c_h \mid \boldsymbol{x}(k)) \tag{6.6}$$

Since any monotonically increasing function of $d_h(k)$ can also be used as a discriminant function and the log-likelihood function is one such convenient variation, we then redefine the discriminant function as

$$d_h(k) = 2\log\{P(c_h|\boldsymbol{x}(k))\} \tag{6.7}$$

In the case of Gaussian distribution, the class conditional probability $p(\boldsymbol{x}(k)|c_h)$ is given in (6.5). From the Bayes' relation in (6.3), the discriminant function becomes

$$\begin{aligned} d_h(k) &= -\left[(\boldsymbol{x}(k)-\boldsymbol{\mu}_h)^T \Sigma_h^{-1} (\boldsymbol{x}(k)-\boldsymbol{\mu}_h)\right] \\ &\quad - \left[N\log(2\pi) + \log\{p(\boldsymbol{x}(k))\}^2\right] - \log\frac{\det.\Sigma_h}{P(c_h)^2} \end{aligned} \tag{6.8}$$

Note that the second term in (6.8) is the same for all classes, i.e. playing no role in the discriminant function, and therefore can be dropped. The last term represents a *class bias*. Hence, finding the largest $d_h(k)$ in (6.6) is the same as finding the largest class conditional probability $p(\boldsymbol{x}(k)|c_h)$, that is, the first term in (6.8) is maximum.

If all classes have equal a priori probabilities $P(c_h)$ (i.e. the same sample population for each class) and if we assume further, for simplicity, that all classes have the same covariance matrix, i.e. $\Sigma_h = \Sigma$, then the negative of the first term in (6.8) is the *normalised distance* from the feature vector $\boldsymbol{x}(k)$ to the mean vector μ_h of the h^{th} class. Thus $\boldsymbol{x}(k)$ is assigned to the h^{th} class if its distance to the class mean vector μ_h is minimum

$$\begin{aligned} D_h(k)^2 &= \left\|\boldsymbol{x}(k) - \boldsymbol{\mu}_h\right\|_{\Sigma^{-1}}^2 \\ &= (\boldsymbol{x}(k) - \boldsymbol{\mu}_h)^T \Sigma^{-1} (\boldsymbol{x}(k) - \boldsymbol{\mu}_h) \end{aligned} \tag{6.9}$$

The distance as defined in (6.9) is known as the *Mahalanobis* distance [95]. If all the feature components in the feature vector are uncorrelated, that is, they are independent of one another, but with different variances, then

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^{\;2}} & 0 & 0 \cdots & 0 \\ 0 & \frac{1}{\sigma_2^{\;2}} & 0 \cdots & 0 \\ & \ddots & & \\ 0 & 0 & \cdots 0 & \frac{1}{\sigma_N^{\;2}} \end{bmatrix} \tag{6.10}$$

Thus if feature component x_i has a smaller spread than component x_j, i.e. $1/\sigma_i^{\;2} > 1/\sigma_j^{\;2}$, we put more emphasis on x_i than x_j. If furthermore, all feature components are uncorrelated and have the same variance, then $\Sigma = \sigma^2 \boldsymbol{I}$ and $D_h(k)^2$ becomes a simple *Euclidean* distance

$$D_h(k)^2 = \frac{(\boldsymbol{x}(k) - \boldsymbol{\mu}_h)^T (\boldsymbol{x}(k) - \boldsymbol{\mu}_h)}{\sigma^2} \tag{6.11}$$

We note that the discriminant requires the measurement of distance, and patterns are classified in accordance with the class membership of nearest neighbours or with the nearest class centre. In some cases, discriminants are hypersurfaces defined in the input space, and patterns are classified based on class decision boundaries.

6.3 ARTIFICIAL NEURAL NETWORK APPROACH

Although artificial neural networks (ANN) have a rich history of approximately 60 years since their first publication by McCulloch and Pitts [96] in 1943, they did not become popular until about 20 years ago. Basically, there are three entities to characterise an ANN [94]. They are:

1. Network topology, or interconnection of neural "units",
2. Characteristics of individual units or artificial neurons, and
3. Strategy for pattern learning or training.

The operation of ANNs is based on some organisation principles such as learning, generalisation, adaptivity, and distributed representation and computation in the network weights. ANNs are typical of globally generalising networks which have the capability of robust learning inference and generalisation from the training data. The most commonly used family of ANNs for pattern classification applications are the feedforward networks (also known as multilayer perceptrons). These networks are organised into layers and have unidirectional connection between layers. Another popular network used for data clustering and feature mapping is the Self-Organising Map (SOM) [97,105].

6.3.1 Multilayer, Feedforward Network Structure

The feedforward network is composed of a hierarchy of processing units that are organised into a series of two or more mutually exclusive sets of neurons or layers. The first layer is the input layer that is used for applying input values to the network. The last layer is the output layer where the final state of the network is read. Between these two layers, there is zero or more layers of hidden units. Weighted links connect each unit in one layer to those in the next-higher layer. Figure 6.1 shows the architecture of a two-layers (of processing units) feedforward network. The network has N units for the input layer, H processing units at the hidden layer, and M processing units at the output layer.

A commonly used function for the activation functions of the feedforward network are the sigmoid function:

$$f(u) = \frac{1}{1 + \exp\left(-\frac{u + \theta_h}{\theta_0}\right)} \qquad (6.12)$$

where θ_h is a threshold or bias, whose effect is to shift the activation function along the horizontal axis by a value of θ_h, and θ_o is the scaled parameter that modifies the shape of the sigmoid. Other functions are also used by some researchers, namely, the tangential function,

$$f(u,b,c) = b\tanh(cu) \tag{6.13}$$

where b and c are two constants, and the linear activation function (i.e. $f(u) = bu$, b is a constant).

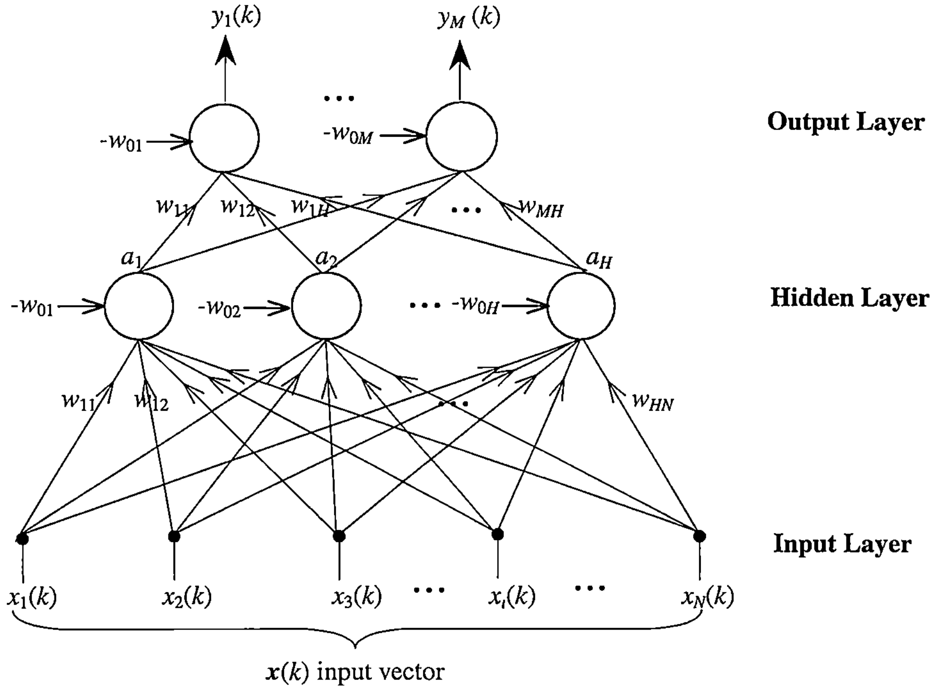


Figure 6.1: A schematic depiction of a two-layer feedforward network

The output of a neuron is calculated by first summing the weighted inputs to produce an internal activation, *net*, then by applying the activation function. A neuron is normally connected to a bias that is included as a weight connected to a fixed input of a value of –1. If we denote the index of this fixed bias is $i = 0$, then for an k^{th} input sample, the activation output of the h^{th} node in the hidden layer is

$$net_h(k) = \sum_{i=0}^{N} w_{hi} x_i(k) \tag{6.14}$$

and

$$a_h(k) = f_h(net_h(k)) \tag{6.15}$$

For a two-layer feedforward network, the activation output of the m^{th} node in the output layer that takes the activation outputs of the hidden layer $\{a_h(k)\}$ as its input, is

$$net_m(k) = \sum_{h=0}^{H} w_{mh} a_h(k), \quad \text{and} \quad y_m(k) = f_m(net_m(k)) \tag{6.16}$$

Considering the complexity requirement of the multilayer feedforward network, it is natural to ask how many layers and how many units should be in the hidden layer(s). A simple argument is provided in [93,98], in which a network of three-layers can form any arbitrarily complex decision region, and for a network of N input features, ($2N$+1) processing units in the (single) hidden layer are capable of modelling the problem. It is however, difficult to choose the best network in any practical problem.

6.3.2 Training the Feedforward Network

There are a number of methods for training feedforward networks such as backpropagation training [99], recursive least-squares (Kalman) based training [100], conjugate-gradient training [101], Newton's method [102] and its modifications (e.g. Marquadt-Levenberg algorithm [103]). Each method has its own advantages and disadvantages. In fact, the backpropagation learning algorithm is perhaps the most popular training method for feedforward networks, and is presented here.

Backpropagation learning is a gradient descent method based on the Least Mean Square (LMS) algorithm. There are two computational passes that are made in the learning phase of network training. At first, a sample is presented to the network and a forward pass calculates the activation output of each neuron. Then the error of the actual output of the network compared to the desired response is propagated through the network in a backward pass for adjusting the connection weights.

For an input pattern $x(k)$, the square of error of the network output is defined by

$$E(k) = \frac{1}{2}\sum_{m=1}^{M}[t_m(k) - y_m(k)]^2 \qquad (6.17)$$

where $t_m(k)$ is the target or desired response of the network at the m^{th} output node for the k^{th} training sample, and the factor ½ is inserted for latter mathematical convenience. We then have the average square error of K training steps, $\{x(k)\}_{k=1\ K}$

$$E = \frac{1}{2K}\sum_{k=1}^{K}\sum_{m=1}^{M}[t_m(k) - y_m(k)]^2 \qquad (6.18)$$

or in vector

$$E = \frac{1}{2K}\sum_{k=1}^{K}[\boldsymbol{t}(k) - \boldsymbol{y}(k)]^2 \qquad (6.19)$$

Similar to the LMS algorithm, the aim of backpropagation training is to minimise the cost function given by the average square error in (6.18). This can be done by applying the weight corrections $\Delta w_{mh}(k)$ that are proportional to the error gradient $-\partial E(k)/\partial w_{mh}$, i.e.

$$\Delta w_{mh}(k) = -\eta\frac{\partial E(k)}{\partial w_{mh}} \qquad (6.20)$$

where η is the learning rate. Formula (6.20) is known as the *delta rule* [99]. The partial derivative $\partial E(k)/\partial w_{mh}$ can be evaluated using the chain rule

$$\frac{\partial E(k)}{\partial w_{mh}} = \frac{\partial E(k)}{\partial net_m(k)} \cdot \frac{\partial net_m(k)}{\partial w_{mh}} \qquad (6.21)$$

Using (6.16) we have

$$\frac{\partial net_m(k)}{\partial w_{mh}} = \frac{\partial}{\partial w_{mh}}\sum_{h=0}^{H} w_{mh}a_h(k)$$
$$= a_h(k) \qquad (6.22)$$

Now if we define $\delta_m(k) = -\partial E(k)/\partial net_m(k)$, then (6.20) can be rewritten as

$$\Delta w_{mh}(k) = \eta\delta_m(k)a_h(k) \qquad (6.23)$$

To compute $\delta_m(k)$, again we use the chain rule expressed in term of output $y_m(k)$ as

$$\delta_m(k) = -\frac{\partial E(k)}{\partial y_m(k)} \cdot \frac{\partial y_m(k)}{\partial net_m(k)} \tag{6.24}$$

From (6.18) and (6.16), the two factors are obtained as follows:

$$\frac{\partial E(k)}{\partial y_m(k)} = -(t_m(k) - y_m(k)) \tag{6.25}$$

and

$$\frac{\partial y_m(k)}{\partial net_m(k)} = f_m^{'}(net_m(k)) \tag{6.26}$$

Hence, for any output-layer node m, the change in weight is

$$\begin{aligned} \Delta w_{mh}(k) &= \eta \delta_m(k) a_h(k) \\ &= \eta (t_m(k) - y_m(k)) f_m^{'}(net_m(k)) a_h(k) \end{aligned} \tag{6.27}$$

For units that are not output units, there is no available desired output. We then need a method for estimating the factor $\partial E(k)/\partial w_{hl}$ that is used to update the connection weights. In this case, we still have:

$$\begin{aligned} \Delta w_{hl}(k) &= -\eta \frac{\partial E(k)}{\partial w_{hl}} \\ &= -\eta \frac{\partial E(k)}{\partial net_h(k)} \cdot \frac{\partial net_h(k)}{\partial w_{hl}} \\ &= -\eta \frac{\partial E(k)}{\partial net_h(k)} x_l(k) \\ &= \eta \left(-\frac{\partial E(k)}{\partial a_h(k)} \cdot \frac{\partial a_h(k)}{\partial net_h(k)} \right) x_i(k) \\ &= \eta \left(-\frac{\partial E(k)}{\partial a_h(k)} \right) f_h^{'}(net_h(k)) x_l(k) \\ &= \eta \delta_h(k) x_i(k) \end{aligned} \tag{6.28}$$

where

$$\delta_h(k) = -\frac{\partial E(k)}{\partial a_h(k)} f_h^{'}(net_h(k)) \tag{6.29}$$

However, the partial $\partial E(k)/\partial a_h(k)$ cannot be evaluated directly, and needs to be evaluated in terms of quantities in the output layer

$$\begin{aligned} -\frac{\partial E(k)}{\partial a_h(k)} &= -\sum_{m=0}^{M} \frac{\partial E(k)}{\partial net_m(k)} \frac{\partial net_m(k)}{\partial a_h(k)} \\ &= \sum_{m=0}^{M} \left(-\frac{\partial E(k)}{\partial net_m(k)} \right) \frac{\partial}{\partial a_h(k)} \sum_{h=0}^{H} w_{mh} a_h \\ &= \sum_{m=0}^{M} \left(-\frac{\partial E(k)}{\partial net_m(k)} \right) w_{mh} \\ &= \sum_{m=0}^{M} \delta_m(k) w_{mh} \end{aligned} \tag{6.30}$$

The result obtained in (6.30) is incorporated into (6.29) to yield

$$\delta_h(k) = f_h^{'}(net_h(k)) \sum_{m=0}^{M} \delta_m(k) w_{mh} \tag{6.31}$$

Hence, in general the delta terms at an internal node can be evaluated in terms of the delta terms at an upper layer. Starting at the highest layer, i.e. the output layer, expression (6.27) yields the value of $\delta_m(k)$, and then we can propagate the "errors" backward to the lower layers for updating the connection weights.

For *epoch* training, we form an overall correction to the weights after each scan of all pattern pairs in the training set, that is

$$\Delta w_{hi} = \sum_{k=1}^{K} \Delta w_{hi}(k) \tag{6.32}$$

The derivatives of the activation functions can be calculated once given the transfer function. In particular, for a sigmoid transfer function at the hidden layer, we have

$$a_h(k) = f_h\ (net_h(k)) = \frac{1}{1+\exp(-net_h(k))} \tag{6.33}$$

then

$$\frac{\partial a_h(k)}{\partial net_h(k)} = f_h^{'}(net_h(k)) = a_h(k)[1 - a_h(k)] \tag{6.34}$$

and a linear transform function at the output layer gives

$$y_m(k) = f_m(net_m(k)) = net_m(k) \tag{6.35}$$

then

$$\frac{\partial y_m(k)}{\partial net_m(k)} = f_m^{'}(net_m(k)) = 1 \tag{6.36}$$

The backpropagation algorithm is a slow method for training feedforward networks. The rate of correction is controlled by the learning rate η, which when set too large may cause the system to oscillate and prevent the network's convergence. Often in gradient approaches, the learning rate is adjusted as a function of the iteration (e.g. $\eta^{(n)} = \eta^{(0)} / n$), so that it allows large initial connections, yet avoids weight oscillations around the minimum when near the solution.

There are several methods for speeding up the training processes of the backpropagation algorithm. One of the most commonly used methods is the *momentum* method that adds a momentum of weight update in the last iteration (n^{th}) to the correction weight in the current iteration $(n+1)^{\text{st}}$ [99], i.e.

$$\Delta^{(n+1)} w(k) = -\eta\left(\frac{\partial E(k)}{\partial w}\right) + \alpha \Delta^{(n)} w(k) \tag{6.37}$$

where α is the momentum constant and is restricted to the range $0 \leq \alpha < 1$, and is usually chosen to be between 0.9 and 0.99 for a high degree of momentum. In (6.37), the second term with a positive α, guides the correction weight at step $(n+1)^{\text{st}}$ to the same direction as at step n^{th}. This momentum term may prevent oscillations in the system and may help it to escape local minima of the error function in the training process.

6.3.3 Self-Organising Mapping (SOM) Algorithm

In [104,105], an alternative neural learning structure is proposed to perform a dimensionality reduction, in which the feature space is converted to yield *topologically ordered* similarity maps or clustering diagrams. In addition, the network uses a lateral unit interaction function to implement a form of local competitive learning. The mapping is achieved autonomously by the system, i.e. unsupervised training.

A one-dimensional (1-D) configuration of units that form feature dimensionality reducing maps is shown in Figure 6.2. Each unit receives the input pattern $\boldsymbol{x} = (x_1, x_2, \ldots x_N)^T$ in parallel. The mappings can be generalised to higher dimensions; for example, a 2-D topology yields a planar map indexed by a 2-D coordinate system.

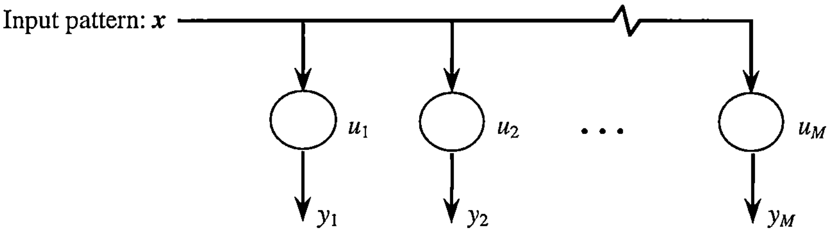


Figure 6.2: A one-dimensional topology mapping configurations

Since each unit u_m in the network receives the input pattern, $\boldsymbol{x} = (x_1, x_2, \ldots x_N)^T$, in parallel, they have the same number of weights, $\boldsymbol{w}_m = (w_{m1}, w_{m2}, \ldots w_{mN})^T$, as the dimension of the input vector. Given a large, unlabeled training set, the self-organise network adapts its neural clusters to reflect input pattern similarity. Then the overall structure may be viewed as "an array of matched filters, which competitively adjust unit input weights on the basis of the current weights and goodness of match" [94]. Each unit in the network competes with other units and tries to become a matched filter.

At a training iteration, an input pattern $\boldsymbol{x}(k)$ is presented to the network. A distance measure $D_m(k)$ between the pattern $\boldsymbol{x}(k)$ and $\boldsymbol{w}_m$, $m \in [1,M]$ is computed. The distance measure can be an inner product measure (i.e. correlation), Euclidean distance or another suitable measure that reflects the similarity of the input pattern. The Euclidean distant is normally used, and a *matching phase* is used to define a "winner" unit u_c, with its corresponding weight w_c so that

$$\|x(k)-w_c(k)\| = \min_m\{\|x(k)-w_m(k)\|\} \tag{6.38}$$

Hence, at iteration k, c is the index of the best matching unit for the input pattern $x(k)$. By defining a *topological neighbourhood* $N_c(k)$ of the winning unit u_c, units that are in $N_c(k)$ are considered active to the input pattern $x(k)$. Other units that are not in $N_c(k)$ are considered inactive to this input. Then units in the currently defined cell $N_c(k)$ are affected through the global network *updating phase* as [97]

$$w_m(k+1) = \begin{cases} w_m(k)+\eta(k)[x(k)-w_m(k)], & m \in N_c(k) \\ w_m(k), \; m \notin N_c(k) \end{cases} \tag{6.39}$$

where $\eta(k)$ is the learning rate at the iteration k. The update moves the weight vectors of the winning unit and of units in its neighbourhood towards the input vector $x(k)$.

The result of the accuracy of the mapping is dependent on the choices of the topological neighbourhood $N_c(k)$, learning rate $\eta(k)$, and the number of iterations [97,105]. The learning rate $\eta(k)$ should start with a value close to 1, and gradually decrease as the number of iterations increases. Similarly, the neighbourhood $N_c(k)$, starts with a large size, and then reduces the size for large k to allow fine-adjustment phase. As this is an unsupervised training, it usually requires a large training set as well as the number of iterations.

6.4 RADIAL BASIS FUNCTION APPROACH

In [106,107] a network structure, known as Radial Basis Function (RBF) network, of locally tuned nodes in the hidden layer is proposed. The network has only a *local* learning capability and is suitable for those patterns that have clear statistical distributions. Since it has many fewer connecting weights to be updated compared to the multilayer perceptrons, it is much *faster for training*. In fact, in an RBF network, the training of the two layers is decoupled while in an MLP network the training is iteratively coupled together.

6.4.1 The RBF Network Structure

Figure 6.3 shows a typical RBF neural network which has H processing nodes in the hidden layer and M summing nodes in the output layer. The N-dimensional feature vector serves as an input sample. The network has only one hidden layer and the fact that the hidden nodes receive input directly from the input layer, that is without having to calculate the weighted sums, makes it much faster to train than a backpropagation network of comparable size.

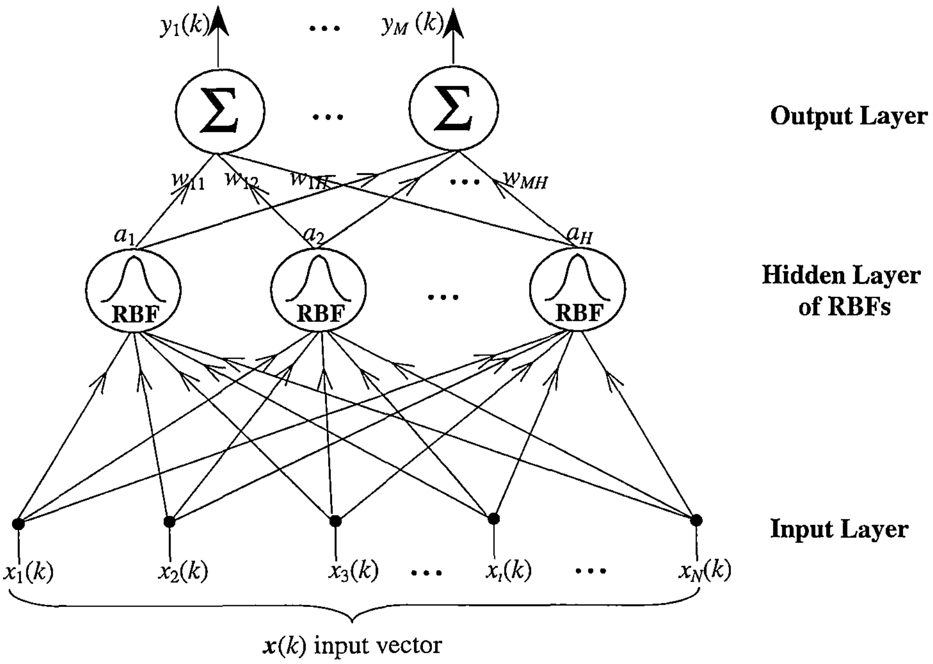


Figure 6.3: General structure of a Radial Basis Function Neural Network.

The non-linear *activation* function of the hidden nodes in an RBF neural network is *non-monotonic* in contrast to the monotonic sigmoid activation function of multilayer perceptrons. This function, also called the *receptive field* of the node, is usually a multi-dimensional symmetric Gaussian function centred on each node. The receptive fields of neighbouring nodes overlap. The dimension of the Gaussian activation function is equal to the dimension of the input data vector. For an input pattern $x(k)$, the activation output from the h^{th}, hidden node centred at μ_h is given by [106,107],

$$a_h = \exp\left(\frac{-\left\| \boldsymbol{x}(k) - \boldsymbol{\mu}_h \right\|^2}{\sigma_h^{\;2}} \right) \tag{6.40}$$

where σ_h is the distance scaling parameter which determines over what radial distance in the input space the node will have a significant influence. σ_h is also known as the *width* or the *spread* of the node. Note that σ_h has the similar function as the standard deviation in a normal probability distribution. Furthermore, $D_h(k) = \left| \boldsymbol{x}(k) - \boldsymbol{\mu}_h \right|$ is the Euclidean distance from the k^{th} data point to the centre of the h^{th} node given in (6.11), and $\left(D_h(k) / \sigma_h \right)$ is the scaled distance measured in terms of the width of the node.

The output layer of the RBF network consists simply of *linear summation* units with linear activation. The network output from the m^{th} node due to the data input vector $x(k)$ is therefore

$$y_m(k) = \sum_{h=1}^{H} w_{mh} a_h \tag{6.41}$$

where w_{mh} is the coefficient or the weight from the h^{th} hidden node to the m^{th}, $1 \leq m \leq \mathrm{M}$, output node. When the activation of the hidden nodes is Gaussian as in (6.40), the network output is

$$y_m(k) = \sum_{h=1}^{H} w_{mh} \exp\left(\frac{-\left\| \boldsymbol{x}(k) - \boldsymbol{\mu}_h \right\|^2}{\sigma_h^{\;2}} \right) \tag{6.42}$$

The output formula in (6.42) is exactly the same as the decision function of a Bayesian detector in a communication receiver, in which μ_h and $x(k)$ is the desired received signal and the actual received signal respectively and the variance of channel noise is $\sigma^2 = \sigma_h^2 / 2$. A Gaussian RBF neural network can therefore be used to realise a Bayesian detector [108,109].

However, in most practical applications, the distributions of features are normally different in different dimensions of the N-dimensional input space. Therefore, the Euclidean distance that incorporates with the scaling width for each node as in (6.40) is not flexible enough. We then generalise the activation to a non-symmetrical Gaussian receptive field by using the square of Mahalanobis distance in the Gaussian function as in (6.9). The activation output of the hth hidden node due to input sample $\boldsymbol{x}(k)$ in (6.40) becomes [91,92]

$$a_h(k) = \exp[-(\boldsymbol{x}(k) - \boldsymbol{\mu}_h)^T \Sigma_h^{-1} (\boldsymbol{x}(k) - \boldsymbol{\mu}_h)] \tag{6.43}$$

Σ_h is the distance scaling matrix of the node's receptive field that provides the width or the spread of influence of the node and is simply the covariance matrix of the training samples assigned to, or captured by, the h^{th} node cluster.

To guarantee the influence of the classification when moving across decision boundaries and to provide a relative meaning of confidence level for the classification, a constant is multiplied to all standard deviations $\{\sigma_{hi}\}$ so that the maximum activation output at decision boundaries has a value of less than 0.5. We can then reject a sample if it is classified to a class but with small confidence level. This is also an advantage of RBF networks over the networks of sigmoid transfer function for dealing with outliers.

In general, the choice of a non-linear activation function in an RBF network is not crucial for its performance. However, the performance of an RBF network critically depends upon the chosen centres [106]. Too few centres cause the network to be not capable of generating a good approximation for the model, while too many centres cause it to fit misleading variations due to ill-defined or noisy data. To generate the centres, an unsupervised clustering technique on the training set is used in [107], as there is no requirement of the class output. The technique provides no control over the model complexity. A direct approach to the model complexity issue is to select a subset of centres from a larger set which can be the entire training set. This approach is adopted in [110,111] by starting with an empty network and then adding centres one at a time until the model has accounted for a sufficient fraction of the variance of the data. An indirect approach [112,113] to controlling the model complexity is to use all the training set data as centres, then to use weight decay or ridge regression to reduce the effective number of free parameters, thus reducing overfit. The authors in [121,122,126] propose a technique that is based on *regression trees* [114,115,116] to generate RBF centres and their widths. The input space is divided into hyperrectangles organised into a binary tree that minimises the residual error between model and data.

In the following we present two typical methods for training the node centres and their width in an RBF network. One method uses unsupervised training known as *k-means clustering* technique [107], and the other uses a regression tree, whose advantage, besides its speed, is the possibility of interpreting decision rules in terms of individual features.

6.4.2 Unsupervised Training for RBF Networks

In applications where there is no prior knowledge about the desired states of the system, or no teacher is available, unsupervised training is the only choice. The training leaves the hidden nodes to compete with one another. The node, which responds strongest to the input training sample, wins the competition and obtains the training for its parameters. The parameters for an RBF neural network are determined in three steps. First, the centres $\{\mu_h\}$ of the hidden nodes can be determined by a k-means clustering technique. Second, the distance scaling matrix Σ_h of each node's receptive field is the covariance matrix of the training samples captured by the h^{th} node cluster. Then finally, the connection weights of the output layer are determined by a *multiple linear regression*.

k-Means Clustering of node centres

The *k-means clustering* algorithm finds a set of cluster centres and associated partition boundaries to best separate the training data into subsets or clusters. Each cluster centre becomes the centre of a node in the hidden layer of the RBF network. The algorithm finds a local minimum in the total squared Euclidean distances, $E_{k\text{-}mean}$, between the training data points $\boldsymbol{x}(k)$, $k = 1,2 \ldots K$, and the cluster centres μ_h, $h = 1,2,\ldots H$, that they are assigned to, i.e.

$$E_{k-mean} = \sum_{k=1}^{K} \min_{h} \left\| \boldsymbol{x}(k) - \boldsymbol{\mu}_h \right\|^2 \tag{6.44}$$

The algorithm achieves the above least squared (LS) minimisation iteratively using the following steps:

Step 1: Choose H points randomly among the given batch of K training data points (multidimensional samples) $\boldsymbol{x}(k)$, $k = 1,2 \ldots K$, and assign these to *initial* centres of H clusters $\mu_1, \mu_{2,} \;.\, \mu_H$.

Step 2: Assign each training sample to the cluster with centre nearest to it. This results in H clusters.

Step 3: Calculate the average position (centroid) of the training points for each of H clusters.

Step 4: Update $\{\mu_h\}$ to these new cluster centres (i.e. the centroids).

Step 5: Calculate $E_{k\text{-}mean}$ from (6.44) above.

Repeat steps 2 – 5 until $E_{k\text{-}mean}$ converges to an acceptable small level.

Note that techniques are available to speed up the convergence by a better initialisation of the hidden node centres in Step 1. If there is some prior knowledge about the clustering distribution of the training set, then the node centres may be strategically placed, making use of the prior knowledge.

Determination of connection weights of the output layer

Since the M output nodes are simple *linear* summation units (linear activation function) taking inputs from the H hidden nodes, the training can be done by the classical *linear least squares regression* technique. If all K training samples are available, the MxH weight matrix $\boldsymbol{w}$ can be simply *computed* so that the norm-2 of the output error is minimised, i.e.

$$E = \sum_{k=1}^{K} \left(\boldsymbol{T}(k) - \boldsymbol{wa}(k)\right)^2 \text{ minimised} \tag{6.45}$$

where T is the desired (target) MxK output matrix as the result of applying K training samples to the network, and a is the HxK activation matrix (output) from the hidden layer. This requires

$$\sum_{k=1}^{K} \left[\frac{\partial}{\partial w} \left(\boldsymbol{T}(k) - \boldsymbol{wa(k)}\right)^T \left(\boldsymbol{T}(k) - \boldsymbol{wa}(k)\right) \right] = 0 \tag{6.46}$$

Or in matrix form,

$$\boldsymbol{w} = \boldsymbol{Ta}^T(\boldsymbol{aa}^T)^{-1} \tag{6.47}$$

However, if real-time training is required or if the nodes in the output layer have a *non-linear activation* function, then iterative training using conventional gradient-search methods such as the LMS algorithm or Recursive-Least-Square (RLS) algorithms, may be used.

6.4.3 Supervised Training for RBF Networks

Since the training of the two layers in an RBF network is decoupled, a regression tree can be used to train the centres of the nodes in the hidden layer [121,122]. The connection weights of the output layer are determined by the multiple linear regression that is described above in the unsupervised section. Most techniques for implementation of decision trees decompose the training space into pairwise disjoint regions. Ideally each region contains samples of a single class.

Generating the Regression Tree

Using a regression tree, a binary tree divides the training space recursively into two and approximates the function in each half by the average output value of the data it contains [114]. A binary tree is shown in Figure 6.4, in which the root nodes of the tree $\{\mathbf{T}_4,\mathbf{T}_6,\mathbf{T}_7,\mathbf{T}_9,\mathbf{T}_{10},\mathbf{T}_{11}\}$ are the smallest hyperrectangles that include all of the training data, $\mathbf{T}_1 = \{\boldsymbol{x}(k)\}_{k=1\ K}$.

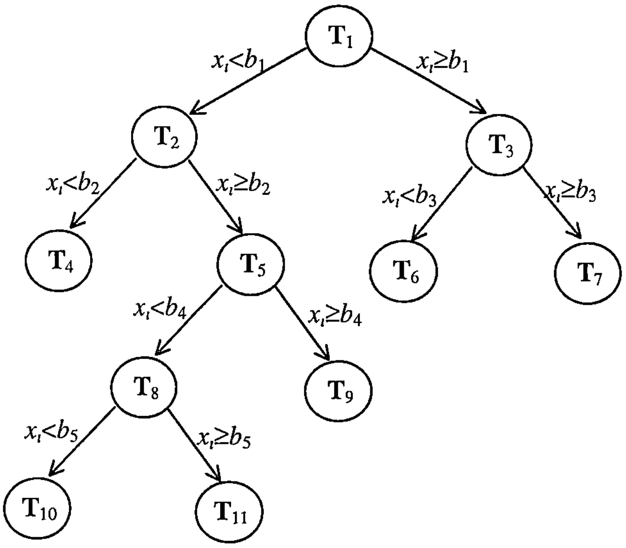


Figure 6.4: Training space is divided into hyperrectangles organised into a binary tree

At a bifurcation, a training set $\mathbf{T} = \{\boldsymbol{x}(k)\}_{k=1\cdot P}$ is divided into left and right subsets, $\mathbf{T}_L$ and $\mathbf{T}_R$, on either side of a boundary b in one of the dimensions i such that

$$\begin{aligned} \mathbf{T}_L &= \{\boldsymbol{x}(k) : x_i(k) < b\}, \\ \mathbf{T}_R &= \{\boldsymbol{x}(k) : x_i(k) \geq b\} \end{aligned} \tag{6.48}$$

The mean output value of each side of the bifurcation is

$$\bar{y}_L = \frac{1}{P_L}\sum_{k\in\mathbf{T}_L} y(k),$$
$$\bar{y}_R = \frac{1}{P_R}\sum_{k\in\mathbf{T}_R} y(k) \tag{6.49}$$

where P_L and P_R are respectively the number of samples in the left and right subset. Then the residue square error between the model and data after the bifurcation is

$$E(i,b) = \frac{1}{P}\left(\sum_{k\in\mathbf{T}_L}(y(k)-\bar{y}_L)^2 + \sum_{k\in\mathbf{T}_R}(y(k)-\bar{y}_R)^2\right) \tag{6.50}$$

The subsets (children) $\mathbf{T}_L$ and $\mathbf{T}_R$ of the root node $\mathbf{T}$ are created by finding the dimension i and a boundary b such that $E(i,b)$ is minimised. This can be done by a simple discrete search over N dimensions and P cases. Similarly, the children of the root node are split recursively. We terminate the process at a node if there is a sufficient degree of class purity within the node (i.e. the square error of data within the node smaller than an $E_{\min}$), or if the bifurcation of the node creates children that contain fewer than $P_{\min}$ samples. The training space is thus divided into hyperrectangles organised into a binary tree that minimises the residual error between model and data.

The regression tree technique is less sensitive to irrelevant attributes as they do not usually appear in the bifurcation of the regression tree. Then the network needs only to connect to relevant attributes, which reduces the links in the network.

Transforming Tree Nodes into RBFs

Suppose the input space is bounded, and the boundaries are given by the maximum values of the attributes in the training set. In [121], there are three options for the transforming of the partitioning regions (hyperrectangles) produced from the regression tree to the node centres $\{\mu_h\}$ of the RBF's hidden node:

1. If one and only one side of the region lies on the border of the input space, then μ_h is placed at the centre of this side.

2. If two adjacent sides of the region lie on the borders of the input space, then μ_h is placed into the corner defined by these sides.

3. If the region borders all sides with other regions, then μ_h is placed in the geometric centre of the region.

The sizes of the Gaussian function $\{\sigma_{hi}\}$ are calculated based on the sizes of the corresponding hyperrectangles so that the activation outputs $\{a_h\}$ of the hidden layer at the hyperrectangles' borders are always the same. If r_{hi} is the distance from the centre of the h^{th} node to its border in the i^{th} dimension, a new parameter α can be introduced so that

$$\sigma_{hi} = \alpha \, r_{hi} \tag{6.51}$$

The parameter α is the same for all nodes and for all dimensions. This holds the ratio between r_{hi} and σ_{hi} constant, thus keeping the same response $\{a_h\}$ at the borders of the hyperrectangles.

6.5 CONCLUSION

Depending on the available data and the properties of the feature vector, different suitable classifiers can be designed. In practice, the selection of a classifier is a difficult problem and is usually based on the choice of those which happened to be available, or best known to the user.

In [88], three different approaches used to design a classifier are identified. The simplest and most intuitive approach to classifier design is based on the concept of similarity, in which patterns can be classified by template matching or minimum distance measure using a few prototypes per class. Self-organise mapping or some other advanced techniques for computing prototypes such as vector quantisation [117,118], or learning vector quantisation (LVQ) [97], are based on this concept. Some techniques used in [134,135] employ LVQ as a classifier for classifying the

PQ disturbances, in which the DWT coefficients of the PQ disturbances are used as the input feature vector.

The second category of classifiers is based on probabilistic concept. It requires priori knowledge on class conditional densities, or at least an estimation of the densities so that an optimal decision rule for the classification can be produced. It is therefore suitable for patterns that have clear distributions between their classes and are obvious invariants of some kind of transformation of the signal (e.g. shift invariant).

The third approach used for designing pattern classifiers is to construct decision boundaries directly by optimising the error criterion. Feedforward networks or MLPs are examples of this type where the training procedures aim to minimise the mean squared error between the classifier outputs and the predetermined target values. These networks, however, are slow in learning and suffer from the possibility of being trapped in local minima of the chosen optimisation cost function. Moreover, the *monotonic* nature of the hidden layer's activation in MLPs gives rise to hyperplane decision boundaries. This, together with global learning ability, makes these networks more prone to *extrapolation error* (i.e. when a test sample falls outside the range of the training set). Whereas in an RBF, the *non-monotonic* nature of the radial basis function allows the network to produce robust hypersphere decision surfaces, classification error in the RBF networks is usually due to *decision error* alone (i.e. due to stochastic overlap of the classes), while error in MLP networks is due to both decision and extrapolation sources. An interesting comparative study of the performance of RBF networks compared to MLP networks is reported in [119].

In practice, since the collections of power quality disturbances from transmission lines are costly and time consuming, limited data is available for training. In order to generalise the problem from the small set of training data, a construction of pattern recognition techniques that are invariant to specified transformations of the input data is required. The techniques lead to the use of a statistical approach rather than a "black box" neural network. This is also a reason why we use the radial basis function network to utilise the most information from the statistics distribution of the shift-invariant features that we designed for the PQ disturbance classification problem.

Despite the advantages of the RBF network as a classifier for the classification of power quality disturbances, it has the weakness of having only a local learning capability and a limited learning inference from the training data. In fact, hidden units in an RBF network receive input directly from the input layer, and are simply "weighting" the input features by the node's scaling parameters $\{\sigma_{hi}\}$. This has been proved as not being sufficient to enhance the discriminant of relevant or dominant features or to eliminate the effect of irrelevant features for each particular class [123,124,125]. In the next chapter, we propose some algorithms to improve the classification process of an RBF network, and optimise the decision boundary in the training of the network.

Chapter 7

OPTIMAL LEARNING FOR PATTERN CLASSIFICATION IN RBF NETWORKS

7.1 INTRODUCTION

In an RBF network, the crucial concern is the selection of cluster centres $\mu_h = \{\mu_{hi}\}$ and their widths $\{\sigma_{hi}\}$ so that the model can fit closely to the training space. However, current techniques give suboptimum positions of cluster centres and their widths. For example, for the *k-means* (unsupervised) clustering technique in [120], since it does not require the knowledge of the output (targets) the technique cannot give the optimum cluster centre positions if the training space is highly overlapped between their classes. In the regression tree technique used in [121,122], the centres are simply placed at the middles of the hyperrectangles.

Also, in many applications, some features of a pattern are more important or more discriminating than other features, e.g. the formant frequencies of a voiced sound, the dominant components in a principal component analysis etc. The pattern matching gives more weight to these components in the feature vector. In a typical RBF network, the node's widths $\{\sigma_{hi}\}$ also play the role of weights for the input features. They are however, not sufficient to enhance the discriminant of relevant or dominant features or to eliminate the effect of irrelevant features for each particular class [123,124,125].

In this chapter, we propose to modify the structure of the RBF network by introducing weighting for the input features (in contrast to the direct connection of the input to the hidden layer of a conventional RBF) so that the training space in the RBF network is adaptively separated by the resultant decision boundaries and class regions. The estimation of the input weights can be carried out by one of the two techniques: the *knowledge-based* technique [123,124] and the training technique that trains the network as for a single layer perceptron together with the clustering process [125]. In this way the network has the ability to deal with complicated problems that have a high degree of interference in the training data, and achieves a higher classification rate over the current classifiers using a conventional RBF.

7.2 THE PROPOSED RBF NETWORK

7.2.1 The Network Structure

Figure 7.1 shows the proposed RBF network with the additional weight matrix in the input layer which is in contrast to the direct connection of the input to the hidden layer of a conventional RBF. The network has H processing nodes in the hidden layer and M summing nodes in the output layer. The input sample is an N-dimensional vector.

Similar to a conventional RBF network, the non-linear activation function of the hidden nodes is non-monotonic and normally is a multi-dimensional Gaussian function centred on the node. The dimension of the Gaussian activation function is equal to the dimension of the input data vector. The output of the h^{th} hidden node due to input sample $\boldsymbol{x}(k)$ is therefore given as

$$a_h(k) = \exp[-D_h{}^2(k)] \tag{7.1}$$

where $D_h^2(k)$ is the squared distance from the input $\boldsymbol{x}(k)$ to the h^{th} node centre $\boldsymbol{\mu}_h$ ($1 \leq h \leq H$).

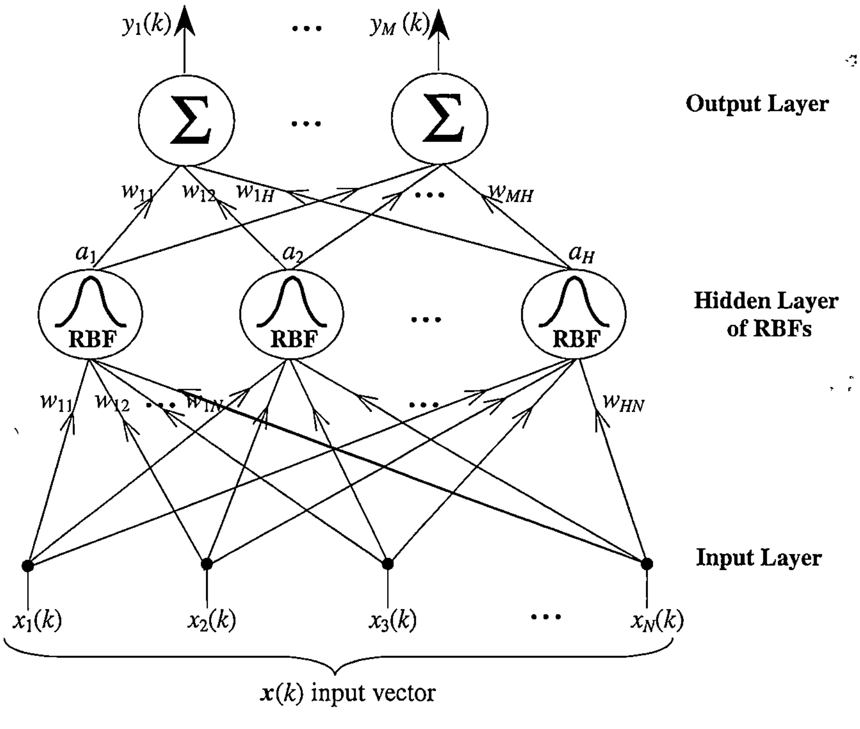


Figure 7.1: The proposed RBF network with the input layer weights

We define a distance function, which includes the feature weights, from an input vector $\boldsymbol{x}(k)$ to the h^{th} class, centred at μ_h as

$$D_h{}^2(k) = \sum_{i=1}^{N} \frac{w_{hi} \mid x_i(k) - \mu_{hi} \mid^2}{\sigma_{hi}^2} \tag{7.2}$$

It is obvious from (7.2) that the input layer connection weights are equivalent to the feature weights of the input vector (or equal to the square of the latter, to be precise). Note that, for a given problem the samples $\boldsymbol{x}(k)$ are known, the cluster centres μ_h and their widths $\{\sigma_{hi}\}$ can be determined by one of the current techniques (e.g. we used the regression tree [121,122] in this paper). As a result, the set of weights $\{w_{hi}\}$ in (7.2) is the only parameter that can be trained in the RBF network so that the training space is adaptively separated by the resultant decision boundaries and class regions. We can verify that by incorporating the weights $\{w_{hi}\}$ into the scaling widths $\{\sigma_{hi}\}$, then obtaining a new "effective" width

$$(\sigma_{hi})_{new} = \frac{\sigma_{hi}}{\sqrt{w_{hi}}} \tag{7.3}$$

Since $\{w_{hi}\}$ are feature weights, their values are restricted to positive numbers. The output layer of the RBF network consists simply of linear summation units with linear activation. The network output from the m^{th} node due to the data input vector $\boldsymbol{x}(k)$ is given by (6.41) as

$$y_m(k) = \sum_{h=1}^{H} w_{hm} a_h(k) \tag{7.4}$$

7.2.2 Initialisation of the RBF Network

Initialisation of the RBF network involves the determination of the network structure (i.e. the number of nodes on each layer) and the estimation of each cluster centre and width in each hidden node. The number of nodes of the input layer and output layer are simply the size of the input vector and the number of the output respectively. The crucial concern pointed out in [106] is the choice of centres (i.e. the number of centres or the number of hidden nodes and their positions). As used in [121,122,126], the simplest and quickest way is based on the regression tree technique to initialise the RBF centres and their widths. As shown in Section 6.4.3, the technique divides the input space into hyperrectangles organised into a binary tree that minimises the residual error between model and data. From the tree nodes, we then perform a transformation into hidden nodes of the RBF network. We use the algorithm as in [122] (but different to the algorithm used in [121]) to initialise our network by placing the node centres at the centres of the hyperrectangles specified by the tree nodes. The originality of our training technique is that all three parameters - μ_h, $\{\sigma_{hi}\}$ and $\{w_{hi}\}$ are optimally determined together and concurrently adjusted during training iterations to maximise the discriminant between classes, thus minimising the classification error [123,124,125].

7.3 KNOWLEDGE-BASED TECHNIQUE FOR TRAINING FEATURE WEIGHTS

The fundamental idea of the *knowledge-based* technique is to make use of *experiences* obtained from earlier problem solving situations in a similar context [127,128]. Applying to our problem, we propose to use the knowledge from training data to train the feature weights $\{w_i\}_{i=1\cdot N}$. This means that in terms of the discriminant function, each feature contributes an equivalent effect to all classes, i.e. $w_{hi} = w_i$, $\forall h \in [1,H]$. The distance in (7.2) is thus reduced to

$$D_h^{\,2}(k) = \sum_{i=1}^{N} w_i \frac{|x_i(k) - \mu_{hi}|^2}{\sigma_{hi}^{\,2}} \tag{7.5}$$

and the feature weights in (7.5) are normalised so that $\sum w_i = 1$.

7.3.1 Construction of the Knowledge-Base

We start with the construction of a knowledge-base by examining the interclass relative distribution of the feature components $\{x_i\}_{i=1\,N}$ of x and ranking these components in the order of their interclass discriminating power for every pair of classes. Like the decision tree method, this can be done by first finding a decision threshold $b_{pq}(i)$ in a dimension i for any pair of classes p and q so that the error, $E_{pq}(i)$ of the model on each side of the decision threshold is minimised. Then the order of the features' interclass discriminant power $\{i\}$ between classes p and q is simply the inverse order of the minimum error $E_{pq}(i)_{\min}$.

In the following example, we construct a knowledge-base to classify four types of PQ disturbances: impulse transient (IT), high frequency capacitor switching (HF), low frequency capacitor switching (LF) and aperiodic notch (NT). We select a feature vector of five components $\boldsymbol{x} = (\bar{s}, \sigma_s, \bar{\alpha}, \sigma_\alpha, L)$ (their definitions are presented later in Chapter 8), whose distributions in the training set are respectively shown in Figure 7.2 to 7.6. Figure 7.2 shows the average scale $\bar{s}$, Figure 7.3 for the standard deviation of the scale σ_s, Figure 7.4 for the average Lipschitz exponent $\bar{\alpha}$, Figure 7.5 for the standard deviation of the Lipschitz exponent σ_α and Figure 7.6 for the disturbance duration L. The training set contains 134 samples, in which the first 29 samples are IT disturbances, the next 35 samples are HF disturbances, then comes to 35 LF disturbances and the last 35 samples are NT disturbances.

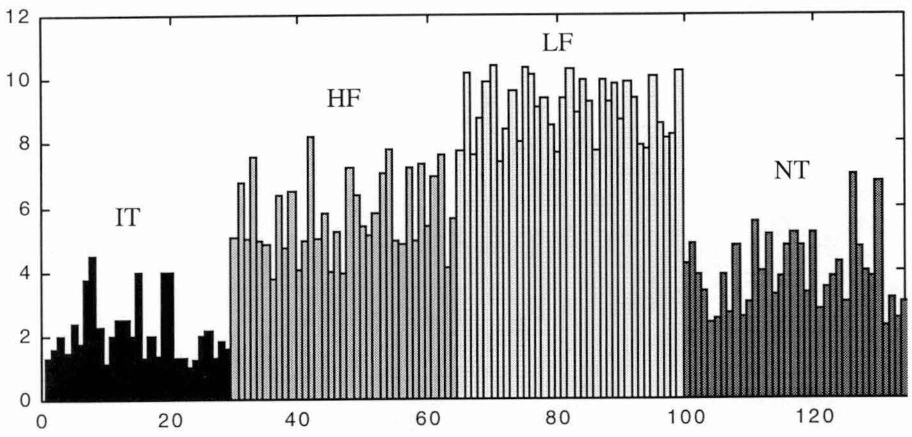


Figure 7.2: Distribution of the average scale $\bar{s}$ in the training set

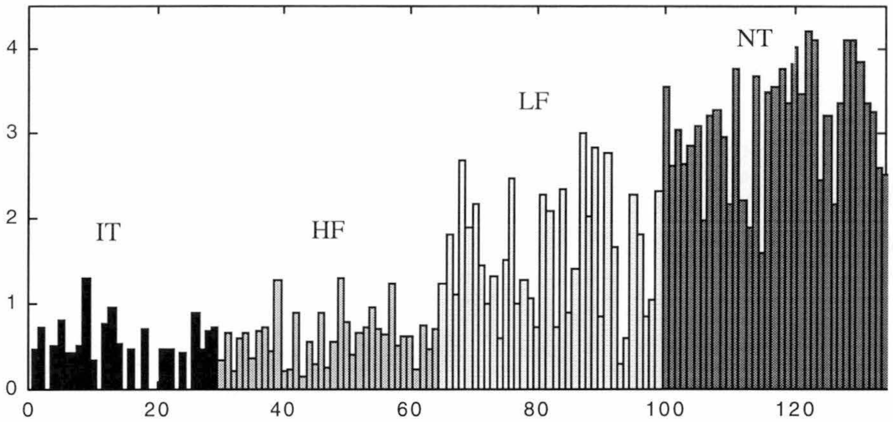


Figure 7.3: Distribution of the standard deviation of the scale σ_s in the training set

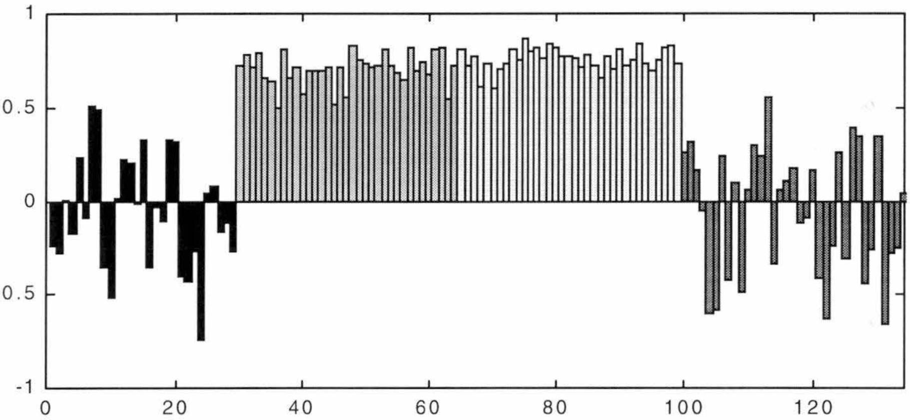


Figure 7.4 Distribution of the average Lipschitz exponent $\overline{\alpha}$ in the training set

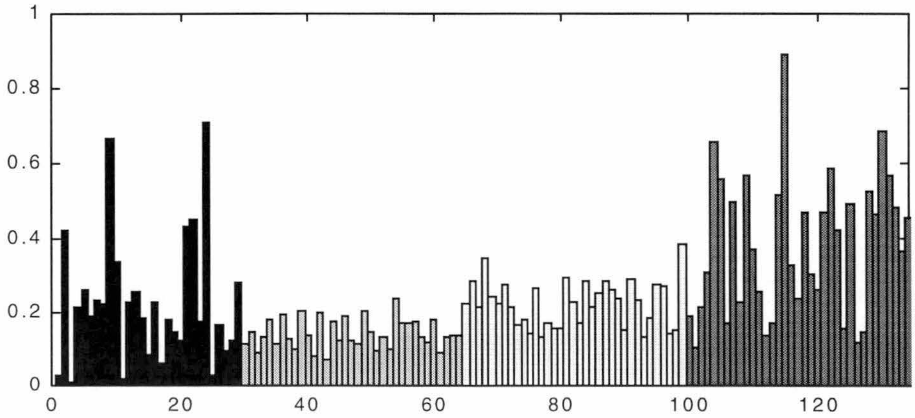


Figure 7.5: Distribution of the standard deviation of the Lipschitz exponent σ_α in the training set

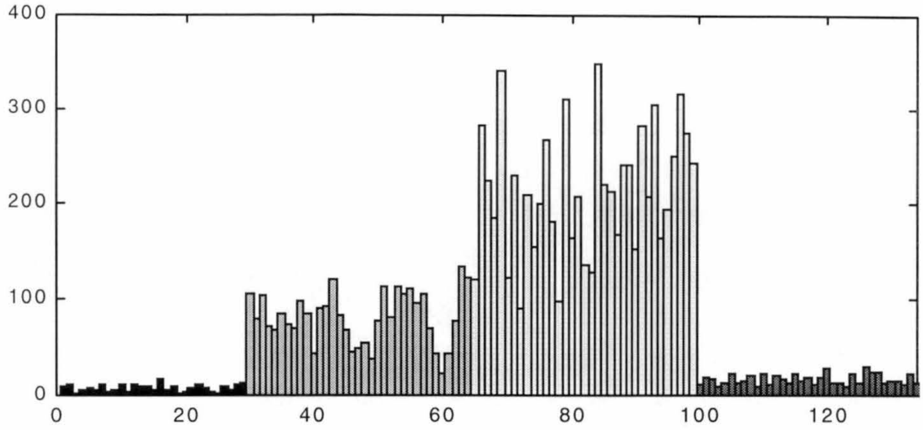


Figure 7.6: Distribution of the disturbance duration L in the training set

By examining the interclass relative distribution of the five feature components $(\bar{s}, \sigma_s, \bar{\alpha}, \sigma_\alpha, L)$ in the training samples presented in the above figures, we can rank these components in the order of their interclass discriminating power for every pair of disturbances as shown in Table 7.1 below.

Pair of disturbance types	Interclass discriminating power of features (decreasing from left to right)
IT ⇔ HF	$L,\ \bar{\alpha},\ \bar{s},\ \sigma_\alpha,\ \sigma_s$
IT ⇔ LF	$L,\ \bar{s},\ \bar{\alpha},\ \sigma_s,\ \sigma_\alpha$
IT ⇔ NT	$\sigma_s,\ L,\ \bar{s},\ \bar{\alpha},\ \sigma_\alpha$
HF ⇔ LF	$\bar{s},\ L,\ \sigma_s,\ \bar{\alpha},\ \sigma_\alpha$
HF ⇔ NT	$L,\ \sigma_s,\ \bar{\alpha},\ \bar{s},\ \sigma_\alpha$
LF ⇔ NT	$L,\ \bar{\alpha},\ \bar{s},\ \sigma_s,\ \sigma_\alpha$

Table 7.1: Ranking of feature components in the order of their interclass discriminating power.

The idea of training feature weights from this knowledge-base is that features are competitive in gaining their weights based on their interclass discriminating power for every pair of disturbances. Depending on the number of mis-assigned samples between clusters of each pair of classes, we make an increment on the feature weights for those having high interclass discriminating power and make a decrement on the feature weights for those having low interclass discriminant power.

7.3.2 Training the Feature Weights

In order to calculate the amount of weight update after each updating of the cluster centres and their widths in ***Step 4*** below, we determine the total number of mis-assigned samples between clusters of each pair of classes. At step j of the iteration, let $e_{pq}^{(j)}$ be the percentage of mis-assigned samples between class p and class q, and $w_{pq}^{(j)}$ be the feature weight vector, whose feature elements are arranged in the order of interclass discriminating features between class p and class q (Table 7.1). Then the update value of feature weights $w_{pq}^{(j+1)}$ at step (j+1) is [124]

$$w_{pq}^{(j+1)} = w_{pq}^{(j)} + \beta e_{pq}^{(j)} \boldsymbol{A} \tag{7.6}$$

where β is the learning rate, $\boldsymbol{A}$ is a vector that defines the amount of weights adjustment in w_{pq}, and that its elements are in a decreasing order which is based on the order of discriminating features of each pair of classes, e.g. A = [1; 0.5; 0; -0.5; -1] for a space of five input features. In (7.6), the term e_{pq} allows a faster training rate at the beginning when there is a large error, and reduces the training rate when the clustering error is smaller. During the training, we ensure that all feature weights $\{w_\iota\}$ are positive and their sum is normalized to 1.

As mentioned above, the training is an iterative algorithm, which produces simultaneously the RBF cluster centres μ_h, their scaling widths $\{\sigma_{h\iota}\}$ and the feature weights $\{w_\iota\}$ by the following steps.

Step 1- Initialisation: Using the regression tree technique, we initialise the node centres and their widths corresponding to the centres of the tree nodes and the size of the Gaussian $\{\sigma_{h\iota}\}$ as in (6.51). The initial feature weights $\{w_i\}$ are the same for all features, i.e. w_ι=1/N.

Step 2: Assign each training sample to the cluster with the centre *nearest* to it according to the distance in (7.5). This results in H clusters.

Step 3: Calculate the average position (centroid) and the variance (width) σ_{hi}^2 for each of the H clusters.

Step 4: Update $\boldsymbol{\mu}_h$ to these new cluster centres and width σ_{hi}^2 to its new values.

Step 5: Calculate the clustering error for all pairs of disturbance types.

Step 6: Adjust feature weights $\{w_i\}$ according to the look-up Table 7.1 and equation (7.6).

Repeat Steps *2 to 6* ***until*** the iteration converges to a minimum clustering error.

After the training, we obtain an optimal weighting level for each feature, as well as the node centres and their widths. There are some limitations on this training technique – these are the requirement for the knowledge of the training data, and the assumption that the feature weights are the same for all classes. In most practical problems, the relative importance of features is different for different classes, hence, the generalised form of the feature weight is a matrix, i.e. $\{w_{hi}\}$. In the next section, we present a training technique that performs such the generalisation.

7.4 GENERALISED TECHNIQUE FOR TRAINING THE INPUT LAYER WEIGHTS

We propose to train the input weights $\{w_{hi}\}$ as in a single-layer perceptron, which only involves the input layer and hidden layer of the RBF in the training [125]. It turns out that we need to define a concept label of each training input $\boldsymbol{x}(k)$ for the output of the hidden layer since the concept label of $\boldsymbol{x}(k)$ is for outputs of the output layer but is not available at the hidden layer.

In an RBF network, the hidden layer is the most important layer providing most of the clustering power of the classification process. The output layer of the RBF network consists simply of linear summation units with linear activation. This layer provides only a minimal contribution of the classification. Consequently, for an input sample $\boldsymbol{x}(k)$ that belongs to cluster c^{th}, its classification output is correct with a high confidence rate only if its distance to the centre of a class h^{th}, $D_h^2(k)$, is smallest for $h = c$ and larger for $h \neq c$. This means that the activation output $a_h(k)$ will respond the most at the node $h = c$ and respond less at the other nodes $h \neq c$ (the smaller the response of these nodes will result in higher discrimination). For this reason, for the k^{th} training sample we normalise the activation $\{a_h(k)\}$ of all nodes to their highest value $(a_h(k))_{\max}$ and produce a new variable $o_h(k)$ whose values vary in a range of [0,1]

$$o_h(k) = \frac{a_h(k)}{(a_h(k))_{\max}} \tag{7.7}$$

Now we can have a concept label of the k^{th} input sample for the hidden layer, that is the target $t_h(k)$ of the normalised activation output $o_h(k)$. For the correct hidden node c^{th}, the desired target $t_c(k)$ is '1', and $t_h(k)$ is '0' at other hidden nodes h^{th} $(h \neq c)$.

Note that, if there is overlapping between classes in the training space, we cannot guarantee the purity of samples within a cluster. However its majority samples belong to the class that it is assigned to (depending on the distribution, a class can be represented by more than one cluster in the training space). Then in the training process, the output target of the hidden node that associates with that cluster is '1' for the majority of samples plus the samples of that class within a neighbourhood. For other samples this target is '0'.

The output error of the h^{th} hidden node in response to the k^{th} training sample is

$$e_h(k) = t_h(k) - o_h(k) \tag{7.8}$$

and the squared error of the network in response to the k^{th} pattern is

$$E(k) = \frac{1}{2}\sum_{h=1}^{H}(t_h(k) - o_h(k))^2 \tag{7.9}$$

Using a gradient descent approach, we minimise the squared error by iteratively adjusting the weights according to [99,129]

$$\Delta w_{hi}(k) = -\eta \frac{\partial E(k)}{\partial w_{hi}} \tag{7.10}$$

where η is a positive constant, referred to as the learning rate, which determines the portion of weight change that will be used for the correction. Depending on the characteristics of the error surface, different values of learning rates can be used. The derivative $\partial E(k)/\partial w_{hi}$ can be evaluated using the chain rule

$$\frac{\partial E(k)}{\partial w_{hi}} = \frac{\partial E(k)}{\partial o_h(k)} \frac{\partial o_h(k)}{\partial a_h(k)} \frac{\partial a_h(k)}{\partial (D_h(k))^2} \frac{\partial (D_h(k))^2}{\partial w_{hi}} \tag{7.11}$$

Each partial in (7.11) can be evaluated as following. From (7.9),

$$\frac{\partial E(k)}{\partial o_{hi}} = -(t_h(k) - o_h(k)) \tag{7.12}$$

Since all the activation outputs $\{a_h(k)\}$ are normalised in (7.7) to their highest value $(a_h(k))_{\max}$, we can consider the value $(a_h(k))_{\max}$ as a constant of each k^{th} training sample rather than a variable of $\{a_h(k)\}$. So that the partial

$$\frac{\partial o_h(k)}{\partial a_h(k)} = \frac{1}{(a_h(k))_{\max}} \tag{7.13}$$

From (7.1) and (7.2)

$$\frac{\partial a_h(k)}{\partial (D_h(k))^2} = -a_h(k) \tag{7.14}$$

and

$$\frac{\partial (D_h(k))^2}{\partial w_{hi}} = \frac{| x_i(k) - \mu_{hi} |^2}{{\sigma_{hi}}^2} \tag{7.15}$$

Finally (7.11) becomes

$$\frac{\partial E(k)}{\partial w_{hi}} = o_h(k)(t_h(k) - o_h(k)) \frac{| x_i(k) - \mu_{hi} |^2}{{\sigma_{hi}}^2} \tag{7.16}$$

Then the updating weight formula in (7.10) can be written as

$$\Delta w_{hi}(k) = -\eta o_h(k)(t_h(k) - o_h(k))\frac{|x_i(k) - \mu_{hi}|^2}{\sigma_{hi}^2} \tag{7.17}$$

For *epoch* training, we form an overall correction to the weights after each scan of all pattern pairs in the training set, that is

$$\Delta w_{hi} = \sum_{k=1}^{K} \Delta w_{hi}(k) \tag{7.18}$$

Like the back-propagation algorithm presented in the previous chapter, the learning rate is adjusted as a function of the iteration (e.g. $\eta^{(n)} = \eta^{(0)}/n$), so that it allows large initial connections, yet avoids weight oscillations around the minimum when near the solution. In order to speed up the training process, at $(n+1)^{st}$ iteration, a momentum of weight update in the last iteration (n^{th}) is added to the correction weight as in (6.37), i.e.

$$\Delta^{(n+1)} w_{hi}(k) = -\eta\left(\frac{\partial E(k)}{\partial w_{hi}}\right) + \alpha\Delta^{(n)} w_{hi}(k) \tag{7.19}$$

This momentum term may prevent oscillations in the system and may help it to escape local minima of the error function in the training process.

After an epoch training of the weights, each training sample is assigned to the centre nearest to it according to the distance in (7.2). From this new set of clusters, each cluster centre μ_h and its width $\{\sigma_{hi}\}$ are updated, respectively, to the mean and the variance of the samples in its new cluster. The training is repeated until it converges.

Determination of the connection weights at the output layer in the network is similar to that of a conventional RBF network, where the classical linear least square regression technique is used to give

$$w = \boldsymbol{T}\mathbf{a}^T(\mathbf{aa}^T)^{-1} \tag{7.20}$$

7.5 NUMERICAL DEMONSTRATION & CONCLUSION

With the effective initialisation by the regression tree and the efficient clustering of the shift-invariant data, we have found that a training sample population of about 25-35 per hidden node is adequate for the RBF network for the knowledge-based training technique to converge after 6-8 iterations, and the network with the generalised training technique to converge after 3-5 iterations without being trapped into local minima. The classification results by different RBF networks are shown in the three tables below. Table 7.2 shows the classification results for a conventional RBF network (no feature weights), whose network parameters are initialised by the regression tree. Table 7.3 shows the classification results of the network with the knowledge-based training for the feature weights $\{w_i\}$, and Table 7.4 is for the network with the generalised training for the input layer weights $\{w_{hi}\}$. Due to the overlapping in distribution of the features between different disturbance types, the conventional RBF network has a high overall classification error rate of 7.7% (i.e. 15 out of 196 testing samples). In comparison, the network trained by the knowledge-based technique has a significant improvement in the classification error rate, which is only 3.6% (i.e. 7 out of 196 testing samples). Finally, as expected we have the lowest classification error rate (of 3.1%) with the network that is trained to obtain optimal features weights for each individual hidden node, together with the optimisation of the node centres and their widths. To compare the performance with a feedforward network, the same training and testing data sets are used for a backpropagation network of the same network size. The classification results of the back-propagation network are shown in Table 7.5, which has an overall classification rate of 4.6%.

Output Input	*IT*	*HF*	*LF*	*NT*	Total tested samples	Error rate [%]
IT	42	1	0	3	46	8.7
HF	0	47	3	0	50	6.0
LF	0	1	49	0	50	2.0
NT	1	4	2	43	50	14.0

Table 7.2: Classification results for a conventional RBF initialised by a regression tree

Output Input	*IT*	*HF*	*LF*	*NT*	Total tested samples	Error rate [%]
IT	43	0	0	3	46	6.5
HF	0	49	1	0	50	2.0
LF	0	0	50	0	50	0.0
NT	0	1	2	47	50	6.0

Table 7.3: Classification results for the RBF network with the knowledge-based training technique

Output Input	*IT*	*HF*	*LF*	*NT*	Total tested samples	Error rate [%]
IT	43	1	0	2	46	6.5
HF	0	50	0	0	50	0.0
LF	0	2	48	0	50	4.0
NT	0	0	1	49	50	2.0

Table 7.4: Classification results for the RBF network with the generalised training technique

Output / Input	*IT*	*HF*	*LF*	*NT*	Total tested samples	Error rate [%]
IT	44	0	0	2	46	4.3
HF	0	50	0	0	50	0.0
LF	0	4	46	0	50	8.0
NT	2	0	1	47	50	6.0

Table 7.5: Classification results for the backpropagation network

We have successfully demonstrated that there is a significant improvement in the classification results when appropriate feature weights, node centres and their widths are found for the RBF network. This increases the discriminant level between classes and obtains optimal decision boundaries in the training space in terms of minimising the mean squared error. As the result, the classification of the RBF network with the generalised training technique achieves the best overall classification error rate by only 40% of that for the conventional RBF network trained with regression tree and 67% of that for the backpropagation network.

The regression tree technique is less sensitive to irrelevant attributes as they usually do not appear in the bifurcation of the regression tree. Thus reduces the number of links in the network by only connecting to relevant features. This is however reduces the classification rates of the network since less relevant features, when incorporated with appropriate feature weights to different classes, do in fact increase the discriminant level between classes.

If the training of the input weights is involved, it takes longer to train the RBF network. However, the time taken to train the proposed RBF network is still much faster than that of a backpropagation network of the same network size. This is because the training of the input and output layer in the proposed RBF network is decoupled, and the initialisation of the network by the regression tree technique allows the network to train with a minimal number of training epochs without being trapped into local minima.

Chapter 8

APPROPRIATE SIGNAL PROCESSING TOOLS FOR THE CLASSIFICATION OF POWER QUALITY DISTURBANCES

8.1 APPROPRIATE SIGNAL PROCESSING TOOLS

In order to classify different types of disturbances that may present on a power supply, we aim to employ appropriate signal processing tools, which have the capability of measuring or at least closely extracting information on the three most important PQ disturbance attributes. They are *spectral content, duration and magnitude* whichever is appropriate for each category of power quality disturbances as shown in Chapter 2. In particular, the spectral content and the magnitude attributes are used for classifying steady state PQ disturbances, while the spectral content, the duration and some particular structures of the disturbance are used for classifying transient PQ disturbances.

Since disturbances in power supply range from the sustained long-duration, low and steady frequency type such as voltage sag, voltage swell, dc offset and voltage fluctuation, to the very short-duration and very high frequency type such as impulses, notches, and oscillatory switching transients, any linear expansion in a single basis is not flexible enough to characterise different types of disturbances. Then different appropriate signal processing techniques must be used in order to precisely measure the disturbance characteristics of different types [65,66].

It is well known that Fourier transform is best for the analysis of stationary signals, e.g. the 50 Hz main supply and its harmonics. However, it does not have the temporal characteristic to cope with sharp changes and discontinuities in signals, and suffers from the lack of time localisation. Therefore, it is not really suited to the analysis of non-stationary signals. On the other hand, wavelet-based techniques are sensitive to sudden changes in amplitude and are used to extract local characteristics such as edges, discontinuities and instantaneous frequencies in transient disturbances. Using wavelet transform, a signal can be analysed locally in both the time and frequency domains.

This Chapter presents a method that employs a combination of Fourier-based and wavelet-based transforms in the detection and classification of PQ disturbances so that the method has the ability to classify a wide range of power disturbances [66,71]. Moreover, for achieving a stable and efficient classification process, we aim to use appropriate signal processing tools that allow us to precisely measure the disturbances characteristics and present them in a compact and translation-invariant feature vector. In our work, we consider 10 types of disturbances.

1. Impulse transient
2. Low frequency capacitor switching
3. High frequency capacitor switching

4. Aperiodic notch
5. Interruption
6. Sag (dip)
7. Swell
8. Harmonics
9. Voltage fluctuation
10. Power frequency variation

We propose to use wavelet-based analysis and the RBF classifier to classify the first four types of disturbances as they have fast variations (high frequencies and non-stationeries) and most of them have short duration. The remaining six types, which contain low and steady frequency components, are classified using Fourier transform and rule-based expert systems.

8.2 FOURIER TRANSFORM FOR STEADY STATE DISTURBANCES

The Discrete Fourier Transform (DFT) of a discrete-time signal $f(n)$ is defined as:

$$F(k)=\frac{1}{N}\sum_{n=0}^{N-1} f(n)W_N^{nk} \qquad n,k=0,1,\ldots,N-1 \tag{8.1}$$

where $W_N = e^{-2\pi/N}$ and W_N^k is known as the twiddle factors of the DFT. The twiddle factors are periodic and define N equally spaced points around the unit circle at frequency increments of ω_s/N, where ω_s is the sampling rate of the input signal sequence. Therefore, the set of frequency samples, which defines the spectrum $F(k)$, is given on a frequency axis whose discrete frequency locations are given by

$$\omega_k = k\frac{\omega_s}{N} \qquad k=0,1,\ldots,N-1 \tag{8.2}$$

Therefore, using a sampling rate ω_s of 12.8kHz and capturing a 1024 samples length, this gives a frequency resolution of 12.5Hz in the transform domain. The sampling theory gives $\|F(k)\| = \|F(N-k)\|$, $0 \leq k \leq N/2$, which are both the responses of the frequency component ω_k.

Two properties of DFT, *linearity* and *circular shift* provide the ability to analyse each particular component in the signal and the periodic components. The linearity property is a key property that allows us to compute the DFTs of several different signals and determine the combined DFT via the summation of the individual DFTs. Hence, the system output frequency response can be easily evaluated for specific frequency components. For a signal of M components, this property is given by

$$DFT\left[\sum_{\iota=1}^{M} \alpha_\iota f_\iota(n)\right] = \sum_{\iota=1}^{M} \alpha_\iota F_\iota(k) \tag{8.3}$$

where $F_\iota(k) = DFT[f_\iota(n)]$ and α_ι are arbitrary constants. The circular shift is a shift in a finite-length sequence viewed as a periodic extension of the finite sequence. If $\{F(k)\}$ is the DFT of $\{f(n)\}$, then the DFT of $\{f(n+m)\}$ is given by

$$DFT\{f(n+m)\} = \{W_N^{-km} F(k)\} \tag{8.4}$$

This implies that the DFT of $\{f(n+m)\}$ has the same amplitude but a different phase to the DFT of $f(n)$. Consequently, all periodic components whose frequency is a multiple of 12.5Hz (this includes 50Hz power frequency and its harmonics), have their DFT magnitudes unchanged but with a simple phase shift. These DFT magnitudes are also proportional to their corresponding signal amplitudes in the time domain.

We now explain in detail the classification of Type 5 to Type 10 disturbances. We take the Fourier transform of 1024 samples of power signal, which is sampled at 12.8 kHz. This gives a frequency resolution of 12.5 Hz in the Fourier transform domain. We define $F(k)$ to be the Fourier transform coefficient at the frequency component of $12.5k$ Hz, $k = 0,1,2,\ldots N/2$. Using the DFT coefficients of disturbances, we establish the following six rules (one for each type of disturbance) to classify the disturbances of Type 5 to Type 10.

Rule 1: If $\|F(4)\| \leq 0.1$ pu, and zero elsewhere $\Rightarrow$ this is a Type 5 disturbance (voltage interruption).

Rule 2: If $\|F(4)\| \in (0.1$ pu, 0.9 pu), and zero elsewhere $\Rightarrow$ this is a Type 6 disturbance (voltage sag).

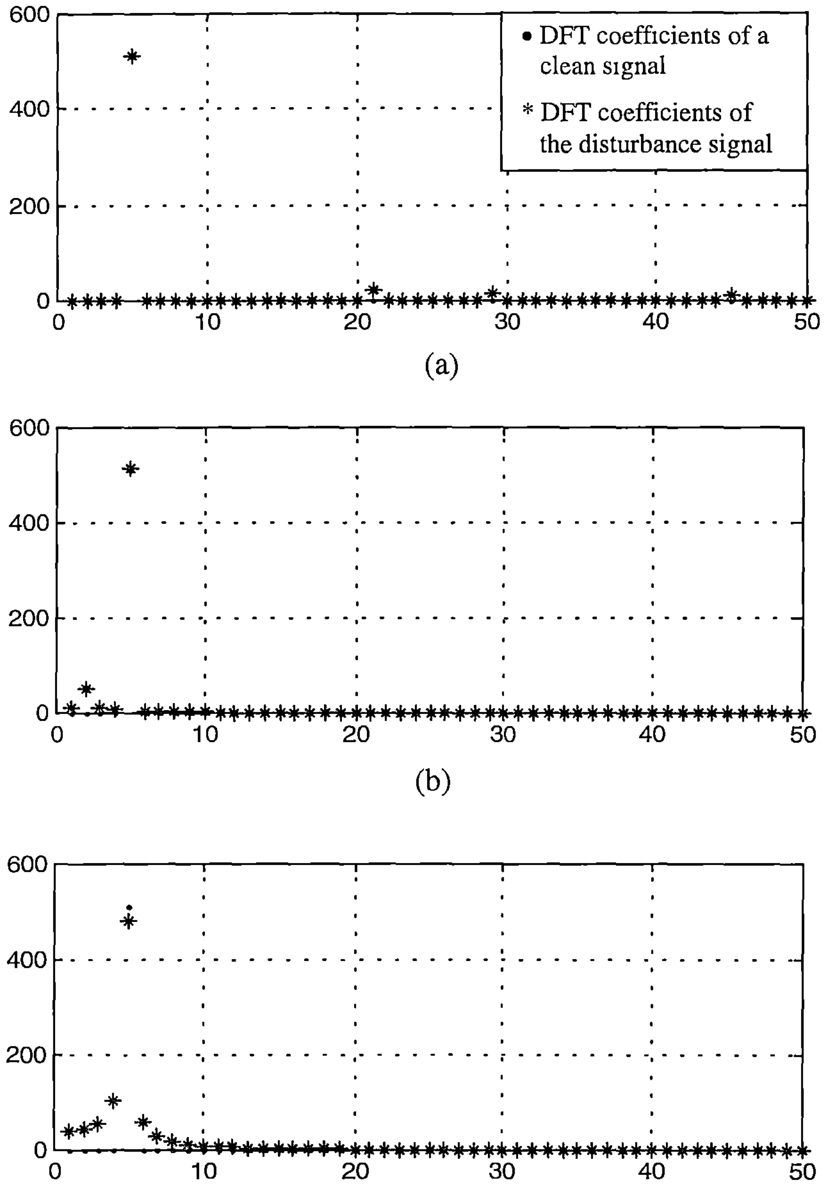
Rule 3: If $\|F(4)\| > 1.1$ pu, and zero elsewhere $\Rightarrow$ this is a Type 7 disturbance (voltage swell).

Rule 4: If $\|F(4)\| \in [0.9$ pu, 1.1 pu], and there is $F(k) >$ a threshold (e.g. 0.05 pu) for $k = 4q$ ($q = 2, 3, 4, \ldots$) and zero elsewhere $\Rightarrow$ this is a Type 8 disturbance (harmonics) and is shown in Figure 8.1 (a).

Rule 5: If $\|F(4)\| \in [0.9$ pu, 1.1 pu], and there is $F(k) >$ a threshold for $k < 4$ and zero elsewhere $\Rightarrow$ this is a Type 9 disturbance (voltage fluctuation) and is shown in Figure 8.1 (b).

Rule 6: If $\|F(3)\|$ and $\|F(5)\| >$ a threshold $\Rightarrow$ this is a Type 10 disturbance (power frequency variation) and is shown in Figure 8.1 (c).

If a signal component has its frequency equalled to a multiple of the frequency resolution (12.5 Hz), then there is a direct relationship between its amplitude and its corresponding DFT coefficient and it does not produce any other DFT coefficient at the other frequency. This is the case for PQ disturbance belonging to any of the four types - voltage sag, voltage swell, interruption or harmonics. Hence, the interpretation of the first four rules, ***Rule 1*** to ***Rule 4*** are straightforward, in which the signal amplitude has a linear relationship with the magnitude of its corresponding DFT coefficient. When the signal component has a frequency different from a multiple of the frequency resolution, it produces a number of DFT coefficients in a neighbour of its frequency. This is the case in voltage fluctuation and power frequency variation disturbances. The voltage fluctuation disturbance, whose frequency is lower than 25 Hz, is normally producing DFT coefficients at frequencies corresponding to $k = 1, 2$ and 3 (i.e. ***Rule 5***), while the power frequency variation disturbance, whose frequency deviates from the 50 Hz, produces DFT coefficients in a neighbourhood of the 50 Hz, (i.e. at least for the two nearest neighbours $F(3)$ and $F(5)$).



(a)

(b)

(c)

Figure 8.1: (a) DFT coefficients of a harmonic disturbance;
(b) DFT coefficients of a voltage fluctuation disturbance;
(c) DFT coefficients of a voltage frequency variation disturbance.

If the disturbance does not belong to these types (Type 5 to Type 10), in the FT domain, the 50 Hz fundamental power component can be removed completely to leave the disturbance for further processing. Thus removes the effects of this large fundamental component on the extraction of features from the non or small power PQ disturbances of Type 1 to Type 4.

8.3 FEATURES EXTRACTION USING WTMM

After removing the 50 Hz fundamental power component, the voltage magnitude and the duration of the disturbance signal *d*(*n*) can be measured correctly in the time domain. A WTMM analysis is carried out for the transient disturbance signal, through which we aim to measure the frequency spectrum and any particular signal structure that may contain in each disturbance type. In the time-scale domain, chains or contours of WTMM are traced from fine to coarse scales for the detection of singularities and irregular structures in the disturbance signal. By following each chain the GMM for each inflection point in the disturbance signal *d*(*n*) is detected. Since most short chains and chains that have weak GMMs are due to noise, we discard them and extract the following two parameters from each remaining chain: the scale $s_{\max}$ in (4.33), at which the GMM occurs, and the Lipschitz exponent α of the disturbance signal at the originating location of the chain. The latter can be approximated from the slope of $\log_{s_0}(|Wd|_{\max})$ versus j ($s_j = s_0^j$) at the low-scales end of the chain, i.e.

$$\log_{s_0}(|Wd(s,t)|_{\max}) \leq \log_{s_0}(A) + (\alpha + 1/2)j \tag{8.5}$$

We found that the scale step s_0 of 1.25 is sufficiently fine to measure closely the feature of the PQ disturbance of Type 1 to Type 4. The WTMM analysis is then carried out for 13 scales $j=0$ to 12, that at the sampling frequency of 12.8 kHz provide a measurement to a frequency range of 440 Hz to 6.4 kHz. Finally, for each disturbance we calculate its four characteristic features from its WTMM:

The average of the scales at which the GMM occur, $\{s_{\max}(p)\}_{p=1 \cdot P}$, is

$$\bar{s} = \frac{1}{B}\sum_{p=1}^{P} \left|Wd_{\max}(p)\right|^2 s_{\max}(p) \tag{8.6}$$

where $Wd_{\max}(p)$ is the WT coefficient of the p^{th} GMM, and $B = \sum_{p=1}^{P} \left|Wd_{\max}(p)\right|^2$ is the normalised factor. This gives a measurement on the disturbance average frequency. The variance of the GMM scales can then be given by

$$\sigma_s{}^2 = \frac{1}{P}\sum_{p=1}^{P} \left(s_{\max}(p) - \bar{s}\right)^2 \tag{8.7}$$

The average value of the Lipschitz exponents, $\{\alpha(p)\}$,

$$\bar{\alpha} = \frac{1}{B}\sum_{p=1}^{P} \left|Wd_{\max}(p)\right|^2 \alpha(p) \tag{8.8}$$

And the variance of the Lipschitz exponents

$$\sigma_\alpha{}^2 = \frac{1}{P}\sum_{p=1}^{P} \left(\alpha(p) - \bar{\alpha}\right)^2 \tag{8.9}$$

By adding the disturbance duration L, the complete feature vector characterising each disturbance is defined by [92]

$$x = (\bar{s}, \sigma_s, \bar{\alpha}, \sigma_\alpha, L) \tag{8.10}$$

In Chapter 7, Figure 7.2 to Figure 7.6 show the distributions of the five WTMM feature components $(\bar{s}, \sigma_s, \bar{\alpha}, \sigma_\alpha, L)$ for the four transient disturbance types (IT, HF, LF and NT). These features are extracted from the training set of 134 samples.

Beside the two 'clear' overall distributing features $\bar{s}$ and L, the standard deviation of the scale σ_s gives a higher discriminant level for the two classes IT and NT, and the average Lipschitz exponent $\bar{\alpha}$ gives a clear indication on whether a disturbance belongs to two classes IT and NT or it belongs to the other two class HF and LF. Hence different features provide different discriminating levels to different disturbance types. This agrees with our proposal of adding the input feature weights to a RBF classifier so that the network optimises the contribution of each feature to each individual class in terms of the discriminant function.

8.4 FEATURES EXTRACTION USING MATCHING PURSUIT

Another method that we use in this thesis for processing the four types of transient PQ disturbances is a so-called adaptive signal representation tool called matching pursuit [75]. Given a redundant dictionary of waveforms, we decompose a signal into a linear expansion of these waveforms, which are selected in order to best match the signal structure. In particular, using the dictionary of Gabor functions g_γ, $\gamma = (2^j, p, k)$, for $j \in [0,\log_2 N]$ and $(p,k) \in [0,N\text{-}1]^2$, presented in Section 5.3.3, we decompose any given disturbance into dominant atoms in the frequency-time distribution which are used to characterise the type of power quality disturbance. As shown in Chapter 5, matching pursuit decomposition with the Gabor dictionary can approximate more than 95% of the disturbance energy after only five iterations. We then extract from the five most dominant Gabor atoms (i.e. the first five iterations) of each disturbance to get its four following characteristic features:

The average value of the five window sizes, which provides information on the size of the fluency or the regular behaviour in the signal structure

$$\bar{s} = \frac{1}{B}\sum_{m=0}^{4} a_m^{\ 2}\, 2^{j_m} \tag{8.11}$$

where a_m is the m^{th} matching pursuit coefficient, and B is also the normalised factor, and is given by $B = \sum_{m=0}^{4} a_m^{\ 2}$. The variance of the five window sizes,

$$\sigma_s^{\ 2} = \frac{1}{5}\sum_{m=0}^{4}\left(2^{j_m} - \bar{s}\right)^2 \tag{8.12}$$

The average value of the modulating frequency,

$$\overline{\xi} = \frac{1}{B}\sum_{m=0}^{4} a_m{}^2 \xi_m \tag{8.13}$$

where the m^{th} modulating frequency is given by $\xi_m = \frac{\omega_s k_m}{2N}$, and ω_s = 12.8 kHz. Finally, the variance of the modulating frequency is

$$\sigma_\xi{}^2 = \frac{1}{5}\sum_{m=0}^{4}\left(\xi_m - \overline{\xi}\right)^2 \tag{8.14}$$

For each disturbance, we add the disturbance duration L to its feature vector to make a feature vector of five components as in [83]

$$\boldsymbol{x} = (\overline{s}, \sigma_s, \overline{\xi}, \sigma_\xi, L) \tag{8.15}$$

The following figures, Figure 8.2 to Figure 8.6 show the distribution of each corresponding feature value for the four transient disturbance types in the same training set of 134 samples.

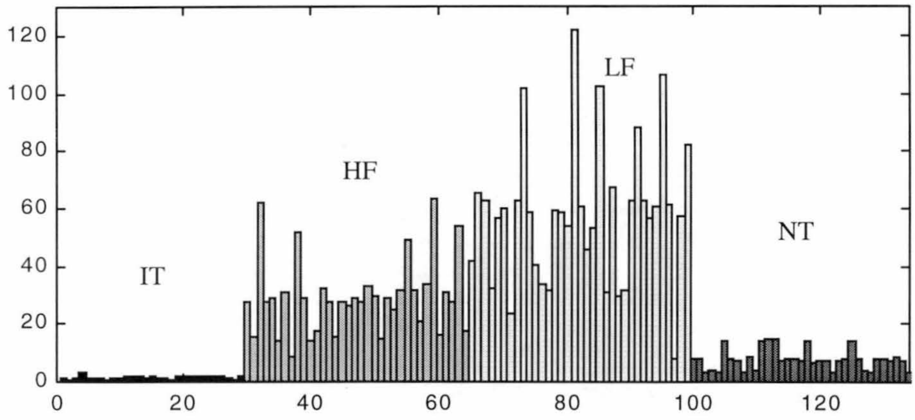


Figure 8.2: Distribution of the average window size $\overline{s}$ in the training set

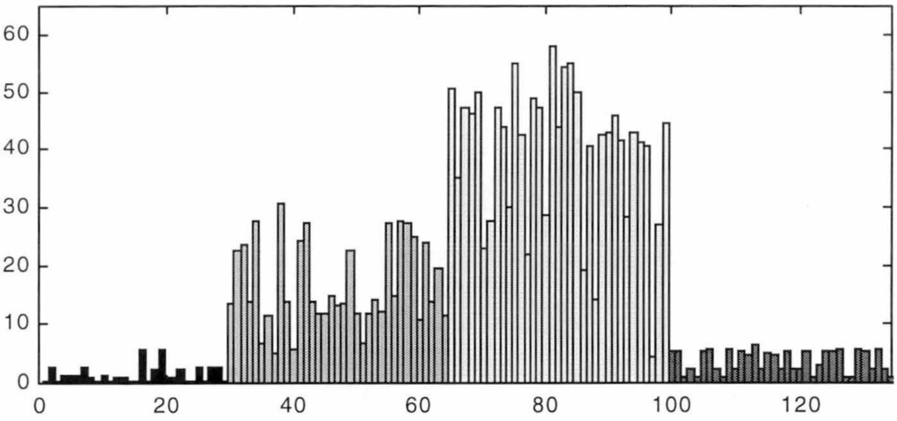


Figure 8.3: Distribution of the standard deviation of the window size σ_s in the training set

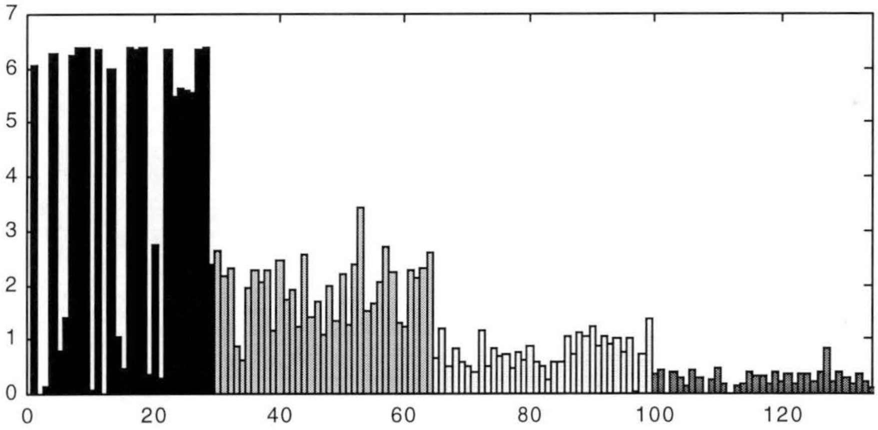


Figure 8.4: Distribution of the average modulating frequency $\bar{\xi}$ in the training set

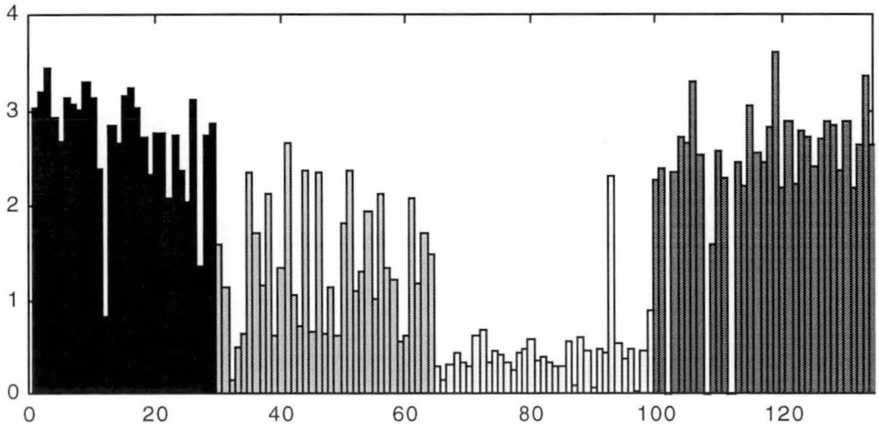


Figure 8.5: Distribution of the standard deviation of the modulating frequency σ_ξ in the training set

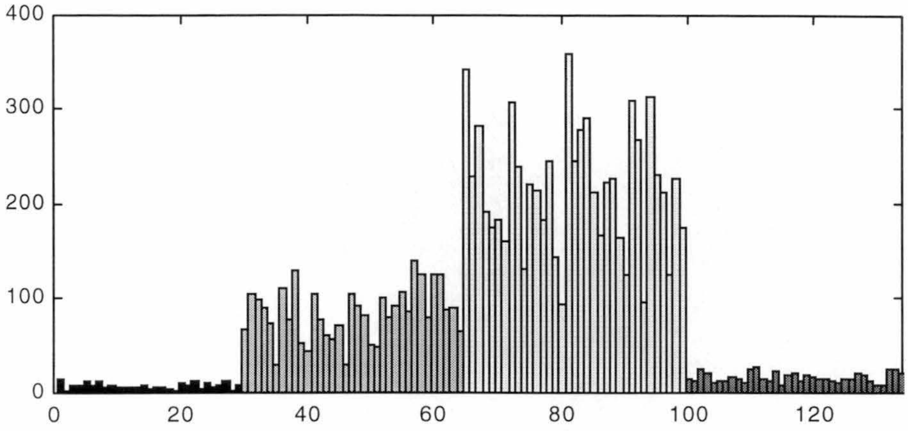


Figure 8.6: Distribution of the disturbance duration L in the training set

We include the modulating frequency in the feature vector since it provides some information on the local disturbance frequencies. However, is does not give a correct measurement on these frequencies, and it can be seen from Figure 8.4 and Figure 8.5 that there are a lot of variances in its value which makes its distribution highly overlapped between different disturbance types. In fact, this is the result of the windowed effect on the modulating frequency, which can approximate well the disturbance frequency with a long window, but with a narrow window, the modulating frequency itself gives little indication on the true disturbance frequency. This is regarded to the time-frequency resolution in a Heisenberg box.

In order to improve our previous work in [83], here we replace the features on the modulating frequency ($\overline{\xi}$ and σ_ξ) by other appropriate features that give a better estimation of the frequencies contained in the disturbance. These new features take both the effects of the modulating frequency ξ_m and the window size s_m to form a measurement of the average value and the variance on the frequency energy distribution ($\overline{E}$ and $\sigma_E{}^2$), which can be obtained from the time-frequency energy in (5.33) as in the following

$$\overline{E} = \frac{1}{B}\sum_{m=0}^{4} a_m{}^2 \overline{\omega}_m \tag{8.16}$$

where the average frequency value $\overline{\omega}_m$ of the m^{th} atom is obtained by discretising the frequency ω in Ω steps such that

$$\overline{\omega} = \frac{1}{C}\sum_{\omega=0}^{\Omega}\omega \exp\left(-\frac{2s_m^{\ 2}}{\omega_s}\left(\frac{\omega_s\omega}{2\Omega} - \xi_m\right)^2 \right) \tag{8.17}$$

in which C is the normalise factor and is given by

$$C = \sum_{\omega=0}^{\Omega}\exp\left(-\frac{2s_m^{\ 2}}{\omega_s}\left(\frac{\omega_s\omega}{2\Omega} - \xi_m\right)^2 \right) \tag{8.18}$$

Finally, the variance of the frequency energy distribution is

$$\sigma_E^{\ 2} = \frac{1}{5}\sum_{m=0}^{4}\left(\overline{\omega}_m - \overline{E}\right)^2 \tag{8.19}$$

The distributions of the two new features $\overline{E}$ and σ_E in the same training set are shown respectively in Figure 8.7 and Figure 8.8.

As we see from Figure 8.4 and Figure 8.7, the average frequency energy distribution $\overline{E}$ is more appropriate then the modulation frequency alone in measuring the disturbance frequency, and it provides a clearer distribution and has higher discriminant between different types of disturbances.

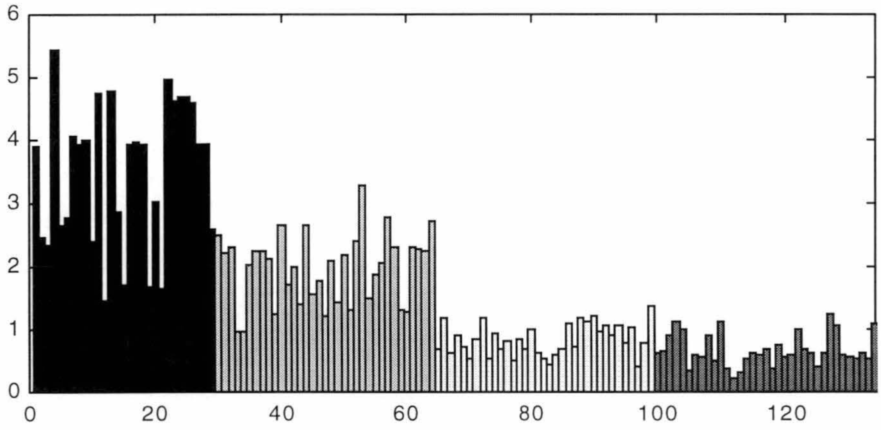


Figure 8.7: Distribution of the average frequency energy distribution $\overline{E}$ in the training set

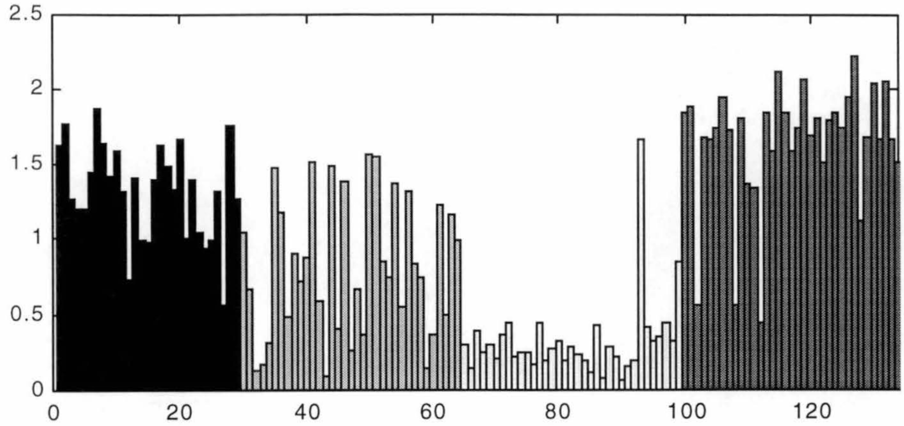


Figure 8.8: Distribution of the standard deviation of the frequency energy distribution σ_E in the training set

8.5 AUTOMATIC CLASSIFICATION OF PQ DISTURBANCES

The block diagram shown in Figure 8.9 explains the proposed process for the automatic recognition of PQ disturbances. We first take the DFT of 1024 samples of a 50Hz power signal that has been sampled at 12.8KHz (i.e. 4 cycles). In the Fourier domain, we use Condition A (explained below) to test for the presence of disturbances of Type 5 to Type 10. If there is a disturbance of these types, the detection process based on the Fourier coefficients in different frequency bands will classify the type of disturbance present in the power signal as described in Section 8.2. If Condition A is not satisfied, that is, there are no disturbances of Types 5 to 10, the 50Hz fundamental power component can be removed completely leaving the disturbance component for subsequent analysis. This is done by first setting the 50Hz Fourier coefficient to zero, and then by taking the IDFT to recover the disturbance component $d(n)$ in the time domain. A small voltage threshold is used to detect the 'clean' power signal situation if $d(n)$ is found insignificant. Otherwise we proceed to extract the features of the disturbance signal $d(n)$ using either the WTMM or the matching pursuit technique as described in Section 8.3 and 8.4. For classifying the four transient disturbance types from the extracted feature vector, we use an RBF classifier with the generalised training for the feature weights (Chapter 7).

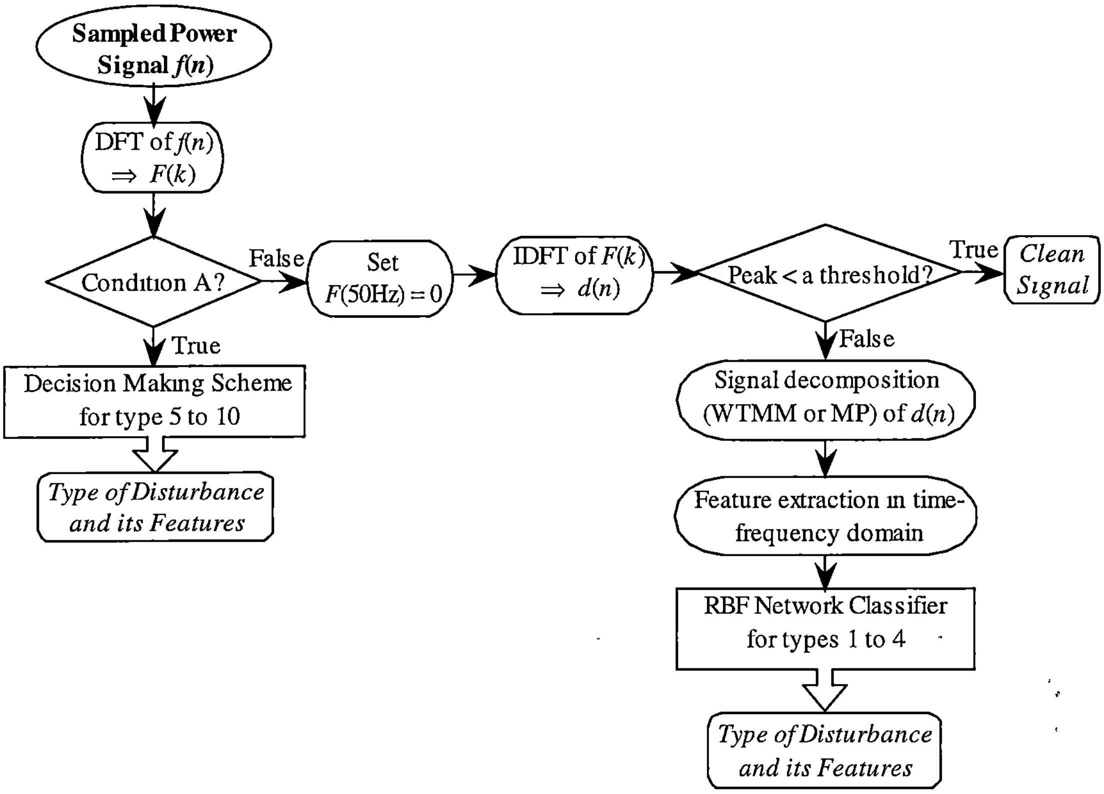

Sampled Power Signal *f*(*n*)
DFT of *f*(*n*) ⇒ *F*(*k*)
Condition A?
False
True
Set *F*(50Hz) = 0
IDFT of *F*(*k*) ⇒ *d*(*n*)
Peak < a threshold?
True
Clean Signal
False
Decision Making Scheme for type 5 to 10
Type of Disturbance and its Features
Signal decomposition (WTMM or MP) of *d*(*n*)
Feature extraction in time-frequency domain
RBF Network Classifier for types 1 to 4
Type of Disturbance and its Features

Figure 8.9: Block diagram of the proposed method for the automatic classification of the 10 types of disturbances under consideration

Condition A: In Fourier transform domain, we use three conditions to test for the presence of disturbance Type 5 to 10. If the disturbance belongs to any of these types then there is at least one of the three following conditions is true. This indicates the satisfaction of Condition A.

Condition 1: $\|F(4)\| \notin$ [0.9 pu, 1.1 pu].

Condition 2: For $k < 4$, there is $\|F(k)\| \geq$ a threshold (e.g. 0.05 pu).

Condition 3: For $k = 4q$, there is $\|F(k)\| \geq$ a threshold and for all $k \geq 8$, $k \neq 4q$ (q = 2, 3, 4,...), $\|F(k)\|$ are insignificant.

In fact, Condition 1 implies a variation of amplitude of the power signal or a frequency drift from its 50 Hz. Condition 2 implies a frequency drift from the 50 Hz power frequency or a voltage fluctuation. And Condition 3 indicates the presence of harmonics in the signal.

8.6 RESULTS, COMPARISON & CONCLUSION

In this section we present the classification results of the four transient disturbance types (IT, HF, LF and NT) using the two feature extraction techniques (WTMM and the matching pursuit) and using different classifiers. We also show that our classification techniques achieve a superior recognition rate over the current automatic disturbance classification techniques. This is because of the two step improvement in our classification method. The first improvement is the extraction of disturbance features by appropriate signal processing tools from which we obtain an efficiency and translation invariant feature vector. The second improvement is the designing of an appropriate classifier which maximises the discriminant function between different disturbance types.

8.6.1 Classification Results

Performing the WTMM decomposition on transient PQ disturbances, we then extract the disturbance feature vector of five components, $\boldsymbol{x} = (\bar{s}, \sigma_s, \bar{\alpha}, \sigma_\alpha, L)$. Using this feature vector, the classification results by different classifiers are shown in Table 7.2 to Table 7.5 in Chapter 7. The technique obtains the smallest classification error rate of 3.1% with the RBF network of generalised training.

For classification results using the matching pursuit technique, we extract the disturbance feature vector, $\boldsymbol{x} = (\bar{s}, \sigma_s, \bar{\xi}, \sigma_\xi, L)$, from its five most dominant matching pursuit atoms [83]. The classification results by different classifiers are shown in the three tables below. Table 8.1 shows the classification results for a conventional RBF network, whose network parameters are initialised by the regression tree. Table 8.2 shows the classification results for a backpropagation network of the same network size (5 input nodes, 4 hidden nodes and 4 output nodes), and Table 8.3 is for the RBF network with the generalised training for the input layer weights $\{w_{hi}\}$. With this classifier, the technique also achieves the smallest classification error rate of 3.1%.

Output Input	*IT*	*HF*	*LF*	*NT*	Total tested samples	Error rate [%]
IT	44	0	0	2	46	4.3
HF	2	45	3	0	50	10.0
LF	0	4	46	0	50	8.0
NT	0	3	0	47	50	6.0

Table 8.1: Classification results with the matching pursuit feature vector $(\bar{s}, \sigma_s, \bar{\xi}, \sigma_\xi, L)$ by a conventional RBF network.

Output Input	*IT*	*HF*	*LF*	*NT*	Total tested samples	Error rate [%]
IT	45	0	0	1	46	97.8
HF	0	50	0	0	50	100
LF	0	6	44	0	50	88
NT	1	1	0	48	50	96

Table 8.2: Classification results with the matching pursuit feature vector $(\bar{s}, \sigma_s, \bar{\xi}, \sigma_\xi, L)$ by the backpropagation network.

Output / Input	*IT*	*HF*	*LF*	*NT*	Total tested samples	Error rate [%]
IT	44	0	0	2	46	4.3
HF	0	50	0	0	50	0.0
LF	0	4	46	0	50	8.0
NT	0	0	0	50	50	0.0

Table 8.3: Classification results with the matching pursuit feature vector $(\bar{s}, \sigma_s, \bar{\xi}, \sigma_\xi, L)$ by the RBF network of generalised training technique

Now we replace the two features $\bar{\xi}$ and σ_ξ in the feature vector by the two features of the frequency energy distribution $\bar{E}$ and σ_E which gives a better estimation of the frequencies contained in the disturbance. The classification results with this new feature vector $x = (\bar{s}, \sigma_s, \bar{E}, \sigma_E, L)$ by the RBF network of generalised training technique are shown in Table 8.4. As expected, this new feature vector does improve the classification rate, as the error rate is down to nearly 2.0% (i.e. 4 out off 196 testing samples).

Output / Input	*IT*	*HF*	*LF*	*NT*	Total tested samples	Error rate [%]
IT	45	1	0	0	46	2.2
HF	0	50	0	0	50	0.0
LF	0	2	48	0	50	4.0
NT	1	0	0	49	50	2.0

Table 8.4: Classification results with the new matching pursuit feature vector $(\bar{s}, \sigma_s, \bar{E}, \sigma_E, L)$ by the RBF network of generalised training technique.

8.6.2 Other Automatic PQ Disturbance Recognition Techniques

Wavelet analysis is recently proposed in the literature as a new tool for monitoring PQ problems. However, much of the work done in the power quality area deal with these problems either from the detection and localisation point of view or from a data compression framework [130,131,132,133], and a limited number of them deal with real classification methodologies that can be used to classify different PQ disturbances as in [134,135,136,137,138,141].

In [137], PQ disturbance signals sampled at 16.4 kHz are decomposed into 12 levels DWT. Then using the total energy in each of the 12 DWT levels, the authors propose to detect and classify different types of disturbance. The method thus only provides a very coarse approximation of the disturbance frequency range, while leaving out many detailed and important characteristics of the disturbance.

Authors of [134,135,136,138] propose almost the same methodology for the classification of PQ disturbances, even though there are some variations in the designing of classifiers and the selecting of disturbance types to be classified. These classification methods also use DWT to pre-process the disturbance signal, and the remaining classification tasks rely on the classification ability of large neural networks, which normally involve a Learning Vector Quantisation (LVQ) network architecture for the calculation of similarity. Since the DWT coefficients, which are very large and lack translation-invariant, are used for the feature vector, the network requires a long training time and cannot achieve a high classification rate unless a large and comprehensive training set is made. This is the reason for the use of large training populations (in the order of hundreds of training samples for each disturbance type involved in the classification) in these methods [134,138].

Using the same training set and the same testing set as in our classification approach, the classification results for the four transient disturbance types by the method used in [135,136] are shown in Table 8.5. Due to a small training set (29 to 35 samples per class) and the feature vector of the DWT being redundant and lacking translation-invariant, the method produces a low classification rate, which has an overall error rate of 11.2% (i.e. 22 out of 196 testing samples).

Output Input	*IT*	*HF*	*LF*	*NT*	Total tested samples	Error rate [%]
IT	45	0	0	1	46	2.2
HF	3	45	2	2	50	14.0
LF	0	7	41	2	50	18.0
NT	5	0	0	47	50	10.0

Table 8.5: Classification results for the method used in [135,136].

There is a significant improvement in our classification technique, in which the overall classification error rate for our method is only 18% (more than 5 times smaller) of that for the method used in [135,136]. We have demonstrated that signal decomposition and features extraction are essential and necessary for studying and classifying different PQ disturbances. Depending on the application and the types of signal characteristics, particular features are extracted so that they can review the discriminant between different types. Poor feature extraction techniques result in large and redundant feature vectors, thus taking a longer time for the system to learn and reduce the classification rate.

Chapter 9

CONCLUSIONS & SUGGESTIONS FOR FURTHER RESEARCH

9.1 CONCLUSIONS

Automatic power quality disturbance classification is discussed in this thesis. There has been an increasing incidence of misadventure on the PQ supplied to the electric utilities and their customers as more and more equipment is used that is sensitive to variations in power supply.

The current practice of recognising and studying the possible cause of the PQ problem is performed manually, which is highly inefficient and costly. This makes imperative the need for automatic disturbance classification methods to replace the current visual inspection.

PQ disturbances can vary in a wide range from sustained long-duration, low and steady frequency to very short-duration, sudden and high frequency, which normally involves different analysing methods for transient disturbances and for steady state disturbances. The analysing of the transient disturbances is often more difficult and more challenging than analysing the steady state disturbances and need to be done in time domain. These transient types can be efficiently analysed by the WT as the WT has the ability to analyse a signal locally in both time and frequency domain. Unfortunately, the conventional orthogonal or biorthogonal WT is not a translation invariant representation, which is its major weakness for pattern recognition applications.

There are several wavelet techniques that provide translation invariant properties in their representations. However, many of these techniques obtain this property by entailing high oversampling rates which make them inefficient. The WTMM technique and the matching pursuit technique presented in Chapter 4 and Chapter 5 respectively are two most suitable analysis techniques for transient disturbances.

Most digital filters are not bandlimited. The analysing of the transient disturbances is then affected by the large fundamental power component even though their frequencies are far apart. In order to have a clear spectrum of disturbance, in our technique the disturbance is isolated first before making any further signal decomposition. This can be done in the Fourier transform domain.

By keeping only the modulus maxima of a continuous WT, the WTMM obtains a multiscale translation invariant representation that is described in Chapter 4. The position of GMMs and the values of the Lipschitz exponent provide an estimation about the local signal frequencies and the local signal characteristics, which can be used as a compact, shift-invariant feature vector for the classification of transient disturbances.

The other efficient decomposition technique presented in Chapter 5 is the matching pursuit, which can closely approximate the disturbance by only the first few iterations. The parameters of the selected atoms (i.e. the window sizes and the modulation frequencies) provide a meaningful description of the local signal structures and provide a measurement of the local signal frequencies. Unfortunately, due to the windowed effect, the modulation frequency itself does not give a correct measurement of the local frequency. A more appropriate feature that includes both the modulation frequency and the window size is found to give a measurement on the frequency energy distribution and is described in Chapter 8.

One the decomposition techniques can represent the disturbance by an efficient and compact feature vector that has a clear distribution between different disturbance types, a statistical pattern recognition approach is more suitable for the classification of these types. We then employ an RBF network for this task. The training of this network is much faster than that of a backpropagation network as the training for the parameters in its two layer are decoupled. However, there is a weakness in a conventional RBF network in that it has only a local learning capability and a limited learning inference from the training data. Therefore, for a problem where the distribution in the feature vector is highly overlapped between different classes, this network produces a lower classification rate than that of backpropagation network of the same network size. By modifying the RBF network structure, in which we add the input layer weights to the network, and propose two new training procedures by either the knowledge base technique or by the generalised training techniques, the classification results by this network are significant improved. The overall classification error rate for this network is less than half of that for the conventional RBF network and is one third smaller than that for the backpropagation network shown in Chapter 7.

Finally, in Chapter 8 we make a comparison between our classification techniques with other current automatic disturbance classification techniques. Our classification techniques achieve a superior recognition rate over the current automatic disturbance classification techniques, in which we achieve the overall classification error rate of more than five times lower than that for the method used in [135,136]. This is because of the two steps improvement in our classification method. The first improvement is the extraction of disturbance features by appropriate signal processing tools from which we obtain an efficiency and translation invariant feature vector. The second improvement is the designing of an appropriate classifier which maximises the discriminant function between different disturbance types.

9.2 SUGGESTIONS FOR FURTHER RESEARCH

Although our classification methodology presented in Chapter 8 significantly improves the classification results of transient PQ disturbances, the implementation of either the WTMM or the matching pursuit is relatively slow. This makes the method unable to monitor and analyse the PQ online. At the current time this is not a problem since most of the classification tasks are done offline, but it will be in the near future when the monitoring task and classifying task are coupled together and require to be done online for a faster procedure. Hence, a faster signal decomposition technique is needed for the classification of PQ disturbances.

For steady state PQ disturbances, in Chapter 8 we present a method to study their steady state characteristics based on their Fourier transform coefficients. However, these disturbances do not always have 'valid' steady state behaviours during the recording data, and this makes the Fourier analysis inaccurate. For example, some steady state disturbances may only exist for as short as half of the power frequency cycle, or some may not start right at the beginning of the recorded data, but somewhere within the recorded data, or even contain transient before the voltage dropping off in a voltage sag. Hence, extra signal processing tasks are required for achieving a more stabile classification of these steady state disturbances. The disturbance interval needs to be determined first before applying the DFT, or some other detection techniques can be used such as peak detection or RMS calculation, in which the latter is often used in power system monitoring [10,11].

Although the main concern of this thesis is the classification of PQ disturbances, in particular the classification of the four transient disturbance types studied in this thesis, the technique can apply further for other applications in other fields as well.

For example, this technique can be used for recognition of human speech, in which formant frequencies of a voiced sound can be either reviewed from the GMM of the WTMM or estimated from the selected atoms of the matching pursuit.

REFERENCES

[1] K.H. Lee and J. M. Schneider, "Rockport Transient Voltage Monitoring System: Analysis and Simulation of Recorded Waveforms", *IEEE Transactions on Power Delivery*, vol. 4, no. 3, pp. 1794-1805, Jul. 1989.

[2] C. J. Kim and B. D. Russell, "Classification of Faults and Switching Events by Inductive Reasoning and Expert System Methodology", *IEEE Transactions on Power Delivery*, vol. 4, no. 3, pp. 1631-1637, Jul. 1989.

[3] M. Y. Chow, et al., "Incipient Fault Detection in DC Machines Using a Neural Network", *Proceedings of the 22nd Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 706-709, 1988.

[4] S. Ebron, et al., "A Neural Network Approach to the Detection of Incipient Faults on Power Distribution Feeders", *IEEE Transactions on Power Delivery*, vol. 5, no. 2, pp 905-911, Apr. 1990.

[5] K. Yabe, "Power Differential Method for Discrimination Between Fault and Magnetizing Inrush Current in Transformer," *IEEE Transactions on Power Delivery*, vol 12, no. 3, Jul. 1997.

[6] A.K. Ghosh, D.L. Lubkeman, "The Classification of Power System Disturbance Waveforms using a Neural Network Approach", *IEEE Transactions on Power Delivery*, vol. 10, no. 1, pp. 109-115, Jan. 1995.

[7] G.T. Heydt and A.W. Galli, "Transient Power Quality Problems Analysed Using Wavelet," *IEEE Transactions on Power Delivery*, vol. 12, no. 2, Apr. 1997.

[8] D.E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA.: Addison-Wesley, 1989.

[9] K.S. Narendra and M.A.L. Thathachar, *Learning Automata - An Introduction*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[10] L. Wessels and E. Barnad, "Avoiding false local maxima by proper initialization of connections," *IEEE Transaction on Neural Networks*, vol. 3, no. 6, pp.899-905, 1992.

[11] M.H.J. Bollen, *Understanding Power Quality Problems: Voltages Sags and Interuptions*. IEEE Inc., New York, 2000.

[12] R.C. Dugan, M.F. McGranaghan and H.W. Beaty, *Electrical Power Systems Quality*. McGraw-Hill, New York, 1996.

[13] D. T. Rizy, "Transient and Harmonic Voltages Associated with Automated Capacitor Switching on Distribution Systems", *IEEE Transactions on Power Systems*, vol. PWRS-2, no. 3, pp. 713-723, Aug. 1987.

[14] J. Arrillaga, S. Chen and N.R. Watson, *Power System Quality Assessment*. J. Wiley & Sons, 2000

[15] IEEE Project 1346 Working Group, "Electric Power System Compatibility with Industrial Process Equipment, part 1: Voltage Sags," *IEEE Industrial and Commercial Power Systems Technical Conference*, Irvine, C.A., USA, pp. 261-266, May 1994.

[16] D. Gabor, "Theory of Communication," *Journal of the IEE*, vol. 93, pp. 429-457, 1946.

[17] P. Franklin, "A Set of Continuous Orthogonal Functions," *Math. Anal.*, vol. 100, pp. 522-529, 1928.

[18] J. Littlewood and R. Paler, "Theorems of Fourier Series and Power Series," *Proceedings of London Math. Soc.*, vol. 42, pp. 52-89, 1937.

[19] I. Daubechies, "Orthonormal Bases of Compactly Supported Wavelets," *Comm. in Pure and Applied Math.*, vol 41, no. 7, pp. 909-996, 1988.

[20] S. Mallat, "A Theory for Multiresolution Signal Decomposition: the Wavelet Representation", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674-693, June 1989.

[21] I. Daubechies, *Ten Lecture on Wavelets.* SIAM Applied Mathematics, 1991.

[22] R.N. Bracewell, *The Fourier Transform and Its Applications.* New York, McGraw-Hill Inc., 1986.

[23] O. Rioul and M. Vertterli, "Wavelet and Signal Processing," *IEEE Signal Processing Magazine*, vol. 8, pp. 14-38, Oct. 1991.

[24] G. Kaiser, *A Friendly Guide to Wavelets.* Boston, Birkhauser, 1994.

[25] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding.* Englewood Cliffs, New Jersey: Prentice Hall, 1995.

[26] C.K. Chui, *An Introduction to Wavelets.* New York, Academic Press, 1992.

[27] S. Mallat, *A Wavelet Tour of Signal Processing.* 2nd Edition, New York, Academic Press, 1998.

[28] R. Duffin and A. Schaeffer, "A Class of Non-harmonic Fourier Series," *Transaction American Math. Soc.*, vol. 24, pp. 263-277, 1952.

[29] A. Grossmann, "Transforms associated to Square Integrable Group Representations," *Journal Math. Phys.*, pp. 2473-2479, 1985.

[30] M. Antonini, M. Barlaud, P. Mathieu and I. Daubechies, "Image Coding using Wavelet Transform," *IEEE Transactions on Image Processing*, vol. 1, pp. 205-220, Apr. 1992.

[31] P.J. Burt and E.H. Adelson, "The Laplacian Pyramid as a Compact Image Code," *IEEE Transactions on Communication*, pp. 532-540, Apr. 1983.

[32] A. Croisier, D. Esteban and C. Galand, "Perfect Channel Splitting by Use of Interpolation, Decimation, Tree Decomposition Techniques," *International Conference on Information Sciences/Systems*, Patras, pp. 443-446, Aug. 1976.

[33] S. Mallat, "Multiresolution Approximation and Wavelet Orthonormal Bases of $\mathbf{L}^2(R)$," *Transactions on America Mathematics Society*, vol. 315, pp. 69-87, Sep. 1989.

[34] Y. Mayer, *Wavelets and Operators.* Advanced mathematics, Cambridge University Press, 1992.

[35] S. Mallat, "An Efficient Image Representation for Multiscale Analysis," *Proceedings of Machine Vision Conference*, Lake Taho, Feb. 1987.

[36] R.E. Crochiere and L.R. Rabiner, "Optimum FIR Digital Filter Implementations for Decimation, Interpolation and Narrowband Filtering," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, pp. 444-456, Oct. 1975.

[37] R.E. Crochiere, "Subband Coding", *Bell System Technical Journal*, vol. 60, pp. 1633-1654, Sep. 1981.

[38] R.E. Crochiere and L..R. Rabiner, "Multirate Processing of Digital Signal," *Advanced Topics in Signal Processing*, Prentice Hall Signal Processing Series, Englewood Cliffs, New Jersey: Prentice Hall, 1988.

[39] R.R. Coifman and Y. Meyer, "The Discrete Wavelet Transform," preprint, Department of Mathematics, Yale University, 1991.

[40] M. Vetterli, "Splitting a signal into subsampled channels allowing perfect reconstruction," *Proceedings IASTED Conference on Applied Signal Processing and Digital Filter*, Paris, Jun. 1995.

[41] M. Vetterly, "Filter banks allowing perfect reconstruction," *Signal Processing*, vol. 10, pp. 219-244, Apr. 1986.

[42] G. Strang and T. Nguyen, *Wavelet and Filter Banks.* Wellesley-Cambridge Press, 1996.

[43] A. Cohen, I. Daubechies and J. Feauveau, "Bi-orthogonal bases of compactly supported wavelets," *Comm. Pure & Appl. Math.*, vol. 45, pp. 485-560, 1992.

[44] W. Sweldens, "The Lifting Scheme: A Construction of Second Generation Wavelets," *SIAM Journal of Math. Anal.*,vol. 29, no. 2, pp.511-546, Mar. 1998.

[45] I. Daubechies and W. Sweldens, "Factoring Wavelet Transforms into Lifting Steps," *The Journal of Fourier Analysis and Applications*, vol. 4, no.3, pp. 247-269, 1998.

[46] N. Saito and G. Beylkin, "Multiresolution Representations Using the Autocorrelation Function of Compactly Supported Wavelets," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, Dec. 1993.

[47] R. Kronland-Martinet, J. Morlet and A. Grossman, "Analysis of Sound Patterns Through Wavelet Transform," *International Journal on Pattern Recognition and Artificial Intelligent*, vol. 1, no. 2, pp. 273-310, 1987.

[48] O. Rioul and P. Duhamel, "Fast Algorithms for Discrete and Continuous Wavelet Transforms," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 569-586, Mar. 1992.

[49] J.C. Pesquet, H. Krim, H. Carfantan and J.G. Proakis, "Estimation of Noisy Signals Using Time-Invariance Wavelet Packets," *Proceedings of Asilomar Conference*, Monterey, CA, USA, pp. 31-34, Nov. 1993.

[50] I. Cohen, S. Raz and D. Malah, "Shift Invariance Wavelet Packet Bases," *Proceedings 20th IEEE International Conference on Acoustics, Speech and Signal Processing*, Detroit, Michigan, pp. 1081-1084, May 1995.

[51] S. Mallat, "Zero-Crossings of a Wavelet Transform," *IEEE Transaction on Information Theory*, vol. 37, no. 4, pp. 1019-1033, Jun. 1991.

[52] R. Hummel and R. Moniot, " Reconstructions from Zero Crossing in Scale Space," *IEEE Transactions on AS Signal Processing*, vol. 37, no. 12, pp. 2111-2130, Dec. 1989.

[53] Z. Berman and J.S. Baras, "Properties of the Multiscale Maxima and Zero-Crossings Representations," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, Dec. 1993.

[54] S. Mallat and W.L. Hwang, "Singularity Detection and Processing with Wavelets," *IEEE Transaction on Information Theory*, vol. 38, no. 2, pp. 617-643, Mar. 1992.

[55] S. Mallat and S. Zhong, "Characterisation of Signals from Multiscale Edges," *IEEE Transactions PAMI*, vol. 14, no. 7, pp. 710-732, Jul. 1992.

[56] S. Chen and D. Donoho, "Atomic Decomposition by Basis Pursuit," *SPIE International Conference on Wavelets*, San Diego, Jun, 1995.

[57] G. Davis, S. Mallat and Z. Zhang, "Adaptive Time-Frequency Decompositions," *Optical Engineering*, vol. 33, no. 7, pp. 2183-2191, Jul. 1994.

[58] S.G. Mallat, "A Theory of Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.11, pp. 674-693, Jul. 1989.

[59] G. Strang, "Wavelets and Dilation Equations: A Brief Introduction," *SIAM Review*, vol. 31, pp. 614-627, Dec. 1989.

[60] E.P. Simoncelli, W.T. Freeman, E.H. Adelson and D.J. Heeger, "Shiftable Multiscale Transform," *IEEE Transactions on Information Theory*, vol. 38, pp. 587-607, Mar. 1992.

[61] P.P. Vaidyanathan, "Quadrature Mirror Filter Banks, M-Band Extension and Perfect-Reconstruction Techniques," *IEEE ASSP Magazine*, pp. 4-19, Jul. 1987.

[62] M. Holschneider, R. Kronland-Martinet, J. Morlet and P. Tchamitchian, "A Real-Time Algorithm for Signal Analysis with the Help of Wavelet Transform." preprint for CPT, CNRS LUMINY Marseilles, 1988.

[63] J. Canny, "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligent*, vol. 8, pp. 679-698, 1986.

[64] D. Marr, *In Vision*, New York, W.H. Freeman, 1982.

[65] D.T. Nguyen and T.A. Hoang, "Detection of Disturbances on Electricity Supply Using Wavelets," *Proceedings of Australian Universities Power Engineering Conference and IEAust Electric Energy Conference AUPEC/EECON'99*, Darwin, Australia, pp. 231-235, 26-29 Sep. 1999.

[66] T.A. Hoang and D.T. Nguyen, "Appropriate Processing for the Classification of Power Quality Disturbances," *Proceedings of Australian Universities Power Engineering Conference AUPEC 2000*, Brisbane, Australia, pp. 196-201, 24-27 Sep. 2000.

[67] A. Liew and D.T. Nguyen, "Uniqueness Issue of the Wavelet Transform Modulus Maxima Representation and a Least Square Reconstruction Algorithm," *Electronics Letters*, vol. 31, no. 20, pp. 1735-1736, Sep. 1995.

[68] A.W.C. Liew, N.F. Law and D.T. Nguyen, "A Direct Recontruction Method for the Wavelet Transform Extrema Representation," *IEE Proceedings – Vision, Image and Signal Processing*, 1996.

[69] S. Zhong, *Edges Representation from Wavelet Transform Maxima*, Ph.D. Thesis, New York University, Sep. 1990.

[70] A.W.C. Liew, *Multiscale Wavelet Analysis of Edges*, Ph.D. Thesis, University of Tasmania, Nov. 1996.

[71] T.A. Hoang and D.T. Nguyen, "Classification of Power Quality Disturbances using Wavelets", under review by *IEEE Transactions on Power Delivery*, Oct. 2001.

[72] N. Delprat, B. Escudie, P. Guillemain, R. Kronland-Martinet, P. Tchamichian and B. Torresani, "Asymptotic Wavelet and Gabor Analysis, Extraction of Instantaneous Frequencies," *IEEE Transactions in Information and Theory*, vol. 36, no. 2, pp. 644-664, Mar. 1992.

[73] R.A. Carmona, W.L. Hwang and T. Torresani, "Characterization of Signals by the Ridges of their Wavelet Transform," *Technical Report*, Department of Maths, University of California at Irvine, CA92717, Mar. 1995.

[74] J. H. Friedman and W. Stuetzle, "Projection pursuit regression", *Journal of the American statistical Association*, vol. 76, pp. 817-823, 1981.

[75] S. Mallat and Z. Zhang, "Matching Pursuit with Time-Frequency Dictionaries", *IEEE Transactions in Signal Processing*, vol. 41, no. 12, pp. 3397-3415, Dec. 1993.

[76] G. Davis, S. Mallat and Z. Zhang, "Adaptive Time-Frequency Approximations with Matching Pursuits", In *Wavelet: Theory, Algorithms and Application*, Eds. C.K. Chui, L. Montefusco, L. Puccio, pp. 271-293, San Diego: Academic Press, 1994.

[77] G.M. Davis, S. Mallat and Z. Zhang, "Adaptive Time-Frequency Decompositions," *SPIE Journal of Opt. Engin.*, vol. 33, no. 7, pp. 2183-2191, Jul. 1994.

[78] M.R. McClure and L. Carin, "Matching Pursuits with a Wave-Based Dictionary," *IEEE Transactions on Signal Processing*, vol. 45, no. 12, pp. 2912-2927, Dec. 1997.

[79] P.J. Phillips, "Matching Pursuit Filters Applied to Face Identification," *IEEE Transactions on Image Processing*, vol. 7, no. 8, pp. 1150-1164, Aug. 1998.

[80] N. Neff and A. Zakhor, "Very Low Bite-rate Video Coding Based on Matching Pursuit," *IEEE Transactions on Circuit Systems for Video Tech.*, vol. 7, no. 1, pp.158-171, Feb. 1997.

[81] P.J. Durka, D. Ircha and K.J. Blinowska, "Stochastic Time-Frequency Dictionaries for Matching Pursuit," *IEEE Transactions on Signal Processing*, vol. 49, no. 3, pp. 507-510, Mar. 2001.

[82] P.J. Durka and K.J. Blinowska, "In Pursuit of Time-Frequency Representation of Brain Signals," *Time Frequency and Wavelets in Biomedical Signal Processing*, Metin Akay, Ed., IEEE Press Series in Biomedical Engineering, pp. 389-406. IEEE press, New Jersey, 1997.

[83] T.A. Hoang and D.T. Nguyen, "Matching Pursuit for the Recognition of Power Quality Disturbances," *Proceedings 33rd Annual IEEE Power Electronic Specialists Conference PESC'02*, Cairns, Australia, pp. 1791-1796, 23-27 Jun. 2002.

[84] Y.C. Pati, R. Rezaifar and P.A. Krishnapasad, "Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition," in *27th Asilomar Conference on Signal, Systems and Computers*, Nov. 1993.

[85] G.M. Davis, S. Mallat and M. Avelanedo, "Greedy Adaptive Approximations," *Journal of Constr. Approximation*, vol. 13, pp. 57-98, 1997.

[86] L. Cohen, "Time-Frequency Distributions: a Review," *Proceedings of the IEEE*, vol. 77, no. 7, pp. 941-979, Jul. 1989.

[87] S. Qian and D. Chen, "Signal Representation via Adaptive Normalized Gaussian Functions", *IEEE Transactions on Signal Processing*, vol. 36, no. 1, Jan. 1994.

[88] A.K. Jain, R.P.W. Duin and J. Mao, "Statistical Pattern Recognition: A Review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no.1, pp. 4 -37, Jan. 2000.

[89] J.A. Anderson, A. Pellionisz and E. Rosenfeld, *Neurocomputing 2: Directions for Research*, MIT Press, Cambridge CA, 1990.

[90] B. Ripley, U. Bornndorff-Nielsen, J. Jensen and W. Kendal, *Networks on Chaos: Statistical and Probabilistic Aspects*, Chapman and Hall, 1993.

[91] T.A. Hoang and D.T. Nguyen, "Wavelet-based Classification of Power Quality Disturbances using Radial Basis Function Networks," *Proceedings 6th IASTED International Multi-Conference on Power and Energy Systems*, Marina Del Rey, USA, pp. 282-287, 13-15 May 2002.

[92] T.A. Hoang and D.T. Nguyen, "Training Radial Basis Function Networks for Wavelet-Based Classification of Power Quality Disturbances," *Proceedings 4th IASTED International Conference on Signal and Image Processing SIP 2002*, Hawaii, USA, Paper # 359-112, 12-14 Aug. 2002.

[93] Y.H. Pao, *Adaptive Pattern Recognition and Neural Network.* Addison-Wesley Publishing Company, Inc., 1989.

[94] R. Schalkoff, Pattern Recognition: Statistical, Structural and Neural Approaches. John Wiley & Sons, Inc., 1992.

[95] C.H. Chen, *Statistical Pattern Recognition.* Hayden, Washington, D.C. 1973.

[96] W.S. McCulloch and W. Pitts, "A Logical Calculus of the Ideas Immanent in Nervous Activity," *Bulletin of Mathematical Biophysics*, vol. 5, pp. 115-131, 1943.

[97] T. Kohonen, *Self-Organising Maps.* Springer Series in Information Sciences, vol. 30, Berlin, 1995.

[98] R.P. Lippman, "An Introduction to Computing with Neural Networks," *IEEE ASSP Magazine*, vol. 4, pp. 4-22, 1987.

[99] D.E. Rumelhart and J.L. McClelland, Parallel Distributed Processing – Exporations in the Microstructure of Cognition, vol. 1: Foundations. MIT Press, Cambridge, Mass., 1986.

[100] F. Palmieri, "Sound Localization with a Neural Network Trained with the Multiple Extended Kalman Algorithm", *International Joint Conference on Neural Networks*, Seattle, Washington, vol. 1, pp. 125-131, 1991.

[101] A. H. Kramer and A. Sangiovanni-Vincentelli, *Efficient Parallel Learning Algorithms for Neural Networks*. Advances in Neural Information Processing Systems 1, Morgan Kaufmann, California, 1989.

[102] R. Battiti, "First and Second-Order Methods for Learning: Between Steepest Descent and Newton's Method", *Neural Computation*, vol. 4, pp. 141-166, 1992.

[103] M.T. Hagan and M.B. Menhaj, "Training Feedforward Networks with the Marquardt Algorithm", *IEEE Transactions on Neural Networks*, vol. 5, no. 6, Nov. 1994.

[104] T. Kohonen, "Self-Organised Formation of Topologically Correct Feature Map," *Biological Cybernetics,* vol. 43, pp.59-69, 1982.

[105] T. Kohonen, *Self-Organization and Associative Memory,* SpringerVerlar, Berlin, 1984.

[106] D.S. Broomhead and D. Lowe, "Multivariate Function Interpolation and Adaptive Network," *Complex Systems*, pp. 321-355, 1988.

[107] J. Moody and C.J. Darken, "Fast Learning in Units of Local-turned Processing Units," *Neural Computation*, vol. 1, no. 2, pp. 281-294, 1989.

[108] S. Chen, B. Mulgrew, and P.M. Grant, "A Clustering Technique for Digital Communications Channel Equalisation using Radial Basis Function Networks," *IEEE Transactions on Neural Networks*, Vol. 4, No. 4, pp.570-579, July 1993.

[109] S. Spengenberg, "The RBF Network Receiver," http://www.ee.ed.ac.uk/~ssp/project/html/, 1995.

[110] S. Chen, E.F.N. Cowan and P.M. Grant, "Orthogonal Least Square Algorithm for Radial Basis Function Networks," *IEEE Transaction on Neural Networks*, vol. 2, no. 2, pp. 302-309, 1991.

[111] M.J.L. Orr, "Regularisation in the Selection of Radial Basis Function Centres," *Neural Computation*, vol. 7, no. 3, pp. 606-623, 1995.

[112] J.E. Moody, S.J. Hanson and R.P. Lippmann, *An Analysis of Generalisation and Regularisation in Nonlinear Learning Systems*. Morgan Kaufmann Pub., San Mateo, California, 1992.

[113] D.J.C. MacKay, "Bayesian Interpolation," *Neural Computation,* vol. 4, no. 3, pp. 415-447, 1992.

[114] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, *Classification and Regression Tree*. Wadsworth, California, 1984.

[115] P. A. Chou, "Optimal Partitioning for Classification and Regression Tree," *IEEE Transaction Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 340-354, 1991.

[116] J.R. Quinlan, *C4.5: Program for Machine Learning*. Morgan Kaufmann Pub., San Mateo, California, 1993.

[117] K.L. Oehler and M.R. Gray, "Combining Image Compression and Classification Using Vector Quantisation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 5, pp. 461-473, 1995.

[118] Q.B. Xie, C.A. Laszlo and R.K. Ward, "Vector Quantisation Technique for Nonparametric Classifier Design," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 15, no. 12, pp. 1326-1330, 1993.

[119] J.A. Leonard and M.A. Kramer, "Radial Basis Function Networks for Classifying Process Faults," *IEEE Control Systems*, pp. 31-38, Apr. 1991.

[120] T.J. Moody and C.J.Darken, "Fast Learning in Networks of Locally Tuned Processing Units," *Neural Computation,* vol.1, pp.151-160, 1989.

[121] M. Kubar, "Decision Trees Can Initialise Radial Basis Function Networks," *IEEE Transactions on Neural Networks*, vol. 9, no. 5, pp.813-821, 1998.

[122] M. Orr, J. Hallam, K. Takezawa, A. Murray, S. Ninomiya, M. Oide and T. Leonard, "Combining Regression Trees and Radial Basis Function Networks," *Int. Journal of Neural System*, vol. 6, no. 6, pp. 453-465, 2000.

[123] T.A. Hoang and D.T. Nguyen, "A Comprehensive Training for Wavelet-based RBF Classifier for Power Quality Disturbances," to appear in *Proceedings of IEEE TENCON'02*, Beijing, China, 28-31 Oct. 2002.

[124] T.A. Hoang and D.T. Nguyen, "Improving the Training of Radial Basis Function Network for Classification of Power Quality Disturbances," *IEE Electronic Letters*, vol. 38, no. 17, pp. 976-977, Aug. 2002.

[125] T.A. Hoang and D.T. Nguyen, "Optimal learning for patterns classification in RBF networks," *IEE Electronic Letters*, vol. 38, no. 20, pp. 1188 –1190, Sep. 2002.

[126] M. Kubar and I. Ivanova, "Initialisation of RBF Networks with Decision Trees," *Processding of the 5th Belgian-Dutch Conference Machine Learning, BENELEARN'95*, pp. 61-70, 1995.

[127] M.M. Richter, *The Knowledge Contained in Similarity Measurements*. Invited Talks on the ICCBR-95, 1995. http://wwwagr.informatik.uni-kl.de/~lsa/CBR/Richericcbr95remarks.html.

[128] M. Lenz, H.D. Burkhard, P. Pirk, E. Auriol and M. Manago, "Case Based Reasoning for Diagnosis and Decision Support," *Artificial Intelligence Communication*, vol. 9, no. 3, pp.138-146, 1996.

[129] R.J. Schalkoff, Pattern Recognition: Statistical, Structural and Neural Approaches, John Wiley & Sons, Inc. 1992.

[130] S. Santoso, E.J. Powers, W.M. Grady and P. Hofmann, "Power Quality Assessment via Wavelet Transform Analysis," *IEEE Transactions on Power Delivery*, vol. 11, no. 2, Apr. 1996.

[131] P. Pillay, A. Bhattacharjee, "Application of Wavelets to Model Short Term Power System Disturbances," *IEEE on Power Systems*, vol. 11, no. 4, Nov. 1996.

[132] L. Angrisani, P. Daponte, A. D'Apuzzo and A. Testa, "A Measurement Method based on the Wavelet Tranform for Power Quality Analysis", *IEEE Transactions on Power Delivery*, vol. 13, no. 4, pp. 990-998, Oct. 1998.

[133] S. Santoso, E.J. Powers and W.M. Grady, "Power Quality Disturbance Data Compression using Wavelet Transform Methods," *IEEE Transactions on Power Delivery*, vol. 12, no. 3, pp. 1250-1256, Jul. 1997.

[134] S. Santoso, E.J. Powers, W.M. Grady and A.C. Parsons, "Power Quality Disturbance Waveform Recognition Using Wavelet-based Neural Classifier, Part 1: Theoretical foundation, Part 2: Application," *IEEE Transactions on Power Delivery*, vol. 15, no. 1, pp. 222-254, Jan. 2000.

[135] J.S. Huang, M. Negnevitsky and D.T. Nguyen, "Wavelet Based FSCL-LVQ Neural Networks for Power Quality Disturbance Classification," *Proceedings of the IASTED International Conference on Power and Energy Systems*, Las Vegas, USA, Nov. 8-10, 1999.

[136] J.S. Huang, M. Negnevitsky and D.T. Nguyen, "A Neural-Fuzzy Classifier for Recognition of Power Quality Disturbance," *IEEE Transaction on Power Delivery*, vol. 17, no. 2, pp. 609 –616, Apr. 2002.

[137] A.M. Gaouda, M.M.A. Salama, M.R. Sultan and A.Y. Chikhani, "Power Quality Detection and Classification Using Wavelet-Multiresolution Signal Decomposition," *IEEE Transactions on Power Delivery*, vol. 14, no. 4, Oct. 1999.

[138] B. Perunicic, M. Mallini, Z. Wang and Y. Liu, "Power Quality Disturbance Detection and Classification Using Wavelets and Artificial Neural Networks," *Proceedings of 8th International Conference On Harmonics and Quality of Power*, vol. 1, pp. 77 –82, 1998.

[139] D.T. Nguyen and T.A. Hoang, "Analysis of Power Transient Disturbances using Wavelet Transform Modulus Maxima Technique," *Proceedings of Australian Universities Power Engineering Conference AUPEC 2000*, Brisbane, Australia, pp. 190-195, 24-27 Sep. 2000.

[140] D.T. Nguyen and T.A. Hoang, "Classification of Power Quality Disturbances Using Radial Basis Function Networks," *Proceedings of International Power Quality Conference IPQC 2002*, Singapore, 21-25 Oct. 2002.

[141] M. Negnevitsky, M.J. Ringrose, J. Huang and T.A. Hoang, "Neuro-Fuzzy Classifier for Recognition of Power Quality Disturbances, Quality and Reliability of Supply", *Proceedings of 6th International Transmission and Distribution Conference and Exhibition*, Brisbane, Australia, 11-14 Nov. 2001.