NOISE REDUCTION AND SOURCE RECOGNITION OF PARTIAL DISCHARGE SIGNALS IN GAS-INSULATED SUBSTATION

JIN JUN

NATIONAL UNIVERSITY OF SINGAPORE

2005

NOISE REDUCTION AND SOURCE RECOGNITION OF PARTIAL DISCHARGE SIGNALS IN GAS-INSULATED SUBSTATION

JIN JUN

(B. ENG)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2005

ACKNOWLEDGEMENT

It is in great appreciation that I would like to thank my supervisor, Associate Professor Chang Che Sau, for his invaluable guidance, encouragement, and advice in every phase of this thesis. It would have been an insurmountable task in completing the work without him.

I would like to extend my appreciation to Dr. Charles Chang， Dr. Toshihiro Hoshino and Dr. Viswanathan Kanakasabai for their valuable advice on this research project.

Acknowledgement is also towards to Toshiba Corporation, Japan for its support on this project.

I would like to thank my wife and my parents for their love, patience, and continuous support along the way.

Thanks are also given to the Power System Laboratory Technician Mr. H. S. Seow, for his help and cooperation throughout this research project.

Last but not least, I would like to thank my friends and all those, who have helped me in one way or another.

PAPERS WRITTEN ARISING FROM WORK IN THIS THESIS

1. C.S. Chang, J. Jin, C. Chang, Toshihiro Hoshino, Masahiro Hanai, Nobumitsu Kobayashi, "Separation of Corona Using Wavelet Packet Transform and Neural Network for Detection of Partial Discharge in Gas-insulated Substations," *IEEE Trans. Power Delivery*, vol. 20, no. 2, pp. 1363 –1369, April 2005..
2. C.S. Chang, J. Jin, S. Kumar, Qi Su, Toshihiro Hoshino, Masahiro Hanai, Nobumitsu Kobayashi, "Denoisng of Partial Discharge Signals in Wavelet Packets Domain," *IEE Proc. Science, Measurement and Technology*, vol. 152, no. 3, pp. 129-140, May 2005.
3. C.S. Chang, J. Jin, C. Chang, "Online Source Recognition of Partial Discharge for Gas Insulated Substations Using Independent Component Analysis," accepted and will appear in *IEEE Transactions on Dielectrics and Electrical Insulation*, Sep. 2005.
4. J. Jin, CS. Chang, C. Chang, T. Hoshino, M. Hanai and N. Kobayashi, "Classification of Partial Discharge for Gas Insulated Substations Using Wavelet Packet Transform and Neural Network," accepted and will appear in *IEE Science Measurement and Technology*, Nov. 2005.
5. C.S. Chang, J. Jin, Toshihiro Hoshino, Masahiro Hanai, Nobumitsu Kobayashi, "De-noising of Partial Discharge Signals for Condition Monitoring of GIS," *Proc. of International Power Quality Conference 2002*, Singapore, vol. 1, pp 170-177.
6. C.S. Chang, J. Jin, C. Chang, Toshihiro Hoshino, Masahiro Hanai, Nobumitsu Kobayashi, "Optimal Selection of Parameters for Wavelet-Packet-Based Denoising of UHF Partial Discharge Signals," *Proc. of Australasian Universities Power Engineering Conference 2004*, paper number 38, Australia.

7. C.S. Chang, R.C. Zhou, J. Jin, “Identification of Partial Discharge Sources in Gas-Insulated Substations,” *Proc. of Australasian Universities Power Engineering Conference 2004*, paper number 50, Australia.

TABLE OF CONTENT

SUMMARY

A PD is a localized electrical discharge that partially bridges the insulation between conductors. It causes progressive deterioration of the insulation and eventually leads to catastrophic failure of the equipment. Measurement and identification of PD signal are thus crucial for the safe operation and condition-based maintenance of Gas-insulated Substations (GIS). However, high-level noises present in the signals limit the accuracy of diagnoses from such measurements. Hence, denoizing of PD signals is usually the first issue to be accomplished during PD analysis and diagnosis.

In the first part of this thesis, a “wavelet-packet” based denoizing method is developed to effectively suppress the white noises. A novel variance-based criterion is employed to select the most significant frequency bands for noise reduction. Parameters associated with the denoizing scheme are optimally selected using genetic algorithm.

Using the proposed method, successful and robust denoizing is achieved for PD signals having various noise levels. Successful restoration of the original waveforms enables the extraction of reliable features for PD identification.

Traditionally, phase-resolved methods are employed for PD source recognition and corona noise discrimination. Although the methods have been extensively applied to diagnose the insulation integrity of high-voltage equipments such as generator, transformer and cable, they have significant limitations when applied to GIS in terms

of speed and accuracy. Therefore, new methods are developed in the second part of this thesis to solve the problems with phase-resolved methods.

To improve the efficiency and accuracy of PD identification, various PD features are extracted from the measured UHF signals. The first category of PD features, namely *ICA_Feature* is extracted using Independent Component Analysis (ICA). The method is seen to reduce the length of the feature vector significantly. Thus improvement on the efficiency of the classification is achieved. Using *ICA_Feature*, successful identification of PD is achieved with limitation of small "between-class" margins due to the time-domain nature of ICA.

Features extracted using wavelet packet transform (*WPT_Feature*) form the second category of PD features. A statistical criterion, known as *J* criterion is employed to ensure that the features with the most discriminative power are selected. Taking advantage of the additional frequency information equipped with wavelet packet transform, *WPT_Feature* exhibits a large margin between feature clusters of different classes, which indicates good classification performance.

Owing to the compactness and high quality of the extracted features, successful and robust PD identification is achieved using a very simple MLP network. Particularly, MLP with WPT-based pre-processing achieves 100% correct classification on test and on data obtained from different PD to sensor distances. This verifies the robustness of the WPT-based feature extraction. Moreover, both the WPT and ICA based PD diagnostic methods are potentially suitable for online applications.

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The background of this research is introduced first. The importance of partial discharge (PD) detection, PD measurement system in gas-insulated-substation (GIS), various noise reduction methods for PD signals and the methods for PD source recognition are reviewed. The objectives, scope and contributions to knowledge of the research are described. Finally, an outline of the thesis is given.

1.1 BACKGROUND OF THE RESEARCH

A significant trend in the development of electrical power equipment over the years has been the increase of equipment operating voltage. This has given rise to the need for more reliable insulation systems and subsequently the need to detect the degradation of such systems through diagnostic measurements. In the past couple of years, increasing attention has been paid to the development of such tools. Among the various diagnostic techniques, partial discharge (PD) measurement is generally considered crucial for condition-based maintenance, as it is nondestructive, non-intrusive and can reflect the overall integrity of the insulation system. Thus, a good understanding of the PD phenomenon is the basis of this diagnostic system.

A PD is a localized electrical discharge that partially bridges the insulation between conductors [1]. PD may happen in a cavity, in a solid insulating material, on a surface or around a sharp edge subjected to a high voltage. An electrical stress that exceeds the local field strength of insulation may cause the formation of PD. Each discharge event damages the insulation material through the impact of high-energy electrons or accelerated ions. This could, with time, lead to the catastrophic failure of the equipment. PD occurring in insulation systems may have different natures depending on the type of defect. Since the degree of harmfulness of PD depends on its nature [2], recognition of the PD source is fundamental in insulation system diagnosis.

1.1.1 Introduction to Gas-insulated Substation

Over the last 30 years, gas-insulated substations (GIS) have been used increasingly in transmission systems due to their many advantages over conventional substations which include space saving and flexible design, less field construction work resulting in shorter installation time, reduced maintenance, higher reliability and safety, and excellent seismic tolerance characteristics. Aesthetics of a GIS are far superior to that of a conventional substation due to its substantially smaller size. Therefore, GIS has become an indispensable part of transmission networks for many years. Fig. 1.1 shows an indoor GIS of 230 kV located at Senoko Road, Singapore.

Fig. 1.1 A 230 kV indoor GIS in Singapore

GIS is a very complicated system that consists of busbars, arresters, circuit breakers, current and potential transformers, and other auxiliary components as illustrated in Fig.

1.2. These components are enclosed in a grounded metal enclosure which is filled with sulfur hexafluoride (SF_6). Epoxy resin spacers are used to hold the conductor in place within the enclosure as shown in Fig. 1.3.

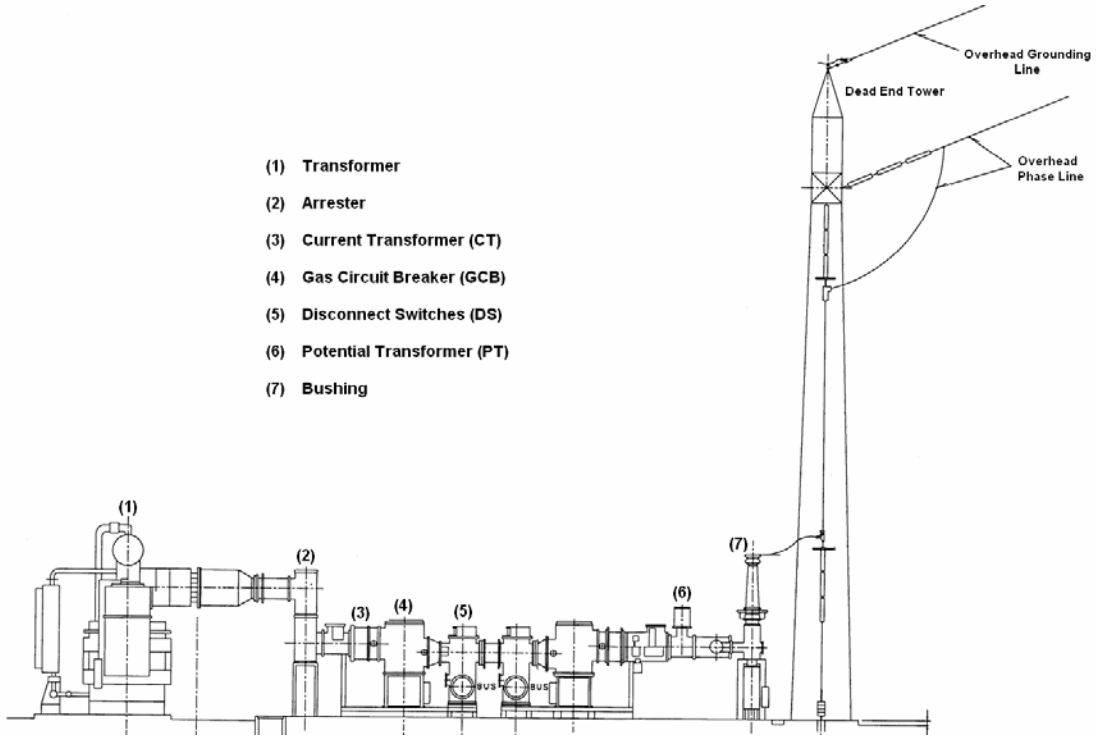


Fig. 1.2 Sectional view of the structure of a 300 kV GIS

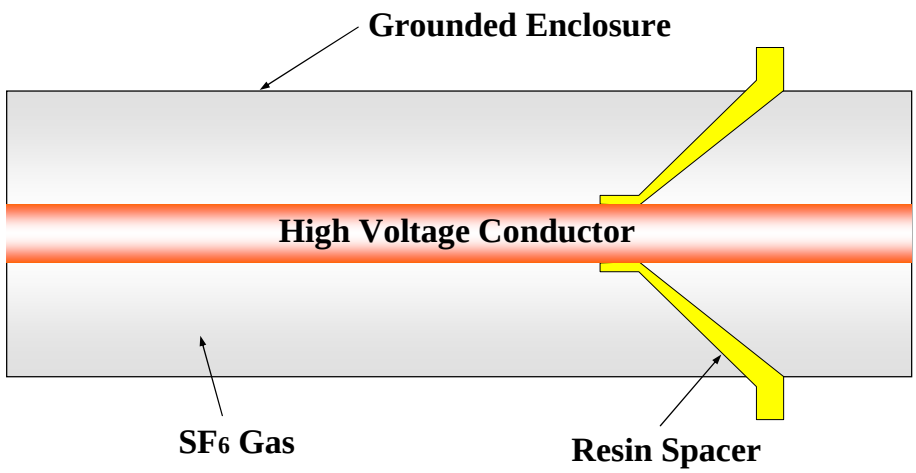


Fig.1.3 GIS test chamber

1.1.2 Condition Monitoring of Gas-insulated Substation

It is crucial to maintain electrical equipment in good operating condition and prevent failures. Traditionally, routine preventive maintenance is performed for such purposes. With the increasing demands on the reliability of power supply, the role of condition monitoring systems become more important, as reliance on preventive maintenance done at a predetermined time or operating interval will be reduced and maintenance is only carried out when the condition of the electrical equipment warrants intervention. This will give the user financial benefits of reduced life cycle costs, improved availability due to fault prevention and the ability to plan for any outages required for maintenance [77].

Traditionally, various methods have been developed for condition monitoring of electrical equipment such as transformer, generator and GIS. Gas-in-oil analysis and on load tap changer monitoring are the key techniques for transformer condition monitoring [78]. The classical monitoring techniques applied in power generators include vibration and air-gap flux monitoring [79]. For GIS, the parameters to be monitored include partial discharge, gas density, gas quality, voltage, current, circuit breaker (CB) position, CB contact erosion, CB spring status and surge arrester leakage current. Among these parameters, CB position and contact erosion have been monitored to prevent failure [80-81].

In recent years, there has been a great deal of new development in GIS monitoring techniques, among which partial discharge detection [3-7] is found to be the most important method as PD is an indicator of all dielectric failures in the initial stages. This thesis focuses on the detection and identification of PD activities in GIS.

1.1.3 PD in SF_6

Sulfur hexafluoride (SF_6) gas has been used as a popular insulation material since its dielectric strength is twice as good as air and it also offers excellent thermal and arc interruption characteristics [28]. However, conducting particles may cause PD in SF_6 and lower the breakdown voltage of a GIS considerably. The likely causes of such contamination are debris left from the manufacturing and assembly process, mechanical abrasion, movement of the central conductor under load cycling and vibration during shipment. Even with a very high level of quality control, it appears that a certain level of particulate contamination is unavoidable. Therefore, investigation of PD activities in SF_6 is imperative for the condition monitoring of GIS.

The common defects in GIS include free conducting particles, surface contamination on insulating spacers and protrusions on conductor [7-10] as illustrated in Fig. 1.4. These defects enhance the local electric field, leading to partial discharge and ultimately a complete breakdown. Corona, which is regarded as an important source of noise is also reviewed in this section.

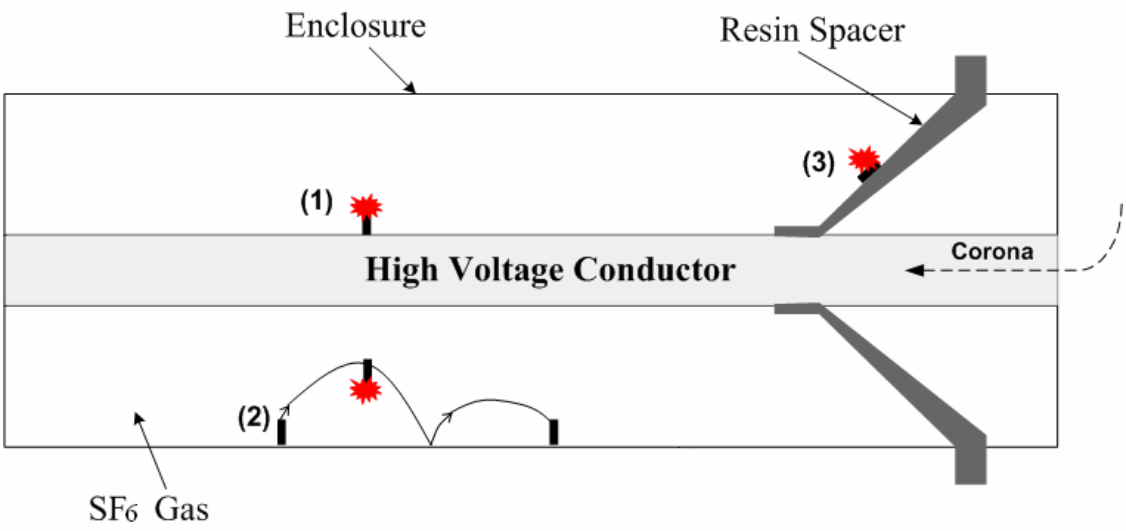


Fig. 1.4 Common defects in GIS. (1) protrusion on conductor, (2) free conducting particle, (3) particle on spacer surface.

Free Conducting Particles

Contamination of GIS with metallic particles occurs either in the field, during operation or during assembly in the plant. The particles can reduce the breakdown voltage significantly due to partial discharge. Therefore, it is of great interest to identify such defects through analysis of PD signals.

When a free conducting particle, such as a piece of swarf, is exposed to the electric field in a GIS, it becomes charged and experiences an electrostatic force. The electrostatic force may be sufficient to overcome the particle's weight, so that the particle moves under the combined influence of the electric field and gravity. The particle may return to the enclosure at any point on the power frequency wave and a "dancing" motion is observed. When the particle moves, it periodically makes contact with the grounded enclosure, and a discharge occurs with every touch. The breakdown

occurs when the particle approaches, but is not in contact with the busbar. There is a critical particle-to-busbar spacing where the system breakdown voltage is a minimum. Apart from the movement of the particle, there are a number of factors that affect the degree of harmfulness of a free particle, such as the shape and size of the particle, applied voltage level, etc. Long, thin and wire-like particles are more likely to trigger breakdown than spherical particles of the same material [8].

As breakdown will only occur when a particle is lifted and approaches the busbar, various techniques have been developed for permanently deactivating or removing particles from the active region during high voltage testing [85, 86]. For instance, an adhesive can be employed at the low field enclosure in conjunction with a low field trap. Other techniques for preventing particle movement include applying insulating coatings on the enclosure, using magnetic fields and coating the particles with a dielectric layer [86]. Although probability of breakdown is reduced due to the above-mentioned measures which decrease the number of free particles in the chamber, particle-initiated breakdown is still unavoidable in GIS due to the particles generated during operation.

Particle on Spacer Surface

A free metallic particle tends to migrate towards a spacer surface under the influence of the applied field [30]. Electrostatic forces or grease on the particle may then attract the particle to the surface, which could lead to a partial discharge. Thus, the gas-insulator interface is often considered as the weak point in a high voltage system [29]. During the design of such a system, the maximum operating voltage is often limited by

the voltage rating of insulating supports rather than the dielectric strength of the SF_6 gas. This voltage rating is highly dependent on surface conditions and the presence of any contamination which may initiate partial discharge. Sources of contamination include fixed metallic particles, grease and trapped charge [10].

A particle on the spacer is in contact with a surface that will store charge near the particle ends. The accumulated charges can then lead to high field concentration on the surface of spacer. Therefore, particles on the spacer can reduce the flashover voltage significantly.

Protrusion on Conductor

A sharp metallic protrusion on a busbar enhances the local electric field. If the local electric field exceeds some critical value, there is a localized breakdown of the SF_6 gas which causes discharges that could lead to complete breakdown. This type of defect is usually considered to be the most critical one that defines the critical PD level [29].

For a protrusion on the busbar, three distinct phases of discharge activities can be identified namely diffuse glow, streamer and leader discharge. However, the glow discharge is not detectable using UHF measurement as the PD current magnitude is small and the frequency components are too low for UHF excitation. On the other hand, leader discharge is only observed at high voltages prior to breakdown. Hence, PD data is measured from streamer phase in this work.

Air Corona

Corona is a discharge phenomenon that is characterized by the complex ionization which occurs in the air surrounding high voltage transmission line conductors outside the GIS at sufficiently high levels of conductor surface electric field. It is usually accompanied by a number of observable effects, such as visible light, audible noise, electric current, energy loss, radio interference, mechanical vibrations, and chemical reactions. Corona signals propagate through the busbar and are detected by the sensors.

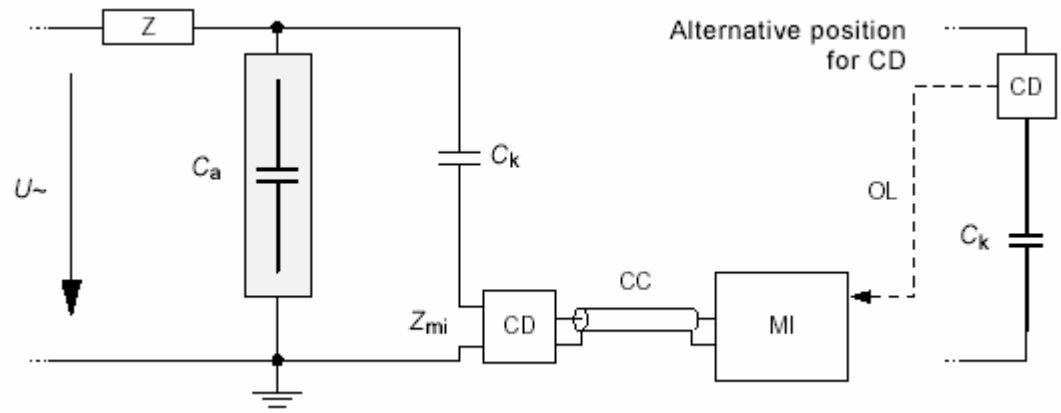
1.1.4 PD Measurement in Gas-insulated Substation

It is well known that GIS breakdown is invariably preceded by PD activities inside the GIS chamber. Therefore, detection and identification of PD activities allow action to be taken at the appropriate time so that potential failure may be prevented. To ensure safety operation, the GIS should be checked for partial discharge during its commissioning tests, and then monitored continuously while in service to reveal any potential fault condition.

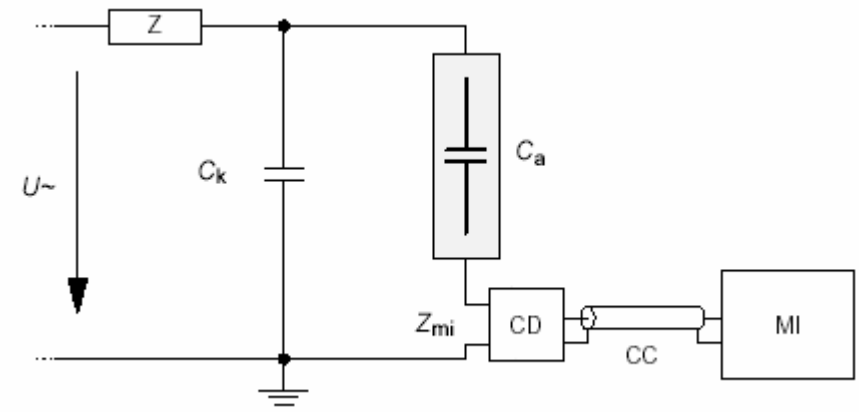
Associated with PD activity in GIS are a number of phenomena which may be monitored. These include light output, chemical by-products, acoustic emission, electrical current and UHF resonance. In the acoustic method, vibration transducers are attached on the outside of the GIS chambers. They are then able to detect the pressure waves caused by PD. However, too many transducers would be needed if a complete GIS is to be monitored in service. Alternatively, optical measurements have the advantage of great sensitivity, but they are unsuited for practical use because of the large number of optical couples needed. Efforts have also been made on detecting

chemical changes in SF_6, but this technique appears to be too insensitive for PD detection in GIS [3].

For many years, the conventional electrical method, IEC 270, has been well developed and widely used in detecting PD activities in cables, transformers, generators, and other equipment. The typical frequency range of this type of measurement is 40 kHz to 1 MHz. Fig. 1.5 shows the typical measurement circuit of the IEC 270 method. A coupling capacitor is placed in parallel with the test object and the discharge signals are measured across the external impedance.



(a)



(b)

Fig. 1.5 PD measurement circuit of IEC 270 method
(a) Coupling device in series with the coupling capacitor; (b) Coupling device in series with the test object

U~: High-voltage supply

Z_{mi}: Input impedance of measuring system

CC: Connecting cable

OL: Optical link

Ca: Test object

Ck: Coupling capacitor

CD: Coupling device

MI: Measuring instrument

Z: filter

One of the main advantages of this method is that a very broad scale of experience has been obtained through years of practical applications. In addition, the measurement can be calibrated to assure that the same result is obtained from two different systems that are used to measure the same sample. However, there are three major drawbacks associated with this method which make it inappropriate to be applied in GIS [3-6]. Firstly, the IEC 270 method needs an external coupling capacitor which is not normally provided in GIS. Hence, the method can not be employed on the GIS in service. Secondly, the sensitivity of the method depends on the ratio of the coupling capacitance to the capacitance of the test object. The total capacitance of a GIS is large. Therefore, the method has insufficient sensitivity for a complete GIS. Thirdly, such a low frequency method is not suitable for field application on GIS as a result of excessive interferences as shown in Fig. 1.6.

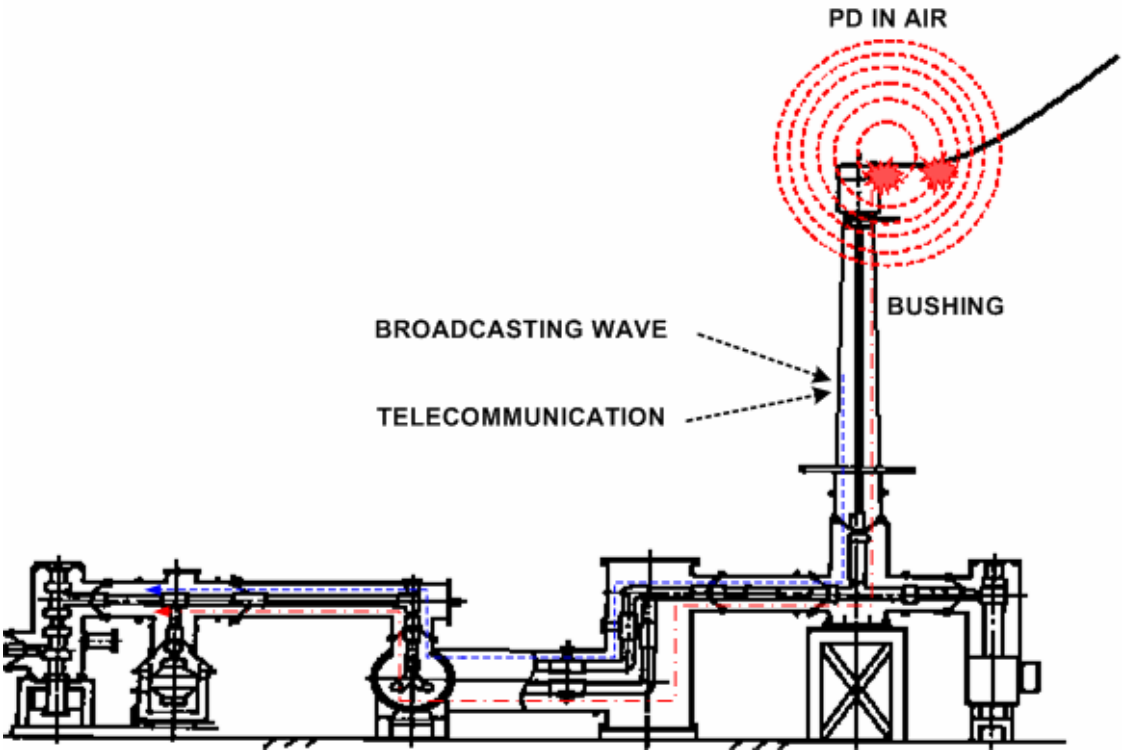


Fig. 1.6 Various noises travel through the GIS conductor via bushing

To address the abovementioned issues, ultra-high-frequency (UHF) method was introduced for PD measurement in GIS [2, 5-6] and is adopted in this study. The UHF ranges from 300 MHz to 1.5 GHz. This technique involves the use of coupling sensors for extracting the UHF resonance signals that are excited by PD current occurring at a defect site within the GIS. Since the UHF signals propagate throughout the GIS with relatively little attenuation, it is sufficient to fit sensors at intervals of about 20 m along the chambers to achieve a sufficiently high sensitivity. In addition, UHF method possesses better noise suppression capability than IEC 270 method due to its high operating frequency. According to the time domain properties, the noises encountered during on-site PD measurement in GIS can be broadly divided into three classes: sinusoidal continuous noise, white noise and stochastic pulse-shaped noise [11-12]. The sinusoidal continuous noises include radio broadcasting, power frequency, harmonic, and so on. These interferences have a frequency range from power

frequency up to VHF ranges (30 MHz to 300 MHz). However, they do not produce electromagnetic waves within UHF ranges (300 MHz to 1.5 GHz). Thus sinusoidal continuous noises can not be detected by the UHF sensor and are not considered in this study. . However, the other two types of noise contain both low frequency and high frequency components. Thus, advanced noise reduction techniques have to be developed for suppressing the residual noises in UHF signals.

1.1.5 Overview of the UHF PD Monitoring System for GIS

Based on UHF PD measurement, a PD monitoring system usually consists of several functional components as shown in Fig. 1.7. The function of each component is briefly described as follows [82]:

1. UHF Measurement.

   Data acquisition is usually performed through internal or external UHF sensors. The recorded data are then transferred and stored on a PC hard drive for further analysis.

2. Noise reduction.

   It is well-known that environmental noises present on the GIS site would cause distortion in the measured signals. Therefore, sufficient noise suppression is a pre-requisite for any on-site PD evaluation and analysis.

3. Partial discharge fingerprints construction.

To achieve effective insulation diagnosis, it is highly desired to extract discriminative features from the original UHF signals. Examples of PD fingerprints include phase-resolved PD patterns and point on wave.

4. Air corona discrimination.

   Air corona is the most important form of interference in the PD monitoring system of GIS. Therefore, discrimination between SF_6 PD and air corona is the basis for PD source recognition and location.

5. PD source recognition.

   The degree of harmfulness is dependent on the type of defect. Thus, identifying the source of SF_6 PD is crucial for risk assessment.

6. PD location.

   Once a critical SF_6 PD is detected, it should be located quickly so that it can be corrected in time.

7. Alarm or message.

   When a harmful PD is detected, it is desired that some form of alarm is triggered, such as sound or light. In the case of recognition of source and location, a message may be displayed, indicating the type of defect or the distance between PD site and the measurement point. Based on the message and the operating conditions, risk assessment can be done by an engineer or an expert system that have the complete knowledge of the GIS.

In many commercial PD monitoring systems for GIS, some of the components, such as PD location are not included. This may be due to the lack of practical methods and the

complicated structures of GIS. In such commercial systems, the UHF signals created by partial discharge are detected by couplers positioned throughout the substation. The signals are then passed via coaxial cables to a local processing unit where they are amplified, filtered and digitized. Subsequently, the processed data is transferred and saved in a central PC, where a PD diagnostic software is usually installed. By running the software, various PD patterns are built for data obtained from each sensor and used by an experienced engineer or artificial intelligence software to assess the risk of defects in GIS.

In this thesis, various components of a PD monitoring system, namely noise reduction, feature extraction, air corona discrimination and source recognition have been featured as illustrated in Fig. 1.7.

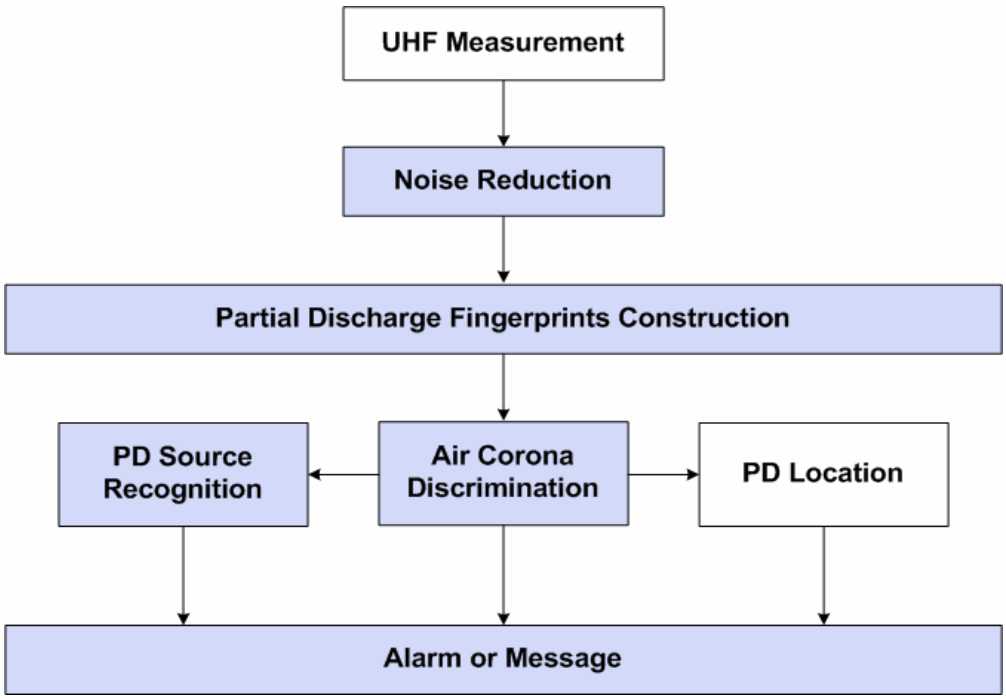


Fig. 1.7 A typical PD monitoring system

1.1.6 The Necessity of Noise Reduction and Discrimination

Although an increase of the signal to noise ratio (SNR) can be achieved to some degree by using UHF measurement as discussed in Section 1.1.4, the noises present in the signals are still too massive to achieve accurate diagnosis from such measurements [23]. This limitation can cause delays in employing appropriate remedial measures, leading to further deterioration of the GIS insulation or a total breakdown.

White noises widely exist in the high voltage laboratory and on site. They are Gaussian distributed in time domain and uniformly distributed in frequency domain. Therefore, it is impossible to effectively eliminate white noise using any time or frequency methods. Fig. 1.8 shows a measured UHF PD signal buried in excessive white noise. It can be seen that the PD signal has been distorted and it is impossible to gauge the condition of the insulation based on such a signal.

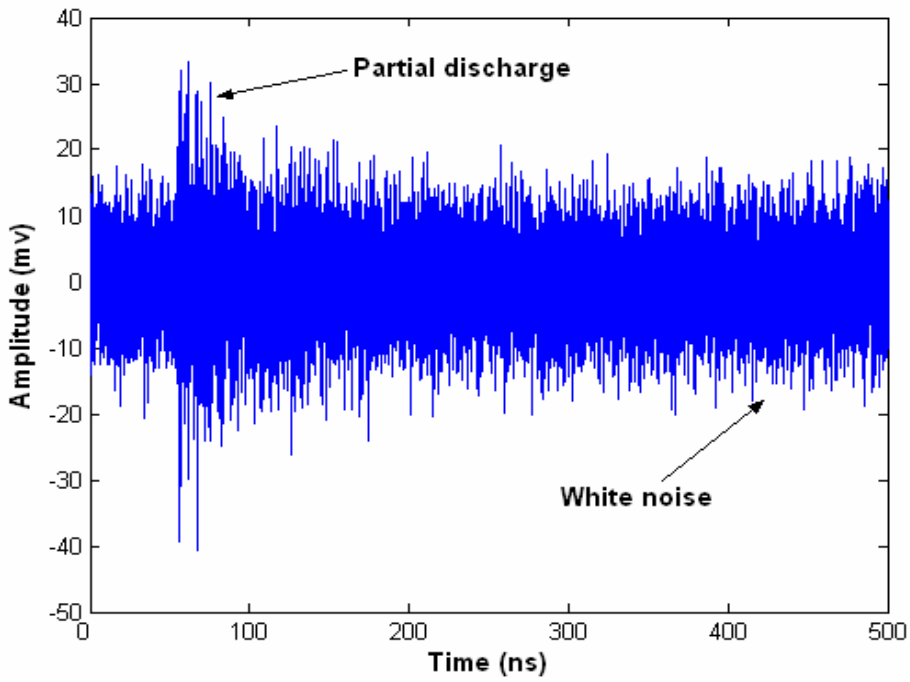


Fig. 1.8 Partial discharge signal buried in white noises

Air corona occurs in the form of stochastic pulse-shaped noise at the bushing of the GIS. It is therefore not so harmful to GIS insulation. However, the signal is usually so intense that enough UHF components are fed into the busbar to give an unacceptably high noise level. It is difficult to distinguish this kind of interference due to the similarities between SF_6 PD and air corona. The amplitudes of corona signals are often comparable to or even bigger than those of PD as illustrated in Fig. 1.9. Therefore, discrimination of air corona is crucial for PD detection and source recognition.

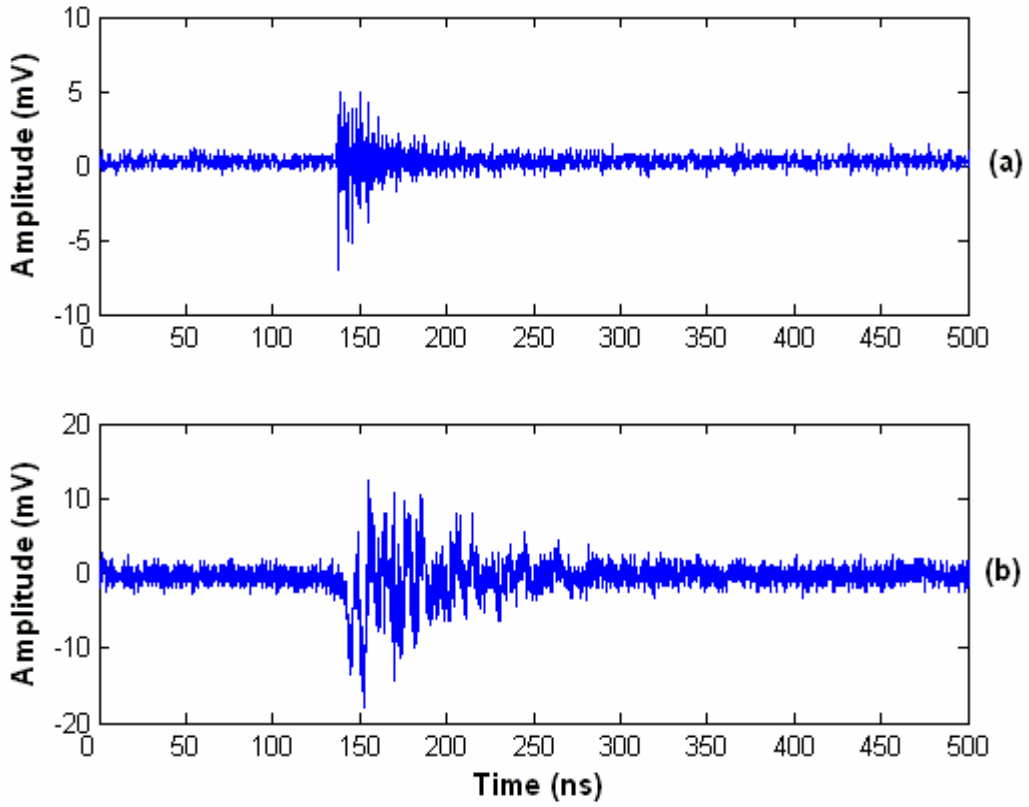


Fig. 1.9 Comparison of SF_6 PD and air corona. (a) SF_6 PD; (b) air corona.

1.1.7 The Necessity of PD Source Recognition

When PD is detected in the insulation system of GIS, it is crucial to identify the type of the defect promptly, as the degree of harmfulness of PD is dependent on its source [87].

As distinct from partial discharge occurring in solid or liquid dielectrics for generators and transformers, PD in SF_6 exhibits unique breakdown characteristics as illustrated in Fig. 1.10. It can be seen that both PD inception and breakdown voltage increase with the gas pressure in region I. In region II, breakdown voltage decreases with increasing pressure, while inception voltage keeps going up. Above a critical pressure P_c, breakdown voltage is seen to coincide with inception voltage, meaning that PD in SF_6 leads to breakdown very fast. This suggests that the PD diagnostic system must be able to detect and identify the PD source in time so that breakdown can be prevented. However, the widely adopted PD diagnosis method, namely phase-resolved PD (PRPD) pattern analysis requires a long time for signal measurement and formation of PRPD patterns. Thus, it may not meet the requirement for GIS application. In addition, this approach can not be applied to DC power transmission system, where phase reference is not available. With the increasing application of DC transmission, PD identification in such systems becomes more and more important. There is therefore an urgent need to develop a new method for fast and reliable classification of SF_6 PD. Detailed review of PRPD pattern analysis and its application is given in Section 1.3.

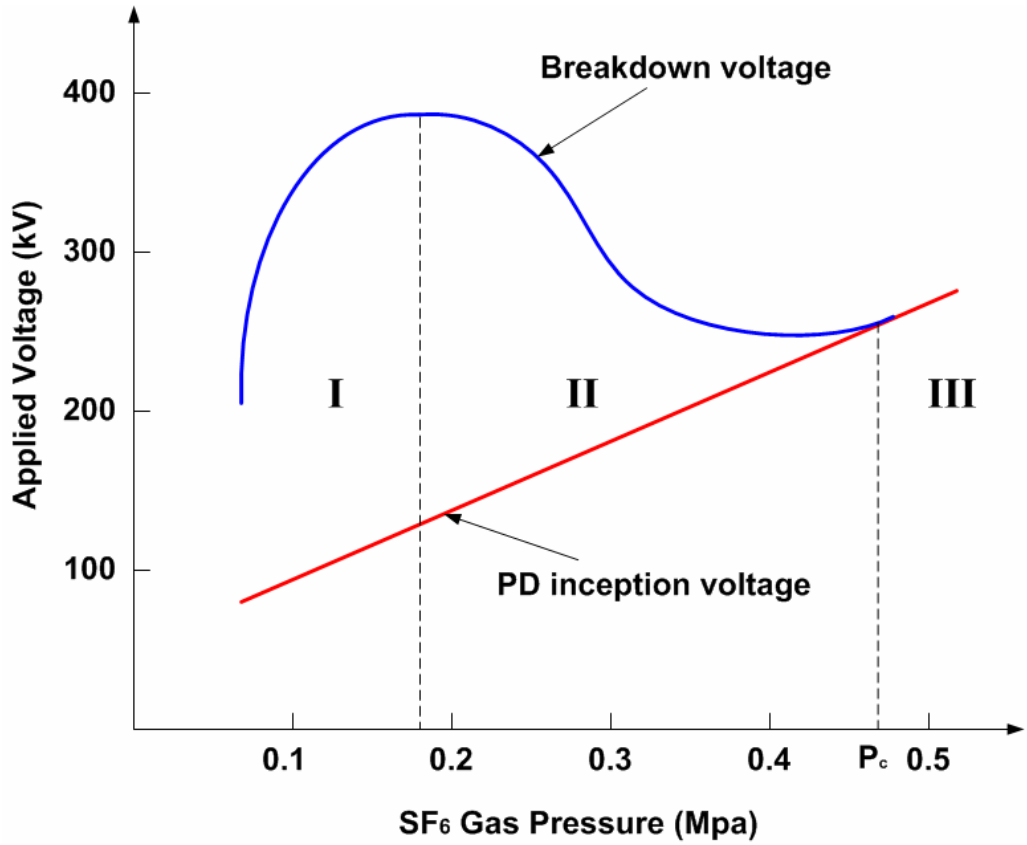


Fig. 1.10 Breakdown characteristics of SF_6

1.2 REVIEW OF NOISE REDUCTION AND DISCRIMINATION

In this section, previous works on reduction of white noise and discrimination of corona are reviewed.

1.2.1 Removal of White Noise

Firstly, methods of eliminating white noises are reviewed. In this thesis, denoizing refers to the process of suppressing white noises.

The various techniques for white noise reduction include filtering, spectral analysis and Wavelet Transform (WT) [13], among which filtering and spectral analysis are

based on Fast Fourier Transform (FFT). Fast Fourier Transform and its inverse give a one-to-one relationship between the time domain and the frequency domain [14]. Although the spectral content of the signal is easily obtained using the FFT, information in time is however lost. Fig. 1.11 shows the FFT of a measured PD signal. As illustrated in Fig. 1.11 (b), FFT only gives the frequency components of the PD signal. Since white noises are uniform distributed in frequency domain, it is impossible to remove white noises using FFT without significant distortion in the original PD signal. Therefore, additional time information is crucial for PD signal denoizing and detection due to its non-periodic and fast transient waveform in time domain.

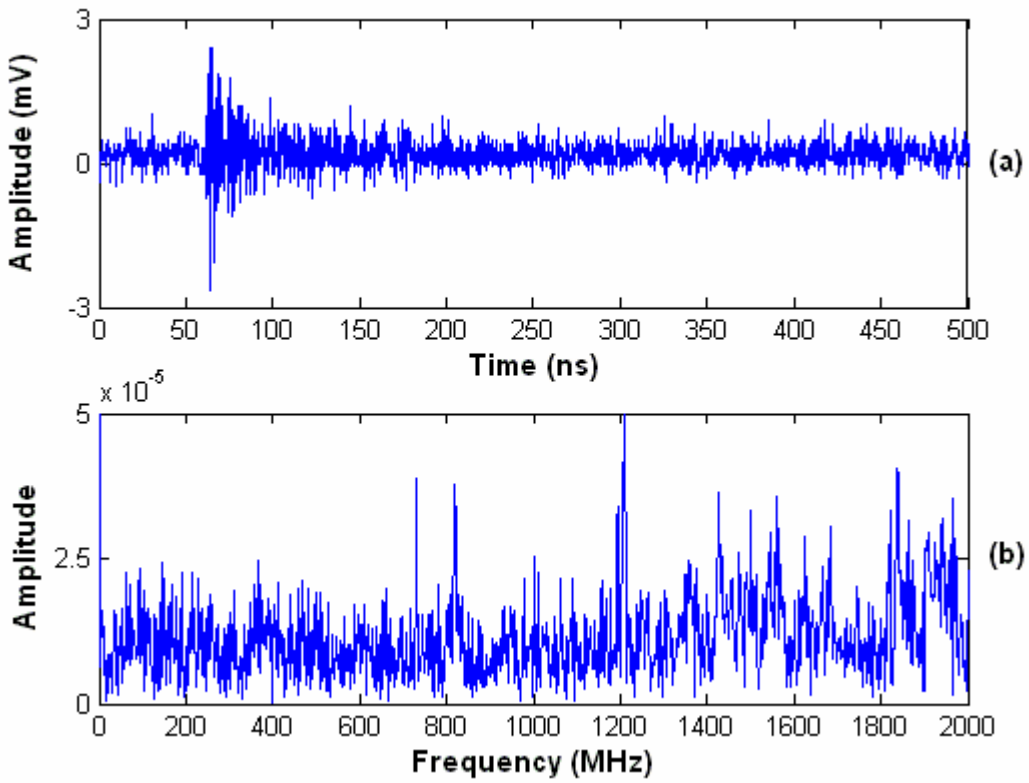


Fig. 1.11 Fast Fourier Transform of UHF PD signal (a) PD signal; (b) FFT of (a).

In recent years, wavelet transform has been proposed as an alternative to Fourier

Transform [13], [15-17] for PD signal denoizing. Wavelets are functions that satisfy certain mathematical requirements and are used in representing data or other functions. Using their practical implementation known as wavelet filter banks, discrete wavelet transform (DWT) maps the data into different frequency components, and then studies each component with a resolution matched to its decomposition level. As illustrated in Fig. 1.12, DWT processes PD signal at different time-frequency resolutions so that both frequency and time characteristics can be studied simultaneously. In addition, the energy of PD signal is concentrated in a few large decomposition coefficients, while the energy of white noise is spread among all coefficients in wavelet domain, resulting in small coefficients [83, 84]. Therefore, it is feasible to remove white noises in wavelet domain with little distortion by employing a thresholding method. DWT thus suppresses white noise within the PD signals more effectively than Fourier based methods.

Although DWT has advantages over traditional Fourier methods in analyzing PD signals, there is still a drawback with DWT, namely the poor frequency resolution at high frequencies as shown in Fig. 1.12. It can be seen that only the low frequency components are decomposed further at each level. The high frequency components, such as “D1”, are however used for denoizing without further decomposition. It has therefore caused difficulties in estimating the noise components at high-frequency subbands due to the low frequency resolution. In particular, when the measured PD signal has a very low signal to noise ratio (SNR), the wavelet transform based methods could have a poor performance. On the other hand, Wavelet Packet Transform (WPT) overcomes the shortcoming with DWT by further splitting the high frequency components as well, which gives much finer resolution in high frequencies.

Therefore, a WPT-based method that automatically determines noise levels in various frequency components is developed in this research project to address the issues with DWT-based methods as reviewed below.

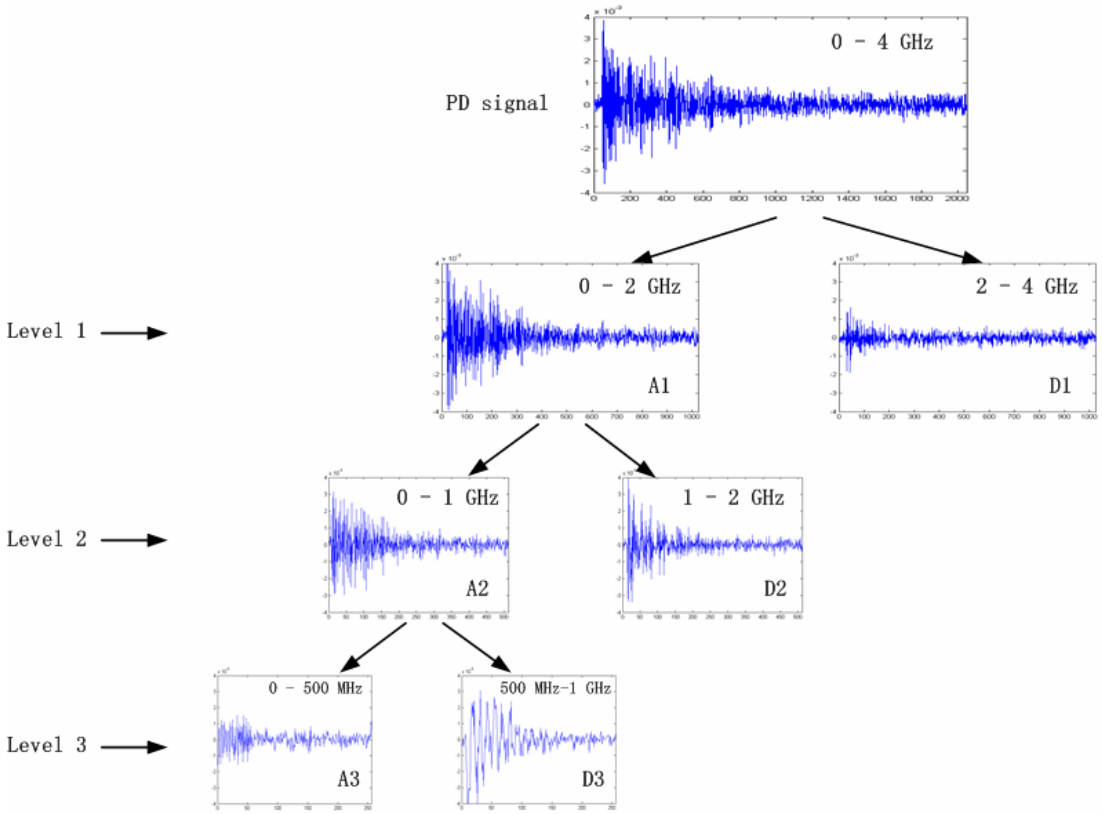


Fig. 1.12 Discrete Wavelet Transform of PD signal

Various denoizing methods are discussed in [13] with a special focus upon the wavelet-based method. The method first decomposes the PD signal into several detail components, each containing a set of decomposition coefficients. Subsequently, components that are dominated by noises are discarded. Thresholding is then performed on the decomposition coefficients of retained components, followed by the reconstruction of the denoized signal. Although the feasibility of applying wavelet transform to PD signal denoizing is studied, the denoizing performance in terms of signal-to-noise ratio and distortion is however not fully investigated as only graphic

results are presented without any numerical calculation. Furthermore, the selection of detail components for reconstruction is based on observation, which is not robust for all applications. Therefore, an automated method should be developed.

In [15], a DWT-based approach is employed to denoise PD signals. A global threshold that based on standard deviation is used to remove noise components in all frequency bands. However, noise components at various frequency bands can have different standard deviation. Therefore, the method with a global threshold can encounter problems when applied on-site.

In [16-17], the issues associated with the wavelet-based PD denoizing methods, such as wavelet selection and threshold estimation are investigated. However, one threshold is applied to all detail coefficients at the first decomposition level that corresponds to high-frequency bands. Noise levels corresponding to high-frequency bands could be different. Thus, further investigation of time-frequency features at high-frequency bands should be required for PD signal denoizing.

1.2.2 Discrimination of Corona Interference

Discrimination of corona from SF_6 PD is another important issue to be addressed. In [18-19], a wavelet-based method is employed to suppress the corona noise. The method first decomposes the signal measured from IEC 270 method into components corresponding to non-overlapping frequency bands. Subsequently, the resulted components are examined for PD or corona domination by observation or a specific criterion derived from the frequency characteristics of PD and corona. Results show

that the method works well on the data obtained from the low-frequency measurement. However, the frequency contents of PD and corona signals obtained from UHF measurement are overlapped. This means that it is difficult to determine whether a component is dominated by PD or corona. Therefore, the method may not work on UHF resonance signal. Moreover, the method can not be applied online as the discrimination process is not automatic.

In [20], a method based on phase-resolved pulse-height analysis is proposed to separate corona from PD signal. The method is however not applicable to UHF signal, as the fingerprint is derived from PD charge which is not available from UHF measurement.

Methods based on neural networks are proposed in [21-23] to classify PD and corona. Using the measured signals or phase-resolved PD patterns as input, various neural network structures are constructed and trained for discrimination of corona. These methods however do not provide a detailed discussion on feature extraction, which is crucial for neural network design and its classification performance. Moreover, the neural networks employed in [21-23] have very complicated structures, which prevent them from online application due to the slow response. Hence, there comes the need to develop a new scheme for discrimination of corona and PD.

1.3 REVIEW OF PARTIAL DISCHARGE SOURCE RECOGNITION

Traditionally, the approach using phase-resolved PD (PRPD) patterns has been widely employed to monitor partial discharge activities [23-25]. Here the total charge transferred during a discharge and the time or ac phase at which the discharge occurs are measured. In addition, the total number of PD events occurring within a time interval is counted. Based on these parameters, PRPD pattern analysis investigates the PD magnitude and/or PD repetition rate in relation to voltage ac cycle, which is equally divided into a certain number of windows. Typical PRPD patterns, accumulated over a number of cycles, are shown in Figs. 1.13 and 1.14.

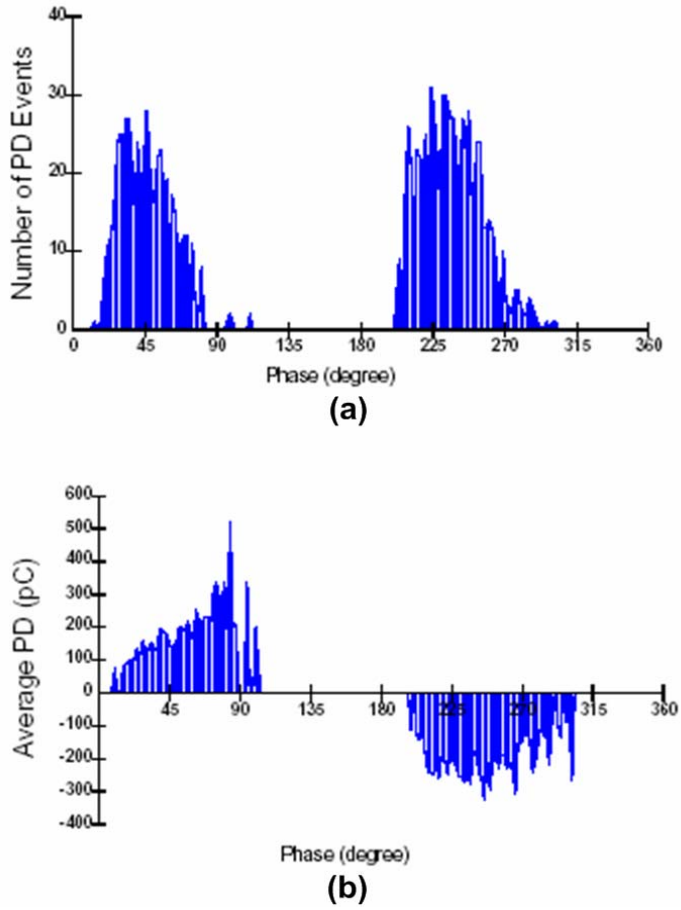


Fig. 1.13 Two-dimensional PRPD patterns (a) PD repetition rate against phase; (b) PD amplitude against phase

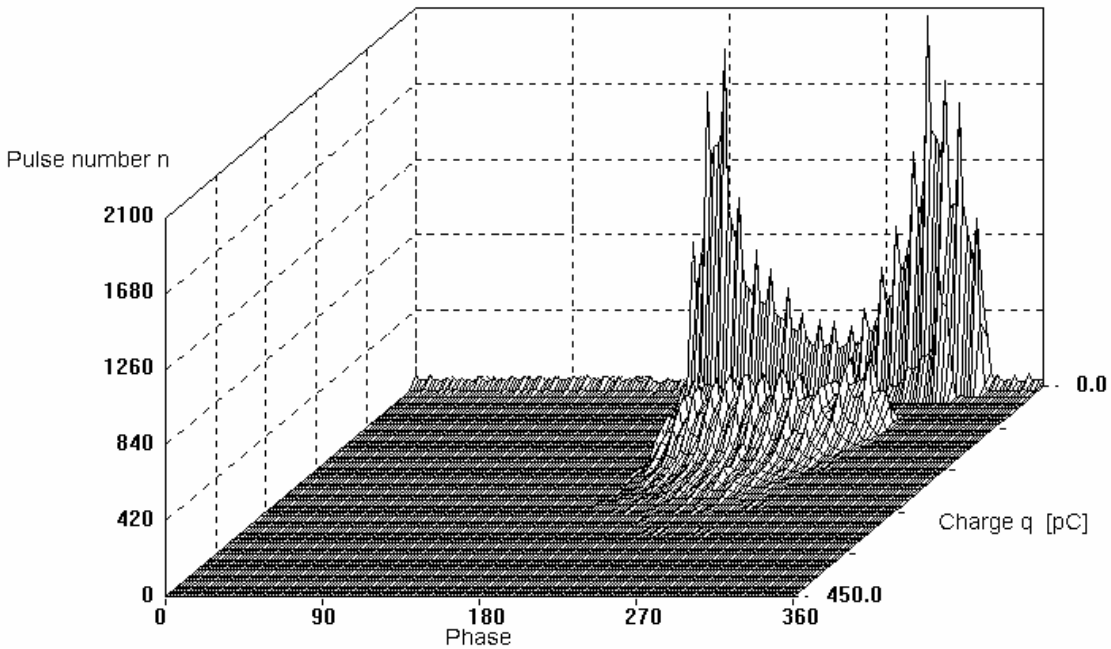


Fig. 1.14 Three-dimensional PRPD pattern

A variation of PRPD known as point-on-wave (POW) analysis is also commonly employed in UHF PD source recognition in GIS [3, 26-27]. POW is different from PRPD in that only a specified frequency range is scanned for PD occurrence. In other words, it is a narrow-band approach. The PD amplitude is then recorded with respect to the phase angle to build up the POW over a large number of power cycles.

In [3, 23-27], features are extracted from the PRPD or POW patterns using envelop extraction, statistical methods, orthogonal transforms, unsupervised neural networks or fractals method. Subsequently, various classification schemes are developed to identify defects based on the extracted features. However, results of these methods show large classification error due to the variety of the patterns produced by defects of the same type as shown in [26]. Another major drawback with these approaches is that they require signals measured within a few seconds or even longer to form the PRPD or POW patterns before feature extraction and classification. On the other hand, PD can progress very quickly from initiation to breakdown in GIS, particularly in high-pressure SF_6 for working voltages at 300 kV and above. In addition, more than one type of PD can take place in the GIS chamber during the forming PRPD or POW patterns [3]. This has resulted in inaccurate PRPD or POW patterns and lead to further misclassification. There is therefore an urgent need to develop a fast and reliable diagnosis method for source recognition of PD.

1.4 OBJECTIVES AND CONTRIBUTIONS OF THE THESIS

Through the background review, the traditional denoizing and source recognition methods are considered to be insufficient to provide fast and reliable diagnosis of insulation system in GIS. Thus, in contrast to the PRPD- or POW-based methods, a novel scheme based on UHF signals with duration of several hundred nanoseconds is developed in this thesis as shown in Fig. 1.15. As data are collected in much shorter windows, the possibility of encountering more than one type of discharge signals during measurement and subsequent classification is very small. In addition, the short data acquisition time enables the development of fast PD diagnosis system which can be potentially applied online. Therefore, the problems with PRPD- and POW-based methods are basically solved through the use of UHF signal directly.

1.4.1 Objectives of the Project

As reviewed in Section 1.1.3, it is hard to achieve reliable PD diagnosis if signals with high level of white noises are employed in the classification process. Regarding the issue of corona noise discrimination, since it is a classification problem in nature, it can be considered together with the source recognition of SF_6 PD. Moreover, the PD fingerprints derived from UHF signals have to be established as little work has been done in this area. Therefore, following objectives are set for this thesis:

(1) To develop an effective denoizing method that is able to suppress excessive white noise and restore the original PD signal with little distortion.

(2) To establish a wide range of PD parameters from UHF signals as a solid base for current and future work on PD pattern recognition.

(3) To select features with the largest discriminating power to form compact and high-quality PD fingerprints, so that the speed and classification performance are improved significantly.

(4) To investigate the robustness of the PD features on various measuring conditions.

As UHF PD measurement is employed in this research instead of the traditional IEC 270 measurement, modeling of the UHF PD signal involves modeling of signal propagation in GIS using numerical transient electromagnetic field analysis, which is another area of research. Therefore, modeling of UHF PD signal is not included in this research.

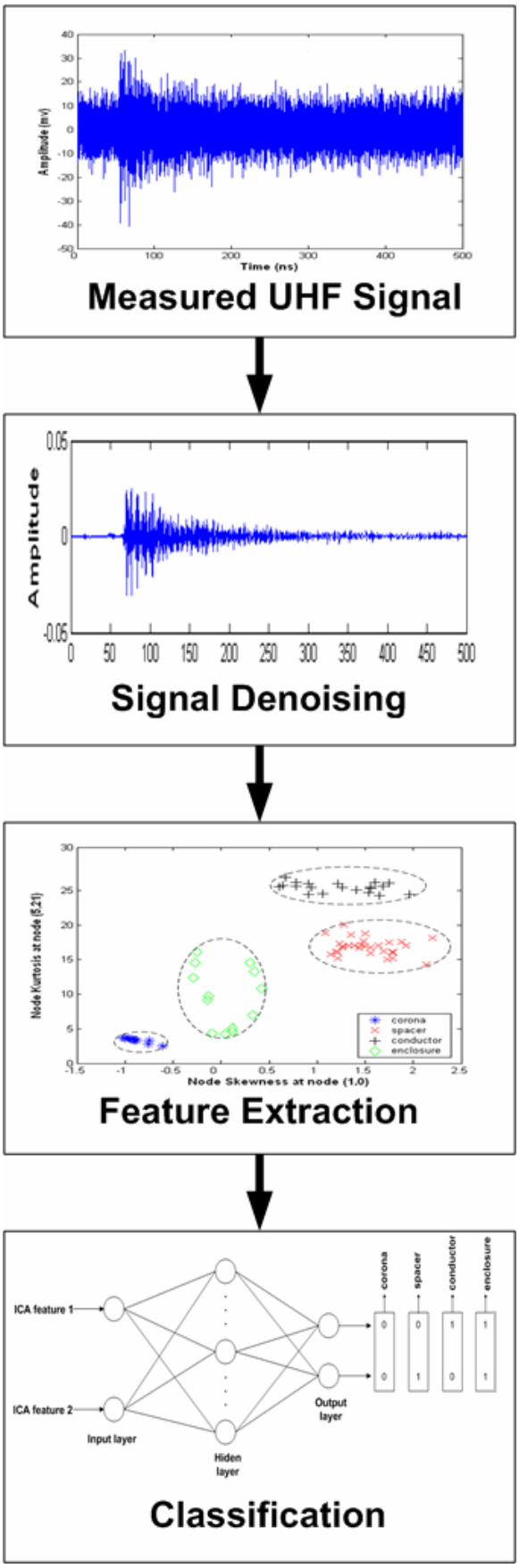


Fig. 1.15 PD diagnosis procedures

1.4.2 Author's Main Contributions

The contributions of this project are summarized as follows:

(1) To build a novel PD diagnosis software system based on UHF signals with short duration, so that the speed and classification accuracy can be greatly improved. The new method is also promising for other applications such as PD diagnosis in DC power transmission system, where phase reference is not available. All the algorithms developed in this thesis have been tested with 256 sets of data measured in the laboratory of TMT&D Co.

(2) To develop a novel wavelet-packet-based method for effective PD signals denoizing.

(3) To optimize the parameters of wavelet-packet-based denoizing method to achieve best denoizing performance.

(4) To introduce new waveform-based PD fingerprints to classify PD source of different types.

1.5 OUTLINE OF THE THESIS

The overall structure of this thesis is illustrated in Fig. 1.16. Content of each chapter is briefly described as follows:

Chapter 1 provides brief background information about PD and its measurement in GIS. Previous works on noise reduction and source recognition of PD signals are reviewed. Based on this, the objectives of current project are outlined with the contributions made by the author.

Chapter 2 studies the denoizing of UHF PD signals using wavelet packet transform. A novel variance-based criterion is developed to select the best tree from wavelet packet decomposition tree for improving the denoizing. Selection of other denoizing parameters is also studied based on overall performance. Results from different denoizing methods are presented and compared.

Chapter 3 addresses the issue of optimal parameters selection for wavelet-packet-based denoizing. A method based on genetic algorithm is proposed to automatically optimize the set of denoizing parameters. Denoizing performance of the optimized parameters is compared with those obtained in Chapter 2.

Chapter 4 and Chapter 5 develop novel methods for PD feature extraction based on UHF signals with short duration. In Chapter 4, a time-domain technique known as Independent Component Analysis (ICA) is employed to perform the feature extraction. ICA is first introduced through a comparison with the well-known Principal Component Analysis. Subsequently, ICA-based feature extraction method is described followed by experimental results.

Chapter 5 proposes a time-frequency domain method for PD feature extraction, which is based on the wavelet packet transform. Firstly, the wavelet-packet-based method is described followed by a discussion of parameters selection for feature extraction purpose. Then numerical results are presented and the necessity of denoizing is justified. Lastly, the relation between wavelet-packet PD features and Fast Fourier Transform (FFT) PD features is clarified.

Chapter 6 implements a simple multilayer perceptron (MLP) neural network to classify PDs based on the extracted PD features. Firstly, a general introduction to neural networks is given. Secondly, training and test of the MLP is studied with discussions on the network parameters selection. Lastly, the usefulness and effectiveness of the extracted features are proved by results of comparative studies.

Chapter 7 investigates the robustness of selected PD features on data measured under various conditions. A general scheme for ensuring the robustness of PD identification within the test GIS section is first described, and is followed by its implementation in ICA- and wavelet-based methods. Numerical results are then presented and discussed.

Chapter 8 contains the conclusions and recommendations for future work.

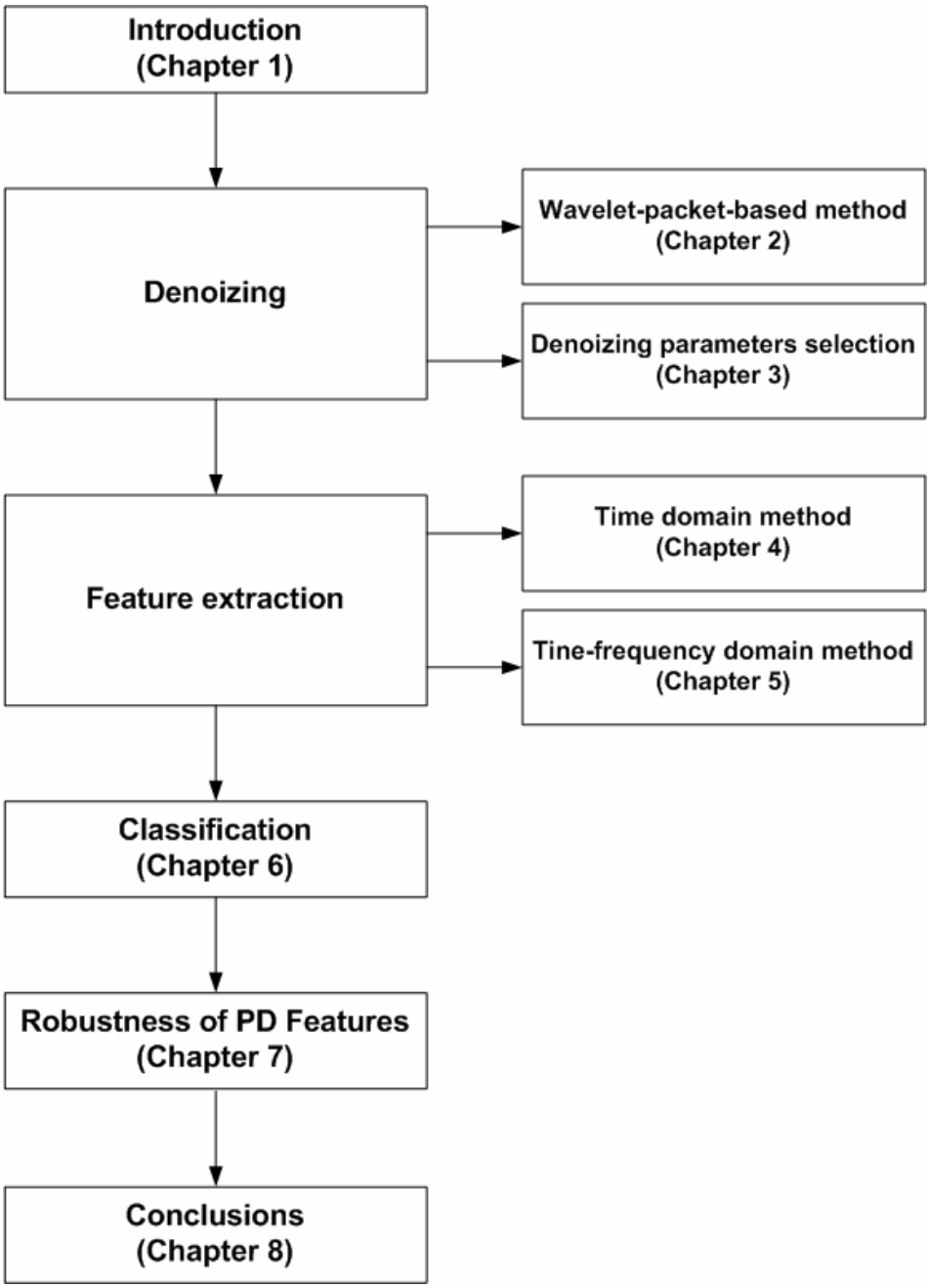


Fig. 1.16 Overall structure of this thesis

CHAPTER 2

DENOIZING OF PD SIGNALS IN WAVELET PACKET DOMAIN

In Chapter1, the background information about PD and its measurement has been introduced. Previous research on noise reduction and PD source recognition has been reviewed and a novel PD diagnosis scheme has been proposed. In this chapter, denoizing of UHF PD signals using wavelet packet transform is studied. First, wavelet packet transform and the general wavelet-packet-based denoizing scheme are briefly reviewed. Secondly, the proposed denoizing scheme is described with special emphasis on a novel approach for best tree selection. Lastly, numerical results are presented and discussed.

2.1 INTRODUCTION

As reviewed in Chapter 1, wavelet-based methods do not perform well in denoizing PD signal due to the poor frequency resolution at high frequencies with wavelet transform. On the other hand, the wavelet packet transform (WPT) [31] describes a rich library of bases (wavelet packets) with an arbitrary time-frequency resolution for overcoming the drawback. By applying linear superposition of wavelets, desirable properties of orthogonality, smoothness, and localization of the mother wavelets are retained.

Based on WPT, a general method was proposed in [31] and implemented in a software package [42] for signal denoizing. However, the method is found in this work not applicable to PD signals in terms of noise level reduction and restoration of the original waveform, as it was only developed and tested on standard waveforms, such as sine waves. The major drawback of the method is that the criterion employed for selecting PD dominated decomposition components may cause loss of critical PD information, leading to poor denoizing performance. An outline of the general method and its shortcomings is given in Section 2.2.2 and 2.2.3 respectively.

To address the above-mentioned issue with the general denoizing method, a novel variance-based criterion is proposed in Section 2.3.2 for selecting the most effective components from the wavelet-packet-decomposition tree. Moreover, a scheme is proposed in the flowchart of Fig. 2.1 for determination of the "best" choice of denoizing parameters, such as wavelet filters, decomposition level and thresholding parameters, in terms of noise reduction and original signal restoration. A

comprehensive database containing 256 data records was built for developing and verifying the new denoizing method as well as the new PD source identification methods, which will be discussed in chapters 4 to 7. Data were collected by TMT&D from a test section of an 800 kV GIS [89], where PD of various types and locations were initiated by applied voltages of various values. Details of the equipment specifications and experimental set-up are given in Appendix A. Numerical results are shown in Section 2.4 to compare the performance of various denoizing parameters and methods, where signal-to-noise-ratio (SNR) and correlation coefficient (CC) are employed to evaluate noise reduction and signal restoration respectively.

In Fig. 2.1, a mechanism is also proposed for verifying the performance of determined denoizing parameters on new data by dividing the measured signals into a training set and a test set, using which a genetic-algorithm-based method is developed in Chapter 3 to optimize the entire set of denoizing parameters.

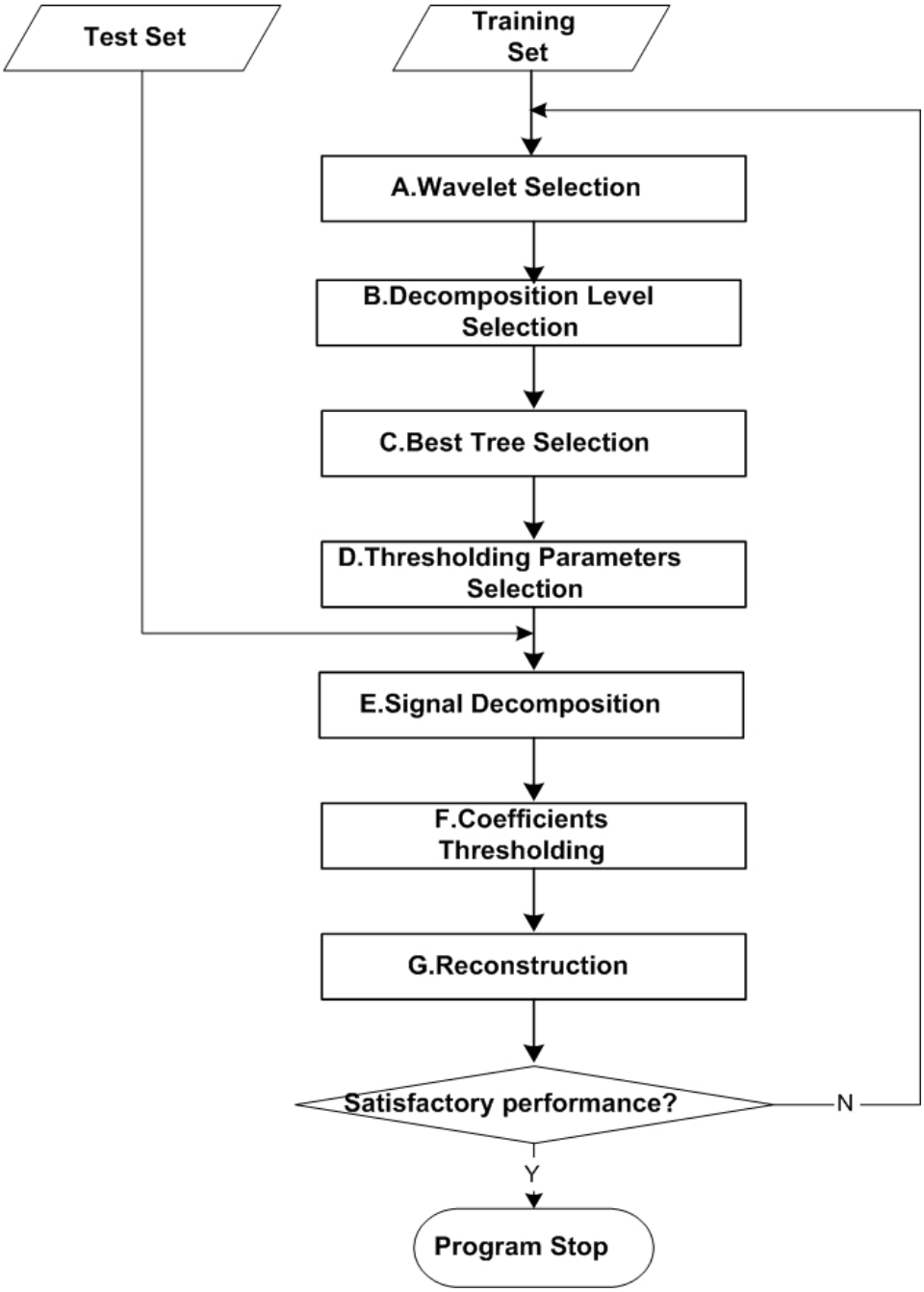


Fig. 2.1 Proposed denoizing scheme

2.2 WAVELET PACKET TRANSFORM AND THE GENERAL WAVELET-PACKET-BASED DENOIZING METHOD

2.2.1 Introduction to Wavelet Packet Transform

Wavelet packet transform (WPT) is a direct expansion from the DWT pyramid tree algorithm (Fig. 2.2(a)) to a binary tree (Fig. 2.2(b)), where each branch of the tree has two sub-branches. It is the generalization of DWT in that both the low-pass and the high-pass output undergo splitting at the subsequent level. Therefore, WPT is seen to have the capability of partitioning the high-frequency bands to yield better frequency resolution. The equations of WPT under level j are defined as:

$$\omega_{j+1,2n}(k) = \sum_m h(m)\omega_{j,n}(2^j m - k) \tag{2.1}$$

$$\omega_{j+1,2n+1}(k) = \sum_m g(m)\omega_{j,n}(2^j m - k) \tag{2.2}$$

where h, g are the low-pass and high-pass decomposition filter respectively. $\omega_{j,n}(k)$ represents the kth decomposition coefficient at node (j,n), namely the nth node of level j. Fig. 2.3 shows the 3D plot of the decomposition coefficients corresponding to the WPT binary tree of Fig. 2.2(b).

The complete binary tree resulted from WPT contains many nodes. It follows that the terminal nodes (leaves) of every connected binary subtree of the complete tree form an orthogonal basis of the signal space. Therefore, to achieve the best denoizing performance, there is a need of choosing the best nodes subset (best tree) for

representing a signal in wavelet packet domain. A review on the DWT and the generalized WPT is given in Appendix B.

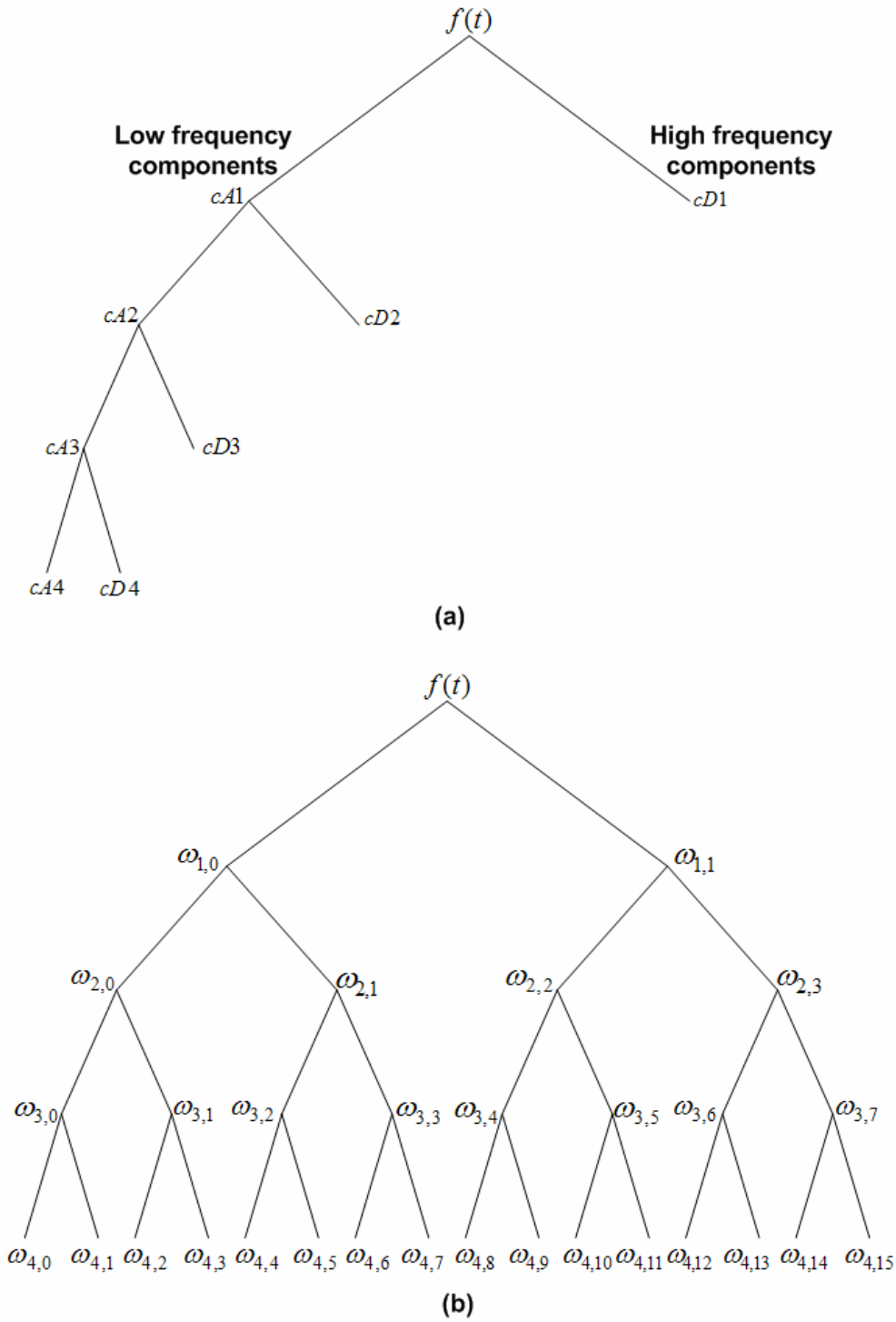


Fig. 2.2 The decomposition tree structure of (a) DWT and (b) WPT

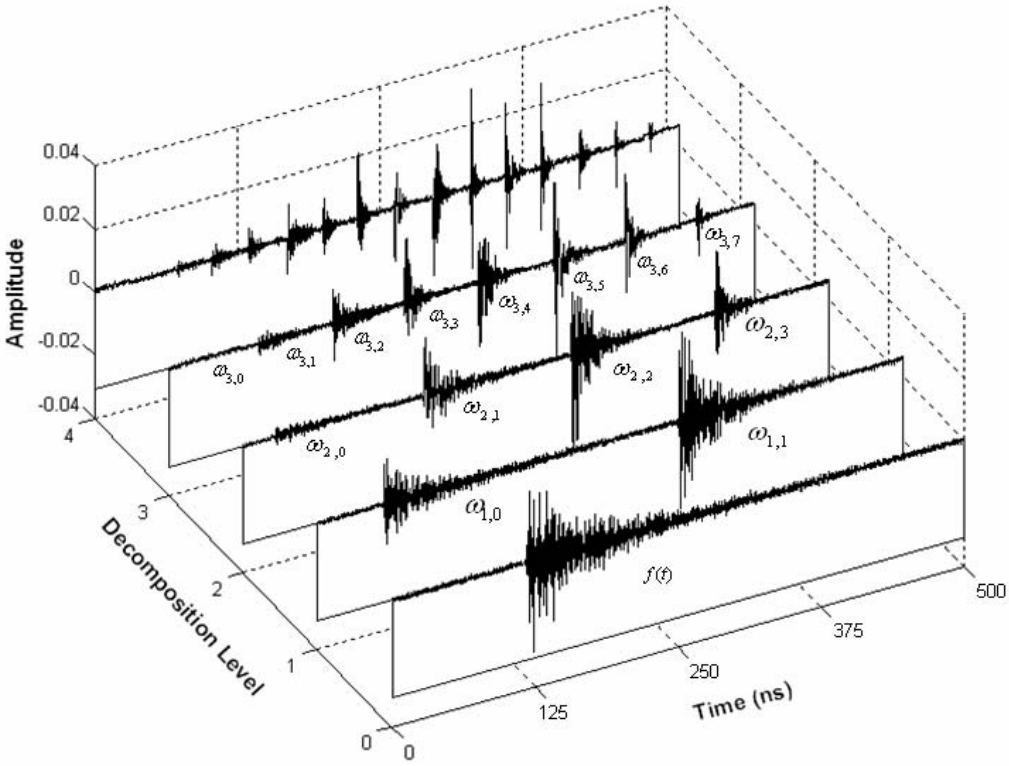


Fig. 2.3 3D plot of decomposition coefficients in WPT tree

Typical applications of WPT include biomedical engineering [32-33], signal [34] and image [35] processing. Recently, WPT has been successfully applied to various fields in power system, such as power system disturbances [36-38], energy measurement [39] and fault identification [40]. However, only a limited number of publications on the application of WPT to PD analysis have been reported. In [41], WPT was employed to compress PD data.

2.2.2 Introduction to the General Denoizing Method

A brief introduction of the general method is given in this section. Fig. 2.4 shows the procedure of the denoizing method.

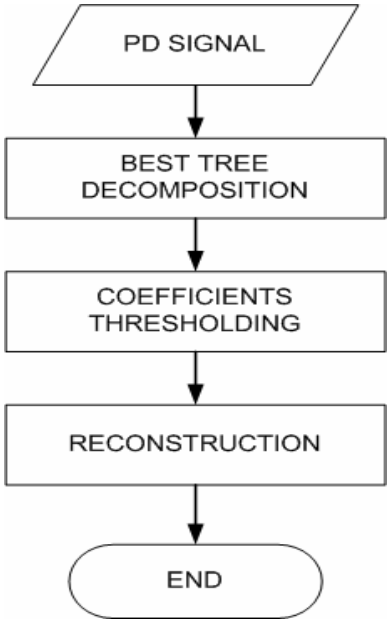


Fig. 2.4 Procedure of the standard denoizing method

The standard method is started by creating a “father” node from a given PD signal. Then the best tree decomposition (splitting process) is carried out as follows:

(1) Compute the entropy of the decomposition coefficient vector of the "father" node based on a predetermined entropy function. Denote the entropy value [42] by C_f.

(2) Split the "father" node into two "child" nodes by one-step-DWT using a predetermined wavelet.

(3) Compute the entropies of the decomposition coefficient vectors of the "child" nodes, denoted by C_{c1} and C_{c2} respectively.

(4) Compare C_f with the sum of C_{c1} and C_{c2}. If C_f is larger, the "child" nodes are kept. Otherwise, the "child" nodes are discarded.

(5) Choose the next node at the current decomposition level as the "father" node and go to step (2). If all the nodes at the current level have been split, go to next level and select the leftmost node as the "father" node. Then go to step (2). If the last node of level J-1 has been examined where J is the specified decomposition level, the process stops.

Many entropy functions can be used in the above process, such as Shannon entropy, logarithm of the "energy" entropy, threshold entropy, and so on [42]. The Shannon entropy is used in the present experiment due to its proven suitability for wavelet packet analysis [43].

After decomposition, white noises are removed in wavelet packet domain by thresholding of the decomposition coefficients. Finally, the denoized signal is reconstructed by wavelet packet reconstruction.

2.2.3 Shortcomings of the General Method

The method in [31] provides optimal representation of a signal by minimizing the mean-square-error for a given set of data. It however does not provide an optimal choice of nodes for denoizing weak PD signals that are corrupted by high-level noises due to significant loss of PD information during the splitting process, as described below. The splitting stops prematurely and both of the "child" nodes are discarded when the entropy of the "father" node is smaller than the sum of the entropies of the two "child" nodes. There is no checking on the entropy of individual child nodes.

This would cause information loss representing the features of the PD. In addition, the best tree structure resulted from the splitting has to be constructed every time when a new PD signal is presented. This is inefficient as the tree structure can be determined from a set of typical PD signals and kept unchanged for all the signals that are going to be processed. Thus, a more efficient PD denoizing strategy is required to address these issues.

2.3 A NEW WAVELET-PACKET-BASED DENOIZING SCHEME FOR UHF PD SIGNALS

2.3.1 Introduction

A novel variance-based criterion is developed for selecting the best tree from wavelet-packet-decomposition tree for denoizing PD signals. The comprehensive scheme proposed in the flowchart of Fig. 2.1 is further described as follows. Measured PD and corona signals are first divided into two sets, namely the training and test sets for selecting and verifying the denoizing parameters respectively. The training set is used to determine the optimal parameters required for the remaining denoizing process. The optimal wavelet for the wavelet packet decomposition is first selected, and followed by the selection of decomposition level. The selection of best decomposition tree is then performed. Parameters related to thresholding are set. The test set is entered at a much later part of the proposed scheme of Fig. 2.1. The process of signal decomposition and coefficients thresholding are applied to both the training and test sets. Finally, the denoized signal is reconstructed and the denoizing performance is evaluated by signal-to-noise ratio (SNR) and Correlation Coefficient. Another round of training will be carried out, should the post-denoizing performance be below a pre-

determined performance level. The method is seen to capture the features of PD signals better than the earlier methods [13, 15-17, 42] and thus has a better denoizing performance.

2.3.2 Parameters Setting for Denoizing

In order to achieve the best denoizing performance, it is crucial to set the parameters associated with the denoizing scheme properly. However, since PD signals corresponding to various defects exhibit different characteristics such as waveform and frequency content, optimal parameters for signals of one class may not perform well on the signals of other classes. For instance, wavelet "db4" achieves good performance on corona signals but fails to denoise SF_6 PD signal of free particle. Therefore, signals of each class should ideally have their own set of optimal parameters. In practice, however, the class information is unknown at first. Thus the parameters should be set by using a set of training signals with all existing types of PD and corona signals, so that they can denoise all types of signals relatively well. With this in mind, a training set that contains 24 UHF signals, 6 from each class of PD and corona signals, is constructed to determine all the parameters except the best tree structure. The best tree structure is determined using an extended training set of size 48, which contains the original training set and 24 white noise signals. Details of finding the best parameters are discussed in the following subsections.

A. Selection of wavelet for wavelet packet decomposition (WPD)

There are two important issues for the WPD that affect the denoizing performance, namely: the selections of optimal wavelet and decomposition level.

The first task to be accomplished with the training set is to identify the optimal wavelet (Fig. 2.4), which best describes a set of PD signals. In this thesis, a method based on *minimum-prominent-decomposition coefficients* [44] is extended to choose the optimal wavelet from a set of candidate wavelets, such as Daubechies, Symlets, Coiflets and Biothogonal wavelets. The flowchart of the method is shown in Fig. 2.5.

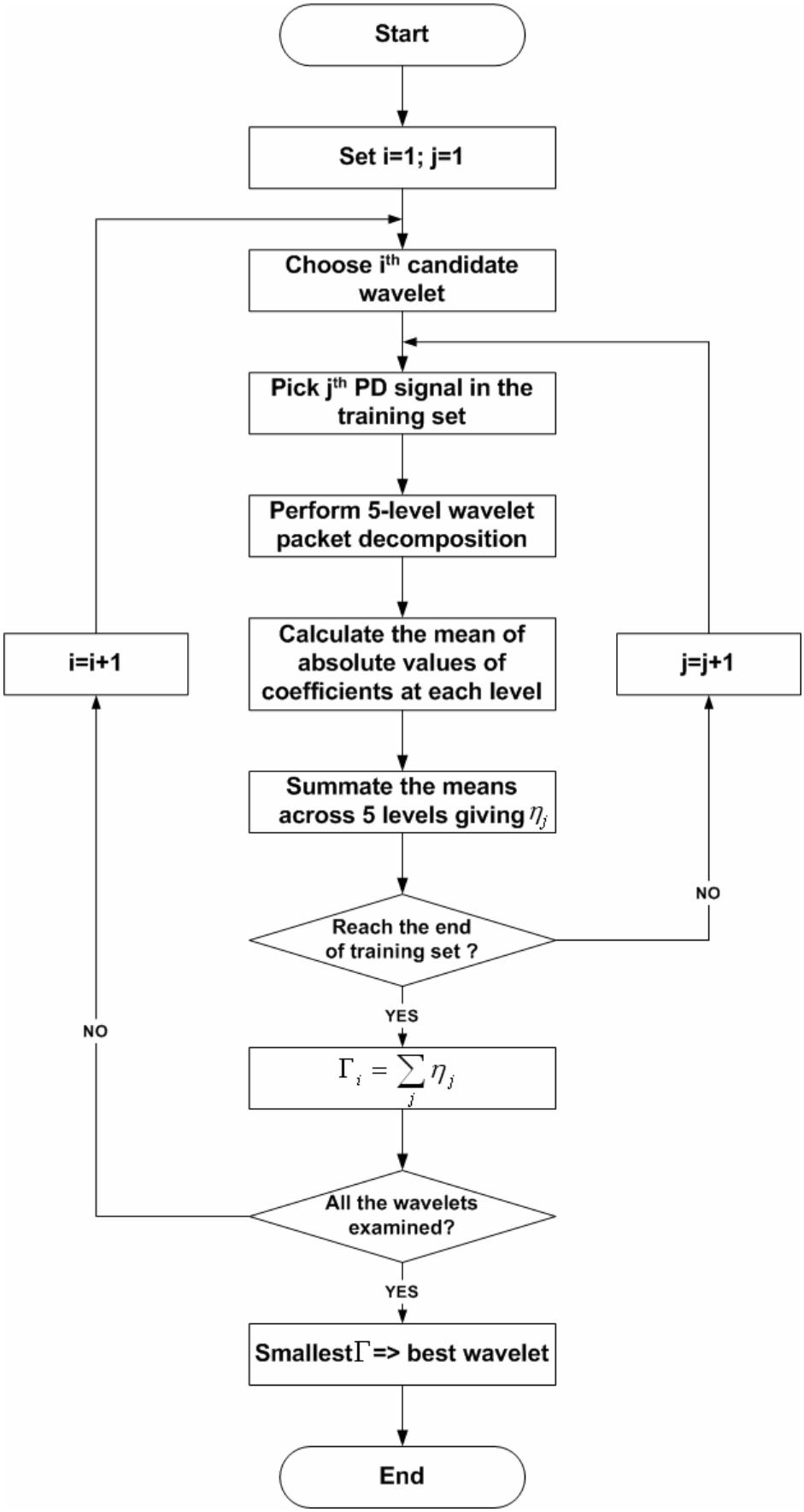


Fig. 2.5 Flowchart of best wavelet selection

For each candidate wavelet, the method first decomposes the jth PD signal of the training set into wavelet packet domain down to a predetermined level of 5 as shown in Fig. 2.6. Secondly, the mean value of the absolute values of detail coefficients is calculated for each decomposition level and then summated across all the five decomposition levels forming η_j. The value η is computed for all the other signals in the training set and summated to give Γ. The value of Γ indicates how closely the candidate wavelet is describing the PD signals. A small Γ indicates good performance of the candidate wavelet. The procedure is then applied to all the other wavelets. The wavelet giving the lowest Γ is chosen as the best wavelet. As a result, the 'sym8' wavelet is obtained from the training set. The effectiveness of the above procedure is illustrated in Fig. 2.7. As observed, the shape of the selected wavelet, which results in the smallest Γ, best represents the PD signal that is resulted from a free particle. Similar results are obtained on the other type of PD and corona signals.

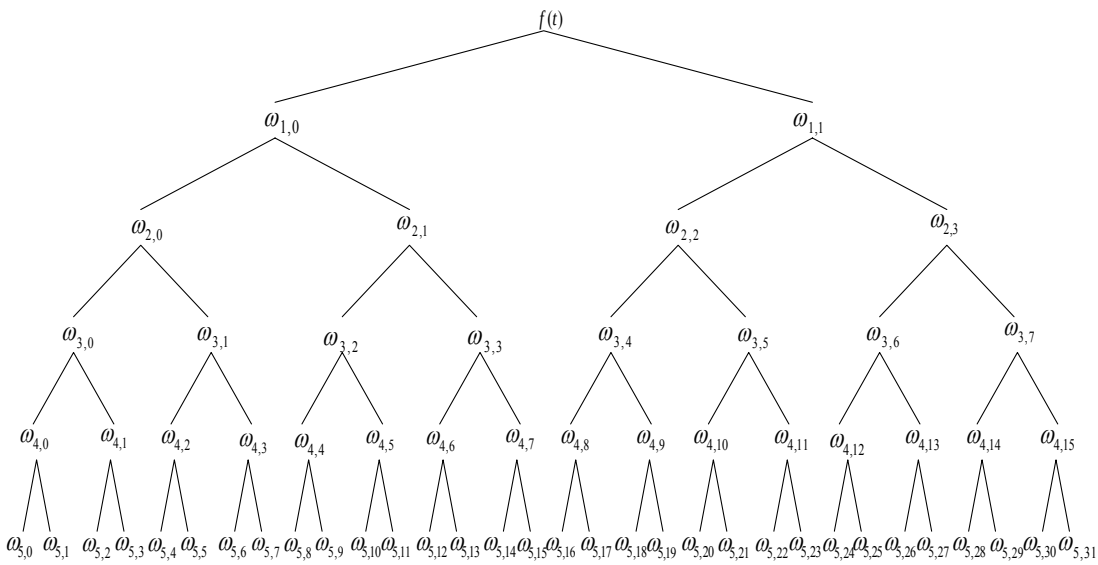


Fig. 2.6 WPD tree structure with a decomposition level of 5

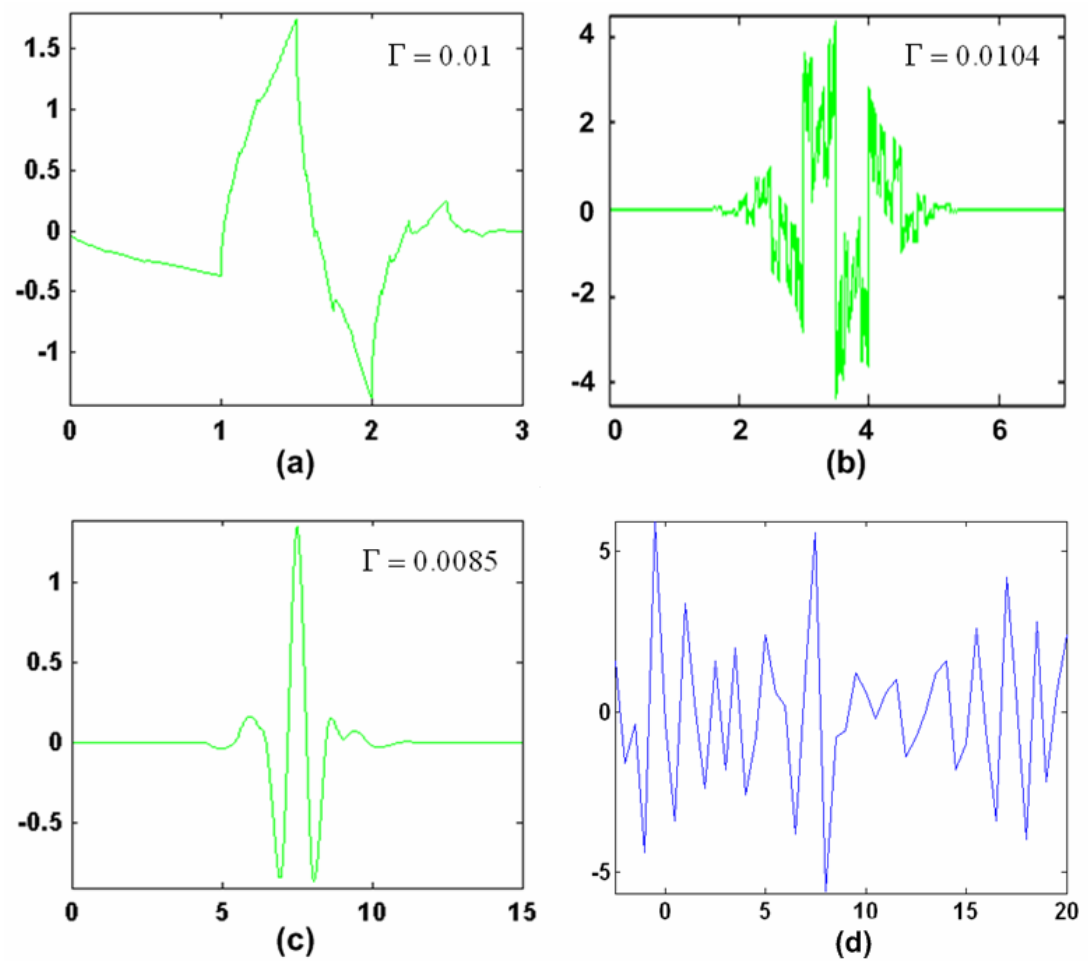


Fig. 2.7 Comparison of wavelets (a) db2; (b) bior3.3; (c) sym8; (d) PD signal.

B. Selection of decomposition level for denoizing

After its selection, the best wavelet performance at different decomposition levels is evaluated using the signal-to-noise ratio (SNR) and Correlation Coefficient (CC). SNR is a measure of signal strength relative to background noise. The ratio is usually measured in decibels (dB). On the other hand, CC is a measure of similarity between denoized and original PD signals. Therefore, to effectively suppress the noises and restore the original PD signal with little distortion, large values of SNR and CC are

desired. As a result, a decomposition level of 5 is selected from the evaluation. Numerical results leading to the selection of the optimal wavelet and the decomposition level are further discussed in Section 2.4.

C. Proposed method for best tree selection

In order to effectively denoise PD signals, it is crucial to prune the original WPD (wavelet-packet-decomposition) tree of Fig. 2.6. The objective is to retain the "effective" nodes to best characterize the PD signals in the training set and to remove the "non-effective" nodes that are highly corrupted by white noise. The tree structure after pruning will be used for denoizing signals of both the training and test sets.

To evaluate the effectiveness of the nodes, a "union tree" is first constructed as in Fig. 2.8. Each node of the union tree is the union of the corresponding nodes in the WPD trees of all the signals in the extended training set, which consists of 24 PD signals and 24 white-noise signals. For convenience, nodes of the union tree are numbered as in Fig. 2.9.

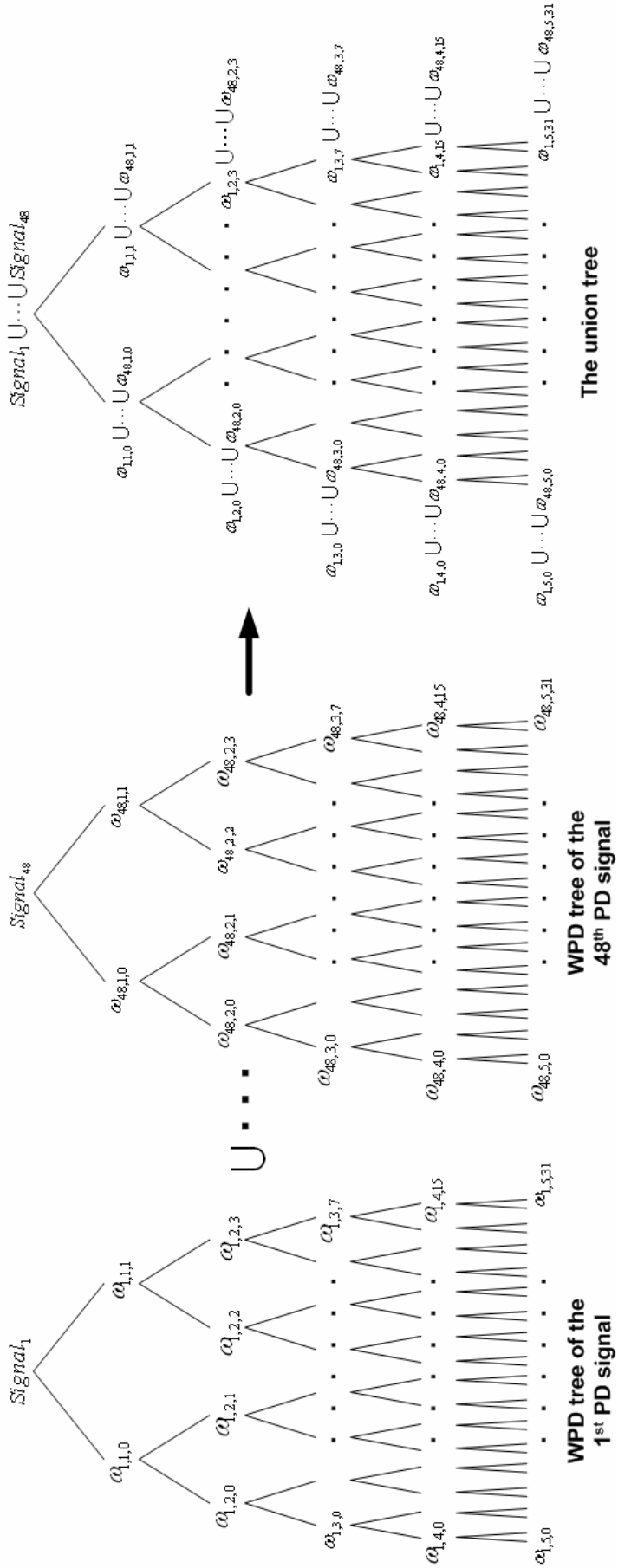


Fig. 2.8 Construction of the union tree

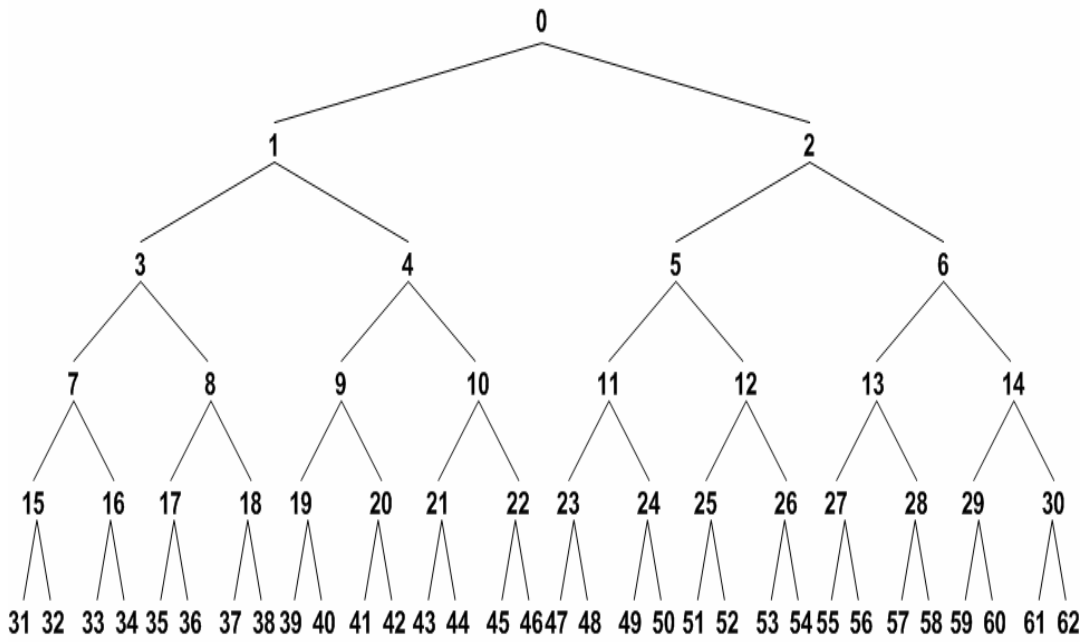


Fig. 2.9 Numbered union tree

A performance index is then required to measure the level of white noise at each node during the best tree selection. Figs. 2.10 (a) and (b) show the wavelet-packet-decomposition coefficients of a measured PD signal and a white noise signal respectively. Each grid in the figure represents a node of original WPD tree. It can be seen that the decomposition coefficients of white noise have small and similar magnitude in all the nodes, while decomposition of PD signal results in large coefficients in the PD-dominated nodes. Therefore, if a node of the original WPD tree is dominated by all the PD signals in the extended training set, then the coefficients in the corresponding node of the union tree have the largest standard deviation as shown in Fig. 2.11(a). Fig. 2.11(b) shows the case where the node is partially dominated by PD and (c) illustrates a noise-dominated node. It is seen that the standard deviation of the coefficients of a node in the union tree, which is defined as *global standard*

deviation, reflects the degree of PD domination of the node. It is thus computed for each node of the union tree to evaluate its effectiveness.

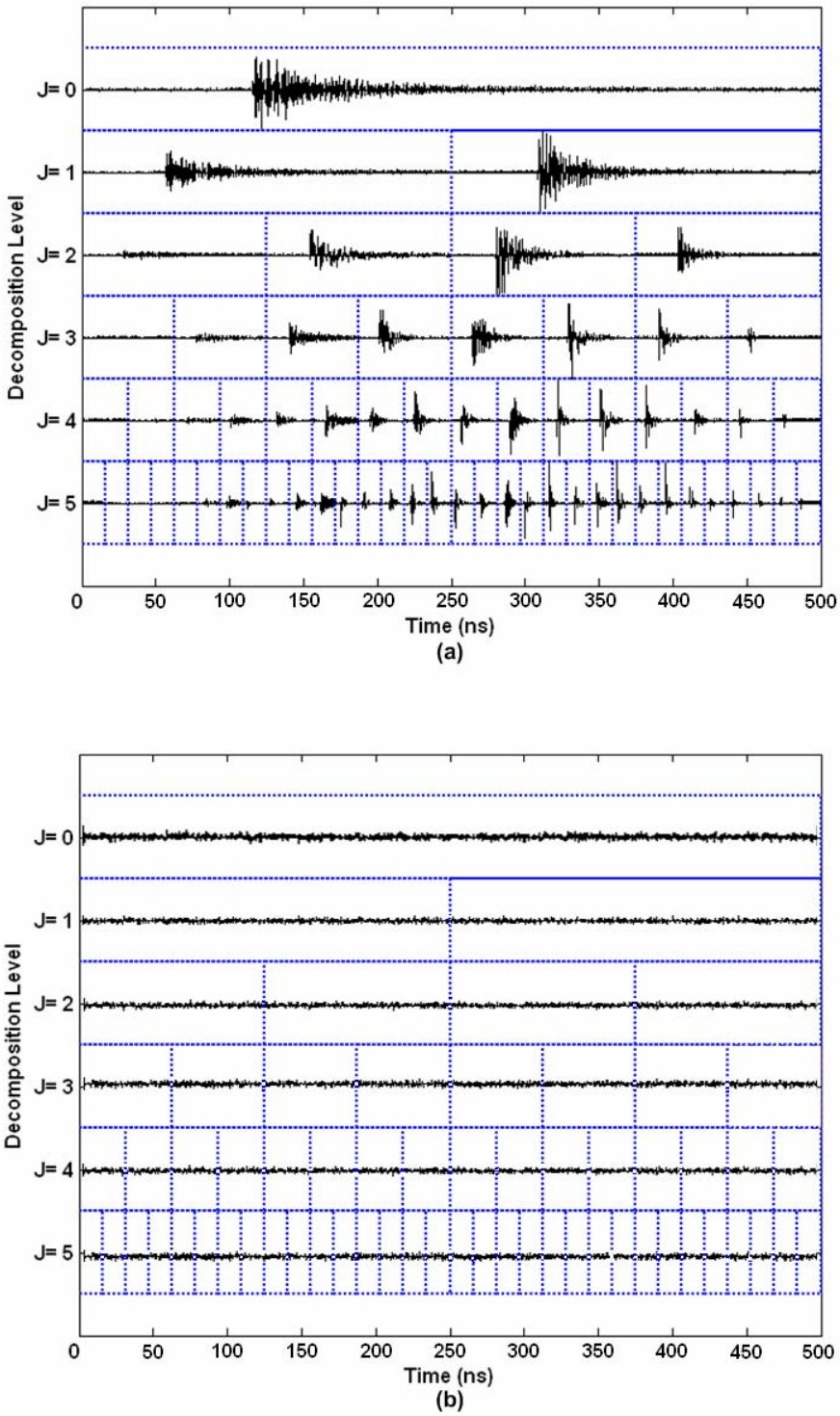


Fig. 2.10 Wavelet-packet-decomposition coefficients of (a) PD signal; (b) white noise signal.

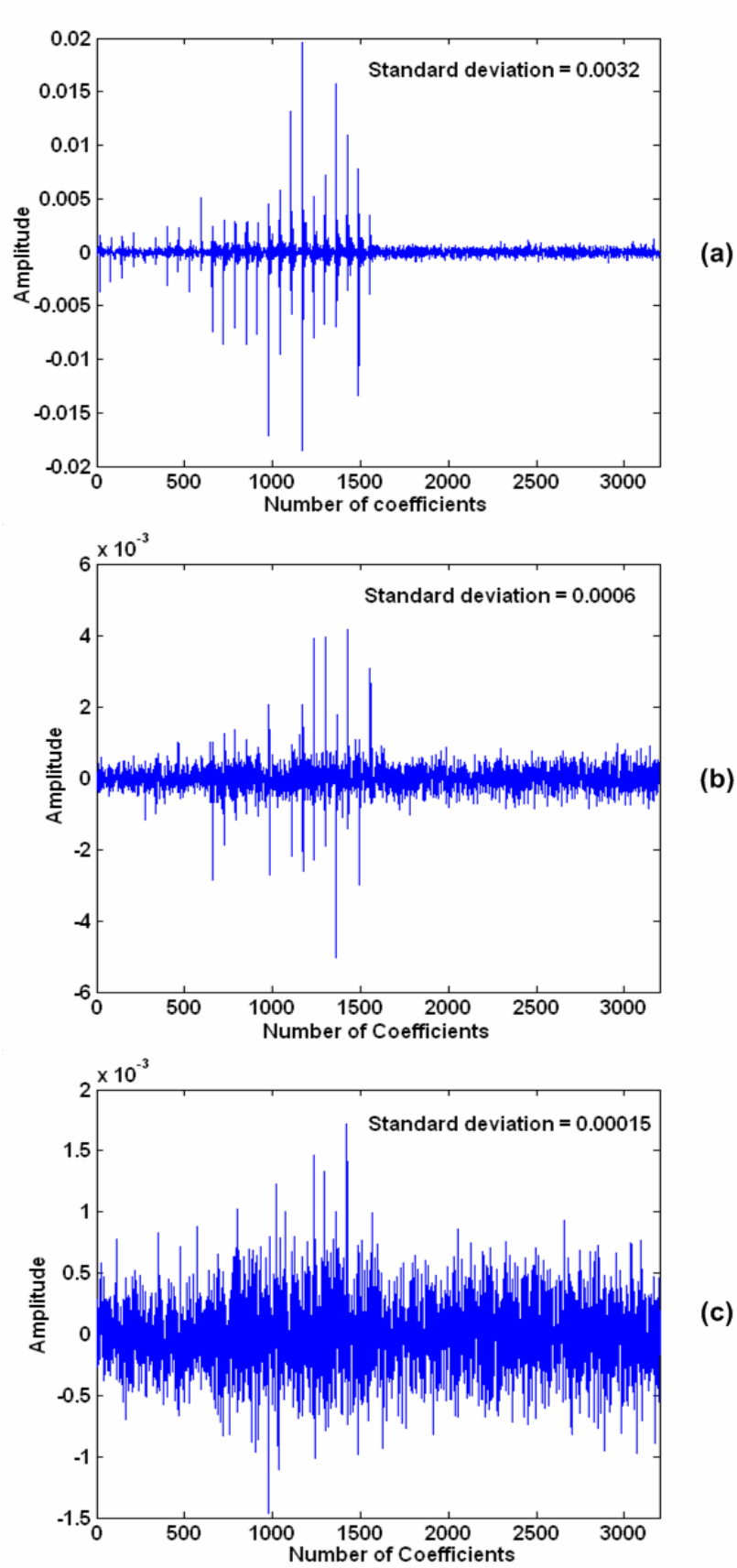


Fig. 2.11 Nodes of the union tree (a) node 50 – dominated by PD; (b) node 53 – partially dominated by PD; (c) node 34 – dominated by noise.

The global standard deviation λ_n for the nth node of the union tree is given as:

$$\lambda_n = \sqrt{\frac{1}{M-1}\sum_{k=1}^{M}(c_n^k - \mu_{c_n})^2} \tag{2.3}$$

where

c_n = the decomposition coefficient vector of nth node of the union tree.

μ_{c_n} = the mean of c_n.

M = the number of coefficients in n^{th} node.

n = number of nodes. Runs from 1 to 62 for a decomposition level of 5.

Fig. 2.12 shows the calculated global standard deviations for nodes of the union tree. Nodes with small global standard deviations that are marked with (*) in Fig. 2.12 are thus considered white-noise corrupted and to be removed from the original WPD tree. Only nodes with large global standard deviations that are marked with (o) are retained in the best tree structure due to strong PD domination.

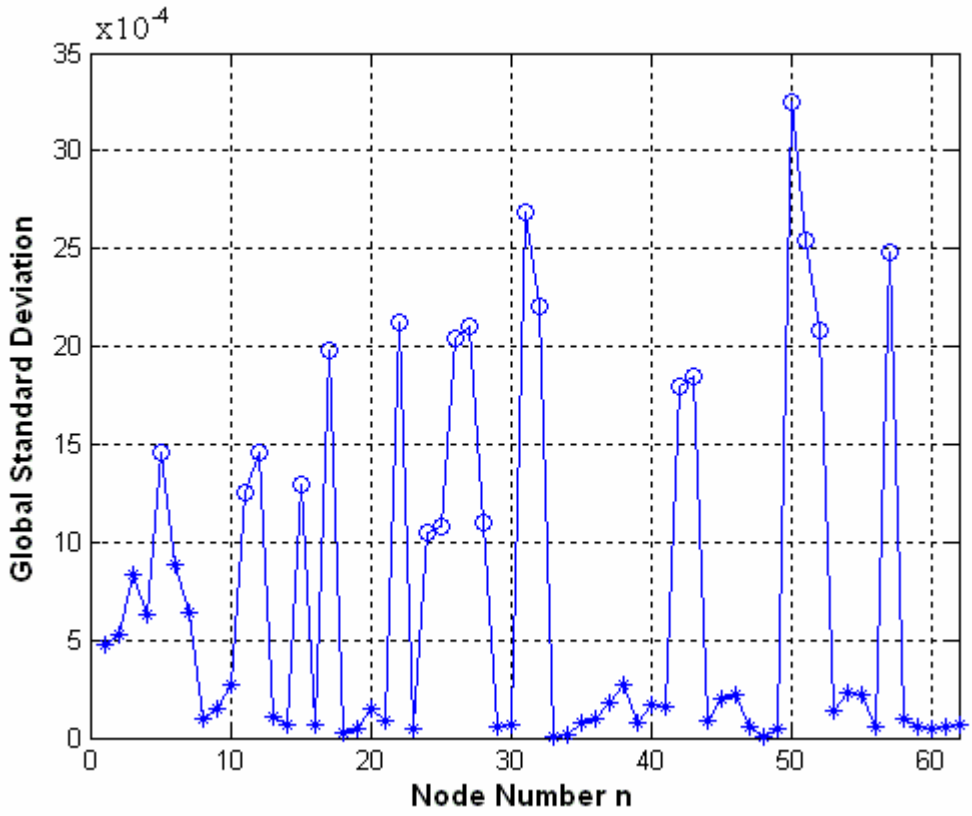


Fig. 2.12 Global standard deviations on each node of the union tree

Aside from having large global standard deviations, nodes retained from the above procedure must meet the *orthogonality condition* [45]. The method of bi-directional priority registration (BPR) is proposed here to meet the condition, using which a complete pruning of the original WPD tree is performed to obtain the best tree as follows:

(1) Calculate for each node in the union tree its global standard deviation as in Fig. 2.12. Rank the nodes in descending order of the magnitude of their global standard deviations.

(2) Remove those nodes from the ranking in (1), whose global standard deviations are below a predetermined value (set to 0.001 in this study based on extensive

study).

(3) Starting from $i = 1$ on the node with highest global standard deviation.

(4) Trace back the family tree of node i, and remove all father node(s) from the current ranking.

(5) Remove all the child nodes of node i from the ranking.

(6) Descend to the next node in the current ranking, $i = i+1$. Go to step 7 if it goes beyond the end of ranking. Otherwise go to (4).

(7) The resulted ranking will provide the best tree structure.

Fig. 2.13 shows the obtained best tree, using which denoizing of PD signals is carried out. Comparative studies of the overall denoizing performance with other proposed methods are presented in Section 2.4.

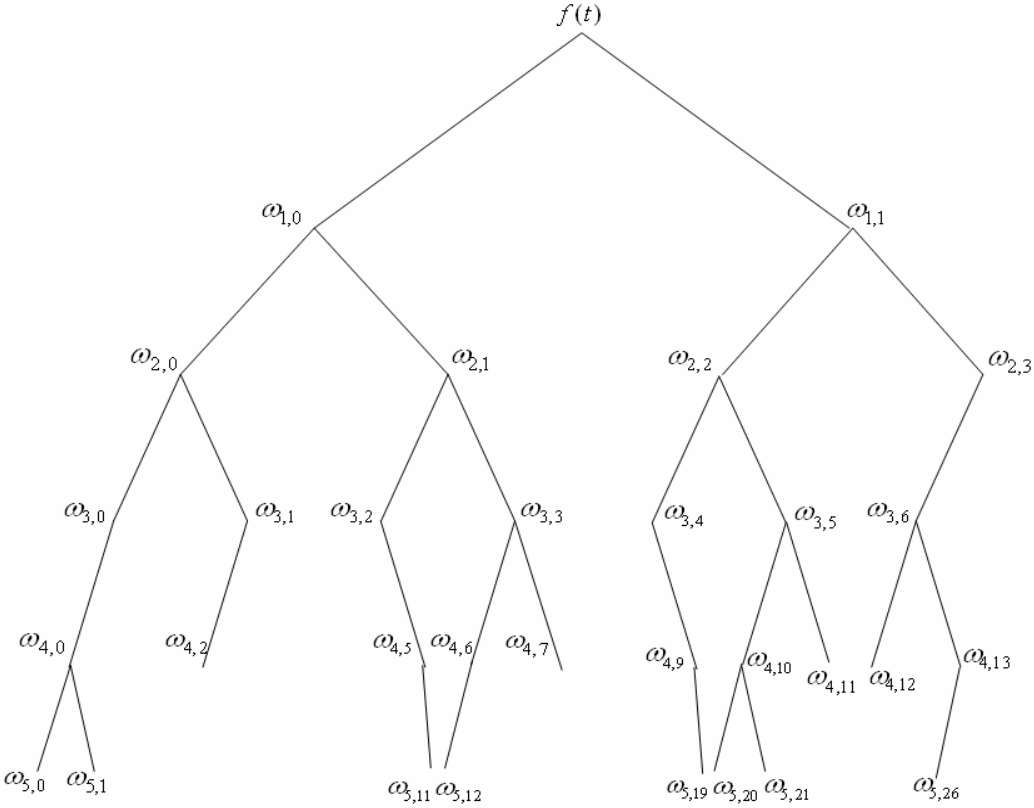


Fig. 2.13 Best decomposition tree structure

D. Thresholding parameters selection

In the denoizing scheme, denoizing is carried out by first removing the white-noise corrupted nodes from the original WPD tree. Further denoizing is carried out by applying thresholding to the decomposition coefficients of each retained node in the best tree. Note that the energy of white noise presented in the measured signal will be spread out evenly among all coefficients, resulting in small decomposition coefficients. On the other hand, the energy of the underlying PD signal will be compacted into a small number of large decomposition coefficients. Based on this idea, either the soft or hard thresholding [46-47] can be used to suppress the noise further.

Hard thresholding removes all decomposition coefficients, which are below a certain threshold value. In addition to hard thresholding, soft thresholding shrinks all remaining coefficients according to some linear law.

Fig. 2.14 shows results from soft and hard thresholding the decomposition coefficients of node (4,7) of the best decomposition tree. Fig. 2.14(a) shows coefficients before thresholding. The large coefficients in Fig. 2.14(a) represent PD components whereas the remaining coefficients represent the white noise. Figs. 2.14(b) & (c) show the processing results of soft and hard thresholding respectively.

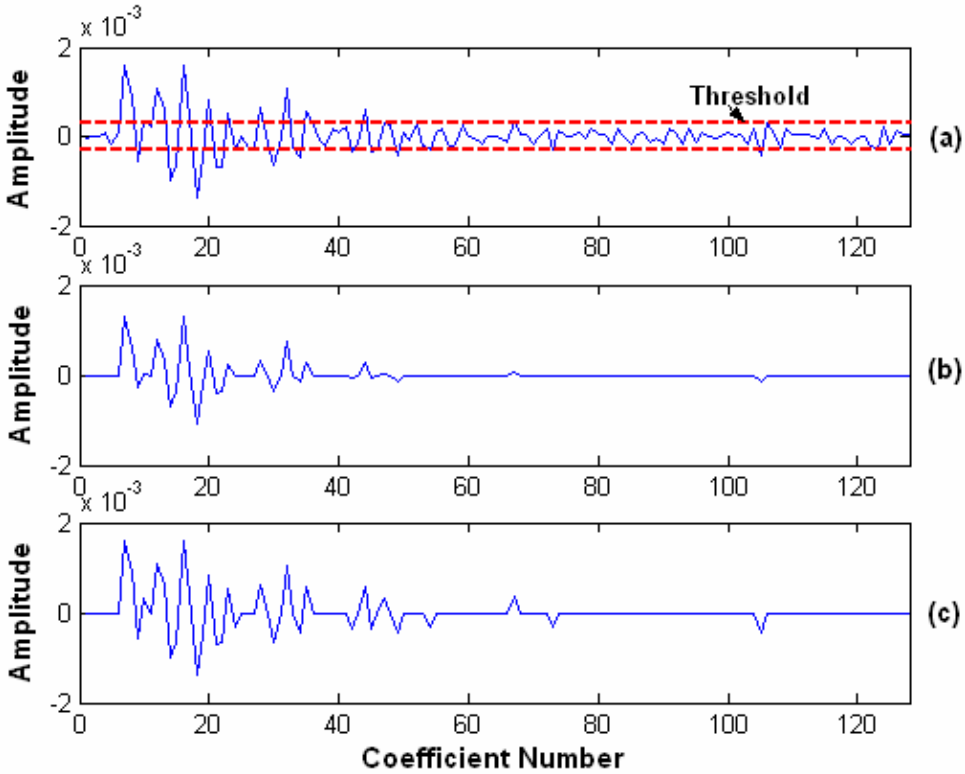


Fig. 2.14 Coefficients thresholding (a) original decomposition coefficients at node (4,7); (b) after soft thresholding; (c) after hard thresholding.

In the present application, determination of the threshold is a crucial issue. Algorithms for calculating the threshold include Stein's unbiased risk estimate, fixed form threshold, minmax criterion and a mixed selection rule [48]. The chosen selection rule is a mixture of the first two algorithms, namely Stein's unbiased risk estimate and fixed form threshold. The noise level of the signal is first estimated. If the SNR is small, fixed form threshold is employed as Stein's unbiased risk estimate is not effective in such cases. Otherwise, Stein's unbiased risk estimate is used to calculate the threshold. The mixed selection rule is adopted here due to its proven suitability for signals with different SNRs [48].

2.3.3 Denoizing of PD Signals

A. Signal decomposition and coefficients thresholding

After the parameters are set, the PD signals are first decomposed using the selected wavelet filters and best tree structure. Starting from the original signal (topmost node), the decomposition is performed by high-pass or low-pass filtering followed by downsampling process as shown in Fig. 2.15. According to the best tree structure, this process is repeated for other nodes in the best tree from top to bottom.

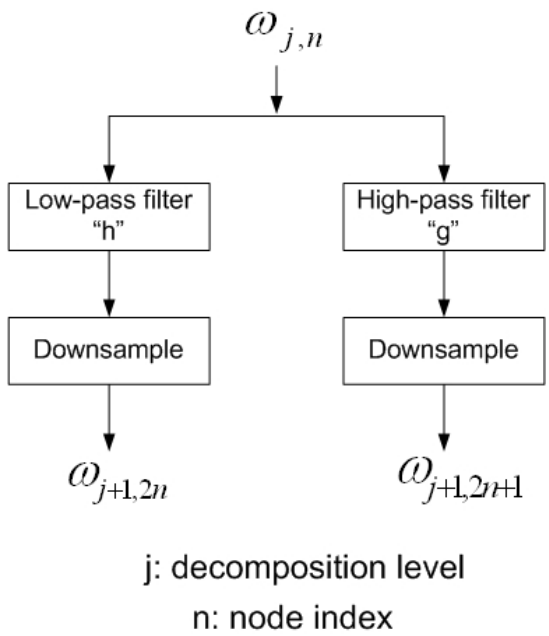


Fig. 2.15 One-step decomposition

The decomposition coefficients are then processed by thresholding using parameters determined in Section 2.3.2 (D). As illustrated in Fig. 2.14, the coefficients are processed by either soft or hard thresholding using the threshold that is calculated based on the determined threshold calculation rule.

B. Wavelet packet reconstruction

After thresholding, the decomposition coefficients of the terminal nodes in the best tree are used to reconstruct the denoized signal. As illustrated in Fig. 2.16, reconstruction is the inverse process of decomposition. It starts from the terminal nodes and ends in the topmost node (denoized signal). The algorithm of reconstruction is given by:

$$\omega_{j,n}(k) = \sum_m H(m-2k)\omega_{j+1,2n}(m) + \sum_m G(m-2k)\omega_{j+1,2n+1}(m) \quad (2.4)$$

where H,G are reconstruction filters and $\omega_{j,n}(k)$ is the kth coefficient at node (j,n). The denoized signal is the sum of all the components reconstructed from the terminal nodes in the best tree.

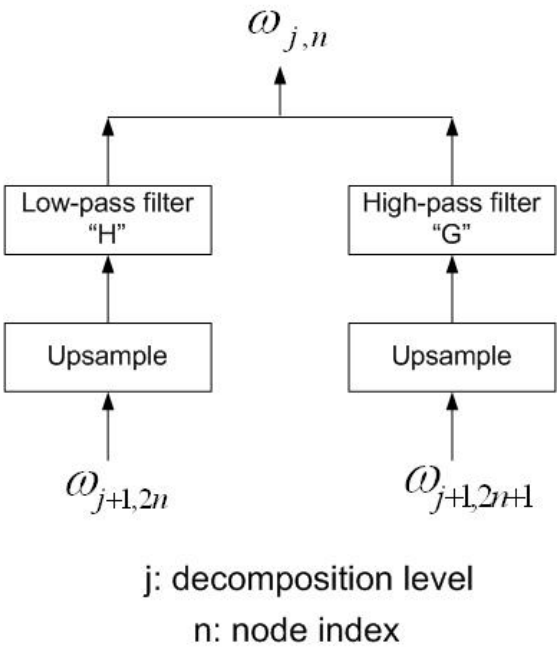


Fig. 2.16 One-step reconstruction

C. Performance testing

After the denoized signal is reconstructed, denoizing performance is assessed. If the performance on training set is satisfactory and the assessment on test set is better than or close to the average performance on the training set, the parameters determined in Section 2.3.2 are accepted. Bad performance is probably due to:

(1) Signals in training set are not able to cover the variety of the PD waveforms. Therefore, more PD signals have to be measured under the same condition as

the under-performed signals and used to extend the training set.

(2) Denoizing parameters are selected individually. Therefore, there is no guarantee of optimal selection of the complete set of parameters. To solve this problem, a method optimising the entire set of parameters is developed in Chapter 4.

2.4 RESULTS AND DISCUSSIONS

Results obtained from various choices of denoizing parameters are presented and discussed in this section. The signal-to-noise-ratio (SNR) and correlation coefficient (CC) as in equations (2.5) & (2.6) are employed to evaluate the denoizing performance.

$$SNR = 10 * \log_{10}\left(\frac{Energy(R)}{Energy(R-Y)}\right) \tag{2.5}$$

$$CC = \frac{\sum_{i=0}^{N-1}(Y(i)-\overline{Y})(R(i)-\overline{R})}{\sqrt{\sum_{i=0}^{N-1}(Y(i)-\overline{Y})^2\sum_{i=0}^{N-1}(R(i)-\overline{R})^2}} \tag{2.6}$$

where Y and R denote the denoized and original PD signals respectively. $\overline{Y}$ and $\overline{R}$ denote the mean values of Y and R respectively.

Due to the limitation of space, only denoizing results of PD signals resulted from free particle are shown in this section. Similar results are obtained for other types of PD. Fig. 2.17 shows a typical noise-free PD signal (free particle) obtained with noise control in a shielded laboratory. To verify the effectiveness of the proposed method, signals of various SNR are generated by superimposing artificial white noises of different levels on the noise-free signal. As the noise-free signal and noise content are known in advance, SNR and CC can be calculated accurately. Apart from the generated signals, results obtained from measurement without noise control are also presented in Section 2.4.4.

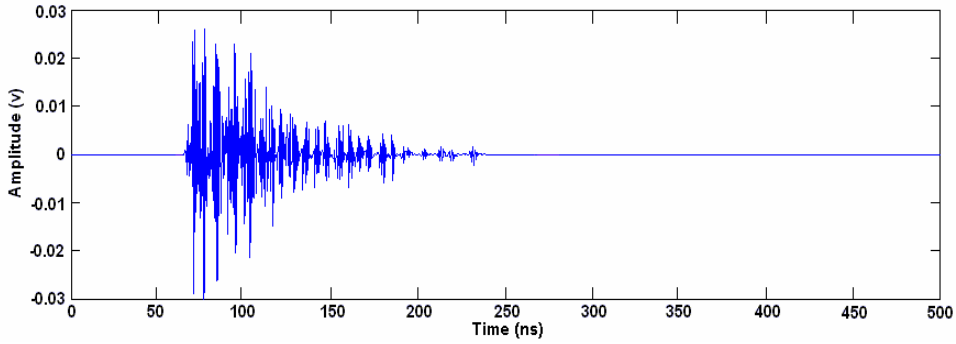


Fig. 2.17 Original PD signal

2.4.1 Wavelet and Decomposition Level Selection

To verify the effectiveness of the wavelet selection method described in Section 2.3.2 (A), performance of candidate wavelets is compared in Table 2.1 for a PD signal having SNR of 0dB. The*'sym8'* wavelet is seen to achieve the largest SNR and CC after denoizing, which confirms the effectiveness of the wavelet selection method described in Section 2.3.2 (A).

Table 2.1 Impact of wavelet filters on SNR and Correlation Coefficient

Wavelet	SNR after denoizing (dB)	Correlation Coefficient
‘db2’	15.2	0.86
‘db4’	15.7	0.86
‘db6’	16.8	0.90
‘db8’	17.5	0.92
‘db10’	16.9	0.90
‘sym2’	15.6	0.88
‘sym4’	16.0	0.90
‘sym6’	17.6	0.92
‘sym8’	**18.3**	**0.96**
‘sym10’	17.9	0.94
‘coif2’	16.2	0.89
‘coif3’	16.4	0.90
‘coif4’	15.8	0.87
‘coif5’	16.0	0.88

Figs. 2.18 & 2.19 show the impact of decomposition level on the denoizing performance. Both SNR and CC after denoizing hardly increase when the decomposition level gets beyond 5. Similar results are obtained for PD signals having different SNRs.

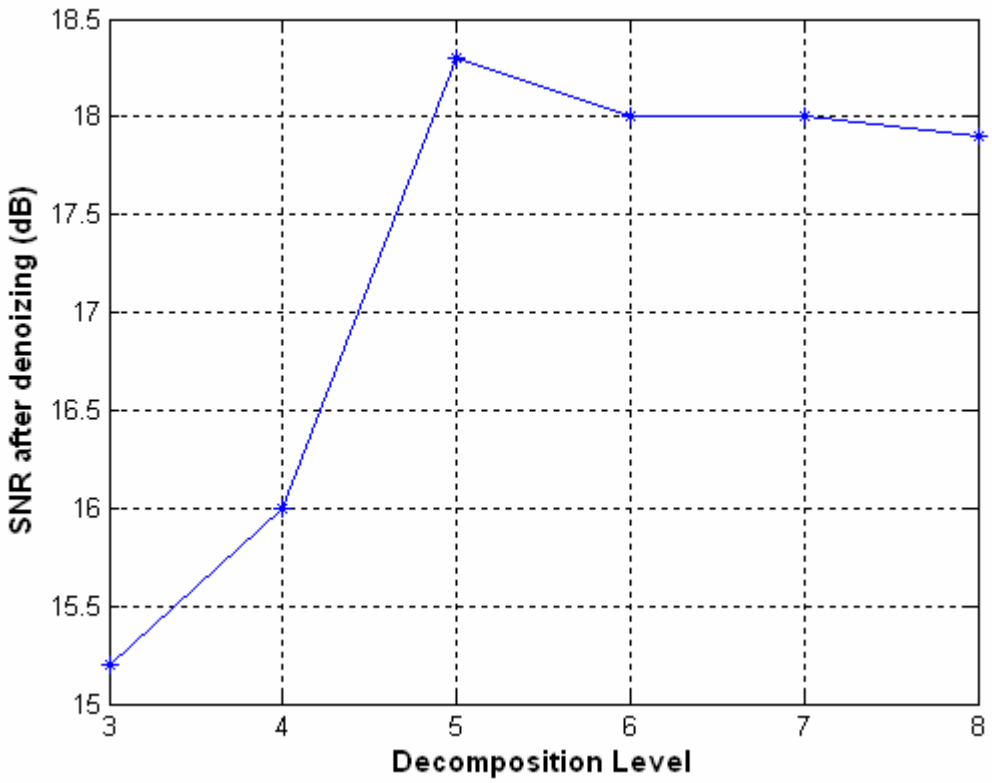


Fig. 2.18 Impact of decomposition level on SNR

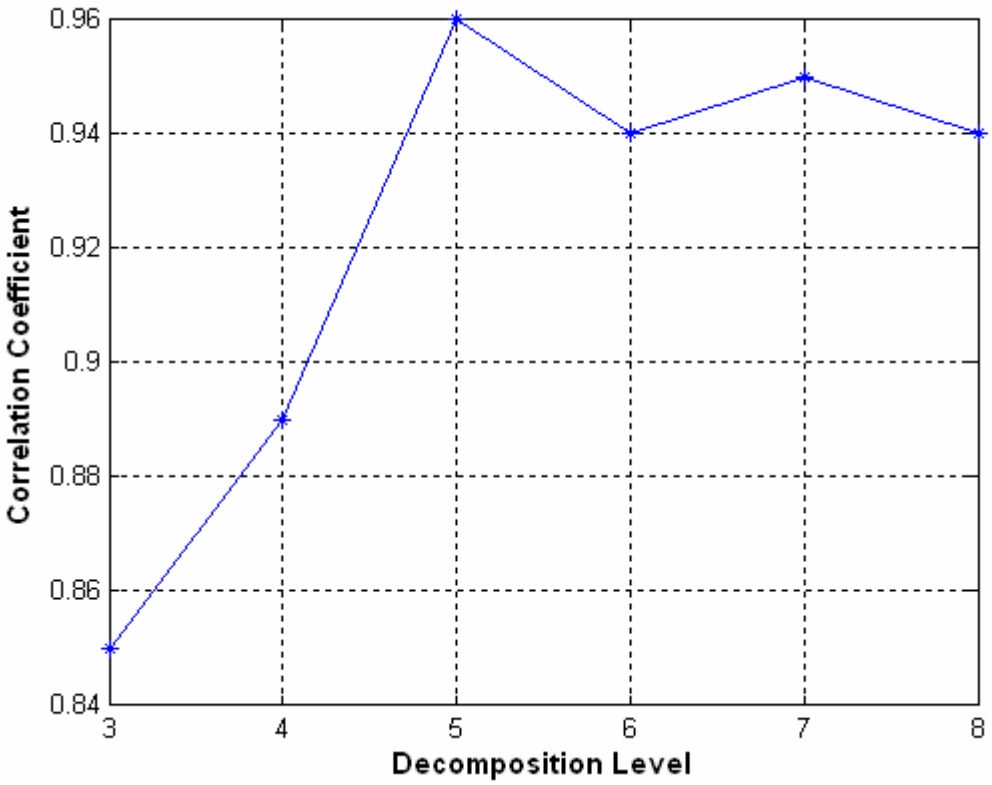


Fig. 2.19 Impact of decomposition level on Correlation Coefficient

2.4.2 Best Tree Selection

Three methods for forming the decomposition tree structure are compared, namely: the DWT-based method, the standard entropy-based-WPT method (Ent-WPT) (Section 2.2.1) and the proposed variance-based-WPT method (Var-WPT). In Figs. 2.20-2.22, PD signals having different noise levels are studied. As shown in Fig. 2.17, PD occurs solely between 65 ns and 230 ns. In all cases, wavelet-packet-based methods lead to tree structures, which better perform than that from the wavelet-transform-based method due to the higher frequency resolution in high-frequency subbands. Among the wavelet-packet-based methods, the tree structure formed by the Var-WPT method is seen to remove the noise more effectively than that from the Ent-WPT method for all three noise levels. Even in the most severe case where the noise energy is ten times PD energy, the Var-WPT method effectively suppresses the noise and restores the original PD signal. Although the DWT-based method and Ent-WPT method are effective to some extent, their performance is much inferior as in Table 2.2. The Var-WPT method leads to the largest SNR and CC after denoizing for all three noise levels. This shows that the Var-WPT method outperforms the other two methods on both noise reduction and PD signal restoration.

The Var-WPT method is seen to increase the SNR values of all PD signals to a very narrow range after denoizing. Similar observation is made on the CC values. These results suggest that the performance of Var-WPT method is robust for PD signals of different noise levels.

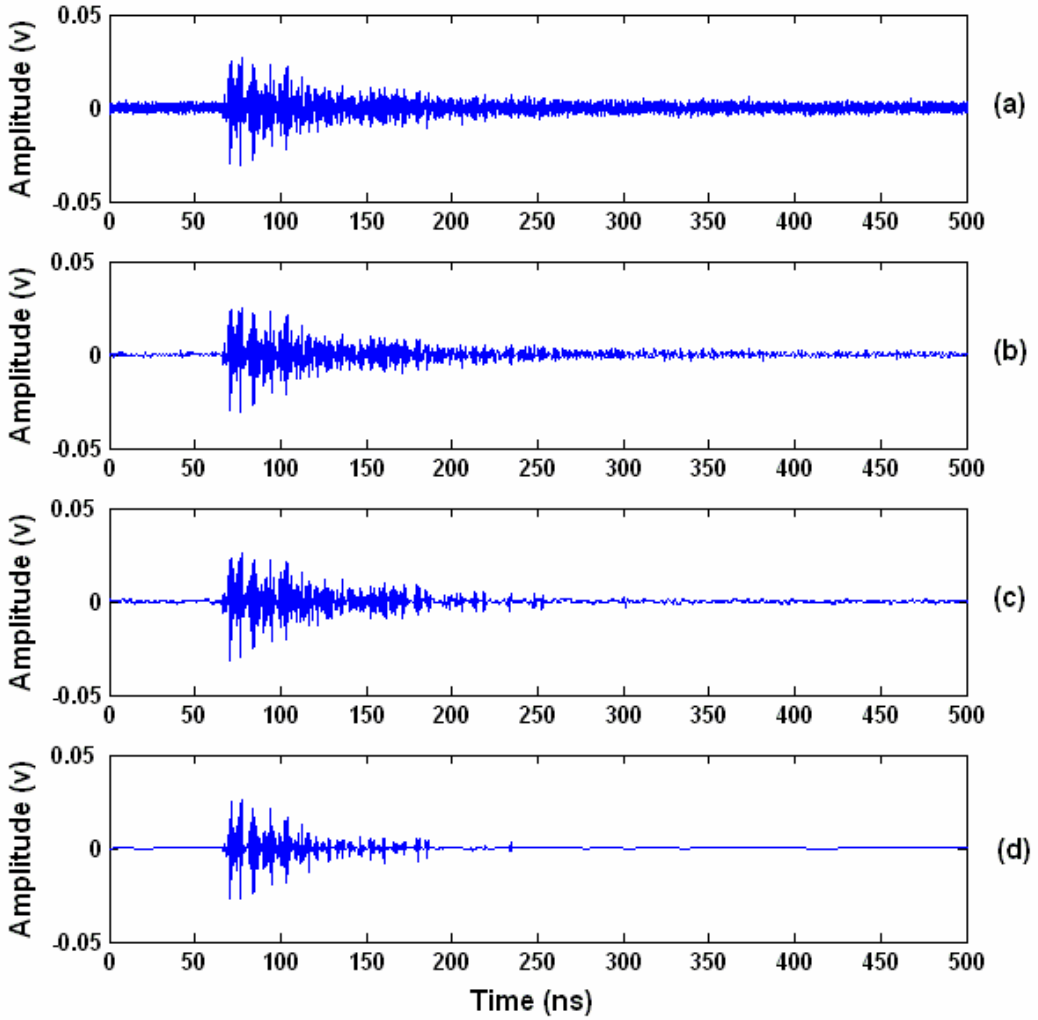


Fig. 2.20 A comparison of the denoizing performance for PD signal with SNR=10 dB. (a) Noisy signal; (b) result of DWT-based method; (c) result of Ent-WPT method; (d) result of Var-WPT method

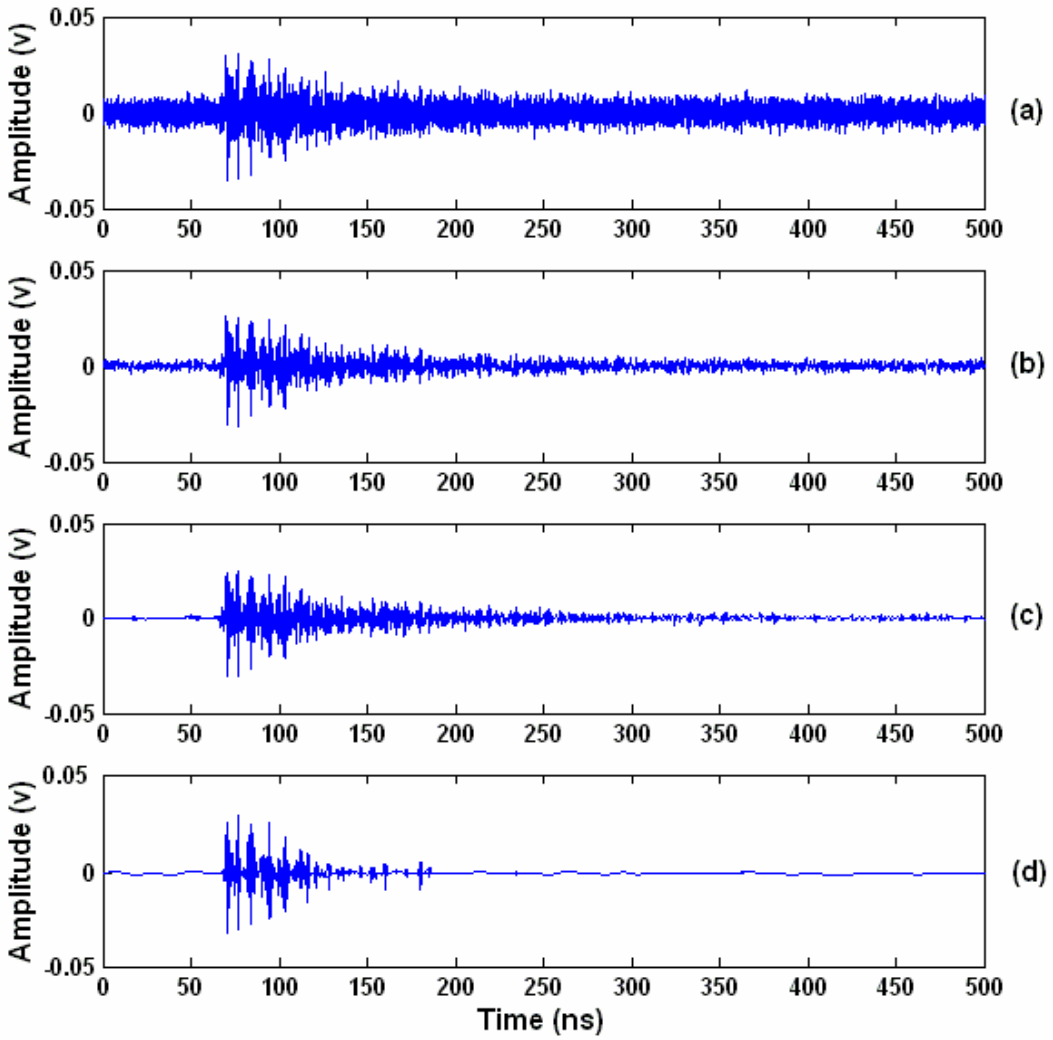


Fig. 2.21 A comparison of the denoizing performance for PD signal with SNR=0 dB (a) Noisy signal; (b) result of DWT-based method; (c) result of Ent-WPT method; (d) result of Var-WPT method

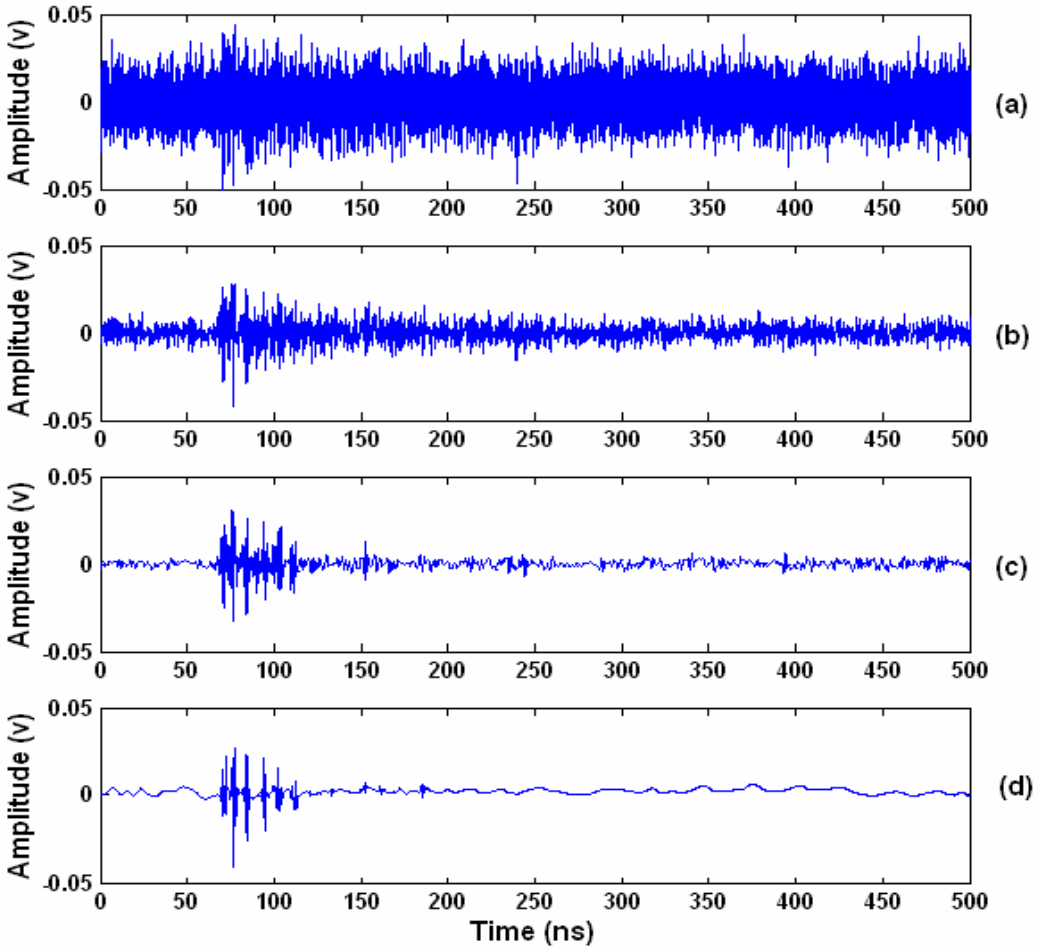


Fig. 2.22 A comparison of the denoizing performance for PD signal with SNR= -10 dB (a) Noisy signal; (b) result of DWT-based method; (c) result of Ent-WPT method; (d) result of Var-WPT method

Table 2.2 Comparison of SNR and CC values of different methods

SNR of Noisy PD Signals	Denoizing Approach	SNR of Denoized PD Signals (dB)	Correlation Coefficient
SNR = 10 dB	DWT	15.2	0.86
	Ent-WPT	15.6	0.88
	Var-WPT	19.0	0.98
SNR = 0 dB	DWT	9.8	0.82
	Ent-WPT	12.5	0.87
	Var-WPT	18.3	0.96
SNR = -10 dB	DWT	2.0	0.69
	Ent-WPT	8.8	0.84
	Var-WPT	17.5	0.93

2.4.3 Thresholding Parameters Selection

Impact of the threshold calculation rule (Section 2.3.2 (D)) is illustrated in Table 2.3. Both the SNR and CC after denoizing take high values from the use of the mixed selection rule, beyond those from other methods. Thus, the effectiveness of mixed selection rule to determine the threshold value is verified.

Table 2.3 Impact of threshold calculation rule on SNR and Correlation Coefficient

Algorithm	SNR of noisy PD signal (dB)	SNR after denoizing (dB)	Correlation Coefficient
1	-5	8.5	0.84
	0	11.4	0.86
	5	15.0	0.90
2	-5	13.9	0.89
	0	13.3	0.89
	5	13.8	0.88
3	-5	4.2	0.78
	0	12.1	0.86
	5	19.7	0.97
4	-5	17.6	0.94
	0	18.3	0.96
	5	20.2	0.98

1: Stein's unbiased risk estimate; 2: fixed form threshold; 3: minimax criterion; 4: mixed selection rule

Performances of the soft and hard thresholding are compared in Fig. 2.23. Fig. 2.23(a) shows a noisy PD signal. Figs. 2.23(b) and (c) show the denoizing results by applying *soft* and *hard* thresholding respectively. The correlation coefficients resulted from *soft*

and *hard* thresholding are 0.86 and 0.93 respectively, which indicate the effectiveness of the latter method over that of the former. The better performance of the hard thresholding is also confirmed by the observation of Figs. 2.23(b) and (c), which is seen to result in less distortion than soft thresholding. Hence, hard thresholding is used in all studies.

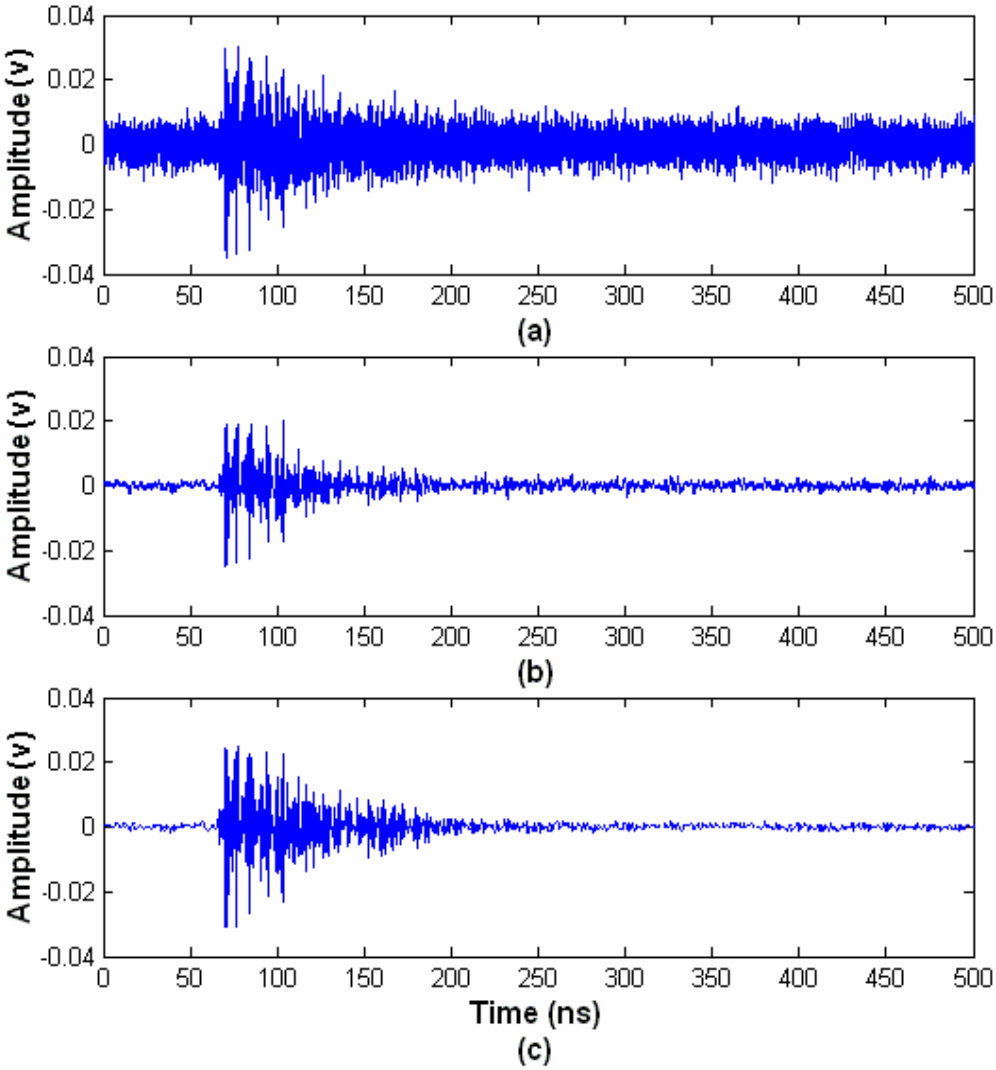


Fig. 2.23 Denoizing results of soft and hard thresholding
(a) Noisy PD signal; (b) result of soft thresholding method; (c) result of hard thresholding method

2.4.4 Performance on PD Signal Measured Without Noise Control in Laboratory

Fig. 2.24 shows denoizing result of a typical PD signal measured without noise control. As observed, the measured signal in Fig. 2.24 (a) exhibits similar waveform to those generated artificially. The Var-WPT method with properly selected parameters is seen to suppress the noises effectively.

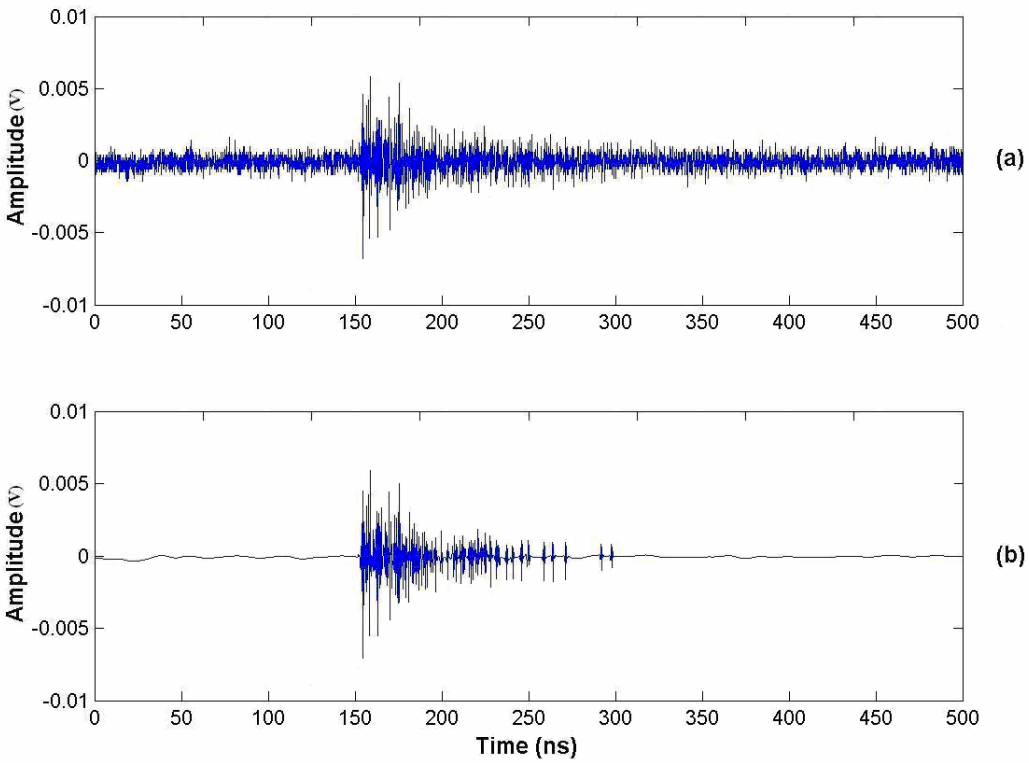


Fig. 2.24 Denoizing result of PD signal measured without noise control
(a) Measured signal; (b) denoized signal

2.5 CONCLUDING REMARKS

Denoizing of PD signals is the first issue to be accomplished during PD detection and diagnosis. In this chapter, a novel variance-based criterion is employed to construct the best tree from wavelet packet tree for PD signals denoizing. Experimental results indicate that the implementation of the Var-WPT method results in successful restoration of PD signals during denoizing with a significant reduction in the noise level. Results show that the proposed method offers better denoizing compared to DWT and WPT with the standard entropy-based criterion. Furthermore, the method is robust for PD signals having various SNR levels and restores weak PD pulses from high noises.

Besides the best tree, selection of other parameters associated with the denoizing scheme is also studied and discussed. However, the parameters are considered separately, which may result in bad overall performance. Thus, optimal selection of a complete set of parameters is further investigated in Chapter 3.

CHAPTER 3

OPTIMAL SELECTION OF PARAMETERS FOR WAVELET-PACKET-BASED DENOIZING

In this chapter, a method based on genetic algorithm (GA) is developed to address the issue of optimal denoizing parameters selection. It begins with a summary of the parameters to be optimized, followed by the construction of fitness function. Subsequently, the GA optimization method is described with detailed discussion on its control parameters. Lastly, numerical results are presented and compared with those obtained in Chapter 2.

3.1 INTRODUCTION

To achieve good denoizing, it is crucial to select the denoizing parameters optimally, such as mother wavelet, decomposition level and thresholding related parameters. Although some denoizing results are presented in [13, 15], there is very little discussion about how to select the optimal parameters. Hence, a general solution of finding the optimal parameters is highly desirable. In [16], the cross-correlation coefficient is used as a criterion for wavelet selection and the estimation of threshold is discussed. However, the parameters are individually considered and the selection of decomposition level is not studied. Moreover, the selection of wavelet is just based on the simulated signals. Therefore, the method proposed in [16] does not guarantee the optimal choice of parameters for denoizing measured PD signals.

In Chapter 2, a method based on *minimum-prominent-decomposition coefficients* is proposed to select the best wavelet. Other parameters are selected based on subsequent assessment of denoizing performance. However, there is no guarantee of optimal selection of the complete set of parameters as they are considered individually rather than holistically. Moreover, considering parameters individually tends to be time-consuming, as the selection process is often not automatic. To overcome these drawbacks, an optimization method is required to automatically optimize the entire set of parameters resulting in the best denoizing performance. Among a few Evolutionary Algorithms, such as Genetic Algorithm (GA), Genetic Programming (GP), Evolution Strategy (ES) and Evolutionary Programming (EP), GA is chosen for this application due to its simple concept and easy implementation. Moreover, GA has been proved to be sufficient for this application by experimental results in Section 3.6.

3.2 DESCRIPTION OF THE PROBLEM

The wavelet-packet-based denoizing scheme as in Fig. 2.1 is used to denoise PD signal. Before denoizing of PD signals, parameters associated with the denoizing scheme must be determined first (blocks A-D of Fig. 2.1). These parameters include wavelet, decomposition level, best tree structure, soft or hard thresholding, threshold estimation rule and threshold processing rule. The last three parameters are required for thresholding (block D). Among the parameters, the construction of best tree structure has been studied and a variance-based method is proposed in Chapter 2. The method is adopted here for constructing the best tree. GA is employed to select the remaining parameters to further improve the denoizing by searching through all possible combination of the parameters.

Table 3.1 shows the parameters to be optimized. Four wavelet families, namely Daubechies wavelets, Symmlet wavelets, Coiflet wavelets, and Biorthogonal wavelets are short-listed for selection due to their proven applicability [42, 45]. Total number of candidate wavelets is thus sixty-four. The decomposition level to be selected is from 1 to 8.

Table 3.1 Parameter ranges

Parameter	Range of Parameter	Subtotal
Wavelet	Daubechies (db) 1-22, Symmlet (sym) 1-22, Coiflet (coif) 1-5, Biorthogonal (bior) 1-15	64
Decomposition Level	1-8	8
Soft or Hard Thresholding	Soft thresholding, hard thresholding	2
Threshold Estimation Rule	Stein's unbiased risk estimate, fixed form threshold, minmax criterion, mixed estimation rule	4
Threshold Processing Rule	No processing, global processing, node dependant processing	3

3.3 DENOIZING PERFORMANCE MEASURE AND FITNESS FUNCTION

To effectively denoise PD signal, the performance of the set of parameters used must be evaluated by some common criteria. The objectives of denoizing are to effectively suppress the noises and restore the original PD signal with little distortion. The signal-to-noise-ratio (SNR) and correlation coefficient (CC) as in equations (2.5) & (2.6) are thus employed to evaluate the performance.

As illustrated in Fig. 3.1, SNR and CC are sometimes conflicting. Their combination is therefore used in the GA fitness function for consistent evaluation of the overall denoizing performance.

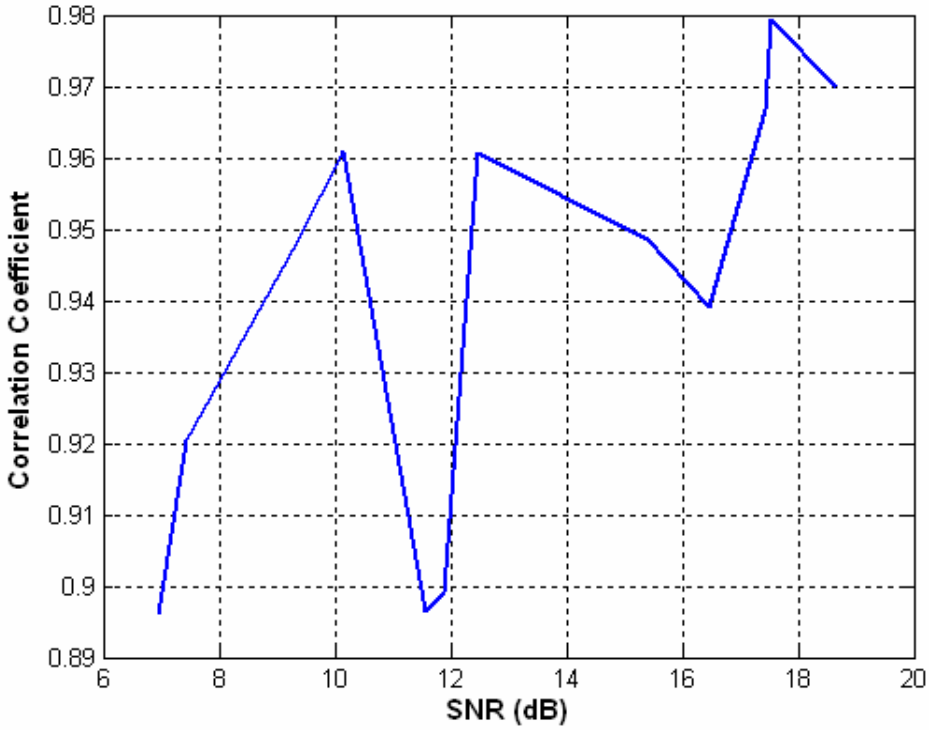


Fig. 3.1 Relation between SNR and CC

The original definition of SNR of equation (2.5) allows negative values to be taken due to the logarithmic computation, which makes it impossible to be used in the GA fitness function. Therefore, another version of SNR (m_SNR) is defined as

$$m\_SNR = \frac{Energy(R)}{Energy(R-Y)}, \tag{3.1}$$

where Y and R denote the denoized and original PD signals respectively. Obviously, the value of m_SNR is always positive. Subsequently, the GA fitness function corresponding to each signal in the training set is defined as the combination of m_SNR and the original CC, which may take various forms such as:

$$g = m\_SNR * CC \tag{3.2}$$

or

$$g = m\_SNR + CC \tag{3.3}$$

However, GA is not able to converge when fitness function in equation (3.2) is used. Therefore, only equation (3.3) is considered as the fitness function. Since the m_SNR usually takes a much larger value (about twenty times) than CC, the fitness values calculated by the above formulas are governed by m_SNR. Therefore, only a high signal-to-noise-ratio is guaranteed by optimizing the fitness function in equation (3.2) or (3.3). The correlation coefficient is however neglected during GA optimization. As a result, the obtained parameters may lead to effective suppression of noise, but large distortion could be observed. To tackle this problem, the fitness function of equation (3.3) is modified as:

$$g = 0.05 * m\_SNR + CC \tag{3.4}$$

where the coefficient of 0.05 is used to set the two components of g in the same range. Considering all signals in the training set, the GA fitness function is finally:

$$fitness = \frac{1}{N}\sum_{i=1}^{N} g(i) \tag{3.5}$$

where N is the number of signals in the training set.

3.4 PARAMETER OPTIMIZATION BY GA

In this section, GA is first reviewed briefly. Subsequently, application of GA in finding the optimal denoizing parameters is investigated, followed by the discussion of GA control parameters selection.

3.4.1 Brief Review of GA

GA is a global search method utilizing the principle of natural selection and genetics. The method starts from a randomly generated population (potential solutions) whose performance is evaluated by a fitness function. Based on the evaluation, a new population is created from the process of reproduction, crossover and mutation. The process is iterated until the stop criteria are met [49]. A comprehensive review of GA theory is given in Appendix C.

As an optimization method, GA has the advantages of flexibility imposed on the search space, easy implementation, fast convergence, and so on. GA has been successfully applied to many fields in electric power engineering [50-52]. Recently, it has also been applied to PD analysis [53-55]. In [53-54], GA is used to optimize the parameters of classifiers for PD pattern recognition. In [55], GA is applied to calculate the optimal parameters of a transformer model.

3.4.2 GA Optimization

For GA optimization, the denoizing parameters shown in Table 3.1 must be represented in binary form. Therefore, they are coded in a string of 14 binary bits as in Fig. 3.2.

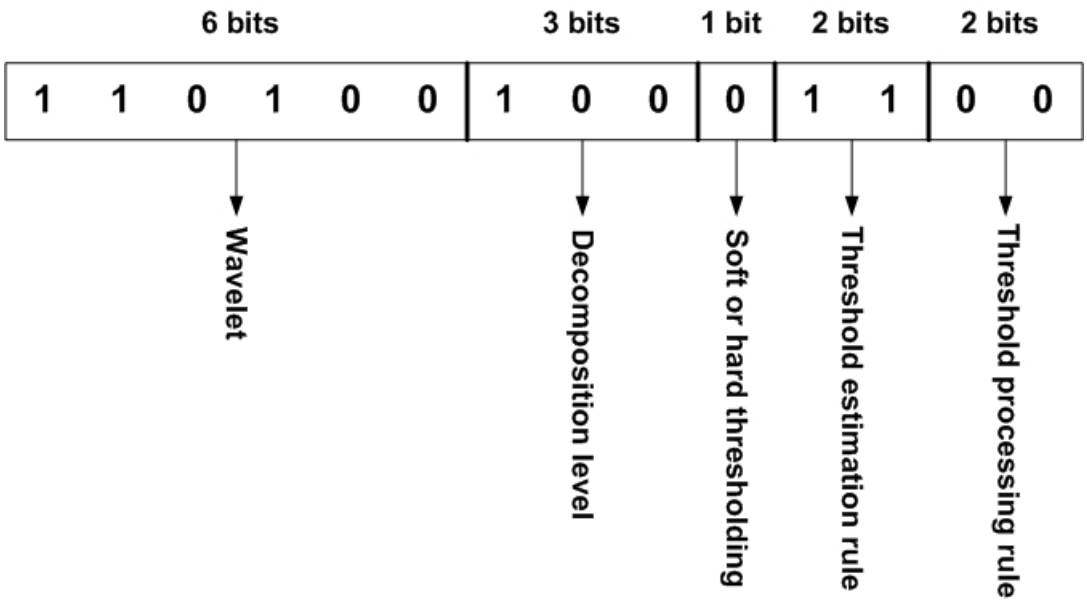


Fig. 3.2 GA coding string

For the implementation of GA, the roulette wheel approach is adopted here in reproduction. The single-point crossover is applied to randomly paired sub-strings with a probability *Pc*. To ensure diversity during evolution, mutation is performed for each bit in the population with a probability *Pm*.

The GA flowchart for denoizing parameters optimization is shown in Fig. 3.3 and a description of the major steps is as follows:

(1) Prepare the training set that is the same as that used in Chapter 2.

(2) Randomly generate an initial population.

(3) Denoise each PD signal of the training set using the parameters determined by each individual of the current population.

(4) Calculate the fitness of each individual on the entire training set by taking the mean of its fitness on each signal and save the best solution.

(5) If the stop criterion is met, use the best solution so far as the optimal one and end the program. Otherwise, continue step (6).

(6) Create intermediate population by copying the individuals of current population in proportion to their fitness.

(7) Apply crossover and mutation to the individuals of the intermediate population to create the next generation, and then go to (3).

3.4.3 Selection of Control Parameters for GA

There are a number of control parameters associated with the application of GA, such as the population size (*Np*), crossover probability (*Pc*) and mutation probability (*Pm*). It is crucial to investigate the influences of these parameters, as they have significant impact on the performance of GA.

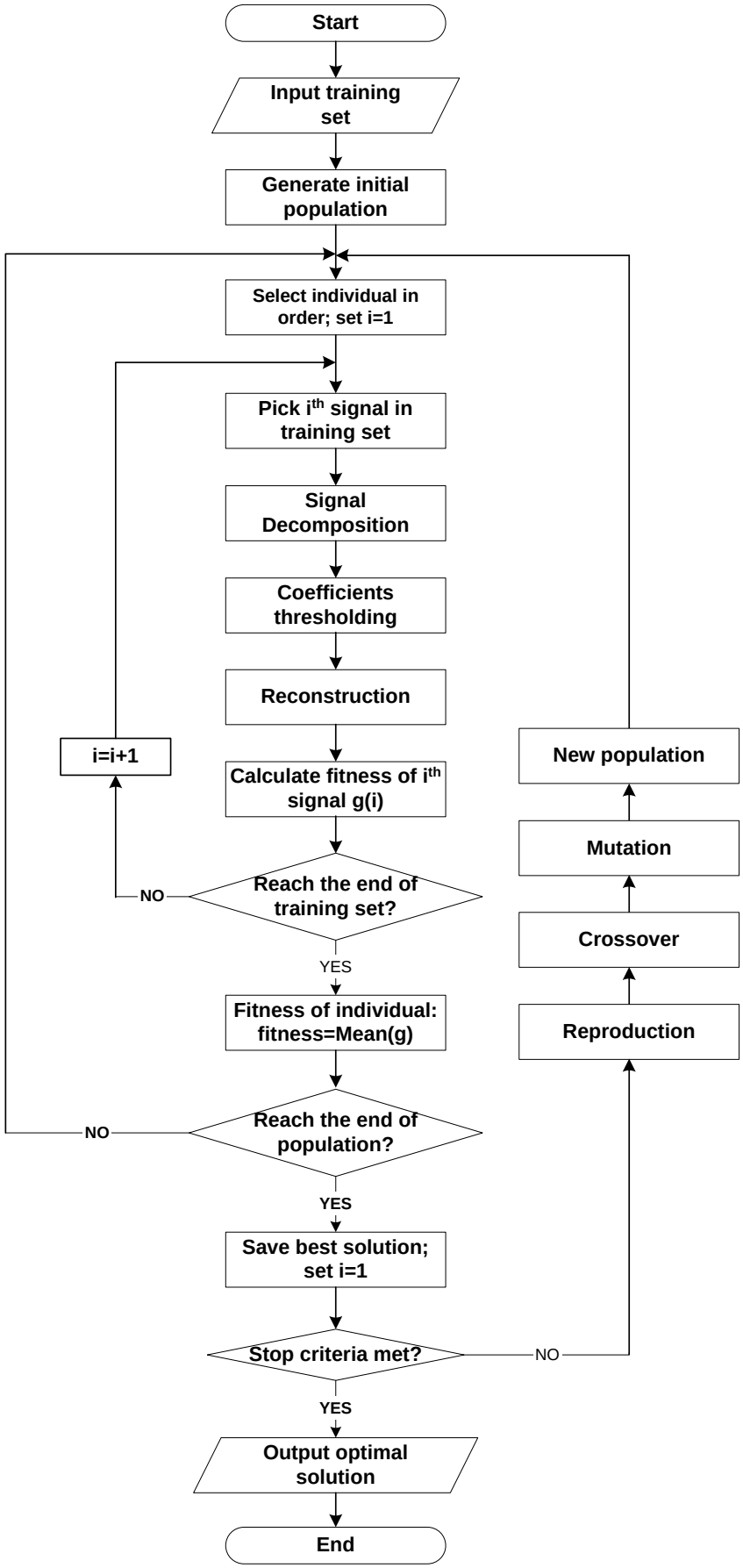


Fig. 3.3 GA flowchart

A. Population size *Np*

The population size of GA defines the number of candidate solutions in each generation. Choosing a suitable population size is a fundamental consideration for GA application. If the size of population is too small, GA may converge prematurely due to the insufficient information given on the searching space. On the other hand, a large population requires more evaluations per generation, which may result in an unacceptably slow rate of convergence. In this study, a relatively small population size (*Np*=8) is employed first. Then, the population size is increased until a consistent solution is found.

Fig. 3.4 shows the performance of GA using population size of 8, 16 and 40. It can be seen that GA converges to a sub-optimal solution when a small population size (*Np*=8) is employed. In the cases of *Np*=16 and *Np*=40, similar performance is achieved, which is better than the case of *Np*=8.

Table 3.2 shows the computation time of GA with various *Np*. As observed, the computation time is proportional to *Np*. Although more iterations are required for the case of *Np*=16 than that of *Np*=40, GA converges faster in the former case, as less evaluations are performed at each iteration. In a word, the population size of 16 leads to a good tradeoff between performance and computation time, and thus is chosen for the optimization task in this study.

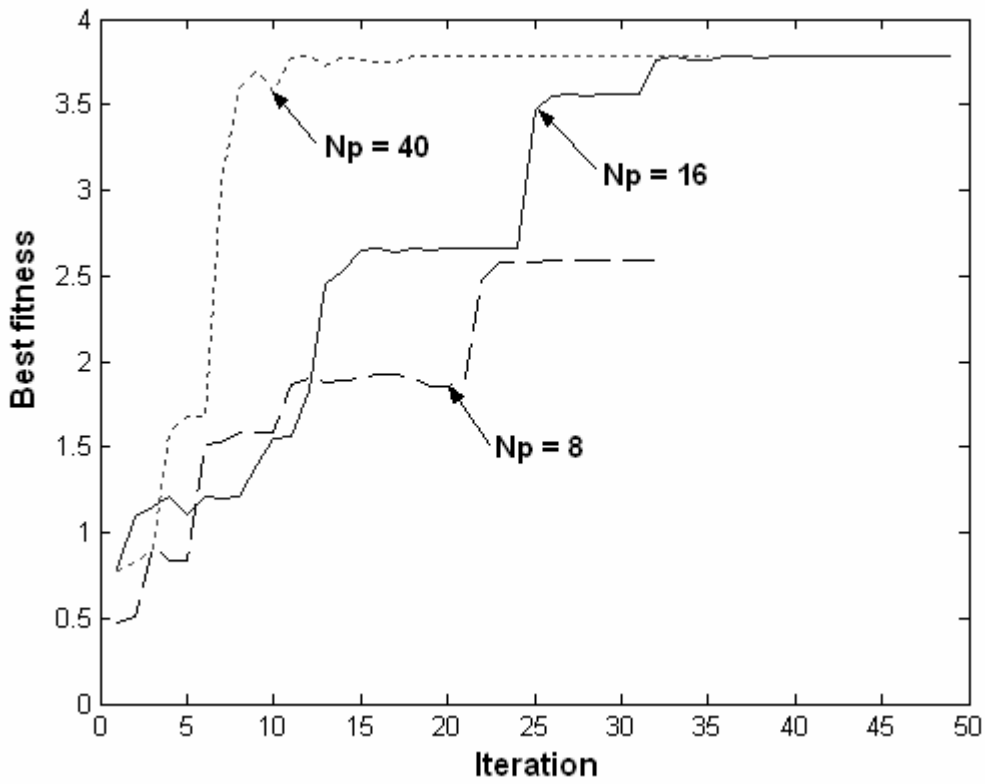


Fig. 3.4 Effect of population size Np

Table 3.2 Computation time of GA with various population sizes

Population size (Np)	Iterations	Computation time (sec)
8	32	102
16	48	306
40	35	675

B. Crossover probability (*Pc*)

The crossover probability controls the frequency with which the crossover operator is applied. The higher the crossover probability, the more quickly new individuals are introduced into the population. If an unnecessary high crossover probability is taken,

the individuals with good performance may be discarded and the improvement of performance may not be achieved. On the contrary, if the crossover probability is too low, the search may stagnate prematurely due to the low exploration rate. Thus, a proper crossover probability must be selected experimentally.

Fig. 3.5 illustrates the effect of using different crossover probability in the GA optimization. It can be seen that GA with *Pc* of 0.75 gives the best performance. In the other two cases, where *Pc* takes 0.95 and 0.55 respectively, GA converges to much lower fitness values. Thus, *Pc* is set to 0.75 for all the subsequent experiments.

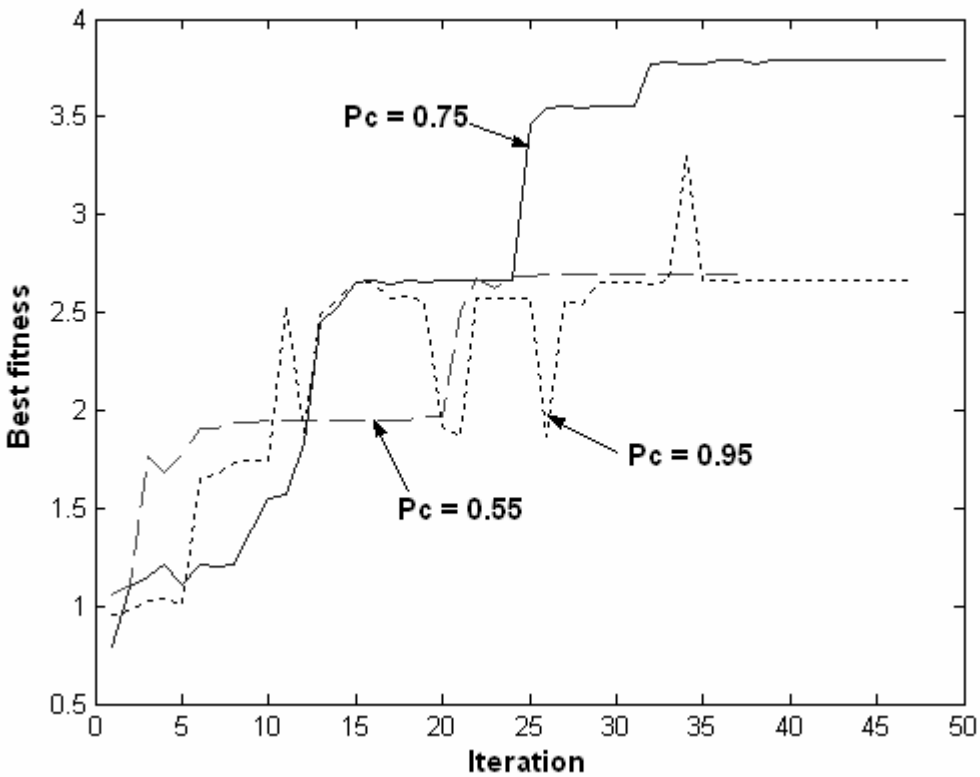


Fig. 3.5 Effect of crossover probability (fixed *Pm* = 0.15, *Np* = 16)

C. Mutation probability (*Pm*)

Mutation is another operator applied to the individuals to create a new generation. It increases the variability of the new generation to prevent GA from stagnating on local extreme. The selection of mutation probability is problem dependent. For many problems, a low mutation rate is suggested, as a high level of mutation could yield an essentially random search [49, 56]. However, a growing number of works indicate that mutation plays a more important role for certain applications and thus a high mutation probability is required [57, 58]. In this thesis, *Pm* is determined by comparative studies.

Fig. 3.6 illustrates the performance of GA with various *Pm*. It is seen that a mutation probability of 0.15 leads to the best performance. Neither a higher Pm (=0.3) or a lower Pm (=0.01) gives satisfactory result. Therefore, Pm=0.15 is chosen for the optimization.

D. Other issues related to GA application

The choice of initial population has impact on GA convergence. GA could converge sub-optimally with bad starting point. Since initial populations are generated randomly, one solution to this problem is to run GA several times to check consistency.

Another issue related to GA optimization is the criteria used to stop the GA program. In this study, two criteria are adopted as follows:

(1) When the maximum number of generations (*Ns*) is reached, the GA program stops. *Ns* is set to 1000 in this study.

(2) GA stops when the best fitness saturates over a number of generations.

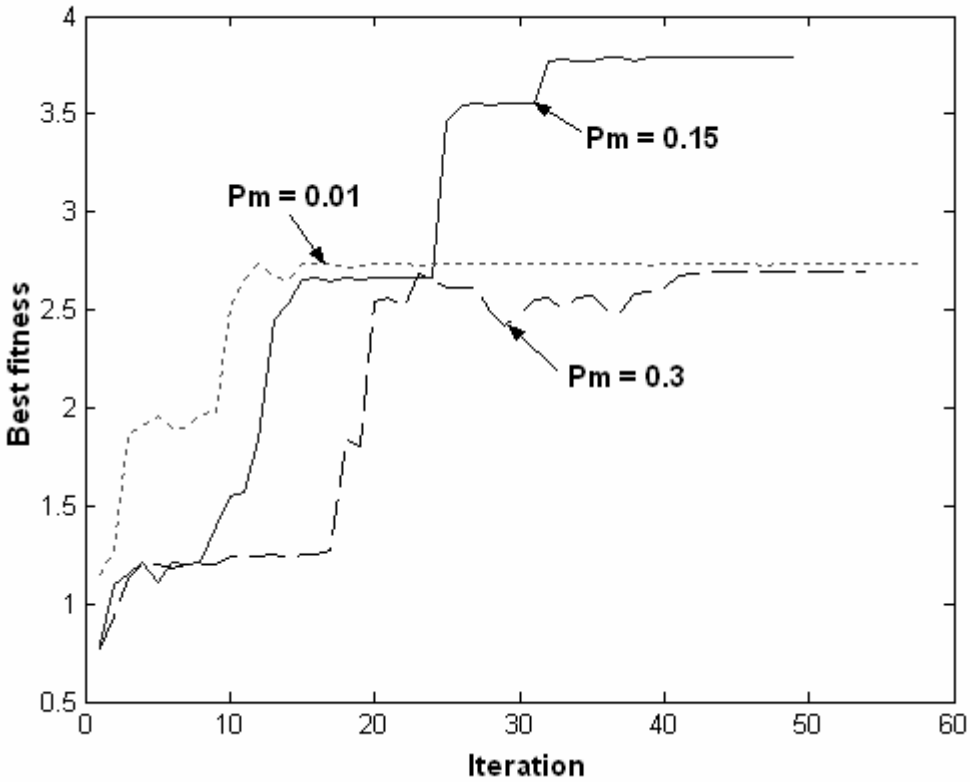


Fig. 3.6 Effect of mutation probability (fixed *Pc* = 0.75, *Np* = 16)

3.5 PERFORMANCE TESTING

After parameters optimization using the training set, the performance of the parameters is assessed on the test set. If the assessment is better than or close to the average performance on the training set, the obtained parameters are accepted. Otherwise, possible reasons for having bad performance are as follows:

(1) The signals in training set are not able to cover the variety of the PD waveforms. Therefore, more PD signals that belong to the same class as the under-performed signals have to be measured and used to extend the training set.

(2) GA could have converged sub-optimally due to badly chosen GA parameters. Therefore, GA parameters have to be adjusted.

After proper measures are taken, GA is executed with the updated parameters and training set (Fig. 3.3).

3.6 RESULTS AND DISCUSSIONS

In this section, results from GA are presented and compared with those obtained from the method presented in Chapter 2. The same training and test set as in Chapter 2 is used here.

Fig. 3.7 shows the convergence of GA and the denoizing performance using intermediate parameters obtained during convergence. GA takes 48 iterations and about five minutes on the Pentium-IV to converge. It improves the denoizing effectively and continuingly during convergence. The choice of the GA fitness function and control parameters is thus verified. As observed, the denoizing performance is improved as the fitness value increases.

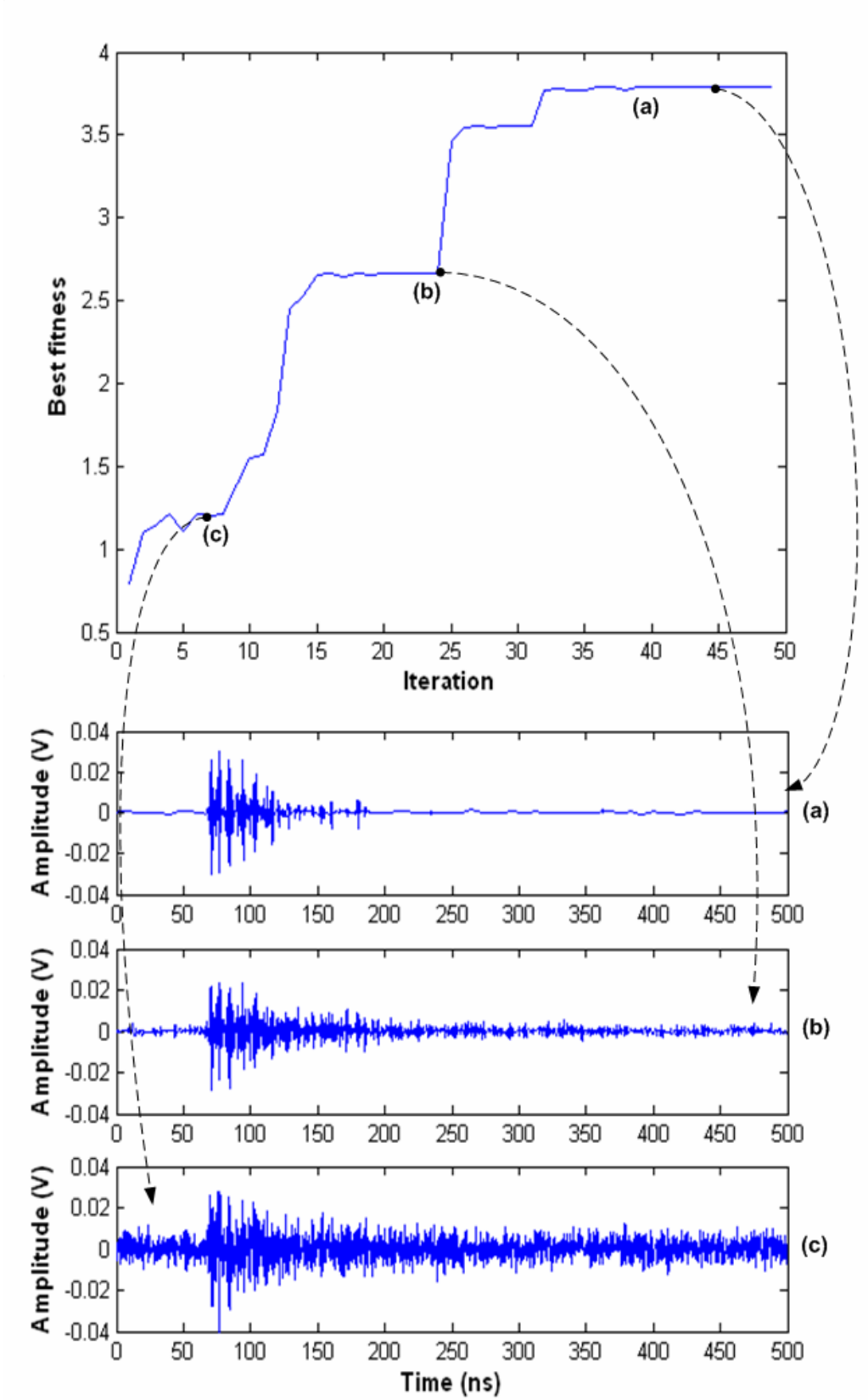


Fig. 3.7 GA convergence and denoizing performance of intermediate parameters

Table 3.3 shows the parameters obtained at intermediate stages of convergence. Stage (a) corresponds to the highest fitness value (convergence), whose parameters are optimal for the given set of training data. Parameters obtained from Chapter 2 with the same training set are shown in Table 3.4. It can be seen that the decomposition level and thresholding method obtained by stage (a) and the method in Chapter 2 are the same while other parameters are different. Stage (a) and the method in Chapter 2 both recommend the same wavelet family (Symmlet), but different members of the family. This indicates that the *minimum-prominent-decomposition coefficients* method as adopted in Chapter 2 is effective although not optimal. In all study cases, the Symmlet family fits the PD signals better than other wavelet families.

Table 3.3 GA intermediate parameters

	fitness	Wavelet	Decomposition level	Soft or hard thresholding	Threshold estimation rule	Threshold processing rule
(a)	3.8	sym6	5	hard	fixed form threshold	node dependant processing
(b)	2.7	coif2	5	soft	mixed estimation rule	node dependant processing
(c)	1.2	db10	8	hard	Stein's unbiased risk estimate	global processing

Table 3.4 Parameters obtained from the method in Chapter 2

Wavelet	Decomposition level	Soft or hard thresholding	Threshold estimation rule	Threshold processing rule
sym8	5	hard	mixed estimation rule	global processing

The GA-based method and the method in Chapter 2 are further compared in Fig. 3.8, with Fig. 3.8 (a) showing the noisy PD signal. Figs. 3.8 (b) & (c) show the denoized signals using parameters obtained by the method in Chapter 2 and GA respectively. As observed, parameters obtained by GA suppress the noise and restore the original PD signal far more effectively. The SNR values correspond to Fig. 3.8 (b) & (c) are 16.7 and 19.1 and CC values are 0.93 and 0.97 respectively. These results confirm the better performance of the parameters obtained by GA. Similar results are obtained from other signals taken from the test and training sets.

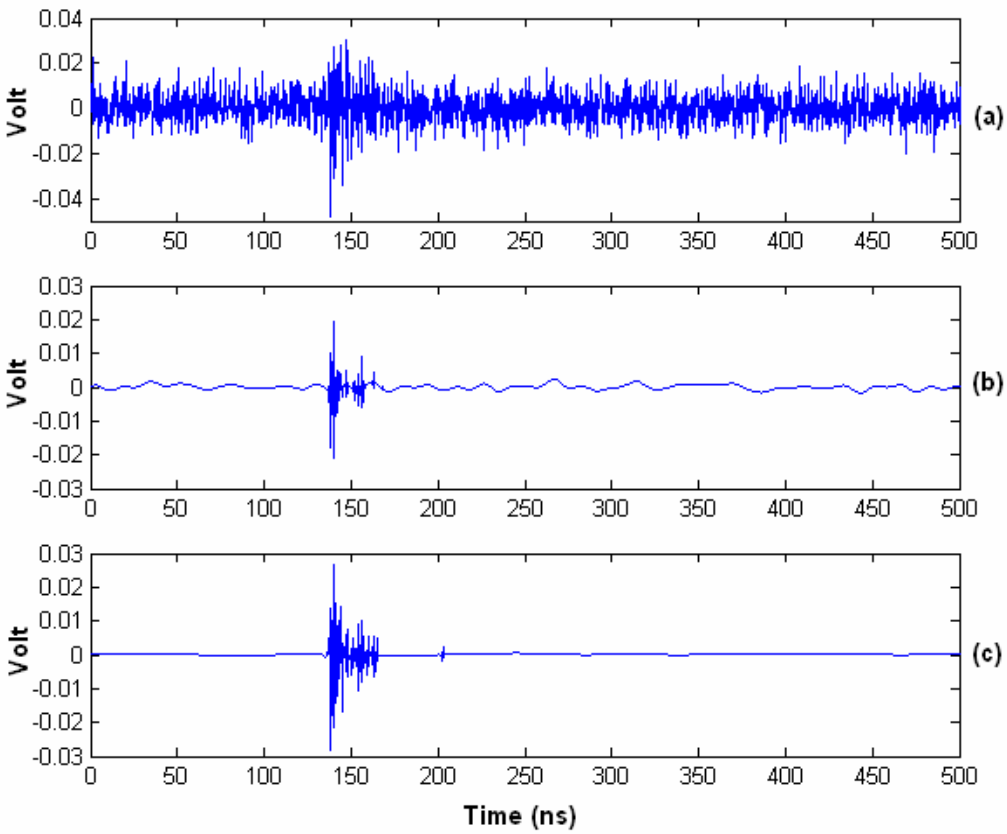


Fig. 3.8 Performance comparison of GA and the method in Chapter 2

3.7 CONCLDING REMARKS

The performance of the denoizing scheme is largely dependent on how the scheme parameters are determined. In this chapter, a GA-based method is developed to optimize the parameters associated with the wavelet-packet-based denoizing scheme. Numerical results indicate that the GA-based method ensures optimal denoizing in terms of successful restoration of the original PD signal with significant reduction in the noise level. The method enables automatic and fast determination of parameters. Denoized signals can then be used to develop a reliable diagnosis system for recognizing corona and SF_6 PD resulted from various defects.

CHAPTER 4

PD FEATURE EXTRACTON BY INDEPENDENT COMPONENT ANALYSIS

This chapter explores the application of Independent Component Analysis (ICA) in PD feature extraction. To ensure reliability of the extracted features, a process known as pre-selection is first introduced. Secondly, Independent Component Analysis is reviewed through a comparison with the well-known Principal Component Analysis. Subsequently, ICA-based feature extraction method is described with discussions on the selection of parameters for implementing ICA. Lastly, numerical results are presented and discussed.

4.1 INTRODUCTION

For condition monitoring of GIS, it is crucial to recognize the source of the harmful PD activities in SF_6 and the unharmful air corona in a fast and reliable manner. The key component of such a PD diagnosis system is to extract the most effective and reliable PD features from the measured raw data, so that satisfactory performance can be achieved in the subsequent classification task. Fig 4.1 illustrates various methods for extracting PD features. As reviewed in Chapter 1, the traditional PRPD and POW approaches have noticeable limitations in terms of speed and classification performance. Therefore, methods using UHF signals measured within hundreds of nanoseconds are developed for PD identification in this study. In this chapter, time-domain techniques namely independent component analysis (ICA) and principal component analysis (PCA) are employed to perform the feature extraction. In Chapter 5, a wavelet-packet-based method is proposed for extracting the most discriminating features from time-frequency domain. Using the features extracted by ICA- or wavelet-packet-based method, a neural network is trained and tested in Chapter 6 for classifying a new set of measured data. Data measured one metre away from PD source as in Table A.1 are employed in Chapters 4,5 and 6 for developing the PD identification system. The robustness of extracted PD features on data measured from other PD-to-sensor distances is investigated in Chapter 7, where a re-selection and re-training scheme is proposed.

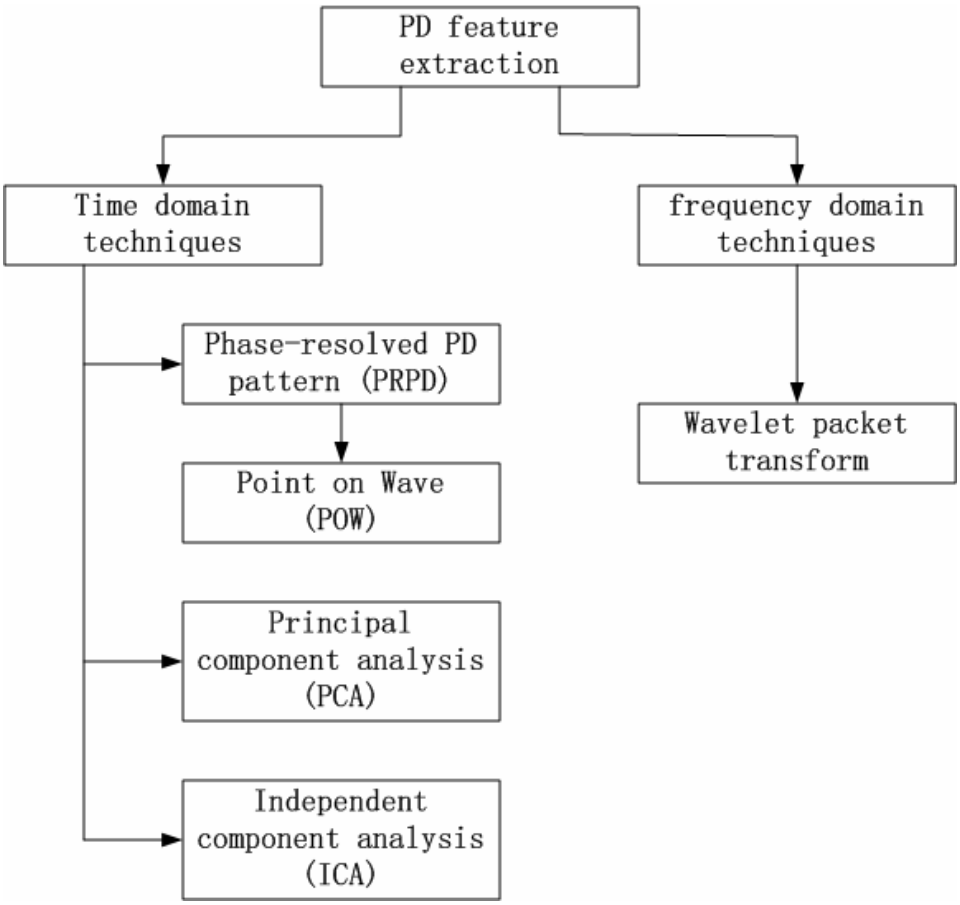


Fig. 4.1 Methods for extracting PD features

The ICA-based PD feature extraction is illustrated in Fig. 4.2. In the current study, the original waveforms of UHF signals are crucial for source recognition, as the feature extraction and classification are based on the time-domain signals only. However, due to the excessive white noises, the original waveforms are often distorted or even buried under the noise. In Chapters 2 and 3, the problem of white noise has been successfully tackled by applying the wavelet packet denoizing on each measured waveform as shown in Fig. 4.2, which makes the subsequent recognition of PD source an easier task to be accomplished.

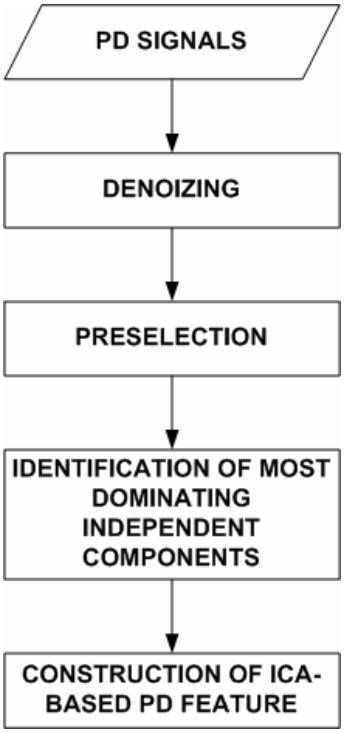


Fig. 4.2 Flowchart of ICA-based PD feature extraction

Air corona is often regarded as another form of noise in PD monitoring system of GIS. Since corona signal is very similar to SF_6 PD signal, it often leads to misclassification, which may result in wrong decision. Therefore, it is of great importance to correctly classify PD and corona. To reduce the response time of the PD diagnosis system, source recognition of SF_6 PD and the discrimination of corona and SF_6 PD are considered together in this study, so that no second judgment is needed. **In the following text, "PD identification" refers to classification of all types of SF_6 PD as well as air corona, except specified.**

Another issue related to the waveform-based PD identification, as illustrated in Fig. 4.3, is the *time shift* of PD signal. Figs. 4.3 (a) and (b) show two sections of the measured PD signal. They are captured by two windows with the same length but a shift in time. In practice, the time shift between measured signals is caused by changes in noise

levels during measurement or setting of the oscilloscope. Since the statistical measures used in this study such as negentropy, kurtosis and skewness are subject to time translation, the values of these measures are different for signals in Figs. 4.3 (a) and (b). Such a difference may cause difficulties in extracting PD features and the subsequent classification task. Hence, a process known as pre-selection (Fig. 4.2) is employed to cancel the "time shift" effect by capturing a segment with a predetermined length starting from the initial surge of the signal. The process thus ensures the signals to have the same set of features upon a signal pattern with all possible time shifts. Details of the pre-selection process are given in Section 4.2.

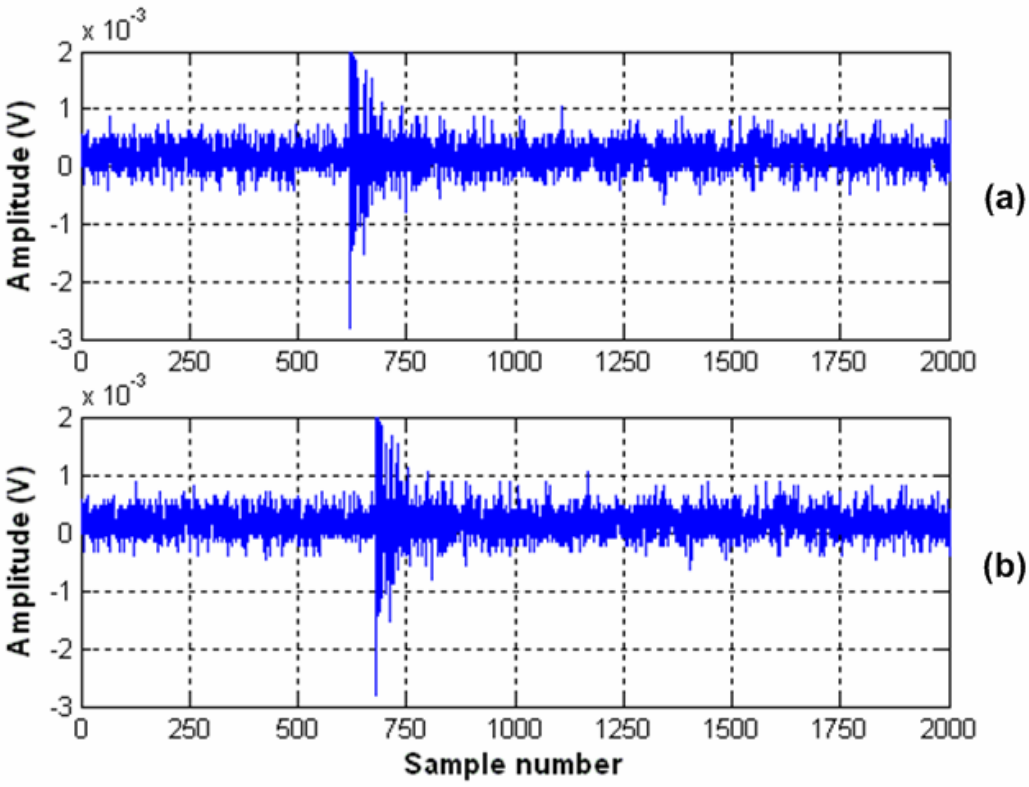


Fig. 4.3 Signal shift in time

After denoizing and pre-selection, the PD identification task is performed in two steps, namely feature extraction and classification. Each set of pre-selected signals has

typically a length of 1000. It is highly desirable to compress the pre-selected signal to a smaller working set (features) in order to improve the efficiency of PD identification without sacrificing much of the discriminating power of the original signal. In this chapter, a time-domain technique known as Independent Component Analysis (ICA) is employed to perform the data compression as shown in Fig. 4.2. The compressed data set, known as the *ICA_feature*, is formed by projecting the pre-selected signal onto the directions of independent components. Using the compressed working set, classification of PD is carried out by a neural network (Chapter 6). Denoizing of PD signals has been studied in previous chapters. Salient features of the other blocks in Fig. 4.2 are discussed in the following sections.

4.2 PRE-SELECTION

To perform the pre-selection, a threshold determined by background noise level is employed to detect the starting point of PD event (big oscillation) as shown in Fig. 4.4. Since most of the white noises have been removed during the process of denoizing as shown in Fig. 4.4(b), it is feasible to detect the starting point by applying a fixed threshold (=0.5 mV). The length of the pre-selected signal is set to 1000 points to capture the entire waveform of PD event. Fig. 4.5(b) shows a typical pre-selected UHF signal that is used in the following feature extraction process.

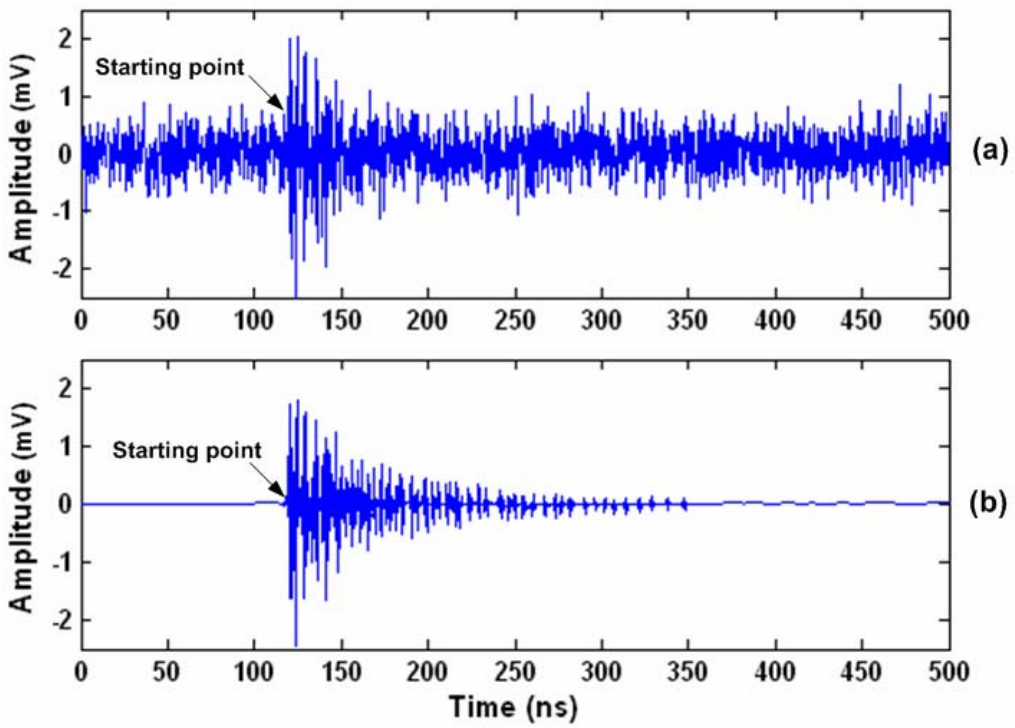


Fig. 4.4 Detecting the starting point of PD event (a) measured signal; (b) denoized signal

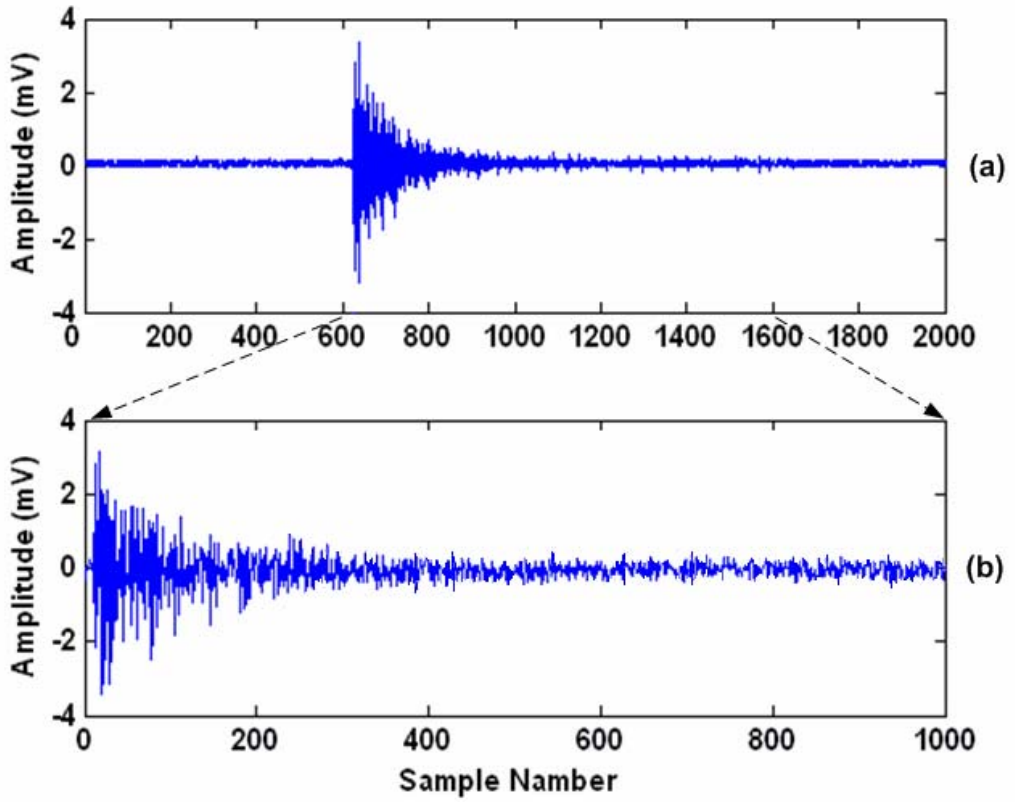


Fig. 4.5 Pre-selection of UHF signal (a) before pre-selection; (b) after pre-selection

4.3 REVIEW OF INDEPENDENT COMPONENT ANALYSIS

Independent component analysis (ICA) is a linear transformation method, which transforms the observed signals into statistically independent components [59-60]. ICA has been applied to image processing [61-62], biomedical engineering [63] and signal processing in radio communications [64]. It has also been applied to load estimation in electric power system [65], where ICA is used to separate the individual customer load profiles from the branch flows. In this research, ICA is used in the new application of feature extraction.

4.3.1 Comparison of PCA and ICA

Principal component analysis (PCA) involves a mathematical procedure that transforms a number of correlated variables into a smaller number of uncorrelated variables known as *principal components*. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. Thus, the objectives of PCA are as follows:

1. To reduce the dimensionality of the data set.
2. To identify meaningful underlying features of the given data set.

The mathematical technique used in PCA is called eigen analysis. A comprehensive review of PCA is given in [66].

ICA can be considered as a generalization of PCA. Both ICA and PCA linearly transform the measured signals into independent or principal components, which are ranked in descending order according to the variance of their corresponding projections. The key difference between ICA and PCA is however in the nature of components obtained. The goal of PCA is to obtain principal components, which are uncorrelated. However, components obtained from ICA are statistically independent, which is a stronger condition than uncorrelated in terms of independency of the components. Separability of features in the measured data is affected by factors such as the frequency response of sensor, the PD source and path of propagation, which are statistically independent. A comparison of the numerical results from ICA and PCA are given in Section 4.5, which clearly favor the former.

4.3.2 Introduction to ICA

Fig. 4.6 illustrates the basic form of ICA, which denotes the process of taking a set of measured signal vectors, X, and extracting from them a set of statistically independent components, Y. Thus, the ICA problem is formulated as

$$Y = WX \tag{4.1}$$

where W is the transformation matrix.

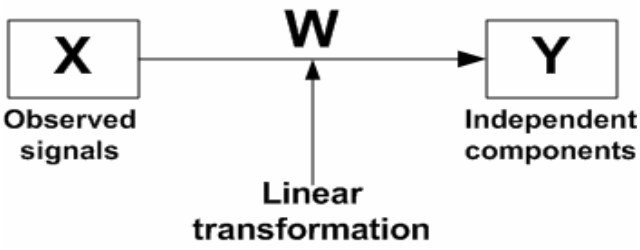


Fig. 4.6 Schematic representation of ICA

In (4.1), both the independent components Y and matrix W are unknown. Therefore, the independent components must be found iteratively by maximizing the independency with respect to W. In this study, an algorithm known as FastICA is adopted for implementing the ICA [67]. According to the Central Limit Theorem, the independency of components can be measured from the statistical property, known as "nongaussianity". In FastICA, a criterion known as "negentropy" is employed to be a quantitative measure of nongaussianity. Maximizing the negentropy with respect to W results in the independent components. Figs 4.7-4.10 show an example that demonstrates the effectiveness of FastICA and the negentropy criterion. Fig. 4.7 shows the two basic signals that are generated independently. The basic signals are then linearly combined to simulate the measured signals (**X**) as illustrated in Fig. 4.8. Using **X** as the input of FastICA, the independent components are estimated one by one. As shown in Figs. 4.9-4.10, the independent components are found in four and three iterations respectively by maximizing the negentropy (J). As observed, the estimated components are almost the same as the original ones. Thus, the effectiveness of FastICA for finding independent components is verified. Key features of ICA and its implementation - FastICA are reviewed in Appendix D.

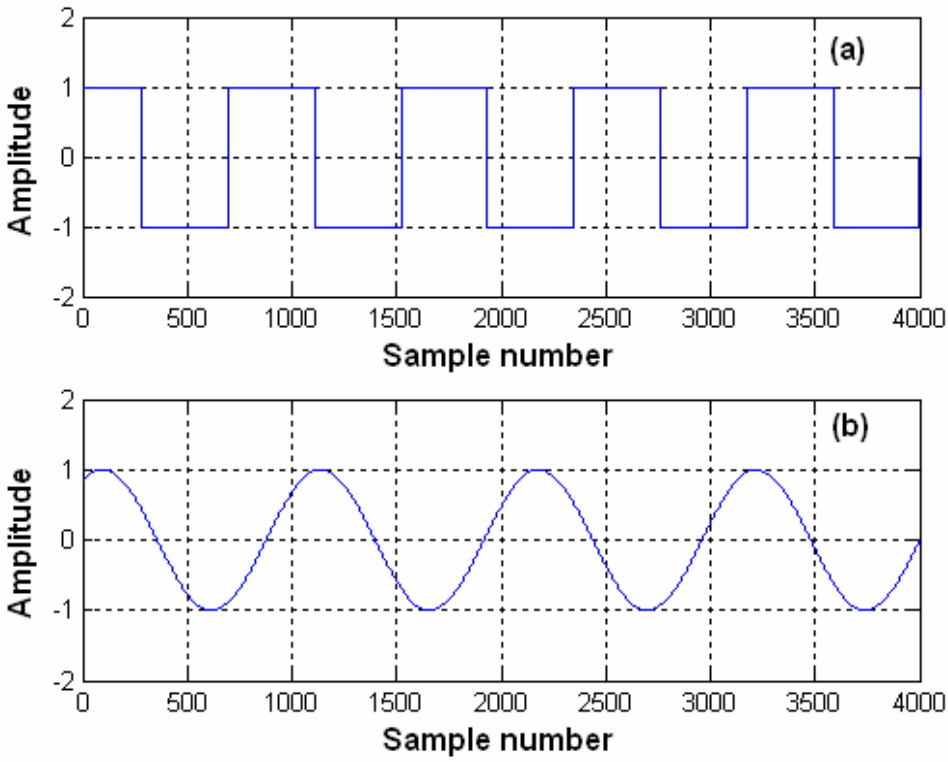


Fig. 4.7 Basic signals

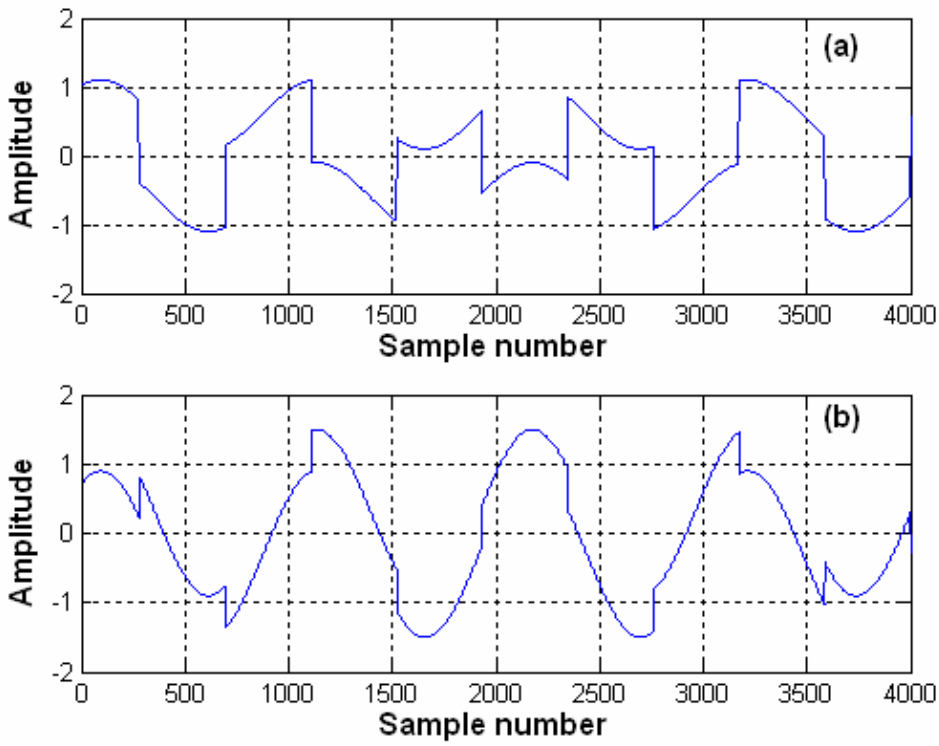


Fig. 4.8 Measured signals (**X**)

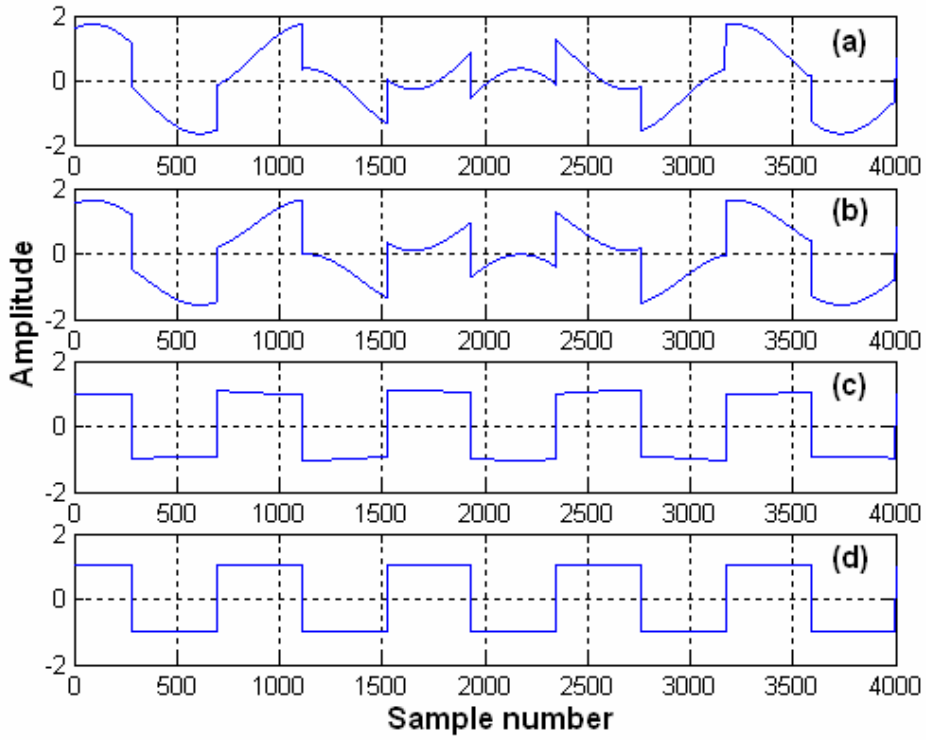


Fig. 4.9 Process of finding the first independent component (a) 1st iteration (J= 4.2797); (b) 2nd iteration (J= 5.7788); (c) 3rd iteration (J= 8.0597); (d) 4th iteration (J= 11.1297).

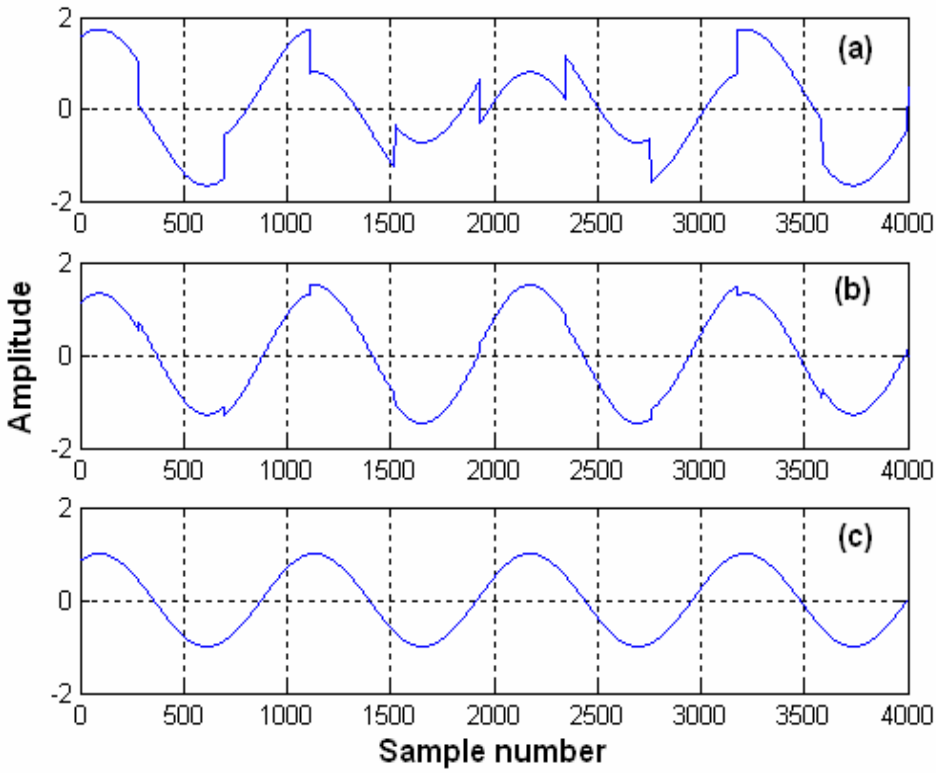


Fig. 4.10 Process of finding the second independent component (a) 1st iteration (J= 4.6197); (b) 2nd iteration (J= 7.4563); (c) 3rd iteration (J= 10.9805).

4.4 FEATURE EXTRACTION BY ICA

The process of ICA-based feature extraction is carried out in two stages:

1. Identification of most dominating independent components.
2. Construction of ICA-based PD features.

The process is carried out with the aim of reducing the length of the working data for subsequent PD identification to be automated by a neural network (Chapter 6).

4.4.1 Identification of Most Dominating Independent Components

The most dominating independent components for compressing the pre-selected signals are identified. The FastICA algorithm (Appendix D) is adopted to first find all the independent components from a chosen set of eight pre-selected signals. The total number of independent components is the same as the number of chosen signal sets. The chosen signal sets and the obtained independent components are shown in Fig. 4.11 and Fig. 4.12 respectively.

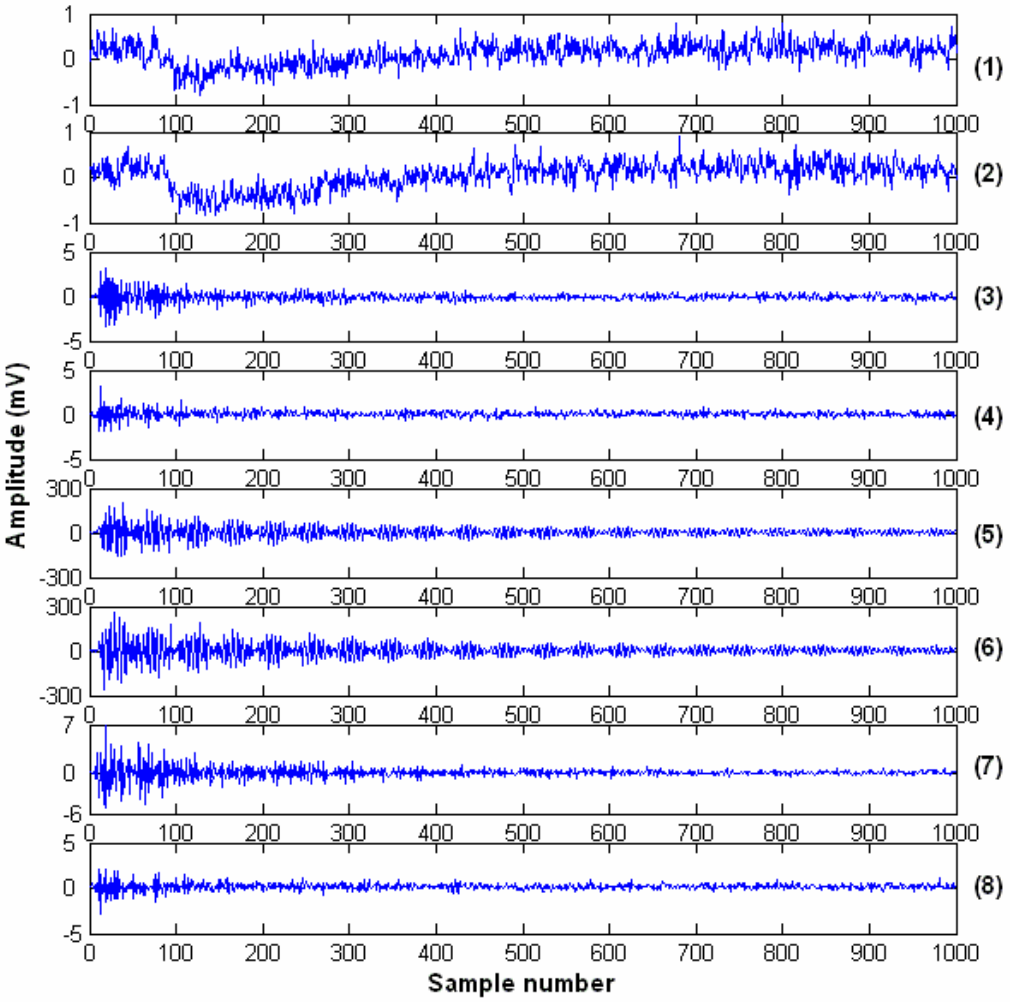


Fig. 4.11 Chosen signal sets for calculating independent components (1)-(2) corona; (3)-(4) particle on the surface of spacer; (5)-(6) particle on conductor; (7)-(8) free particle on enclosure.

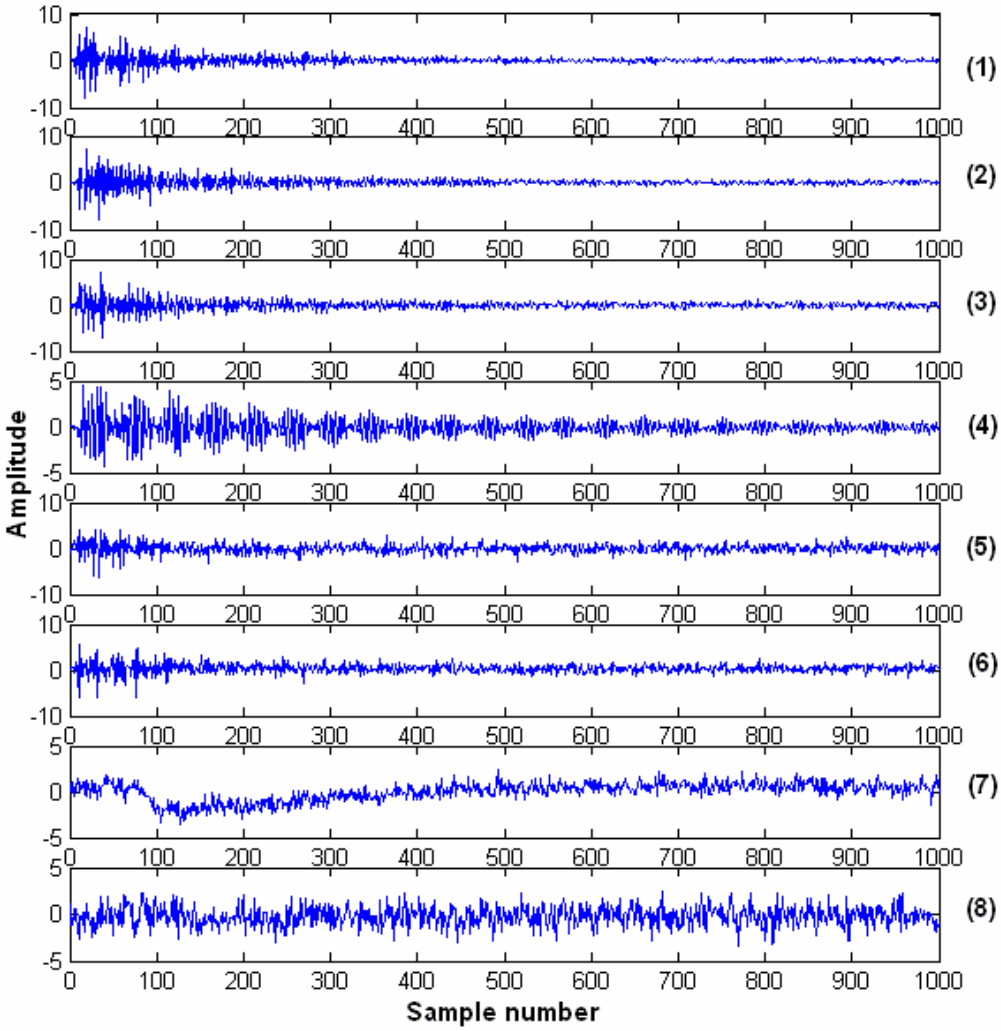


Fig. 4.12 Independent components obtained from FastICA

Each chosen set of signals $\mathbf{x}_i$, ***i=1,2,…,8*** is thus a linear combination of the independent components:

$$\mathbf{x}_i = \sum_{j=1}^{8} a_{i,j} \cdot \boldsymbol{ICAPD}_j \qquad i = 1, 2, \ldots 8 \tag{4.2}$$

where

ICAPD$_j$ = the jth independent component obtained by FastICA that has a size of 1*1000. j runs from 1 to 8.

$a_{i,j}$ = the projection of ith signal set (**x**$_i$) on the direction of jth component. Thus $a_{i,j}$ form a vector of 1*8 for each signal **x**$_i$.

Subsequently, the variance of the projections onto the pth independent component is defined as

$$Var_p = \frac{1}{7}\sum_{i=1}^{8}(a_{i,p} - \mu_p)^2 \tag{4.3}$$

where

$a_{i,p}$ = the projection of ith signal set on the direction of pth component.

μ_p = the mean of the vector $[a_{1,p}, a_{2,p}, ...a_{8,p}]$.

In Fig. 4.12, all ***ICAPD***$_j$ are ranked in descending order according to the variance of their corresponding projections as shown in Table 4.1.

Table 4.1 Variance of projections of all the eight independent components

Independent Components	Variance of the projections
ICAPD₁	*0.2028*
ICAPD₂	*0.1885*
ICAPD₃	*0.0329*
ICAPD₄	*0.0228*
ICAPD₅	*0.0215*
ICAPD₆	*0.0188*
ICAPD₇	*0.0164*
ICAPD₈	*0.0067*

Following the same idea used in PCA-based method, any ***ICAPD*** with small variance (<0.05 in this thesis) in the corresponding projections is discarded for having negligibly small discriminating information. As a result, only the first two independent components in Fig. 4.12 are retained to represent the set of 8 chosen signals.

4.4.2 Construction of ICA-based PD Feature

Altogether 80 measured signals are to be compressed by projecting them onto the two most dominating independent components by the following equation:

$$ICA\_Feature_{m,n} = \boldsymbol{All\_Signal}_m \bullet \boldsymbol{ICAPD}_n^T, \quad \text{m=1,2,...,80; n=1,2.} \tag{4.4}$$

where

$\boldsymbol{All\_Signal}_m$ = the mth set of measured data each of a size 1*1000.

$\boldsymbol{ICAPD}_n^T$ = the transpose of the nth component $\boldsymbol{ICAPD}_n$ and has a size of 1000*1.

m = the number of measured data sets, which runs from 1 to 80.

n = the number of most dominating independent components, which runs from 1 to 2.

The size of the extracted feature set "*ICA_Feature*" is thus 80 * 2 that is much smaller than the size of pre-selected signal sets 80*1000.

4.4.3 Selection of Control Parameters for FastICA

Associated with the FastICA algorithm, there are a number of control parameters to be determined, such as the number of input signals, approximation of negentropy and the stopping criteria. It is crucial to investigate the influences of these parameters, as they have significant impact on the performance of FastICA.

A. Number of Input Signals (Number of Independent Components)

The number of input signals, that is the same as the number of independent components resulted from FastICA, must be set properly to ensure the correctness of the obtained independent components and fast convergence of the algorithm.

If the number of inputs is too small, there will not be enough information of PD signals for FastICA to compute the independent components correctly. On the other hand, if there are too many inputs, it will take longer time for the algorithm to converge. In

addition, since only the most dominating components are useful for the subsequent feature construction task, it is not necessary to compute too many independent components as most of them result in projections with small variances.

Since there are four classes of signals under investigation, the number of inputs should be at least four to cover the varieties of the measured signals. Based on waveforms of the typical signals, the number of inputs is set to eight (two from each class) to make a good tradeoff between accuracy of the resulted components and the convergence speed.

B. Approximation of Negentropy

As introduced in Section 4.3.2, negentropy is employed in FastICA as a measure of nongaussianity to maximize the independency between components. However, it is computationally very difficult to calculate negentropy directly, as an estimate of the probability density function is required [59]. Therefore, it is highly desired to use simpler approximations of negentropy.

In general, the approximation of negentropy for a random vector **t** is formulated as

$$J(t) \approx [E\{G(t)\} - E\{G(v)\}]^2 \tag{4.5}$$

where

E = expectation operator.

v = a Gaussian variable of zero mean and unit variance.

G = any non-quadratic function [59].

Therefore, choosing function G differently results in different approximations of negentropy. As suggested in [67], the following choices of G have proved very useful in many applications.

$$G_1 = -\exp(-u^2/2)$$

$$G_2 = \log(\cosh(u)) \tag{4.6}$$

$$G_3 = \frac{1}{4}u^4$$

$$G_4 = \frac{1}{3}u^3$$

where u is the component vector under investigation. These functions are conceptually simple, robust and fast to compute. Thus, their performances on PD signals are studied and compared in this thesis.

To compare the performances of the approximated negentropies, the sum of variances of projections onto the first two independent components, denoted by ϑ, is employed

as the evaluation criterion. The larger the ϑ value, the better the performance of the corresponding approximated negentropy in terms of discriminative power. Following procedure is then used to compare the approximated negentropies with different function G.

(1) Use the chosen set of signals as input of FastICA as in Section 4.4.1. Set i=1.

(2) Set G_i as the function used to calculate the approximated negentropy in FastICA algorithm.

(3) Run FastICA to find all the independent components.

(4) Compute the variances (Var_1^i , Var_2^i) of the projections onto the first two independent components using equation 4.3.

(5) Compute $\vartheta_i = Var_1^i + Var_2^i$.

(6) Set i=i+1. If i<5, go to (2).

(7) Find the best G that results in the largest ϑ, namely $G_{opt} = \max_{G_i}(\vartheta_i)$.

Table 4.2 shows the performances of approximated negentropies with different G functions. It can be seen that G_1 achieves the largest ϑ value that indicates the best discriminative ability. G_1 is thus adopted in the process of finding the independent components.

Table 4.2 Variances of projections and ϑ corresponding to different G functions

Function	Var_1	Var_2	ϑ
G_1	0.2082	0.1885	0.3913
G_2	0.2092	0.1387	0.3479
G_3	0.2016	0.0914	0.293
G_4	0.2022	0.1142	0.3164

C. Stop Criteria

Since FastICA is an iterative algorithm, some criteria must be applied to stop the program. In this thesis, two criteria are adopted as follows:

(1) The algorithm stops when the maximum number of iterations is reached. It is set to 1000 in this study.

(2) FastICA stops when the change of components saturates over a number of iterations.

The FastICA program stops when either of the above criterions is met.

4.5 RESULTS AND DISCUSSIONS

In this section, low-dimensional feature spaces formed by ICA-based feature extraction method are first presented and compared with those constructed by PCA-based method. Subsequently, the impact of white noise levels on the feature clusters and the convergence performance of FastICA algorithm are illustrated.

4.5.1 Comparison of PCA- and ICA-based Methods

Results from the ICA-based feature extraction are presented and compared with results from PCA-based method. The effectiveness of using the most dominating independent component (1st ranked) is shown in Fig. 4.13 (a). The effect of using a less dominating independent component (6th ranked) is shown in Fig. 4.13 (b), which shows poor separability among PD sources. This indicates that the independent components, with large variances in the corresponding projections, capture the fundamental characteristics of SF_6 PD and corona. Thus the features associated with these components are able to discriminate the defects effectively.

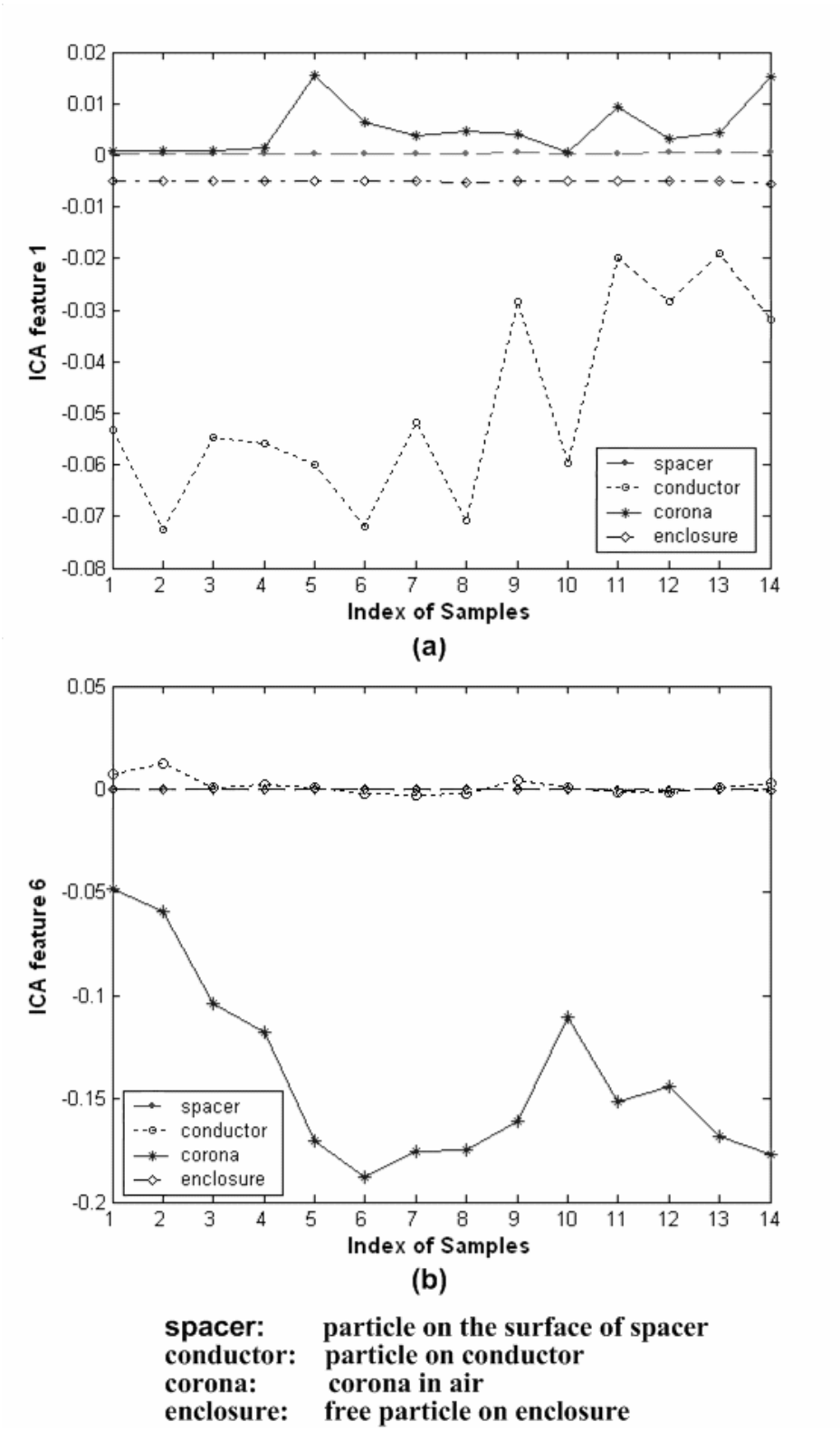


Fig. 4.13 ICA features corresponding to (a) **ICAPD**$_1$ and (b) **ICAPD**$_6$

To compare the performances of ICA- and PCA-based methods, feature extraction using PCA is carried out based on the following procedure:

(1) Use PCA to find the most dominating *principal components*, which result in the largest variances in the corresponding projections.

(2) Project 80 pre-selected signals onto the two most dominating *principal components*, which is similar to the process described in Section 4.4.2.

Figs. 4.14 (a) and (b) show the two most dominating independent components, while the most dominating principal components are illustrated in Figs. 4.14 (c) and (d). It is seen that the components obtained by ICA and PCA are quite different. This indicates that although there are some seeming similarities between PCA and ICA, they are essentially different statistical methods.

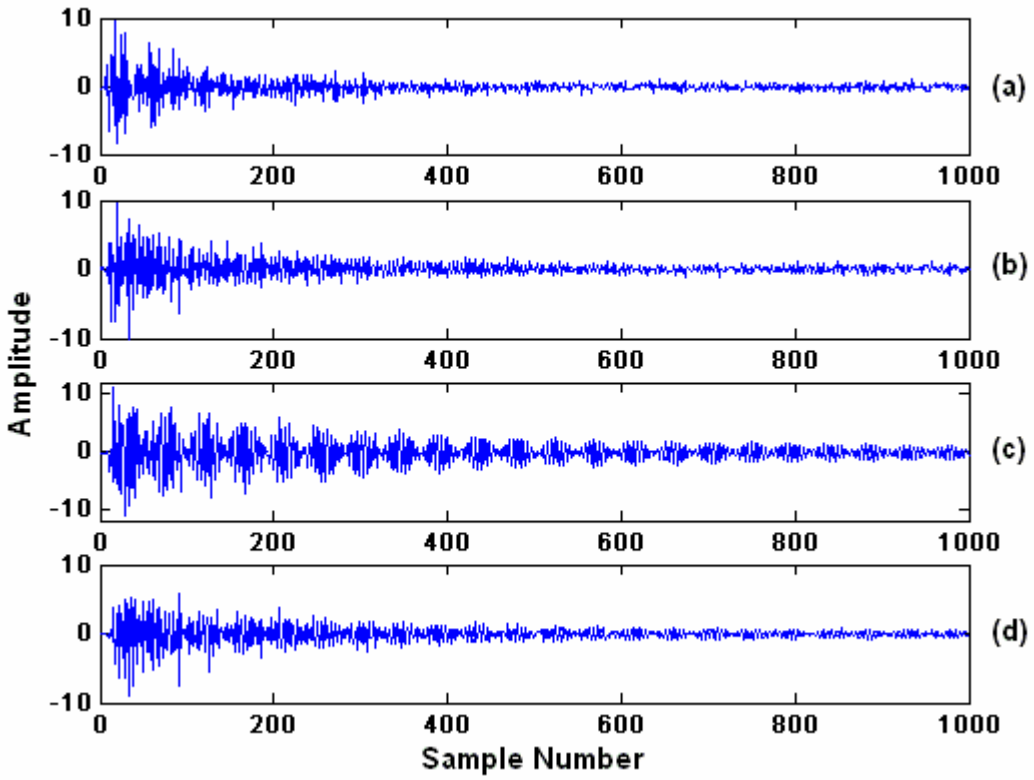


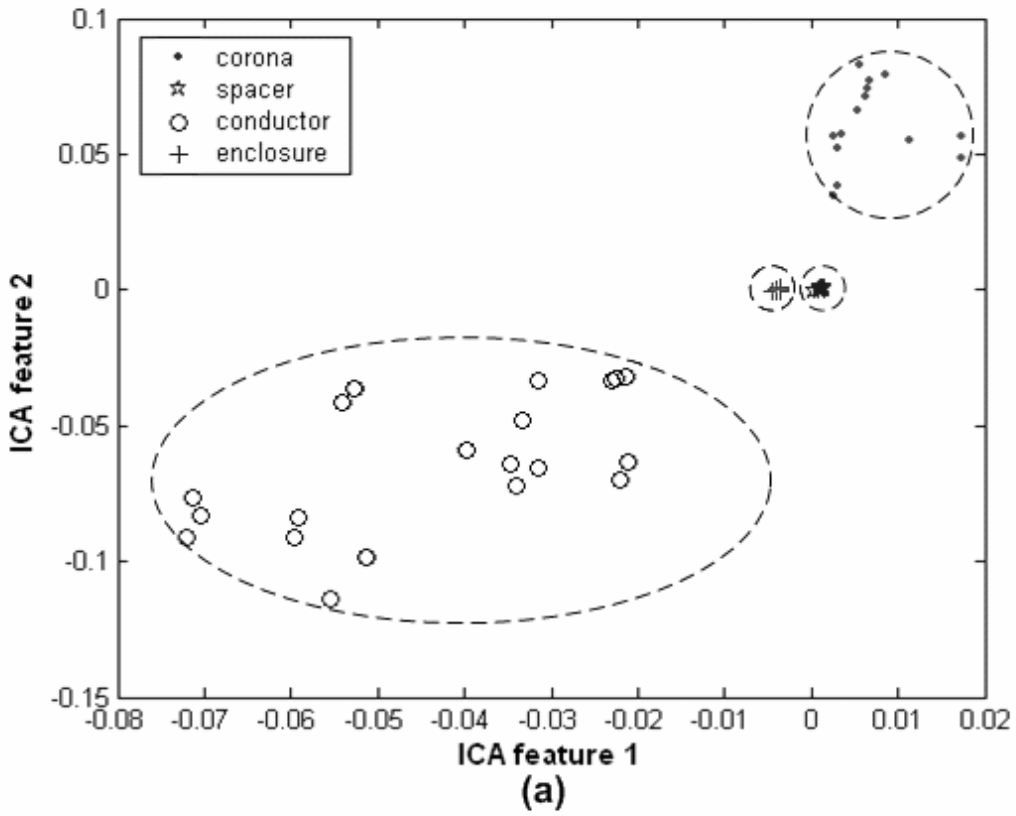
Fig. 4.14 Most dominating (a)-(b) independent components and (c)-(d) principal components

The performances of PCA- and ICA-based methods are first compared in Table 4.3. As observed, both of the variances obtained from independent components take much larger values than those obtained from principal components. This suggests that the features extracted by ICA-based method should lead to better classification due to more discriminative power introduced by independency of the features.

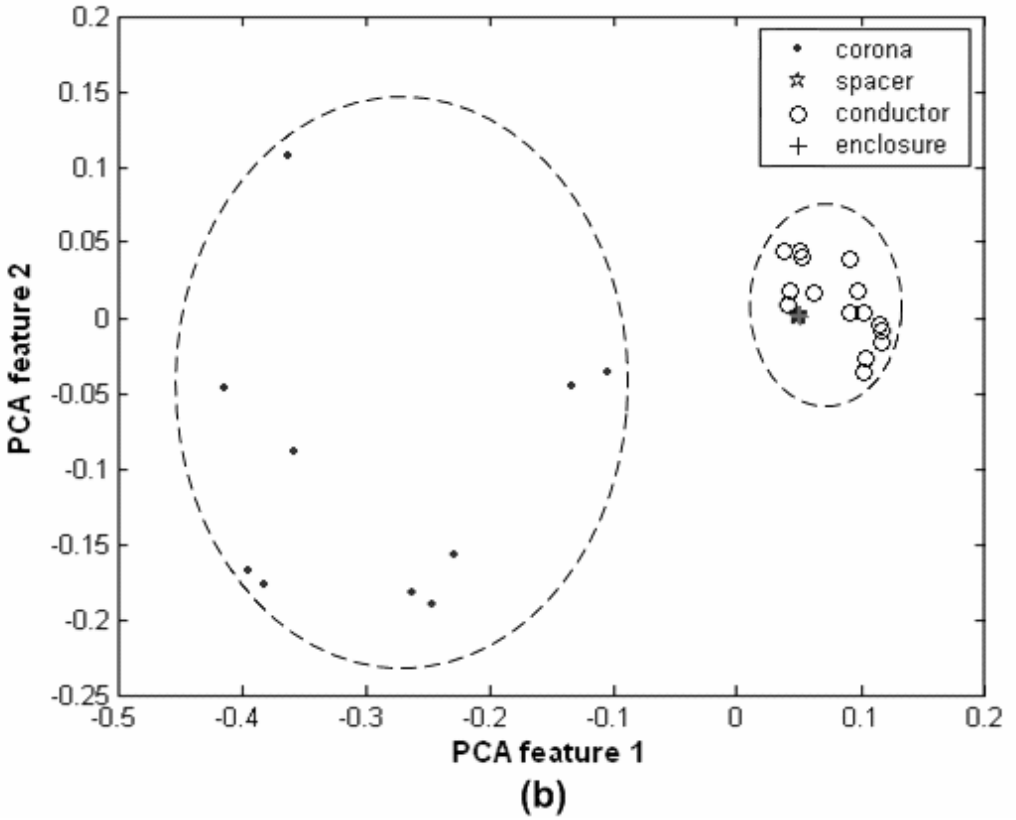
Table 4.3 Variances of projections onto the most dominating independent and principal components

	Var_1	Var_2
Independent components	0.2082	0.1885
Principal components	0.1698	0.0893

Fig. 4.15 further compares the performance of ICA- and PCA-based feature extraction. Features obtained from ICA are seen to cluster distinctly according to the four sources, although clusters corresponding to “spacer” and “enclosure” are close to each other due to the similarity of the two types of PD as shown in Figs. A.3 (b) and (d). Features of “spacer”, “conductor” and “enclosure” resulted from PCA are seen to overlap with each other. This indicates that the ICA-based feature extraction outperforms PCA-based method due to superior statistical properties of the former components.



(a)



(b)

Fig. 4.15 Feature clusters formed by (a) ICA features (b) PCA features

4.5.2 Need for Denoizing

In this section, the need for first removing white noises is demonstrated by investigating the impact of different background noise levels on the results of ICA-based feature extraction.

Table 4.4 shows the average convergence time of FastICA when signals of different SNR levels are used as its input. It can be seen that the convergence time gets longer as the noise level gets higher. The convergence time increases significantly due to the more computation time required in the process of maximizing negentropy. In the worst case, where the SNR of input signals is -5, the algorithm is not able to converge within the pre-determined maximal iteration.

Table 4.4 Average convergence time

Noise Level	SNR=17 (after denoizing)	SNR=0	SNR= -5
Convergence Time (s)	1.911	3.207	* 183

*: In this case, FastICA is not able to converge in 1000 iteration. (Section 4.4.3 – C). Convergence is observed at 9800 iteration.

Fig. 4.16 illustrates the feature clusters obtained from ICA-based method with input signals of different noise levels. As shown in Fig. 4.16 (a) where the SNR of input signals is 0, features of “spacer” and “enclosure” are seen to overlap with each other, although features of “corona” and “conductor” are still well separated. The worst case (SNR= -5) is shown in Fig. 4.16 (b), where the features are all mixed up. It is impossible to discriminate PD source correctly using these features. Thus, it is imperative to remove the white noises before the features are extracted.

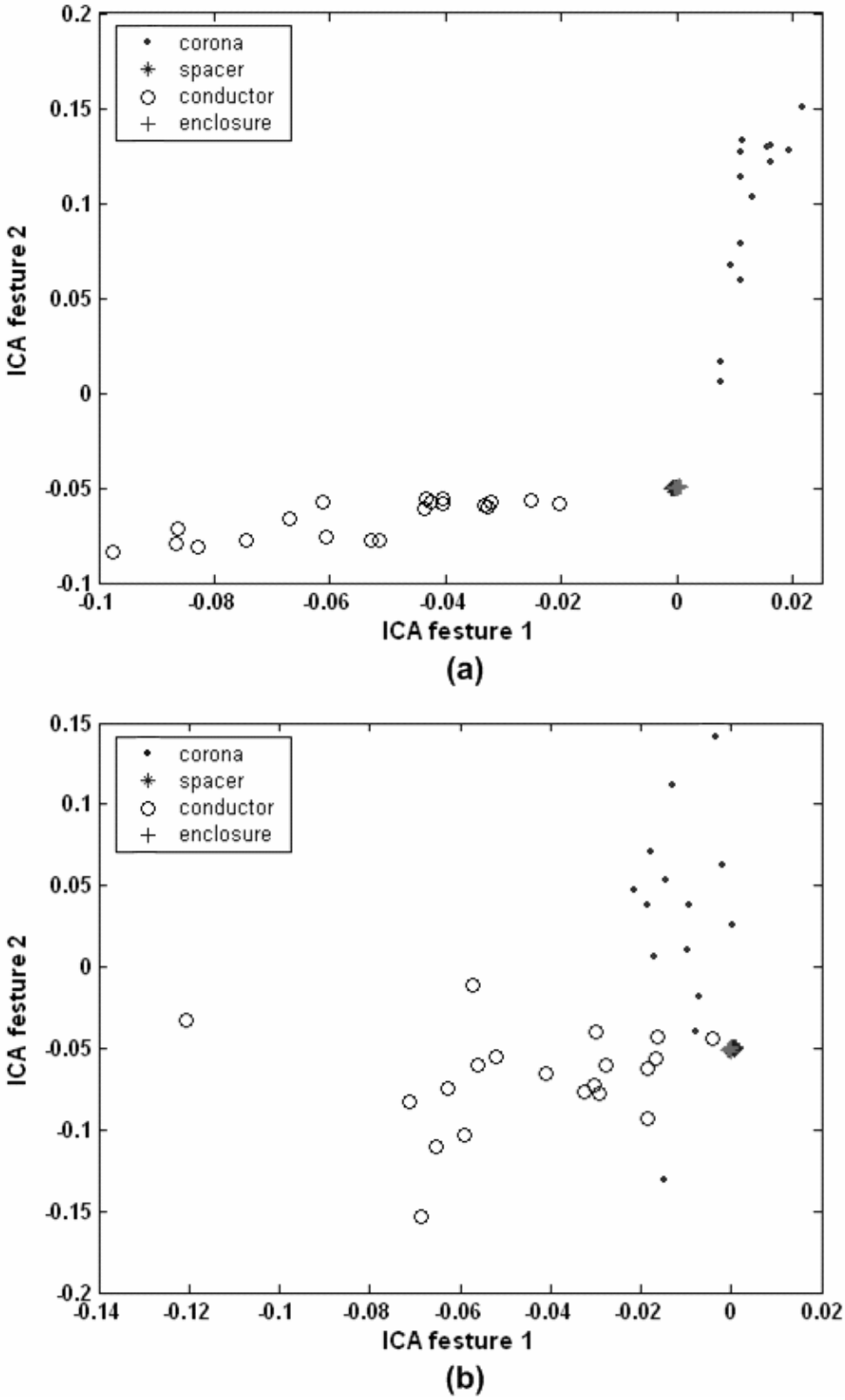


Fig. 4.16 Feature clusters formed by ICA-based method. Noise level of input signals is (a) SNR=0; (b) SNR= -5.

4.6 CONCLUDING REMARKS

In order to improve the efficiency and accuracy of PD identification, it is crucial to extract the most dominating features of measured UHF resonance signals. In this chapter, a method using Independent Component Analysis is developed for such purpose. Experimental results show that the extracted features form distinct clusters according to different sources, which indicates that good classification performance may be achieved by using such features. White noises present in the measured signals are seen to have deteriorated the discrimination ability of the extracted features. The importance of denoizing is thus verified.

CHAPTER 5

PD FEATURE EXTRACTION BY WAVELET PACKET TRANSFORM

In previous chapter, a typical time-domain method, namely ICA-based method is developed for extracting PD features. However, the method forms feature clusters with small margin between “enclosure” and “spacer”. To extract features with higher quality, a time-frequency-domain method, which is based on the wavelet packet transform, is proposed in this chapter. Firstly, the wavelet-packet-based method is described, followed by discussions of parameters selection for feature extraction purpose. Secondly, numerical results are presented and the necessity of denoizing is justified. Lastly, the relationship between PD features extracted by Wavelet Packet Transform and Fast Fourier Transform is discussed.

5.1 INTRODUCTION

In Chapter 4, ICA-based PD feature extraction method is developed with limited success. Although features resulted from ICA form distinct clusters, the margin between the clusters of “spacer” and “enclosure” is too small to ensure a low misclassification rate on new data. The reason of having close clusters is that the time domain signals of the two types of PD have similar waveforms. As a result, the features extracted by ICA, which is a time domain method, tend to be close to each other. To solve this problem, not only time domain but also frequency domain information should be considered.

One advantage of using wavelet-based techniques to decompose a signal is that wavelet transform allows us to examine different time-frequency resolution components in a signal. Therefore, more effective features may be extracted by using such techniques including discrete wavelet transform and wavelet packet transform. Wavelet packet transform of a signal results in a full decomposition tree that offers better frequency resolution than the partial tree formed by discrete wavelet transform. Therefore, in this chapter, a wavelet-packet-based scheme is proposed to extract PD features as shown in Fig. 5.1. The first two blocks in the scheme, namely denoizing and pre-selection have been discussed in previous chapters. Salient features of the other blocks are discussed in the following sections.

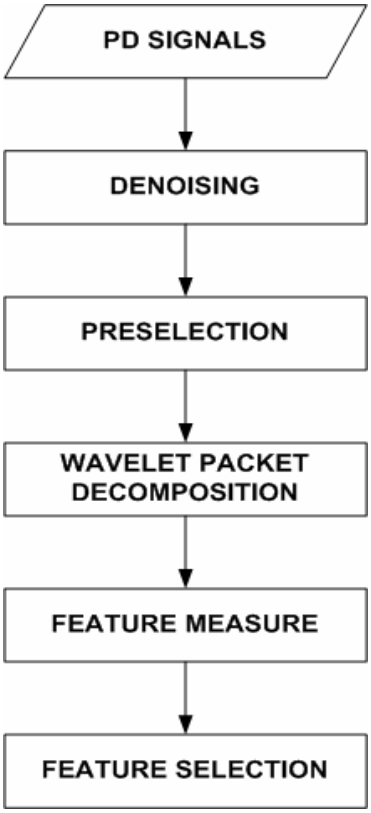


Fig. 5.1 Flowchart of wavelet-packet-based PD feature extraction scheme

5.2 WAVELET-PACKET-BASED FEATURE EXTRACTION

In this section, the major steps of wavelet-packet-based feature extraction method, namely wavelet packet decomposition, feature measure and feature selection, are described.

5.2.1 Wavelet Packet Decomposition

To extract characteristic information from time domain UHF signals, they are first decomposed into the wavelet packet domain, forming wavelet-packet-decomposition (WPD) trees. Since there are totally 80 UHF signals used for developing the method,

80 WPD trees are formed by performing the decomposition. The wavelet packet decomposition is set on a decomposition level of 5 (Fig. 5.2) and the “db9” wavelet packets based on the effectiveness of the obtained features. The selection of decomposition level and wavelet filters is discussed in Section 5.3.

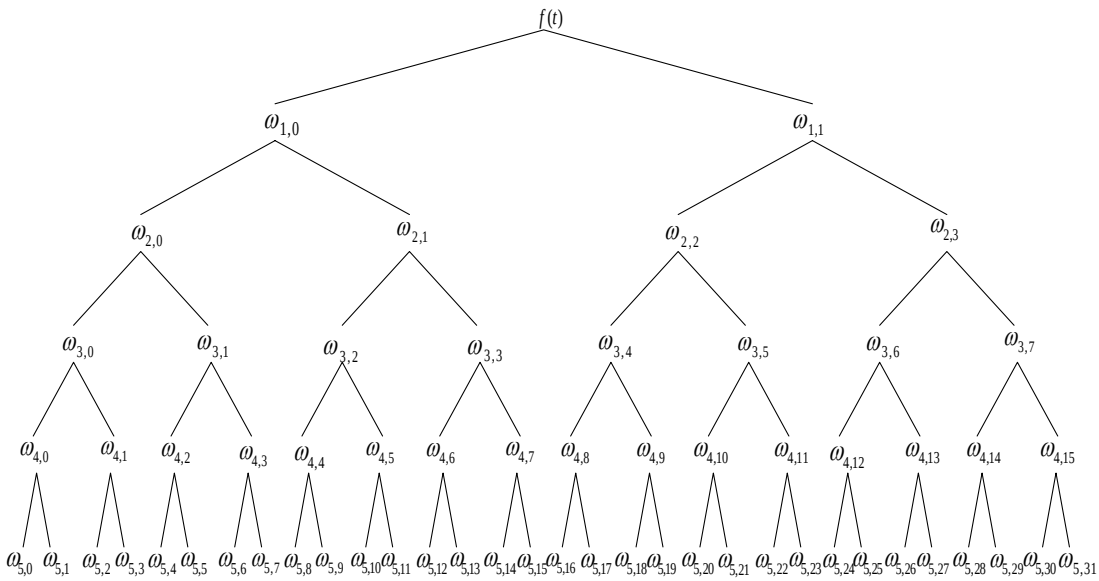


Fig. 5.2 WPD tree of level 5 (Copy of Fig. 3.8 for reference)

Each node in the WPD tree represents a set of decomposition coefficients which correspond to a certain frequency band as shown in Fig. 5.3. The topmost node contains the pre-selected signal which has a sampling frequency of 4 GHz. According to the Nyquist theory, the highest frequency content contained in the nodes is up to 2 GHz, namely half of the sampling frequency f_0. Therefore, one level of decomposition results in two nodes that have spectra of 0-1 GHz ($0-\frac{f_0}{4}$) and 1-2 GHz ($\frac{f_0}{4}-\frac{f_0}{2}$) respectively. As illustrated in Fig. 5.3, frequency span of each father node is the union of that of its child nodes.

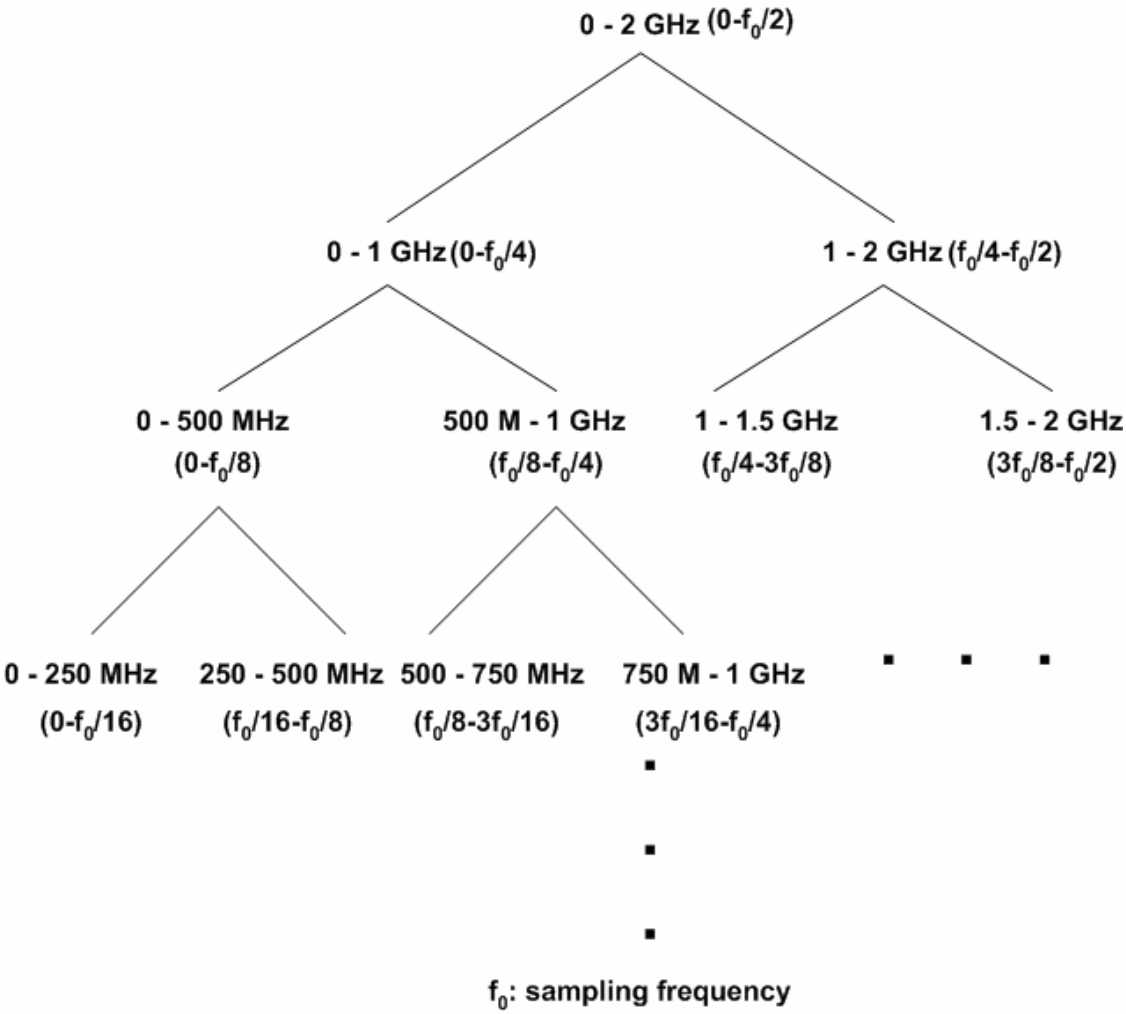


Fig. 5.3 Frequency span of nodes in the WPD tree

5.2.2 Feature Measure

Wavelet packet decomposition enables time-frequency analysis of the PD signals based on the decomposition coefficients. However, direct manipulation of a whole set of decomposition coefficients is prohibitive as the space normally has very high dimensionality. For instance, a five-level WPD (Fig. 5.2) of a pre-selected signal results in 5000 (5*1000) coefficients. Therefore, appropriate features must be defined based on the WPD coefficients to reduce the dimensionality and retain the time-

frequency characteristics of the decomposition coefficients. Features defined according to nodes known as *node feature* are discussed in this section.

A. Node kurtosis

Kurtosis is a statistical parameter describing the shape of a data distribution. It is a measure indicating whether a data distribution is more or less peaky than the normal distribution. As shown in Fig. 5.4, data with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data with low kurtosis tend to have a flat top near the mean rather than a sharp peak.

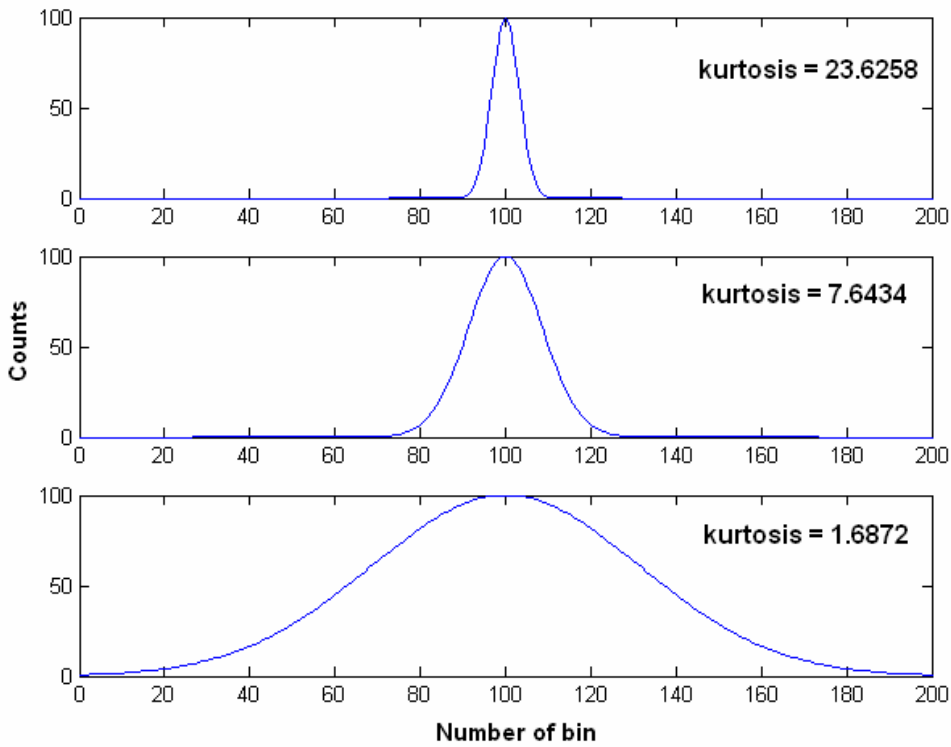


Fig. 5.4 Data distribution with different kurtosis values

Node kurtosis is defined as the kurtosis of the decomposition coefficients of each node (j,n) in the WPD tree as in equation 5.1.

$$K_{j,n} = \frac{\sum_{k} (\varpi_{j,k,n} - \mu_{j,n})^4}{(N_{j,n} - 1)\sigma_{j,n}^4} - 3 \tag{5.1}$$

where

$K_{j,n}$ = node kurtosis of node (j,n).

$\omega_{j,n}$ = the WPD coefficients vector corresponding to node (j,n) in the decomposition tree.

$\varpi_{j,k,n}$ = the kth coefficient of node (j,n).

$N_{j,n}$ = the length of the coefficients vector $\omega_{j,n}$.

$\mu_{j,n}$ = mean value of coefficients vector $\omega_{j,n}$.

$\sigma_{j,n}$ = standard deviation value of coefficients vector $\omega_{j,n}$.

Since normal distribution has a kurtosis value of three, the minus three in the above equation means normalization according to normal distribution.

B. Node skewness

Skewness is another distribution-shape-related statistical parameter. It characterizes the degree of asymmetry of a distribution around its mean. As illustrated in Fig. 5.5, skewness is zero for a symmetrical distribution, positive if it is heavier towards the left-hand side and negative if it is heavier towards the right-hand side.

Node skewness is defined as the skewness of decomposition coefficients of each node (j,n) as in equation 5.2.

$$S_{j,n} = \frac{\sum_k (\varpi_{j,k,n} - \mu_{j,n})^3}{(N_{j,n} - 1)\sigma_{j,n}^3} \tag{5.2}$$

where

$S_{j,n}$ = node skewness of node (j,n).

The other variables in the above equation have the same meaning as in equation 5.1. Comparing equation 5.1 with equation 5.2, it is seen that they have similar structure in mathematical formula. The difference is only in the order of formula, where kurtosis has an order of 4 and skewness is of order 3. However, they have completely different statistical property.

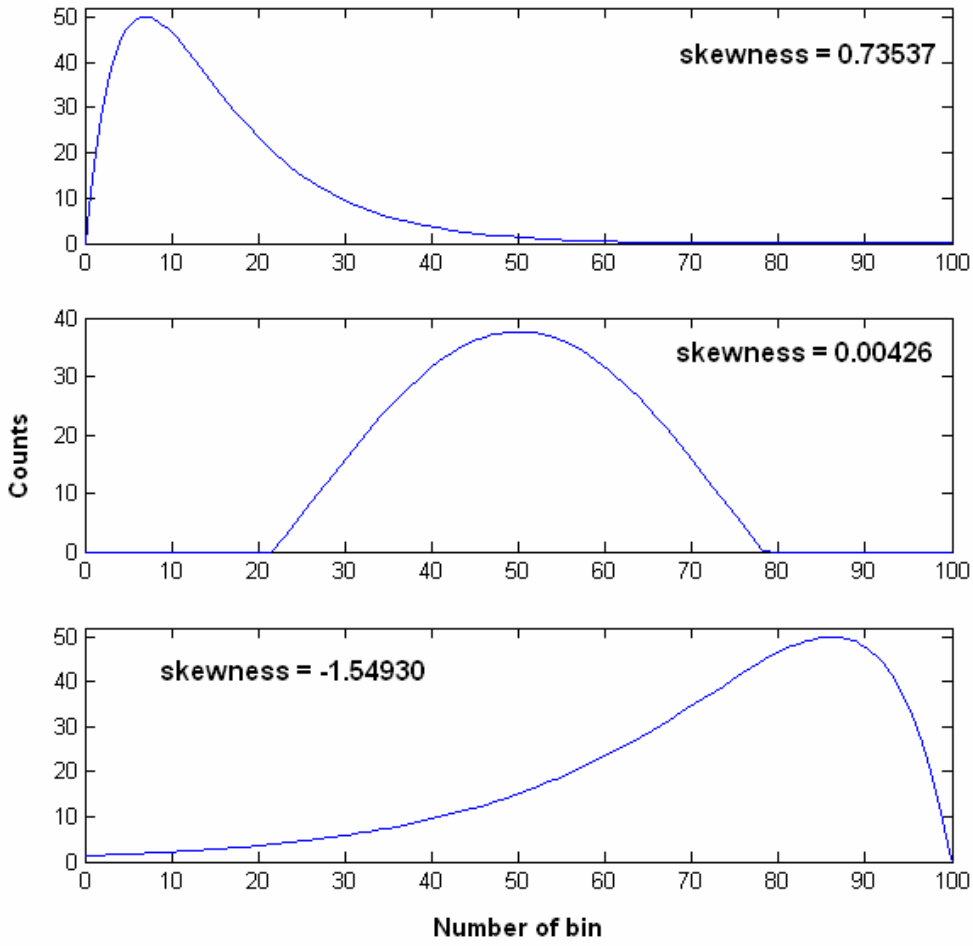


Fig. 5.5 Data distribution with different skewness values

Taking advantage of the time information provided by wavelet packet transform, node kurtosis and node skewness describe the distribution shape of the decomposition coefficients locally in a specified frequency band at each node. They enable detailed time-frequency analysis of the UHF signals. Thus, they are considered as important *local* features for PD identification.

C. Node energy

The wavelet packet power spectrum provides us with information about the local spectral content of the signal. The local wavelet packet power spectrum corresponding to each node (j,n) is defined as

$$P_{j,n} = \frac{1}{N}\left|\omega_{j,n}\right|^2 \tag{5.3}$$

where

$\omega_{j,n}$ = the WPD coefficients vector corresponding to node (j,n) in the decomposition tree.

N = length of the signal.

To reduce the computation complexity, the normalization factor 1/N in (5.3) is omitted in our analysis. The modified wavelet spectrum is named as “node energy” [68], and is denoted as

$$E_{j,n} = \left|\omega_{j,n}\right|^2 \tag{5.4}$$

D. Node median and node mean

Mean and median are two types of measures for central tendency. Median is a measure of the "middle" of the data. For an odd number of data points arranged in ascending order, median is actually the middle value, and for an even number of data points it is

the value halfway between the two middle data points. Mean is computed by adding all the numbers in the set and dividing the sum by the number of elements added. For a given set of data, these measures may be very close or may be quite different, depending on how the data are distributed.

Node median and node mean are defined in the same way of the previous node features. They are computed by taking the median and mean of the decomposition coefficients of each node as in equation 5.5 and 5.6 respectively.

$$Med_{j,n} = \begin{cases} y_{(N+1)/2} & \text{if N is odd} \\ \frac{1}{2}(y_{N/2} + y_{N/2+1}) & \text{if N is even} \end{cases} \tag{5.5}$$

where

y = sorted coefficients vector of node (j,n).

N = length of the coefficients vector of node (j,n).

$$M_{j,n} = \frac{1}{N}\sum_{k=1}^{N} \varpi_{j,k,n} \tag{5.6}$$

where

$\varpi_{j,k,n}$ = the kth coefficient of node (j,n).

Node kurtosis, node skewness, node energy, node median and node mean are computed for each node in a WPD tree. As illustrated in Fig. 5.6, these calculated features form five feature trees, namely the kurtosis tree, skewness tree, energy tree, median tree and mean tree, in association with each WPD tree. For example, each node of the energy tree contains the energy value of the coefficients in the corresponding node of WPD tree. Since each feature tree contains 62 nodes, the total number of node features for a PD signal is 310 (=62*5), which is much smaller than the number of WPD coefficients (=5000).

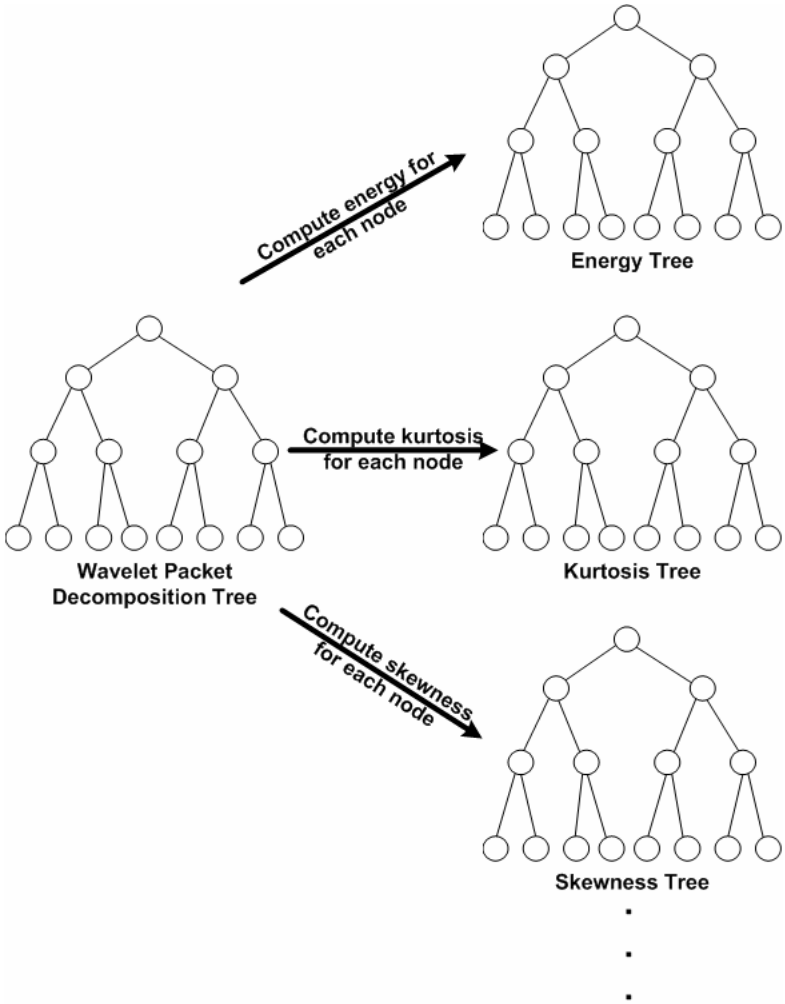


Fig. 5.6 Construction of feature trees

5.2.3 Feature Selection

One of the crucial issues in classification is the curse of dimensionality [69]. Therefore, a low-dimensioned feature space is highly desired to ease the design of classification system and improve its generalization properties. Although the node features extracted from the WPD coefficients have reduced the number of features, the dimensionality of the feature space is still too high to achieve satisfactory speed and classification performance. In addition, the existence of undesired features makes the classification unnecessarily difficult. Therefore, feature space must be further reduced by discarding the features that have little discrimination information. Only those features that preserve maximum class separability are selected to be used in the classification process. In this study, the criterion based on within- and between-class scatter is modified to be the measure of discrimination ability of individual node features.

The *within-class scatter value* (S_w) measures the scatter of feature vectors of different classes around their respective mean values. The *between-class scatter value* (S_b) is defined as the scatter of the conditional mean values around the overall mean value. In this thesis, the S_w and S_b of a node feature of type t for an L-class problem are defined as follows:

$$S_w(j,n)_t = \sum_{c=1}^{L} \frac{N_c}{N} \sigma_c^2(j,n)_t \tag{5.7}$$

$$S_b(j,n)_t = \sum_{c=1}^{L} \frac{N_c}{N} \left(\eta_c(j,n)_t - \eta(j,n)_t \right)^2 \tag{5.8}$$

where

t = the type of feature such as energy, kurtosis, and so on.

$\sigma_c^2(j,n)_t$ = the variance of features of type t at node (j,n) across the signals belonging to class c.

$\eta_c(j,n)_t$ = mean value of features of type t at node (j,n) for class c.

$\eta(j,n)_t$ = mean value of features of type t at node (j,n) for all signals.

N_c = the number of signals belonging to class c.

N = the number of total signals that is 80 in this study.

Then a criterion, known as J criterion for feature selection is defined as:

$$J(j,n)_t = \frac{S_b(j,n)_t}{S_w(j,n)_t} \tag{5.9}$$

The between-class scatter value indicates how far the features of different classes are separated. On the other hand, the within-class scatter value shows the compactness of the feature cluster corresponding to each class. In order to have a good separability for classification, large *between-class scatter* and small *within-class scatter* are desired. Therefore, a large $J(j,n)_t$ value indicates that features of type t at node (j,n) form a good feature set.

To illustrate and verify the effectiveness of the *J* criterion, equations 5.7 and 5.8 are simplified by considering the 2-class case as follows:

$$S_w(j,n)_t = C_1\sigma_1^2(j,n)_t + C_2\sigma_2^2(j,n)_t \tag{5.10}$$

where $C_1 = \frac{N_1}{N}$ and $C_2 = \frac{N_2}{N}$ are constants. $\sigma_1^2(j,n)_t$ and $\sigma_2^2(j,n)_t$ are the variances of features of type *t* at node (j,n) for the two class respectively.

$$S_b(j,n)_t = C_3\left(\eta_1(j,n)_t - \eta_2(j,n)_t\right)^2 \tag{5.11}$$

where $C_3 = \frac{N_1 N_2}{N^2}$ is a constant. $\eta_1(j,n)_t$ and $\eta_2(j,n)_t$ are mean values of features of type *t* at node (j,n) for class 1 and 2 respectively.

It is seen from equations (5.10) and (5.11) that S_w and S_b are in proportion to the sum of the variances and the distance of the means respectively. Therefore, the smaller the variances and the larger the distance of means, the better the features' class separability.

The effectiveness of the *J* criterion is illustrated in Fig. 5.7. Fig. 5.7 (a) shows the case where the feature clusters have means that are far from each other, but they are still not well-separated due to their large variances. On the other hand, the means of feature clusters in Fig. 5.7 (b) are too close to have a good separability, although the clusters

are compact. Fig. 5.7 (c) is the worse case where the mean values are close and variances are large. As observed, the feature clusters are almost overlapped. An example of good separability is shown in Fig. 5.7 (d), where feature clusters with compact distribution are separated in the distance. Therefore, it can be concluded that a small S_w and a large S_b lead to good features for classification. Thus the use of *J* criterion is justified.

To select the best features, *J* values of all the 310 (62*5) nodes in the feature trees are calculated using the *J* criterion. Features with the largest *J* values are selected to be the input of the neural network (Chapter 6).

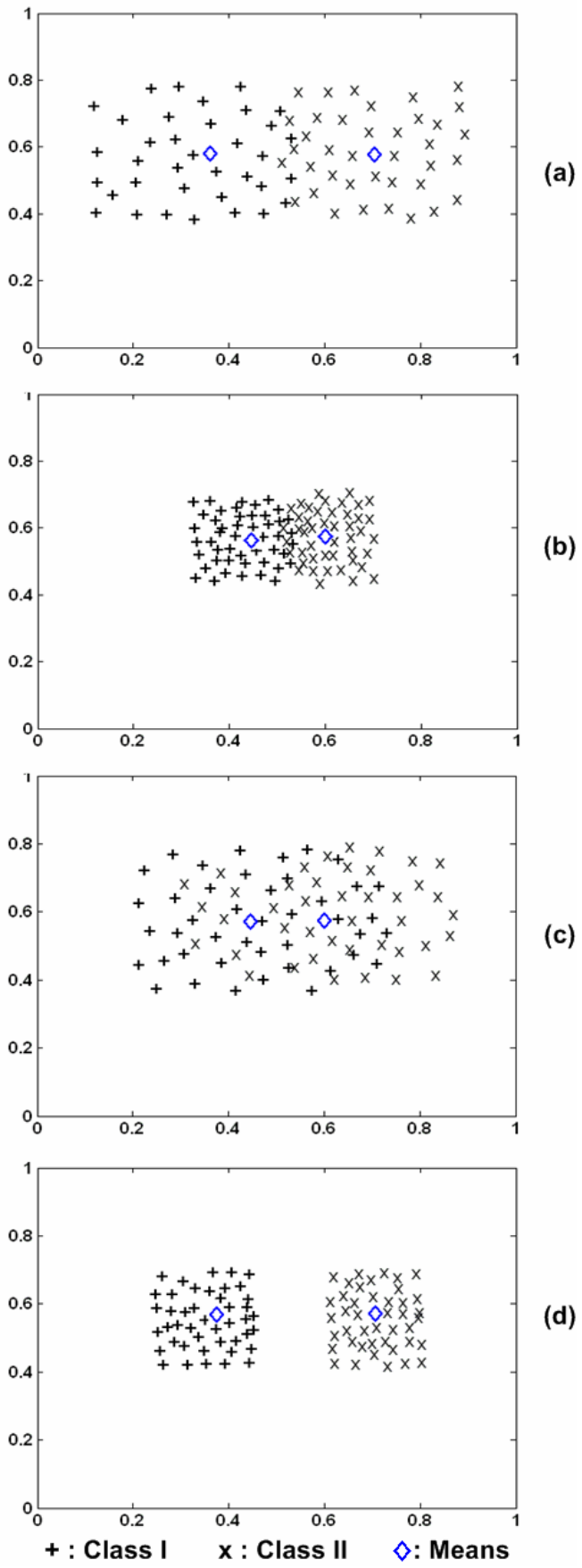


Fig. 5.7 Effectiveness of the *J* criterion

5.3 DETERMINATION OF WPD PARAMETERS

Associated with the wavelet packet decomposition, there are two parameters to be determined, namely decomposition level and wavelet filters. These parameters have significant impact on the feature calculation and selection. Thus, the selection of these parameters is investigated in this section.

5.3.1 Level of Decomposition

As the time-frequency features are defined according to nodes of WPT tree, the number of candidate features is proportional to the number of nodes in the decomposition tree. Therefore, a low decomposition level results in less candidate features, which may not include the best features. Thus, it is preferred to apply a decomposition level as high as possible.

On the other hand, when decomposition level gets higher, the algorithm will get slow dramatically. Therefore, it is crucial to select a suitable decomposition level that makes a good tradeoff between number of candidate features and the speed. Table 5.1 shows the effect of choosing different decomposition level. It can be seen that a decomposition of 5 achieves sufficient number of features as well as acceptable speed. Therefore, a decomposition of 5 is used for feature extraction.

Table 5.1 Selection of decomposition level

Decomposition level	Number of features	Time (min)
1	10	0.5
2	30	2.6
3	70	7.0
4	150	10.5
5	310	25.5
6	630	123.6
7	1270	425.0
8	2550	935.5

5.3.2 Best Wavelet for Classification Purpose

Criteria used to measure the suitability of a wavelet are application dependent. In Chapters 2 and 3, "minimum prominent decomposition coefficients" and denoizing performance indicators such as SNR and CC are employed as the wavelet selection criteria for denoizing. However, these criteria do not reflect the classification ability of a wavelet, as class information is not considered in the selection process.

For classification, the wavelet which leads to maximal separation of classes in the feature space is the best choice. Therefore, the J criterion defined in Section 5.2.3 is used to select the best wavelet. The procedure leading to the determination of best wavelet is as follows:

(1) Select a wavelet from a set of candidate wavelets that have not been examined. Set the decomposition level to 5.

(2) Perform wavelet packet decomposition on all 80 data as in Section 5.2.1.

(3) Construct feature trees according to Section 5.2.2.

(4) Calculate J values for all the nodes in five types of feature trees according to Section 5.2.3.

(5) Summate the first five largest J values and denoted as J_{sum}.

(6) If all the candidate wavelets have been examined, go to (7). Otherwise, go to (1).

(7) Compare J_{sum} and the largest J values corresponding to different wavelets and choose the one with the largest J_{sum} value.

Using above procedure, largest J values and J_{sum} corresponding to candidate wavelets are computed and shown in Table 5.2. It can be seen that the use of wavelet "db9" results in the largest J_{sum}, which in turn leads to the most discriminating features. The best wavelet for denoizing, namely "sym6" wavelet is seen to have an inferior performance in terms of discrimination ability. Thus, "db9" is employed in the feature extraction process.

Table 5.2 Largest J values corresponding to candidate wavelets

wavelet	largest J	2nd largest	3rd largest	4th largest	5th largest	J_{sum}
db1	11.1472	8.9926	8.9214	5.1801	5.0376	39.2790
db2	8.7717	6.4077	5.1951	4.6426	4.5084	29.5255
db3	8.6536	8.53	7.4543	7.2913	6.8793	38.8084
db4	11.6558	10.6842	9.0882	8.7225	7.7441	47.8948
db5	10.0822	8.8847	8.3091	7.1419	5.1006	39.5185
db6	8.9007	7.6649	6.6189	5.6495	5.5355	34.3695
db7	9.2279	9.0119	7.5341	6.633	6.3431	38.7501
db8	9.065	8.4025	7.749	7.2971	7.2311	39.7446
db9	**12.1435**	**11.8492**	**8.6009**	**8.5927**	**8.0492**	**49.2355**

db10	11.0247	8.2442	7.4261	6.5047	6.0645	39.2642
sym4	9.162	8.9113	8.7878	7.9445	6.5349	41.3406
sym5	9.3107	8.4964	8.0997	7.4842	7.3544	40.7455
sym6	9.0327	8.9521	8.1455	7.2787	6.302	39.7110
sym7	8.8172	8.4305	6.9262	5.8794	5.5216	35.5750
sym8	9.0529	8.391	8.3408	8.2732	8.241	42.2990
sym9	9.4052	8.1913	6.2384	5.5362	5.1012	34.4723
sym10	9.196	7.1077	6.1544	6.068	5.6584	34.1846
coif1	11.3313	9.2107	8.0681	7.6744	7.2843	43.5689
coif2	11.03	10.1664	9.5728	8.7226	7.9131	47.4049
coif3	9.0115	8.8234	7.6481	7.1127	7.0737	39.6694
coif4	8.8934	8.4061	8.0776	7.2761	6.6866	39.3399
coif5	8.9401	6.8749	6.5338	6.1461	5.2337	33.7286

5.4 RESULTS AND DISCUSSIONS

Results obtained from the wavelet-packet-based feature extraction method are presented and discussed in this section. The effectiveness of the extracted features is first verified. Subsequently, impact of wavelet and white noise levels is investigated. Lastly, the relationship between node energy and power spectrum is clarified.

5.4.1 Effectiveness of Selected Features

Extracted by the wavelet-packet-based method, ten features (*WPT_feature*) with the largest *J* criterion values are summarized in Table 5.3. It is seen that seven out of ten selected features are distribution-shape-related features, namely node kurtosis and node skewness. This indicates that the distribution-shape-related node features are more effective in PD identification.

The frequency ranges of selected features show that both high-frequency and low-frequency decomposition coefficients contain discriminating information. Particularly, the selection of features defined on nodes at the right-hand side of WPD tree, such as (5,21), (5,19) and (5,20), suggests that wavelet packet transform is more suitable than discrete wavelet transform for this study, as these nodes do not exist in the tree structure formed by discrete wavelet transform.

As shown in Table 5.3, the feature with the largest J value is the node kurtosis of node (5,21) that corresponds to frequency range of 1.3125 to 1.375 GHz. This means that the sharpness of decomposition coefficients distribution of the particular frequency range exhibits the largest difference between signals of SF_6 PD as well as air corona.

Table 5.3 Features extracted by wavelet-packet-based method (*WPT_feature*)

serial no.	feature	*J* value	frequency range (Hz)
1	$(5,21)_{kurtosis}$	12.1435	1.3125 G – 1.375 G
2	$(1,0)_{skewness}$	11.8492	0 – 1 G
3	$(5,1)_{energy}$	8.6009	62.5 M – 125 M
4	$(5,19)_{skewness}$	8.5927	1.1875 G – 1.25 G
5	$(5,0)_{kurtosis}$	8.0492	0 – 62.5 M
6	$(3,0)_{kurtosis}$	7.7291	0 – 250 M
7	$(5,20)_{median}$	7.5266	1.25 G – 1.3125 G
8	$(5,11)_{skewness}$	6.6111	687.5 M – 750 M
9	$(4,0)_{skewness}$	6.5075	0 – 125 M
10	$(4,2)_{energy}$	6.1892	250 M – 375 M

The effectiveness of the extracted features is shown in Figs. 5.8 – 5.10. Fig. 5.8 shows the number of wavelet-packet-decomposition coefficients whose values fall into evenly partitioned ranges. Taking Fig. 5.8(a) as an example, the first range is [-0.02, -0.018], the second range is [-0.018,-0.016], the third range is [-0.016,-0.014], and so on. There is one decomposition coefficient falling into [-0.02,-0.018] (first range) as shown in Fig. 5.8(a). Fig. 5.8 illustrates the distribution of air corona and SF_6 PD at node (5,21) that is selected by the maximal class separability criterion. These distributions exhibit different shapes and distribution-related features associated with the decomposition coefficients at node (5,21) should be well separated.

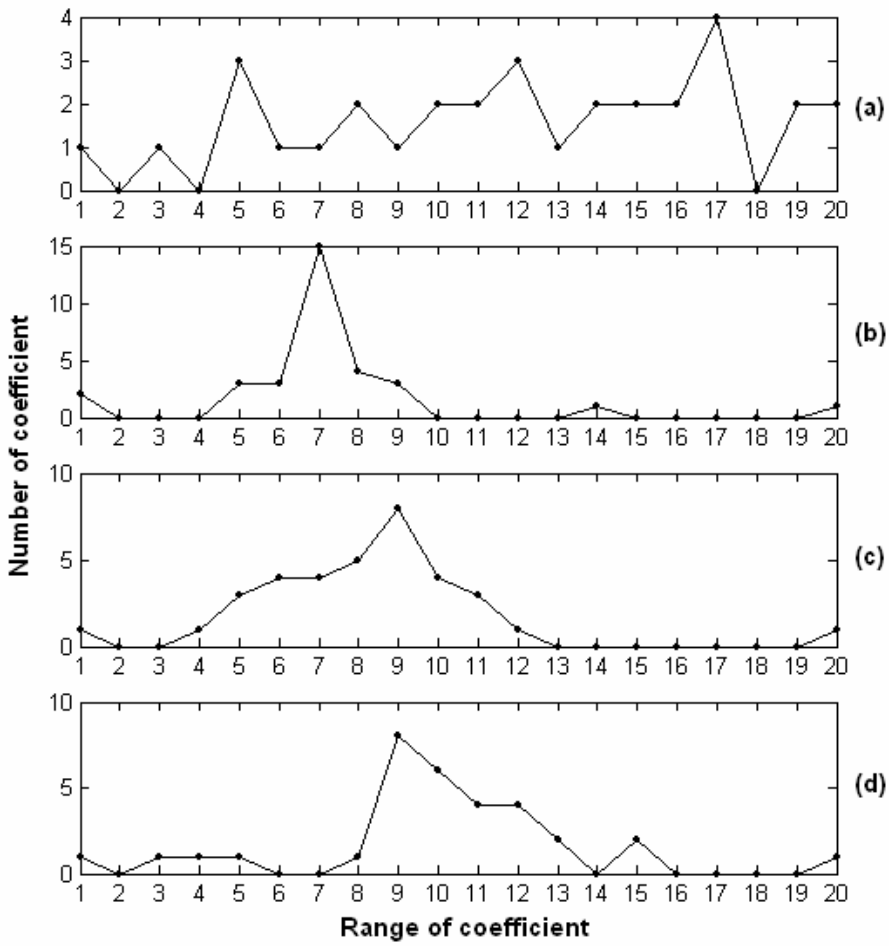


Fig. 5.8 Distribution of wavelet-packet-decomposition coefficients at node (5,21) corresponding to (a) air corona; (b) particle on the surface of spacer; (c) particle on conductor; (d) free particle on enclosure

Figs. 5.9 (a) and (b) show the kurtosis values of wavelet-packet-decomposition coefficients of SF_6 PD and air corona at node (5,21) and (4,15) respectively, while $J(5,21)_{kurtosis}$ is much larger than $J(4,15)_{kurtosis}$. As observed, the kurtosis values corresponding to "conductor", "spacer", "enclosure" and "corona" samples are well separated at node (5,21), and not as well separated at node (4,15). This justifies the use of J criterion for selecting the features.

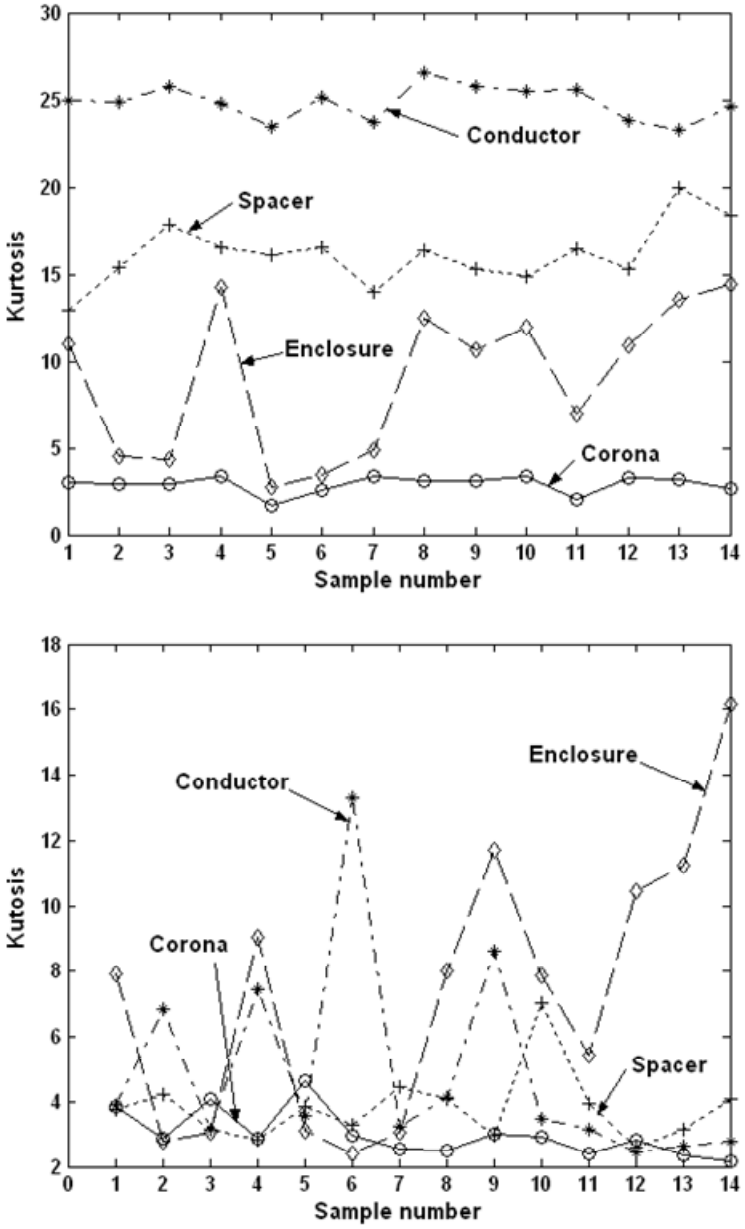


Fig. 5.9 Kurtosis values of wavelet-packet-decomposition coefficients of UHF signals (a) at node (5,21); (b) at node (4,15)

Figs. 5.10 and 5.11 demonstrate the feature clusters formed by the first and last two pairs of extracted features in two-dimensional spaces respectively. As observed, features in Fig. 5.10 are better separated than in Fig. 5.11 due to the greater J values of the first four features. In Figs. 5.11 (a) and (b), overlapping of feature clusters is observed, which indicates inferior classification performance. Thus, the use of J criterion value as the indicator of separability is verified.

Moreover, it is seen that the margin between feature clusters in Fig. 5.10 (a) is much larger than that of ICA-formed feature space as in Fig. 4.15 (a). This suggests that WPT-based method outperforms ICA-based method due to the additional frequency information. The effectiveness of selected features will be further studied in Chapters 6 and 7.

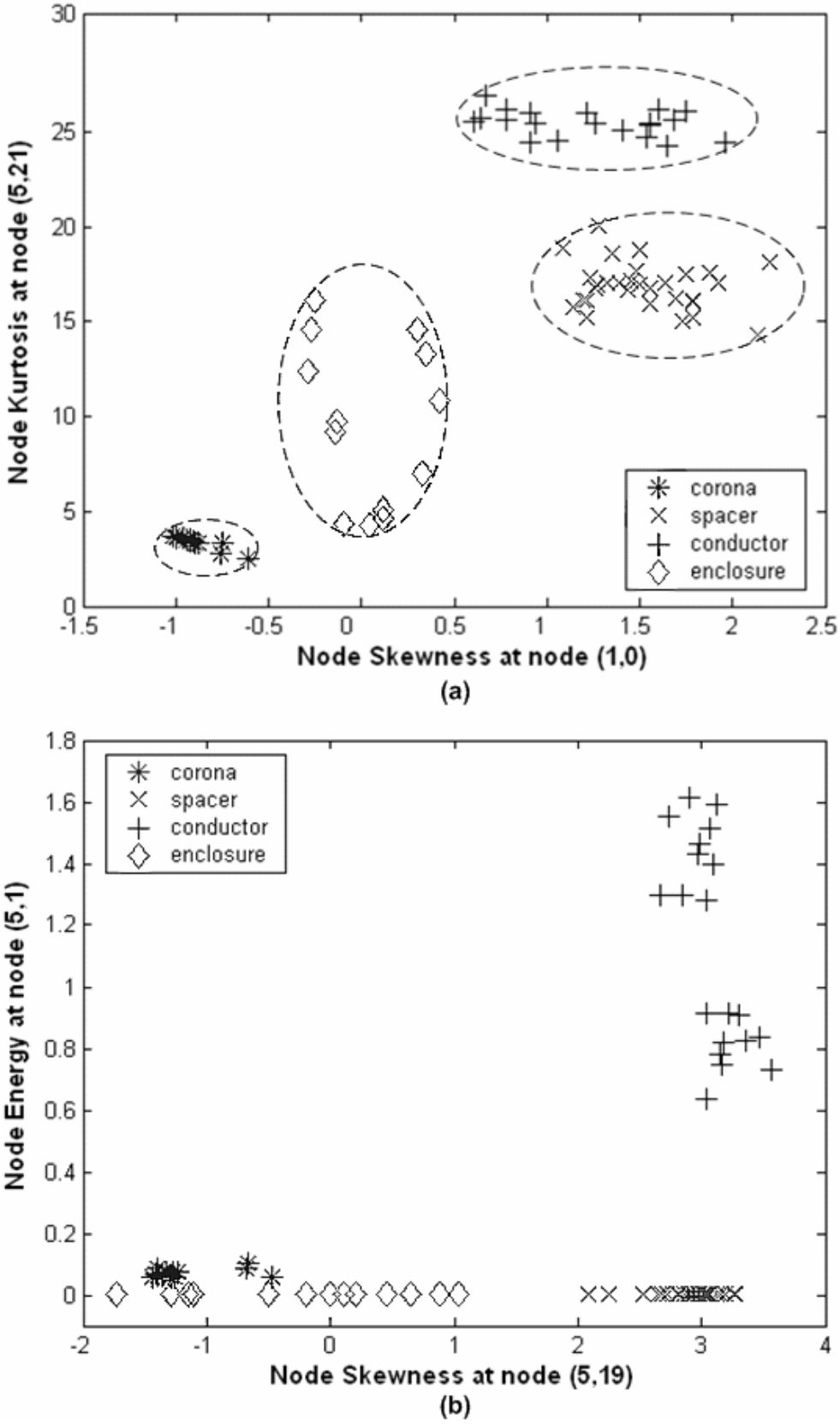
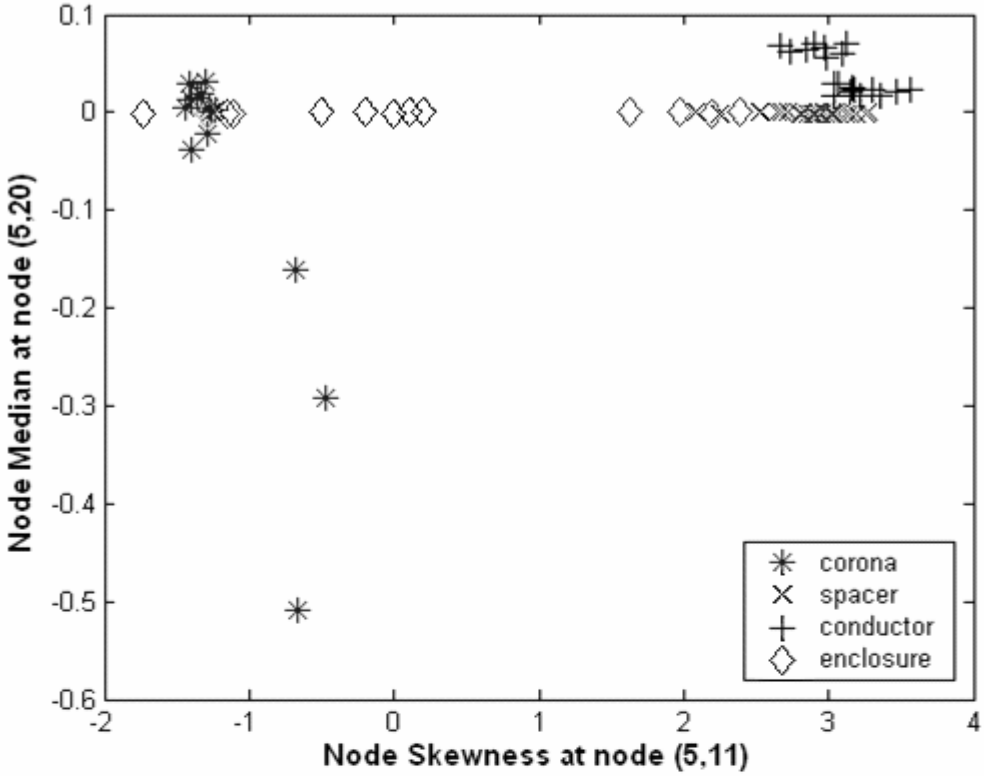
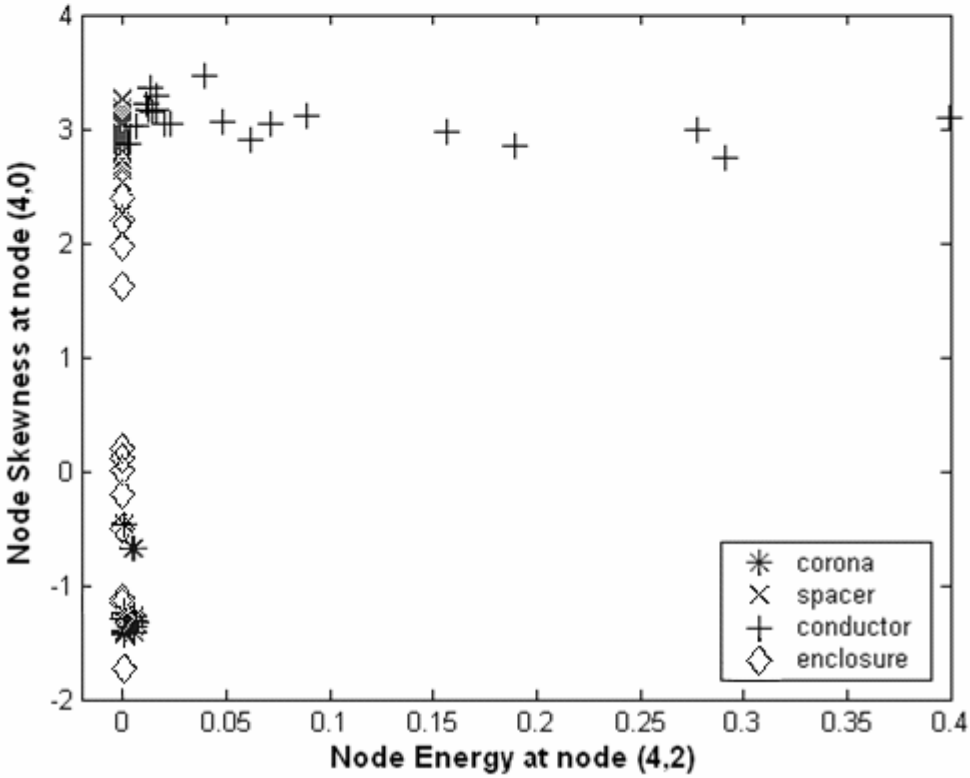


Fig. 5.10 Feature spaces formed by wavelet-packet-based method. (a) 1st and 2nd selected features; (b) 3rd and 4th selected features



(a)



(b)

Fig. 5.11 Feature spaces formed by wavelet-packet-based method (continue). (a) 7th and 8th selected features; (b) 9th and 10th selected features

5.4.2 Impact of Wavelet Selection

In Section 5.3.2, a method based on *J* criterion is employed to select the best wavelet for feature extraction. As a result, the “db9” wavelet is selected by the method for having the best discrimination ability. The impact of the choice of different wavelet filters on the effectiveness of selected features is further discussed in this section by comparative study.

Table 5.4 shows the best features obtained from “sym6” and “db9” wavelet. It can be seen that the wavelets result in the selection of completely different node features. Figs. 5.12 (a) and (b) further illustrate the feature spaces resulted from “sym6” and “db9” wavelet respectively. It can be seen that the features extracted by “sym6” are not as well-separated as those extracted by “db9”. This indicates that although “sym6” is the best wavelet for denoizing, it is not suitable for feature extraction. Thus, the use of *J* criterion is further verified as “sym6” gives a smaller J_{sum} value than “db9” as in Table 5.2.

Table 5.4 Features extracted by “sym6” and “db9”

wavelet	best features	*J* value	frequency range (Hz)
“sym6”	$(4,10)_{kurtosis}$	9.0327	1.25 G – 1.375 G
	$(4,2)_{skewness}$	8.9521	250 M – 375 M
“db9”	$(5,21)_{kurtosis}$	12.1435	1.3125 G – 1.375 G
	$(1,0)_{skewness}$	11.8492	0 – 1 G

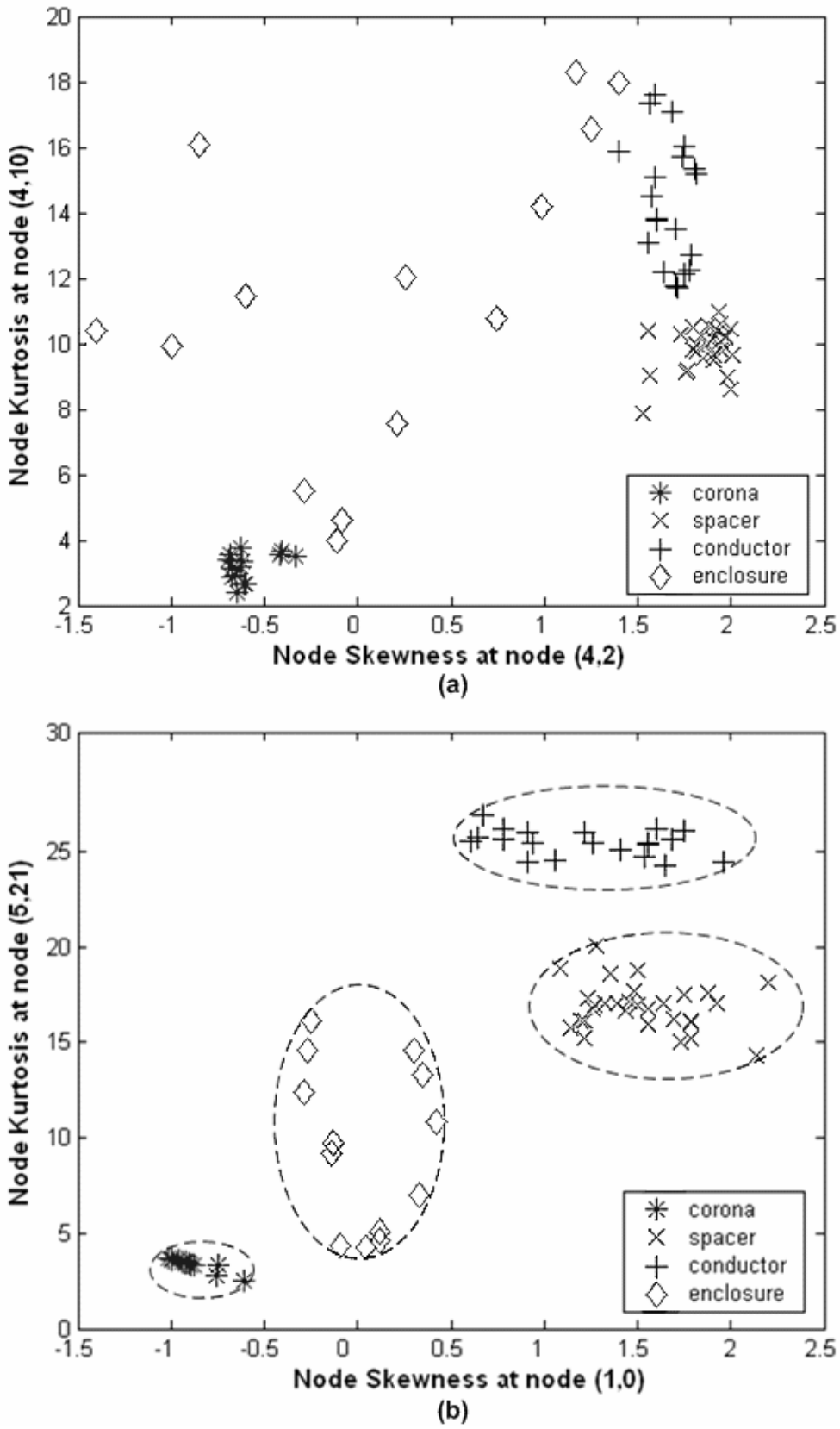


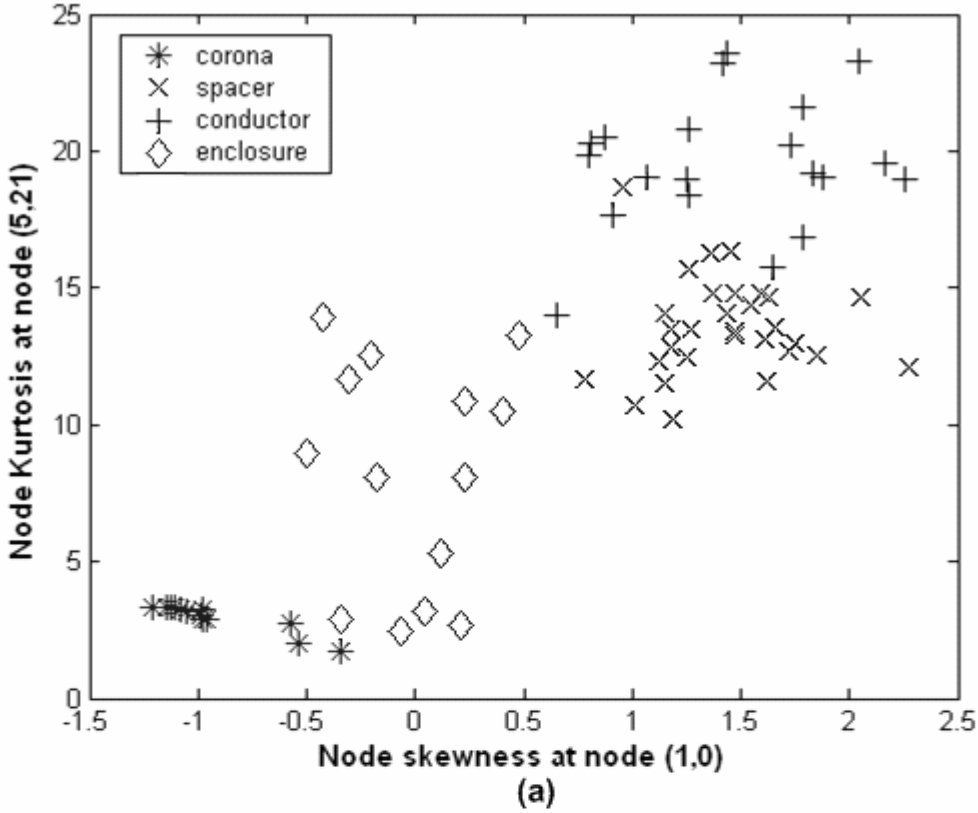
Fig. 5.12 Feature spaces formed by the best features obtained from (a) “sym6” wavelet; (b) “db9” wavelet

5.4.3 Need for Denoizing

The impact of background noise on the performance of wavelet-packet-based feature extraction is studied in this section.

Figs. 5.13 (a) and (b) illustrate the impact due to *medium* background-noise insertion (SNR=0) and *high* background-noise insertion (SNR=-5) on separability of the features, which have been extracted using the "db9" wavelet with denoized data (SNR=17). As shown in the feature clusters of [$(5,21)_{kurtosis}$, $(1,0)_{skewness}$], the features of different classes are seen to become more and more overlapped, as the noise level gets higher and higher.

To investigate the impact of noise levels on the feature extraction process, signals of different SNRs are employed for calculating node features and forming the feature spaces. As illustrated in Table 5.5, fewer features defined on high frequency band are selected when signals corrupted by high level noises are employed in the wavelet-packet-based feature extraction. This indicates that the node features computed from decomposition coefficients of high frequencies are more affected by noises. Furthermore, it is seen that the J values of obtained features are smaller than those in Table 5.3, where denoized signals are used. This suggests that denoizing improves discriminative ability of the extracted features.



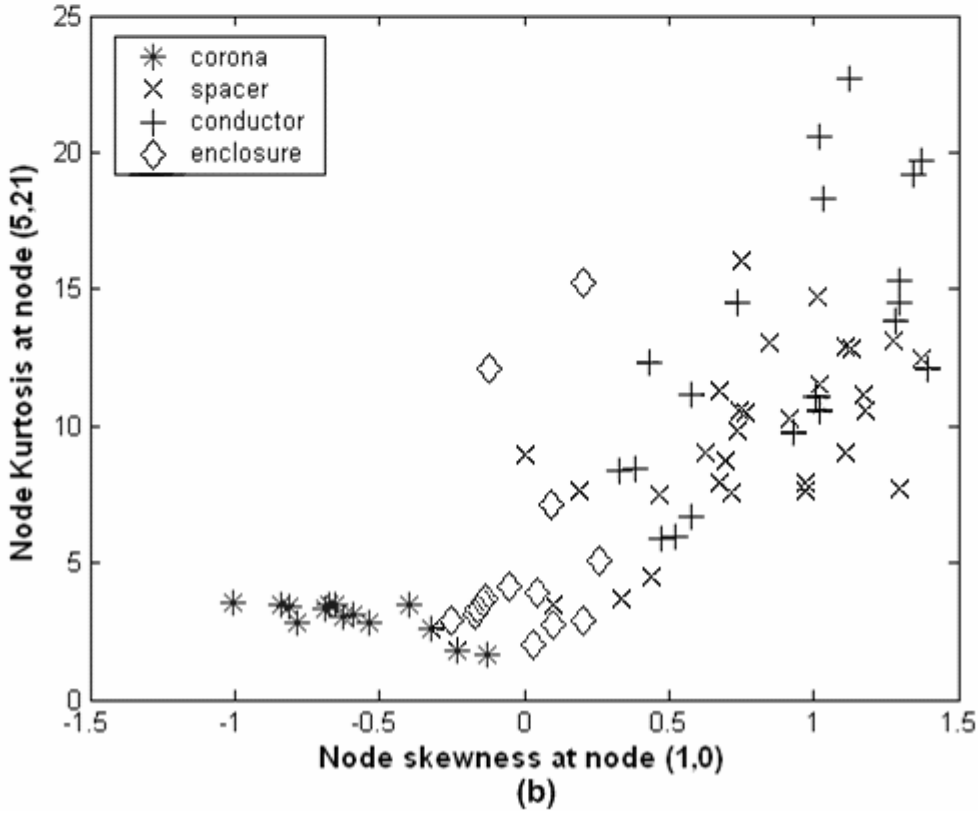


Fig. 5.13 Impact of noise levels on the features selected in Section 5.4.1. (a) SNR=0; (b) SNR=-5.

Table 5.5 Features extracted from signals of different SNR levels

Serial no.	SNR = 0		SNR = -5	
	feature	*J* value	feature	*J* value
1	$(1,0)_{skewness}$	9.5232	$(4,0)_{skewness}$	5.1781
2	$(5,0)_{kurtosis}$	8.5458	$(3,0)_{kurtosis}$	4.1951
3	$(5,1)_{energy}$	7.9701	$(2,0)_{kurtosis}$	4.1788
4	$(3,0)_{skewness}$	7.5755	$(5,1)_{energy}$	4.1379
5	$(4,0)_{skewness}$	6.6046	$(5,0)_{kurtosis}$	3.8543
6	$(3,0)_{kurtosis}$	4.965	$(2,0)_{skewness}$	3.479
7	$(5,4)_{kurtosis}$	4.8702	$(3,0)_{skewness}$	3.3888
8	$(5,21)_{kurtosis}$	4.8486	$(5,4)_{kurtosis}$	3.1177
9	$(4,2)_{energy}$	4.4572	$(5,0)_{energy}$	3.1093
10	$(2,0)_{kurtosis}$	4.1856	$(4,2)_{energy}$	3.0972

The feature spaces are then constructed using features with highest *J* values as highlighted in Table 5.5. Figs. 5.14 (a) and (b) show the best feature spaces obtained from signals with SNR levels of 0 and -5 respectively. It is seen that the features extracted from such signals are not well separated in both feature spaces. Furthermore, as the noise level gets higher, the quality of obtained feature clusters gets worse. Therefore, it is crucial to suppress white noises present in the measured signals before feature extraction and classification.

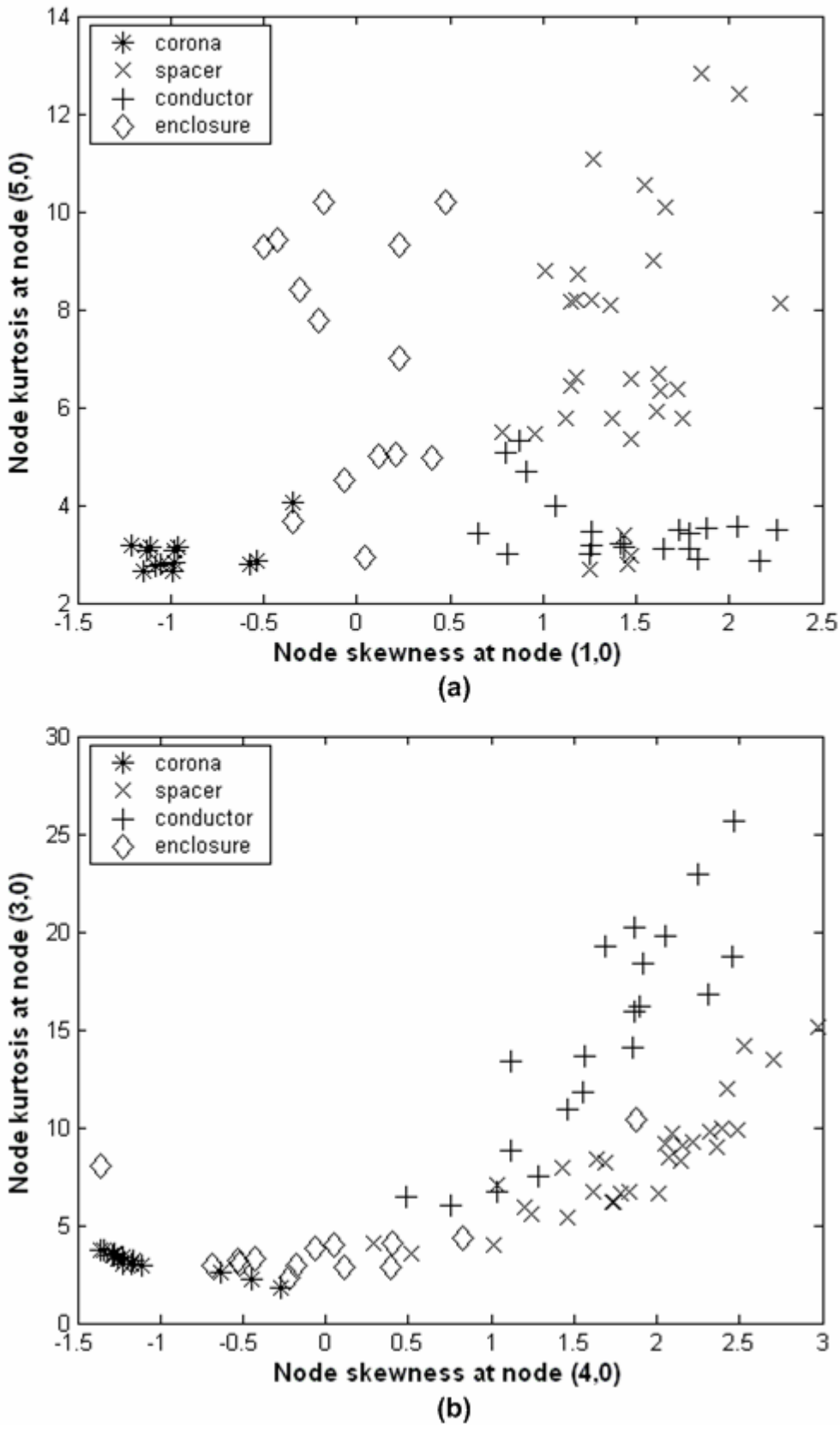


Fig. 5.14 Feature spaces obtained from signals of different SNR levels. (a) SNR=0; (b) SNR=-5.

5.4.4 Relationship between Node Energy and Power Spectrum

As each node in the WPD tree contains decomposition coefficients of certain frequency band, node energy represents energy of the corresponding frequency band in wavelet domain. Therefore, there is a need to clarify the relationship between energy in wavelet domain and in Fourier domain.

To investigate the relationship between node energy and energy in Fourier domain, the power spectrum of a PD signal of type “spacer” is first built using Fast Fourier Transform (FFT) as shown in Fig. 5.15. Subsequently, energy values in Fourier domain are calculated for 62 frequency bands corresponding to the nodes of WPT tree. They are computed from the power spectrum by summing up the square of FFT coefficients of each frequency band, forming *FFT_energy* (1*62). *FFT_ energy* is then compared with node energy that is computed from wavelet-packet-decomposition coefficients (Section 5.2.2 C). As illustrated in Fig. 5.16, node energy is almost the same as *FFT_ energy*. Therefore, it can be concluded that the Fourier domain energy analysis is equivalent to node energy analysis, which is seen to be not sufficient for PD identification as shown in Fig. 5.10 (b). The time-frequency information equipped with wavelet packet transform is thus crucial for the current study.

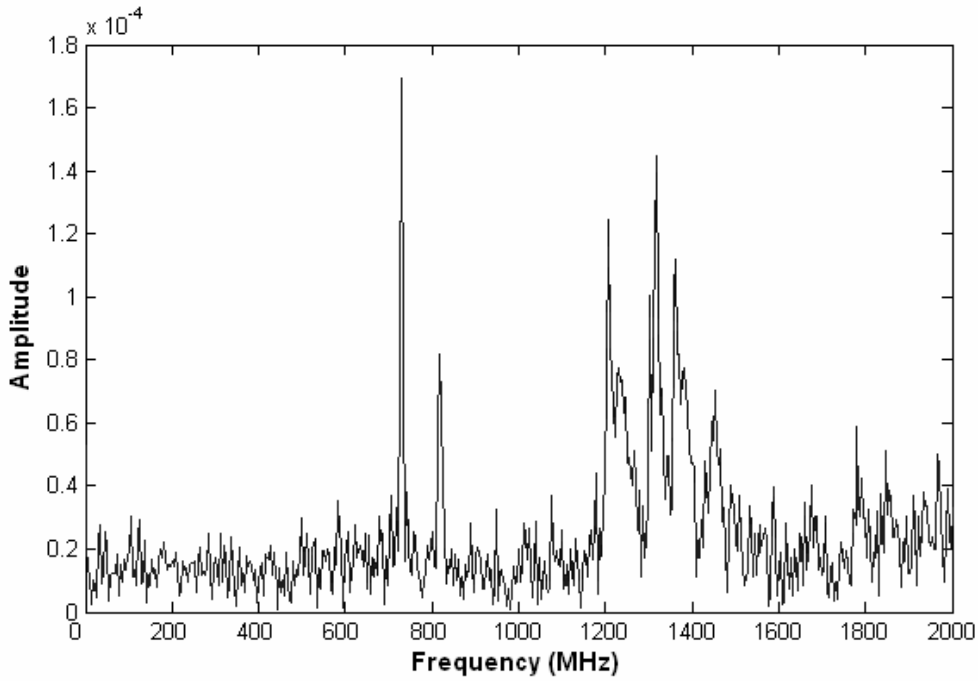


Fig. 5.15 Power spectrum obtained from FFT

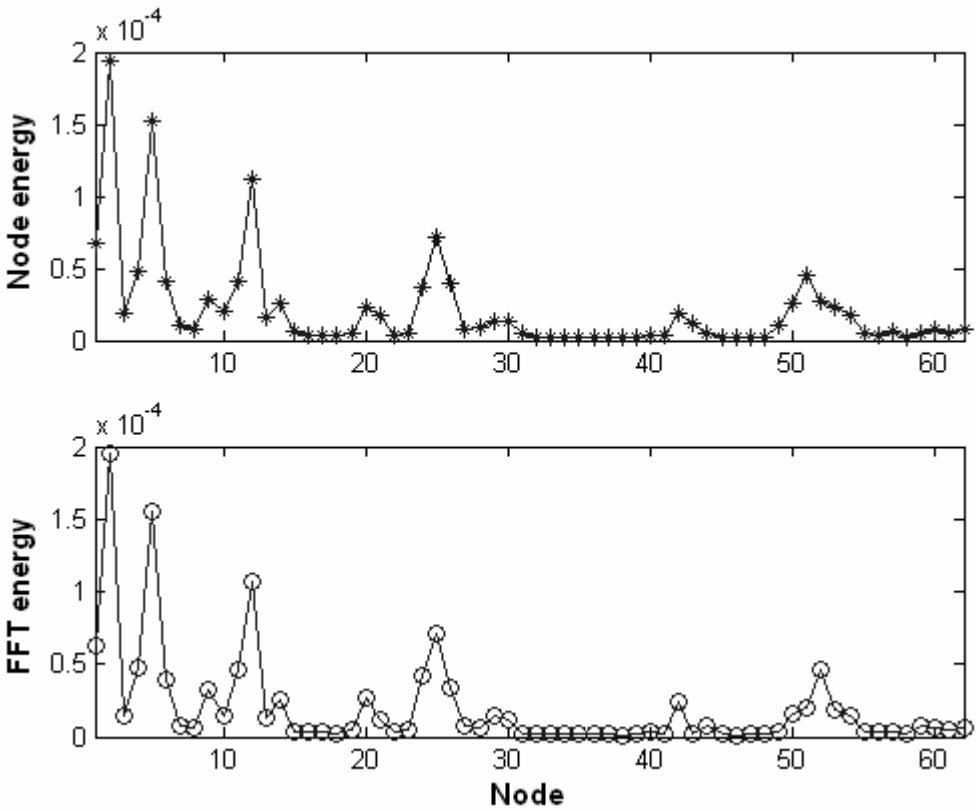


Fig. 5.16 Comparison of node energy and *FFT_energy*

5.5 CONCLUDING REMARKS

This chapter proposes a novel wavelet-packet-based feature extraction method to tackle the difficulties encountered by ICA-based time domain method. Results show that the feature clusters formed by the wavelet-packet-based method exhibit much larger between-class margin than ICA-based method, which indicates a better classification performance.

Comparative studies on features extracted from data with different noise levels show that high level of white noises worsens the performance of the features. Among features derived from decomposition coefficients, distribution-shape-related node features are seen to be more effective than the other node features, such as node energy. Further investigation of the relationship between node energy and power spectrum reveals that Fourier domain energy analysis is equivalent to node energy analysis. Thus, it can be concluded that wavelet-packet-based method outperforms methods solely in time or frequency domain due to its time-frequency characteristics.

CHAPTER 6

PARTIAL DISCHARGE IDENTIFICATION USING NEURAL NETWORKS

In previous chapters, high quality partial discharge features, namely *ICA_feature* and *WPT_feature*, have been established from UHF signals through denoizing and feature extraction. Based on the feature clusters as illustrated in Fig. 5.10 (a), PD identification can be performed by experienced engineers. However, it is difficult to evaluate the measured data by humans when the database gets larger and larger. On the other hand, it has been found that the artificial neural networks perform more effective and reliable classification than engineers, especially when multilayer perceptron (MLP) neural network is employed [23, 26, 72]. Thus, a MLP neural network with a back-propagation (BP) learning rule is implemented in this chapter to automatically classify a new set of measured data among SF_6 PD and air corona. Firstly, training and test of the MLP is studied with discussions on the network parameters selection. Subsequently, the usefulness and effectiveness of the extracted features are proved by results of comparative studies.

6.1 CLASSIFICATION USING MLP NETWORKS

In the past decades, several network architectures such as multilayer perceptron [26], self-organizing map [70] and modular neural network [71] have been adopted to classify PD sources of different types. In [72], three different types of neural networks, namely multilayer perceptron, self-organizing map and learning vector quantization network are studied and compared. In this study, multilayer perceptron (MLP) is chosen due to its proven powerfulness and effectiveness for PD classification [72].

A brief introduction to MLP networks is first given in this section. Subsequently, the construction and training of MLP are discussed. Lastly, the generalization issue of MLP networks is studied.

6.1.1 Brief Introduction to MLP

A multilayer perceptron is a network of simple neurons called perceptrons. MLP consists of an input layer, one or more hidden layers and an output layer of neurons, which perform the processing tasks through a nonlinear activation function. Each neuron has many inputs but only one output that is applied to every neuron in the next layer. Each connected pair of neurons is associated with an adjustable *weight*. The MLP network is trained using the back-propagation algorithm, which modifies the weights to get desired output by means of the gradient search technique.

There are three distinctive characteristics of the multilayer perceptron:

1. There is a nonlinear activation function associated with each neuron and the function must be smooth. The presence of nonlinearities is important because otherwise the input-output relation of the network could be reduced to that of a single-layer perceptron.

2. The network contains one or more layers of hidden neurons, which enable the network to learn complex tasks by extracting progressively more meaningful features from the input vectors.

3. The neurons are fully interconnected so that any element of a given layer feeds all the elements of the next layer.

It is through the combination of these characteristics together with the ability to learn from experience through training that the MLP derives its computing power. A review of MLP is given in [66].

6.1.2 Constructing and Training of MLP

To achieve the best classification performance, MLP must be properly constructed and trained with a suitable algorithm. The parameters to be determined when constructing and training a MLP include number of hidden layers, type of neuron, number of neurons in input, hidden and output layer, training algorithm and training stopping criteria. The selection of these parameters has significant impact on the performance of MLP network. Thus, details of selecting these parameters are discussed in this and next section.

A. Number of Hidden Layers

In general, the more hidden layers MLP contains, the more powerful the MLP is. However, too many hidden layers will slow down the MLP. In addition, unnecessarily large number of hidden layers may result in overfitting to the training data, which could lead to a bad classification performance on new data [66]. On the other hand, as the PD classification problem has been significantly simplified by using the extracted features, MLP with one hidden layer is seen to be powerful enough for current application. Thus, the number of hidden layers is set to one.

B. Number of Neurons in Input, Hidden and Output Layer

In this study, the classification problem involves four classes, namely “spacer”, “conductor”, “enclosure” and corona. Therefore, the number of output neurons is set to two to represent all the classes as shown in Table 6.1. Since the outputs of MLP rarely give exactly the target of 0 or 1 on each output neuron, the PD pattern is deemed to have been correctly classified if the error on each output neuron is within 0.2. For instance, if the output of MLP is (0.88, 0.15) when a signal of particle on conductor is presented (ideally the output should be (1,0)), it is treated as correctly classified.

The number of neurons in input layer equals to the number of features used as the input of MLP. Therefore, it is determined in Section 6.3 by comparative studies on the performance of using different number of extracted features.

As the number of neurons in hidden layer is closely related to the generalization issue of MLP, it will be discussed in the next section.

Table 6.1 Representing four classes by two output neurons

Classes	Output of 1st neuron	Output of 2nd neuron
Corona	0	0
Spacer	0	1
Conductor	1	0
Enclosure	1	1

C. Type of Neuron

The type of a neuron is characterized by the type of activation function used in the neuron. There are three functions commonly employed in MLPs, namely log-sigmoid, tan-sigmoid and the linear function as shown in Fig. 6.1. For this study, the log-sigmoid function is preferred as the relationship between input and output of MLP is nonlinear and output of 0 or 1 is expected on the neurons in output layer. Thus, log-sigmoid type neurons are employed in all of the layers.

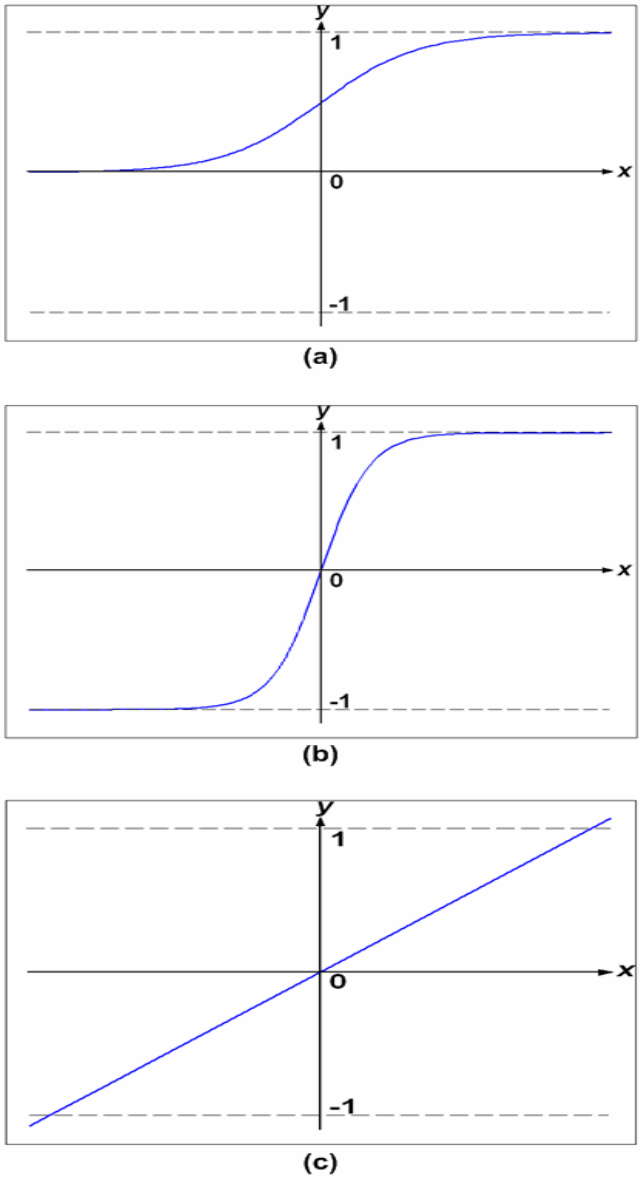


Fig. 6.1 Activation functions. (a) log-sigmoid; (b) tan-sigmoid; (c) linear.

D. Training Algorithms

There are quite a few back-propagation algorithms available to be used to train the MLP. Table 6.2 shows the algorithms compared in this study. A comprehensive review of these algorithms is given in [73].

Table 6.2 Training algorithms

Algorithms	Description
Basic gradient descent (traingd)	Weights and biases are updated in the direction of the negative gradient of the performance function.
Gradient descent with momentum (traingdm)	A variation of the basic gradient descent algorithm. Momentum allows the network to ignore small features in the error surface. Thus, it prevents the network from getting stuck in a local minimum.
Adaptive learning rate (traingda)	Another variation of the basic gradient descent algorithm. The learning rate changes during the training.
Adaptive learning rate with momentum (traingdx)	A combination of adaptive learning rate and momentum.
Resilient back-propagation (trainrp)	The sign of the gradient is used to determine the direction of the weight update. The size of the weight update changes according to the sign of gradient for successive iterations.
Conjugate gradient (trainscg)	Weight update is performed along conjugate direction.
Quasi-Newton (trainbfg)	An alternative to the conjugate gradient method. It often converges faster than conjugate gradient method.
Levenberg-Marquardt (trainlm)	A variation of Quasi-Newton method.

Fig. 6.2 compares the convergence performance of the training algorithms. It can be seen that MLP is not able to converge within 1000 epochs when trained with ‘traingd’, ‘traingdm’ and ‘traingda’. On the other hand, the resilient back-propagation (‘trainrp’) algorithm is seen to achieve the best convergence and thus adopted in this study. Details of the resilient back-propagation algorithm are given in Appendix E.

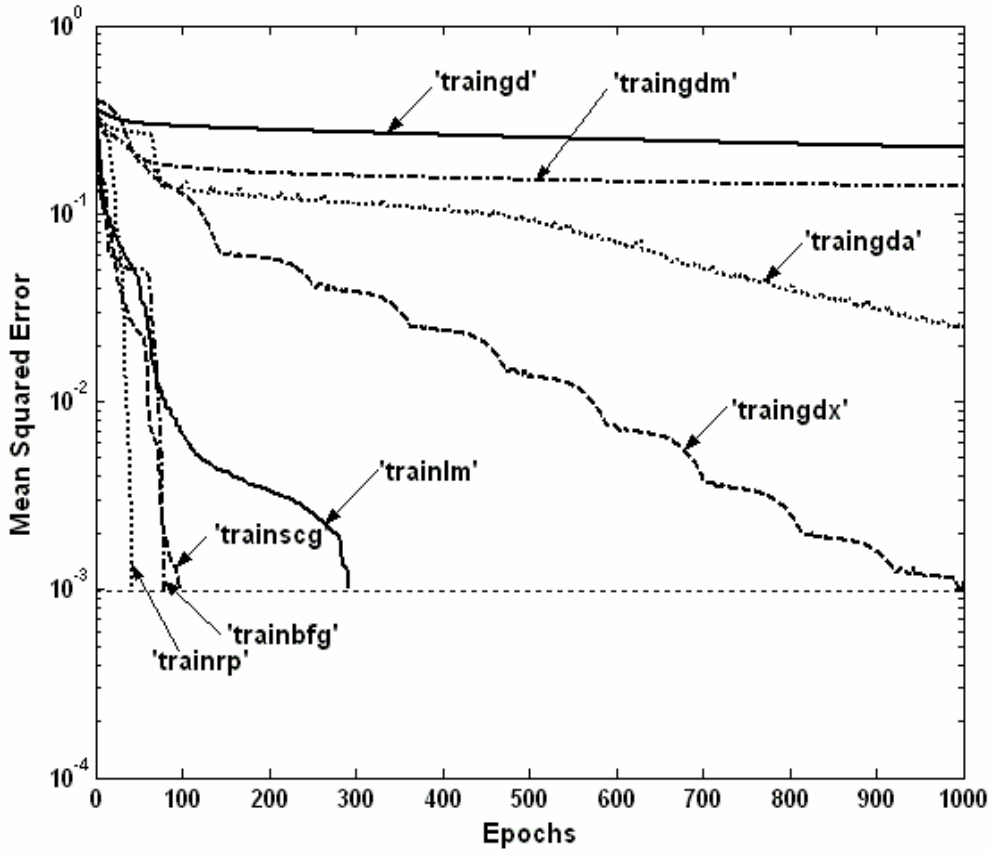


Fig. 6.2 Performance of training algorithms

E. Training Stopping Criteria

Training of the MLP stops when either of the following criteria is met.

(1) When the maximum number of iterations is reached. It is set to 1000 in this study.

(2) When the mean squared error (MSE) between the network outputs and the target outputs drops below the goal, which is set to 0.01 in this study.

F. The Used MLP

To perform PD identification, a three-layer (one hidden layer) MLP network with a back-propagation training algorithm known as resilient back-propagation is adopted to achieve fast convergence during training. Fig. 6.3 shows the structure of the used MLP.

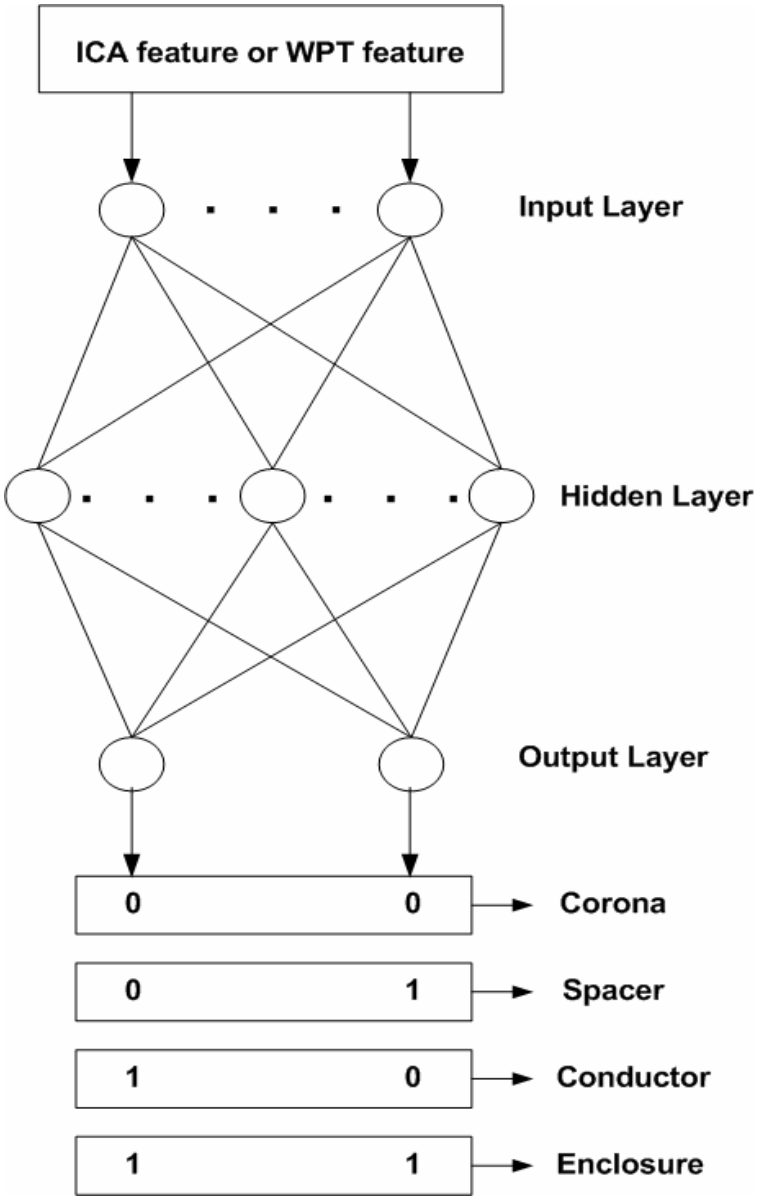


Fig. 6.3 Three-layer MLP for classification

After extensive studies, the configuration of the MLP network is set as in Table 6.3. It can be seen that a very simple MLP is able to perform PD identification successfully due to the high quality of the extracted features.

Table 6.3 Parameters of the used MLP

Parameters	Setting
Type of neuron	Log-sigmoid
Number of neurons in output layer	2
Number of neurons in input layer	2 (when *ICA_feature* is used)
	3 (when *WPT_feature* is used)
Number of neurons in hidden layer	5 (when *ICA_feature* is used)
	7 (when *WPT_feature* is used)

6.1.3 Generalization Issue of MLP

The objective of designing a neural network classifier is to achieve correct classification of new data after training. Therefore, it is crucial to ensure minimum generalization errors when designing the MLP. Generalization is influenced by three factors:

(1) the size and dimension of the training set,

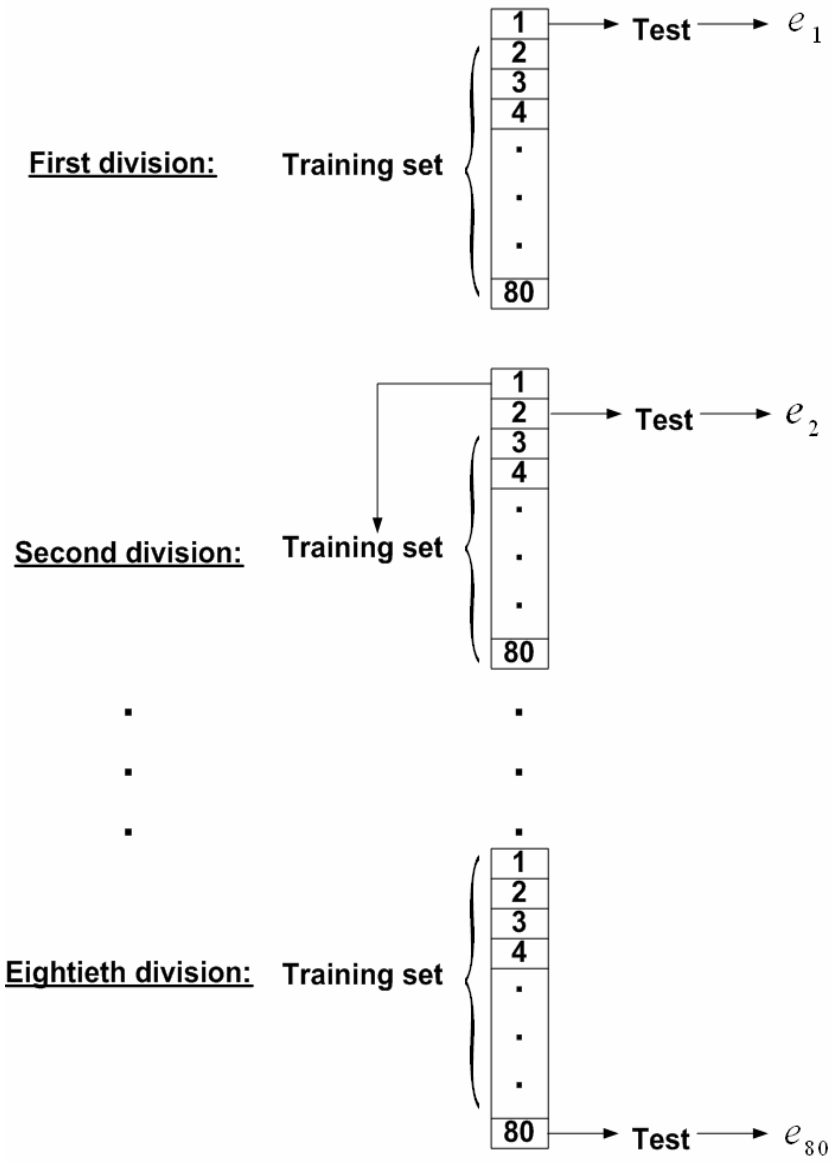
(2) the architecture of the neural network, and

(3) the physical complexity of the problem at hand [66].

Clearly, the third factor is application-oriented. As far as the first factor is concerned, an effective feature extraction, such as the ICA-based or WPT-based schemes, will ensure good generalization by reducing the length of each training vector in the training set.

The extracted feature set (*ICA_Feature* or *WPT_feature*) is usually divided into two sets for determining the weights during the MLP training and estimation of generalization error during testing. One way of forming the training and test sets is to randomly divide the ensemble into two sets. A better method for estimating the generalization error, known as "leave-one-out", is chosen to avoid the possible bias introduced by relying on any particular test or training set after division. The method is chosen because it maximizes the size of the training set by employing all the 80*N (N denotes the length of each feature vector) data for training the MLP weights.

As illustrated in Fig. 6.4, the method first splits the feature set (size of 80*N) into a training set (size of 79*N) and a test set (size of 1*N). Then the MLP is trained using the 79*N training set and tested with the 1*N test set. The mean squared error on test set is calculated and denoted as e_1. The above process is then applied to all the other combinations of training and test sets. As a result, 80 values of mean squared errors (e_1, $e_2 \cdots e_{80}$) of the test sets are obtained. Subsequently, the generalization error E_{test} is calculated by averaging (Fig. 6.4). Once the generalization error is computed, training is re-applied on the 80*N data set to determine the MLP weights.



$$E_{test} = \frac{e_1 + e_2 + \ldots + e_{80}}{80}$$

Fig. 6.4 Illustration of the “leave-one-out” approach

Generalization of MLP also depends on the number of neurons in the hidden layer. If there are not enough neurons in the hidden layer, the MLP network may not have sufficient discriminative power to correctly classify the signals. On the other hand, if too many neurons are used in hidden layer, the MLP may overfit the training data, leading to large error on the new data. Therefore, experiments are also carried out with

different numbers of hidden neurons. The number, which gives the smallest generalization error, is chosen for classification (Section 6.3).

6.2 RESULTS AND DISCUSSIONS

Experimental results using various features as input of MLP are presented and compared. Determination of the best MLP network structure is investigated by comparative studies.

6.2.1 Using Pre-selected Signals as Input

To justify the effectiveness of the feature extraction schemes, classification performance of MLP that uses the pre-selected signals as input is first studied. Without performing feature extraction, the number of input neurons is the same as the length of pre-selected signal, namely 1000.

The best number of hidden neurons is chosen according to the minimum generalization error calculated by the “leave-one-out” method as described in Section 6.2.3. Table 6.4 summarizes the results obtained from using different number of hidden neurons. The generalization error obtained from using different number of hidden neurons is shown in Fig. 6.5. It can be seen that the MLP with 14 hidden neurons offers the best generalization performance with respect to both the mean squared error and number of misclassified patterns. Even in the best case, however, there are still seventeen patterns out of eighty not classified correctly during testing.

After determining the structure of MLP, it is trained using all the 80*1000 data. As illustrated in Fig. 6.6, the training converges in 70 epochs, taking 58.6 seconds on Pentium-IV.

Table 6.4 Generalization performance of MLP using pre-selected signals as input

Number of neurons in hidden layer	Averaged convergence epochs	Generalization mean squared error	Number of Misclassified patterns on test
2	1281	0.0669	23/80
4	161	0.0467	20/80
6	85	0.0434	19/80
8	85	0.0396	19/80
10	82	0.0387	18/80
12	79	0.0369	17/80
14	**73**	**0.0320**	**17/80**
16	63	0.0375	17/80
18	59	0.0382	17/80
20	51	0.0396	18/80
22	48	0.0386	18/80
24	47	0.0401	18/80
26	45	0.0421	19/80
28	49	0.0392	18/80

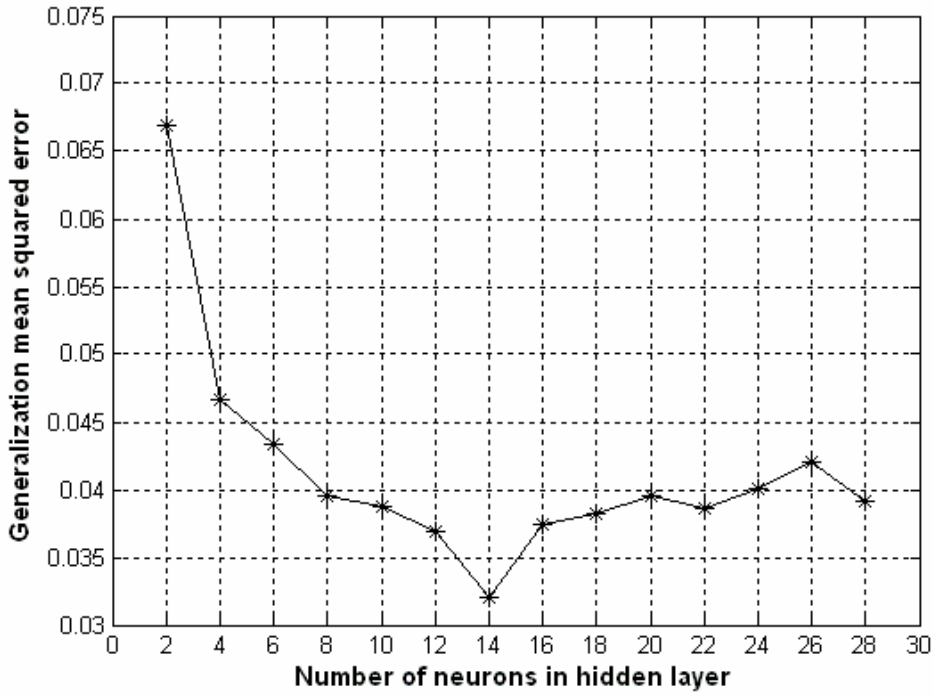


Fig. 6.5 Generalization error of using pre-selected signals as input

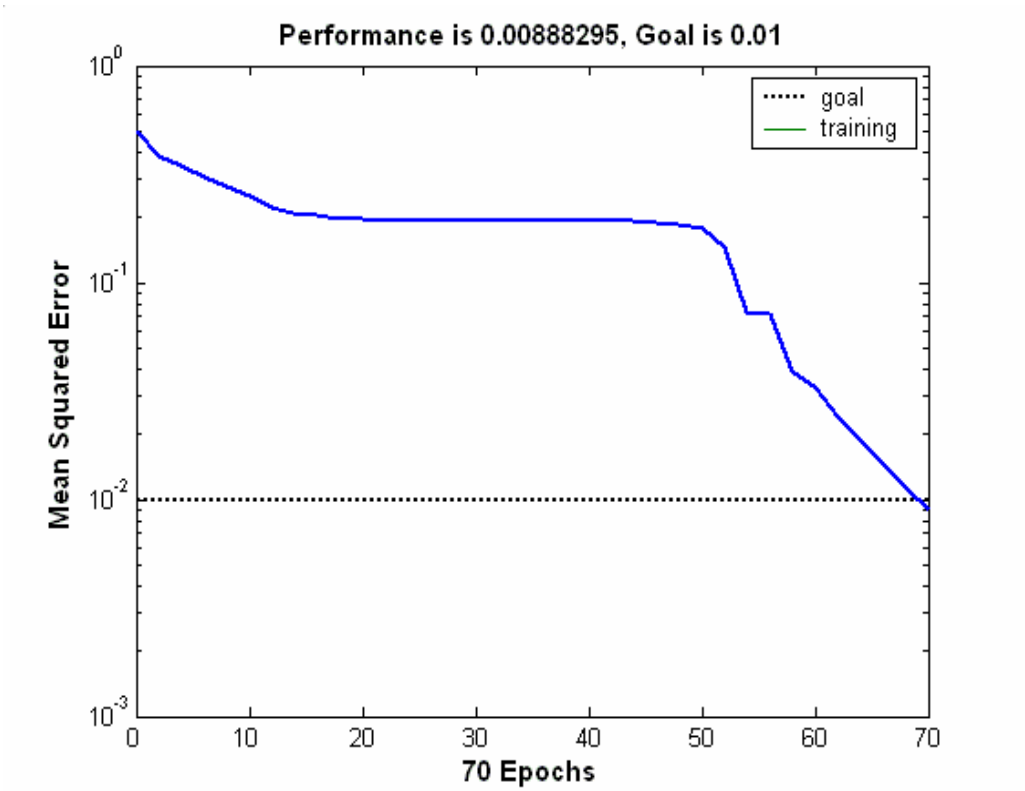


Fig. 6.6 Mean squared error during training when using pre-selected signals as input

6.2.2 Using ICA_feature as Input

Using *ICA_feature* as input, the MLP has two input neurons, which correspond to the two most dominating independent components. The impact of number of hidden neurons is summarized in Table 6.5. The generalization error of using *ICA_feature* is illustrated in Fig. 6.7. As observed, the best generalization performance is achieved when the number of hidden neurons is set to 5. In the best case, there are two patterns misclassified on test set, which is much better than the result obtained from using pre-selected signals without data compression. In addition, misclassification only occurs among SF_6 PD. There is no pattern of corona misclassified as SF_6 PD, and vice versa.

Table 6.5 Generalization performance of MLP using *ICA_ feature* as input

Number of neurons in hidden layer	Averaged convergence epochs	Generalization mean squared error	Number of Misclassified patterns on test
2	408	0.0522	11/80
3	125	0.0284	5/80
4	101	0.0223	3/80
5	**85**	**0.0121**	**2/80**
6	80	0.0156	2/80
7	78	0.0145	2/80
8	75	0.0230	3/80
9	72	0.0175	2/80
10	72	0.0234	3/80
11	70	0.0219	3/80
12	73	0.0258	4/80
13	70	0.0218	3/80
14	69	0.0245	3/80
15	71	0.0229	3/80

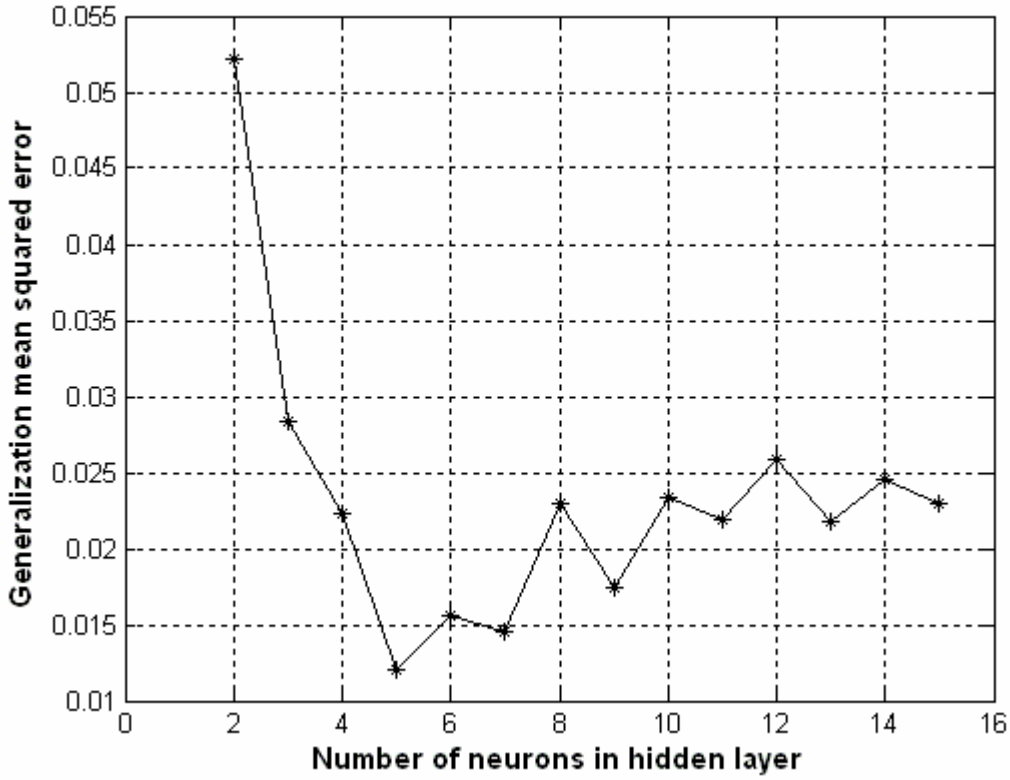


Fig. 6.7 Generalization error of using *ICA_feature* as input

Using the 80*2 feature set, training of the MLP converges in 82 epochs as shown in Fig. 6.8, which takes one second on Pentium-IV.

The performance of using additional independent components (>2) is also studied and the results are summarized in Table 6.6. It can be seen that using additional independent components does not seem to improve the performance of the MLP in terms of speed and classification accuracy due to the dominance of the two most dominating independent components.

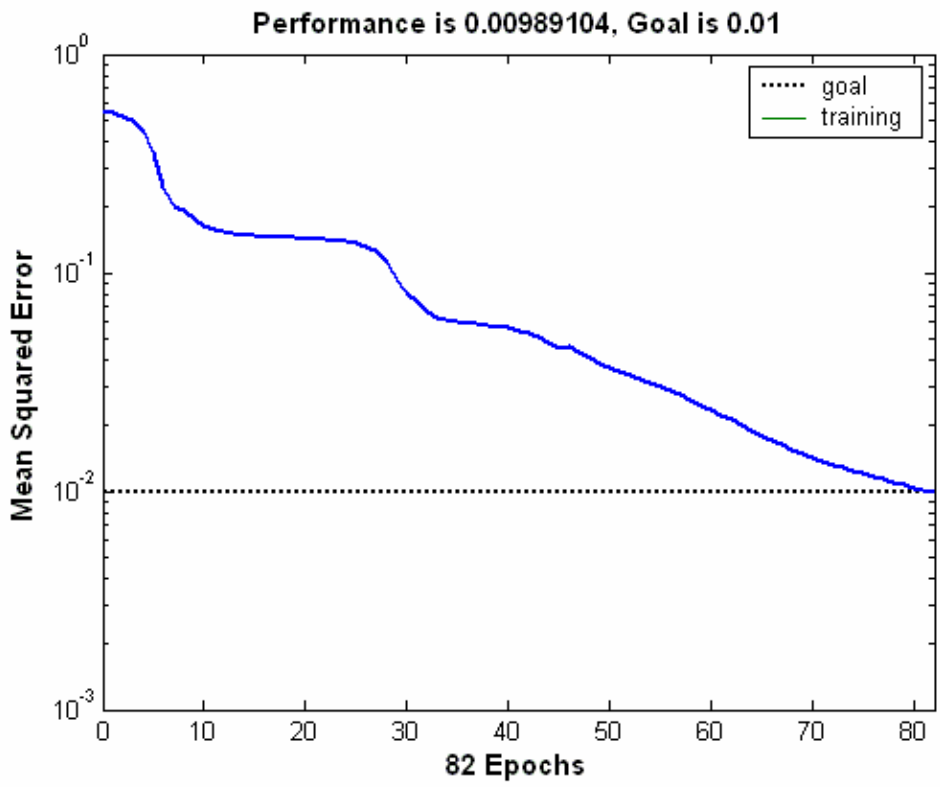


Fig. 6.8 Mean squared error during training when using *ICA_feature* as input

Table 6.6 Performance of using more independent components

Number of used independent components	Number of neurons in input layer	Best number of neurons in hidden layer	Training convergence time (s)	Generalization MSE	Number of Misclassified patterns on test
3	3	5	1.26	0.0146	2/80
4	4	5	1.51	0.0139	2/80
5	5	9	1.69	0.0136	2/80
6	6	5	1.37	0.0181	3/80
7	7	7	1.35	0.0130	2/80
8	8	5	1.83	0.0203	3/80

6.2.3 Using WPT_Feature as Input

Based on comparative studies, the number of input neurons of MLP is set to four, which corresponds to the first four WPT features, namely $(5,21)_{kurtosis,}$ $(1,0)_{skewness}$, $(5,1)_{energy}$ and $(5,19)_{skewness}$. Table 6.7 shows the generalization performance of various network structures using the first four *WPT_feature* as the network input. As illustrated in Fig. 6.9, the best generalization performance is achieved when the hidden layer consists of seven neurons. In this case, minimal-mean-squared error is achieved and no pattern of test set is misclassified.

Table 6.7 Generalization performance of MLP using the first four *WPT_feature*

Number of neurons in hidden layer	Averaged convergence epochs	Generalization mean squared error	Number of Misclassified patterns on test
2	408	0.0236	3/80
3	152	0.0221	3/80
4	70	0.0118	1/80
5	64	0.0114	0/80
6	45	0.0115	0/80
7	**41**	**0.0098**	**0/80**
8	39	0.0102	0/80
9	37	0.0116	1/80
10	34	0.0110	0/80
11	31	0.0114	0/80
12	30	0.0112	0/80
13	29	0.0115	0/80
14	28	0.0112	0/80
15	28	0.0106	0/80

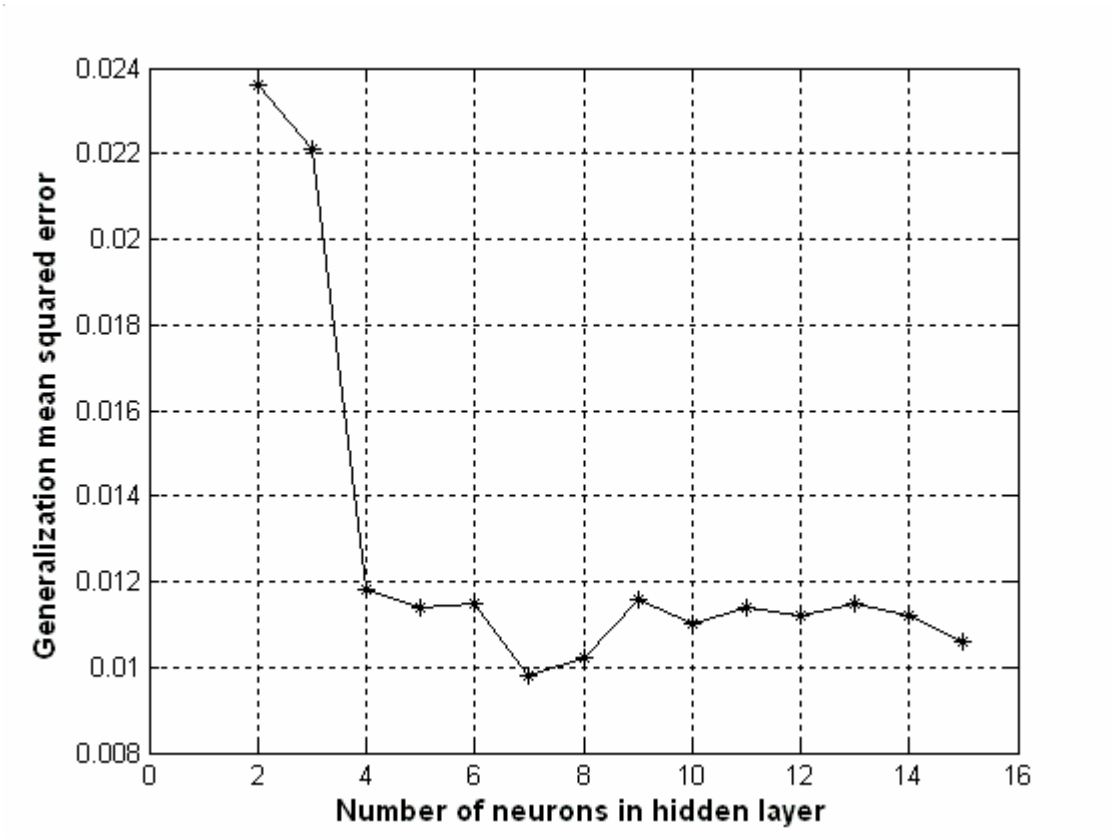


Fig. 6.9 Generalization error of using *WPT_feature* as input

Using the 80*4 feature set, training of the MLP converges in 40 epochs as shown in Fig. 6.10. It takes 1.02 second on Pentium-IV.

The performance of using different number of WPT features as input is also studied. The MLP is not able to converge during training when only one feature is used as the input of MLP. Thus, at least two features are required to classify PD. Table 6.8 shows the classification performance of using two features chosen from Table 6.2 as the input of MLP. It can be seen that the features with higher J values result in better classification. This verifies the use of J criterion for selecting the most effective features.

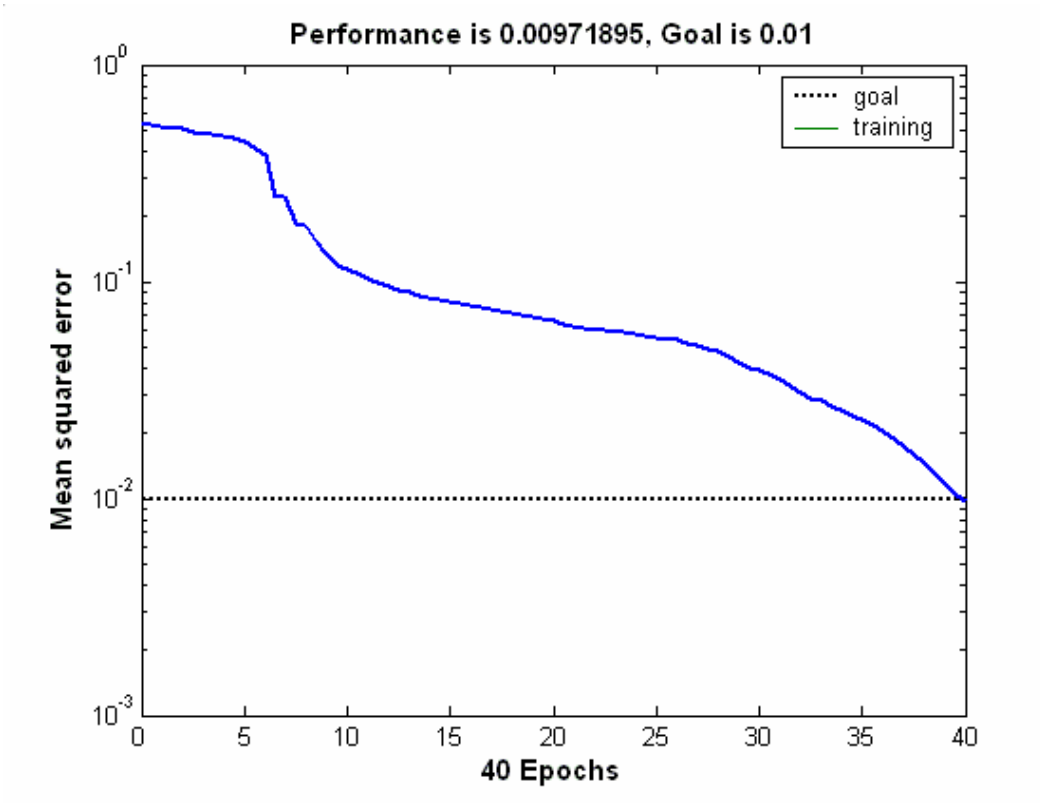


Fig. 6.10 Mean-squared error during training when using *WPT_feature* as input

Table 6.8 Classification performance of features in Table 6.2

Input of MLP	Training convergence time (s)	Generalization MSE	Number of Misclassified patterns on test
1st & 2nd feature	0.95	0.0111	0/80
3rd & 4th feature	1.14	0.0115	0/80
5th & 6th feature	5.344	0.0118	1/80
7th & 8th feature	18.872	0.0230	3/80
9th & 10th feature	22.094	0.0280	4/80

The effectiveness of additional features is investigated as shown in Table 6.9. Using the first two features in Table 6.2 as the benchmark, the performance of adding other features is evaluated by the improvement of generalization. It is seen that only the third and fourth features that have large J values improve the classification performance. Therefore, the J value of the fourth feature (=8.5927) is defined as the critical J value (J_{cr}) to determine the effectiveness of a feature. Table 6.10 shows the performance of using different number of WPT features as input. In coincidence with the results in Table 6.9, the first four features leads to the best performance in terms of generalization MSE as highlighted. Using additional features does not seem to improve the performance of the MLP. Therefore, the first four features in Table 6.2 are selected for PD classification.

Table 6.9 Performance improvement by the additional feature

Additional input of MLP	*J* value of the additional feature	Generalization MSE	Improvement of generalization MSE	Number of Misclassified patterns on test
3rd feature	8.6909	0.0098	0.0013	0/80
4th feature	**8.5927**	**0.0102**	**0.0009**	**0/80**
5th feature	8.0492	0.0113	-0.0002	0/80
6th feature	7.7291	0.0114	-0.0003	0/80
7th feature	7.5266	0.0114	-0.0003	0/80
8th feature	6.6111	0.0115	-0.0004	0/80
9th feature	6.5075	0.0115	-0.0004	0/80
10th feature	6.1892	0.0117	-0.0006	0/80

Table 6.10 Performance of using different number of WPT features

Number of WPT features	Number of neurons in input layer	Best number of neurons in hidden layer	Training convergence time (s)	Generalization MSE	Number of misclassified patterns on test
2	2	5	0.95	0.0111	0/80
3	3	7	1.04	0.0098	0/80
4	**4**	**7**	**1.02**	**0.0096**	**0/80**
5	5	6	1.005	0.0112	0/80
6	6	5	1.036	0.0113	0/80
7	7	6	1.005	0.0112	0/80
8	8	10	1.12	0.0113	0/80
9	9	11	1.088	0.0114	0/80
10	10	9	1.026	0.0113	0/80

6.2.4 Performance Comparison

Table 6.11 compares the performance of using different type of PD features as input of MLP. As observed, both speed and the generalization performance are much better when the input vectors are first reduced in length by ICA- or WPT-based feature extraction before feeding into MLP. The MLP using *WPT_feature* is seen to outperform that using *ICA_feature* due to the larger margin between feature clusters formed by WPT.

As illustrated in Table 6.11, MLPs using *WPT_feature* and *ICA_feature* take only 0.186 s and 0.164 s respectively to identify a new set of data. The methods are therefore potentially suitable for online applications.

Table 6.11 Comparison of performance of using different type of features

Input type	Generalization MSE	Training convergence time (sec)	*Time needed to classify a new set of data (sec)
Pre-selected signals	0.0320	58.6	2.541
ICA_feature	0.0121	1	0.164
WPT_feature	0.0098	1.02	0.186

*: Including all the processes, namely denoizing, feature extraction and MLP classification

Table 6.12 compares the performance of the method developed in this research with methods proposed in other published works. In [3, 23], phase-resolved (PRPD) patterns are used as the PD features. Thus, at least a few seconds are required to form the patterns. In addition, the computing time of the denoizing and classification algorithm has to be added to the total identification time in [3, 23]. During the forming PRPD patterns, more than one type of PD can take place in the GIS chamber, which may lead to further misclassification as indicated by '<' in Table 6.12.

Table 6.12 Comparison of performance of different identification methods

Method	Correct classification rate	Speed (sec)
In this thesis	100%	0.186
In reference [3]	< 95%	> 1
In reference [23]	< 85%	> 1

6.3 CONCLUDING REMARKS

In this chapter, a MLP neural network is implemented in a computer program to improve the reliability and speed of PD identification and automate the classification process. Results show that MLP with a simple structure is able to classify PD successfully due to the compactness and high quality of the features extracted by ICA- or WPT-based method. Comparative studies indicate that ICA- and WPT-based feature extraction improve the performance of MLP. Particularly, MLP with WPT-based preprocessing achieves 100% correct classification on test, which verifies the effectiveness of the WPT-based feature extraction. Moreover, both the WPT- and ICA-based methods correctly classify between corona and SF_6 PD. This verifies the noise rejection capability of these methods.

CHAPTER 7

PERFORMANCE ENSURENCE FOR PD IDENTIFICATION

This chapter proposes a general scheme for ensuring the robustness of PD identification within the test GIS section. The scheme is first described, followed by its implementation in ICA- and WPT-based methods. Numerical results are then presented and discussed.

7.1 INTRODUCTION

In previous Chapters 4, 5 and 6, the methods of feature extraction and PD identification are developed and verified for data measured one metre away from PD source within the test GIS section as described in Appendix A. When applied outside the test GIS section, features extracted from the above database may not work well due to excessive changes in GIS configuration, sensor type, rated voltage, SF_6 gas pressure, sampling rate and etc. Robustness of the extracted features and proposed classifier should however be ensured for all PD activities within the test GIS section. The scheme as in Fig. 7.1 is thus designed for re-selection of the features and re-training of the proposed classifier, should the variations of measurement conditions in the test GIS section be excessive. As PD can occur at any position within the GIS chamber, the impact of PD-to-sensor distance is focused in this Chapter. A comprehensive database containing 176 data records as shown in Table A.3 are measured for verifying the features extracted by ICA-based and WPT-based method. Salient features of the scheme are discussed in the following section. Numerical results showing the robustness of the PD features are presented and discussed in Section 7.3.

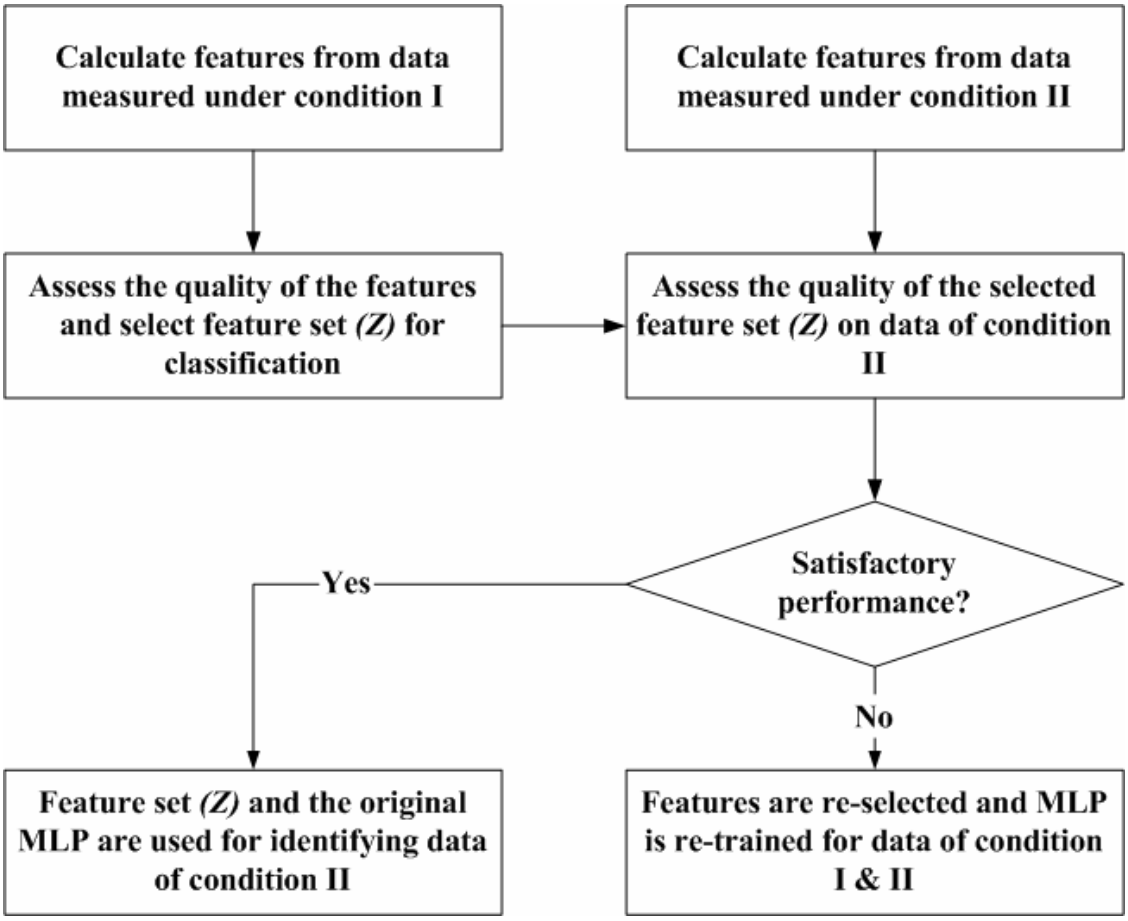


Fig. 7.1 General scheme for selecting features for PD identification
Condition I: Measurement at one metre away from PD source
Condition II: Measurement at other distances

7.2 PROCEDURE FOR ENSURING ROBUSTNESS OF CLASSIFICATION

According to Fig. 7.1, the general procedure for ensuring robustness of PD classification is given as follows:

1. Calculate PD features using ICA-based or WPT-based method for data measured one metre away from the PD sources (Condition I).
2. Assess the effectiveness of features by their classification capability on data with one metre PD-to-sensor distance, forming feature set (*Z*).

3. Calculate features using ICA-based or WPT-based method for data measured at various other distances (Condition II).
4. Assess the effectiveness of features in feature set (Z) by their classification capability on data measured under Condition II.
5. If satisfactory performance is obtained in step 4, feature set (Z) and the original MLP are employed for identifying data measured under Condition II. Otherwise, go to step 6.
6. Features are re-selected and MLP is re-trained using all the data of one metre as well as other distances.

Re-selection of *ICA_feature* and *WPT_feature* for assuring the robustness of PD identification is discussed in the following sections. Details of ICA- and WPT- based feature extraction methods are given in Chapter 4 and 5 respectively. After feature re-selection, MLP must be re-trained using the re-selected features according the procedure described in Chapter 6.

7.2.1 Re-selection of *ICA_feature*

To re-select features from the extended database that consists of 80 data with one metre PD-to-sensor distance and 176 data of other distances, the most dominating independent components are first identified from the extended database using FastICA. The input of FastICA consists of a chosen set of twelve signals with all PD types and all PD-to-sensor distances as shown in Fig. 7.2. The obtained independent components are illustrated in Fig. 7.3.

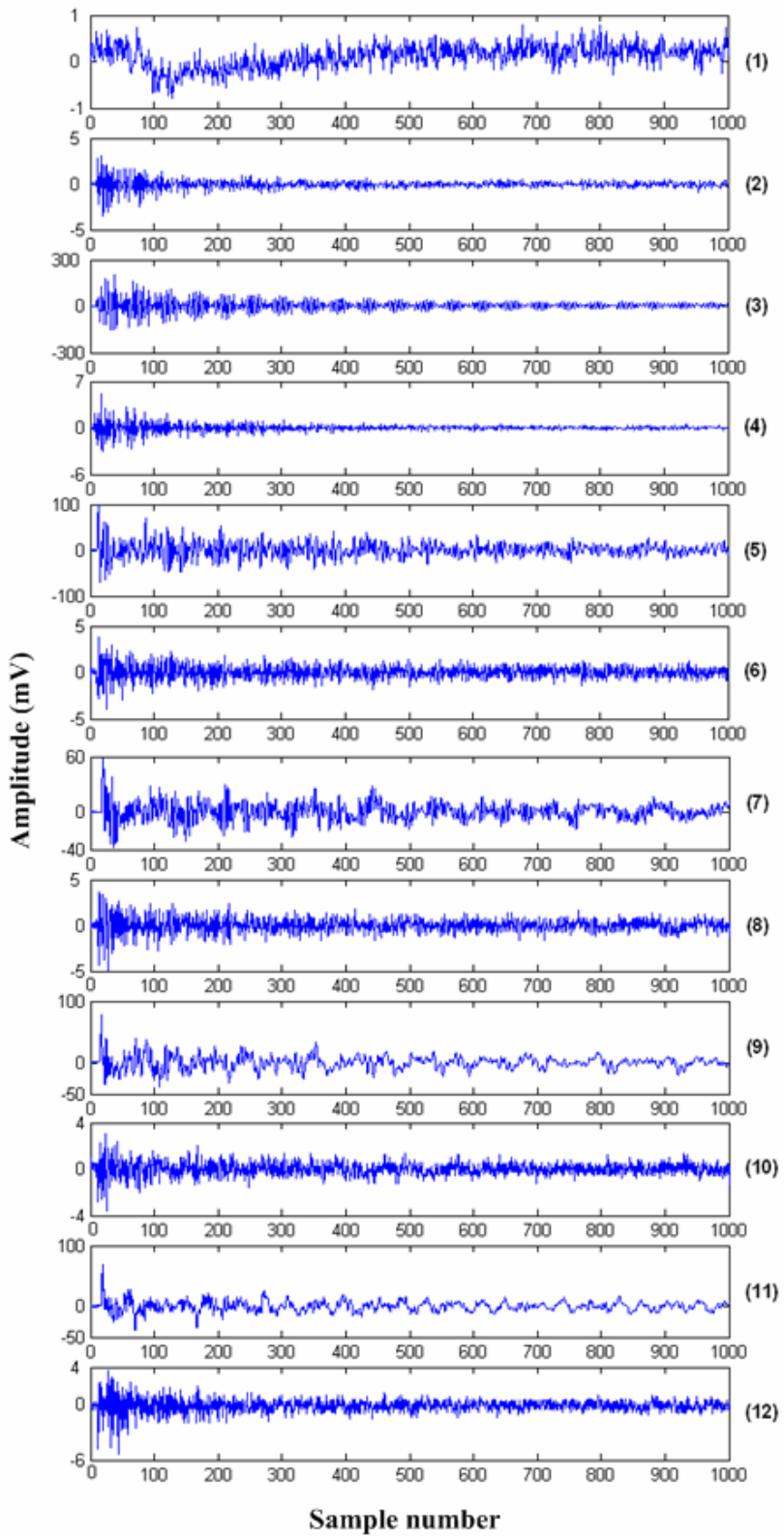


Fig. 7.2 Chosen signal sets for calculating independent components from extended database (1)-corona; (2)- particle on the surface of spacer; (3),(5),(7),(9),(11)- particle on conductor; (4),(6),(8),(10),(12)- free particle on enclosure.
PD-to-sensor distance: (1)-(4) one metre ; (5)-(6) 2.5 m; (7)-(8) 4.6 m; (9)-(10) 6 m; (11)-(12) 7.8 m.

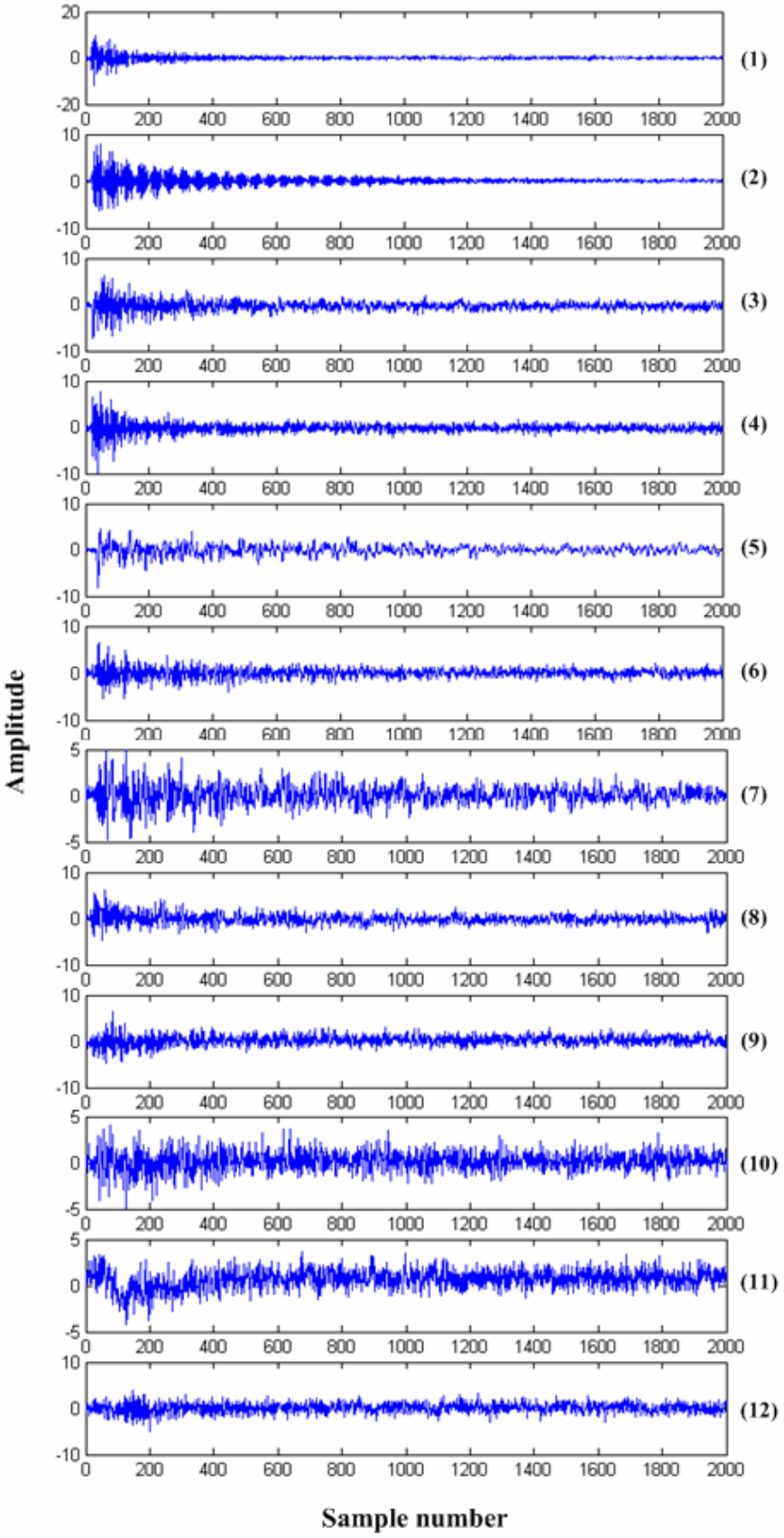


Fig. 7.3 Independent components obtained from FastICA for extended database

To identify the most dominating independent components, the variance of their corresponding projections are calculated for the set of twelve signals according to equation 4.3 and shown in Table 7.1. As highlighted in Table 7.1, two independent components with highest variances in the corresponding projections are selected for calculating ICA features for all the 256 sets of data according to equation 4.4. As a result, the size of the extended *ICA_feature* is 256*2. The classification performance of the re-selected *ICA_feature* is evaluated in Section 7.3.1 (B).

Table 7.1 Variance of projections of the independent components in Fig. 7.3
(For signals of all PD types and all PD-to-sensor distances)

Independent Components	Variance of the projections
ICAPD$_1$	***0.2465***
ICAPD$_2$	***0.1763***
ICAPD$_3$	*0.0489*
ICAPD$_4$	*0.0224*
ICAPD$_5$	*0.0208*
ICAPD$_6$	*0.0200*
ICAPD$_7$	*0.0158*
ICAPD$_8$	*0.0089*
ICAPD$_9$	*0.0075*
ICAPD$_{10}$	*0.0068*
ICAPD$_{11}$	*0.0064*
ICAPD$_{12}$	*0.0047*

7.2.2 Re-selection of *WPT_feature*

All the 256 sets of data with all PD types and all PD-to-sensor distances are first decomposed into wavelet packet domain, forming 256 wavelet packet decomposition (WPD) trees. The 'db9' wavelet is verified to be the most effective wavelet for classification for the extended database as highlighted in Table 7.2. It results in the largest J_{sum} that indicates the best discriminating capability. The level of decomposition is set to 5 according to Section 5.3.1.

Table 7.2 Largest *J* value of candidate wavelets for extended database

wavelet	largest *J*	2nd largest	3rd largest	4th largest	5th largest	J_{sum}
db1	10.0856	9.1245	8.3664	6.0312	5.6563	39.264
db2	7.0786	6.9758	5.0234	5.012	4.2765	28.3663
db3	8.6245	8.4653	7.3424	7.2756	6.6654	38.3732
db4	11.1456	10.0475	9.4579	8.9878	8.1575	47.7963
db5	10.1421	8.8542	8.4636	7.4263	4.7445	39.6307
db6	9.5325	8.0945	6.7543	5.2351	4.5754	34.1918
db7	9.3223	8.8945	7.5641	6.8753	6.0985	38.7547
db8	9.0435	8.2873	7.8633	7.3546	7.1468	39.6955
db9	**12.2098**	**11.2892**	**8.941**	**8.6021**	**7.955**	**48.9971**
db10	11.0021	8.2235	7.3621	6.5431	5.9878	39.1186
sym4	9.1456	8.9673	8.8043	7.9253	6.5454	41.3879
sym5	9.2978	8.5003	8.1023	7.4675	7.2564	40.6243
sym6	9.0298	8.9465	8.1423	7.3034	6.2344	39.6564
sym7	8.8168	8.4289	6.9312	5.8256	5.4896	35.4921
sym8	8.9234	8.3765	8.3234	8.2745	8.2405	42.1383
sym9	9.3869	8.2023	6.2406	5.4852	5.0984	34.4134
sym10	9.1923	7.2342	6.0967	6.0574	5.6456	34.2262
coif1	10.9939	9.3241	8.0456	7.6574	7.1121	43.0431
coif2	10.8934	10.0252	9.4344	8.7687	7.6733	46.795
coif3	8.992	8.7686	7.6546	7.0675	6.7832	39.2659
coif4	8.8914	8.2463	8.0422	7.1389	6.3574	38.6762
coif5	8.9131	6.7842	6.5368	6.0797	5.4356	33.7494

Subsequently, node features defined in Section 5.2.2 namely node kurtosis, node skewness, node energy, node median and node mean are calculated for all nodes in WPD trees, forming feature trees as illustrated in Fig. 5.6. The classification capability of node features are then evaluated using *J* criterion that is defined in equation 5.9. Table 7.3 shows the node features with the highest *J* values. Comparing with Table 5.2, it can be seen that the extracted features are identical and only their sequence in the tables are slightly different. This suggests that WPT features are robust for data having different PD-to-sensor distances. In addition, the first four features in Table 7.3 have J values larger than the critical J value (J_{cr}) defined in Section 6.3.3, which indicates good classification capability. Classification performance of the features in Table 7.3 is further assessed in Section 7.3.2 (B).

Table 7.3 Features extracted from extended database using WPT

serial no.	feature	*J* value	frequency range (Hz)
1	$(5,21)_{kurtosis}$	12.2098	1.3125 G – 1.375 G
2	$(1,0)_{skewness}$	11.2892	0 – 1 G
3	$(5,1)_{energy}$	8.941	62.5 M – 125 M
4	$(5,19)_{skewness}$	8.6021	1.1875 G – 1.25 G
5	$(5,0)_{kurtosis}$	7.955	0 – 62.5 M
6	$(5,11)_{skewness}$	7.3879	687.5 M – 750 M
7	$(5,20)_{median}$	7.1258	1.25 G – 1.3125 G
8	$(3,0)_{kurtosis}$	6.9012	0 – 250 M
9	$(4,0)_{skewness}$	6.326	0 – 125 M
10	$(4,2)_{energy}$	6.2094	250 M – 375 M

7.3 RESULTS AND DISCUSSIONS

The robustness of PD features extracted by ICA- and WPT-based method is verified in this section using the proposed scheme as in Fig. 7.1.

7.3.1 Robustness of ICA-based Feature Extraction

The performance of original *ICA_feature* and MLP is first assessed on extended database. Subsequently, results of re-selected features and re-trained MLP are presented and discussed.

A. Using original *ICA_feature* and MLP

ICA features for each set of new data are calculated by projecting it onto the two most dominating independent components obtained in Section 4.4.1 using equation 4.4. As a result, 176*2 features are obtained from new data. Fig. 7.4 shows the original feature clusters together with the features calculated from the new data. It can be seen that the original cluster boundaries are still valid for all the four cases. This indicates that the impact of PD-to-sensor distance is not significant. However, when the distance between PD source and sensor extends to 7.8 m, the margin between feature clusters of “enclosure” and “spacer” gets small, which may affect the classification performance of neural network. Therefore, re-selection of *ICA_feature* and re-training of MLP may be required.

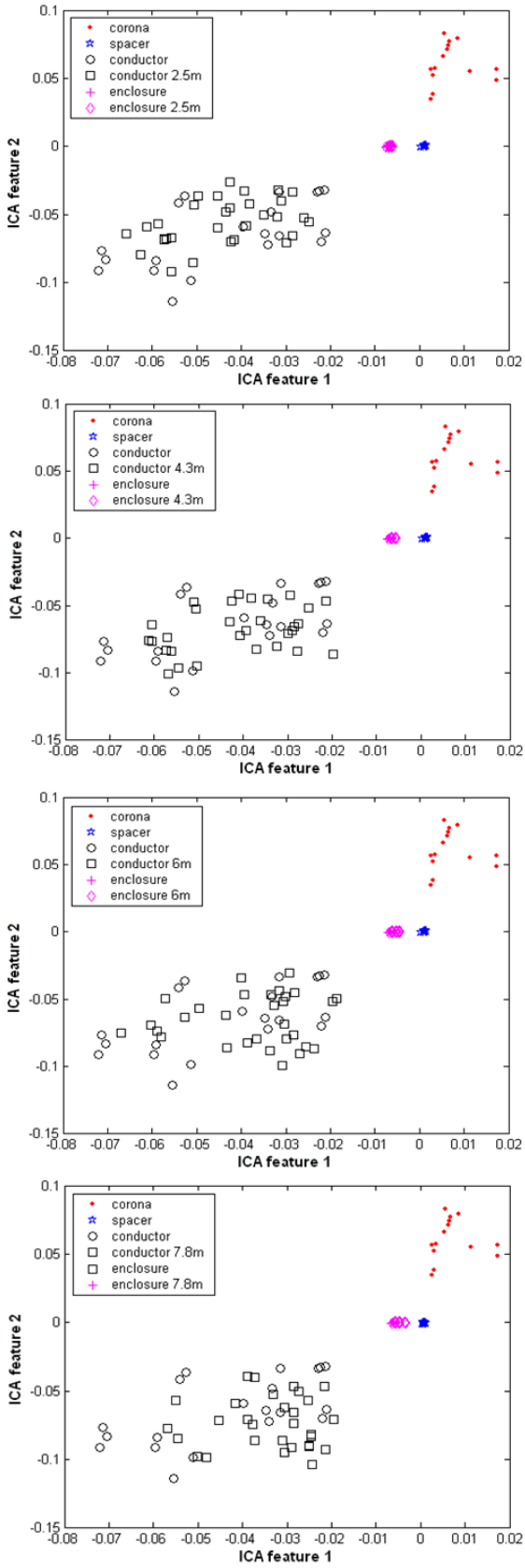


Fig. 7.4 Impact of distance between PD source and sensor on original *ICA_feature* (a) 2.5 m; (b) 4.3 m; (c) 6 m; (d) 7.8 m.

The performance of MLP trained with the original ICA features as in Chapter 6 is investigated with data obtained from the four PD-to-sensor distances. As illustrated in Table 7.4, an overall classification performance of 93.75 % is achieved. In the worst case, where the PD-to-sensor distance is 7.8 m, six out of fifty patterns are misclassified. In all the misclassified cases, patterns of “enclosure” are classified as “spacer”. This may be due to the small margin between ICA feature clusters of “enclosure” and “spacer” as shown in Fig. 4.15.

Table 7.4 Performance of original MLP with *ICA_feature* on data having different PD-to-sensor distances

Distance (m)	Number of misclassified patterns	Correct classification rate
2.5	1/50	98 %
4.3	2/38	94.7 %
6	2/38	94.7 %
7.8	6/50	88 %
Subtotal	11/176	93.75 %

Table 7.5 shows the MLP performance on data with different PD-to-sensor distances using more independent components. It can be seen that using additional independent components does not improve the performance of the MLP in terms of overall and worst case correct classification rate.

Table 7.5 Performance on data with different PD-to-sensor distances using more independent components

Number of used independent components	Overall correct classification rate	Correct classification rate in the worst case
3	93.75 %	88 %
4	93.75 %	88 %
5	93.75 %	88 %
6	93.18 %	86 %
7	93.18 %	86 %
8	93.18 %	86 %

B. Re-selection of *ICA_feature* and re-training of MLP

To improve classification performance of the MLP, ICA features are re-selected from the extended database according to the procedure described in Section 7.2.1. Fig. 7.5 shows the feature clusters obtained from the re-selected *ICA_feature* with a size of 256*2. It is seen that the features of different classes are better separated in Fig. 7.5 than in Fig. 7.4, which indicates improvement on classification.

Using re-selected features as input, the MLP is re-trained and re-tested on the extended database. During re-training, the convergence speed and network structure remain the same as in Chapter 6. On the other hand, the performance of the updated MLP on testing has been improved as shown in Table 7.6. As observed, the most obvious improvement is obtained for the case of 7.8 metre PD-to-sensor distance. In addition,

the overall performance is also improved by 3.4%. It is shown in Table 7.7 that using additional independent components does not improve the performance of the re-trained MLP.

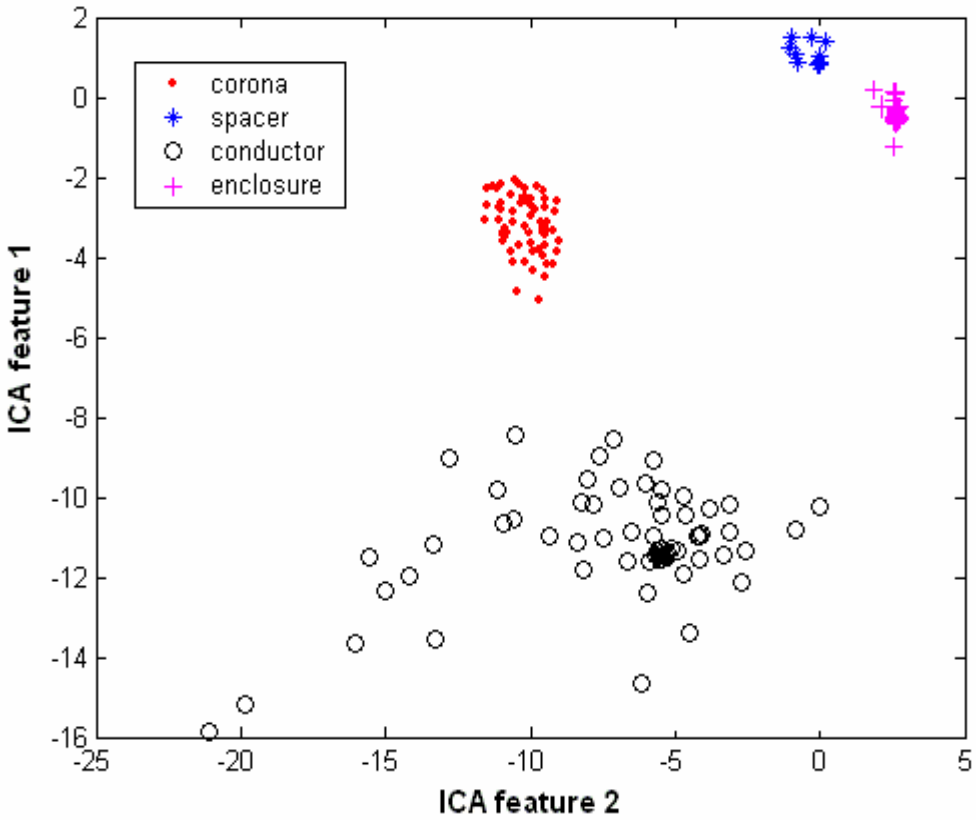


Fig. 7.5 Feature clusters formed by re-selected *ICA_feature* for extended database

Table 7.6 Generalization performance of re-trained MLP with re-selected *ICA_feature*

Distance (m)	Number of misclassified patterns	Correct classification rate
1	2/80	97.5%
2.5	1/50	98 %
4.3	1/38	97.4 %
6	1/38	97.4 %
7.8	2/50	96 %
Subtotal	7/256	97.3 %

Table 7.7 Performance of re-trained MLP using more independent components

Number of used independent components	Overall correct classification rate	Correct classification rate in the worst case (distance = 7.8 m)
3	97.3 %	96 %
4	97.3 %	96 %
5	97.3 %	96 %
6	97.3 %	96 %
7	97.3 %	96 %
8	97.3 %	96 %
9	96.9%	94 %
10	96.9%	94 %
11	96.9%	94 %
12	96.9%	94 %

7.3.2 Robustness of WPT-based Feature Extraction

The impact of PD-to-sensor distance on the WPT features is studied using data measured from various PD-to-sensor distances. Node features selected by the wavelet-packet-based method are calculated for the 176 sets of additional data measured from other PD-to-sensor distances as shown in Table A.2.

A. Using original *WPT_feature* and MLP

Fig. 7.6 shows the original features calculated from data measured one metre away together with the features calculated from the new data. The updated feature clusters are seen to be robust and well segregated for all these four distances. This suggests that the feature extraction method is robust for data having different PD-to-source distances.

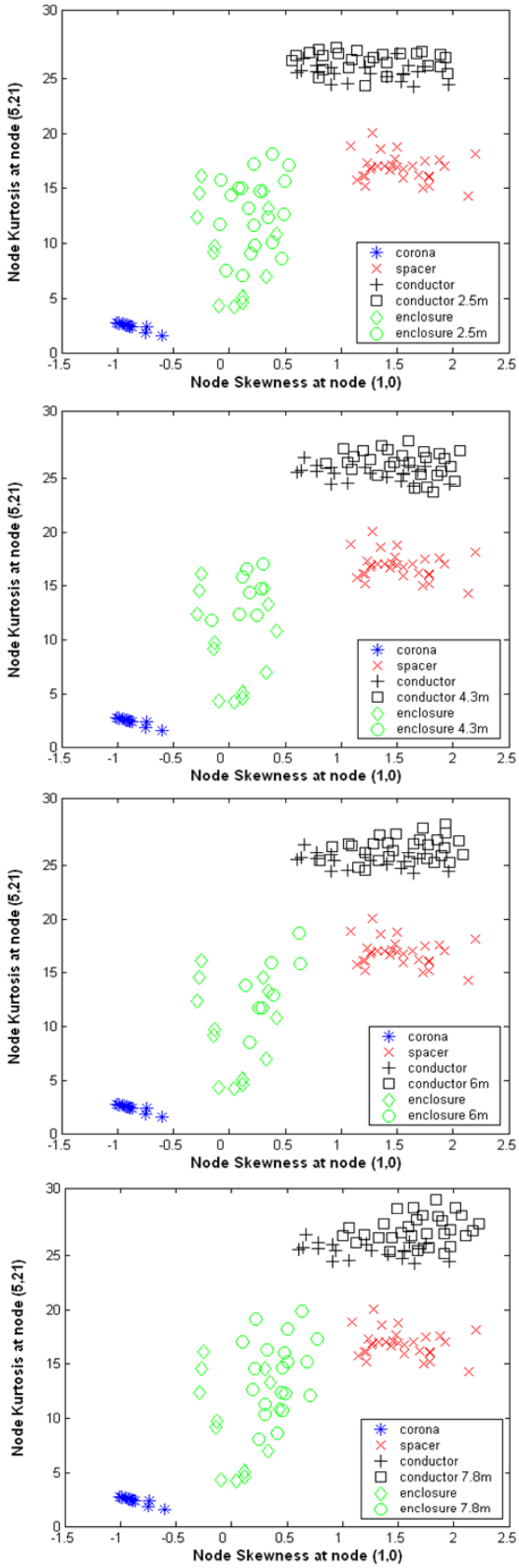


Fig. 7.6 Impact of distance between PD source and sensor on original *WPT_feature*. (a) 2.5 m; (b) 4.3 m; (c) 6 m; (d) 7.8 m.

Table 7.8 shows the corresponding J values for these four distances, which are higher or close to J_{cr} (Chapter 6) indicating a good classification performance.

Table 7.8 Updated J values of the selected features

	1st feature	2nd feature	3rd feature	4th feature
2.5 (m)	12.1328	11.8245	8.6005	8.5927
4.3 (m)	12.1134	11.8109	8.6003	8.5925
6 (m)	11.9907	11.7854	8.5896	8.5925
7.8 (m)	11.9124	11.7565	8.5899	8.5922

The performance of MLP trained with the data measured one metre away from source is tested with data obtained from the four PD-to-sensor distances. As shown in Table 7.9, an overall performance of 98.3% has been achieved, which is better than that obtained from original ICA-based MLP. In addition, only two patterns are misclassified in the worst case.

Table 7.9 Generalization performance of the original MLP on data with different PD-to-sensor distance

Distance (m)	Number of Misclassified patterns	Correct Classification Rate
2.5	0/50	100 %
4.3	0/38	100 %
6	1/38	97.4 %
7.8	2/50	96 %
Subtotal	3/176	98.3 %

B. Re-selection of *WPT_feature* and re-training of MLP

As shown in Tables 7.3 and 5.2, the re-selected WPT features are the same as the original features. However, improvement in classification can be achieved by re-training the MLP using extended database. As shown in Table 7.10, the re-trained MLP is able to classify all the data correctly, regardless of the changing of PD location. Thus, it can be concluded that WPT-based method outperforms ICA-based method in terms of classification accuracy.

Table 7.10 Generalization performance of re-trained MLP with *WPT_feature*

Distance (m)	Number of misclassified patterns	Correct classification rate
1	0/80	100 %
2.5	0/50	100 %
4.3	0/38	100 %
6	0/38	100 %
7.8	0/50	100 %
Subtotal	0/256	100 %

7.4 CONCLUDING REMARKS

In this chapter, a general scheme is proposed for ensuring the robustness of PD identification within the test GIS section. Re-selection of features and re-training of MLP are employed for quality assurance. Numerical results show that the proposed scheme of re-selection and re-training improves the performance of both ICA- and WPT-based classifiers. In particular, the re-trained WPT MLP achieves 100% correct classification on all the data, regardless of the changing of PD location.

CHAPTER 8

CONCLUSIONS AND FUTURE WORK

This chapter concludes the study on PD denoizing and identification in GIS system which has been presented in the former chapters. Based on the results of this research, the conclusions are summarized and followed by recommendations for future work.

8.1 CONCLUSION

GIS has been used worldwide for many years because of its low maintenance and compact size. This has made it an attractive option in many applications. However, on the downside, GIS has problems relating to the sharp deterioration of the dielectric strength of its insulation gas (SF_6) due to PD. On the other hand, PD is caused by the extreme field intensity being built around the sharp edge of small particles which may attach to the bus conductor, the enclosure or the insulation spacer. In industry applications, these faults could be attributed to mechanical faults during manufacture, protrusions on the enclosure, the HV conductor as well as free moving particles. Hence, the extreme field intensity caused by particles may produce PD inside the GIS, which may lead to the failure of the system.

Preventing the failure of a GIS requires a reliable and efficient PD measuring and diagnostic technique, which is able to detect and identify signals from harmful defects. Thus, a prompt warning message can be given before the breakdown occurs. However, the two major issues associated with such diagnostic systems, namely influence of noise and the extraction of effective features from measured data, must be addressed to achieve a successful diagnosis of PD activities in GIS. In this thesis, a novel PD diagnostic system is developed based on UHF signals with special emphasis on denoizing and feature extraction from the PD signal.

8.1.1 Denoizing of PD Signals

In practice, it is impossible to achieve reliable diagnosis of insulation in a highly noisy environment. Hence, denoizing of PD signals is usually the first issue to be accomplished during PD analysis and diagnosis.

In this research project, a “wavelet-packet” based method with a novel variance-based criterion is employed to construct the best tree to denoise the UHF signal. The new criterion automatically selects the most PD dominated components from the wavelet-packet-decomposition tree for signal reconstruction. This leads to good denoizing performance. Various methods were developed for selecting parameters associated with the denoizing scheme, such as wavelet filters and decomposition level. Among them, the method based on the genetic algorithm is able to optimally select a complete set of parameters by evaluating the performance of the parameters holistically. SNR and correlation coefficient are employed for selecting denoizing parameters to ensure restoration of the original PD signal during denoizing with a significant reduction in the noise level.

It has been shown that the proposed method offers better denoizing compared to DWT and WPT with the standard entropy-based criterion. Using the proposed method, successful and robust denoizing is achieved for PD signals having various SNR levels. Successful restoration of the original waveform facilitates the subsequent pre-selection process and enables extraction of reliable features for PD identification.

In this research, external corona discharge is considered as one of the typical pulse-shaped noises and addressed in this thesis. In practical GIS, if other pulse-shaped

noises, such as switching over-voltages, are present and produce significant signals within UHF ranges, the MLP neural network will label them as 'unknown signals'. In such cases, further investigation of the noises maybe required. However, drastic changes should not be required for the proposed method.

8.1.2 Feature Extraction for PD Source Recognition

Traditionally, phase-resolved methods such as PRPD are employed for PD source recognition and corona noise discrimination. Although these methods have been extensively applied in industries to evaluate the insulation integrity of HV equipments such as generator, transformer and cable, they have significant limitations when applied to GIS in terms of accuracy and speed. Hence, new methods are developed in this research project to solve the problems with phase-resolved methods.

Various PD features are derived from UHF signals and form a solid basis for current and future work on PD identification. The first category of PD features, namely *ICA_Feature* is extracted in the time domain using Independent Component Analysis. Using *ICA_Feature*, successful identification of PD is achieved with limitation of small "between-class" margins due to the time-domain nature of ICA. White noise present in the measured signals is seen to reduce the discriminating capability of the extracted features. This shows the importance of denoizing. When the distance between PD source and UHF sensor varies, re-selection of the *ICA_feature* and re-training of MLP are seen to have improved the correct classification rate to 97.3%, which ensures the robustness of the proposed method.

Features extracted in the time-frequency domain using the wavelet packet transform (*WPT_Feature*) form the second category of PD features. Taking advantage of the additional frequency information included with the wavelet packet transform, *WPT_Feature* exhibits a large margin between feature clusters of different classes, which indicates good classification performance. Among subcategories of *WPT_Feature*, "distribution-shape" based node features are more effective than other node features such as node energy. Based on this it can be concluded that the wavelet-packet-based method outperforms methods which operate solely in the time or frequency domain (FFT) due to its time-frequency characteristic. The best wavelet for feature extraction is "db9", which is different from that used for denoizing namely "sym8". This indicates that the selection of wavelet is application-dependent. Investigation of the impact of noise levels on the effectiveness of features confirms that denoizing is crucial for reliable feature extraction and classification. For various PD-to-sensor distances, the same set of features is selected by WPT-based method. However, re-training of the MLP improves the classification performance, which verifies the re-selection and re-training scheme for quality assurance.

Owing to the compactness and high quality of the extracted features, successful and robust PD identification is achieved using a very simple MLP network. Particularly, MLP with WPT-based preprocessing achieves 100% correct classification on all PD activities at all location within the given GIS configuration after re-training. This verifies the robustness of the WPT-based feature extraction. The methods developed in this project can be used either as a stand-alone system or as a supplement to the existing PRPD system to improve its performance. Moreover, both the WPT- and ICA-based PD diagnostic methods are potentially suitable for online applications.

8.2 RECOMMENDATIONS FOR FUTURE WORK

Although significant progress has been made in achieving better diagnosis of insulation integrity of GIS, there is still space for further expansion and improvement:

(1) Other PD-causing Defects in GIS

Major PD-causing defects [7, 10] in SF_6 have been considered in this research. However, PD may also be caused by cavity or metallic intrusion within an epoxy resin support barrier. Although the possibility of encountering these defects is very low in practice [90], further investigation of these defects may be required to develop a comprehensive PD diagnostic system.

The amplitude and rise time of PD current pulses produced by defects in solid differ from those produced in SF_6 due to the different nature of the insulation material [7, 90, 91]. On the other hand, the shape of PD current pulses determines the waveform of corresponding UHF signals [92]. Thus, UHF signals excited by PD in solid and PD in SF_6 should have very different waveforms and time-frequency characteristics. This indicates that good classification may be achieved without drastic changes on the methods developed in this thesis. Re-selection of features and re-training of MLP may be required to achieve satisfactory identification.

Apart from PD-to-sensor distance, dimension and shape of the particle may affect the measured UHF signals. In this research, however, a typical particle which can cause PD of critical amplitude without leading to immediate breakdown is employed to

simulate the defects. Although the dimension and shape of the defect may change the shape of PD pulse, this has no significant influence on the basic principle of the proposed techniques. Re-selection of features and re-training of MLP may be required to achieve satisfactory identification.

(2) Speed Improvement

In this research project, the entire PD denoizing and identification scheme is developed using Matlab language on a PC platform. Since Matlab is an interpreted language instead of a compiled language (such as C), its speed will always lag behind that of a custom program written in a language like C. Therefore, converting the Matlab programs into C or C++ will shorten the response time of the diagnosis system. Further improvement of the speed may be achieved by implementing the scheme on a Digital Signal Processor (DSP).

(3) Extension to Other GIS Configurations

The new PD denoizing and identification methods are developed and tested for a simple GIS configuration, which consists of a straight-through busbar, enclosure and two spacers. However, there are more complicated GIS configurations such as T junction, gas circuit breaker and disconnector in practical GIS systems. Therefore, the performance of the methods developed in this project should be verified for these configurations. Further development of the proposed methods may be required on new measured data to ensure satisfactory performance for the practical GIS system.

(4) Study on PD Location

In contrast to (2), measured PD data can be further classified in terms of source location. Once a harmful PD is detected and recognized, it is crucial to locate it in the GIS tank in a fast manner, so that necessary maintenance can be arranged promptly. Although the location of PD source can be roughly determined through identification of the defect type, it is not sufficient to provide obvious guidance for maintenance and repair due to the complicated structures and huge size of GIS. Therefore, precisely locating the PD source based on UHF measurement should be further investigated in the future study. To determine PD location, data measured from one channel is not sufficient, as the time delay information is crucial to the location problem. Hence, data that is synchronously measured from at least two channels is fundamental for future work.

REFERENCES

[1] IEC Publication 60270, Ed.3.0. "High-voltage test techniques – Partial discharge measurement", 2001.

[2] Judd M.D., Farish O., Hampton B.F., "The excitation of UHF signals by partial discharges in GIS," *IEEE Trans. On Dielectrics and Electrical Insulation*, vol. 3, no. 2, pp. 213-227, Apr 1996.

[3] Pearson J.S., Farish O., Hampton B.F., Judd M.D., Templeton D., Pryor B.M., Welch I.M., "Partial discharge diagnostics for gas insulated substations," *IEEE Trans. On Dielectrics and Electrical Insulation*, vol. 2, no. 5, pp. 893-905, Oct 1995.

[4] Hampton B.F., Meats R.J., "Diagnostic measurements at UHF in gas insulated substations," *IEE Proc. Generation, Transmission and Distribution*, vol. 135, no. 2, pp. 137-145, Mar 1988.

[5] Kurrer R., Feser K., "The application of ultra-high-frequency partial discharge measurements to gas-insulated substations," *IEEE Trans. On Power Delivery*, vol.13, no. 3, pp. 777 –782, July 1998.

[6] Nicholas de Kock, Branko Coric and Ralf Pietsch, "UHF PD detection in gas-insulated switchgear – suitability and sensitivity of the UHF method in comparison with the IEC 270 method," *IEEE Electrical Insulation Magazine*, vol. 12, no.6, pp. 20-26, Nov/Dec 1996.

[7] Baumgartner R., Fruth B., Lanz W., Pettersson K., "Partial discharge - Part X: PD in gas-insulated substations – measurement and practical considerations," *IEEE Electrical Insulation Magazine*, vol. 8, no.1, pp. 16-27, Jan/Feb 1992.

[8] Sellars A.G., Farish O. and Hampton B.F., "Assessing the risk of failure due to particle contamination of GIS using the UHF technique," *IEEE Trans. On Dielectrics and Electrical Insulation*, vol. 1, no. 2, pp. 323-331, April 1994.

[9] Sellars A.G., Farish O. and Peterson M.M., "UHF detection of leader discharges in SF_6," *IEEE Trans. On Dielectrics and Electrical Insulation*, vol. 2, no. 1, pp. 143-154, Feb. 1995.

[10] Sellars A.G., Farish O. and Hampton B.F., "Characterising the discharge development due to surface contamination in GIS using the UHF technique," *IEE Proc. Science, Measurement and Technology*, vol. 141, no. 2, pp. 118-122, Mar 1994.

[11] I. Shim, J. J. Soraghan, W. H. Siew, "Digital Signal Processing Applied to the Detection of Partial Discharge: An Overview," *IEEE Electrical Insulation Magazine*, vol. 16, no.3, pp.6–12, May/June 2000.

[12] U. Kopf and K. Feser, “Rejection of Narrow-band Noise and Repetitive Pulses in On-site PD Measurements,” *IEEE Trans. On Dielectrics and Electrical Insulation*, vol. 2, no. 6, pp. 1180-1191, Dec 1995.

[13] I. Shim, J. J. Soraghan, W. H. Siew, “Detection of PD utilizing digital signal processing methods. Part 3: Open-loop noise reduction,” *IEEE Electrical Insulation Magazine*, vol. 17, no.1, pp.6–13, Jan/Feb 2001.

[14] D. Sundararajan, *Digital signal processing: theory and practice***.** New Jersey: World Scientific, 2003.

[15] J. Ramirez-Nino, S. Rivera-Castaneda, V. R. Garcia-Colon, and V. M. Castano, “Analysis of partial discharges in insulating materials through the wavelet transform,” *Computational Materials Science*, vol. 9, pp. 379-388, 1998.

[16] X. Ma, C. Zhou, and I. J. Kemp, “Automated wavelet selection and thresholding for PD detection,” *IEEE Electrical Insulation Magazine*, vol.18, no.2, pp. 37 -45, Mar/Apr 2002.

[17] X. Ma, C. Zhou, and I. J. Kemp, “Interpretation of Wavelet Analysis and Its Application in Partial Discharge Detection,” *IEEE Trans. On Dielectrics and Electrical Insulation*, vol. 9, no. 3, pp. 446-457, June 2002.

[18] L. Satish, B. Nazneen, “Wavelet-based Denoising of Partial Discharge Signals Buried in Excessive Noise and Interference,” *IEEE Trans. On Dielectrics and Electrical Insulation*, vol. 10, no. 2, pp. 354-367, Apr 2003.

[19] L. Angrisani, P. Daponte, G. Lupo, C. Petrarca and M. Vitelli, “Analysis of Ultrawide-band Detected Partial Discharges by Means of a Multiresolution Digital Signal-processing Method,” *Measurement*, vol. 27, pp. 207-221, 2000.

[20] M. Hikita, T. Kato and H. Okubo, “Partial Discharge Measurements in SF_6 and Air using Phase-resolved Pulse-height Analysis,” *IEEE Trans. Dielectrics and Electrical Insulation*, vol. 1, no. 2, pp. 276 -283, Apr. 1994.

[21] Borsi H., Gockenbach E., Wenzel D., “Separation of partial discharges from pulse-shaped noise signals with the help of neural networks,” *IEE Proc. Science, Measurement and Technology*, vol. 142, no. 1, pp. 69-74, Jan 1995.

[22] Borsi H., “A PD measuring and evaluation system based on digital signal processing,” *IEEE Trans. Dielectrics and Electrical Insulation*, vol. 7, no. 1, pp 21-29, Feb. 2000.

[23] Kranz H.-G., “Fundamentals in computer aided PD processing, PD pattern recognition and automated diagnosis in GIS,” *IEEE Trans. Dielectrics and Electrical Insulation*, vol. 7, no. 1, pp. 12-20, Feb. 2000.

[24] S. Meijer, E. Gulski and J. J. Smit, “Pattern Analysis of Partial Discharges in SF_6 GIS,” *IEEE Trans. Dielectrics and Electrical Insulation*, vol. 5, no. 6, pp. 830 - 842, Dec. 1998.

[25] W. Ziomek, M. Reformat and E. Kuffel, “Application of Genetic Algorithms to Pattern Recognition of Defects in GIS,” *IEEE Trans. Dielectrics and Electrical Insulation*, vol. 7, no. 2, pp. 161-168, Apr. 2000.

[26] D. J. Hamilton, J. S. Pearson, “Classification of partial discharge sources in gas-insulated substations using novel preprocessing strategies,” *IEE Proc. Science, Measurement and Technology*, vol. 144, no. 1, pp. 17-24, Jan 1997.

[27] J. S. Pearson, B. F. Hampton and A. G. Sellars, “A Continuous UHF Monitor for Gas-insulated Substations,” *IEEE Trans. Electrical Insulation*, vol. 26, no. 3, pp. 469 -478, June 1991.

[28] Hugh M. Ryan, *High voltage engineering and testing.* IEE, 2001.

[29] Sudarshan T.S. and Dougal R.A., “Mechanisms of surface flashover along solid dielectrics in compressed gases: a review,” *IEEE Trans. On Electrical Insulation*, vol. 21, no. 5, pp. 727-746, 1986.

[30] Chakrabarti A.K., Van Heeswijk R.G., Srivastava K.D., “Free particle initiated 60 Hz breakdown at a spacer surface in a gas insulated bus,” *IEEE Trans. On Electrical Insulation*, vol. 24, no. 4, pp. 549-560, 1989.

[31] Mladen Victor Wickerhauser, Adapted wavelet analysis from theory to software, Wellesley, MA: A.K. Peters, c1994.

[32] Leman H., Marque C., “Rejection of the maternal electrocardiogram in the electrohysterogram signal,” *IEEE Trans. Biomedical Engineering*, vol. 47, no. 8, pp. 1010-1017, Aug 2000.

[33] Shen M., Sun L., Chan F.H.Y., “Method for extracting time-varying rhythms of electroencephalography via wavelet packet analysis,” *IEE Proc. Science, Measurement and Technology*, vol. 148, no. 1, pp. 23 -27, Jan 2001.

[34] Carnero B., Drygajlo A., “Perceptual speech coding and enhancement using frame-synchronized fast wavelet packet transform algorithms,” *IEEE Trans. Signal processing*, vol. 47, no. 6, pp. 1622 -1635, Jun 1999.

[35] Zixiang Xiong, Ramchandran K., Orchard M.T., “Wavelet packet image coding using space-frequency quantization,” *IEEE Trans. Image Processing*, vol. 7, no. 6, pp. 892 -898, Jun 1998.

[36] Hamid, E.Y.; Kawasaki, Z.-I. , “Wavelet-based data compression of power system disturbances using the minimum description length criterion,” *IEEE Trans. Power Delivery*, vol. 17, no. 2, pp. 460 –466, April 2002.

[37] Jaehak Chung; Powers, E.J.; Grady, W.M.; Bhatt, S.C., “Power disturbance classifier using a rule-based method and wavelet packet-based hidden Markov model,” *IEEE Trans. Power Delivery*, vol. 17, no. 1, pp. 233 –241, Jan. 2002.

[38] Littler T.B., Morrow D.J., “Wavelets for the analysis and compression of power system disturbances,” *IEEE Trans. Power Delivery*, vol. 14, no. 2, pp. 358 -364, Apr 1999.

[39] Hamid, E.Y.; Mardiana, R.; Kawasaki, Z.-I., “Method for RMS and power measurements based on the wavelet packet transform,” *IEE Proc. Science, Measurement and Technology*, vol.149, no. 2, pp. 60 –66, Mar. 2002.

[40] Xianguing Liu, Pei Liu, Shijie Cheng, “A wavelet transform based scheme for power transformer inrush identification,” Power Engineering Society Winter Meeting, 2000. vol. 3, pp. 1862 -1867, Jan 2000.

[41] X. Ma, C. Zhou, and I. J. Kemp, “Wavelets for the analysis and compression of partial discharge data,” Electrical Insulation and Dielectric Phenomena Conf., Annual Report, pp. 329 –334, 2001.

[42] M. Misiti, Y. Misiti, G. Oppenheim, and J. Poggi, Wavelet Toolbox For Use with MATLAB, The MathWorks, Inc., 1996.

[43] Ronald R. Coifman and Mladen Victor Wickerhauser, "Entropy-Based Algorithms for Best Basis Selection," *IEEE Trans on Information Theory*, vol. 38, no. 2, pp. 713 – 718, March 1992.

[44] B. Castro; D. Kogan; A.B. Geva, “ECG feature extraction using optimal mother wavelet,” The 21st IEEE Convention of the Electrical and Electronic Engineers in Israel, 2000, pp. 346 –350.

[45] S.G. Mallat, “A theory for multiresolution signal decomposition: the wavelet representation,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.11, no. 7, pp. 674-693, July 1989.

[46] D.L. Donoho, “Denoising by soft thresholding,” *IEEE Trans. Information Theory*, vol. 41, no. 3, pp. 613-627, May 1995.

[47] D.L. Donoho and I.M. Johnstone, “Adapting to Unknown Smoothness via Wavelet Shrinkage,” *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1200-1224, Dec. 1995.

[48] G. Dai-fei, Z. Wei-hong, G. Zhen-ming and Z. Jian-qiang, “A study of wavelet thresholding denoising,” Signal Processing Proceedings 5th International Conference on WCCC-ICSP 2000, August 2000, vol.1, pp. 329-332.

[49] Darrell Whitley, “A genetic algorithm tutorial,” *Statistics and Computing*, vol. 4, pp. 65-85, 1994.

[50] Sinha N., Chakrabarti R., Chattopadhyay P.K., **“**Evolutionary programming techniques for economic load dispatch**,”** *IEEE Trans. Evolutionary Computation*, vol. 7, no. 1 , pp 83-94, Feb. 2003.

[51] Gerbex S., Cherkaoui R., Germond A.J., “Optimal location of multi-type FACTS devices in a power system by means of genetic algorithms**,”** *IEEE Trans. Power Systems*, vol. 16 , no. 3, pp 537–544, Aug. 2001.

[52] Nicolaisen J., Petrov V., Tesfatsion L., “Market power and efficiency in a computational electricity market with discriminatory double-auction pricing**,”** *IEEE Trans. Evolutionary Computation*, vol. 5 , no. 5, pp 504–523, Oct. 2001.

[53] W. Ziomek, M. Reformat and E. Kuffel, “Application of Genetic Algorithms to Pattern Recognition of Defects in GIS,” *IEEE Trans. Dielectrics and Electrical Insulation*, vol. 7, no. 2, pp. 161-168, Apr. 2000.

[54] Guangning Wu, “A neural network used for PD pattern recognition in large turbine generators with genetic algorithm,” in the 2000 IEEE International Symposium on Electrical Insulation, Anaheim, CA USA, pp. 1-4, Apr. 2000.

[55] Asghar Akbari, Peter Werle, Hossein Borsi, and Ernst Gockenbach, “Transfer Function-Based Partial Discharge Localization in Power Transformers: A Feasibility Study,” *IEEE Electrical Insulation Magazine*, vol.18, no.5, pp. 22 -32, Sep/Oct 2002.

[56] J. J. Grefenstette, “Optimization of Control Parameters for Genetic Algorithms,” *IEEE Trans. Systems, Man, and Cybernetics*, vol. 16, no. 1, pp. 122-128, Jan./Feb. 1986.

[57] D. B. Fogel and J. W. Atmar, “Comparing genetic operators with Gaussian mutations in simulated evolutionary process using linear systems,” *Biological Cybernetics*, vol. 63, no. 2, pp. 111-114, 1990.

[58] Beer, R. D., and Gallagher, J. C. “Evolving dynamical neural networks for adaptive behavior”, *Adaptive Behavior*, vol. 1, pp. 91-122, 1992.

[59] A. Hyvarinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4, pp. 411--430, 2000.

[60] P. Comon, “Independent component analysis - a new concept?” *Signal Processing*, vol. 36, pp. 287-314, 1994.

[61] Te-Won Lee, Lewicki M.S., “Unsupervised image classification, segmentation, and enhancement using ICA mixture models,” *IEEE Trans. Image Processing*, vol. 11, no. 3, pp. 270 -279, Mar 2002.

[62] Bartlett M.S., Movellan J.R., Sejnowski T.J., “Face recognition by independent component analysis,” *IEEE Trans. Neural Networks*, vol. 13, no. 6, pp. 1450 - 1464, Nov 2002.

[63] Semmlow J.L., Weihong Yuan, “Components of disparity vergence eye movements: application of independent component analysis,” *IEEE Trans. Biomedical Engineering*, vol. 49, no. 8, pp. 805 -811, Aug 2002.

[64] Nordberg J., Nordholm S., Grbic N., Mohammed A., Claesson I., "Performance improvements for sector antennas using feature extraction and spatial interference cancellation," *IEEE Trans. Vehicular Technology*, vol. 51, no. 6, pp. 1685 -1698, Nov 2002.

[65] Huaiwei Liao, Niebur D., "Load profile estimation in electric transmission networks using independent component analysis," *IEEE Trans. Power Systems*, vol. 18, no. 2, pp. 707 -715, May 2003.

[66] Simon Haykin, Neural Networks: A Comprehensive Foundation. Prentice-Hall, Inc., 1999.

[67] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Networks*, vol. 10, pp. 626-634, May 1999.

[68] Yen G.G., Lin K.-C., "Wavelet packet feature extraction for vibration monitoring," *IEEE Trans. Industrial Electronics*, vol. 47, no. 3, pp. 650-667, June 2000.

[69] K. Fukunaga, Introduction to Statistical Pattern Recognition. New York: Academic, 1992.

[70] Yu Han and Y.H. Song, "Using Improved Self-Organizing Map for Partial Discharge Diagnosis of Large Turbogenerators," *IEEE Trans. energy conversion*, vol. 18, no. 3, pp. 392 – 399, Sep. 2003.

[71] Tao Hong and M.T.C. Fang, "Detection and Classification of Partial Discharge Using a Feature Decomposition-Based Modular Neural Network," *IEEE Trans. instrumentation and measurement*, vol. 50, no. 5, pp. 1349 – 1354, Oct. 2001.

[72] E. Gulski, A. Krivda, "Neural Network as a Tool for Recognition of Partial Discharges," *IEEE Trans. Electrical Insulation*, vol. 28, no. 6, pp. 984 -1001, Dec. 1993.

[73] Howard Demuth, Mark Beale, Neural Network Toolbox for Use with MATLAB, The MathWorks, Inc., 2001.

[74] David E. Goldberg, "Genetic Algorithms in Search, Optimization, and Machine Learning," Addison-Wesley Publishing Company, Inc., Massachusetts, 1989.

[75] H. Ding, A. A. EI-Keib and R. Smith, "Optimal clustering of power networks using genetic algorithms," *Electric Power Systems Research*, vol. 30, pp. 209-214,1994.

[76] A. Hyvarinen, "Survey on independent component analysis," *Neural Comput. Surveys*, vol. 2, pp. 94-128, 1999.

[77] Y. Han and Y.H. Song, "Condition Monitoring Techniques for Electrical Equipment – A Literature Survey," *IEEE Trans. On Power Delivery*, vol.18, no. 1, pp. 4 –13, January 2003.

[78] C. Bengtsson, “Status and trends in transformer monitoring,” *IEEE Trans. On Power Delivery*, vol.11, no. 3, pp. 1379 –1384, July 1996.

[79] P.J. Tavner and J. Penman, Condition Monitoring of Electrical Machines. Research Studies Press, Ltd., 1987.

[80] Y. Ohshita, A. Hashimoto, and Y. Kurosawa, “A diagnostic technique to detect abnormal conditions of contacts measuring vibrations in metal enclosures of gas insulated substations,” *IEEE Trans. Power Delivery*, vol. 4, pp. 2090–2094, Oct. 1989.

[81] Y. Mukaiyama, I. Takagi, J. Izumi, T. Sekiguchi, A. Kobayashi, and T. Sumikawa, “Investigation on abnormal phenomena of contacts using disconnecting switch and detachable bus in 300 kV GIS,” *IEEE Trans. Power Delivery*, vol. 5, pp. 189–195, Jan. 1990.

[82] A. Girodet, S. Meijer and J.J. Smit, “Development of a partial discharge analysis method to assess the dielectric quality of GIS,” CIGRE paper 15-106, Paris, 2002.

[83] A. Grossmann, “Wavelet transforms and edge detection,” in S. Albeverio et al., editor, *Stochastic Processes in Physics and Engineering*, pp. 149-157. D. Reidel Publishing Company, 1988.

[84] Xiao-Ping Zhang and Mita D. Desai, “Adaptive Denoising Based on SURE Risk,” *IEEE Signal Processing Letters*, vol. 5, no. 10, pp. 265-267, Oct. 1998.

[85] Morcos M.M., Anis H., Srivastava K.D., “Particle-initiated corona and breakdown in GITL systems**,”** *IEEE Trans. Electrical Insulation*, vol. 24, no. 4, pp.561 – 571, Aug. 1989.

[86] Morcos M.M., Ward S.A., Anis H., “On the detection and control of metallic particle contamination in compressed GIS equipment,” Electrical Insulation and Dielectric Phenomena Conference, vol. 2, pp. 476 – 480, 1998.

[87] Steven A. Boggs, “Partial discharge testing of gas insulated substations,” *IEEE Trans. Power Delivery*, vol. 7, pp. 499–506, Apr. 1992.

[88] M. D. Judd; O. Farish; J. S. Pearson and B. F. Hampton, “Dielectric windows for UHF partial discharge detection”, *IEEE Trans. on Dielec. and Elec. Insul*., vol. 8 , no. 6, pp. 953 –958, Dec. 2001.

[89] Toshihiro Hoshino, Kenichi Nojima and Masahiro Hanai, “Real-time PD Identification in Diagnosis of GIS Using Symmetric and Asymmetric UHF Sensors,” *IEEE Trans. Power Delivery*, vol. 19, pp. 1072–1077, July. 2004.

[90] Sellars A.G., Farish O., Hampton B.F. and Pritchard L. S., “Using the UHF technique to investigate PD produced by defects in solid insulation,” *IEEE Trans. On Dielectrics and Electrical Insulation*, vol. 2, no. 3, pp. 448-459, June 1995.

[91] Yonghong Cheng, Chengyan Ren and Xiaolin Chen, “Study on the partial discharge characteristics in different solid and gaseous dielectric by simulation,”

International Symposium on Electrical Insulating Materials, vol. 3, pp. 552-555, 2005.

[92] Sellars A.G., MacGregor S.J. and Farish O., "Calibrating the UHF technique of partial discharge detection using a PD simulator,**"** *IEEE Trans. On Dielectrics and Electrical Insulation*, vol. 2, no. 1, pp. 46-53, Feb. 1995.

APPENDICES

APPENDIX A	UHF Measurement of Partial Discharge in GIS
APPENDIX B	Discrete Wavelet Transform and Wavelet Packet Transform
APPENDIX C	Genetic Algorithm
APPENDIX D	Independent Component Analysis and FastICA Algorithm
APPENDIX E	General Introduction to Neural Networks
APPENDIX F	Resilient Back-propagation Algorithm

APPENDIX A

UHF Measure of Partial Discharge in GIS

Partial discharges produce a series of current pulses with sub-nanosecond durations, and each pulse generates an electromagnetic signal (Fig. A.1) that propagates through the GIS in the UHF range (300 to 1500 MHz). The UHF resonance signals are then picked up by a UHF coupler as in Fig. A.1. In this appendix, the equipment used for PD measurement at TMT&D [89] is first introduced, followed by the experimental set-up.

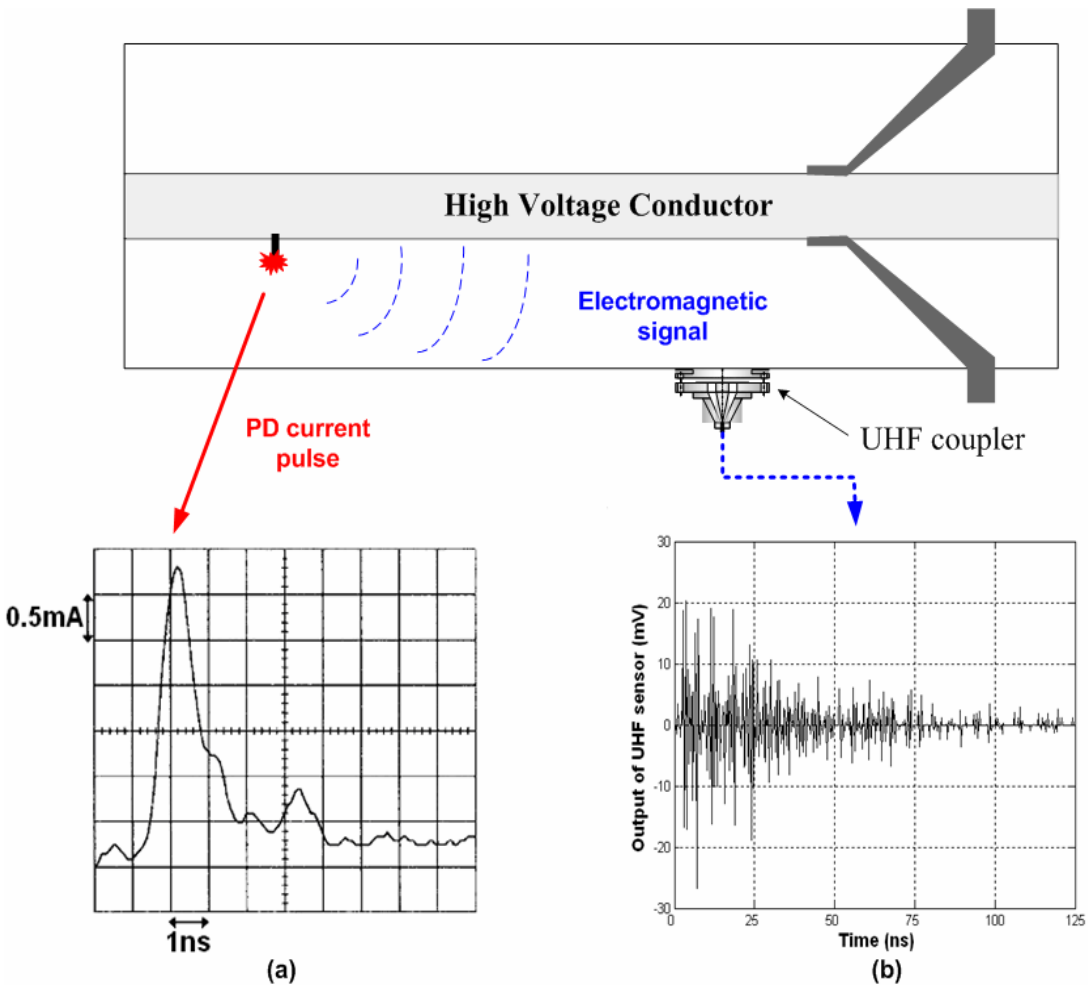


Fig. A.1 Typical UHF signal corresponding to single PD current pulse. (a) PD current pulse; (b) UHF signal results from a PD current pulse shown in (a).

A.1 Equipment Specifications

Equipment	Parameter	Description
Test Chamber	Inner diameter	180 mm
	Outer diameter	880 mm
	Length	10.3 m
	SF_6 pressure	0.2 MPa
	Spacer	cone type (x 2)
Sensor	Type	Conical type UHF coupler
	Frequency	200 MHz to 1.3 GHz
	Sensitivity	0.5 pC
	Inner diameter	43.4 mm
	Outer diameter	100 mm
	Operating	-25 to 70 degree Celsius
	Relative humidity	95% RH
Digital oscilloscope	Model	Tektronix TDS784D
	Bandwidth	1 GHz
	No. of channels	4
	Sampling rate	1 channel: 4 GS/s
		2 channels: 2 GS/s
		3 or 4 channels: 1 GS/s
	Maximum record	8M
Notebook PC	Model	Toshiba Tecra A2
	CPU	Pentium M Processor 715, 1.50 GHz
	Memory	256 MB
	Hard disk	40 GB
	Display	15.0” XGA TFT LCD

Table A.1 Equipment Specifications

A.2 The UHF Sensor

Based on the configuration of Fig. A.2, a conical UHF coupler is employed to detect PD signals. The disk size of the conical coupler must be arranged according to the frequency range of interest, since it determines the frequency characteristics of the coupler. On the other hand, the modes of pulse propagation along a coaxial system are the combination of the transverse electric and magnetic (TEM) mode, the transverse electric (TE) mode and transverse magnetic (TM) mode respectively. According to the configuration of the GIS section under test, PD pulse propagating in TEM mode may peak at around 100 MHz or upwards while pulses in TE or TM modes may peak in the range of 700 – 1100 MHz [88]. However, the mode of propagating pulse is dependent on whether the location of the PD source is on the bus conductor. Therefore, in order to have full coverage over the frequency range of the pulse propagating modes, the coupler with disk diameter of 43 mm is selected for the measurement.

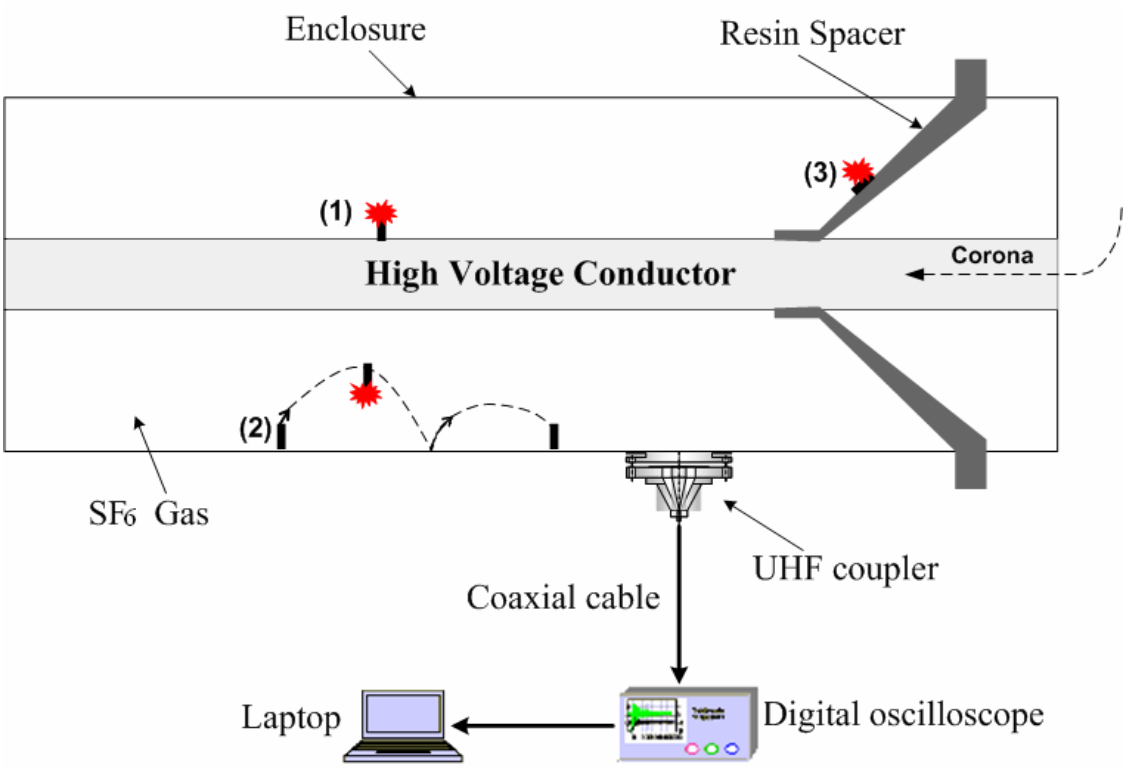


Fig. A.2 The layout of the test setup with a section of an 800 kV GIS

A.3 Experimental Set-up

UHF resonance signals used for the present study are measured from an 800 kV GIS chamber that has a total length of 20 m [89]. The test chamber is formed by isolating a 10.3 m section of the GIS using gas-tight conical epoxy barriers. It is filled with SF_6 gas at 0.2 MPa for the entire test. Power frequency is 50 Hz.

To detect the UHF signals caused by PD, an internal coupler electrode type sensor is incorporated into a hatch cover plate on the side of the test chamber. In addition to the sensor, the measuring system consists of a 3-meter long coaxial cable and a high-speed digital oscilloscope (TDS784D) enabling the system to acquire the high frequency components of the UHF signal as shown in Fig. A.2. The characteristic impedance of the sensor is 50 Ω, which is the same as the characteristic impedance of the cable and the oscilloscope. The triggering voltage of the digital oscilloscope is set to a level well above the background noise, enabling the capture of large UHF signals. The sampling rate of the oscilloscope is fixed at 4 giga-samples per second when measuring and recording the UHF signals.

To generate PD in SF_6, artificial defects are made using an aluminium needle with its length and section diameter of 10 and 0.2 mm respectively. As illustrated in Fig. A.2, the needle is placed on but not fixed to the enclosure to simulate the free particle. For the other two defects, it is either attached to the busbar or spacer surface using the minimum amount of cyanoacrylate adhesive, ensuring that the ends of the needle are clean and in contact with the surfaces. The distance between the needle and the sensor varies from 1 to 7.8 m to study the impact of signal attenuation.

The system is energized using a 2300 kV, 10 MVA single phase metal-clad transformer. Test voltage varies in the range from 40 to 160 kV rms. The PD inception voltages for the defects of free particle, particle on conductor and particle on the surface of spacer are 73, 110 and 158 kV rms respectively.

As illustrated in Fig. A.1, UHF signals excited by a single PD current pulse are measured for this study. The UHF signals usually last for several hundred nanoseconds. Typical waveforms of measured signals (including corona) and their frequency content obtained from Fast Fourier Transform (FFT) are shown in Figs. A.3 and A.4 respectively. In this study, data measured one meter away from the PD source, as shown in Table A.2, are used for developing the denoizing and source recognition method. In addition, the robustness of developed method is verified using data measured from other PD-to-sensor distances as shown in Table A.3.

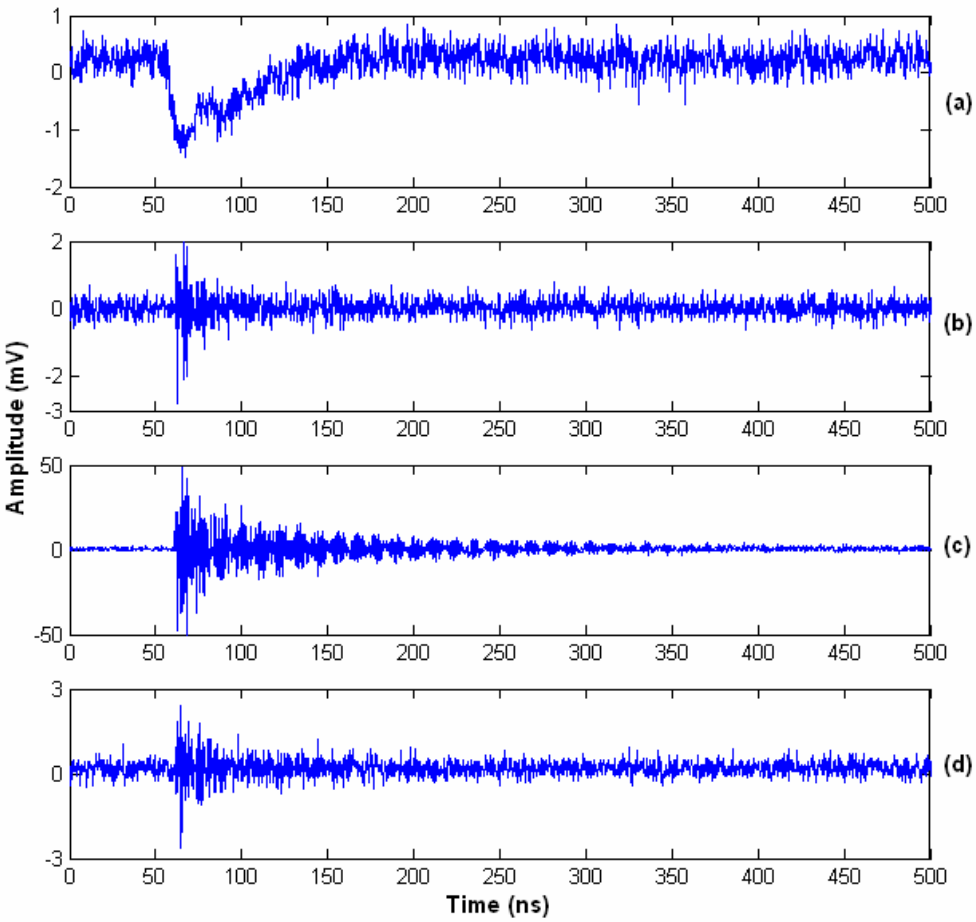


Fig. A.3 Typical waveform of measured signal (a) corona; (b) particle on the surface of spacer; (c) particle on conductor; (d) free particle on enclosure.

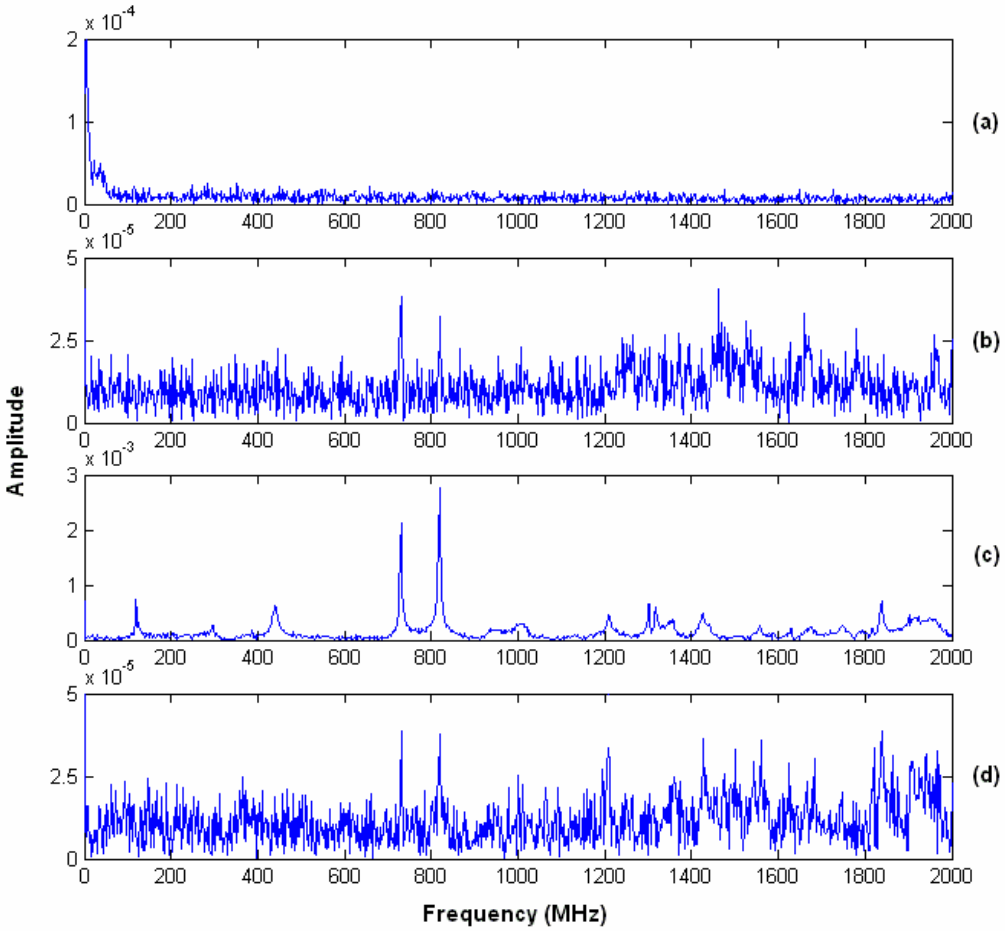


Fig. A.4 Frequency content of measured signal (a) corona; (b) particle on the surface of spacer; (c) particle on conductor; (d) free particle on enclosure.

Table A.2 Data measured one meter away from PD sources

Defect/noise	Number of signals
Corona	14
Particle on the surface of spacer	30
Particle on conductor	20
Free particle on enclosure	16

Table A.3 Data measured from other PD-to-sensor distances

Defect	Distance from PD source to sensor (m)	Number of signals
Particle on conductor	2.5	30
	4.3	30
	6	30
	7.8	30
Free particle on enclosure	2.5	20
	4.3	8
	6	8
	7.8	20

APPENDIX B

Discrete Wavelet Transform (DWT) and Wavelet Packet Transform (WPT)

The Discrete Wavelet Transform of a discrete signal $f(x)$ is defined as

$$\omega_{j,k} = \sum_{x=1}^{N} f(x) \frac{1}{\sqrt{2^j}} \psi\left(\frac{x - k2^j}{2^j}\right) \tag{B.1}$$

where N is the length of the discrete signal $f(x)$. j and k represent the scaling (decomposition level) and shifting (translation) constant respectively. j runs from 1 to j_{max}, which is given by $2^{j_{max}} \leq N$. $\psi\left(\frac{x - k2^j}{2^j}\right)$ is the scaled, shifted wavelet function (baby wavelet) of the original mother wavelet $\psi(x)$. The resultant wavelet coefficients thus reflect the resemblance between the signal and the baby wavelet.

The wavelet function $\psi\left(\frac{x - k2^j}{2^j}\right)$ is comparable to the sine or cosine basis functions in Fourier Transform. There are two characteristics required for any function to be considered as a mother wavelet:

1. The function must have zero average;
2. The function must decay quickly at both ends.

There are actually a large number of functions with such features available. However, the Mallat algorithm of DWT, which has been applied in this research, demands additional requirements as discussed below.

In 1988, a new DWT algorithm, which provides fast wavelet decomposition and reconstruction, was developed by Mallat [45]. Fig B.1 illustrates this wavelet decomposition algorithm. It is actually a classical scheme in the signal processing community, known as a two-channel sub-band coder using the conjugate quadrature filters or quadrature mirror filters (QMF) [45]. It decomposes the original signal $f(x)$ into coefficients of low-frequency (approximation coefficient or cA_i) and high-frequency (detail coefficient or cD_i) components.

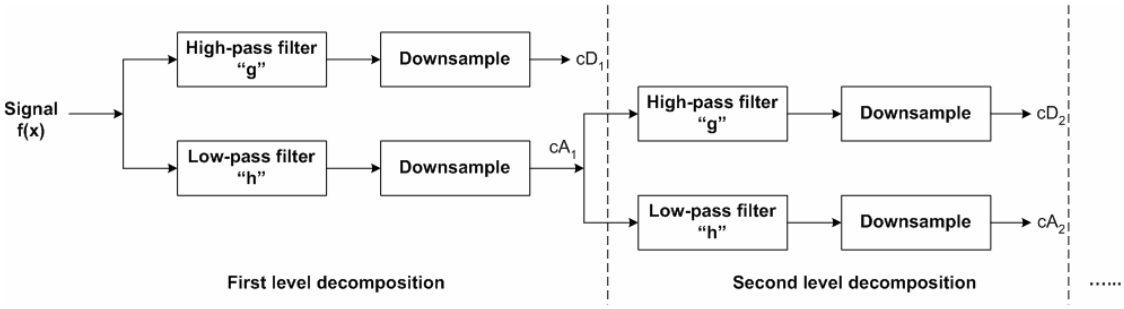


Fig. B.1 Fast DWT algorithm

According to the algorithm, there are two properties that allow the mother wavelet $\psi(x)$ in equation A.1 to have this fast algorithm:

1. Existence of a scaling function $\Phi(x)$;
2. Orthogonal results of the wavelet transform.

Though there are many wavelets available, only several wavelet families possess these properties, such as the Symlet, the Coiflet and the Daubechies.

The scaling function $\Phi(x)$ is used to generate a pair of high-pass and low-pass filters, namely the *g* and *h* in Fig B.1. Using these filters, DWT generates the cA_i and cD_i at

different levels. The decomposition coefficients are obtained by convolving the original signal $f(x)$ (or cA_i) with high-pass filter or lower-pass filter. In this algorithm, when a signal passes through the two filters concurrently, double amount of data will be produced. By discarding every other data coming out of the filters, the signal is downsampled. Though this downsampling process introduces distortion known as aliasing, it has been proved that the effect is completely eliminated by employing the appropriate filters [45].

To reconstruct the original signal, the inverse discrete wavelet transform (IDWT) is carried out involving two steps as the decomposition, namely the upsampling and filtering of the wavelet coefficients. The upsampling process means lengthening a signal component by inserting zeros between samples. Subsequently, the upsampled coefficients will be input into the reconstruction filters to generate the reconstructed signal.

The wavelet coefficient cA_i contains lower half frequency content of the decomposition filter input, and the corresponding cD_i contains the upper half frequency content. In addition, these coefficients is well localized in time domain, so that both time and frequency information of the original signal are kept. Furthermore, the coefficients have greater resolution in time for high frequency components and greater resolution in frequency for low frequency components of a signal. The highest frequency content contained in the wavelet coefficients is up to $\frac{f_0}{2}$, where f_0 is the sampling frequency of the original signal. This limitation is attributed to the Nyquist sampling criterion. Fig. B.2 shows the coverage of the time –frequency plane for the DWT coefficients.

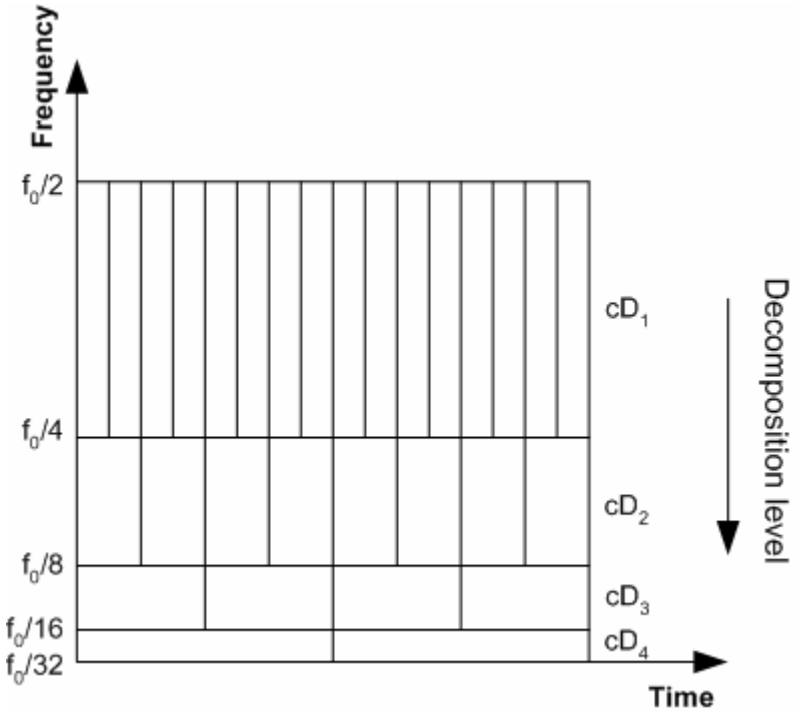


Fig. B.2 The coverage of the time-frequency plane for DWT coefficients

DWT coefficients of four level decompositions are illustrated in Fig. B.2. As observed, cD_1 contains from $\frac{f_0}{2}$ to $\frac{f_0}{4}$ content of the original signal, and has high resolution in time. cD_2 contains from $\frac{f_0}{4}$ to $\frac{f_0}{8}$ content of the original signal, and has lower resolution in time (half that of cD_1). In brief, as the decomposition level increases, the time resolution decreases, while the frequency resolution increases.

The wavelet packet analysis is a generalization of wavelet decomposition that offers a richer signal analysis. In the wavelet decomposition procedure, the process of splitting into low-frequency and high-frequency components is only applied to the approximation components. The detail components are never re-analyzed. In the wavelet packet situation, each detail component is also split into two parts using the same approach as in approximation splitting. This enables the analysis of high

frequency components of the original signal in a higher resolution. Therefore, the wavelet packet transform is applied to denoizing and feature extraction in this research.

APPENDIX C

Genetic Algorithm

Genetic algorithms (GAs) were formally introduced in the United States in the 1970s by John Holland at University of Michigan. They are search algorithms based on the mechanics of natural selection and natural genetics. The fundamental principle is that the fittest member of a population has the highest probability for survival. Generally, GAs have the following components [49]:

1. A genetic representation for potential solutions to the problem;
2. A way to create an initial population of potential solutions;
3. An evaluation function that rates solutions in terms of their fitness;
4. Genetic operators that alter the composition of offspring during reproduction;
5. Values for the various parameters used by GA, such as population size, probabilities of applying genetic operators, and so on.

In each candidate solution, the decision variables to the problem can be binary-coded and concatenated as a string (chromosome). Strings are grouped into sets known as populations. Successive populations are called generations. GAs first form an initial population randomly. Then each string is evaluated to find its fitness by substituting into the fitness function. Based on the merits of different strings, a new set of strings (population) is created using GA operators, namely reproduction, crossover and mutation. The above process is iterated until a pre-specified stop criterion such as the maximum number of generations has been reached. Details of the GA operators are discussed in the following sections.

C.1 Reproduction

The reproduction operator involves choosing a number of individuals according to fitness that will be used for breeding. The purpose of reproduction is to give more reproductive chances to those individuals that have high fitness values. This can be implemented in many ways, such as the roulette wheel selection [74] and tournament selection [75].The roulette wheel selection is adopted in this research.

The idea behind the roulette wheel selection technique is that each individual is given a chance to become a parent in proportion to its fitness. It is called roulette wheel selection as the chances of selecting a parent can be seen as spinning a roulette wheel with the size of the slot for each parent being proportional to its fitness. Obviously those with the largest fitness (slot sizes) have more chance of being chosen. Thus, it is possible for one member to dominate all the others and get selected a high proportion of the time. Roulette wheel selection can be implemented as follows:

1. Sum the fitness of all the population members. Call this *TF* (total fitness).
2. Generate a random number *n*, between 0 and *TF*.
3. Return the first population member whose fitness added to the preceding population members is greater than or equal to *n*.

C.2 Crossover

Crossover is a process that randomly takes two reproduced strings (parents) and exchanges portions of the strings to generate two new strings (offspring) with a

predetermined crossover probability. The purpose of the crossover operator is to combine useful parental information to form new and hopefully better performing offspring. Such an operator can be implemented in the following three ways.

1. Single point crossover.

   The strings of the parents are cut at some randomly chosen common point and the resulting sub-strings are swapped. For instance, if P_1=1 1 0 | 1 0 1 1, P_2=1 0 1 | 0 0 1 0, and the crossover point is between the 3th and 4th bits (indicated by "|"), then the offspring would be O_1=1 1 0 | 0 0 1 0 and O_2=1 0 1 | 1 0 1 1.

2. Two point crossover.

   The strings are thought of as rings with the first and last bit connected, namely wrap-around structure. The rings are cut in two sites and the resulting sub-rings are swapped. For example, consider two strings P_1=1 | 1 0 0 | 0 0 1, P_2=0 | 1 0 1 | 1 1 0, and the crossover points are between 1st and 2nd bits and between 4th and 5th bits. In this case, it generates two strings: O_1=1 | 1 0 1 | 0 0 1 and O_2=0 | 1 0 0 | 1 1 0.

3. Uniform crossover.

   Each bit of the offspring is selected randomly from the corresponding bits of the parents.

The single point crossover is employed in this research.

C.3 Mutation

Selection and crossover alone can obviously generate a large amount of differing strings. However, depending on the initial population chosen, there may not be enough variety of strings to ensure the GA sees the entire problem space. Or the GA may find itself converging on strings that are not quite close to the optimum it seeks due to a bad initial population. Above issues are addressed by introducing a mutation operator into GA. Mutation randomly alters each bit with a small probability, typically less than 1%. This operator introduces innovation into the population and helps prevent premature convergence on a local maximum.

APPENDIX D

Independent Component Analysis and FastICA Algorithm

Independent Component Analysis (ICA) is a statistical technique for finding hidden factors that form sets of measured signals. In the most fundamental ICA model, the measure data are assumed to be linear or nonlinear mixtures of some unknown latent components, and the mixing system is also unknown. The unknown components are assumed to be statistically independent of each other - hence the name *Independent* Component Analysis. ICA algorithms are able to estimate both the unknown independent components and the mixing matrix from the measure data with very few assumptions as follows [59]:

1. The unknown components are assumed *statistically independent*.
2. The unknown components must have nongaussian distributions.
3. The unknown mixing matrix is assumed to be square.

In this research, it is reasonable to make such assumptions, as the factors that affect the measured signals such as sensor response, propagation path and defects are independent and usually nongaussian distributed.

In practice, there are several approaches to find the unknown independent components, which use certain statistical properties of the components, such as nongaussianity, temporal structure, cross-cumulants and nonstationarity [76]. In this research, the

nongaussianity of unknown components is utilized in the implementation of ICA, known as FastICA algorithm.

The nongaussianity of a vector can be measured by its higher-order statistics such as kurtosis, skewness and negentropy. The negentropy is adopted in this thesis due to its proven robustness to noises [59]. However, it is computationally very difficult to calculate negentropy directly, as an estimate of the probability density function is required. Therefore, it is highly desired to use simpler approximations of negentropy. The approximated negentropy for a random vector y is defined as

$$J(y) \approx [E\{G(y)\} - E\{G(v)\}]^2 \tag{D.1}$$

where v is a Gaussian variable of zero mean and unit variance and G is any non-quadratic function.

To find the independent components, the approximated negentropy of the potential solution $w^T x$ is maximized by FastICA which is based on a fixed-point iteration scheme. Denote by *g* the derivative of the function G used in (D.1). Then the FastICA algorithm is given as follows:

(1) Pre-process observed signals to obtain x by centering and whitening.

(2) Let N denote the number of independent components. Set counter t = 1.

(3) Initialize w_t randomly.

(4) Let $w_t \leftarrow E\{xg(w_t^T x)\} - E\{g'(w_t^T x)\}w_t$.

(5) De-correlate outputs by $w_t \leftarrow w_t - \sum_{j=1}^{t-1} (w_t^T w_j) w_j$.

(6) let $w_t \leftarrow w_t / \|w_t\|$.

(7) If not converge, go back to 4.

(8) let t = t + 1.

(9) If $t \leq N$, go back to 3. Otherwise, stop.

In practice, the expectations in FastICA are replaced by their estimates, namely the sample means.

APPENDIX E

General Introduction to Neural Networks

A neural network is an information processing paradigm that was inspired by the way biological nervous systems, such as the brain, process information. The field goes by many names, such as connectionism, parallel distributed processing, neuro-computing, natural intelligent systems, machine learning algorithms, and artificial neural networks. It is an attempt to simulate the multiple layers of simple processing elements called neurons within specialized hardware or sophisticated software. Each neuron is linked to its neighbors with varying coefficients of connectivity that represent the strengths of these connections. Learning is accomplished by adjusting these strengths to cause the overall network to output appropriate results.

The function of neural networks is largely dependent on the network structure that is determined by the way neurons connected. There are basically four types of connections as follows:

1. *Feedforward connections*:

   In this network structure, data from neurons of a lower layer are propagated forward to neurons of an upper layer via feedforward connections. Multilayer perceptron is a typical feedforward neural network.

2. *Feedback Connections*:

Feedback networks bring data from neurons of an upper layer back to neurons of a lower layer. This type of connection is usually employed in neural-network-based controller.

3. *Lateral Connections*:

   Neurons of the same layer are interconnected. One typical example of a lateral network is the self-organizing map.

4. *Time-delayed Connections*:

   Delay elements may be incorporated into the connections to yield temporal dynamics models. They are more suitable for temporal pattern recognitions.

One of the most interesting properties of a neural network is the ability to learn from its environment in order to improve its performance over time. Generally, the learning methods of neural networks can be classified into two categories:

1. *Supervised learning*:

   In supervised learning, the desired output pattern corresponding to an input is presented to the network during training in order to guide learning. The network learns in the training phase by having its weights adjusted such that the actual network output becomes more similar to the desired network output. Thus, the desired output acts as an external teacher in this type of learning.

2. *Unsupervised learning*:

   This type of learning uses no external teacher and is based upon only local information. It is also referred to as self-organization, in the sense that it self-

organizes data presented to the network and discovers their emergent collective properties.

APPENDIX F

Resilient Back-propagation Algorithm

The choice of the learning rate η for the standard back-propagation algorithm in equation E.1, which scales the derivative of the error function, has an important effect on the time needed until convergence is reached.

$$\Delta w_{ij}^{(t)} = -\eta \frac{\partial E}{\partial w_{ij}}(t) \tag{E.1}$$

If η is set too small, too many steps are needed to reach an acceptable solution. On the contrary, a large learning rate will possibly lead to oscillation, preventing the error to fall bellow a certain value.

On the other hand, MLP networks typically use sigmoid transfer functions in the hidden layers. The functions are characterized by the fact that their slope must approach zero as the input gets large. This causes a problem when using steepest descent to train a MLP network with sigmoid functions, since the gradient can have a very small magnitude leading to a small learning rate; and therefore, cause small changes in the weights and biases, even though the weights and biases are far from their optimal values.

The basic principle of Resilient Back-propagation Algorithm is to eliminate the harmful influence of the size of the partial derivative on the learning rate. This algorithm considers the local topology of the error function to change its behaviour. As

a consequence, only the sign of the derivative is considered to indicate the direction of the weight update. The size of the weight change is exclusively determined by a *update-value* $\Delta_{ij}^{(t)}$:

$$\Delta w_{ij}^{(t)} = \begin{cases} -\Delta_{ij}^{(t)} & \text{if } \frac{\partial E^{(t)}}{\partial w_{ij}} > 0 \\ +\Delta_{ij}^{(t)} & \text{if } \frac{\partial E^{(t)}}{\partial w_{ij}} < 0 \\ 0 & \text{else} \end{cases} \tag{E.2}$$

where $\frac{\partial E^{(t)}}{\partial w_{ij}}$ is the summed gradient information over all patterns of the pattern set.

Each *update-value* evolves during the learning process according to its local sight of the error function E. This is based on a sign-dependent adaptation process:

$$\Delta_{ij}^{(t)} = \begin{cases} \eta^{+} * \Delta_{ij}^{(t-1)} & , \quad if \ \frac{\partial E^{(t-1)}}{\partial w_{ij}} * \frac{\partial E^{(t)}}{\partial w_{ij}} > 0 \\ \eta^{-} * \Delta_{ij}^{(t-1)} & , \quad if \ \frac{\partial E^{(t-1)}}{\partial w_{ij}^{(t-1)}} * \frac{\partial E^{(t)}}{\partial w_{ij}} < 0 \\ \Delta_{ij}^{(t-1)} & , \quad else \end{cases} \tag{E.3}$$

where $0 < \eta^{-} < 1 < \eta^{+}$ (η^{-}=0.5, η^{+}=1.2 in this thesis).

Note that the *update-value* is not influenced by the magnitude of the derivatives, but only by the behaviour of the sign of two succeeding derivatives. Every time the partial derivative of the corresponding weight changes its sign, which indicates that the last update is too big and the algorithm has jumped over a local minimum, the *update-*

value $\Delta_{ij}^{(t)}$ is decreased by the factor η^-. If the derivative retains its sign, the *update-value* is slightly increased in order to accelerate convergence in shallow regions. Thus, Resilient Back-propagation Algorithm generally converges much faster than other back-propagation algorithms.