CONDITION CLASSIFICATIONIN UNDERGROUD PIPES BASEDON ACOUSTICAL CHARACTERISTICS

ZAO FENG

PhD

UNIVERSITY OF BRADFORD

2013

Condition Classification in Underground Pipes Based on Acoustical Characteristics

Acoustical characteristics are used to classify the structural and operational conditions in underground pipes with advanced signal classification methods

Keywords: Acoustics
Condition classification
Underground Pipes
Sound intensity
Signal processing
Machine learning

Zao Feng

Submitted for the degree of Doctor of Philosophy

School of Engineering, Design and Technology
University of Bradford

August 2013

Abstract

Keywords: Acoustics, condition classification, pipes,sound intensity, signal processing, machine learning.

“Pattern recognition is a fast-moving and proliferating discipline. It is not easy to form a well-balanced and well-informed summary view of the newest developments in this field. It is still harder to have a vision of its future progress”.

Watanabe in 1972

This thesis is concerned with the development and study of a pattern recognition system for siphon and sewer condition/defect analysis based on acoustic characteristics. Pattern recognition has been studied and used widely in many fields including: identification and authentication; medical diagnosis and musical modelling. Audio based classification and research has been mainly focusing on speech recognition and music retrieval, but few applications have attempted to use acoustic characteristics for underground pipe condition classification. Traditional CCTV inspection methods are relatively expensive and subjective so remote techniques have been developed to overcome this concern and increase the inspection efficiency. The acoustic environment provides a rich source of information about the internal conditions of a pipe. This thesis reports on a classification system based on measuring the direct and reflected acoustic signals and describing the energy spectrum for each condition/pipe defect. A K-nearest neighbour classifier (KNN) and Support vector machines (SVMs) classifier have been adopted to train the classification system to identify sediment and pipe surface defects by comparing the measured acoustic signals with a database containing a range of typical conditions. Laboratory generated data and field collected data were used to train the proposed system and evaluate its ability. The overall accuracy of the system recognizing blockage and structural aspects in each of the series of experiments varies between 70% and 95%.

Acknowledgment

I am very grateful to my supervisors for giving me the opportunity to study under their supervision. I would like to express my deepest gratitude to Professor Kirill V. Horoshenkov for not giving up on me when I was struggling and going through a very rough phase. I wouldn't have done it without your encouragement and guidance. Many many thanks go to Professor Simon Tait for always being so supportive and understanding. Thanks to my colleagues in acoustic group for their help the time we spent together. I would also like to thank all the technicians in the Hydraulic laboratory who helped me with my experiments and fixed every problem I caused. Finally, thanks my family for always being there for me and all the people who have supported and helped me during the time.

Contents

Table of Notations

$\psi(.)$	Wavelet function
$\phi(.)$	Scaling function
j	Wavelet translation factor
k	Wavelet scale factor
cA	Approximation coefficient of wavelet decomposition
cD	Detail coefficient of wavelet decomposition
V_j	Feature space at resolution 2^j
W_j	Space orthogonal to V_j
h_d	low-pass filter
g_d	high-pass filter
σ	Least square error
P_M	Numerator polynomials of degree M
Q_N	Denominator polynomials of degree N
R_N^M	Quotient of P_M and Q_N , Padé approximation
$O(.)$	Landau's big-Oh symbol
$\{a_0, a_1 \ldots a_m\}$	Numerator coefficients of Padé approximation
$\{b_0, b_1 \ldots b_n\}$	Denominator coefficients of Padé approximation
$\mathbf{A}^{\dagger}$	More-Penrose pseudoinverse
$p(x)$	Density function
P	Probability estimation
K	Number of nearest neighbours
V	Sample volume
D_r	Euclidean distance on r -th dimension
$\mathbf{w}$	Weight vector of a classifier
$\mathbf{b}$	Bias of a classifier
Φ	Lagrangian

γ	Hyperplane margin
ξ	Slack variable
C	Soft margin constant
α_i	Lagrange multipliers
$k(x_i, x_j)$	Kernel function
$r_{x,y}$	Cross-correlation coefficient
$h(t)$	Acoustic impulse response
$p(t)$	Acoustic pressure
$u(t)$	Acoustic (particle) velocity
$I(t)$	Acoustic intensity
ρ_0	Density of air
c_0	Sound speed in air
λ	Wavelength
e, E	Acoustic energy and energy spectrum
W	Wavelet entropy
f_s	Sampling frequency
KNN	K-Nearest Neighbours
SVM	Support Vector Machines
$SMOTE$	Synthetic minority oversampling technique

List of Figures

List of Tables

Chapter 1

Introduction to the Project

1.1 Background

A US Environmental Protection Agency (USEPA) report titled “Distribution system Inventory, Integrity and Water Quality” [1] indicates that condition assessment of buried infrastructure is either not used or not used routinely by most utilities, and utilities often have limited data about their systems beyond what was available when the infrastructure was installed. The emphasis on evaluating the condition of underground pipes in the water industry has increased during the past decade. Therefore, it is necessary to identify a range of appropriate techniques that together will provide sufficient information on the condition of pipes to make rational and informed decisions about rehabilitation or replacement. A wide range of both direct and indirect techniques for condition assessment are now available including near field and remote field electromagnetic techniques, acoustic, stray current monitoring, visual inspection, and soil surveys [2].

The structural condition and hydraulic capacity of water pipes deteriorate because of aging in which the physical condition of the pipe changes and so adversely affect the system performance. Therefore, the need for inspecting and assessing the condition of pipes in water and sewer system is increased in order to maintain and upgrade such system. Condition assessment of pipes is challenging compared to other infrastructure assets because they are typically underground and mostly they are inaccessible [3]. Different pipe materials, the type of information and the level of accuracy required, will determine the type of inspection and assessment techniques used. Traditionally, sewer surveys were carried out by sending out inspectors to 'see and touch' the defects inside those man-entry pipes along the network. However, this method, although highly effective at revealing the internal condition and providing certain clues about the external condition, suffers from inefficiency in terms of manpower and has significant health and safety risks. It is obviously impractical for the majority of the smaller pipes that make up the majority of the sewer network [4]. As a result, remote techniques were developed to overcome this concern and greatly increase the inspection efficiency.

Acoustic based techniques are increasingly adopted in condition assessment of utilities since they provide the possibility to measure the location and characteristics of individual defects inside a pipe. Reflected acoustic signals in a sewer pipe can contain vital information about the internal conditions, the signals can be measured at distances sufficiently far from the target. This allows the measurement to be taken in a live pipe which otherwise may not be suitable for inspection with any other monitoring methods. But at the same time, acoustic sensors can pick up noise from other unwanted sources of sound in the vicinity of the sensor. Therefore, it is necessary to apply powerful signal processing techniques to distinguish the effects which can represent the defects from those due to harmless ones.

Most signal analysis instruments utilize a Fast Fourier Transform (FFT) to convert the signals from its time domain representation to its equivalent frequency domain representation and vice versa. It is thought that the frequency spectrum will have a characteristic shape responding to a particular condition pipe condition, but in most of cases this spectrum needs to be processed to remove undesirable noise.

Several modelling approaches have been developed to predict and evaluate the occurrence of sewer blockages and failures. Sewer failures can be modelled following two different possible approaches: physically based or statistical [5]. In general physical models are used where the cost of failure is significant enough to justify the cost of detailed surveys, and the statistically derived models may be applied to less critical sewer pipes for condition monitoring and failure prediction. New data mining techniques proved to be more efficient than classic statistical tools in modelling pipe defects [6]. Statistical based condition modelling requires assessment data are collected and a methodology is available to efficiently extract information from the data. This automated analysis of datasets is performed to determine significant patterns among data. There are many data mining and pattern recognition technologies (Decision Tree, Rule Induction, Statistical analysis, Artificial Neural Networks, etc.), but not all are useful for every type of problem. *Savic et al.* [7] usedclassification tree and rule induction algorithms to derive statistical relationships to predict the likelihood of sewer failure for different pipe classes. *Giustolisiet al*. [8] used Evolutionary Polynomial Regression (EPR) to describe the relationships among data and to discover new knowledge about the factors influence pipe breaks.

This thesis reports on the development and study of the performance of a novel pattern recognition system that enables us to relate the condition of a pipe with its corresponding acoustic characteristics. The assessment of pipe conditions is based on measuring the direct and reflected acoustic signals which are excited in by a pipe with obstructions, sediments or structural defects. The system is capable of identifying sediment and pipe surface defects by comparing the measured acoustic response signals with a signature database covering a number of typical defects.

1.2 Objectives of the Research

The main objectives of the research are:

(1) To develop the feature extraction methods using the acoustic signals collected from water and air filled pipes;

(2) To investigate a range of classification techniques in application to condition classification in pipes;

(3) To develop a new pattern recognition system based on suitable feature extractor and acoustic characteristics;

(4) To study and evaluate the performance of the system in the laboratory and in the field.

The research was carried out in three main stages:

(1) Development of a statistical system for defect pattern classification using acoustic sound pressure level and energy as main features to discriminate different siphon conditions;

(2) Development of new algorithms based on the acoustic intensity characteristics of pipes containing a range of common operational and structural.

(3) Study of the performance of the new classification system through further regularisation of the training datasets.

1.3 Novel Contribution of the Research

(1)Developed a multiple class condition recognition system which is suitedto classify and identify the underground sewer conditions based on acoustic energy characters. The structure of the recognition system is given in Figure 1.1.

(2)Acoustic energy and corresponding spectrum were proved to be informative and be able to provide distinguishable coefficients for pipe condition and defect classification using suitable data fitting tools.

(3) The proposed classification system can be adopted to work with 1-dimensional and 2-dimensional features for binary and multiple class pipe conditions.

(4) The accuracy rates of identifying underground pipe conditions and defects were achieved between 60% and 95% approximately.

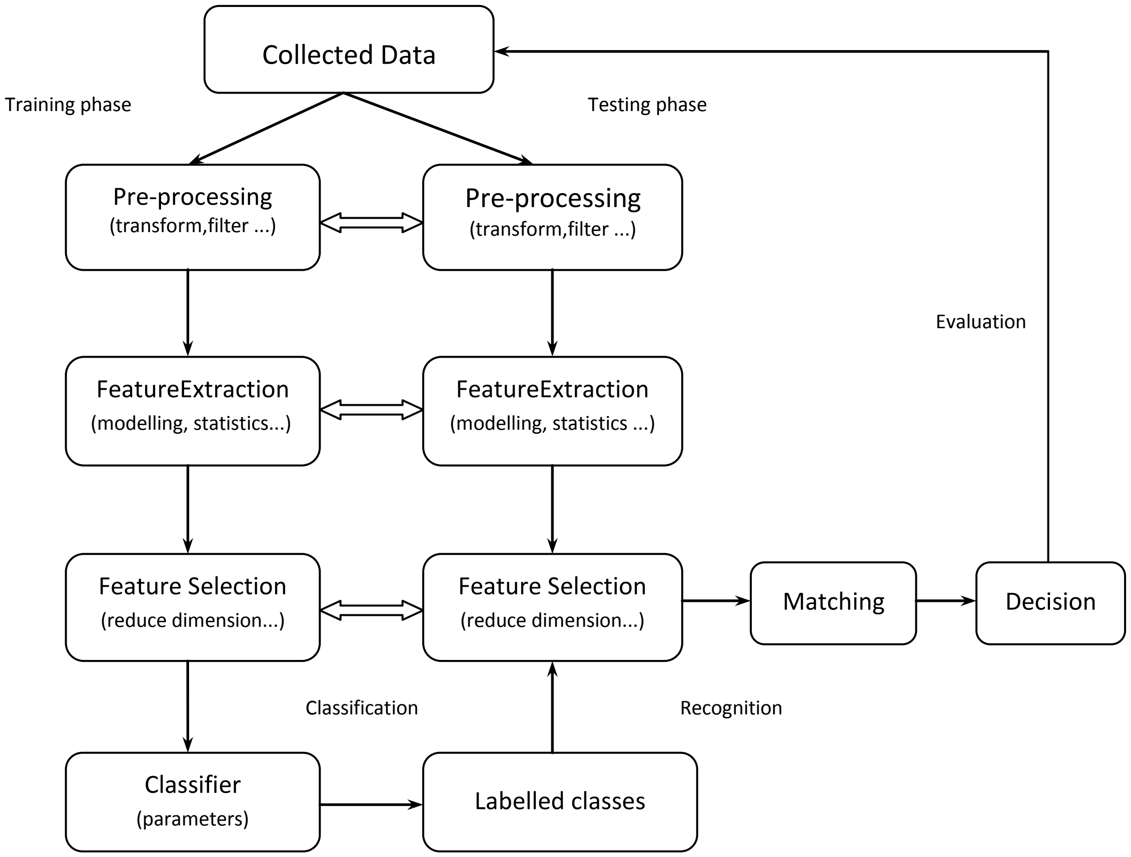


Figure 1.1 The structure of the proposed pattern recognition system

1.4 Structure of the Thesis

This thesis is organised as follows. Chapter 2 presents a literature review of current condition assessment technologies for underground infrastructure, the state-of-the-art feature extraction techniques and classification methods. In Chapter 3, mathematical and theoretical background of a robust pattern recognition system is reviewed, which includes signal pre-processing methods as well as feature extraction techniques and classification algorithms which were adopted in this research. In Chapter 4, details are given of the new experimental facility comprising of a full scale siphon and a set of acoustic sensor device that was developed for data acquisition and subsequent acoustic condition classification. This chapter also details the digital filter and wavelet transform analysis as well as K-nearest neighbours (KNN) classifier algorithm which were used for the purpose of feature extraction and condition classification. In Chapter 5, the results of measurements in a 150mm diameter, 14.4m long clay pipe under a range of simulated conditions. Polynomial and Padé approximation were used to extract features from obtained acoustic signatures, and then KNN and Support Vector Machines (SVMs) were adopted to train and classify the pipe condition. Chapter 6 presents the application of the classification algorithm detailed in Chapter 5 to the data collected in the underground pipes in the field. Here imbalanced learning problems and their solutions are discussed and compared. Chapter 7 presents a summary of this work together with final recommendations for their dissemination for the direction of future work.

References

[1] American Water Works Association, *"Distribution System Inventory, Integrity and Water Quality",* United States Environmental Protection Agency. 2007.

[2] S.B. Costello et al."Underground asset location and condition assessment technologies". *Tunnelling and Underground Space Technology* , vol 22,pp. 524–542. 2007.

[3] T. Hao, et al. "Condition assessment of the buried utility service infrastructure". *Tunnelling and Underground Space Technology* , vol 28, pp.331-344. 2012.

[4] Zheng Liu and Yehuda Kleiner, "State of the art review of inspection technologies for condition assessment of water pipes". *Measurement* , vol 46 issue 1, pp 1-15. 2013.

[5] Dae-Hyun Koo and Samuel T. Ariaratnam, "Innovative method for assessment of underground sewer pipe condition". *Automation in Construction* , vol 15, pp.479-488. 2006.

[6] Vladan Babovic, et al. "A data mining approach to modelling of water supply assets". *Urban Water* , vol 4, pp.401-414. 2002.

[7] D.A. Savic and O. Giustolisi, "A Symbolic Data-driven Technique Based on Evolutionary Polynomial Regression". *Journa of Hydroinformatics* , vol 8, pp.207-222. 2006.

[8] O. Giustolisi, et al. "A multi-model approach to analysis of environmental phenomena". *Environmental Modelling & Software* , Volume 22, Issue 5,pp.674-682. 2007.

Chapter 2

Literature Review

Underneath today's cities exists an extensive and complex network of pipes providing the essential utility services that underpin the modern civilised life. With the ageing of this buried infrastructure and growing demand for the increase in its capacity due to the expansion of the population and the development of new technologies, it is vitally important to monitor and assess their condition throughout their life cycles to avoid major potential failures due to their deterioration. The complexity of the underground pipe networks derives from a great variation in the age, pipe materials and types of pipe design which represent the existing underground infrastructure [1]. This chapter reviews the state-of-the-art methods for condition assessment of underground utilities (especially water and sewage pipelines). Among these, acoustic methods for the inspection of pipes have been used extensively with primary applications related to the quality control and condition monitoring of pipes used in oil and gas industries, the water and sewage industries and chemical engineering [2, 3, 4].

2.1 Condition Assessment Methods of Underground Utilities

The pipe condition assessment can be defined as “the collection of data and information through direct and/or indirect methods, followed by analysis of the data and information, to make a determination of the current and/or future structural, water quality, and hydraulic status of the pipeline” by the US Environmental Protection Agency [5]. Condition assessment methods can be roughly categorized into direct and indirect methods [6]. Direct methods include automated and manual visual inspection, pipe sampling and non-destructive testing. Indirect methods include water audit, flow testing, and measurement of soil resistivity to determine the risk of deterioration.

There is a wide range of direct and indirect techniques for determining the existing condition of a pipeline and the rate of its deterioration. Indirect methods are relatively simple and less costly than direct intrusive methods. Indirect techniques do not require access to either the internal or the external surface of the pipe and, therefore do not disrupt operations or require local excavations. However, indirect methods may not provide the level of detail, timeliness, or confidence required for maintenance and renewal decisions about pipes with a high consequence of failure [6].

2.1.1 Indirect Condition Assessment Techniques for Water Mains

(1) Historical data such as the age of pipe, manufacturer and experience of various pipe materials;

(2) Environmental techniques include a consideration of the chemistry of the water and surrounding soil (e.g. soil conditions, ground water tables, surface conditions). Soil characterization is used to explore the soil parameters relevant to the deterioration of buried pipes. Following are some soil parameters of interest: soil resistivity; pH value; Redox potential; moisture content; shrink/swell capacity; buffering capacity etc [7].

(3) Operational data such as flow, maintenance and repair records. This information, from which pipe and/or network condition can be inferred, coupled with information about potential consequences of failure, is of great value in focusing an investigation strategy to those sections in most need of assessment.

2.1.2 Direct Condition Assessment Methods

2.1.2.1 Visual Inspection Techniques

(1) Closed-Circuit Television (CCTV)

Current alternatives to direct man-entry and visual observation include the collection and inspection of CCTV images, or the use of Light Line surveys. However, these methods are slow, largely subjective and may require a sewer length to be drained or pre-cleaned before inspection, and thus they are expensive [1]. Using CCTV to inspect the interior of pipes was introduced in 1960s. The system consists of a television camera mounted on a tractor and remotely controlled by an operator. The CCTV inspection system has been widely used for decades and the basic principle has remained the same. The obvious advantage of this method is that it provides direct illuminated images of the defects of the pipe’s interior wall, which can be examined in detail by zooming the camera or viewing from different angles by controlling the position of the tractor. The natural limitation of this technique is that the images of the interior wall can only be obtained above the water surface in the case of sewer and water pipes. Since the CCTV tractor travels along the pipeline, unsteady camera movement and lack of geometric references are considered to be further limitations of the technique [8].

Recent research has focused on how to improve the quality of the inspection images, how to improve the interpretation of poor-quality images, and how to improve the automation of the inspection. *Sarshar et al.* [9] proposed a software system to semi-automatically extract historical condition data information from sewer inspection CCTV files, *Cherqui et al*. [10] proposed an algorithm which calibrates dysfunction indicators based on the results of visual inspections. *Yang et al*. [11] proposed the use of a CCTV image quality index to improve the inspection confidence when compared to reference images.

(2) Sewer Scanner and Evaluation Technology (SSET)

In late 1990s, optical scanner and gyroscope techniques were introduced to facilitate pipe interior inspection. The SSET is a flexible non-destructive evaluation data acquisition tool. Unlike CCTV inspection, the SSET device does not need to stop at the defect locations and provide the engineers with the ability to see the total surface of the pipe from one end to the other, which conceptually increases the inspection efficiency [8]. The scanned image is then digitized so that a colour-coded computer generated image of the pipe wall can be obtained.

The major benefit of SSET system over the CCTV technology is that higher quality information is provided for the assessment. However, this techniques still requires manual interpretation of the images and ever higher level of expert assessment. Research in this area has been undertaken to automate the assessment process in order to increase efficiency and interpretation accuracy [12, 13]. In the deployment of SSET in the practical pipeline inspection, commercial systems such as PANORAMO system by RapidView IBAK and SOLO system by RedZone have been deployed in pipeline inspection with high image resolution, inspection speed and efficiency [13].

2.1.2.2 Electromagnetic Methods

(1) Magnetic Flux Leakage (MFL)

The MFL technique is widely acknowledged and used for metallic pipeline inspections. A pipeline inspection gauge (pig) is normally inserted into the system and travels along the pipeline, and it is used to detect and characterize the metal loss defects such as corrosion and cracks on the interior wall of the pipeline [14]. MFL technology is claimed to have good detection capabilities even for small pitting anomalies, attributed to the fact that the MFL pattern registered by the inspection tool is larger than the anomaly itself. MFL technology is therefore potentially suitable for detecting very small pitting defects because even under extremely poor conditions, a magnetic response is still obtainable. However, the use of MFL in water industry is limited to cleaned, unlined pipes and also requires accessibility to the pipes’ exterior [5].

(2) Eddy Current Testing

Eddy current testing tends to be used for smaller diameter metallic pipes, e.g., down to 100 mm diameter pipes [8]. The principle of the eddy current technique is based on the interaction between a magnetic field source and the test material. This interaction induces eddy currents in the test piece, engineers can detect the presence of very small cracks by monitoring changes in the eddy current flow [15].

In eddy current testing, a time varying magnetic field is induced in the pipe by using a magnetic coil with alternating current. This magnetic field causes an electric current to be generated, which in turn produces small magnetic fields around, conducting materials. The smaller magnetic fields generally oppose the original field, which changes the impedance of the magnetic coil. Thus, by measuring the change in impedance of the magnetic coil as it traverses the pipe, different characteristics can be identified [16].

The strength of the eddy current is related to the pipe wall thickness and one drawback of eddy current testing is the dimension of the skin depth that is examined, which is dependent on the induced frequency (e.g. for steel pipes at 50 Hz the skin depth is about 3 mm) [8]. To overcome this problem, the Remote Field Eddy Current (RFEC) method was developed [5]. The RFEC technique uses an internal probe to inspect conducting tubes non-destructively. This method relies on the fact that the remote field signal is larger than the direct eddy current signal measured by the detector coils. As the Remote Field principle works in low frequencies (typical 10Hz ~ 1kHz), the inspection speed is therefore limited [17]. The frequency is adapted to the material and wall thickness and the signal amplitude shows the volume of the defects [17].

(3) Broadband Electromagnetic (BEM)

BEM is a patented technology developed in Australia that is now commercially available. It is currently being utilized in the mineral exploration industry in the search for massive sulphide ore deposits [18]. It uses the equivalent of a continuous range of electromagnetic frequencies to measure the wall thickness of a pipe by sensing the attenuation and phase delay of the signal passed through the pipe wall. Unlike the conventional eddy current technique, which uses a single frequency for testing, the BEM technique transmits a signal that covers a broad frequency spectrum ranging from 50 Hz to 50 kHz [19].

BEM data recorded can reveal the location of perturbations and can only be used on ferrous materials to measure wall thickness, identify and locate metal loss that produces wall thinning or graphitization, and also locate cracks. BEM does not require contact with bare metal, and in water pipes can read pipe condition through a cement lining [5, 18].

(4) Ground Penetrating Radar (GPR)

The first use of GPR for buried objects detection appeared in a German patent by *Leimbach* and *Löwy* in 1911 [20]. In the area of utility service deterioration monitoring, GPR has been used effectively to detect abnormally wet areas within the ground, such as from leaking water pipes, as well as leaking oil from high voltage cables [8]. In GPR surveys, high frequency (typically 1-1000MHz range) electromagnetic waves are transmitted into the ground from an antenna. These pulses propagate through the ground and reflect off sub-surface boundaries and the reflections are detected by a receiving antenna [21]. In general, any object whose electromagnetic properties are different to those of the surrounding soil will reflect a signal. In this way, tunnels, voids, metals and other buried objects can be located.

GPR can be used in a variety of media, including water, soil, rock and pavements. The most significant limitation of GPR's performance is that the pulses lose strength very quickly in conductive, lossy materials such as clay and saturated soils, therefore limiting the depth of penetration [21]. Therefore, research into GPR technologies has been focused on overcoming its drawbacks. *Hata et al*. [22] through their work on antenna design have produced a deep ground penetrating radar capable of surveying at depths up to 5 m in favourable conditions. *Ciochetto* and *Polidoro* [23] have developed an array of antenna system which is an improvement to a more traditional single-antenna instrument. Such arrangement allows for a 3D survey of the area under investigation. Some other research has focused on the interpretation of GPR images, notably the presentation of the output in three-dimensional images. *Conway* and *Bernstein et al.* [24, 25] described a new ground penetrating imaging radar system that creates sharp, three-dimensional images of underground pipelines and other buried objects.

2.1.2.3 Acoustic and Vibration Techniques

(1) Sonar

Sonar is an acoustic detection technology designed to operate under water. In the pipe inspection field, it has been adapted to provide information about elements in the pipe that are submerged below the water line. In sonar surveys, the time of the sound from the point of excitation, through transmission and reflection to the point that it is finally received is measured; the distance from the source to the target can be determined by the speed of the sound in the travelling medium. Such information is used to construct a sonar image from which the condition of the pipe interior can be assessed [8].

The sonar profiling system can be used with different frequencies to achieve different goals [26]. High frequency sonar can provide a higher resolution scan but a high frequency pulse attenuates quickly and therefore has a relatively low penetration capability. In contrast, low frequency sonar has a high penetration capability but it is limited in terms of its scanning resolution. Consequently, high frequency sonar can be suitable for clear water conditions; turbid water with high concentrations of suspended solids may require a lower frequency signal. Small defects are more likely to be observed by a high frequency signal [5].

(2) Impact Echo

Impact echo is a method for non-destructive evaluation typically applied to concrete, masonry materials, stone, plastic and some ceramics. It is based on the use of impact-generated compression waves that travel through the structure and are reflected by internal flaws and external surfaces [5]. The impact echo equation is :

$$T = \frac{V}{2F_p} \tag{2.1}$$

where T is the thickness, V is the wave speed and F_p is the peak frequency.

Impact echo can be used to determine the location and extent of flaws such as cracks and voids, it can also be used to measure the thickness of slabs, plates and hollow cylinders. The testing is conducted by hitting the test surface at a given location with a small instrumented impulse hammer or impactor and recording the reflected wave with a displacement or accelerometer sensor adjacent to the impact location [17]. This method is not limited by pipe size and can be applied both internally and externally only if the testing is executable.

(3) Ultrasound

Ultrasonic guided waves have been used extensively for pipe corrosion assessment [27]. The method employs mechanical stress waves that propagate along an elongated structure while guided by its boundaries. This allows the waves to travel a long distance with little loss in energy. The guided wave modes are generally categorized into 3 groups: torsional, longitudinal and flexural modes. The acoustic properties of these wave modes are a function of the pipe geometry, the material and the frequency.

In Guided Wave Testing of pipelines, an array of low frequency transducers is attached around the circumference of the pipe to generate an axially symmetric wave that propagate along the pipe in both the forward and backward directions of the transducer array. Depending on the type of guided wave, the number of transducers can range between two and four [5]. At location where there is a change of cross-section or a change in local stiffness of the pipe, an echo is generated. Based on the arrival time of the echoes, and the predicted speed of the wave mode at a particular frequency, the distance of a feature in relation to the position of the transducer array can be accurately calculated [28]. The technique is not suitable for pipes in softened materials as the acoustic waves are likely to attenuate significantly and it requires the internal pipe wall to be clean.

A summary of different condition assessment techniques with their applications and limitations of buried infrastructures is presented in Table 2.1.

Table 2.1

Different Condition Assessment Techniques: Applications and Limitations

Technology	**Applicationsand Limitations**
CCTV	• Real time assessment necessary • Subjective to the inspector • Images can only be obtained above the water line
SSET	• Post processing of images possible • Higher efficiency than CCTV • Requires manual interpretation of results
Sonar and Laser system	• Determines internal profile of the pipe along its length • Can measure pipe wall deflection, corrosion loss . • Can be operated in air or water, but not both simultaneously
MFL	• Good for cast iron and steel pipes

	• Access to pipe required • Can detect small defects but difficult for short and shallow defects • Often limited to cleaned and unlined pipes
Eddy Current testing	• Used in smaller diameter cast iron and steel pipes • Access to pipe required • Dimension of skin depth is a problem, RFEC is an improvement
Wave analysis (Ultrasound)	• Can determine location and site of defect • Pipe cleaning prior to inspection •Good detection rates for oil and gas pipelines detecting defects
Impact echo	• Overall condition of the pipe can be assessed • Access to pipe required
Ground Penetrating Radar (GPR)	• Technique successfully applied in pre-stressed concrete pipes • Can determine ground conditions external to the pipe • Can be used at ground surface and in-pipe mode • Requires skilled operator

2.2 Data Analysis and Condition Classification Methods

There is a variety of data analysis techniques to determine the condition assessment data. How to choose the suitable techniques depends on the objectives of the research. Pattern recognition is a machine learning and classification process which can be adopted in many areas. The use of pattern recognition techniques is a new approach for applications where adaptive signal processing methods are conventionally used. The functionality of an automated pattern recognition system can be divided into two basic tasks: the *Description* task generates attributes of an object using *feature extraction* techniques; and the *Classification* task assigns a group label to the object based on those features and a *classifier.*

2.2.1 Signal Processing and Feature Extraction

Feature extraction is the most significant phase of the classification process. In feature extraction, certain transforms or techniques are used to select and generate the features that represent the characteristic of the source signal. Feature arrays of vectors can be generated in time, frequency and time-frequency domain.

The computation of feature vector in time domain is usually simple. One of the methods is based on the energy distribution of the signal where the energy of a short time window of the source signal is used to discriminate between classes [29].Another method named Time Encoded Signal Processing and Recognition (TESPAR) is commonly used in speech waveform encoding to generate features from vehicle acoustic and seismic signals [30]. TESPAR is based on the duration and shape of the portion of the waveform that is between two zero crossings.

Spectral characteristics of acoustic signatures vary significantly among target classes. Feature generation methods based on frequency domain such as Fast Fourier Transform (FFT) and Power Spectral Density (PSD) are commonly used in applications such as speech and vehicle detection and classification [31]. Harmonics can also be used to extract feature vector. Harmonics are the peaks present in spectral domain representation of a signal. Relation between amplitude and phase of these harmonics is used to form the feature vector. These feature vectors are known as Harmonic Line Association (HLA) feature vector [32].

Time-frequency based techniques have been shown to outperform the techniques based on either time- or frequency-only domains. Features extracted in time-frequency domain are the most complete characterization for non-stationary signals as they display the energy distribution of a signal in both time and frequency domains [33]. Mel-frequency cepstral coefficients (MFCC) is the most popular spectral based parameter used in speech recognition due to its advantage of less complexity in implementation of feature extraction algorithm [34]. Short Time Fourier Transform (STFT) is an extension of Fourier Transform allowing for the analysis of non-stationary signals with a fixed resolution. Wavelet Transform (WT) provides multi-resolution time-frequency analysis. A set of wavelet based features can be obtained by calculating the inherent energies of the wavelet packet coefficients of the signal, each of which is related to a certain frequency band [35]. Other analysis including: Wigner-Ville Distribution (WVD), Multidimensional scaling (MDS) and learning vector quantization (LVQ) etc. are feature extraction techniques aim to represent the signal in different mapping [36]. Table 2.2 summarizes the feature extraction methods including the methods discussed above.

Table 2.2

Feature Extraction Methods

Extractor	Property and Comments
Filter-Bank Based	• Parameters sensitive • Criterion adopted for parameters selection • Local spectral energy estimation
Principle Component Analysis(PCA)	• Linear map • Eigenvector based • Good for Gaussian data • Supervised linear map
Linear Discriminative Analysis(LDA)	• Eigenvector based • Better than PCA for classification
Multidimensional scaling (MDS)	• Nonlinear map • Iterative • Sample size limited • Mainly used for 2-dimensional visualization
MFCC	• Nonlinear cepstral analysis • Features are good for automatic speech recognition
Wavelet Transform	• Linear map • Iterative • Good feature localisation • Efficiency depends on the basis selected
Self-Organizing Map (SOM)	• Nonlinear generalization of PCA •Suitable for extracting spaces of lower dimensionality • Iterative

2.2.2 Classification Techniques

Classifiers provide the functions or the rules that divide the feature space into regions, where each region corresponds to a certain class. This process is called classification. Classifiers can be categorized to parametric or non-parametric.

2.2.2.1 Parametric Classifiers

Below are several popular parametric classifiers which have been used in classification based on acoustic characteristics.

(1) Bayesian Classifier

Bayesian classifier is a probabilistic classifier based on Bayes' theorem. The optimal Bayes decision rule assigns a pattern to the class with the maximum posterior probability [37]. Maximum likelihood (ML) is used to estimate the Bayesian classifier parameters: $p(C_i)$ probabilities representing the frequency of class C_i in sample x and $p(x|C_i)$, class probability of x belongs to C_i. Each class is assumed to be normally distributed. Bayesian classifier requires a large number of training set for minimizing the bias [37].

(2) Support Vector Machines (SVMs)

SVMs is a state-of-the-art learning algorithm which was first introduced by Vapnik [38] in 1992. It was initially designed as a binary classifier. SVMs belong to the general category of kernel methods. It has two advantages. Firstly, it has the ability to generate non-linear decision boundaries using methods designed for linear classifiers. Secondly, it makes use of kernel functions which enables the user to apply a classifier to data that has no obvious fixed dimensions in terms of the feature space representation [39]. The effectiveness of SVMs is highly dependent on the selection of the kernel decision function, kernel's parameters, and soft margin parameter. There is no optimal solution to parameters selection for SVMs, prior understanding of the system and repeatedly trials are always required.

(3) Hidden Markov Model (HMM)

Hidden Markov Model (HMM) is an ubiquitous tool for modelling time series data. It is used in almost all the current speech recognition systems and numerous applications of other artificial intelligence and patter recognition [40].

A hidden Markov Model is a tool for representing probability distributions over sequence of observations. More precisely, the HMM is a probabilistic pattern matching technique in which the observations are considered to be the output of stochastic process and consists of an underlying (hidden) Markov chain. It has two components: a finite state Markov chain and a finite set of output probability distribution [41]. There are three basic problems of interest must be solved for the HMM model to be used real world applications [42]:

- Evaluation: with what probability does a given model generate a given sequence of observations? The forward algorithm solves this problem efficiently.

- Decoding : what sequence of underlying (hidden) states most probably generated a given sequence of observations. The Viterbi algorithm solves this problem efficiently.

- Learning : what model most probably underlies a given sample of observation sequences - that is, what are the parameters of such a model. This problem may be solved by using the forward-backward algorithm.

The technique was originally applied to the speech recognition field by *Baker* [43]. Now HMMs are applied in many fields where the goal is to recover a data sequence that is not immediately observable, but other data that depends on the sequence is.

(4) Gaussian Mixture Model (GMM)

Mixture Models are a type of density model which comprise a number of component functions, usually Gaussian. Mixture models are a semi-parametric alternative to non-parametric histograms (which can also be used as densities) and provide greater flexibility and precision in modelling the underlying statistics of sample data [44]. The GMM method is based on a finite mixture probability distribution model. And the method was successfully applied on robust speaker recognition system [45]. GMM provides a robust speaker representation for the difficult task of speaker identification using corrupted, unconstrained speech as reported by *Reynolds* and *Smith* [46]. The models are computationally inexpensive and easily implemented on a real-time platform.

2.2.2.2 Non-Parametric Classifiers

(1) KNN Classifier

KNN is a simple and accurate method for classifying objects based on the majority of the closest training examples in the feature space. The K-nearest neighbour algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbours, with the object being assigned to the class most common amongst its K nearest neighbours. The best choice of k depends upon the data. Generally, a larger value of K reduces the effect of noise on the classification but makes boundaries between classes less distinct. The accuracy of the KNN algorithm can be severely degraded by the presence of noisy or irrelevant features. Therefore, much research effort has been put into selecting or scaling features to improve classification [47]. KNN is implemented in many literatures as a benchmark to evaluate other classifiers [40, 47, 48].

(2) Artificial Neural Network (ANN)

Artificial Neural Network is a bio-inspired network made from neurons and can solve the problems that are hard to be modelled analytically. An important and very useful property of neural network is the ability to learn from examples in a supervisory manner. In most cases an ANN is an adaptive system changing its structure during a learning phase. ANN is used for modelling complex relationships between inputs and outputs or to find patterns in data.

The choice of the network type depends on the problem to be solved. A most commonly used family of neural networks for pattern classification tasks is the feed-forward network, which includes multilayer perceptron and Radial-Basis Function (RBF) networks [49]. These networks are organized into layers and have unidirectional connections between the layers. Another popular network is the Self-Organizing Map (SOM), or Kohonen-Network [50], which is mainly used for data clustering and feature mapping. The utility of artificial neural network models lies in the fact that they can be used to infer a function from observations. This is particularly useful in applications where the complexity of the data or task makes the design of such a function by hand impractical. One drawback of using artificial neural networks is that they require a large diversity of training for real-world operation [51].

(3) Decision Tree

Decision tree is a nonlinear classifier that depends on a multistage decision system, where the classes are sequentially rejected until reach the accepted class. This kind of classifier split the feature space into unique regions, where each region represents a class [52]. The most important feature of Decision tree classifiers is their capability to break down a complex decision-making process into a collection of simpler decisions, thus providing a solution which is often easier to interpret [40].

Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal.Another use of decision trees is as a descriptive means for calculating conditional probabilities. Decision trees are simple to understand and interpret; they require little data preparation and large datasets can be analyzed using standard computing resources in reasonable time [53]. There are also limitations for users if create over-complex trees, then over-fitting could occur and calculations can be overwhelming.

Table 2.3 summarized a few most commonly used classifiers and their properties.

Table 2.3

Classification Methods and Comments

Classifier	Property and Comments
Template Matching	• Assign patterns to the most similar template • Scale (metric) dependent
K-Nearest Neighbours Rule	• Assign patterns to the majority class among K nearest neighbours • Scale dependent
Bayes Rule	• Assign patterns to the class which has the maximum estimated posterior probabilities • Yields simple classifiers for Gaussian distributions • Sensitive to density estimation errors
Decision Tree	• Finds a set of thresholds for a pattern-dependent sequence of features • Iterative training process and needs pruning • Over training sensitive • Nonlinear classification
Support Vector Machines	• Maximizes the margin between the classes by selecting a minimum number of support vectors • Scale dependent • Nonlinear classification function • good generalization performance
Artificial Neural Network	• Iterative optimization of layers of units • Sensitive to training parameters • Nonlinear classification function • Needs regularization

2.3 Learning from Imbalanced Datasets

Most standard classification algorithms were designed based on one assumption that is the datasets are balanced and equally distributed among classes. However, in many real-world domains, class distribution is complex and imbalanced. In a given classification task, the size of datasets has an important role in building a good classifier. Learning from imbalanced datasets will cause machine learning algorithms fail to properly represent the distributive characteristics of the data and perform poorly on the classes contain fewer samples.

One of the common approaches to class imbalance problem is sampling, either randomly or intelligently, for obtaining an altered class distribution. The two basic sampling techniques are random minority oversampling and random majority undersampling. Random sampling is easy to perform but the drawbacks are obvious. In the case of undersampling, removing samples from the majority class may cause the classifier to miss important concepts pertaining to the majority class. With regard to oversampling, since oversampling simply replicates data to the original data set, multiple instances of certain examples become “tied,” leading to overfitting [54]. Numerous intelligent sampling techniques have been developed to improve the performance of random samplings*.* *Kubat* and *Matwi*n [55] proposed a technique called one-sided selection (OSS). One-sided selection attempts to intelligently undersample the majority class by removing majority class examples that are considered either redundant or ‘noisy.’ *Chawla et al*. [56] proposed an intelligent oversampling method called Synthetic Minority Oversampling Technique (SMOTE). SMOTE adds new, artificial minority examples by extrapolating between pre-existing minority instances rather than simply duplicating original examples. *Han et al*. presented a modification of *Chawla et al*.’s SMOTE technique which they call borderline-SMOTE (BSM) [57]. BSM selects minority examples which are considered to be on the border of the minority decision region in the feature-space and only performs

SMOTE to oversample those instances, rather than oversampling them all or a random subset. Cluster-based oversampling (CBOS) proposed by *Jo* and *Japkowicz* [58] attempts to even out the between-class imbalance as well as the within-class imbalance. In this technique, clustering is employed to select the representative training samples to improve the predictive accuracy for the minority class. *Yen* and *Lee* [59]reported that this approach empirically outperforms other undersampling methods. *Yoon* and *Kwek* also proposed to use clustering to reduce the imbalanced ratio, called Class Purity Maximization (CPM) [60].

Besides sampling methods, many other approaches have also been pursued in the imbalanced learning field. Kernel-based learning methods provide state-of-the-art techniques for many of today's data engineering applications. The principles of kernel-based learning are cantered on the theories of statistical learning and Vapnik-Chervonenkis (VC) dimensions [61]. *Zhu* and *Hovy* [62] analyzed the effect of undersampling and oversampling techniques with active learning for imbalanced learning problem. Traditionally, active learning methods are used to solve problems related to unlabeled training data. Similarly to re-sampling, active learning techniques create balanced training datasets at the early stage of the learning process. This technique focus on the query instances nearthe classification boundary rather than selecting randomly by instance. Active learning does not create extra data as in oversampling [63].

Another alternative solution for the imbalanced learning problem is ensemble-learning, in which multiple classifiers are trained from the original data and their predictions are combined to classify new instances [63].

Boosting and Bagging are two widely known ensemble based approaches. Boosting algorithms have been adapted to address the problem with small classes and forced the users to focus more on the difficult samples. At each boosting iteration, the distribution of training data is altered by updating the weight associated with each sample [64]. Most current bagging methods use a similar learning procedure: re-sampling subsets from a given training set, build multiple base classifiers on those subsets, and combining their predictions to make final prediction [65]. Several algorithms based on a variety of sampling strategies are proposed, such as: Roughly balanced (RB) bagging by *Hido* and *Kashima* [66]; underbagging by *Liu et al*. [67]; Overbagging and SMOTEbaggning by *Wang* and *Yao* [68]. Bagging maintains the class distribution of the training set, however, it relies on a simple strategy that is limited for dealing with imbalanced problem, except from changing the bag size and sampling step [63].

2.4 Summary

This chapter firstly reviewed current inspection techniques and technologies towards condition assessment of underground infrastructures. The description of the performance of each technology is provided in Section 2.1. The collection and analysis of relevant data and information is the next paramount step to detect and monitor buried assets. Pattern recognition analysis has experienced a rapid growth in the community. Pattern recognition is the study of how machines can observe the environment, learn to distinguish patterns of interest from their background, and make reasonable decisions about the categories of the patterns. A wide variety of feature extraction and classification methods are given in Section 2.2. The imbalanced learning problem is concerned with the performance of standard classifiers and it has attracted drawn significant attention over years, Section 2.3 provides a brief survey of the state-of-the-art solutions to the problem.

References

[1] T. Hao, et al. "Condition assessment of the buried utility service infrastructure". *Tunnelling and Underground Space Technology* , vol 28, pp.331-344. 2012.

[2] O. Hunaidi and W. T. Chu, "Acoustical Characteristics of Leak Signals in Plastic Water Distribution Pipes". *Applied Acoustics* , vol 58, pp.235-254. 1999.

[3] J. M. Muggleton and M. J. Brennan , "The design and instrumentation of an experimental rig to investigate acoustic methods for the detection and location of underground piping systems". *Applied Acoustics* , vol 69, pp.1101-1107. 2008.

[4] N. Metje, et al. "Mapping the underworld – state-of-the-art review". *Tunnelling and Underground Space Technology* , vol 22, pp.568-586. 2007.

[5] USEPA, *"Condition assessment technologies for water transmission and distribution systems".* 2012.

[6] Zheng Liu and Yehuda Kleiner, "State of the art review of inspection technologies for condition assessment of water pipes". *Measurement* , vol 46 issue 1, pp 1-15. 2013.

[7] Z. Liu, et al. "Exploring the relationship between soil properties and deterioration of metallic pipes using predictive data mining methods". *Journal of Computing in Civil Engineering* , vol 3, pp. 289–301. 2010.

[8] S.B. Costello, et al. "Underground asset location and condition assessment technologies". *Tunnelling and Underground Space Technology* , vol 22, 524–542. 2007.

[9] N. Sarshar, M.R.Halfawy and J. Hengmeechai, "Video Processing Techniques for Assisted CCTV Inspection and Condition Rating of Sewers". *NRCC-50451* . 2009.

[10] F. Cherqui, et al. "CCTV inspection of sewer segments: calibration of performance indicators based on experts' opinions". *11th International Conference on Urban Drainage.* Edinburgh,UK. 2008.

[11] M.D.Yang, et al. "Systematic image quality assessment for sewer inspection". *Expert Systems with Applications* , vol 38, pp.1766-1776. 2011.

[12] M.J. Chae, et al. "Neuro-fuzzy approaches for sanitary sewer pipeline condition assessment". *Journal of Computing in Civil Engineering* , vol 15, pp.4-14. 2001.

[13] D.H.Koo and S.T. Ariaratnam , "Innovative method for assessment of underground sewer pipe condition". *Automation in Construction* , vol 15, 479-488. 2006.

[14] S.Mukhopadhyay and G.P.Srivastava "Characterisation of metal loss defects from magnetic flux leakage signals with discrete wavelet transform". *NDT & E International* , vol 33, pp.57-65. 2000.

[15] D. Atherton, "Remote field eddy current inspection". *IEEE Transactions on Magnetics"* , vol 31, 4142–4147. 1995.

[16] S. Smith, et al. "Using 'intelligent pigs' for successfully assessing the condition of your pipes". *Proceedings of Third Annual Pipeline Conference.* London. 2001.

[17] J.Makar and N. Chagnon, "Inspecting systems for leaks, pits, and corrosion". *Journal American Water Works Association* , vol 91, 36. 1999.

[18] I. Vickridge and Tony Lau, *"Lessons Learnt from Pipeline Condition Assessment In Hong Kong".* Retrieved Aug. 2013, from http://www.zeinnews.com/lessons-learnt-from-pipeline-condition-assessment-in-hong-kong. 2011.

[19] USEPA, *"Innovation and Research for Water Infrastructure for the 21st Century".*Research Plan, Water Supply and water Resources Division. 2007.

[20] D.J.Daniels, *"Ground Penetrating Radar".* London: The Institution of Engineering and Technology. 2004.

[21] G.R.Olhoeft, "Maximising the information return from ground penetrating radar". *Journal of Applied Geophysics* , vol 43, 175-187. 2000.

[22] N.Hata, et al. "Deep ground-penetrating radar technology for surveying buried objects". *Proceedings of the 15th International Conference No-Dig' 97*, pp. 811-819. Taipei, Taiwan. 1997

[23] G.Ciochetto and R. Polidoro, *"Site investigation and output of utilities map using GPR".* Torino: CSELT Technical Reports. 1998.

[24] B.V.Conway, et al. "Bodies of evidence: site surveying with ground radar". *IEE Colloquium (Digest)*, (pp. 115, 21-26). London. 1995

[25] R. Bernstein, et al. "Imaging radar maps underground objects". *Computer Applications in Power. IEEE* , vol 3, pp.20-24. 2000.

[26] M.Eiswirth, et al. "Pipe defect characterisation by multi-sensor systems". *Proceedings of the 18th International Conference No-Dig 2000.* Perth, Western Australia. 2000.

[27] A.Demma, et al. "The reflection of guided waves from notches in pipes: a guide for interpreting corrosion measurements". *Ndt & E International* , vol 37, pp.167-180. 2004.

[28] G. Sposito, et al. "Potential drop mapping for the monitoring of corrosion or erosion". *Ndt & E International* , vol 43, pp.394-402. 2010.

[29] L.Deng, et al. "Analysis and comparison of two speech feature extraction/compensation algorithms". *IEEE Signal processing letters* , vol 12, No.6. 2005.

[30] S.A.Abdusslam. (2011). "Time Encoded Signal Processing and Recognition of Incipient Bearing Faults". *Proceedings of the 17th International Conference on Automation & Computing.* Huddersfield. 2011.

[31] A.A.Sahel., et al. "Speech Recognition from PSD using Neural Network". *International MultiConference of Engineers and Computer Scientist*, (p. vol 1). Hong Kong. 2009.

[32] Mark C. Weliman, et al. "Feature extraction and fusion of acoustic and seismic sensors for target identification". *Peace and Wartime Applications and Technical Issues for Unattended Ground Sensors,* , (p. vol 3081). Orlando.1997.

[33] Qifeng Zhu and Abeer Alwan, "Non-linear feature extraction for robust speech recognition in stationary and non-stationary noise". *Computer Speech and Language* , vol 17, pp.381–402. 2009.

[34] C. Ittichaichareon, et al. "Speech Recognition using MFCC". *International Conference on Computer Graphics, Simulation and Modeling*, (pp. pp. 135-138). Pattaya. 2012.

[35] Ya Wu and R. Du, "Feature extraction and assessment using wavelet packets for monitoring of machine processes" . *Mechanical Systems and Signal Processing* , vol 10, pp. 29-53. 1996.

[36] V. K. Kakar and M. Kandpal, "Techniques of Acoustic Feature Extraction for Detection and Classification of Ground Vehicles". *Emerging Technology and Advanced Engineering* , Vol 3, Issue 22. 2013.

[37] Irina Rish, "An empirical study of the naive Bayes classifier". *Workshop on Empirical Methods in Artificial Intelligence.* 2001.

[38] V.N. Vapnik, et al. "A training algorithm for optimal margin classiesr". *5th Annual ACM Workshop on COLT*, (pp. pp.144-152). Pittsburgh. 1992.

[39] N. Cristianini, N. and J.S. Taylor, *"An Introduction to Support Vector Machines and Other Kernel-based Learning Methods".* Cambridge University Press.2000.

[40] Richard O. Duda, Peter E. Hart and David G. Stork, *"Pattern Classification".* Canada: John & Sons Inc. 2001.

[41] Lawrence R.Rabiner, "A tutorail on HMM and selected applications in speech recognition". *Proceedings of the IEEE* , vol 77, No.2. 1989.

[42] Mark Gales and Steve Yang, "The Application of Hidden Markov Models in Speech Recognition". *Foundations and Trends in Signal Processing* , Vol. 1, No. 3, pp. 195–304. 2007.

[43] J.K.Baker, "The dragon system-an overview". *IEEE Trans. Acoust. Speech Signal Processing* , vol ASSP-23,No.1, pp.24-29. 1975.

[44] G.J.McLachlan and K.E. Basford ,*"Mixture Models: Inference and applications to clustering ".* Marcel Dekker. New York. 1988.

[45] Matthew N. Stuttle, *"A Gaussian Mixture Model Spectral Representation for Speech Recognition".* PhD Thesis. University of Cambridge.2003.

[46] D. A. Reynolds, et al. "Speaker Verification Using Adapted Gaussian Mixture Models". *Digital Signal Processing* , vol 10, pp.19-41. 2000.

[47] Christopher M. Bishop, *"Pattern recognition and machine learning".* Springer. New York .2006.

[48] Hao Zhang, et al. "SVM-KNN: Discriminative Nearest Neighbor Classication for Visual Category Recognition". *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2126-2136. 2006.

[49] A.K. Jain, et al. "Artificial Neural Network: A tutorial". *Computer* , Vol 29 , Issue 3, pp. 31-44. 2002.

[50] T. Kohonen, "Self-Organizing Maps". *Springer Series in Information Sciences*, vol 30. Berlin. 1995.

[51] Anil K. Jain, "Statistical Pattern Recognition: A Review". *IEEE Transaction on pattern analysis and machine intelligence* , vol 22, No.1. 2000.

[52] P.A. Chou, "Optimal Partitioning for Classification and Regression Trees". *IEEE Trans. Pattern Analysis and Machine Intelligence* , vol 13,no. 4, pp. 340-354. 1991.

[53] J.R. Quinlan, *"Programs for Machine Learning".* San Mateo: Morgan Kaufmann. 1993.

[54] Haibo He and Edwardo A. Garcia, "Learning from Imbalanced Data". *IEEE Transactions on Knowledge and data Engineering* , vol 21, No. 9. 2009.

[55] M.Kubat and S. Matwin, "Addressing the Curse of Imbalanced Training Sets: One-Sided Selection". *14th International Conference on Machine Learning*, pp. 179-186. 1997.

[56] N.V. Chawla, et al. "SMOTE: Synthetic Minority Over-Sampling Technique". *Journal of Artificial Intelligence Research* , vol. 16, pp. 321-357. 2002.

[57] H. Han, et al. “Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning". *Intelligent Computing*, pp. 878-887. Berlin. 2005.

[58] T. Jo and N. Japkowicz, “Class Imbalances versus Small Disjuncts". *ACM SIGKDD Explorations Newsletter* , vol 6, No.1, pp.40-49. 2004.

[59] S.J. Yen and Y.S. Lee, " Cluster-based under-sampling approaches for imbalanced data distributions". *Expert systems with applications* , vol 36, pp. 5718-5727. 2009.

[60] K. Yoon, et al. "A data reduction approach for resolving the imbalanced data issue in functional genomics". *Neural computing and applications* , vol 16, pp. 295-306. 2007.

[61] V. N. Vapnik, *"The Nature of Statistical Learning Theory".* New York: Springer. 1996.

[62] J. Zhu and E. Hovy, "“Active Learning for Word Sense Disambiguation with Methods for Addressing the Class Imbalance Problem". *Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 783-790. 2007.

[63] G.H. Nguyen, et al. "learning pattern classification tasks with imbalanced data sets". In *Pattern Recognition.* InTech. 2009.

[64] Y. Sun, M.S. Kamel, and Y. Wang, "Boosting for Learning Multiple Classes with Imbalanced Class Distribution". *Data mining* , pp. 592-602. 2006.

[65] L. Breiman, "Bagging Predictors". *Machine learning* , vol 24, pp. 123-140. 1996.

[66] S. Hido, and H.Kashima, "Roughly balanced bagging for imbalanced data". *SIAM International conference on data mining*, (pp. 143-152). Atlanta. 2008.

[67] A. Liu, et al. "Generative oversampling for mining imbalanced datasets". *2007 Inernational conference on data mining.* Las vegas. USA. 2007.

[68] S. Wang and S. Yao, "Diversity analysis on imbalanced datasets by using ensemble models". *IEEE Symposium on computational intelligence and data mining* , pp. 321-331. 2009.

Chapter 3

Pattern Recognition System and Methods:

The Theory

3.1 Introduction

Automatic (machine) recognition, description, classification and grouping of patterns are important problems in a variety of engineering and scientific disciplines such as biology, statistics, psychology, engineering, computer vision and artificial intelligence. The definition of a pattern is "*as opposite of a chaos; it is an entity, vaguely defined, that could be given a name*" [1]. A pattern could be the ridges of a fingerprint, a handwritten cursive word, a human face, or a speech signal. Given a pattern, its recognition/classification system may contain two tasks: (1) supervised classification in which the input pattern is identified as a member of predefined class; (2) unsupervised classification in which the pattern is assigned to an unknown class. The recognition problem is then being posted as a classification or a categorization task, where the classes are defined either by the system designer (in supervised classification) or are learned based on the similarity of patterns (in unsupervised classification) [2].

Pattern Recognition as a field of study developed significantly in 1960s and to some extent with the growth of research on knowledge-based systems in 1970s and neural networks in 1980s [2]. A large numbers of applications, ranging from the classical ones such as medical diagnosis and automatic character recognition to the more recent ones in data mining such as credit card transaction analysis and biometrics, have attracted considerable research effort with many methods developed and advances made. A common characteristic of a number of these applications is that the available features are not usually suggested by the domain but must be extracted and optimized by data analysis procedures.

The design of a pattern recognition system essentially involves the following four aspects: (1) data acquisition and pre-processing, (2) feature extraction and data representation, (3) feature grouping and pattern classification, and (4) decision making. The outcome of the system can be affected by the choice of sensor(s), pre-processing techniques, feature extraction algorithm and the decision making model. In many of the emerging applications, there is no single approach for classification that is “optimal” and that multiple methods and algorithms have to be used.

A brief description and comparison of several widely used feature extraction methods and pattern classification algorithms are given in this chapter.

3.2 Feature Extraction and Methods

The essential problem of pattern recognition is to identify an object as belonging to a particular group or class. Assuming the objects share common attributes with a particular group more than the others, the task of assigning the object to a group can be accomplished by determining the attributes of the object and identifying the group of which those attributes are most representative.

Given the goal of classifying objects based on their representative attributes, the functionality of an automated pattern recognition system can be divided into two basic tasks: the *Description* task generates attributes of an object using *feature extraction* techniques; and the *Classification* task assigns a group label to the object based on those features and a *classifier*.

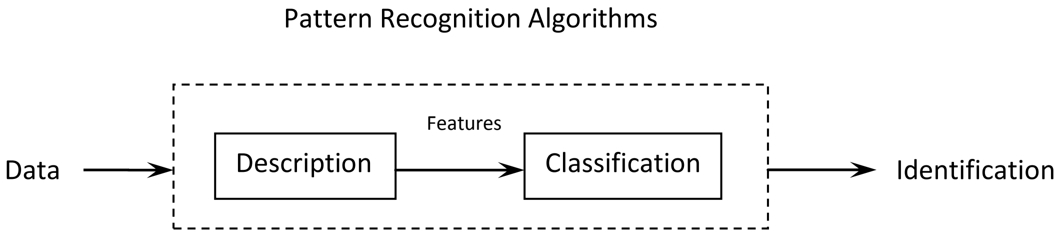


Figure 3.1 A Pattern Recognition System

Various methods have been developed for feature extraction in different fields such as face detection, character recognition, speech recognition and medical image processing. Two particular methods which can be applied to acoustic signals are discussed and compared in this section.

3.2.1 Wavelet Transform (WT)

3.2.1.1 Time-Frequency Signal Analysis

A time-frequency analysis can identify the signal frequency components, reveal the time variant features and is an efficient tool to extract representative information contained in signals. Various time–frequency analysis methods have been proposed and applied to condition classification.The earliest time-frequency method is known as the Spectrogram via Short Time Fourier Transform (STFT).One of the downfalls of the STFT is that it has a fixed resolution, the width of the windowing function relates to how the signal is represented. A wide window gives better frequency resolution but poor time resolution, a narrower window gives good time resolution but poor frequency resolution. This is one of the reasons for the creation of the Wavelet Transform and multi-resolution analysis, which can give good time resolution for high frequency components, and good frequency resolution for low frequency components, which is the type of analysis best suited for many real signals.

Wigner-VilleDistribution (WVD) is apopular nonlinear alternative to the STFT technique which can achieve higher resolution than STFT and give exactly the instantaneous frequency.The Wigner transformation gives good results (high time-frequency resolution) when theexamined signals consist of a small number of higher harmonics. In other cases, thetransformation results include interferences, the so called cross-terms.Currently research is conductedconcerning methods of cross-terms reduction.

In the Wavelet Transform (WT), the mother wavelet can be stretched according to frequency toprovide reasonable window, a long time window is used in low frequency and a short timewindow is used in high frequency.Though the resolution of WT is lower than WVD, the cross-terms don't appear as the WT is linear time-frequency analysis.The result of Discrete Wavelet Transform (DWT) of a continuous signal is a series of wavelet coefficients which represent the degree of correlation between the analyzed signal and the wavelet function at different instances of time; therefore, DWT coefficients contain temporal information of the analysed signal and that is essential for providing useful features for recognition system.

3.2.1.2 Discrete Wavelet Transform (DWT)

The Wavelet Transform (WT) and more particularly Discrete Wavelet Transform (DWT) is one of the computationally efficient techniques for extracting information of non-stationary signals. The DWT was developed as an alternative to the Short Time Fourier Transform (STFT) to improve on the frequency and time resolution of the FT. Instead of providing a fixed time resolution for all the frequencies in the signal spectrum, the DWT is a multi-resolution analysis (MRA) which is able to analyze signals at different frequencies with different resolutions.

The DWT is defined by the following equation:

$$DWT\{X_{j,k}(t)\} = \sum_j \sum_k x(t) \frac{1}{\sqrt{2^j}} \psi(\frac{t-k}{2^j}) \quad (3.1)$$

where $\psi(.)$ is the wavelet function, k is the translation factor and j is the scale parameter, $2^{-j/2}$ is the normalization factor [3].

In order to take advantage of the Wavelet Transform, an efficient computation algorithm and an implementation scheme are needed. *Mallat* [3] solved these problems by introducing the Multi-resolution Analysis (MRA) which is linked to the Perfect Reconstruction filter-bank structures [3, 4].

A signal's approximation at resolution 2^{-j} is defined as an orthogonal projection on a space $V_j \subset L^2(R)$ .The space V_j groups all possible approximations at the resolution 2^{-j}. The orthogonal projection of x on V_j is the function x_j that minimizes distance $\|x - x_j\|$. The detail of a signal at resolution 2^{-j} is the difference between approximations at the resolution 2^{j-1} and 2^j. For a given multi-resolution approximation $\{V_j\}$, there exists a unique function called a scaling function $\phi(t)$. The scaling function plays a role of an averaging function of a low-pass filter in the multi-resolutionanalysis. A problem was raised along with the decomposition process that is "how to cover the signal spectrum all the way down to zero with wavelet spectra" and the scaling function is the solution to it. By introducing the scaling function we have circumvented the problem of the infinite number of wavelets and set a lower bound for the wavelets as the lower limit can never reach zero. The iteration of filtering will stop at the point where the number of samples has become smaller than the length of the scaling filter.

The scaling function $\phi(t)$ is shifted by discrete translations and is dilated by dyadic scale factor $\left\{\phi_{j,k}(t) = 2^{-j/2}\phi(\frac{t-k}{2^j})\right\}$, where $2^{-j/2}$ is a normalization constant. At each scale 2^j the shifted scaling functions constitute a basis that spans a subspace V_j. The orthogonal projection on V_j can be computed by decomposing the signal $x(t)$ in the scaling orthogonal basis. The inner products:

$$a_j[k] = \langle x, \phi_{j,k} \rangle \quad (3.2)$$

represent the discrete wavelet approximation coefficients of the original signal $x(t)$ at scale 2^j. It can also be written as:

$$a_j[k] = \sum_t x(t)\frac{1}{\sqrt{2^j}}\phi(\frac{t-k}{2^j}) = x * \phi_j(2^j k) \quad (3.3)$$

where $\phi_j(t) = 2^{-j/2}\phi(-\frac{t}{2^j})$, symbol $*$ stands for discrete convolution. Let pV_j be the orthogonal projection on the vector space V_j. The approximation signal of $x(t)$ at scale 2^j is equal to:

$$pV_j x(t) = \sum_k \left\langle x, \phi_{j,k} \right\rangle \phi_{j,k} \quad (3.4)$$

The difference between the approximations of a signal $x(t)$ at scales 2^{j-1} and 2^j is called the detail of the signal at scale 2^j. Here the space W_j is orthogonal to V_j, and $V_j \oplus W_j = V_{j+1}$, where $\oplus$ stands for direct sum of two vector spaces. *Mallat* has proven that there exists a function $\psi(t)$, called orthogonal wavelet:

$$\psi_{j,k}(t) = 2^{-\frac{j}{2}}\psi(\frac{t-2^j k}{2^j}) \quad (3.5)$$

The inner products:

$$d_j[k] = \left\langle x, \psi_{j,k} \right\rangle \quad (3.6)$$

represent the detail wavelet coefficients of $x(t)$ at scale 2^j. Let pW_j be the orthogonal projection on the vector space W_j. The detail signal can be implemented as a high-pass filtering of $x(t)$ sampled at rate 2^j , and it equals to:

$$pW_j x(t) = \sum_k \left\langle x, \psi_{j,k} \right\rangle \psi_{j,k} \tag{3.7}$$

A signal $x(t)$ can be fully characterized by its wavelet decomposition and can be written as a sum of its approximation at level L and its details on all levels:

$$x(t) = pV_L x(t) + \sum_{j=1}^{L} pW_j x(t) \tag{3.8}$$

The DWT decomposition can be implemented using a fast pyramidal algorithm related to multirate filterbanks first proposed by Stephane Mallat and it is also called ‘Mallat Algorithm’ [3, 5]. Figure 3.2 shows example of a 3 level wavelet decomposition tree.

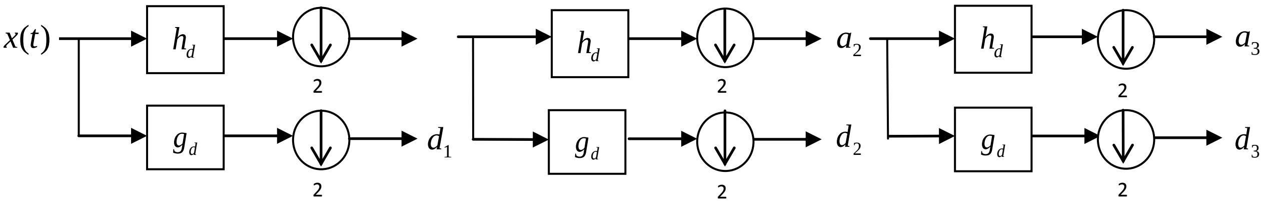


Figure 3.2 Three Level DWT Decomposition Tree

Down-sampling by 2 ($2\downarrow$) following each filter halves the resolution (doubles the scale) is used to avoid the information redundancy. Half-band filters h_d and g_d remove half of the frequencies but leave the scale unchanged, which makes half the number of samples redundant, the down-sampling operation can therefore discard half the samples without any loss of information.

In the algorithm the signal $x(t)$is analysed at different frequency bands with different resolution by decomposing the signal into a coarse approximation and detail information, the coarse approximation is then further decomposed repeating the same wavelet decomposition steps. This is achieved by successive high-pass and low-pass filtering of the original time-domain signal and is defined by the following so-called two-scale equations:

$$d_{j+1}[t] = \sum_k g[k-2t]\left\langle x, \phi_{j,k} \right\rangle = \sum_k g[k-2t]a_j[k]$$

$$= a_j[t] * g_d[2t] \tag{3.9}$$

$$a_{j+1}[t] = \sum_k h[k-2t]\left\langle x, \phi_{j,k} \right\rangle = \sum_k h[k-2t]a_j[k]$$

$$= a_j[t] * h_d[2t] \tag{3.10}$$

The approximation and detail coefficients from one scale can be computed from the approximation coefficients from the previous scale by convolution with the low-pass filter h_d and high-pass filter g_d, respectively, followed by a down sampling with a factor of 2.

3.2.1.3 Wavelet Packet Decomposition (WPD)

Wavelet Packet Decomposition (WPD) is extended from the Wavelet Decomposition (WD). In the DWT, each level is calculated by passing the previous approximation coefficients through high and low pass filters. However, in the WPD, both the detail and approximation coefficients are filtered. For n levels of decomposition the WPD produces 2^n different sets of coefficients (or nodes) as opposed to $(n+1)$ sets for the DWT. However, due to the down sampling process the overall number of coefficients is still the same and there is no redundancy.

The wavelet packet method is a generalization of wavelet decomposition that offers a richer range of possibilities for signal analysis and it allows the best matched analysis to a signal. It provides level by level transformation of a signal from the time domain into the wavelet domain with variable frequency resolution.

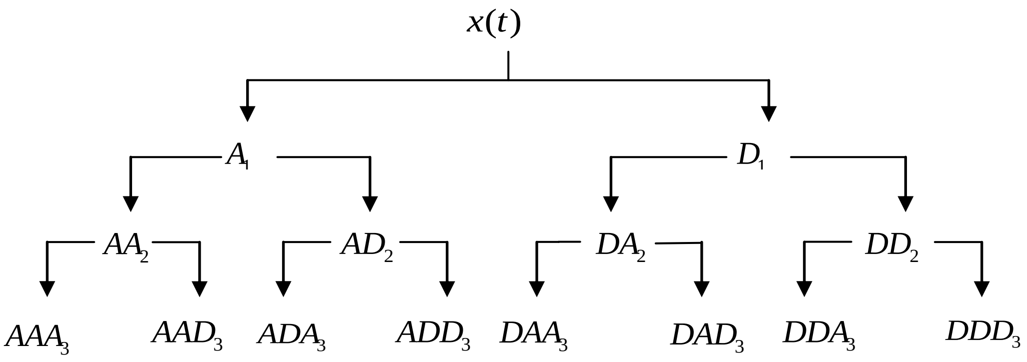


Figure 3.3 Level 3 Decomposition using complete wavelet packet transforms

The top level of the WPD tree is the time representation of the signal. As each level of the tree there is an increase in the trade-off between the time and frequency resolution. The bottom level of a fully decomposed tree is the wavelet representation of the signal. Figure 3.3 shows level 3 decomposition using wavelet packet transform. To define wavelet packets, the coefficients resulting from the decomposition of the signal $x(t)$ are:

$$C_{2m}(2^{j-1}t-k)=\frac{1}{\sqrt{2}}\sum_{l}^{+\infty}h_{l-2k}C_m(2^j t-l) \qquad (3.11)$$

$$C_{2m+1}(2^{j-1}t-k)=\frac{1}{\sqrt{2}}\sum_{l}^{+\infty}g_{l-2k}C_m(2^jt-l) \qquad (3.12)$$

where h_{l-2k} and g_{l-2k} are the previously defined low-pass and high-pass filters, m is the level number of the decomposition, $C_0(t)=\phi(t)$, $C_1(t)=\psi(t)$.

3.2.1.4 Wavelet Family

Several wavelet families are available for signal characterization and the selection of appropriate wavelet is very important for the correct analysis of signals. Depending on the type of signal to be analyzed, the mother wavelet is normally chosen according to the convenience, experience and published studies. For some applications,comparison tools such as cross-correlation can be used to help choose an optimal wavelet function by computing cross correlation coefficient between analysed signal and selected wavelet filter, the wavelet which maximizes the correlation coefficient is considered optimum.

The most popular and commonly used wavelets for signal processing are Daubechies (db), Symlets (Sym) and Coiflets (Coif). Figure 3.4 gives some examples of each of these wavelet families.

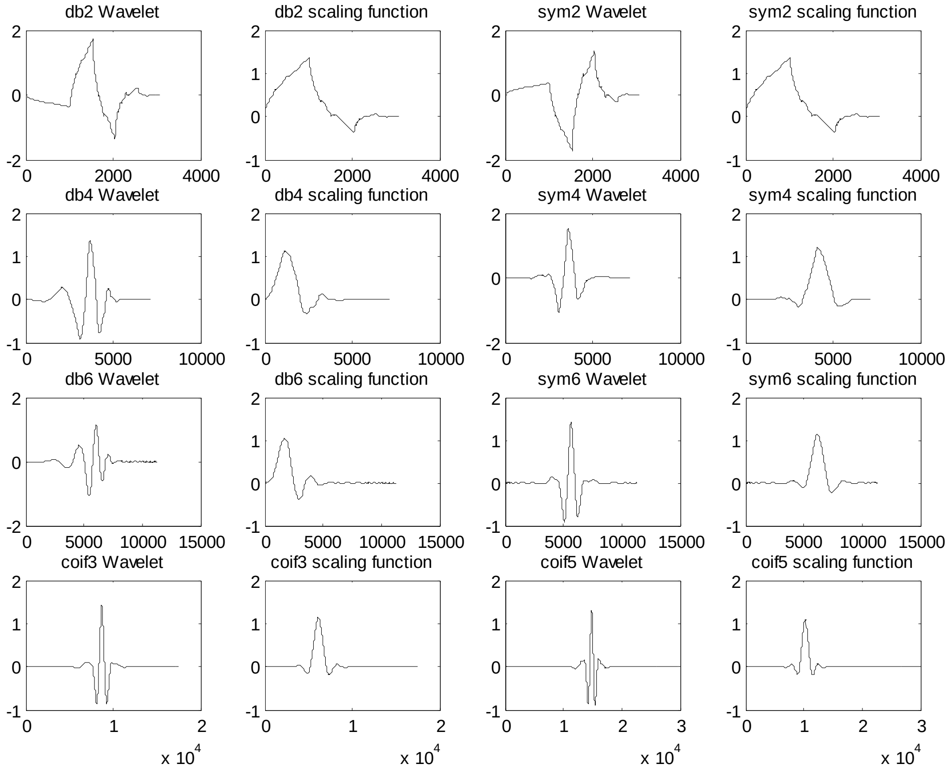


Figure 3.4 Wavelet Families

Table 3.1 General characteristics of popular wavelet families

Family	Daubechies	Symlet	Coiflet	Meyer
Short Name	Db	Sym	Coif	Meyr
Order	N strictly positive integer	N=2, 3…	N=1,2…5	-
Orthogonal	yes	yes	yes	yes
Biorthogonal	yes	yes	yes	yes
Discrete transform	possible	possible	possible	possible
Continuous transform	possible	possible	possible	possible
Fast algorithm	possible	possible	possible	no
FIR filters	possible	possible	possible	possible
Support Width	2N-1	2N-1	6N-1	infinite
Filter Length	2N	2N	6N	[-8 8]
Symmetry	Far from	Near to	Near to	yes

3.2.2 Data Fitting Methods

Data fitting (or parameter estimation) is an important technique used for modelling in many areas of disciplines. It is the process to construct a curve with mathematical function that has the 'best' fit to a series of data points. Fitted curve can be used to summarize the relationship among two or more variables and reveal the hidden patterns which reflect the observed data.

3.2.2.1 Least Squares Polynomial Approximation

The problem can be described as follows: let (x_i, y_i) be the observed quantities. Assume the function f in (3.13) is a $(n-1)-th$ degree polynomial:

$$y \cong f\{x; a_1, a_2, ...a_n\} = a_1 x^{n-1} + a_2 x^{n-2} + ...a_{n-1} x + a_n \tag{3.13}$$

Then the data fitting problem is to solve the system:

$$\begin{bmatrix} x_1^{n-1} & x_1^{n-2} & \cdots & x_1 & 1 \\ x_2^{n-1} & x_2^{n-2} & \cdots & x_2 & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ x_m^{n-1} & x_m^{n-2} & & x_m & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix} \tag{3.14}$$

for the coefficients $\left(a_1, a_2, \ldots a_n\right)$. Usually the system is over-determined, so the least square problem is considered:

$$\sigma = \sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{3.15}$$

where x_i and y_i are known quantities. The error σ gives a measure of how well the function f fits values y. For a straight line fitting, let $f(x) = ax + b$, the goal is to find values of a and b that minimize the error σ. The coefficientswhich produces the smallest value of σ describes the best fit of the observed data and represent the variables' relationship.

To find the minimum of function (3.15), we take the derivative of the error σ with respect to a and b , then set each to zero:

$$\frac{\partial \sigma}{\partial a} = -2\sum_{i=1}^{n} x_i (y_i - ax_i - b) = 0 \tag{3.16}$$

$$\frac{\partial \sigma}{\partial b} = -2\sum_{i=1}^{n} (y_i - ax_i - b) = 0 \tag{3.17}$$

On solving (3.16) and (3.17):

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \tag{3.18}$$

$$a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i \text{ (3.19)}$$

They can be re-written as:

$$\begin{pmatrix} \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i & n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_i x_i y_i \\ \sum_i y_i \end{pmatrix} \tag{3.20}$$

$$(AX = B)$$

The coefficients a and b will be obtained by solving $X = A^{-1}B$. Putting the value of a and b in $f(x)$, we will get the best fit of the observed quantities.

Consider the general form for a polynomial of order j : $f(x) = a_0 + \sum_{k=1}^{j} a_k x^k$. The general expression of error σ using the least squares approach is:

$$\sigma = \sum_{i=1}^{n} (y_i - (a_0 + \sum_{k=1}^{j} a_k x^k))^2 \qquad (3.21)$$

To minimize equation (3.21), we can take the derivative with respect to each of the coefficients in (3.21) $a_0, a_1, \ldots a_k$, $k = 1, \ldots j$ and set these derivatives to zero:

$$\frac{\partial \sigma}{\partial a_0} = -2\sum_{i=1}^{n} (y_i - (a_0 + \sum_{k=1}^{j} a_k x^k)) = 0 \qquad (3.22)$$

$$\frac{\partial \sigma}{\partial a_1} = -2\sum_{i=1}^{n} (y_i - (a_0 + \sum_{k=1}^{j} a_k x^k))x = 0$$

$$\frac{\partial \sigma}{\partial a_2} = -2\sum_{i=1}^{n} (y_i - (a_0 + \sum_{k=1}^{j} a_k x^k))x^2 = 0$$

$$\vdots$$

$$\frac{\partial \sigma}{\partial a_j} - 2\sum_{i=1}^{n} (y_i - (a_0 + \sum_{k=1}^{j} a_k x^k))x^j = 0$$

Then re-write these equations and put into matrix form:

$$\begin{bmatrix} n & \sum_i x_i & \sum_i x_i^2 & \ldots & \sum_i x_i^j \\ \sum_i x_i & \sum_i x_i^2 & \sum_i x_i^3 & \ldots & \sum_i x_i^{j+1} \\ \sum_i x_i^2 & \sum_i x_i^3 & \sum_i x_i^4 & \ldots & \sum_i x_i^{j+2} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_i x_i^j & \sum_i x_i^{j+1} & \sum_i x_i^{j+2} & \ldots & \sum_i x_i^{j+j} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_j \end{bmatrix} = \begin{bmatrix} \sum_i y_i \\ \sum_i x_i y_i \\ \sum_i x_i^2 y_i \\ \vdots \\ \sum_i x_i^j y_i \end{bmatrix} \qquad (3.23)$$

$$(AX = B)$$

The coefficients $a_0, \ldots a_j$ can be obtained by solving $X = A^{-1}B$

3.2.2.2 Padé Approximation

The Padé approximation was firstly made in a systematic study by the French mathematician *Henri Padé* in his thesis, which was a rational approximation to functions given by their power series [6]. He proved the results on their general structure and set out the connection between Padé approximants and continued fractions.

The Padé approximation to $f(x)$ is the quotient of two polynomials $P_M(x)$ and $Q_N(x)$ of degrees M and N, respectively. Notation $R_N^M(x)$ is used to denote this quotient, normalized by $Q_N(0)=1$:

$$R_N^M(x) = \frac{P_M(x)}{Q_N(x)} = \frac{\sum_{i=0}^{M} a_i x^i}{1+\sum_{j=1}^{N} b_j x^j} \qquad (3.24)$$

One of the main applications of Padé approximations is to extract as much information as possible from a power series expansion that is known only to a few terms. Conversion from a Taylor expansion (when $N=0$) to Padé form usually accelerates convergence, and in many cases is often a better approximations of $f(x)$ than a number of terms from its Taylor expansion, for a Padé approximation of $f(x)$ of degree M, it is expected to give results at least as good as its polynomial approximation of degree M.

A Padé approximation to a given data series $y_k = f(x_k)$ can be obtained as follows: finding p and q to satisfy $R_N^M(x) = \frac{P_M(x)}{Q_N(x)}$, the rational function $R_N^M(x)$ is the Padé approximant to the series $f(x)$ if:

$$f(x) - R_N^M(x) = O(x^{m+n+1}) \qquad (3.25)$$

when $x \to 0$ and $Q_N(x) \neq 0$, $O(.)$ is Landau's big-Oh symbol, meaning that the right side is a power series over x^i, beginning with degree , $i = M + N + 1$, up to $i = +\infty$. It can also be written as:

$$P^M(x) = f(x)Q^N(x) + O(x^{m+n+1}) \qquad (3.26)$$

Suppose a and b are $m-th$ and $n-th$ vectors of coefficients of polynomials $p \in P^M$ and $q \in Q^N$, respectively. As a consequence:

$$y_k = f(x_k) \cong \frac{P^M(x_k)}{Q^N(x_k)} = \frac{\sum_{i=0}^{M} a_i (x_k)^i}{\sum_{j=0}^{N} b_j (x_k)^j} \qquad (3.27)$$

if $m \geq n$, equation (3.27) can be written as (3.28) and its matrix form as equation (3.29) with the expansion of $\{x_k\}$:

$$a_0 + (a_1 + b_1 y)x_k + (a_2 + b_2 y)x_k{}^2 + \cdots (a_n + b_n y)x_k{}^n + a_{n+1}x_k{}^{n+1} + \cdots a_m x_k{}^m = y_k \qquad (3.28)$$

$$\begin{bmatrix} 1 & x_1 & y_1x_1 & x_1^2 & y_1x_1^2 & \cdots & x_1^n & y_1x_1^n & x_1^{n+1} & \cdots & x_1^m \\ 1 & x_2 & y_2x_2 & x_2^2 & y_2x_2^2 & \cdots & x_2^n & y_2x_2^n & x_2^{n+1} & \cdots & x_2^m \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \\ 1 & x_k & y_kx_k & x_k^2 & y_kx_k^2 & \cdots & x_k^n & y_kx_k^n & x_k^{n+1} & \cdots & x_k^m \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ b_1 \\ \vdots \\ a_n \\ b_n \\ a_{n+1} \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix} \qquad (3.29)$$

If $m \leq n$, the essence of the matter remains the same although it is displayed in a different form:

$$a_0 + (a_1 + b_1 y)x_k + (a_2 + b_2 y)x_k{}^2 + \cdots (a_m + b_m y)x_k{}^m + (b_{m+1} y)x_k{}^{n+1} + \cdots (b_n y)x_k{}^n = y_k$$

(3.30)

$$\begin{bmatrix} 1 & x_1 & y_1x_1 & x_1^2 & y_1x_1^2 & \cdots & x_1^m & y_1x_1^m & y_1x_1^{m+1} & \cdots & y_1x_1^n \\ 1 & x_2 & y_2x_2 & x_2^2 & y_2x_2^2 & \cdots & x_2^m & y_2x_2^m & y_2x_2^{m+1} & \cdots & y_2x_2^n \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \\ 1 & x_k & y_kx_k & x_k^2 & y_kx_k^2 & \cdots & x_k^m & y_kx_k^m & y_kx_k^{m+1} & \cdots & y_kx_k^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ b_1 \\ \vdots \\ a_m \\ b_m \\ b_{m+1} \\ \vdots \\ b_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix} \quad (3.31)$$

Equations (3.29) and (3.31) can be re-arranged in following form:

$$\begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^m & y_1x_1 & y_1x_1^2 & \cdots & y_1x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^m & y_2x_2 & y_2x_2^2 & \cdots & y_2x_2^n \\ \vdots & \vdots & \vdots & & & & & & \vdots \\ \\ 1 & x_k & x_k^2 & \cdots & x_k^m & y_kx_k & y_kx_k^2 & \cdots & y_kx_k^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \\ b_1 \\ \vdots \\ b_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix} \quad (3.32)$$

The Padé coefficients $[a_0, a_1, \cdots a_m, b_1, \cdots b_n]$then will be obtained by solving a linear algebraic equation $\mathbf{AX} = \mathbf{B}$. That is: $X = A^{-1}B$.

Here, simply to compute the inverse of $\mathbf{A}$. For many situations where the inverse of $\mathbf{A}$ does not exist, the Singular Value Decomposition (SVD) is often used to approximate the inverse which turn a singular problem into a non-singular one. The vector $\mathbf{X}$ in equation (3.32) can be solved for using the transpose of $\mathbf{A}$, i.e.:

$$\mathbf{A^T A X = A^T B} \tag{3.33}$$

$$\mathbf{X = (A^T A)^{-1} A^T B} \tag{3.34}$$

This is the form of the solution in a least-squares sense from standard multivariate regression theory where the inverse of $\mathbf{A}$ is express as:

$$\mathbf{A}^{\dagger} = \mathbf{(A^T A)^{-1} A^T} \tag{3.35}$$

where $\mathbf{A}^{\dagger}$ is called the More-Penrose pseudoinverse, the use of SVD can aid in the computation of the generalized pseudoinverse.

3.3 Classification Methods

Pattern recognition can be seen as a classification process. The task of the classification process is to use the features provided from the feature extraction process (introduced in 3.2) to assign the object to a category. There are two main types of classification: *supervised* classification and *unsupervised* classification. In supervised classification, data samples have given labels and class types which are used as exemplars in the classifier design, sometimes the number of classes must be learned along with the structure of each class. Unsupervised classification refers to situations where the objective is to construct decision boundaries based on unlabeled training data. Unsupervised classification is also known as data clustering, the goal of clustering is to find groups in the data and the features that distinguish one group from another. Clustering techniques can also be used as part of a supervised classification by defining prototypes. Furthermore, in cluster analysis the number of categories or classes may not even be specified, the task is to discover a reasonable categorization of the data (if one exists).

Once a feature selection or classification procedure finds a proper representation, a classifier can be designed using a number of possible approaches. Three different approaches are defined to design a classifier. The simplest and the most intuitive approach to classifier design is based on the concept of similarity: patterns that are similar should be assigned to the same class. Therefore, once an appropriate metric has been established to define similarity, patterns can be classified by template matching or minimum distance classifier using a few prototypes per class. The most straightforward 1-nearest neighbour rule is always used as a benchmark for all the other classifiers based on similarity theory and it does not require any user defined parameters, its classification results are implemented independently [7].

The second main concept used for designing pattern classifiers is based on the probabilistic approach. The optimal Bayesian decision rule assigns a pattern to the class with the maximum posterior probability. In practice, the empirical Bayesian decision rule is used: the estimates of the densities are used in place of the true densities. These density estimates are either parametric or nonparametric. Commonly used parametric models are multivariate Gaussian distributions [8] for continuous features, binomial distributions for binary features, and multinormal distributions for integervalued (and categorical) features. The two well known nonparametric decision rules, the k-nearest neighbours (KNN) rule and the Parzen classifier (the class-conditional densities are replaced by their estimates using the Parzen window approach) [9], while similar in nature, give different results in practice. They both have essentially one free parameter each, the number of neighbour k, or the smoothing parameter of the Parzen kernel. Further, both these classifiers require the computation of the distances between a test pattern and all the patterns in the training set.

The third category of classifiers is to construct decision boundaries directly by optimizing certain error criterion. A classical example of this type of classifier is Fisher's linear discriminate [10] that minimizes the MSE (mean square error) between the classifier output and the desired labels. Decision tree is a special type in this category, which is trained by an iterative selection of individual features that are most salient at each node of the tree [11].One of the most interesting recent developments in classifier design is the introduction of the Support Vector Classifier [12].It is primarily a two-class classifier. The optimization criterion here is the width of the margin between the classes, i.e., the empty area around the decision boundary defined by the distance to the nearest training patterns. These patterns, called support vectors, finally define the classification function.

The classifier is first designed using training samples, and then it is evaluated based on its classification performance on the test samples. The percentage of misclassified test samples is taken as an estimate of the error rate. The error rate of a recognition system must be estimated from all the available samples which are split into training and test sets. There are no good guidelines available on how to divide the available samples into training and test sets. No matter how the data is split into training and test sets, it should be clear that different random splits (with the specified size of training and test sets) will result in different error estimates.

Several most commonly used classifiers are summarized in Table 3.2, many of them represent a family of classifiers.

Table 3.2 Classification methods

Method	**Property**	**Comments**
Template Matching	• Assign patterns to the most similar template, e.g.correlation	The template and metric must be supplied by the user.
1-nearest neighbour rule	• Assign patterns to the class of the nearest training pattern	No training needed; scale dependent
K-nearest neighbour rule	• Assign patterns to the major class among K nearest neighbour using an optimized value K.	Slow testing; scale dependent
Bayesian decision rule	• Assign patterns to the class which has the maximum estimated posterior probability	Sensitive to density estimation error; yields simple classifier function
Logistic classifier	• Maximum likelihood rule for logistic posterior probability	Linear classifier; suitable for mixed data types
Fish linear discriminate	• Linear classifier using MSE optimization	Simple and fast; similar to Bayesian.
Decision tree	• Finds a set of thresholds for a pattern-dependent feature sequence	Fast testing; iterative training; overtraining sensitive
Multilayer Neutral network	• Iterative MSE optimization of two or more layers of features	Nonlinear; slow training; overtraining sensitive
Support vector classifier	• Maximize the margin between classes by selecting a minimum number of support vectors	Scale dependent; nonlinear; good generalization performance

How to select machine learning algorithms for one's classification problem normally depends on the size and the structure of the feature sets and the advantages of some particular algorithms.In this research, we have a reasonable amount of labelled data of each pipe condition that we wish to study, but limited prior knowledge about the domain of the data. One practical choice is to start with simpler algorithm such as Bayes classifier or its extension K-Nearest Neighbours (KNN) which is easy to perform and can be very effective when the training datasets are well-distributed, another appealing character of KNN is that it is non-parametric and can be applied to multi-class applications like ours. It is always necessary to test out a couple different classifiers to compare and achieve higher accuracy. More sophisticated algorithms usually involve more parameters to yield specific decision regions, for example Neural Networks and Support Vector Machines (SVM). When the training datasets are not linearly separable in the feature space, SVM appears to have its advantage to adopt an appropriate kernel and adjust the weights to produce boundaries between classes. Based on above consideration, KNN and SVM were chosen for two types of pipe data classification in this research, but other classifiers may be tested together with different parameters within each algorithm in the future work.Detailed descriptions of these two classifiers are given in the following sections.

3.3.1 K-Nearest Neighbours Classifier (KNN)

The Nearest Neighbour rule is a simple non-parametric decision procedure to classify unknown object into the class of its nearest neighbour. The K-Nearest Neighbours rule is an extension to the Nearest Neighbour approach to use not just one but a set of *K* nearest neighbours in the training data. This rule classifies the sample by assigning it the label which is most frequently presented among the *K* nearest samples. Then a voting scheme (e.g. majority vote) is used to make the decision.

The basic idea behind many of the methods of estimation is a simple probability density function. Suppose n samples $\{x_1, \cdots x_n\}$ are independently and identically distributed according to the probability density function, the probability P of a vector x will fall in a region R is given by:

$$P = \int_R p(x)dx \tag{3.36}$$

Thus P is a smoothed or averaged version of the density function $p(x)$. The probability that K of these n samples fall in the region R is given by the Binominal distribution. If the random variable x follows Binomial distribution with parameters n (total) and K (subset), the probability of having K samples in n measurements are success is given by the probability mass function:

$$P_r(x = K) = \binom{n}{K} p^K (1-p)^{n-K} \tag{3.37}$$

$$\binom{n}{K} = \frac{n!}{K!(n-K)!} \tag{3.38}$$

Now assume that x is a point within R and V is the volume enclosed by R, take x as the centre of V and let it grow until it captures K samples, the estimate for $p(x)$ can be written as:

$$p(x) = \frac{K/n}{V} \tag{3.39}$$

Therefore the general expression of non-parametric density function is:

$$p(x) \cong \frac{K}{nV}\text{ , where}\begin{cases} V \text{ is the volume surrounding } x \\ n \text{ is the total number of examples} \\ K \text{is the number of examples inside } V \end{cases}$$

In applying this result to practical density estimation problems there are two basic approaches can be adopted: (1) fix the volume V and count the number *K* of data points inside V . This leads to the method commonly referred as Kernel Density Estimation(KDE); (2) fix the value of *K* and determine the minimum volume V that encloses *K* data points. This gives rise to the K-Nearest Neighbours (KNN) approach.

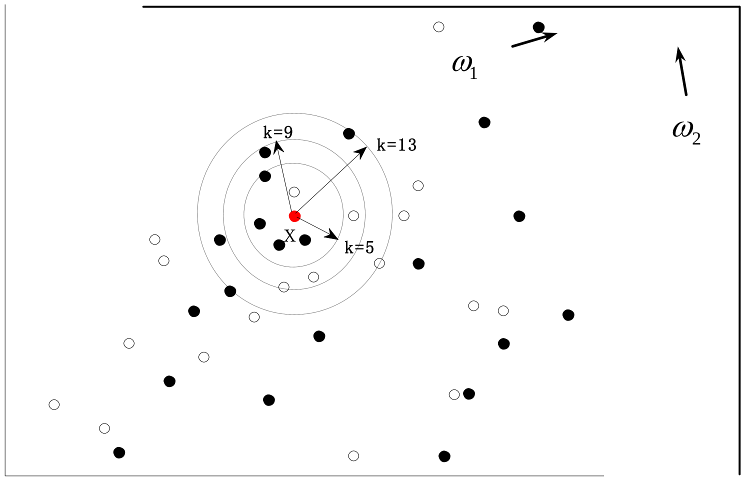


Figure 3.5 KNN classification when k=5, 9 and 13.

Figure 3.5 is an example of KNN classification. Dashed line circles are the corresponding volumes when *K* is chosen to be 5, 9 and 13. The test sample x should be classified to the class which contains more data points inside the volume than the other class among *K* samples. For example: if $K=5$, the test sample x should be assigned to class ω_2 because there are 4 examples belong to ω_2 and only 1 example belongs to ω_1 inside the volume.

The K-Nearest Neighbours classifier relies on a metric or distance function between patterns. Euclidean distance metric in d dimensions is commonly chosen. A metric $D(.,.)$ is merely a function that gives a generalized scalar distance between two argument patterns. Suppose there are n labelled training samples in d dimensions, equation (3.40) can be used to calculate the Euclidean distance to the test sample x to seek *K* closest training samples:

$$D_r(s,x) = \left(\sum_{k=1}^{r} (s_k - x_k)^2 \right)^{1/2} \tag{3.40}$$

where $r \le d$, r is some subset of the full d dimensions.

The main advantage of KNN method is that it leads to a simple approximation of the (optimal) Bayes classifier [13]. Suppose a database with n examples, a hyper-sphere of volume V around test sample x and captures a total of *K* samples among n , in which k_i examples from class ω_i , ω_i are categories that cover all n samples.The likelihood that variable x falls in class ω_i is given by:

$$\mathcal{L}\left(x \mid \omega_1, \omega_2 \ldots \omega_n\right) = P_r\left(\omega_1, \omega_2 \ldots \omega_n \mid x\right) = \prod_{i=1}^{n} P_r\left(x_i \mid \omega\right) \tag{3.41}$$

We wish to estimate the probability that x belongs to one particular class ω_i , the likelihood function (3.41) then can be transformed by introducing equation (3.37) into:

$$\mathcal{L}(x \mid \omega_1, \omega_2 \ldots \omega_n) = \prod_{i=1}^{n} \binom{n}{k} P_r^{\,k} (1 - P_r)^{n-k} \tag{3.42}$$

In practice, it is often more convenient to work with the logarithm of the likelihood function, called the log-likelihood $\ln \mathcal{L}$. Differentiate $\ln \mathcal{L}$ with respect to P_r and set the value to zero to find the value of P_r that maximizes $\ln \mathcal{L}$:

$$\frac{\partial \ln \mathcal{L}(x \mid \omega_1, \omega_2 \ldots \omega_n)}{\partial P_r} = 0 \tag{3.43}$$

$$\hat{P}_r(x \mid \omega_i) = \frac{\sum_{i=1}^{n} k_i}{n_i N} \tag{3.44}$$

$\hat{P}_r$ is the maximum likelihood estimation that variable x belongs to one particular class ω_i , where $n_i \subset N$,are observations enclosed in chosen volume.

The priors can be estimated by:

$$p(\omega_i) = \frac{n_i}{N} \tag{3.45}$$

Therefore the Bayes classifier becomes:

$$P_r(\omega_i \mid x) = \frac{p_r(x \mid \omega_i) p(\omega_i)}{p(x)} = \frac{\frac{k_i}{n_i V} \frac{n_i}{N}}{\frac{K}{NV}} = \frac{k_i}{K} \tag{3.46}$$

Consequently, the category which most frequently presented among *K* examples would be selected for test sample x .

3.3.2 Support Vector Machines (SVMs)

Support Vector Machines (SVMs) is a relatively new supervised classification model introduced and developed by Russian mathematician Vladimir N. Vapnik [14] and his group in 1995. The SVMs method was initially invented to solve binary class problems but they are gaining popularity and development for multiclass classification due to its many attractive features and promising empirical performance.

SVMs belongs to the general classification category of kernel methods [15]. A Kernel method is an algorithm that depends on the data only through inner products, by replacing the inner products with kernels in possibly higher dimensional feature space, flexible representations of data could be obtained. Combining a simple linear discriminate algorithm with the kernels, nonlinear separations of data can be learned efficiently.

3.3.2.1 Linear Support Vector Classification

A linear classifier is defined by a linear discriminate function in the form:

$$f(\mathrm{x}) = \mathrm{w}^{\mathrm{T}}\mathrm{x}+\mathrm{b} \tag{3.47}$$

The vector w is known as weight vector and b is called bias, the inner product between w and x is defined as: $\mathrm{w}^{\mathbf{T}}\mathrm{x} = \sum_i \omega_i x_i$. The set of points x such that $\mathrm{w}^{\mathrm{T}}\mathrm{x}=-\mathrm{b}$ are all points perpendicular to w and go through the origin: a line in two dimensions; a plane in three dimensions and more generally, a hyperplane, which divides the space into two.

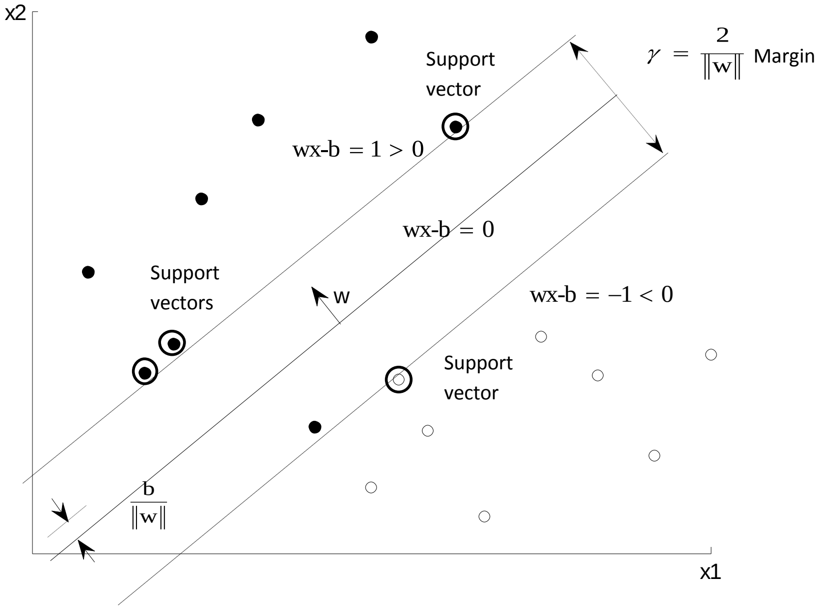


Figure 3.6 Linear binary SVMs classification

Figure 3.6 gives an example of the case $b = 0$. $\{x{:}f(x)=w^T x+b=0\}$ donates the hyperplane which defines the decision boundary between regions, the bias parameter b determines the location of the boundary away from the origin of the space. A classifier with a linear decision boundary is called linear classifier as shown in Figure 3.6. Conversely, when the decision boundary of a classier depends on the data in a non-linear way the classier is said to be non-linear.

Consider a set of training vectors belonging to two classes is of the form $\{x_i, y_i\}$, where $y_i \in \{-1,1\}$, $i = 1,2,\ldots n$. The distance of a point x from the hyperplane $\langle w,b \rangle$ is given by:

$$d(\mathrm{w,b;x}) = \frac{|\mathrm{w}^{\mathrm{T}}\mathrm{x}_i + \mathrm{b}|}{\|\mathrm{w}\|} \tag{3.48}$$

The optimal hyperplane is given by maximizing the margin γ, it is given by:

$$\gamma = \min_{x_i:y_i=-1} d(\mathrm{w,b;x}_i) + \min_{x_i:y_i=1} d(\mathrm{w,b;x}_i)$$

$$= \frac{1}{\|\mathrm{w}\|}\left(\min_{x_i:y_i=-1} |\mathrm{w}^{\mathrm{T}}\mathrm{x}_i + \mathrm{b}| + \min_{x_i:y_i=1} |\mathrm{w}^{\mathrm{T}}\mathrm{x}_i + \mathrm{b}|\right)$$

$$= \frac{2}{\|\mathrm{w}\|} \text{(3.49)}$$

It equivalents to finding $\min\|\mathrm{w}\|$, min$\|\mathrm{w}\|$ is also equivalent to min$\frac{1}{2}\|\mathrm{w}\|^2$ and the use of this term makes it possible to perform Quadratic Programming (QP) optimization [16]. The optimization problem therefore becomes:

$$\min\frac{1}{2}\|\mathrm{w}\|^2 \text{ such that: } y_i(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i + \mathrm{b}) - 1 \geq 0,\ i = 1,2\cdots n \text{(3.50)}$$

Where y_i is the label series of training samples x_i. To allow for errors equation (3.50) is modified with:

$$y_i(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i + \mathrm{b}) \geq 1 - \xi_i \tag{3.51}$$

where $\xi_i \geq 0, i = 1,2\cdots n$ are called slack variables for which the penalty term $\sum_{i=1}^{n}\xi_i$ allow a sample to be in the margin ($0 \leq \xi_i \leq 1$, also called a margin error) or to be misclassified ($\xi_i > 1$). These slack variables ξ_i are basically a measure of the misclassification error. The optimization problem then becomes:

$$\min_{\mathrm{w,b}} \frac{1}{2}\|\mathrm{w}\|^2 + C\sum_{i=1}^{n}\xi_i \qquad (3.52)$$

with the following constrains:

$$y_i(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i + \mathrm{b}) \geq 1 - \xi_i\,,\ \xi_i \geq 0 \qquad (3.53)$$

where the constant $C > 0$ is chosen to maximize the margin and minimize the amount of slack. Equation (3.52) is the formulation called 'soft-margin SVMs' and it was originally introduced by Cortes and Vapink [7, 14]. The solution to the optimization problem of equation (3.53) is given by introducing Lagrange multipliers α_i [17], where $\alpha_i \geq 0$:

$$\Phi(\mathrm{w,b},\alpha) = \frac{1}{2}\|\mathrm{w}\|^2 - \sum_{i=1}^{n}\alpha_i(y_i[\langle \mathrm{w}^{\mathrm{T}}\mathrm{x}_i \rangle + \mathrm{b}] - 1) \quad (3.54)$$

Classical Lagrangian duality enables the primal problem, Equation (3.53), to be transformed to its dual problem, which is easier to solve:

$$\max_{\alpha} W(\alpha) = \max_{\alpha}(\min_{\mathrm{w,b}} \Phi(\mathrm{w,b},\alpha)) \qquad (3.55)$$

The minimum with respect to w and b of the Lagrangian, Φ, is given by:

$$\frac{\partial\Phi}{\partial\mathrm{w}} = 0 \Rightarrow \mathrm{w} = \sum_{i=1}^{n}\alpha_i y_i x_i \qquad (3.56)$$

$$\frac{\partial\Phi}{\partial\mathrm{b}} = 0 \Rightarrow \sum_{i=1}^{n}\alpha_i y_i = 0 \qquad (3.57)$$

Substituting (3.54), (3.56) and (3.57) into (3.55) gives a new dual formulation:

$$\max_{\alpha}\left[\sum\nolimits_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}y_iy_j\alpha_i\alpha_j\mathrm{x}_i^{\mathrm{T}}\mathrm{x}_j\right] \qquad (3.58)$$

with the following constrains:

$$\sum\nolimits_{i=1}^{n}y_i\alpha_i=0,0<\alpha_i<C \qquad (3.59)$$

Solving Equation (3.58) with constraints Equation (3.59) determines the Lagrange multipliers, and the optimal separating hyperplane is given by:

$$\mathrm{w}^{*}=\sum_{i=1}^{n}\alpha_iy_ix_i$$

$$\mathrm{b}^{*}=-\frac{1}{2}\left\langle \mathrm{w}^{*},\mathrm{x}_r+\mathrm{x}_s\right\rangle \qquad (3.60)$$

where $\mathrm{x}_r,\mathrm{x}_s$ are any support vectors from each class satisfying: $\alpha_r,\alpha_s>0$, $y_r=-1,y_s=1$.

The classifier is then given by:

$$f(x)=\mathrm{sgn}(\left\langle \mathrm{w}^{*},\mathrm{x}\right\rangle+\mathrm{b}) \qquad (3.61)$$

The dual formulation of the SVM optimization problem depends on the data only through dot products. The dot product can therefore be replaced with a non-linear kernel function, thereby, performing a non-linear mapping into feature space and the constraints are unchanged:

$$f(x)=\mathrm{sgn}(\sum_{i\in SVs}\alpha_iy_iK(x,x_i)+\mathrm{b}) \qquad (3.62)$$

3.3.2.2 Multiple Class SVMs Classification

The conventional way to extend original binary SVMs classifier to multiple category classifiers is to decompose $M-$class problem into a series of two-class problems, for which One-Against-All is the earliest and one of the most widely used implementations [18]. Suppose there are N training samples $\{(x_1, y_1), \cdots (x_n, y_n)\}$, $x_i \in R^M$ and $y_i \in \{1, 2, \cdots M\}$ is the corresponding class labels. One-Against-All approach constructs M binary SVMs classifiers, each of which separates one class from all the others. The $i-th$ SVM is trained with all the training examples of the $i-th$ class with positive labels (one side of the boundary), and all the others with negative labels (the other side of the boundary). The decision function of $i-th$ SVM $f_i(\mathrm{x}) = \mathrm{w}_i{}^{\mathrm{T}}\phi(\mathrm{x}_i) + \mathrm{b}_i$ solves the following problem:

$$\min_{\mathrm{w},\xi} \frac{1}{2}\|\mathrm{w}\|^2 + C\sum_{i=1}^{n} \xi_j^i \tag{3.63}$$

provided that:

$$\tilde{y}_i(\mathrm{w}_i{}^{\mathrm{T}}\phi(\mathrm{x}_j) + \mathrm{b}_i) \geq 1 - \xi_j^i\,,\ \xi_j^i \geq 0 \tag{3.64}$$

where $\tilde{y}_j = 1$ if $y_j = i$ and $\tilde{y}_j = -1$ otherwise. For an unknown sample x to be classified in class i whose decision function produces the largest value:

$$i = \underset{i=1,2\cdots M}{\arg\max}\ f_i(\mathrm{x}) = \underset{i=1,2\cdots M}{\arg\max}(\mathrm{w}_i{}^{\mathrm{T}}\phi(\mathrm{x}_i) + \mathrm{b}_i) \tag{3.65}$$

Another popular method for multiple class problems is One-Against-One method which is also called pair-wise classification [19]. This method constructs $k(k-1)/2$ classifiers, k is the number of categories. For training samples from the $i-th$ and $j-th$ classes, solve the following binary classification problem:

$$\min_{\mathrm{w},\xi} \frac{1}{2}\left\|\mathrm{w}_{i,j}\right\|^2 + C\sum_{i=1}^{n}\xi_{i,j}\ ,\ \xi \geq 0 \tag{3.66}$$

provided that:

$$\mathrm{w}_{i,j}^{\mathrm{T}}\phi(\mathrm{x}_l) + \mathrm{b}_{i,j} \geq 1 - \xi_{i,j}\text{, if } y_l = i$$

$$\mathrm{w}_{i,j}^{\mathrm{T}}\phi(\mathrm{x}_l) + \mathrm{b}_{i,j} \leq -1 + \xi_{i,j}\text{, if } y_l = j \tag{3.67}$$

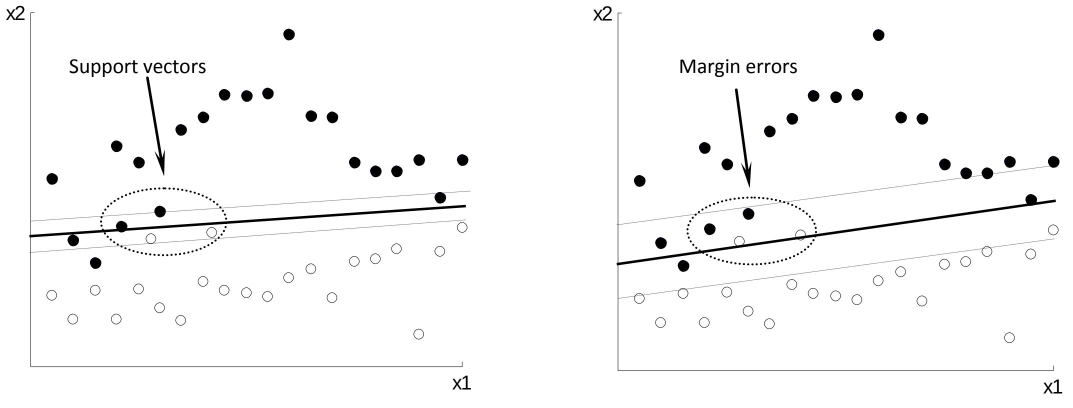
The decision strategy suggested by Friedman [20] is: if $sign\{\mathrm{w}_{i,j}^{\mathrm{T}}\phi(\mathrm{x}) + \mathrm{b}_{i,j}\}$ says x belongs to $i-th$ class then the vote for $i-th$ class is added by one, otherwise $j-th$ is increased by one. Then the sample x is predicted to be in the class with the higher vote. This method is usually slower than the One-Against-All method due to its squared number of classifiers and the complexity of its decision making. However, it is not reasonable to claim which method is always better than the other, the number of classes, the number of training samples and the application constraints all need consideration to choose the suitable strategy.

3.3.2.3 Kernel Models and SVMs Parameters

The effectiveness of SVMs highly depends on the selection of kernel decision function, the kernel's parameters, and the soft margin parameter.

Hyper-Parameters

Parameters α_i and b given in Equation (3.48) are used to train SVMs to find a large margin hyperplane, another parameter called the soft margin constant, C , plays an important role in deciding the boundary of the margin as illustrated in Equation (3.40). For a large value of C , a large penalty is assigned to margin errors as shown in the left of Figure 3.7, those points closest to the hyperplane affect its orientation, as a result the hyperplane comes close to several other data points. When C is decreased as shown in the right of Figure 3.7, those points become margin errors so that the hyperplane's orientation changed, providing a much larger margin for the rest of the data.



Soft margin parameter $C = 10$

Soft margin parameter $C = 2$

Figure 3.7 The effect of parameter C on the decision boundary

Kernel Parameters

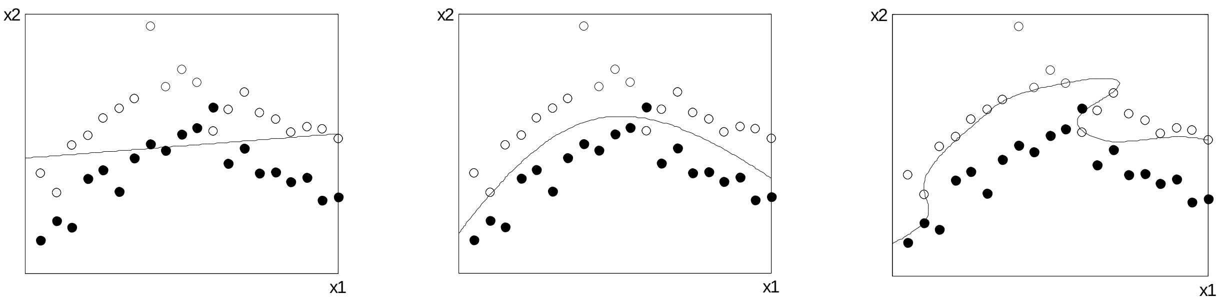
The idea of the kernel function is to enable operations to be performed in the input space rather than the potentially higher dimensional feature space. The computation is critically dependent upon the number of training patterns and to provide a good data separation for a higher dimensional problem requires a smart choice of kernel functions. Table 3.3 summarizes several popular kernel functions and their mathematical forms.

Table 3.3 Some popular Kernels for Classification

Name	**Function**	**Parameter**
Linear	$k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$	-
Polynomial	$k(x_i, x_j) = (\zeta\phi(x_i)^T \phi(x_j) + r)^d, \gamma > 0$	Degree d , ζ , r
Gaussian	$k(x_i, x_j) = \exp(-\frac{1}{2\sigma^2}\left\Vert\phi(x_i) - \phi(x_j)\right\Vert^2)$	Inverse- width σ
Radial Basis Function	$k(x_i, x_j) = \exp(-\gamma\left\Vert\phi(x_i) - \phi(x_j)\right\Vert^2)$	Variance γ
Sigmoid	$k(x_i, x_j) = \tanh(\zeta\phi(x_i)^T \phi(x_j) - r)$	Scaling factor ζ , shifting factor r

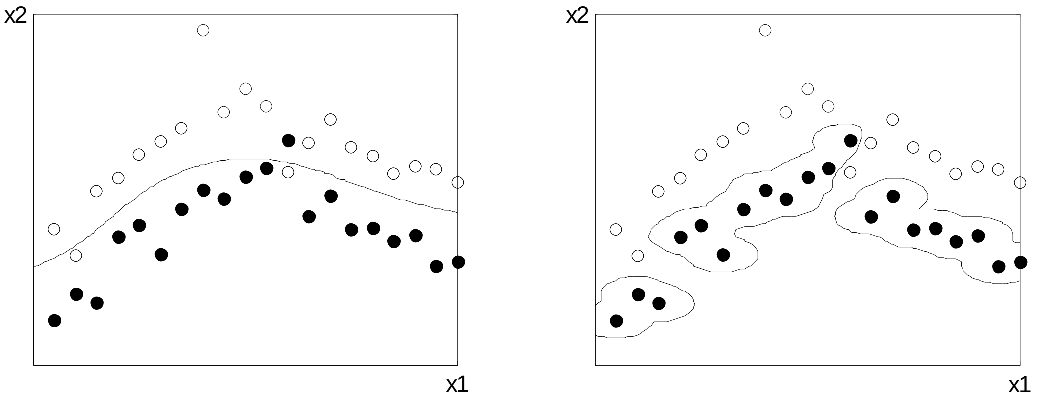
Kernel parameters have significant effects on the decision boundary. The degree d of polynomial kernel and the width parameter σ of Gaussian kernel control the flexibility of the corresponding classifiers. Higher order of polynomial classifier yields decision boundary with greater curvature as can be seen in Figure 3.8. Left figure shows classification using a linear classifier; a 2-nd order polynomial classifier (middle) is already flexible enough to discriminate two classes; the 7-th degree classifier (right) yields a similar boundary but clearly over fitted.

Large value of width parameter σ of Gaussian classifier can lead to overfitting while small value of σ results in a nearly linear decision boundary, Figure 3.9 gives examples of the effect of σ. When σ is small (left), the whole set of data affects the value of discriminant function of a given x, results in a smooth boundary. As σ is increased (right), the locality of the support vector expansion increases resulting in a greater curvature of the decision boundary. As seen from Figure 3.8 and 3.9, if the complexity parameters are too large, overfitting will occur.



$d = 1$, equals to linear $d = 2$, good fit $d = 7$, over-fitted

Figure 3.8 The effect of degree d of polynomial classifier on decision boundary



Gaussian $\sigma = 0.1$ good fit Gaussian $\sigma = 1$ over-fitted

Figure 3.9 The effect of width parameter σ of Gaussian classifier on decision boundary

The frequently asked question is: which kernel should be chosen? There is no simple answer for it. Like most practical problems in machine learning, the answer is data-dependent and several kernels should be tried. A general procedure should be followed: try a linear kernel first, then move to non-linear kernel to see if the classification performance can be improved. The linear kernel provides a useful baseline and it is easy to operate as the only affecting parameter is the soft margin parameter C. Once a result of using a non-linear kernel is available, adjust its corresponding kernel parameters to achieve a better classification result. Comparisons among different kernels and different value of kernel parameters by using independent test data over a reasonable range of problems should be carried out.

3.4 Summary

Pattern recognition has been used for many real world applications and provided satisfying results upon the understanding and procession of knowledge of the system and available algorithms. In this chapter, feature extraction and classification are given detailed description as they are the most important procedures in a pattern recognition system. Following methods are given a throughout presentation of their theoretical aspects: Wavelet Transform (WT) and two data fitting methods: Polynomial and Padé approximation are the main techniques used for feature extraction in the research; K-Nearest Neighbours (KNN) and Support Vector Machines (SVMs) are classification algorithms applied on obtained features. Classification results from Laboratory data and field data will be presented and discussed in the following chapters.

References

[1]Watanabe S. *"*Knowing and guessing : a quantitative study of inference and information*".* New York: John Wiley & Sons. 1969.

[2]Anil K. Jain, R. P. "Statistical Pattern Recognition: A Review". *IEEE Transactions on pattern analysis and machine intelligence* , vol. 22, No. 1,. 2000

[3]Mallat S. "A theory for multiresolution signal decomposition: the wavelet representation". *IEEE Transactions on Pattern Recognition and Machine Intelligence* , vol 11, pp 674–693. 1989

[4] Gilbert S., Truong N. "Wavelets and Filter Banks*".* Wellesley, MA: Wellesley-Cambridge Press.1996

[5] Malla S. "A Wavelet Tour of Signal Processing:The Sparse Way*".* Academic Press. 2008

[6] M.Vajta. "Some Remarks on Pade-approximations". *3rd TEMPUS-INTCOM Symposium.* Veszprém, Hungary.2000

[7]Cover TM, H. P. "Nearest neighbor pattern classification". *IEEE Transactions on Information Theory* , 13 (1): 21–27.1967

[8] Norbert, H. "Invariant tests for multivariate normality: a critical review". *Statistical Papers* , 43 (4): 467–506.2002

[9] Liang Lan,et.al., "An Active Learning Algorithm Based on Parzen Window Classification". *JMLR:Workshop and Conference Proceedings 16*, (pp. 99-112).2011

[10] McLachlan, G. J. *"*Discriminant Analysis and Statistical Pattern Recognition*".* New York, John Wiley & Sons.2004

[11] Safavian, S. R. "A Survey of Decision Tree Classifier Methodology". *IEEE Transactions on Systems, Man, and Cybernetics*, (pp. Vol. 21, No. 3, pp 660-674).1991

[12] Steve R.G. *"*Support Vector Machines for Classification and Regressio*n".* Technical report, University of Southampton.1998

[13] McCallum, Andrew, and Kamal Nigam, "A comparison of event models for Naive Bayes text classification.". *AAAI-98 workshop on learning for text categorization*, Vol. 752, 1998

[14] Vladimir N. V. "Support-vector networks". *Machine Learning* , Volume 20, Issue 3, pp 273-297.1995

[15] Kobe Crammer , "On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines". *Journal of Machine Learning Research* , 265-292.2001

[16] G. Dantzig, *"Linear Programming and Extensions".* Princeton,New Jersey: Princeton University Press. 1963.

[17] I.Vapnyarskii, "Lagrange multipliers",in M. Hazewinkel: *" Encyclopedia of Mathematics".* Springer.2002.

[18]Tristan Fletcher, *"*Support Vector Machines Explained*".* University College London. 2009

[19] Chih Wei Hsu "A Comparison of Methods for Multiclass Support Vector Machines". *IEEE Transactions on neural networks* , vol. 13, No. 2, 2002.

[20] J. Friedman, et al. *"The Elements of Statistical Learning".* New York: Springer. 2001.

Chapter 4

Laboratory Constructed Water-filled Siphon Condition Classification

4.1 Introduction

This chapter presents the results of the laboratory study of the proposed classification and recognition system for siphon conditions. The experiments were carried out using a 450mm diameter concrete siphon constructed in the Hydraulics Laboratory at the University. The temperature was recorded routinely for all measurements and the reproducibility tests were conducted to test the reliability of the system. A number of possible conditions were designed and studied to check the capabilities of the system.

The siphon was filled with water in all the experiments to simulate real live sewer conditions. Acoustic data were collected under three sets of conditions: (i) clean empty siphon; (ii) siphon with a controlled amount of blockage; (iii) siphon with various types of wall damage. The effect of surrounding medium was also studied including air, sand and water. The signal processing of the raw recorded data as the first step of classification was carried out. Acoustic signals were received on the receivers and transformed into acoustic pressure response and recorded with a PC. Discrete wavelet and digital filters were used to decompose and extract representative features from the pressure signals for the further classification analysis. Finally, the classification of siphon conditions was carried out using K-nearest neighbours (KNN). These results of the experiments were used to improve the data analysis methods and condition classification system. The detail of the theories of feature extraction method and classification algorithm used in this part of research can be found in Chapter 3.

4.2 Experimental Design

Within the underground sewerage system there are numerous special structures serving particular needs. Siphons are designed to convey water run-off and sewage flows below deep obstructions where such crossings cannot be attained by a sewer placed on a continuous gradient [1, 2]. Monitoring of the conditions of this infrastructure is important to maintain it timely to avoid failures and resultant flooding. A full-scale siphon has been designed and constructed in the Hydraulics Laboratory in the University of Bradford in order to study the effect of a range of typical conditions, and develop a classification method based on acoustic characteristics to detect sediment level in or damages on the wall of the siphon.

4.2.1 Measurement setup

A 4.2 m long and 2.0 m high siphon using sections of 450mm concrete pipes was installed on a 500mm layer of fine sand in an open top box made of 12mm plywood as shown in Figure 4.1 (left). The siphon was filled up with clean water to the level of 900mm below the top rims of the vertical parts (Figure 4.1 (right)), and the water level remained during all the experiments as reference.

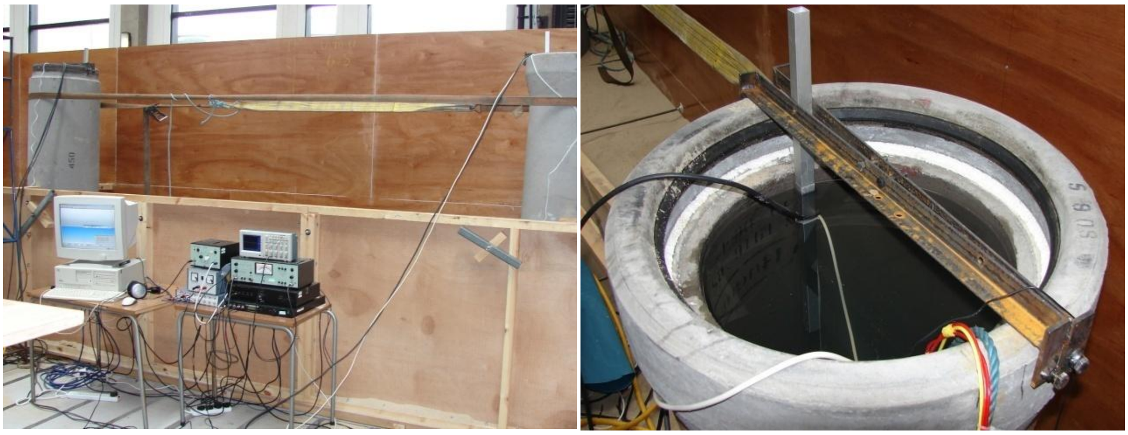


Figure 4.1 Photograph of Siphon Experimental setup

The acoustic instrument consisted of four 25mm hydrophones Type SQ31 by Sensor Technology Inc. (Canada) and one 50mm speaker Type K50WP by Visaton (Germany). Hydrophones H1-H3 were installed in the left leg of the siphon. Hydrophone H4 was installed in the right leg of the siphon 75mm above the speaker and used as a reference receiver (see Figure 4.2). The hydrophones and the speaker were attached securely to two aluminium tubes which were lowered into the opposite legs of the siphon and kept at the same positions in all of the experiments conducted in the siphon.

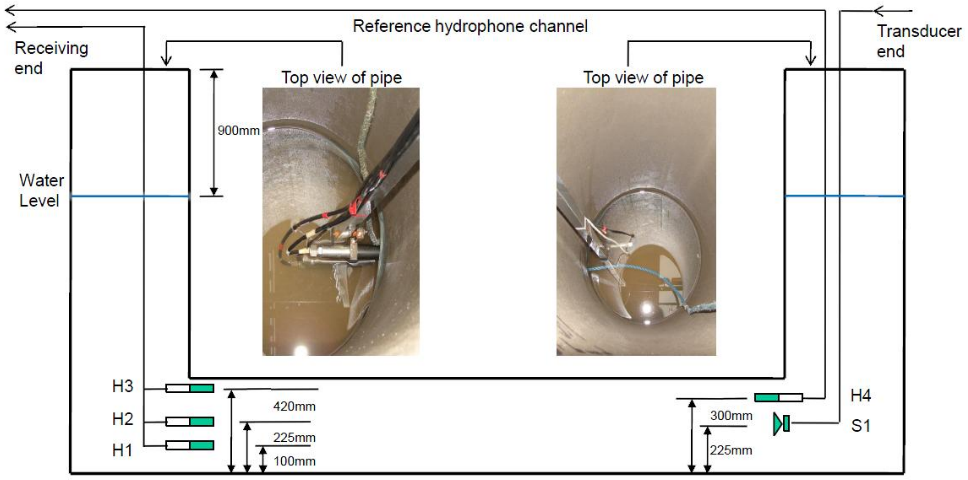


Figure 4.2 Arrangement of sensors in the siphon

The data acquisition and signal processing facilities used in these experiments consisted of: (i) a PC installed with WinMLS software to control the sound card which generated a 10-second sinusoidal sweep in the frequency range of 100 - 6000 Hz; (ii) an 8-channel high-pass hydrophone filter used to remove unwanted low-frequency machinery noise from the signals received on hydrophone H1-H3; (iii) a B&K Type 2610 measuring amplifier and a dual variable filter Kemo VBF 10M filter which were used to condition and filter the signal received on the reference hydrophone in the 100 – 4000 Hz range. In addition, a B&K Type 2708 power amplifier was used to drive the underwater speaker. Rotel Type RA-9708 X Stereo amplifier and headphones were used to control subjectively the quality of the signal produced by the underwater speaker.

WinMLS software controlled the sound card which generated a sinusoidal sweep (chirp) in the frequency range of 100 – 6000 Hz. Sinusoidal sweep (chirp) is widely used excitation signal to measure the transfer function. Chirp-based measurements are considerably less vulnerable to the deleteriouseffect of time variance, they are best suited for outdoor measurements in presence of dynamically rough water surface [3].The signal was repeated 8 times and averaged to increase the signal-to-noise ratio. The sounds received on the four hydrophones were digitised at 22050Hz sampling rate and recorded on the PC. The sinusoidal sweep signals were then deconvolved using WinMLS software so that the acoustic pressure impulse response of the siphon was obtained at the four hydrophone positions at a high signal-to-noise ratio. Typically, three measurements were recorded for each of the experimental conditions studied in this work to provide data for the statistical analysis. In addition, the background noise was regularly recorded to control the levels of noise produced by the other machinery and equipment operated in the laboratory at the time of measurements.

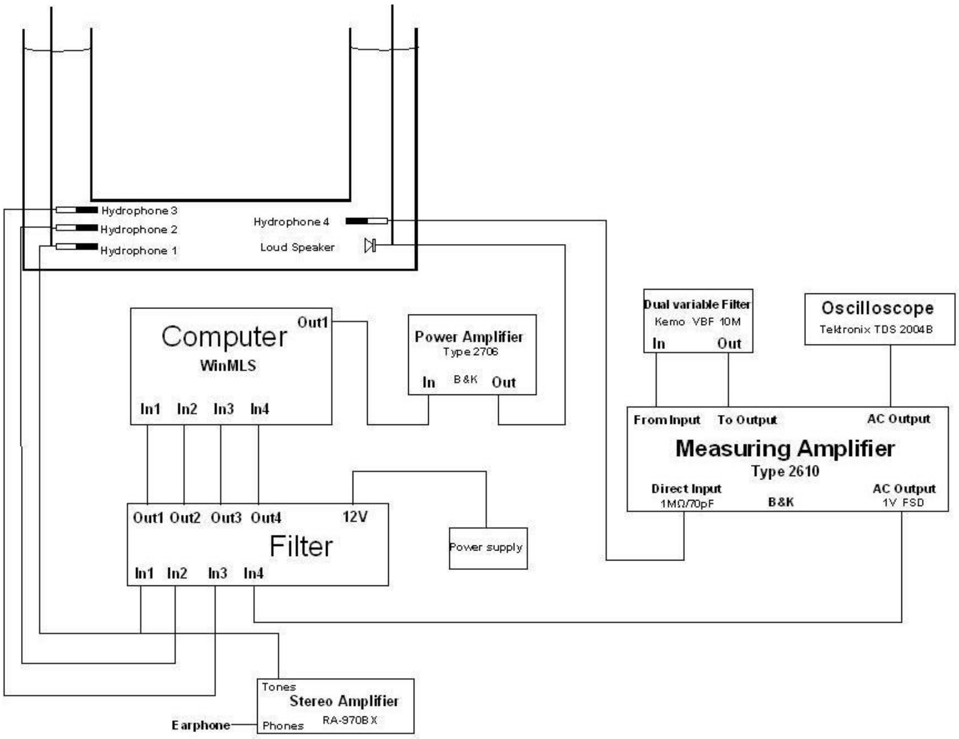


Figure 4.3 A schematic diagram of the data acquisition and analysis system

4.2.2 Laboratory Data Collection

The effects of the following conditions on sound propagation in the siphon were studied: (i) water level in the siphon; (ii) air bubbles effect; (iii) type of the medium surrounding the siphon; (iv) amount of sediment deposited in the siphon; (v) various types of the wall damage.

The original acoustic signals received on hydrophones H1-H3 were de-convolved with WinMLS software to obtain the acoustic pressure response.

These pressure data were used in signal processing and condition analysis, the results of which will be presented in the next section. The purpose of experiments (i), (ii) and (iii) listed above was to find out whether or not conditions (iv) and (v) can be detected acoustically in the presence of noise and environmental uncertainties.

4.2.2.1. Water Level Effect

The purpose of this experiment was to determine the effect of the water level on the acoustic pressure emitted by the speaker into the siphon. The horizontal section of the siphon was fully surrounded by dry sand up to a level of 1m from the bottom of the box. The water level in the siphon was varied between 600mm and 1200mm from the top of the left vertical pipe. Measurements were taken with 100mm difference of water level.

4.2.2.2. Influence of Air Bubbles

Air bubbles are usually present in the turbulent water flow and can strongly affect sound propagation in a certain frequency range. For this purpose a fish tank air pump with a 150mm long porous airstone (Figure 4.4) was installed on the bottom in the middle of the horizontal section of the siphon to study the effect of bubbles on the sound propagation in the siphon. The acoustic field in the siphon was recorded in the presence and absence of bubbles.

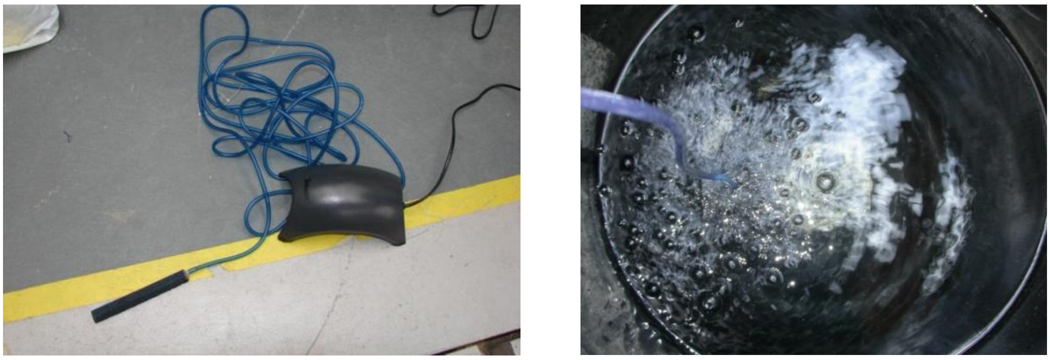


Figure 4.4 Air pump and air stone used in bubble effect experiments

4.2.2.3. Siphon Surrounding Medium

Measurements were taken with the siphon surrounded by different types of medium to study the effect on the sound propagation in the siphon. Surrounding conditions were designed as following: (i) exposed (air medium); (ii) dry sand in variable amount; (iii) water in variable level.

Exposed condition was set as approximately 50% of the horizontal section of the siphon being exposed to air, and it was the initial condition of the sand and water level experiments. Approximately 6 tonnes of sand were added to the box on the reference exposed condition, one tonne at a time in six sequential steps, and equally distributed in the box. Four conditions: exposed; 2 tons; 4 tons and all 6 tons were studied, see Figure 4.5.

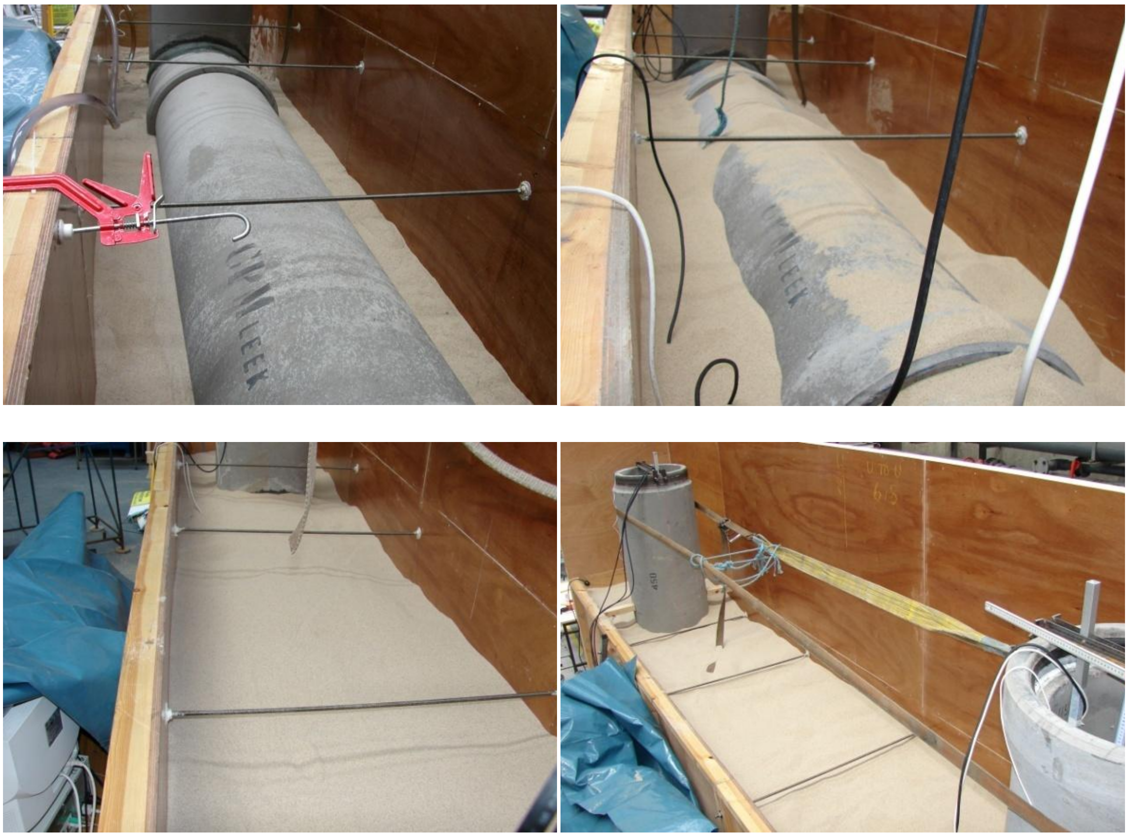


Figure 4.5 Photographs of dry sand conditions studied

Approximately six tonnes of dry sand were then removed from the box so that the siphon condition became exposed again as the initial condition for water level experiments. A water pump working at flow rate of 20 litre/min was used to fill the box with water. During this experiment measurements were carried out at 25 minute intervals which corresponded to approximately 500 litres of water pumped into the box every 25 minutes. Ultimately, the water level in the box was aligned with that inside the siphon (900mm below the open end of the siphon). This experiment required in total approximately 4500 litres of water and lasted 3 hour 45 minutes. These conditions are illustrated by photographs shown in Figure 4.6.

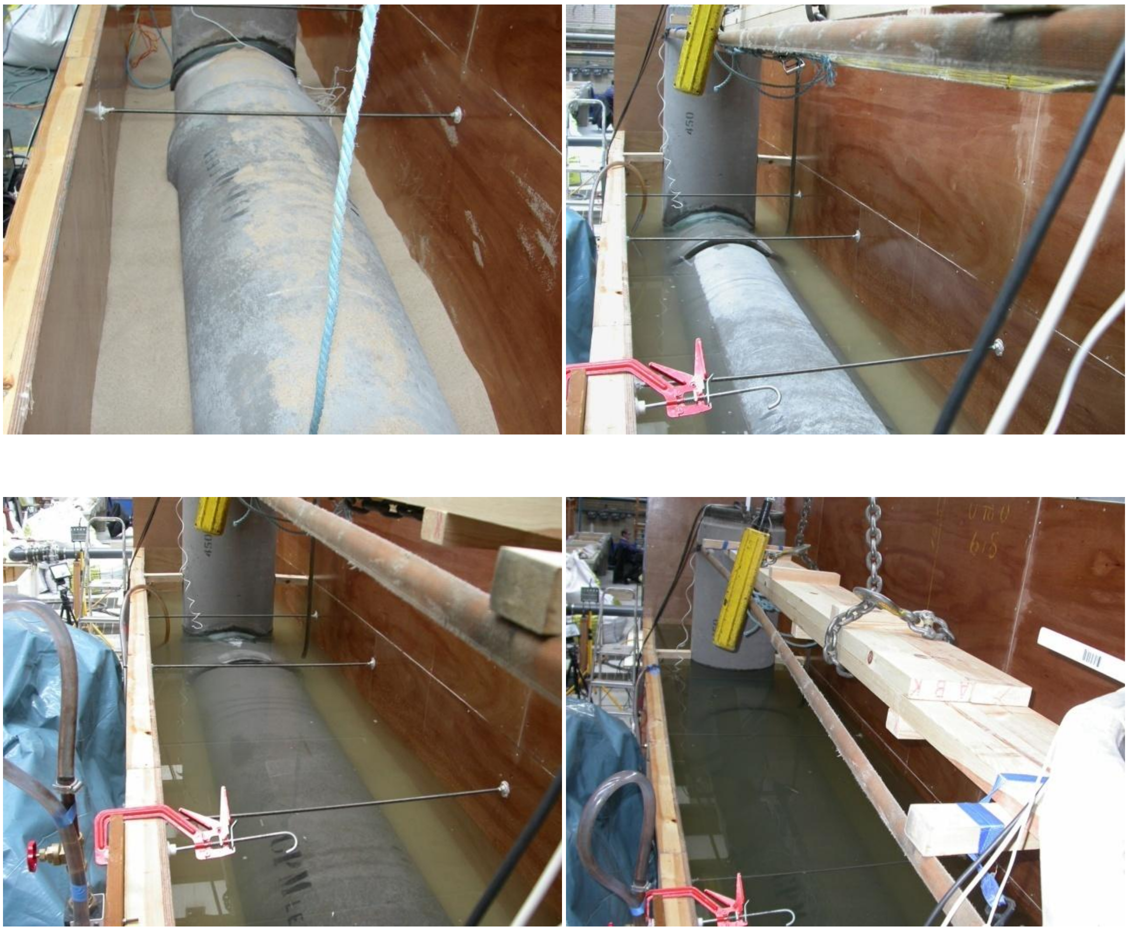


Figure 4.6 Siphon conditions in water level effect experiments

The comparison results of the effects of different amount of surrounding sand and water on the acoustic field in the siphon will be presented in the next ‘signal pre-processing ’ section.

4.2.2.4 Siphon with variable amount of sediment

In this experiment the siphon was fully surrounded by dry sand and filled with water up to the level of 900mm below the top. Ten 5kg acoustically transparent bags were prepared and filled with fine sand. The maximum cross-sectional dimension of one sand bag corresponded to approximately 20% of the pipe cross-section. Several bags at a time were tied to a 9m rope separated by a 300mm distance as shown in Figure 4.7. The number of bags deposited in the siphon by these means varied from 1 to 10. The acoustic sensors were removed and installed again each time when the bags were deposited or taken out of the siphon before the measurements were taken.

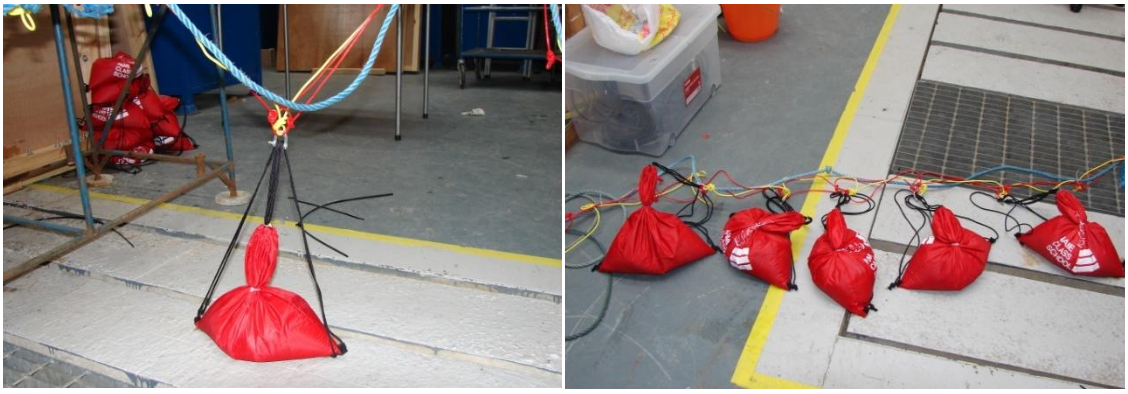


Figure 4.7 Sandbags used in sediment experiments.

4.2.2.5 Wall Damages

The purpose of this experiment was to study the effect of wall damage on the acoustic field in the siphon. Artificial cuts were inflicted to the top of the horizontal section of the siphon. The siphon remained exposed and then surrounded by water up to the level matching with the reference level inside the siphon after the damages were conducted. Measurements were taken in the presence of 6 different damages: (i) 50mm longitudinal cut; (ii) 100mm longitudinal cut; (iii) 200mm longitudinal cut; (iv) 200mm longitudinal and 55mm transversal cuts; (v) 200mm longitudinal and 150mm transversal cuts; (vi) 200mm longitudinal cut and 120mm x 70mm hole (see Figure 4.8).

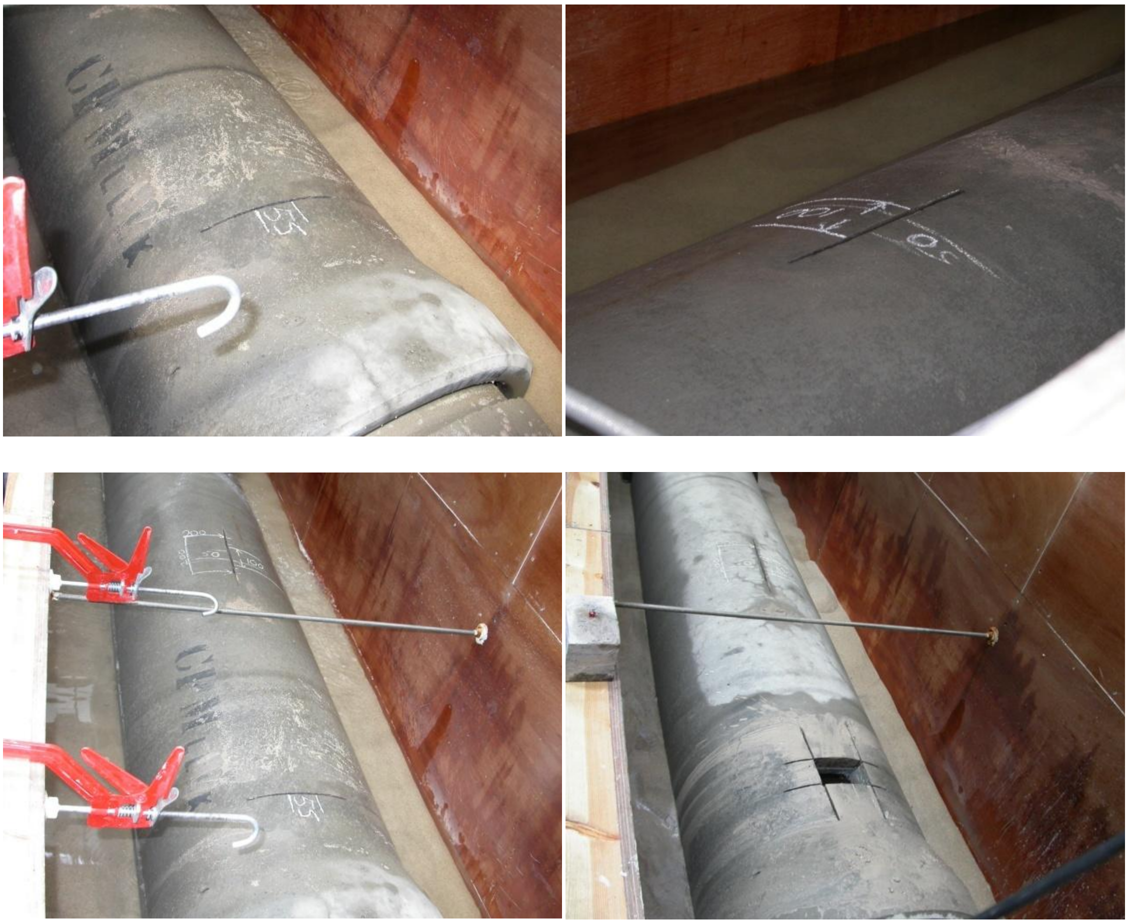


Figure 4.8 Photographs of the artificial damages on the wall

4.3 Signal Pre-processing

The acoustic impulse response of a physical system is a very useful quantity. It contains information of the sound speed, system geometry and conditions at its boundaries [3]. Any change in these properties is reflected in a change in the acoustic impulse response. The acoustic pressure response data were obtained through deconvolution on the original signals received on the hydrophones H1-H3 using the convolution theorem:

$$h_j(t) = \mathrm{Re}\left\{F^{-1}\left[F\{y_j(t)/(F\{x(t)\}+\gamma)\right]\right\} \qquad (4.1)$$

where $x(t)$ is the excitation signal, $y_j(t)$ is the signal recorded on hydrophone j and γ is the regularization factor. Only the real part of the signal was analysed and used in this research, the phase information contained in the signal may also provide useful features for condition analysis and may be considered as future work. Figure 4.9 shows a recorded acoustic signal of an empty siphon and the corresponding pressure response.

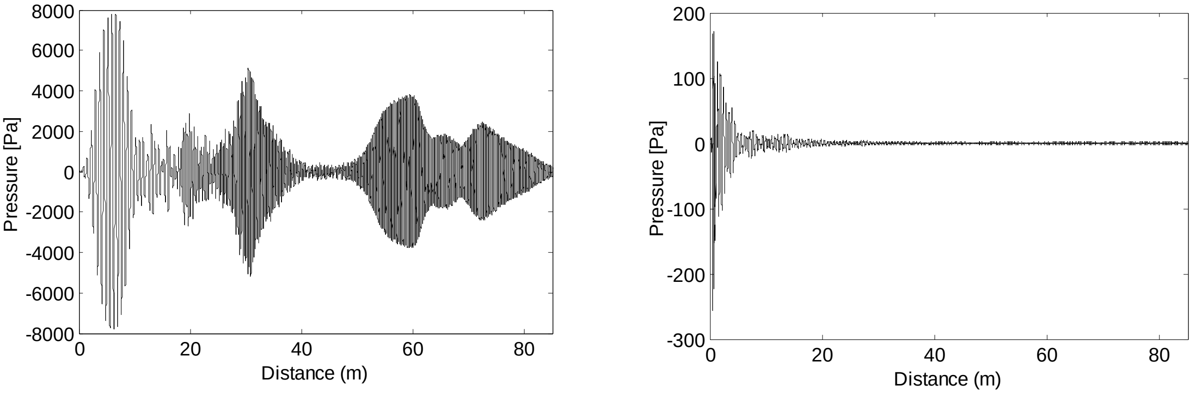


Figure 4.9 Recorded acoustic signal on one hydrophone of clean siphon (left) and its pressure response (right)

Figure 4.10 shows examples of deconvolved impulse response data received on all three receiving hydrophones of the clean siphon and the siphon with one blockage. It was noted that the data received on three hydrophones are remarkably similar and it appears difficult to detect visually the blockage from the impulse response data

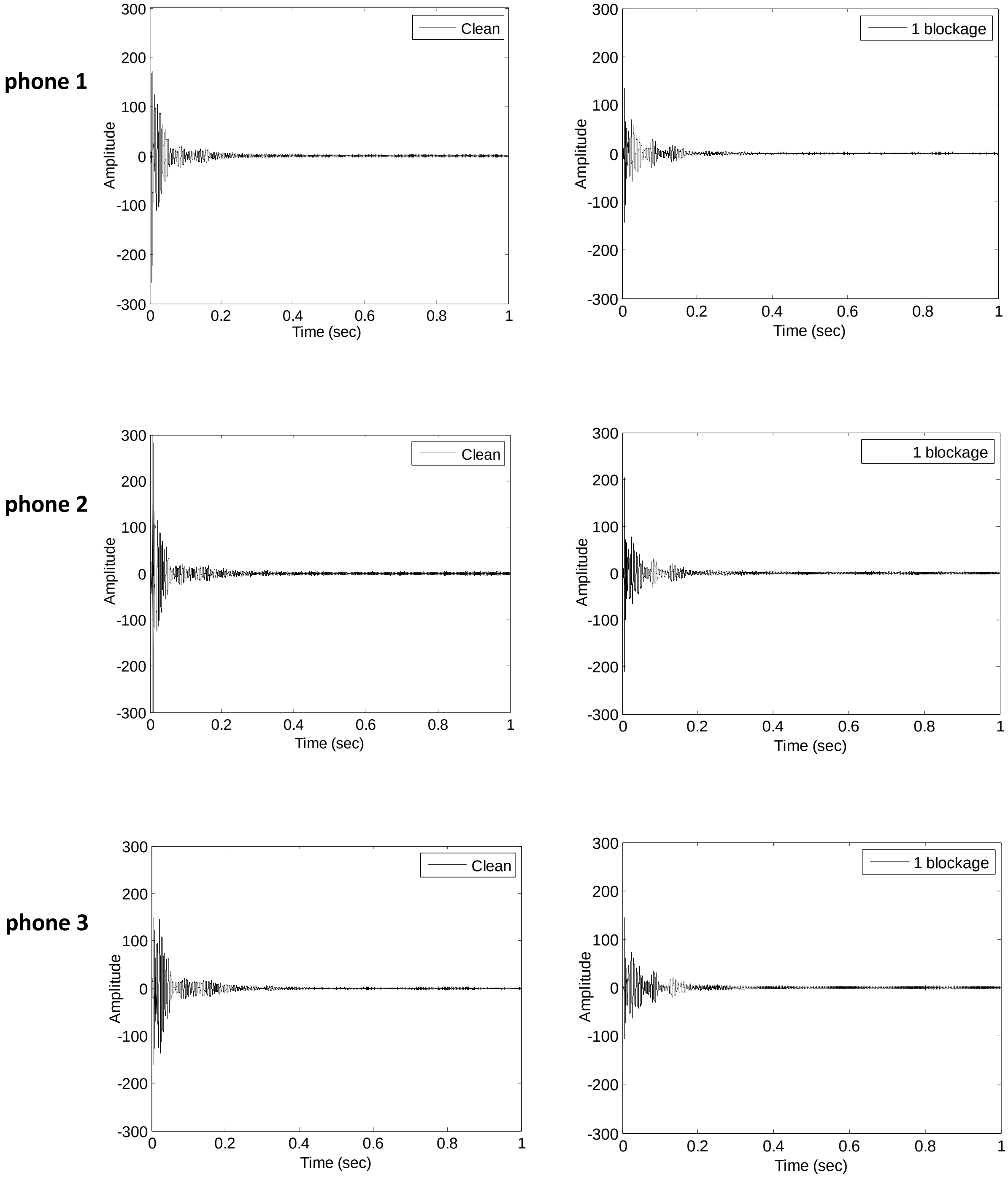


Figure 4.10 Examples of time domain impulse response data received on three hydrophones of two siphon conditions. Left column corresponds to clean siphon condition while right column shows the data for a siphon with one blockage.

A major goal of the data pre-processing part was to modify the measured data obtained from the data acquisition part so that those data could be more suitable for the further processing in feature extraction and classification. It was assumed that the acoustic impulse response carries sufficient dynamic system information and can be used to study the characteristics of the system. However, the impulse response is a broad band signal and only a part of its spectrum is affected by the change in the siphon condition.

The impulse response obtained from Equation (4.1) was then used to determine the sound pressure level as a function of time according to the Schroeder integral [4]:

$$L(t) = 10\log_{10}\left(\frac{1}{\Delta}\int_{t}^{t+\Delta}\hat{h}^{2}(t)dt\right) \tag{4.2}$$

where t is a time instant at which the sound pressure level is calculated, Δ is the duration of the integrating time interval and $\hat{h}(t)$ is the impulse pressure response obtained on one hydrophone in a particular frequency band. The change in the sound field is easier to be detected and observed by measuring the sound pressure level, the SPL data is therefore more suitable than the original impulse response data for further analysis.

4.3.1 Reproducibility of the Experiments Performed

In order to study the stability of the measurement system, reproducibility experiments were performed at a 1 hour interval. The sensors were kept at the same position as shown in Figure 4.1 and the siphon was filled with clean water up to the reference level. No change in the siphon conditions were made during these measurements. The RMS sound pressure level was calculated over 9 hour period using Equation (4.3) and presented in Figure

4.11 (left), where $p(t)$ is the instantaneous pressure. The maximum difference between the RMS sound pressure levels was less than 2.3% during the same time interval.

$$p_{rms} = \sqrt{\frac{1}{T}\int_0^T p^2(t)dt}$$

$$\text{rms}(SPL) = 10\log(\frac{1}{T}\int_0^T p(t)^2 dt) \tag{4.3}$$

Another set of experiments was conducted to determine if the influence of the sensors position in the siphon needs to be taken into account during those experiments which require removing and re-placing the hydrophone array. The temperature and water levels were kept the same and no change in the siphon conditions was made in this experiment. A measurement was taken in the clean siphon, the transmitters were then removed and placed back in the original position with possible small misalignment. The same measurement was then repeated to determine if there was noticeable change on the acoustic pressure data. The correlation coefficient was calculated using Equation (4.4) to measure the similarity between two sets of SPL data x_i and y_i:

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4.4}$$

Where $\bar{x}$ and $\bar{y}$ are the means of x_i and y_i. The correlation coefficients of 6 sets of SPL data obtained before and after reinstalling the sensor back to the original position are presented in Figure 4.11 (right).

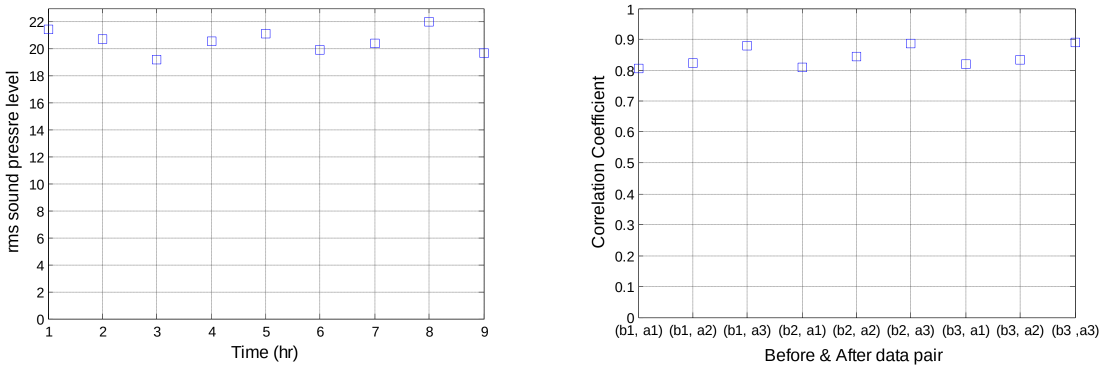


Figure 4.11 Left: RMS sound pressure; right: correlation coefficients of SPL data obtained before (b) and after (a) reinstalling the sensor.

4.3.2 Digital Filters

Digital filters are commonly used to remove unwanted components like noise, or to extract useful parts from the recorded signals at some frequencies. Butterworth, Chebyshev and Elliptic filters are frequently used for digital signal analysis. Their performance can be summarized as following [5]: the Butterworth filter of order 3 is found to provide the best roll-off; the Chebyshev filter is found to provide the most flat response in the design frequency range; the magnitude response of the Elliptic filter is somewhat between that provided by the Butterworth and Chebyshev filters, but it shows the most monotonic variation in the phase response. In this research, the 3rd order Butterworth filter was adopted to process the acoustic pressure response to determine the frequency range within which a change in the siphon condition can be noticed and detected.

Figure 4.12 presents the sound pressure levels obtained from hydrophone 1 for the water levels 200mm, 600mm, 900mm and 1300mm below the top of the siphon. The details of water level experiments are given in section 4.2.3.1. The results show that the water level has a very noticeable effect on the sound pressure particularly in the frequency range below 1000 Hz. This phenomenon is associated with a strong interference between the sound wave incidents on and reflected by the free water surfaces in the vertical legs of the siphon. This effect is also observed consistently in the data recorded on all the other two hydrophones.

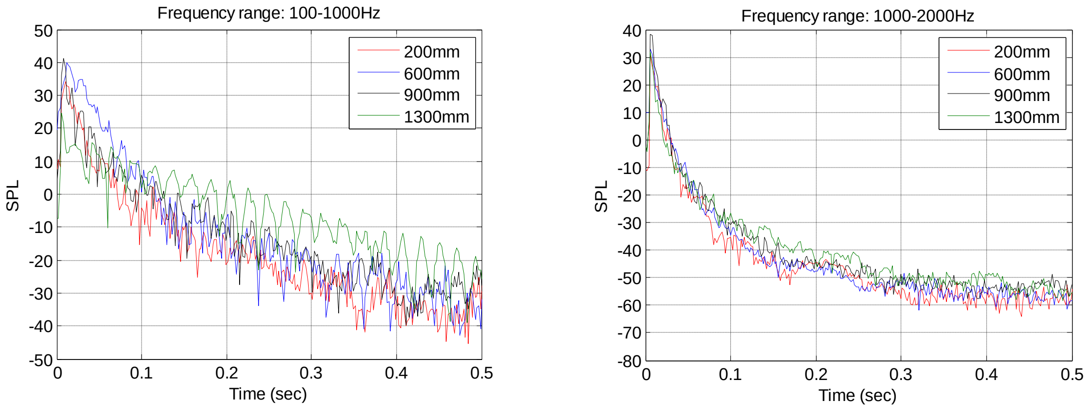


Figure 4.12 Examples of SPL at different water levels in the siphon: 200mm; 600mm; 900mmm and 1300mm at low and high frequency range.

Bubble effect experiment stated in section 4.2.3.2 aims to study the effect of air bubbles on the sound propagation in the siphon. The results of this experiment show that there is a strong similarity in the acoustic field between the presence and the absence of bubbles, Figure 4.13 shows the sound pressure levels with and without the presence of air bubbles. The degree of correlation between these two conditions is 90% calculated using equation (4.4) and is comparable to that which was observed in the reproducibility experiment. Therefore, the effect of bubbles on the sound field in the siphon is relatively small and can be neglected.

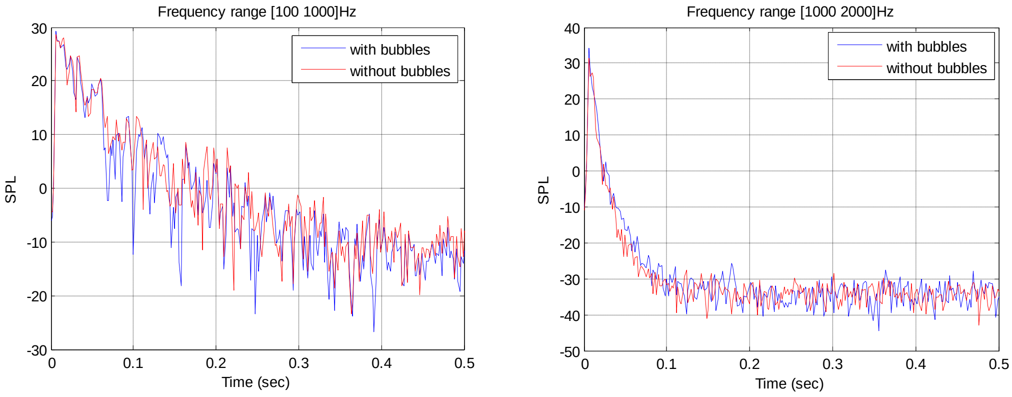


Figure 4.13 Sound pressure level of an empty siphon with and without the presence of air bubbles at low (left [100-1000] Hz) and high (right [1000 2000] Hz) frequency range.

Figure 4.14 and 4.15 show the sound pressure level comparisons between different amount of surrounding dry sand and water effect in two frequency bands, respectively. The sound pressure level results suggest that the effect of different amount of surroundings on the acoustic field in the siphon is progressive, but limited. More difference showed in SPL at lower frequencies. Relatively, more area covered by dry sand, the lower sound pressure level resulted in the siphon; and on the contrary, the higher surrounding water level caused higher interior sound pressure level. Whether the surrounding medium effects should be taken into account in further siphon condition analysis depends on the exact amount and type of the surrounding medium, also the time window that chosen to extract information from.

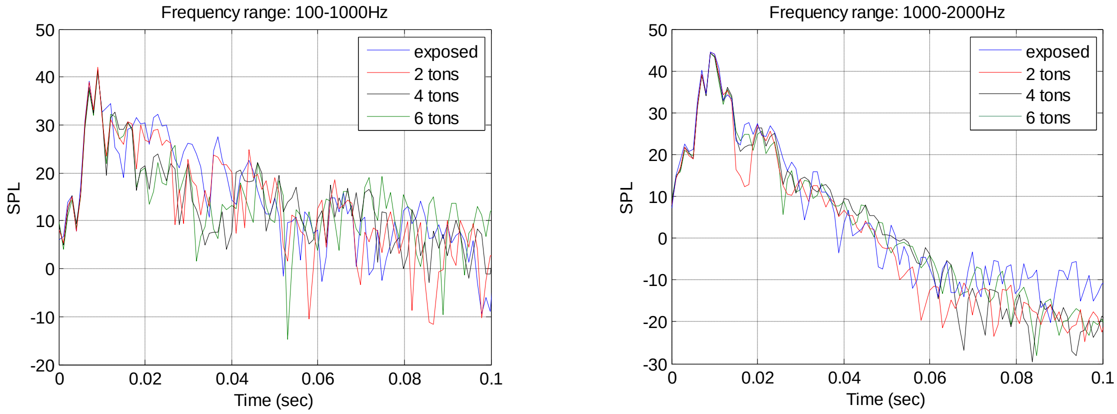


Figure 4.14 Sound pressure level of an empty siphon surrounded by different amount of dry sand at low frequencies (left) and high frequencies (high).

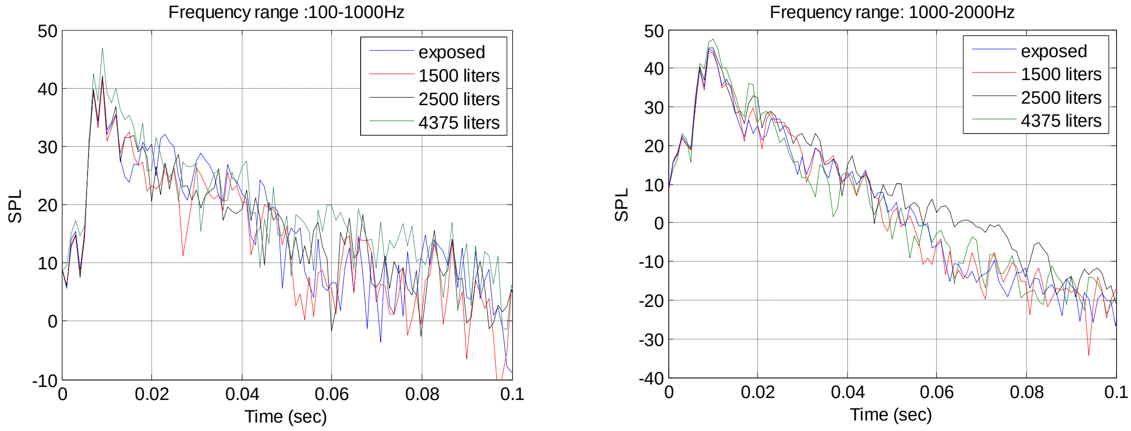


Figure 4.15 Sound pressure level of an empty siphon surrounded by different amount of water at low frequencies (left) and high frequencies (high).

Figure 4.16 shows the sound pressure level of different sediment conditions in several frequency bands. The original pressure response was filtered by 3rd order Butterworth filter in 4 frequency bands from 100 to 2400Hz: 100~600Hz; 600~1200Hz; 1200~1800Hz and 1800~2400Hz , to calculate the sound pressure level. It was found that the sound pressure level did not depend on the amount of sediment when the frequency was higher than 2000Hz, and that the difference between different conditions was mostly distinguishable on the time line from 0 to 0.04 second. The data within that range contained mostly useful characteristics for siphon condition analysis.

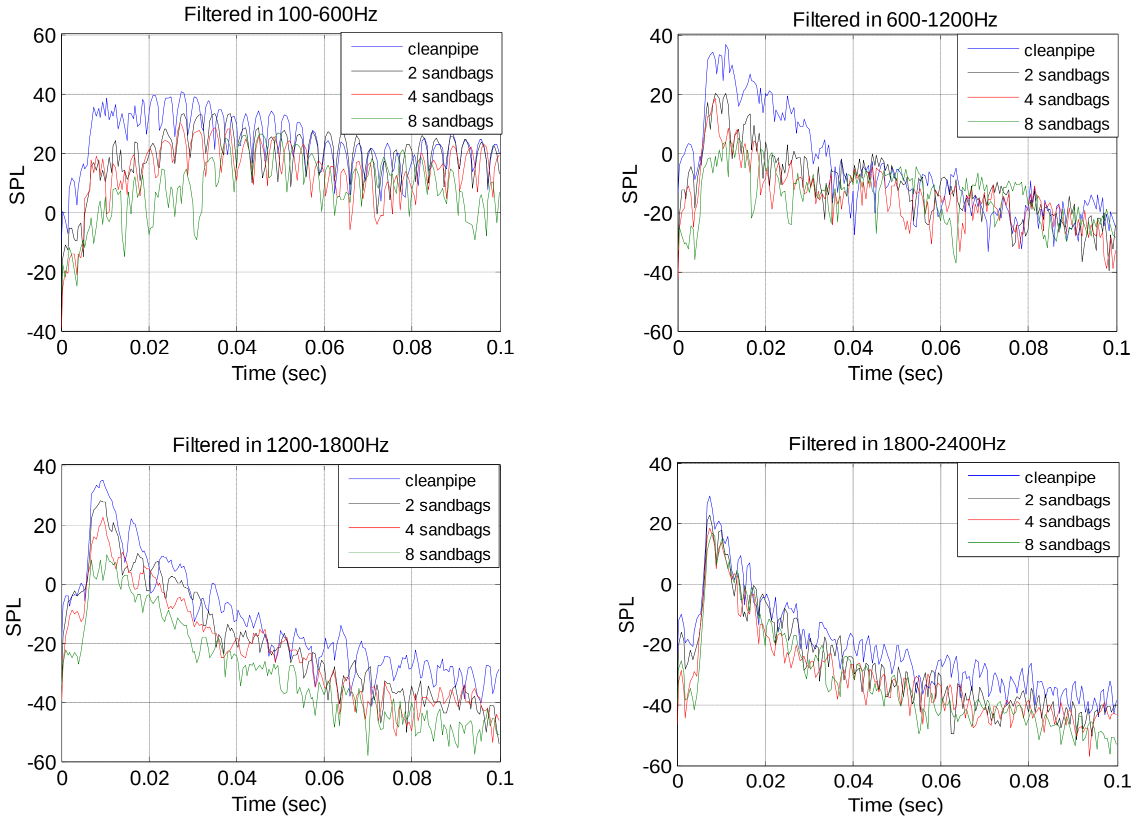


Figure 4.16 Sound pressure levels of different sediment conditions in several frequency bands

Figure 4.17 gives the sound pressure levels of several damage conditions for different frequency bands. The sound pressure level was found not sensitive to damage at frequencies below 1000Hz. From 1000 to 4000Hz range, there is noticeable difference between undamaged condition and damaged conditions. However, the differences between different types of damages were not obvious from the visual examination of these graphs. Choosing time window seems critical for damage condition analysis as it can be seen from these figures that data within 0.02 ~ 0.06 sec appear to be more separable from one type of damage to another.

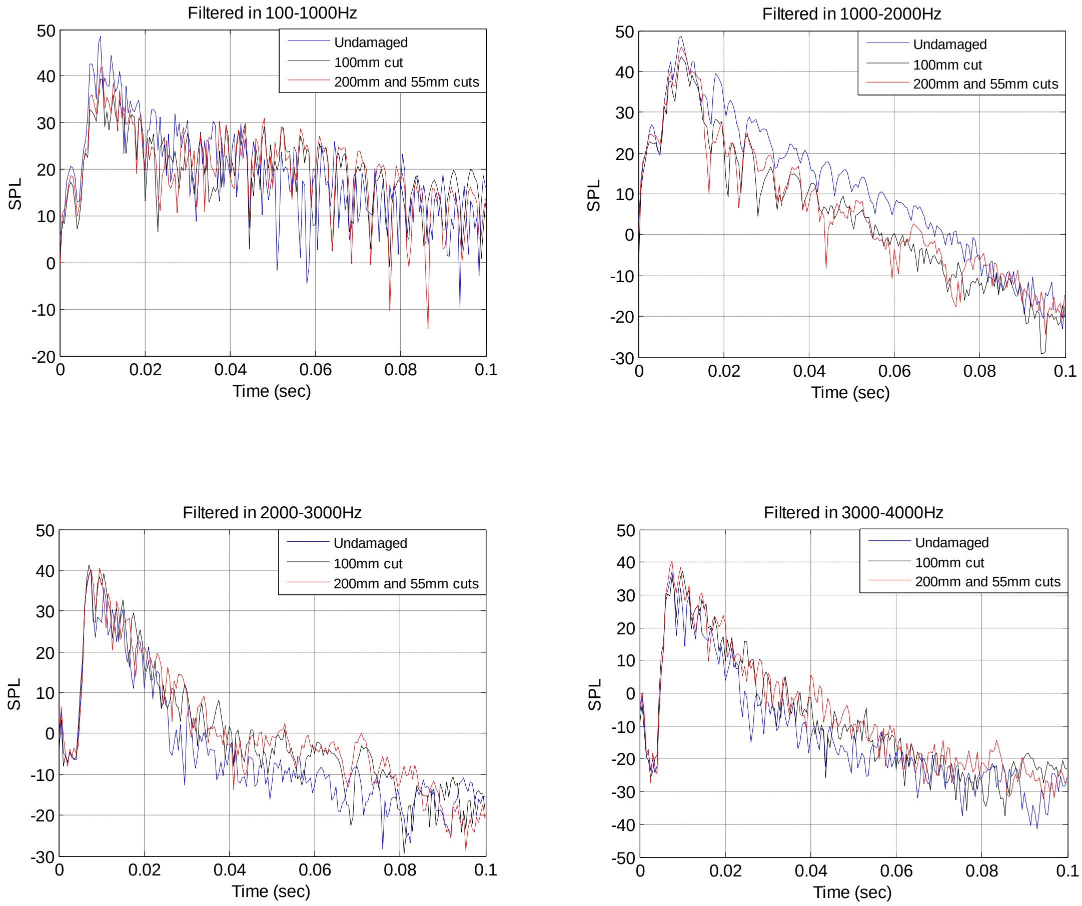


Figure 4.17 Sound pressure level of damage conditions at different frequency bands

4.3.3 Discrete Wavelet Transform

The discrete wavelet transform (DWT) is a powerful tool which has been proven to have a relationship with digital filter banks which can be achieved with a tree algorithm. For many signals, the low-frequency content is the most important part. It is what gives the signal its identity. For this reason, in wavelet analysis, approximations and details are components named after the filtering process corresponding to low frequencies and high frequencies, respectively. The decomposition process can be iterated, with successive approximations being decomposed in turn, so that one signal is broken down into many components. This is called the wavelet decomposition tree. The frequency responses for the decomposed signals for a system with a sampling frequency f_s are shown in Table 4.1. The frequency band on a certain decomposition depth p can be calculated as:

$$\frac{k}{2^p} f_s \square \frac{k+1}{2^p} f_s \qquad (4.5)$$

Where k is the node number in the wavelet tree (Figure 4.18), $k = 0,1,\ldots n$.

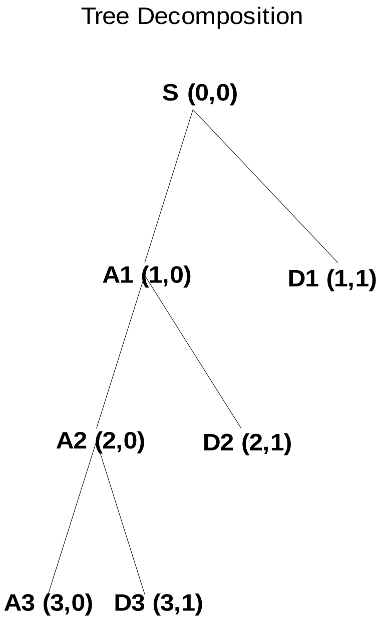


Table 4.1
Frequency bands of DWT decomposition of depth 3

Decomposition level	Frequency bandwidth (Hz)
D_1	$\frac{f_s}{2} - f_s$
D_2	$\frac{f_s}{4} - \frac{f_s}{2}$
D_3	$\frac{f_s}{8} - \frac{f_s}{4}$
A_3	$0 - \frac{f_s}{8}$

Figure 4.18 Wavelet decomposition tree generated with MATLAB

The DWT coefficients represent the degree of correlation between the analyzed signal and the wavelet function at different instances of time. Therefore, DWT coefficients carry useful temporal information about the transient activity of the analyzed signal [6]. Figure 4.19 shows a 3 level wavelet decomposition tree using ‘db4’ as the wavelet function, s is the original analyzed pressure response signal, cA1-cA3 are approximation coefficients of s , cD1-cD3 are detail coefficients of s .

The coefficients can be used to reconstruct approximation and detail signals from their coefficients on any decomposition level, the decomposition and reconstruction phases together finished the signal filtering process. Meanwhile, the DWT coefficients themselves preserve the temporal information about the original signal, they have been proven effective for analysis of non-stationary signals in some applications [6], there have not been much effort in applying DWT on acoustic signal for condition classification, its effectiveness were studied in this research and will be presented in the following sections.

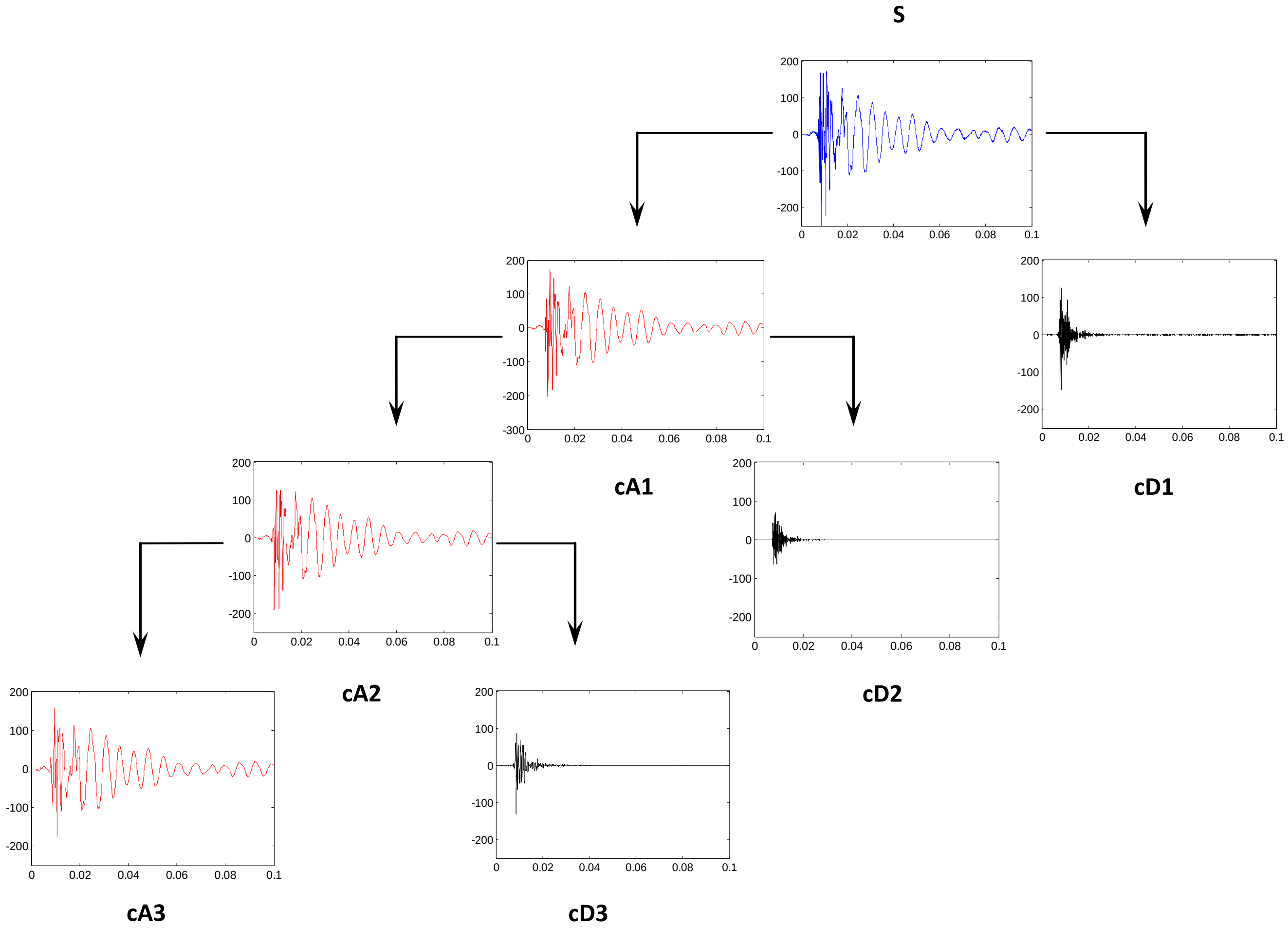


Figure 4.19Discrete Wavelet Decomposition of depth 3 of a clean siphon signal

Different wavelet functions have different effect on the performance of the decomposition process. The performance of the different types of wavelet functions would indicate their suitability to detect transient activity in acoustical signals, which correspond to different siphon conditions.

The waveforms of the wavelet function should be as similar to the transient activity to be detected in the acoustic signals. However, since the optimal waveform for this research is unknown, various types of wavelet function were considered: Daubechies (‘db’), Symlets (‘sym’) and Coiflets (‘coif’). Figure 4.20 displays the approximate shapes of several wavelets.

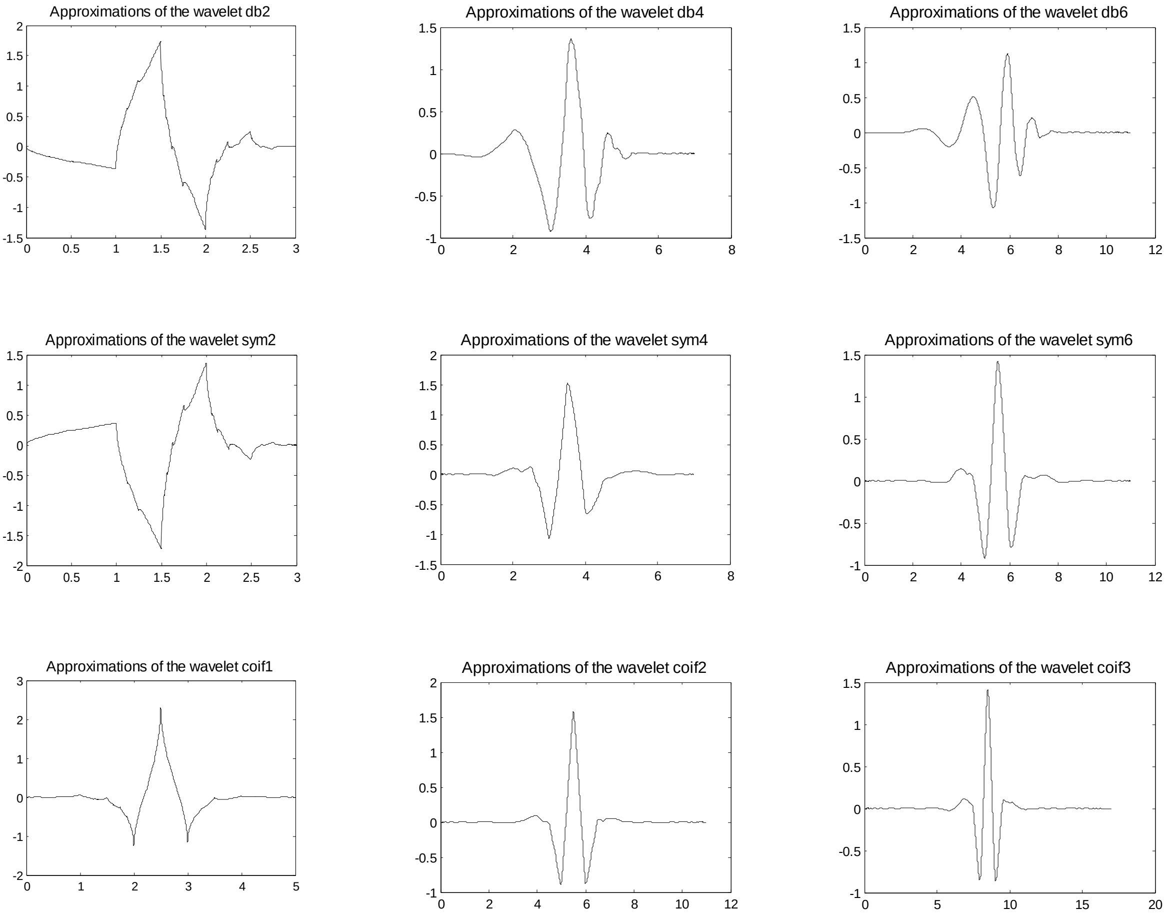


Figure 4.20 Approximations of wavelet functions of ‘db’, ‘sym’ and ‘coif’ in different orders.

Figure 4.21 shows examples of a recorded acoustic signal and the reconstructed signals using wavelet ‘coif2’, ‘sym4’ and ‘db4’. The Coiflets wavelets with order above 2, Symlets and Daubechies wavelets with order above 4 all appear to have similar waveforms as the original recorded signal. The cross-correlation coefficients were calculated between the original signal and each reconstructed signal, the reconstructed signal using wavelet ‘db4’ has the highest similarity. Therefore, Daubechies of order 4 was chosen as the wavelet function.

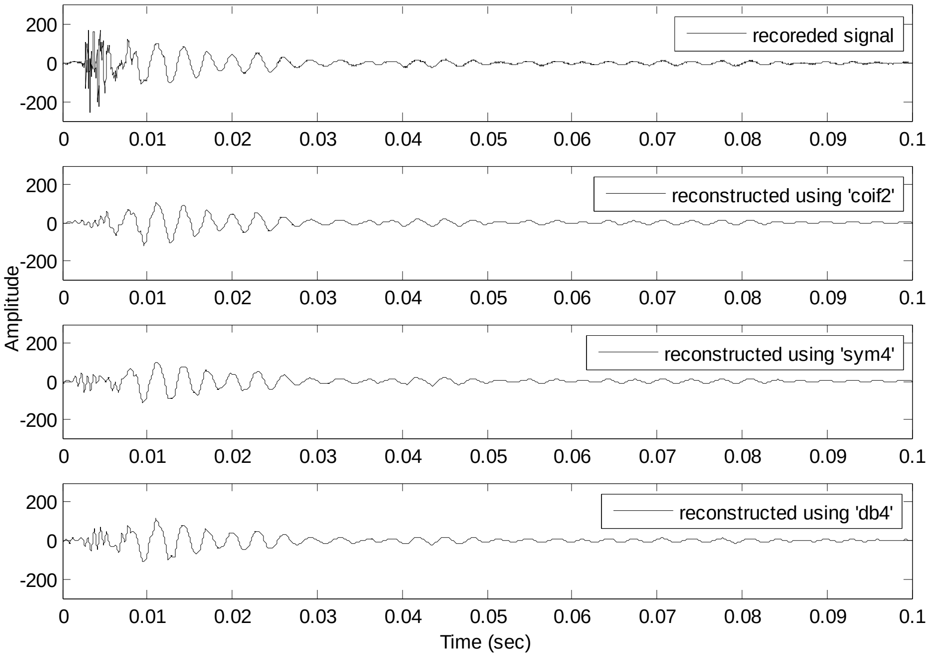


Figure 4.21 Examples of a recorded signal and reconstructed signals using wavelet ‘coif2’, ‘sym4’ and ‘db4’

Daubechies have no analytical formula for scaling and wavelet functions, but its coefficients are available in textbook and literatures for each order [7]. Using ‘db4’ as the wavelet function, the scaling function coefficients are:

$$h_0 = \frac{1+\sqrt{3}}{4\sqrt{2}}$$
$$h_1 = \frac{3+\sqrt{3}}{4\sqrt{2}}$$
$$h_2 = \frac{3-\sqrt{3}}{4\sqrt{2}}$$
$$h_3 = \frac{1-\sqrt{3}}{4\sqrt{2}}$$

The wavelet function coefficients are:

$$g_0 = h_3$$
$$g_1 = -h_2$$
$$g_2 = h_1$$
$$g_3 = -h_0$$

The scaling and wavelet functions are the sum of inner products of the coefficients and four values of the input data$s$, the scaling function is given by:

$$\phi(t) = \sqrt{2}\sum_{k=0}^{2N-1} h_k \phi(2t-k) \qquad (4.6)$$

The wavelet function is given by:

$$\psi(t) = \sqrt{2}\sum_{k=0}^{2N-1} g_k \phi(2t-k) \qquad (4.7)$$

The decomposition high-pass and low-pass filters were then obtained by:

$$L_o\_D = \left\langle \phi_{j,n}(t), \phi_{j-1,k}(t) \right\rangle$$

$$H_i\_D = \left\langle \phi_{j,n}(t), \psi_{j-1,k}(t) \right\rangle \qquad (4.8)$$

Where j,k are scale and shift parameters if applicable, n is the number of data samples. The reconstruction filters are the reverse of the decomposition filters. Figure 4.22 illustrates the ‘db4’ decomposition and reconstruction filters.

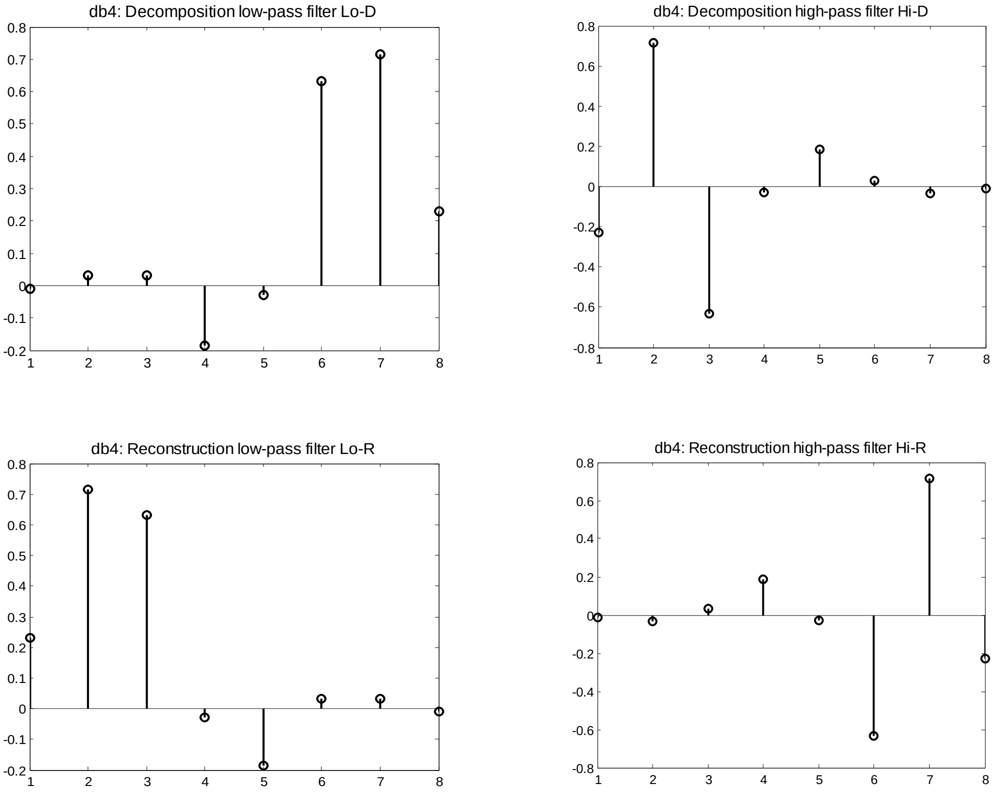


Figure 4.22 db4 wavelet decomposition filters (top) and reconstruction filters (bottom)

Each signal $x(t)$ in space V_j ($V_j = V_{j-1} + W_{j-1}$) can be expressed using the basic functions in each of the spaces:

$$x(t) = \sum_k cA_1(k)\phi_{j-1,k}(t) + \sum_k cD_1(k)\psi_{j-1,k}(t) \quad (4.9)$$

$$= A_1 + D_1$$

The decomposition starts with producing two sets of coefficients at scale $j-1$, the process repeats until reaches the frequency bandwidth as needed. The approximation and detail coefficients are calculated as:

$$cA_1(k) = \langle x(t), \phi_{j-1,k}(t) \rangle \tag{4.10}$$

$$= \left\langle \sum_n cA_0(n)\phi_{j,n}(t), \phi_{j-1,k}(t) \right\rangle$$

$$= \sum_n cA_0(n) \langle \phi_{j,n}(t), \phi_{j-1,k}(t) \rangle$$

$$cD_1(k) = \langle x(t), \psi_{j-1,k}(t) \rangle \tag{4.11}$$

$$= \left\langle \sum_n cA_0(n)\phi_{j,n}(t), \psi_{j-1,k}(t) \right\rangle$$

$$= \sum_n cA_0(n) \langle \phi_{j,n}(t), \psi_{j-1,k}(t) \rangle$$

Alternatively, they also can be written upon substitution with (4.4):

$$cA_1(k) = \sum_n L_o\_D(n-2k)cA_0(n) \tag{4.12}$$

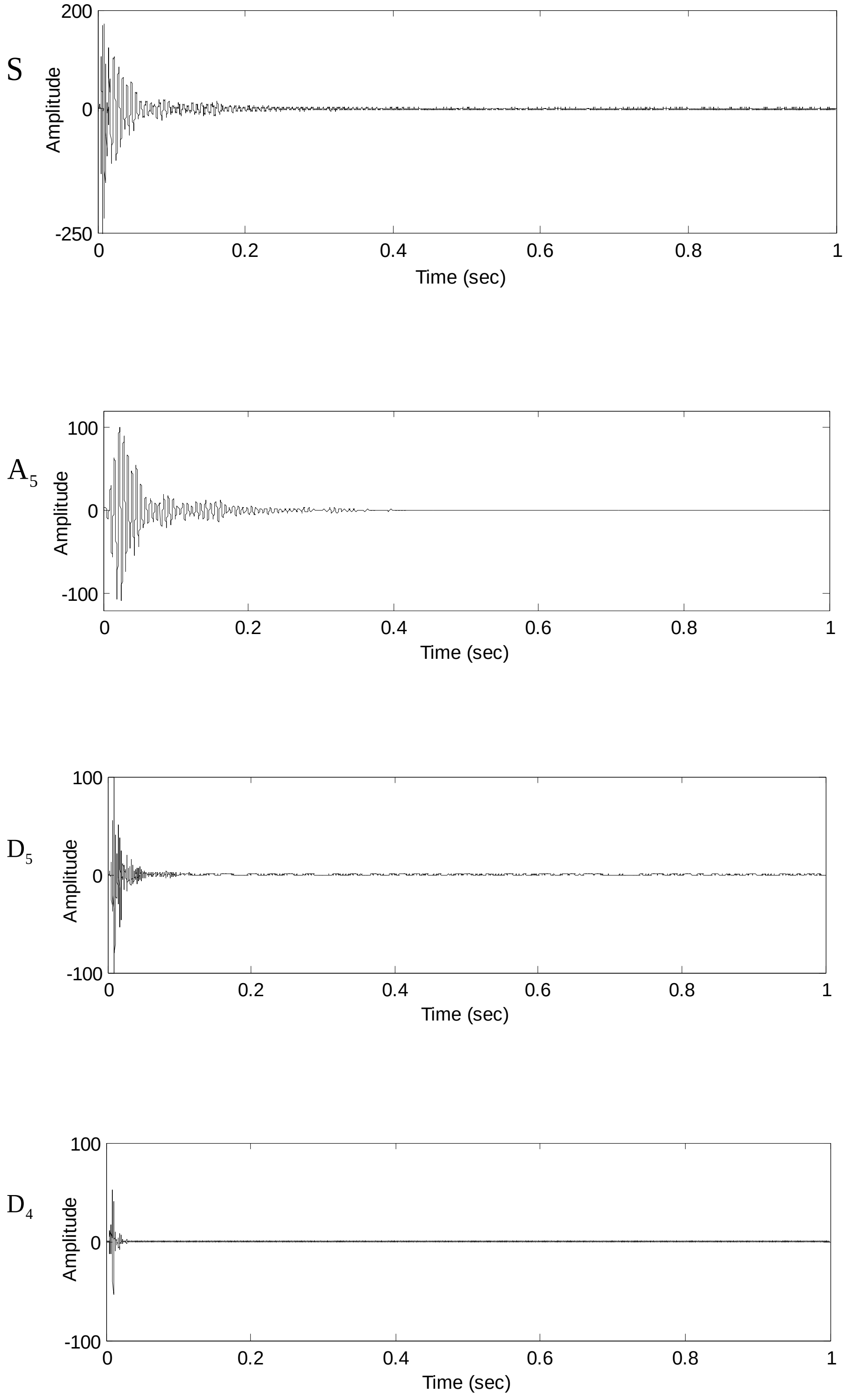
$$cD_1(k) = \sum_n H_i\_D(n-2k)cA_0(n)$$

For a depth= p decomposition, signal $x(t)$ can be written as:

$$x(t) = \sum_k cA_p(k)\phi_{j-p,k}(t) + \sum_k cD_p(k)\psi_{j-p,k}(t) + \sum_k cD_{p-1}\psi_{j-(p-1),k}(t) + \ldots \sum_k cD_1(k)\psi_{j-1,k}(t)$$

$$= A_p + D_p + D_{p-1} + \ldots D_1 \tag{4.13}$$

The sampling frequency of acoustic pressure response in this research is $f_s = 22050$ Hz. For the sediment condition, based on the results from Butterworth filtering, a modified wavelet tree of depth 5 (Figure 4.24) was designed to decompose the original pressure response signals into 4 frequency bands with bandwidth calculated as: $B = f_s/2^5 = 22050/32$ =689Hz. Reconstructed signals with same length are shown in Figure 4.23: S is the original broad band signal; A_5 is the lowest frequency component of S; D_3 is the highest frequency component.



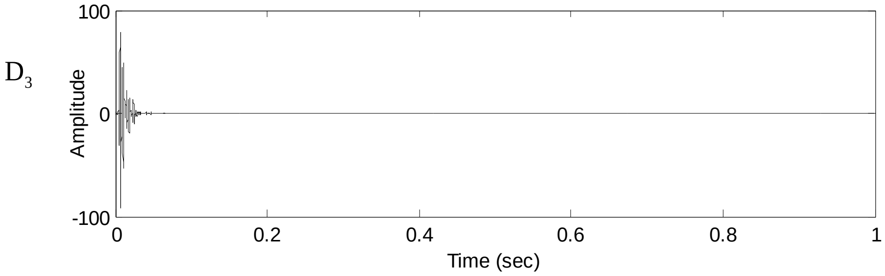


Figure 4.23 Reconstructed filtered signals of clean pipe pressure response used ‘db4’ as wavelet function, 5 depth decomposition, decomposition filter bandwidth=689Hz. **S** is the original signal, A_5 - D_3 : frequency from low to high with maximum frequency =2756Hz.

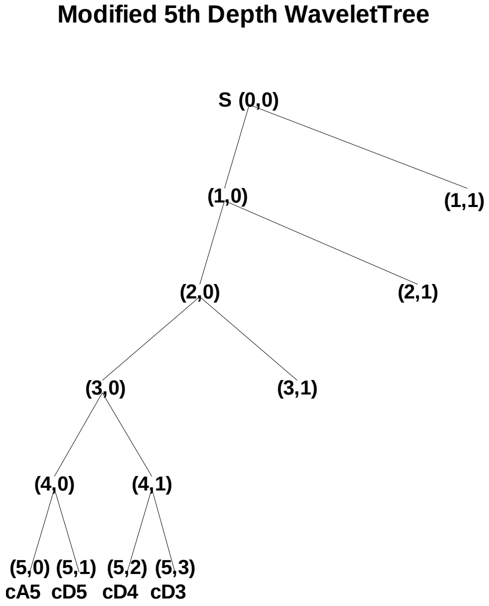


Figure 4.24 Modified ‘db4’ wavelet decomposition tree of 5 depth, each node corresponds to a set of decomposition coefficients, from cA5 to cD3: frequency goes from 100-2756Hz.

The sound pressure level was calculated using reconstructed signals A_5, D_5, D_4, D_3 to compare and determine the most relevant segments for condition identification. Sediment conditions plots are shown in Figure 4.25, the difference with the Butterworth filtering results is that Butterworth filtered signals were not distinguishable when the frequency was higher than 2000Hz, therefore only three frequency bands are useful for the further condition analysis, while the filtered signals obtained from the modified wavelet decomposition tree are showing noticeable difference in a certain time window $0 \le t \le 0.04$ sec at all four frequency bands which supply more data information for condition analysis.

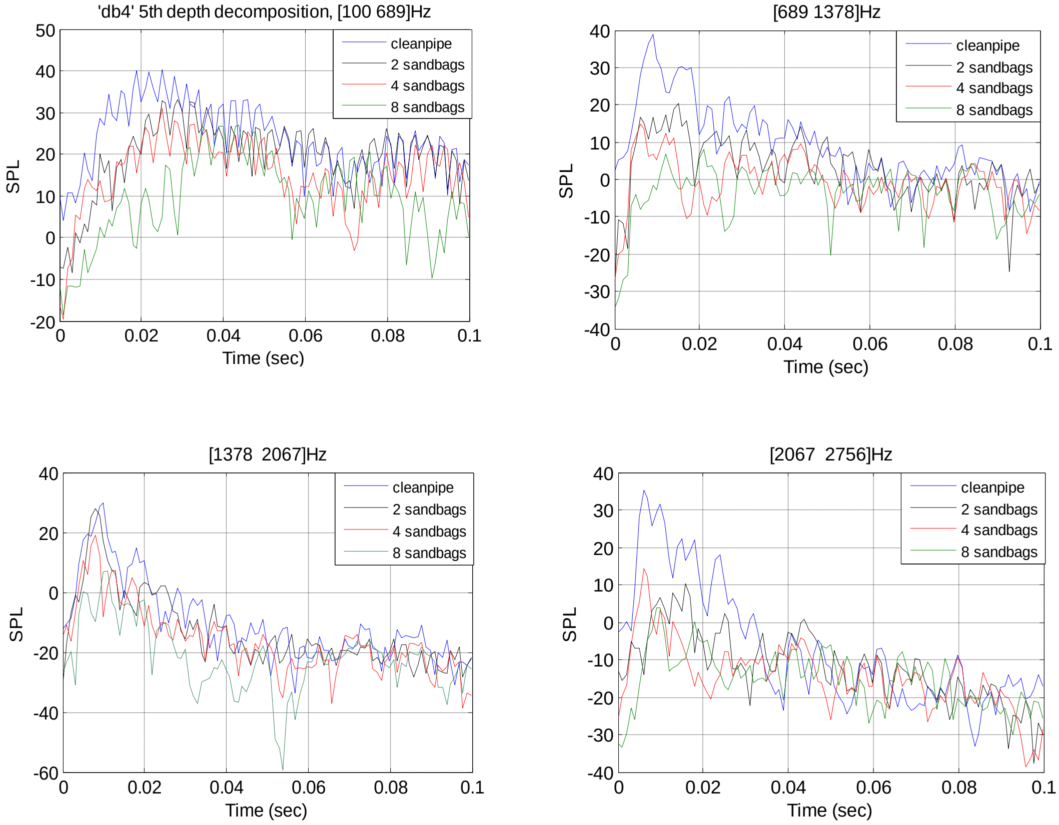


Figure 4.25 Sound pressure level of sediment conditions at 4 frequency bands defined by modified ‘db4’ wavelet decomposition tree as shown in Figure 4.23

The sound pressure level data obtained from acoustic signals filtered by Butterworth filter in damage conditions didn’t show enough difference in the frequency range below 1000Hz between different conditions as stated in section 4.3.1. Same pressure response signals then filtered using the modified wavelet decomposition tree to calculate sound pressure level, it was found that the sound pressure level responded to different types of damage in all four frequency bands in a wide time interval, but the same time window

$0.02 \le t \le 0.06$ sec as determined for digital filtered data is chosen for the consistency of analysis and comparison. Also, the SPL data generally show more differences between different conditions than the SPL data obtained from Butterworth filter. See Figure 4.26.

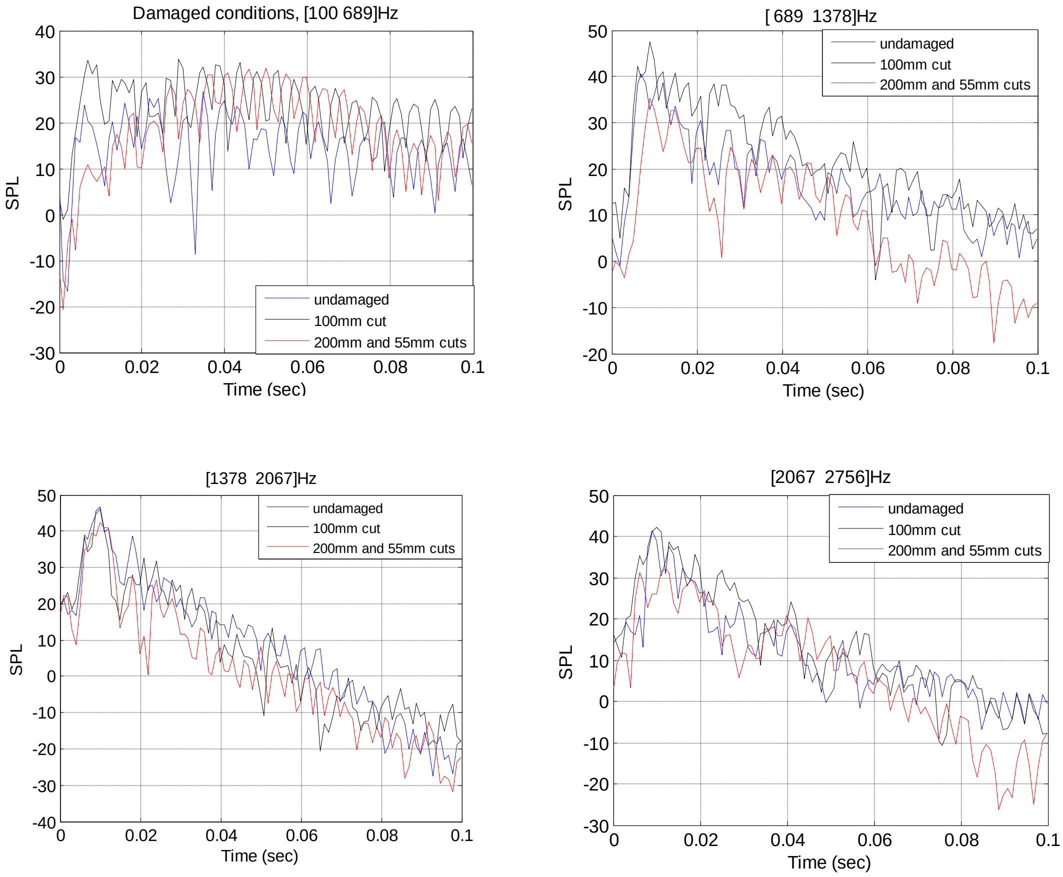


Figure 4.26 Sound pressure level of damage conditions at 4 frequency bands defined by ‘db4’ wavelet decomposition functions

4.4 Feature Extraction and Selection

Acoustic energy contained in the reflected acoustical signals is a valuable quantity to help defining the siphon conditions. For comparison purpose, filtered signals obtained from digital filters and DWT decomposition were used to extract useful features for siphon condition analysis.

The time window chosen for the detection of sediment conditions was $0 \le t \le 0.04$ sec for both Butterworth filtered data and DWT decomposed data. In the case of damage conditions, the time windows were $0.02 \le t \le 0.06$ sec and $0 \le t \le 0.04$ sec for data filtered by Butterworth and DWT, respectively.

4.4.1 Energy Data from Signals Filtered by 3rd Butterworth

The estimation of the acoustic energy in the pressure response at a given instant can be calculated as:

$$E = \int_{t}^{t+\rho} 10^{L(t)/10} dt \tag{4.14}$$

where the times t and $t+\rho$ define the time window, within which the integration was carried out. ρ was the time interval. $L(t)$ was the sound pressure level obtained from Equation (4.2).

The four frequency bands defined by 3rd order Butterworth filter for sediment conditions are: 100~600Hz, 600~1200Hz, 1200~1800Hz and 1800~2400Hz. The time window $0 \le t \le 0.04$ sec contained the most contents of the difference between conditions. Acoustic reflected energy was calculated in the time window in the four frequency ranges for each sediment condition. Table 4.2 presents an example of the energy data sets for a range of sediment conditions. These data were obtained using the pressure response recorded on hydrophone 1, signals recorded on the other two hydrophones were processed in the same manner.

Table 4.2 The acoustic energy as a function of frequency band and amount of porous sediment calculated in time window $0 \le t \le 0.04$ sec				
Relative energy	**100-600Hz**	**600-1200Hz**	**1200-1800Hz**	**1800-2400Hz**
Class1 (clean pipe)	133.9636	12.4720	9.3339	1.0605
Class2 (1bag)	42.3982	1.1934	1.9585	0.4606
Class3 (2bags)	34.3554	0.3877	1.8760	0.3009
Class4 (3bags)	22.0364	0.0772	0.2148	0.3039
Class5 (4bags)	15.0750	0.1247	0.3408	0.1237
Class6 (5bags)	11.3719	0.1224	0.2933	0.0590
Class7 (6bags)	10.9847	0.0931	0.1570	0.0869
Class8 (7bags)	13.6994	0.0715	0.1886	0.2001
Class9 (8bags)	5.2682	0.0214	0.0341	0.1153
Class10 (9bags)	1.3901	0.0174	0.0141	0.0801
Class11 (10bags)	1.0464	0.0244	0.0383	0.0265

Table 4.2 clearly illustrates that increasing the amount of porous sediment in the siphon results in a noticeable decrease in the calculated acoustic energy in the first three frequency bands, the energy contained in the highest frequency band 1800~2400Hz does not seem to follow the same pattern after the number of sandbag was more than 5, which was also verified in Figure 4.10. This information is essential in the pattern analysis of siphon conditions. The energy data of damage conditions were calculated and organized in the same manner as shown in Table 4.3. The frequency bands determined by the 3rd Butterworth filter were: 100~1000Hz; 1000~2000Hz; 2000 ~3000Hz and 3000 ~4000Hz. The time window $0.02 \le t \le 0.06$ sec was used to obtain the energy data in all 4 frequency bands. The energy value wasn't proportional to the size of the damage as suggested by the datasets, however, there was a pattern in energy distribution in three frequency bands from 1000 to 4000 Hz, it didn't apply to the lowest band, which was revealed also in Figure 4.11. As shown in Figure 4.27, in order to remain the most

useful frequency bands for condition classification, the highest frequency band 1800~2400Hz in sediment conditions (the dot line) and the lowest frequency band 100~1000Hz in damage conditions (the dot line) were removed.

Table 4.3 The acoustic energy as a function of frequency band and type of damage calculated in the time window $0.02 \le t \le 0.06$ sec

Relative energy	**100-1000Hz**	**1000-2000Hz**	**2000-3000Hz**	**3000-4000Hz**
Class1 (undamaged pipe)	0.0199	0.0238 $\times 10^{-3}$	0.0012 $\times 10^{-3}$	0.0007 $\times 10^{-3}$
Class2 (50mm cut)	0.0124	0.0712 $\times 10^{-3}$	0.0017 $\times 10^{-3}$	0.0007 $\times 10^{-3}$
Class3 (100mm cut)	0.0033	0.0071 $\times 10^{-3}$	0.0020 $\times 10^{-3}$	0.0006 $\times 10^{-3}$
Class4 (200mm cut)	0.0184	0.5693 $\times 10^{-3}$	0.1998 $\times 10^{-3}$	0.0841 $\times 10^{-3}$
Class5 (200mm&55mm cut)	0.0304	0.8629 $\times 10^{-3}$	0.2959 $\times 10^{-3}$	0.1004 $\times 10^{-3}$
Class6 (200mm&150mm cut)	0.0004	0.0101 $\times 10^{-3}$	0.0040 $\times 10^{-3}$	0.0014 $\times 10^{-3}$
Class7 (200mm&square hole)	0.0683	0.0614 $\times 10^{-3}$	0.0025 $\times 10^{-3}$	0.0017 $\times 10^{-3}$

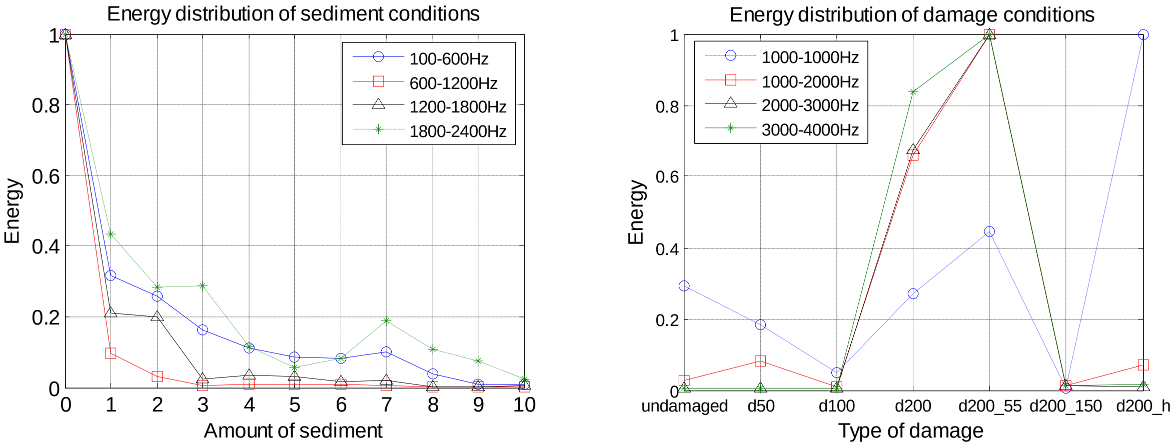


Figure 4.27 Energy distribution as a function of sediment conditions (left) and damage conditions (right).

4.4.2 Wavelet Sub-Band Energy and Entropy

The energy derived from DWT at $j-th$ level is called Sub-band energy, it is given by:

$$E_{j,k} = \sum_{n} \left| C_{x(t)}(j,k) \right|^{2} \tag{4.15}$$

where C is the DWT coefficient, n is the number of samples contained in one band, j,k are the scale and translation variables of the wavelet function, respectively. $j=5$ corresponds to the 5th level of decomposition.

The concept of the entropy been widely used as a measure of the disorder of a system, it can provide additional information about the underlying dynamical process associated with the signal [8]. Shannon entropy [9], which is the average unpredictability in a random variable calculated from DWT for each scale. This parameter is defined as:

$$W_{j} = -\sum_{k} \left| C_{x(t)}(j,k) \right|^{2} \log \left\lfloor \left| C_{x(t)}(j,k) \right| \right\rfloor^{2} \tag{4.16}$$

Then, the wavelet energy-entropy spectrum of $m-th$ decomposition scale of signal $x(t)$, E, can be written as :

$$\mathrm{E}_{m} = \{(W_{1},E_{1}),(W_{2},E_{2}),\ldots(W_{n},E_{n})\}_{m} \tag{4.17}$$

These features are useful to describe temporal information related properties for an accurate representation of a given signal. Table 4.4 displays the DWT sub-band energy and corresponding entropy of sediment conditions. The data were calculated at 5th level of wavelet decomposition using ‘db4’ as the wavelet function. Each set of coefficient represents one frequency range: A_5 (100~689) Hz; D_5 (689~1378) Hz; D_4 (1378~2067)Hz; D_3 (2067~2756)Hz. Time window $0 \le t \le 0.04$ sec was chosen for feature extraction.

Table 4.4 DWT Sub-Band Energy E and Entropy W of Sediment Conditions

Coefficient	A_5		D_5		D_4		D_3	
Condition	E_{A5}	$W_{A5}(\times 10^5)$	E_{D5}	$W_{D5}(\times 10^4)$	E_{D4}	$W_{D4}(\times 10^4)$	E_{D3}	$W_{D3}(\times 10^4)$
Class1 (clean)	114.1477	-2.9827	23.4282	-6.3720	2.3479	-1.5536	9.8314	-3.4295
Class2 (1bag)	41.2702	-2.2180	3.0056	-3.3407	0.6288	-0.8113	1.0286	-1.2762
Class3 (2bags)	35.3020	-2.3198	0.8767	-1.4651	1.4458	-0.9395	0.0547	-0.0657
Class4 (3bags)	23.7594	-1.6903	0.3418	-0.5946	0.0889	-0.1833	0.0330	-0.0269
Class5 (4bags)	15.8800	-1.6027	0.2709	-0.3983	0.1702	-0.2949	0.0607	-0.0977
Class6 (5bags)	12.5985	-1.4567	0.2659	-0.3677	0.2467	-0.4477	0.0779	-0.1164
Class7 (6bags)	12.4730	-1.3337	0.3399	-0.4800	0.1013	-0.1945	0.0378	-0.0366
Class8 (7bags)	15.0681	-1.4051	0.2873	-0.4193	0.0859	-0.1620	0.0246	-0.0125
Class9 (8bags)	6.2219	-0.6192	0.0991	-0.0463	0.0170	-0.0185	0.0130	-0.0007
Class10 (9bags)	1.8587	-0.2833	0.0799	-0.0163	0.0086	-0.0080	0.0123	-0.0001
Class11 (10bags)	1.0030	-0.1814	0.1242	-0.1103	0.0137	-0.0084	0.0174	-0.0002

Table 4.5 DWT Sub-Band Energy E and Entropy W of Damage Conditions

Coefficient	A_5		D_5		D_4		D_3	
Condition	E_{A5}	$W_{A5}(\times 10^5)$	E_{D5}	$W_{D5}(\times 10^5)$	E_{D4}	$W_{D4}(\times 10^5)$	E_{D3}	$W_{D3}(\times 10^5)$
Class1 (Undamaged)	7.2627	-0.8847	69.2964	-1.2671	141.9002	-1.1580	47.6275	-0.8309
Class2 (50mm cut)	27.4188	-2.1608	185.7676	-1.7892	107.7842	-0.9690	98.2164	-1.2466
Class3 (100mm cut)	50.6451	-2.6198	218.5889	-2.1350	120.8663	-1.0416	88.5572	-1.2586
Class4 (200mm cut)	16.6938	-1.4367	9.9978	-0.6185	39.3166	-0.6446	5.5845	-0.4261
Class5 (200&55cut)	28.4780	-1.7510	31.1480	-0.9738	72.1236	-0.7335	13.4458	-0.6910
Class6 (200&150cut)	17.5681	-1.3619	15.8280	-0.7166	78.1741	-0.7308	7.2371	-0.4740
Class7 (200&hole)	45.3661	-2.2533	63.7955	-1.1092	79.1571	-0.7904	27.7935	-0.6900

Table 4.5 illustrates the wavelet sub-band energy and entropy of damage conditions at the 5th decomposition level. Time window $0 \le t \le 0.04$ sec was used to calculate these data.

Figure 4.28 shows the wavelet sub-band energy of 4 DWT coefficients plotted as a function of the corresponding entropy. The energy and entropy value were in log scale in the sediment conditions figure to enhance the difference between wavelet coefficients. Shannon entropy was employed to discrete wavelet coefficients generated by DWT where larger entropy values represent higher process uncertainty and therefore higher complexity.

As shown in Figure 4.28 that the distribution of energy-entropy data obtained from four DWT coefficients followed similar patterns which can reflect the variation of the signal energy in different frequency bands. It also suggests that all four DWT coefficients are useful for condition classification, a classification system can be trained by these datasets to classify siphon conditions with suitable classification algorithms.

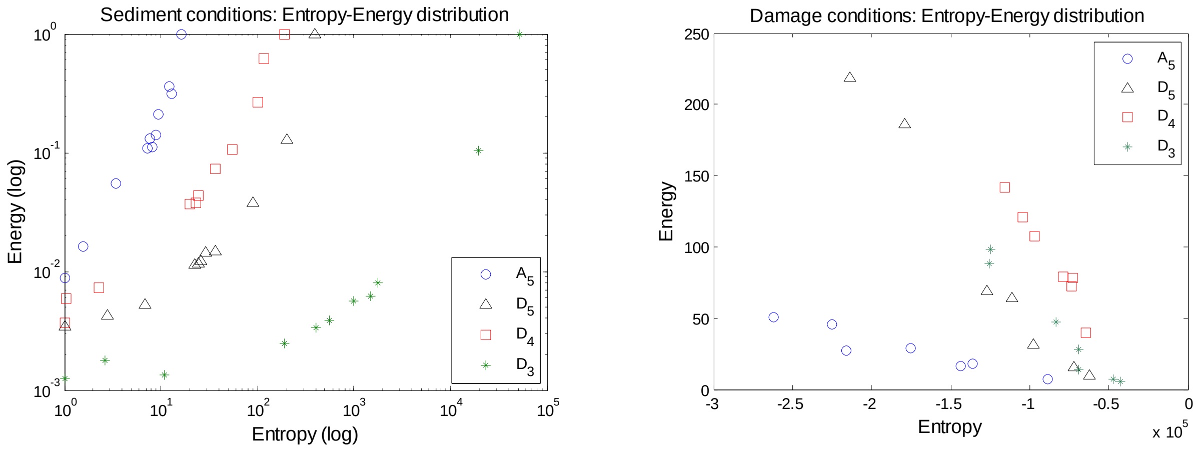


Figure 4.28 Wavelet Entropy-Energy distribution $\vec{E}$ of sediment conditions (left in log scale) and damage conditions (right) at 5th decomposition level.

4.5 Condition Classification and Recognition

Pattern classification is the process after feature extraction in a pattern recognition system. "*Each pattern is a three-part rule, which expresses a relation between a certain context, a problem, and a solution" is the description of a pattern given by* Christopher Alexander [10]. Pattern classification is an organization process of mapping patterns into groups where patterns sharing the same set of properties. The definition of these properties is not fixed and may include criteria such as structure, intent or applicable. Depending on the chosen criteria, a classification scheme can be defined.

Classification can be divided into two categories based on the type of problems: supervised classification (labelled training samples) and unsupervised classification (unlabelled training samples). Supervised classification requires the upfront knowledge of the data, predefined classes and the algorithm to be used before the training process. In unsupervised classification, samples are given as unlabelled, the input feature is assigned to an unknown class using some clustering algorithms. Unsupervised classification is more computer-automated based on the nature of the data, while supervised classification is more closely controlled by the users to specify parameters with a priori information.

A recognition system is operated in two modes: training (machine learning) and classification (testing) (see Figure 4.29). In the training mode, the feature extraction/selection module finds the appropriate features for representing the input data and the classifier is trained to partition the feature space. In the classification mode, the trained classifier assigns the test sample to one of the pattern classes under consideration based on the measured features.

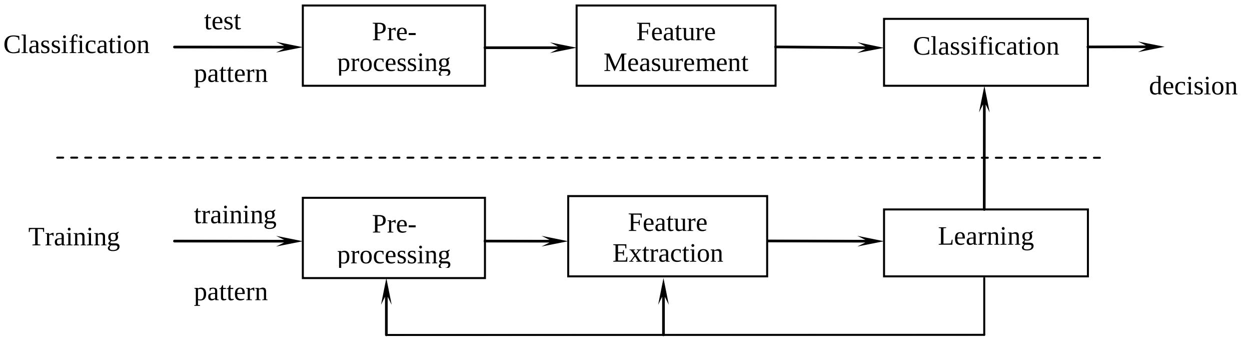


Figure 4.29 Model for statistical pattern recognition

The choice of the classifier is a difficult problem and in practice, it is often based on which classifier happens to be available or best known to the users. The simplest and the most intuitive approach to classifier design is based on the concept of similarity: patterns that are similar in some ways should be assigned to the same class. The second main concept used for designing pattern classifiers is based on the probabilistic approach. The optimal Bayes decision rule assigns a pattern to the class with the maximum posterior probability [11]. One of the well-known nonparametric decision rules is the k-nearest neighbours rule (k-NN), it is closely related to the problem of density estimation in statistics. The similarity between patterns in k-NN classification is the distance between the test sample and training samples, usually the Euclidean distance is used as the similarity measure:

$$d_p(x,x') = \|x-x'\|_p = (\sum_{i=1}^{N} |x_i - x_i'|^p)^{1/p} \qquad (4.18)$$

where x is the test sample, x' is the training sample, x and x' are points in the parameter space $\mathbf{X}=\square^{N}$, p is the dimension of the space.

After the feature extraction and selection procedures through Butterworth filtering and DWT, two sets of energy related data were obtained from raw acoustic data recorded on each hydrophone (H1, H2 and H3) for a range of siphon conditions. Data extracted from hydrophone 1, 2 and 3 were used to train a classifier and new unknown data extracted were used to test the system. When the number of nearest neighbour $k=1$, KNN is also called nearest neighbour which simply assigns the test sample to the class of its nearest neighbour. It is the simplest of all algorithms but very effective because the training error is zero (no overfitting).

The decision rule of the nearest neighbour classifier for a single observation x is:

$$i=\arg\min_{k}\left\{\min_{n=1,\dots N}\left\|x-x'_{n}\right\|^{2}\right\} \tag{4.19}$$

where N is the number of training samples in class i and x'_{n} is the n-th observation from this class. A common extension to the nearest neighbour approach is to use features in more dimensions in a majority vote scheme. The classification is decided by majority vote which assigns the test sample to the class which is the most common amongst its nearest neighbours in the feature spaces.

In general, a majority vote rule is defined basing on probability estimation:

$$C(X)=\arg\max_{i}\sum_{j}\omega_{j}\hat{p}_{ij}\ , \tag{4.20}$$

$$\hat{p}_i = \Pr(X = i \mid N) = \frac{N_i}{N} \tag{4.21}$$

where $\hat{p}_{ij}$ is the probability estimated from the j-th classification rule for the i-th class, ω_j is the optional weights. N_i is the number of the condition estimation X chosen to be class i and N is the total number of all estimations.

4.5.1 Classification Using Features from Digital Filters

A training data matrix was constructed using energy data derived from a range of known siphon conditions: $\mathbf{E} = \{E_{mnl}\}$, each element in this matrix is the value of the acoustic energy determined by the siphon condition (m); frequency band (n) and hydrophone channel (l). In sediment conditions, the training matrix is in the form $11 \times 3 \times 3$ ($m \times n \times l$); in damage conditions, the size of the training matrix is $7 \times 3 \times 3$. 5 sets of data recoded in the siphon with different amount of sediment inside and 5 sets of data recorded in the siphon with some damage on the wall were used for testing the classification system. Each testing dataset is in the form $1 \times 3 \times 3$ containing features extracted from data recorded on 3 hydrophones and filtered in 3 frequency bands for one testing siphon condition.

The Euclidean distance was in fact the absolute distance in a 1-parameter space ($p = 1$in (4.14)) in this part of the analysis, and the distance matrix obtained from one set of testing data and the training datasets is: $\mathbf{D}(i) = \{d\}_{mnl}$, for every $d \in \mathbf{D}$, its class label $i \in \mathfrak{R}^m$. Then the problem of condition recognition can be reduced to finding the minimum $\Re$ of $\mathbf{D}$ with the respect to the condition class label i:

$$\Re = \left\{ i : \min_{i} \mathbf{D}(i) \right\}_{n \times l} \tag{4.22}$$

$\Re$ is the condition label matrix of size $N \times L$, where N is the number of frequency bands and L is the number of hydrophone channels through which the data were collected.

Majority vote was applied to $\Re$ of each testing dataset to select the class number which appeared the most appropriate and assign the testing data to that class. Four outcomes of the majority vote analysis were achieved: (i) the correct result, whereby the majority vote identified the correct condition in the siphon; (ii) a false result, whereby the majority vote identified a wrong condition in the siphon; (iii) an ambiguous result, when no clear decision could be drawn from majority vote because the correct condition number appeared as frequent as another condition number; (iv) a failure to make a decision as all condition numbers appeared equal times.

4.5.2 Classification Using Wavelet Coefficients

Discrete wavelet coefficients were obtained through discrete wavelet transform, sub-band energy and entropy were calculated from the coefficients and the wavelet sub-band energy feature was defined in Equation (4.13). The Euclidean distance between wavelet features derived from training data and testing data is calculated as:

$$d(x, x') = \sqrt{(e - e')^2 + (w - w')^2} \tag{4.23}$$

where e and w are the sub-band energy and entropy of test sample x; e' and w' are the sub-band energy and entropy of training sample x'. The distance matrix was constructed in the same manner as pervious the only difference is the number of frequency band, 4 bands were used in wavelet analysis, the size of training matrix of sediment and damage conditions are: $11\times4\times3$ and $7\times4\times3$, respectively. The size of test data is $1\times4\times3$ corresponding to 4 frequency bands and 3 hydrophones from one unknown siphon condition.

Figure 4.30 is an example of energy features of sediment conditions obtained from two extraction process at the lowest frequency band. It shows how the Euclidean distance between test sample and training samples would be calculated and the nearest neighbour of the test sample should be chosen.

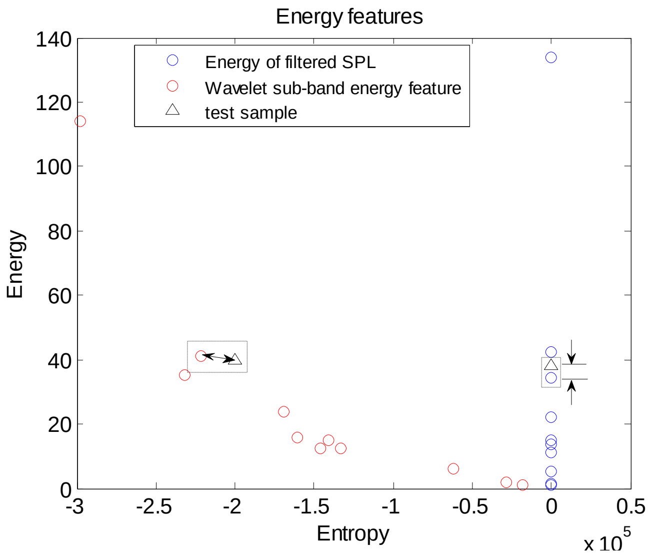


Figure 4.30 Energy features obtained through digital filter and Wavelet sub-band energy features at the lowest frequency band from a range of sediment conditions

Table 4.6 presents the probability of the correct condition classification of 5 sets of test data from sediment and damage conditions using features obtained through digital filter and discrete wavelet transform. In the case of the sediment tests, the probabilities of the correct estimation from using features pre-processed by the digital filters are generally below 60% except the clean siphon condition. It appears to be difficult to classify the sediment condition by the exact number of sandbags inside the siphon, the energy did not change significantly to the change when one sandbag was added or removed. However, the acoustic energy had noticeable decrease when the amount of sediment reached certain level, in this research four conditions were clearly detectable with 100% recognition accuracy rate: (i) 0 (clean); (ii) 1~2sandbags; (iii) 3~7 sandbags; (iv) 8~10 sandbags. Classifying testing sediment condition to above four classes instead of specifying the exact number of sandbags would be more sensible. The results of damage condition estimation are less ambiguous than the results of sediment conditions. Although the energy was not necessarily decreasing with the increase of the size of the wall damage, the difference between different types of damage (as illustrated in Table 4.5) were distinguishable in energy distribution.

The wavelet energy was derived from the DWT coefficients with each set of coefficient corresponding to a frequency range. The wavelet entropy was calculated from these DWT coefficients too, the energy and entropy showed similar pattern to the change of the siphon conditions. The entropy was added to make the energy features two-dimensional, the probabilities of the correct classificationsoftesting samples are given in Table 4.6 and 4.7. Table 4.8 and 4.9 present the classification results of testing samples using Nearest Neighbour classifier. The class label appeared the most frequently resulted in the highest condition estimation probability and therefore will be determined as the class label for the testing sample. The probabilities of each estimation for all testing samples are shown in Figure 4.31 and 4.32.

Table 4.6

Probabilities of correct estimation of siphon sediment condition

Test data index	Feature extractor	
	Filter	DWT
Test bk_1	0.89	1.0
Test bk_2	0.44	0.5
Test bk_3	0.33	0.42
Test bk_4	0.44	0.58
Test bk_5	0.56	0.67

Table 4.7

Probabilities of correct estimation of siphondamage condition

Test data index	Feature extractor	
	Filter	DWT
Test dm_1	0.78	0.92
Test dm_2	0.67	0.67
Test dm_3	0.89	0.83
Test dm_4	0.67	0.75
Test dm_5	0.78	0.83

Table 4.8Numbers of sediment condition class labels appeared among nearest neighbours of two sets of features

	Blockage	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
Filter Features (3x3)	Test 1	8	1	-	-	-	-	-	-	-	-	-
	Test 2	-	-	4	3	2	-	-	-	-	-	-
	Test 3	-	-	-	-	3	2	-	-	4	-	-
	Test 4	-	-	-	-	-	4	4	1	-	-	-
	Test 5	-	-	-	-	-	-	-	-	5	2	2
Wavelet Features (4x3)	Test 1	12	-	-	-	-	-	-	-	-	-	-
	Test 2	-	5	6	1	-	-	-	-	-	-	-
	Test 3	-	-	-	-	5	1	1	5	-	-	-
	Test 4	-	-	-	-	-	5	7	-	-	-	-
	Test 5	-	-	-	-	-	-	-	-	8	2	2

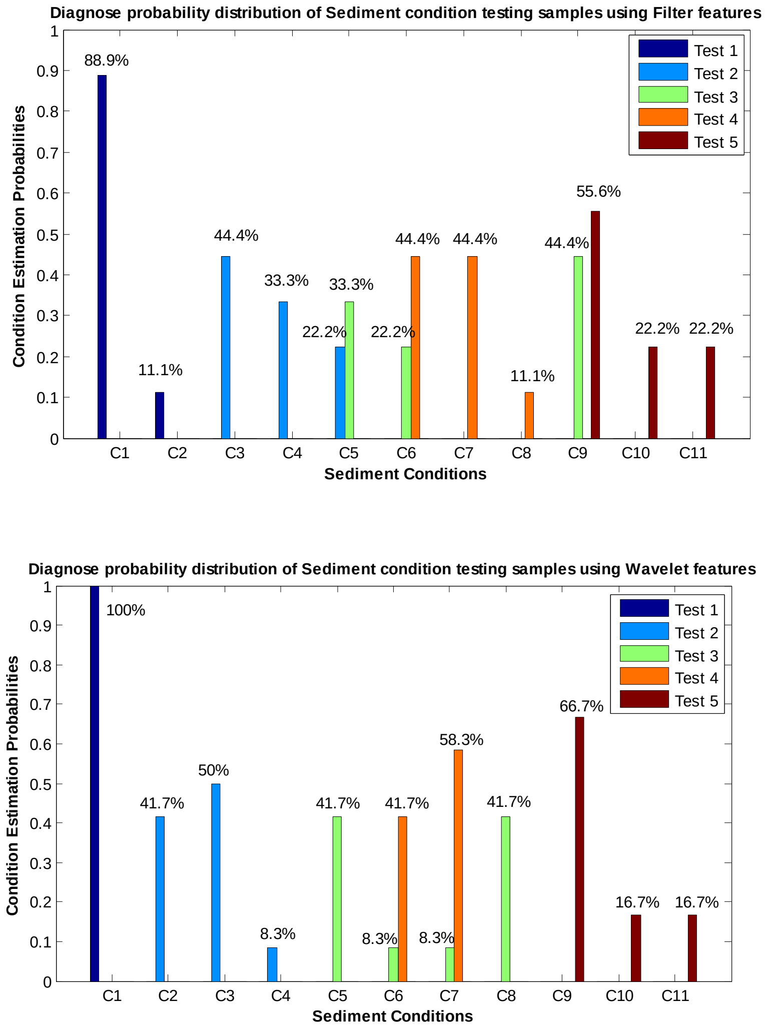


Figure 4.31Total probabilities of sediment condition estimation using 2 sets of features: filter features (above) and wavelet features (bottom).

Table 4.9 Numbers of damage condition class labels appeared among nearest neighbours of two sets of features

	Damage	C1	C2	C3	C4	C5	C6	C7
Filter Features (3x3)	Test 1	7	-	-	2	-	-	-
	Test 2	1	6	-	2	-	-	-
	Test 3	-	1	-	8	-	-	-
	Test 4	-	1	2	-	-	6	-
	Test 5	-	-	-	-	2	-	7
Wavelet Figures (4x3)	Test 1	11	-	-	-	-	-	1
	Test 2	-	8	3	-	1	-	-
	Test 3	-	-	-	10	1	1	-
	Test 4	-	-	-	2	1	9	-
	Test 5	-	-	1	-	1	-	10

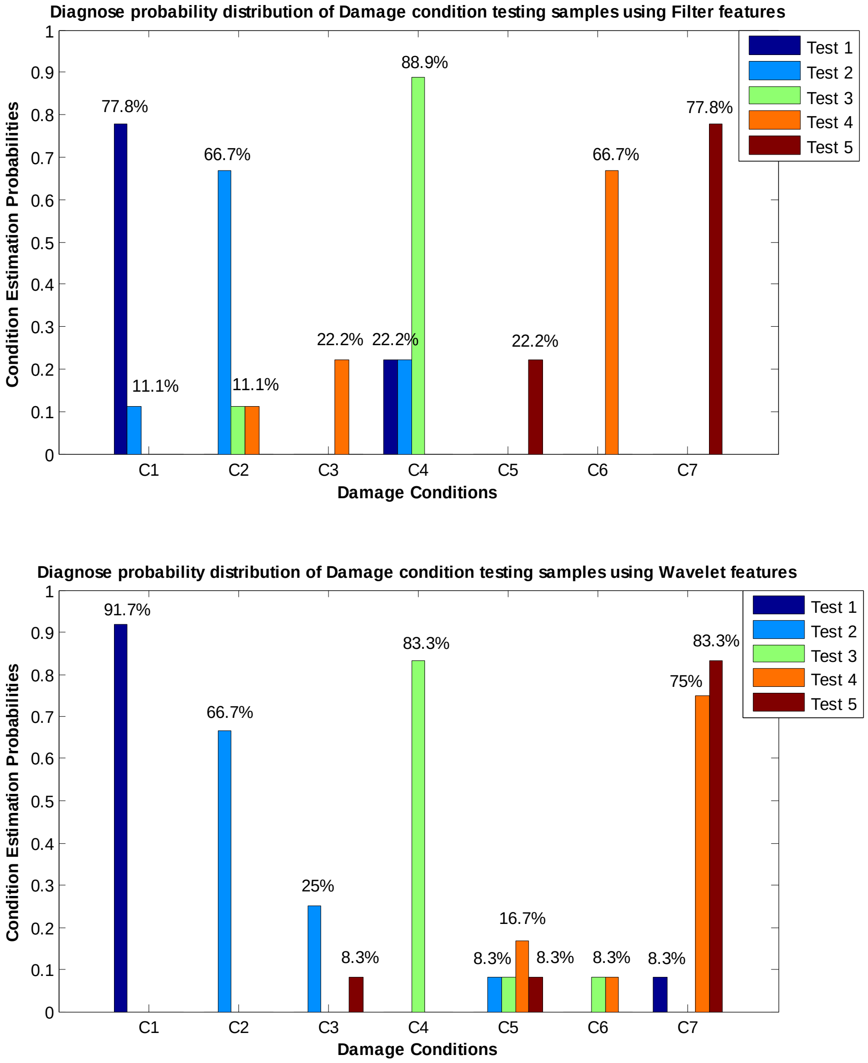


Figure 4.32Total probabilities of damage condition estimation using 2 sets of features: filter features (above) and wavelet features (bottom).

4.6 Summary

Acoustic signals were collected from a range of typical blockage and damage siphon conditions which were recreated in the laboratory. Original broadband signals recorded on 3 hydrophones were filtered by using Butterworth digital filter and discrete Daubechies wavelet transform in several narrow frequency bands. The sound pressure level was calculated so that the acoustic energy and wavelet energy-entropy feature could be determined as a function of time and used as features in the classification analysis.

Nearest neighbour (NN) was used as the classification method to identify the siphon conditions. The acoustic based nearest neighbour classification system is proved to be capable of discriminating different siphon conditions. For sediment conditions, acoustic data in the lower frequency bands contain more useful information than those filtered through the higher frequency bands. It is challenging to classify the siphon condition by the exact amount of sediment, however, the size of the sediment can be limited in a certain range and be classified. Acoustic energy was more sensitive to the change of damages than sediment, damage condition classification results showed higher certainty than sediment conditions. The accuracy of sediment classification was improved by 20% using wavelet features than filter features. Damage classification results are 100% correct for all 5 testing samples using both wavelet and filter features, however, the estimation probabilities leading to the classification decision have shown that wavelet features generally produced higher probabilities for the correct classification than filter features.

Discrete wavelet transform coefficients carry useful information about the transient activity of the acoustic signals. Therefore, DWT coefficients were used to extract energy and entropy as features for condition analysis. Results showed that DWT improved the classification accuracy of the system in general. It can be concluded that the K-NN classification system showed promising performance in condition classification using acoustic energy as features to discriminate a range of sediment and damage siphon conditions. A few factors could affect the classification results are: (i) the time window chosen to calculate sound pressure level; (ii) the frequency bands used to derive energy features; (iii) wavelet function, if use discrete wavelet transform as filterbank.

References

[1]S.B. Costello et al. "Underground asset location and condition assessment technologies". *Tunnelling and Underground Space Technology* , vol 22, pp 524-542. 2007.

[2] Zheng Liu et al. "State of the art review of inspection technologies for condition assessment of water pipes". *Measurement* , vol 46, issue 1, pp 1-15. 2013

[3] Martin Holters et al. "Impulse response measurement techniques and their applicablility in the real world". *Proc of the 12th Digital Audio Effects.* Italy. 2009

[4] M. R. Schroeder, "New method of measuring reverberation time". *Acoustic Society of America* , vol 37, pp 409-412. 1965.

[5] William D. Stanley, *"Digital Signal Processing",* Prentice-Hall Inc. New Jersey, USA.1984.

[6] Gilbert Strang and Truong Nguyen, *"Wavelets and Filter Banks".* Wellesley-Cambridge Press, MA USA. 1996.

[7]I. Daubechies , "Orthogonal bases of compactly supported wavelets". *Communications on Pure, and Applied Mathematics* , vol 41, pp 909-996. 1988.

[8] Robert M. Gray , *"Entropy and Inforamtion Theory".* Springer-Verlag, New York. 1990.

[9]C. E. Shannon, "A mathematical theory of communication". *Bell Systems Technical Journal* , vol 27:pp 379–423 and pp 623–656. 1948.

[10] Christopher Alexander et al. *"A Pattern Language:Towns, Buildings, Construction",* Oxford University Press, Berkeley California. 1977.

[11] I. Rish, "An empirical study of the naive Bayes classifier". *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*. 2001.

Chapter 5

Condition Classification of Laboratory Constructed Partly-Filled Pipe

5.1 Introduction

This chapter presents the results on the application of the pattern recognition methods to determine the conditions in a partially full sewer pipe. For this purpose a special, full scale pipe facility was constructed in the Hydraulic laboratory at the University of Bradford. It was designed to provide as a flexible experimental facility in which a representative range of structural and operational conditions such as lateral connection and sediments could be reproduced in a controlled experiment in the presence and absence of flow [1]. This type of pipe was chosen because it is representative of a large network of small combined sewers found in the UK. This facility was used to study the capabilities of the acoustic equipment for data acquisition that was developed and tested by the acoustic group at the University of Bradford.

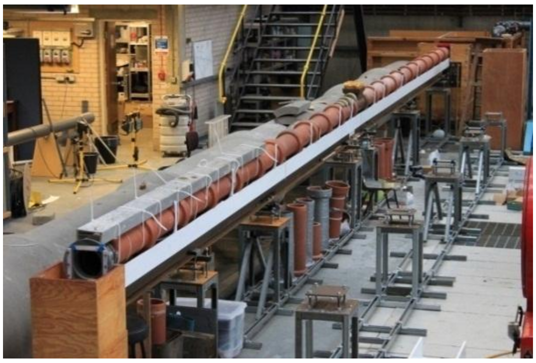
The focus of this study was on the effects of water level and multiple defects on the performance of the condition classification methods. In each of these experiments, 3 categories of pipe condition were studied: (1) pipe end; (2) lateral connection; (3) sediment. The acoustic signals reflected from the conditions simulated in the pipe were recorded and processed using the classification algorithms proposed in Chapter 3.

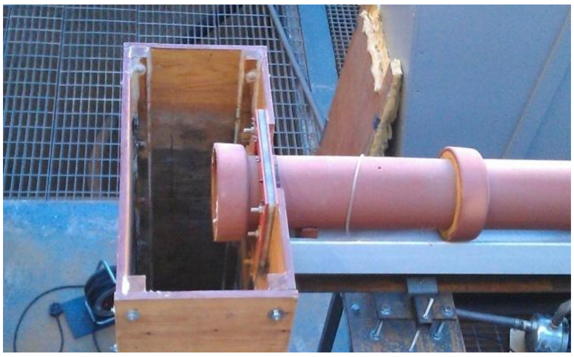
Especially two data fitting algorithms were employed: polynomial and Padé approximations were used as feature extractors whereas Support Vector Machines (SVMs) were used as main is the main classification method.

5.2 Data Acquisition and Pre-Processing

5.2.1 Experimental Set-up

The experimental facility used in this work was a 150mm diameter, 14.4 meters long clay pipe, each of the two ends of this pipe was connected to a rectangular tank which was capable of holding water to allow different level of flow to be set as shown in Figure 5.1. A lateral connection was installed in the middle of the pipe through which different types of blockages (see Figure 5.1) could be deposited at different locations in the pipe to study the capability of the classification algorithms to discriminate between multiple defects and for a range of water levels.





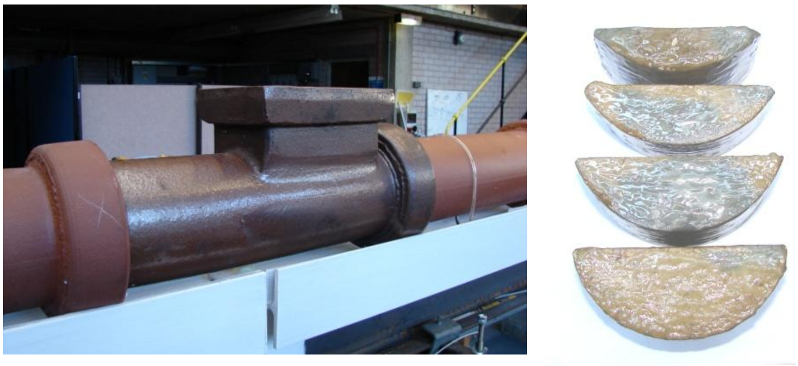


Figure 5.1 The 150mm clay pipe, lateral connection, pipe end and blockages used for sediment simulation

The acoustic system which was used for data collection in this research consisted of four microphones arranged in line on a slim PCB board separated non-equidistantly to optimize the accuracy of the sound intensity measurement, as shown in Figure 5.2. The distance between each pair of these microphones, Δ, were chosen to be less than the wavelength , λ, to allow for the sound pressure gradient measurements which was then used to calculate the sound intensity.

The sensor was inserted through one end of the pipe and attached to the interior wall. The other end remained either open or closed.

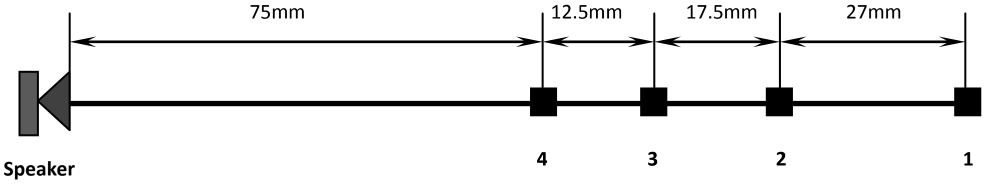


Figure 5.2 Acoustic sensor made of four microphones and one speaker

5.2.2 Acoustic Intensity Response

A sinusoidal sweep within the frequency range from 50 to 7000Hz was used as excitation signal to measure the impulse response of the pipe. This type of signal is considered more suitable for outdoor measurements than periodic pulse and Maximum Length Sequence (MLS) when high SNR is required[2].

The acoustic impulse or frequency response of a physical system is a very useful quantity. This quantity contains detailed information on the system geometry, sound speed and the conditions at its boundaries. Any change in these properties is reflected in a change in the acoustic impulse or the associated frequency response. The impulse response can be obtained by deconvolving the output of the system, the convolution is given by: $y(t) = \int_t h(\tau)x(t-\tau)d\tau$, where $h(\tau)$ is the system impulse response. The acoustic pressure impulse response data could be used to calculate the instantaneous acoustic intensity. The acoustic intensity is equal to the product of the acoustic pressure (a scalar) and particle velocity (a vector). Hence it is a vector quantity, possessing both a magnitude and direction. The relationship of acoustic pressure and intensity is given by following equation:

$$\tilde{\mathbf{I}}(t) = p(t)\mathbf{u}(t) \approx \mathbf{n}\frac{p_1(t)+p_2(t)}{2\Delta\rho_0}\int[p_1(\tau)-p_2(\tau)]d\tau \quad (5.1)$$

where $p_1(t)$ and $p_2(t)$ are acoustic pressure data recorded on a pair of microphones spaced by a distance Δ□ wavelengthλ , $\mathbf{u}(t) = -\frac{1}{\rho_0}\int\frac{\partial p}{\partial \mathbf{n}}d\tau \approx \frac{1}{\Delta\rho_0}\int[p_1(\tau)-p_2(\tau)]d\tau$ is the acoustic (particle) velocity vector in the direction of the normal $\mathbf{n}$ that coincides with the direction of sound wave propagation. Here assume $\rho_0 = 3.43\times10^{-3}\frac{P_a}{T}$ is the density of air, T is the temperature and P_a is the atmosphere pressure.

Unlike the problem of siphon condition classification presented in Chapter 4 which relied on acoustic pressure data for classification, the sewer pipe condition classification was reliant on acoustic intensity data. The acoustic field in the fully filled is inherently reverberant combination of sound waves in the fluid and solid phase which are strongly coupled. As a result, there are multiple reflections and scattering from individual defects that cannot be easily separated. Therefore, the acoustic field in the siphon is strongly diffused so that the use of the acoustic intensity as a vector field does not have any advantages over the acoustic pressure field measurements. On the other hand, the acoustic intensity field in the partially full sewer pipe is a vector quantity for which the direction can be clearly established. This information enables us to develop a system to recognize a plurality of defects based on the acoustic intensity signatures which can be extracted separately for separate defects.

5.2.3 Water Level Effect Experiment and Data Pre-Processing

To understand how much different level of water inside the pipe could affect the acoustic reflected signals and the accuracy of further condition analysis, a set of measurements were implemented with water flowing through the pipe. For this purpose the water level was varied from 0 to 20mm depth to simulate the dry/flow conditions typical for the real underground live sewers.

The following conditions were simulated in these experiments:

(1) clean empty pipe;

(2) empty pipe with an open lateral connection;

(3) a 55mm blockage placed inside the pipe at two locations;

(4) a 55mm blockage placed inside pipe with the effect of an open lateral connection.

For each above condition, 20 sets of data were collected for the water level depth inside the pipe varying from 0 to 20mm with the highest water level corresponding to approximately 10% of the pipe cross-section. The purpose of this experiment was to determine the capability of the pattern recognition system of recognizing different pipe defects under various water levels, so that the performance of the proposed classification algorithms can be evaluated and improved.

The acoustic pressure signals were recorded on the 4-microphone array at the sampling frequency of 15 kHz. The analysis of these signals consisted of deconvolution which was used to obtain the acoustic pressure impulse response containing information on the pipe geometry and operational conditions in the pipe. Figure 5.3 gives examples of raw acoustic pressure data recorded in the empty pipe (left) and the pressure impulse response obtained from its deconvolution (right).

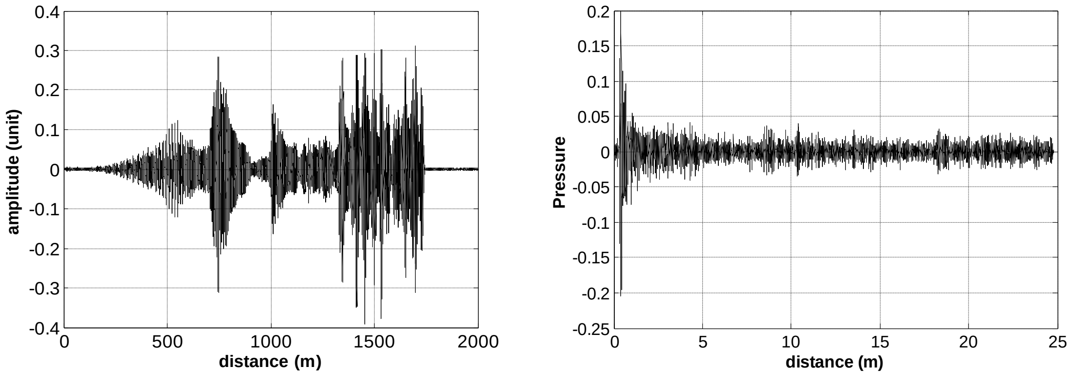
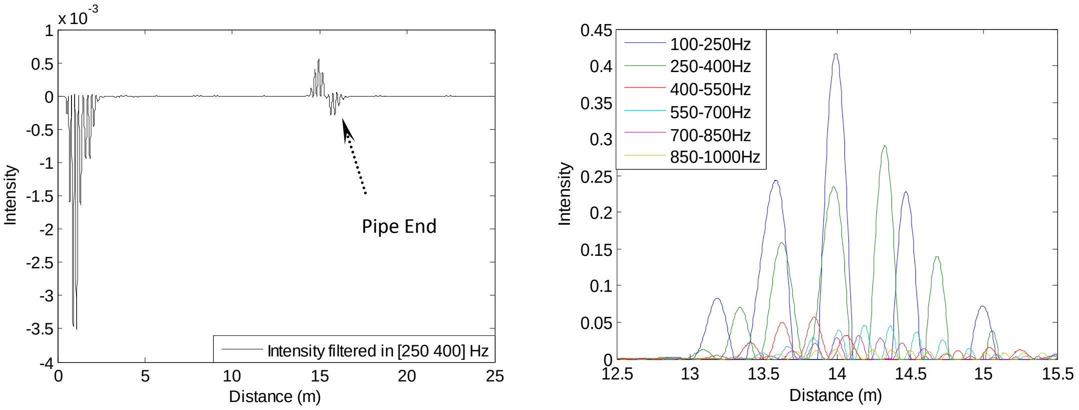


Figure 5.3 Acoustic recorded data and pressure impulse response of clean empty 150mm clay pipe

A Butterworth filter of order 3 was then used to filter the broadband acoustic signals into several narrow bands with a 150Hz bandwidth. The reason for the choice of this filter is given in Ref [1]. The reflections from individual conditions in the pipe were separated in time domain and used in the condition classification process.



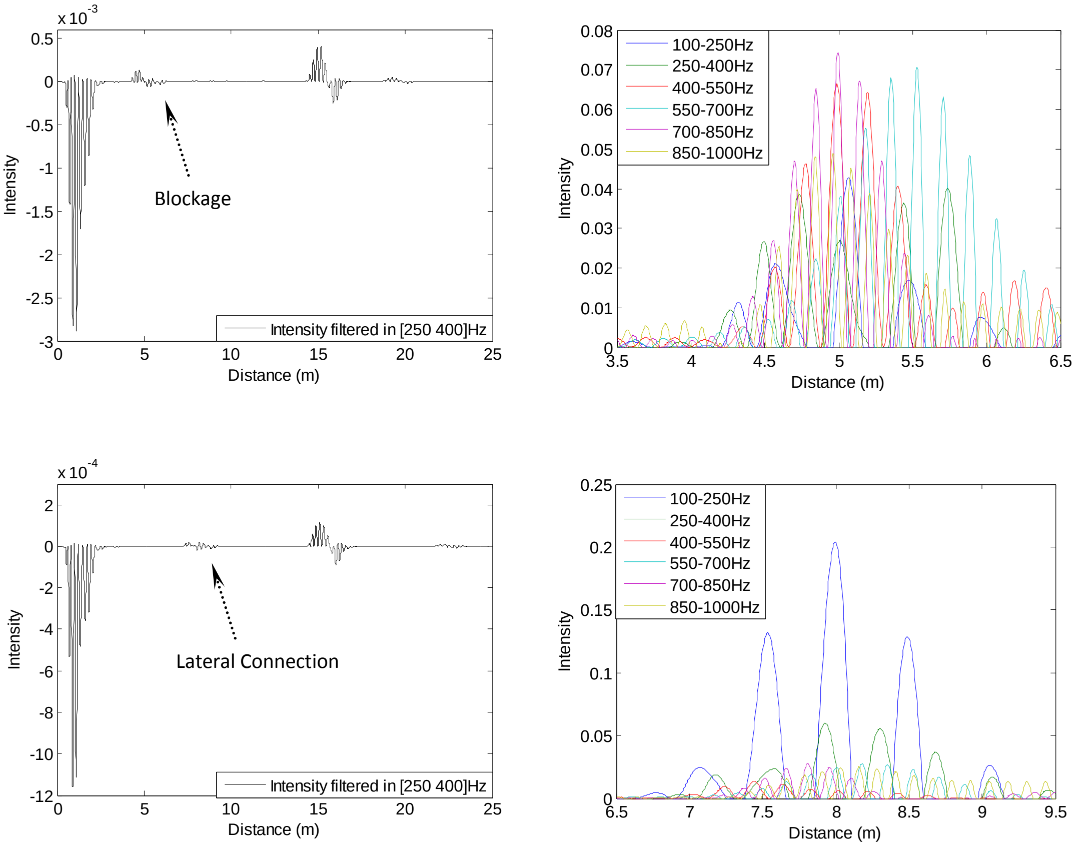


Figure 5.4 Intensity responses (left) and acoustic signatures (right) of defects

Figure 5.4 shows the intensity responses/reflections (left column) in the frequency range 250 to 400 Hz of three conditions from top down: empty pipe; 55mm blockage placed at 5 meters from the sensor and an empty pipe with an open lateral connection at 8 meters from the sensor. These reflections are plotted as a function of the distance in the pipe which was calculated as $d = t \times c$, where t is the time and c is the sound speed in the pipe. It can be seen that part of the intensity spectrum contains clear data reflected from a particular defect in the pipe which can be extracted as signature of the defect. Signatures collected from a range of conditions of each defect can be stored as a database and used in the training process. Figure 5.4 gives examples of acoustic signatures containing 600 sample points (right column) extracted from the left intensity data in frequency range from 100 to 1000 Hz and filtered in 6 frequency bands with a 150Hz bandwidth. A library of intensity signatures extracted from data collected for 4 conditions listed above is built for the next step of condition classification.

Table 5.1 Summary of conditions and extracted signatures of water level effect experiments

Condition	**Signature extracted**		**Number of signatures**
Empty clean pipe	Pipe end		20
Empty pipe with an open Lateral connection	Lateral connection pipe end		20 20
Blockage placed 5m from the source	Blockage pipe end		20 20
Blockage placed 3m before open lateral connection	Blockage lateral connection pipe end		20 20 20
Signature type	Pipe End (PE)	Blockage (BK)	Lateral Connection (LC)
Total amount	80	40	40

Table 5.1 gives a summary of all the conditions that were simulated in these experiments and number of signatures extracted for each of these conditions. The details of the database development, data analysis and classification are given in the following sections.

5.2.4 Experiments with Multiple Defects

The effect on condition classification in the presence of multiple defects was an important part of this research. The pipe conditions in this experiment were designed to study the interaction between the reflections from multiple defects and its effect on the classification algorithm. The conditions and defect signatures used in this experiment are summarized in Table 5.2.

Table 5.2 Summary of conditions and extracted signatures of multi-defect effect experiments

Condition	Signature extracted	No. of signatures	
Empty clean pipe	Pipe end	5	
Empty pipe with an open lateral connection	Pipe end lateral connection	5 5	
40mm blockage placed 4M away from the source	Pipe end Blockage	5 5	
40mm blockage placed 11M away from the source	Pipe end blockage	5 5	
55mm blockage place 4M away from the source	Pipe end blockage	5 5	
55mm blockage placed 11M away from the source	Pipe end blockage	5 5	
40mm blockage placed 3M before open lateral connection	pipe end blockage Lateral connection	5 5 5	
40mm blockage placed 3M after open lateral connection	pipe end blockage Lateral connection	5 5 5	
55mm blockage placed 3M before open lateral connection	pipe end blockage Lateral connection	5 5 5	
55mm blockage placed 3M after open lateral connection	pipe end blockage Lateral connection	5 5 5	
55mm and 40mm blockages placed 4M and 11M away from the source, respectively	Pipe end blockage	5 10	
55mm blockage and 40mm blockage placed 3M before and after the open lateral connection, respectively	Pipe end Blockage Lateral connection	5 10 5	
Signature type	PE	BK	LC
Total amount	60	60	30

For each above condition, 5 sets of measurement were taken. Signatures of defects from each condition were extracted in the same way as stated in section 5.2.3. Some defects were used for training and others were used for proposed classification algorithm.

5.3 Feature Extraction and Selection

The acoustic energy reflected for a defect in the pipe was used and proved to be able to provide enough information for the system to determine the siphon condition as presented in Chapter 4. Since the reflected intensity signatures contained a set of clear data of defect, the reflected energy of each intensity signature in a range of frequency bands is calculated:

$$\bar{I}(t) = (I(t) + |I(t)|)/2 \qquad (5.2)$$

$$e^{+}(t) = \sum_{t=t_1}^{t_2} \bar{I}(t)$$

$$E(f_i) = [e_1^{+}(t), e_2^{+}(t), \cdots e_i^{+}(t)] \qquad (5.3)$$

Where $\bar{I}(t)$ is the positive (reflected) part of the intensity response function using for the pipe condition characterisation, $E(f_i)$ gives the energy spectrum of the acoustic intensity signal reflected from a defect in the pipe in 20 frequency bands below the 1st cut-off frequency of the pipe. In the case of a cylindrical pipe, the 1st cut-off frequency is $f_{\max} \cong \frac{0.5861c_0}{2a}$, where a is the pipe radius and c_0 is the sound speed in air. The acoustic pressure patterns created by the modal acoustic field in a pipe is rather complex [1]. Therefore, this research is focused on the propagation of the fundamental mode in a round pipe. For this purpose we will ensure that the frequencies of sound in our study do not exceed the cut-off frequency.

The reason of choosing 20 frequency bands with relatively bid bandwidth is to capture the energy spectrum variation in the acoustic wave which corresponds to the plane wave in the pipe. Figure 5.5 are the energy-frequency plots of 3 types of defects which were generated using the data obtained from 2 sets of measurement of each condition from water level (Figure 5.5 top figure) and multiple defect (Figure 5.5 bottom figure) experiments. It is clear that there are consistent patterns for each of the three defect types. These are relatively independent from the water level present in the pipe and relatively unaffected by the presence in the pipe some other types of multiple defects. Clearly, the pattern in the acoustic energy frequency spectrum is somewhat unique to a particular defect type and can be used to distinguish between different types of defect under various conditions in the sewer pipe. For this purpose, a suitable data fitting algorithm is required. This topic is discussed in the following sections of this chapter.

The energy spectrum of all defects were normalized by subtracting the mean value of each class to remove the baseline so that the derived pattern of each defect can be fairly compared.

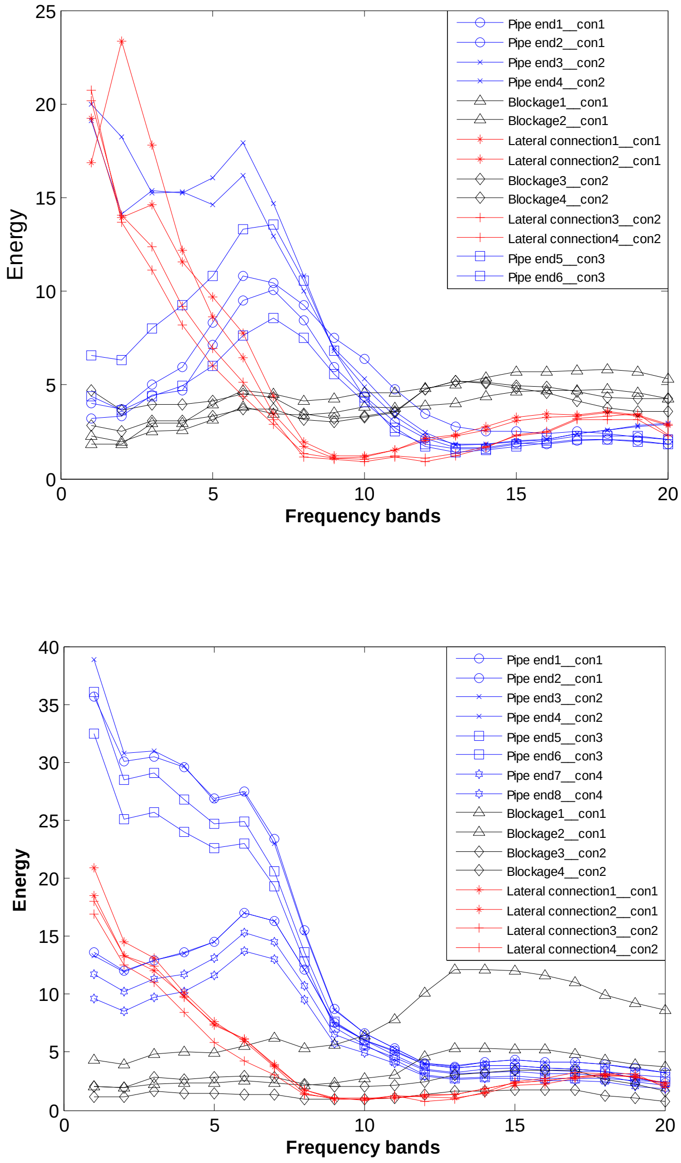
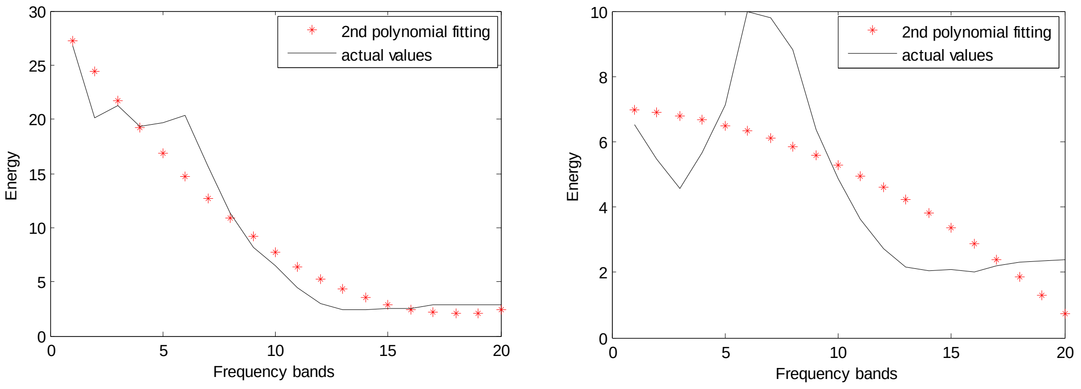


Figure 5.5 Energy-frequency band plots of water level and multiple defects experiments

5.3.1 Least Squares Polynomial Fitting

Polynomial fitting as a spectral approximation algorithm was introduced in Chapter 3. The least-squares error estimate is a criterion that can be used to measure the goodness of a fitting by calculating the square of the separation (residue) between a series of data and its approximation over a given interval. The polynomial order n is the parameter that affects the performance of the fitting. The polynomial coefficients are the classification features to provide spectral pattern information in our case and to distinguish patterns from one to the others.

Figure 5.6 gives 2 sets of polynomial fitting examples: left column is the reflected energy from the pipe end of an empty clean pipe and its polynomial approximations of order 2, 3 and 4 from top to bottom, respectively. The right column in Figure 5.6 presents the reflected energy from the pipe end 3m from which a 55mm blockage is placed. In this case the pipe has an open lateral connection between the blockage and the sensor. The approximation here is made with the polynomial expression of the same orders as shown in the left column in the same figure. Generally, 2nd order polynomial could not describe fully of the behaviour of real data, the 3rd and 4th polynomials fit better but have a very little difference between them. In order to reduce the complexity of classification, a better trade-off is to use a smaller number of features in the form of polynomial coefficients, i.e. a low-order polynomial which could provide a reasonable fitting to the real data.



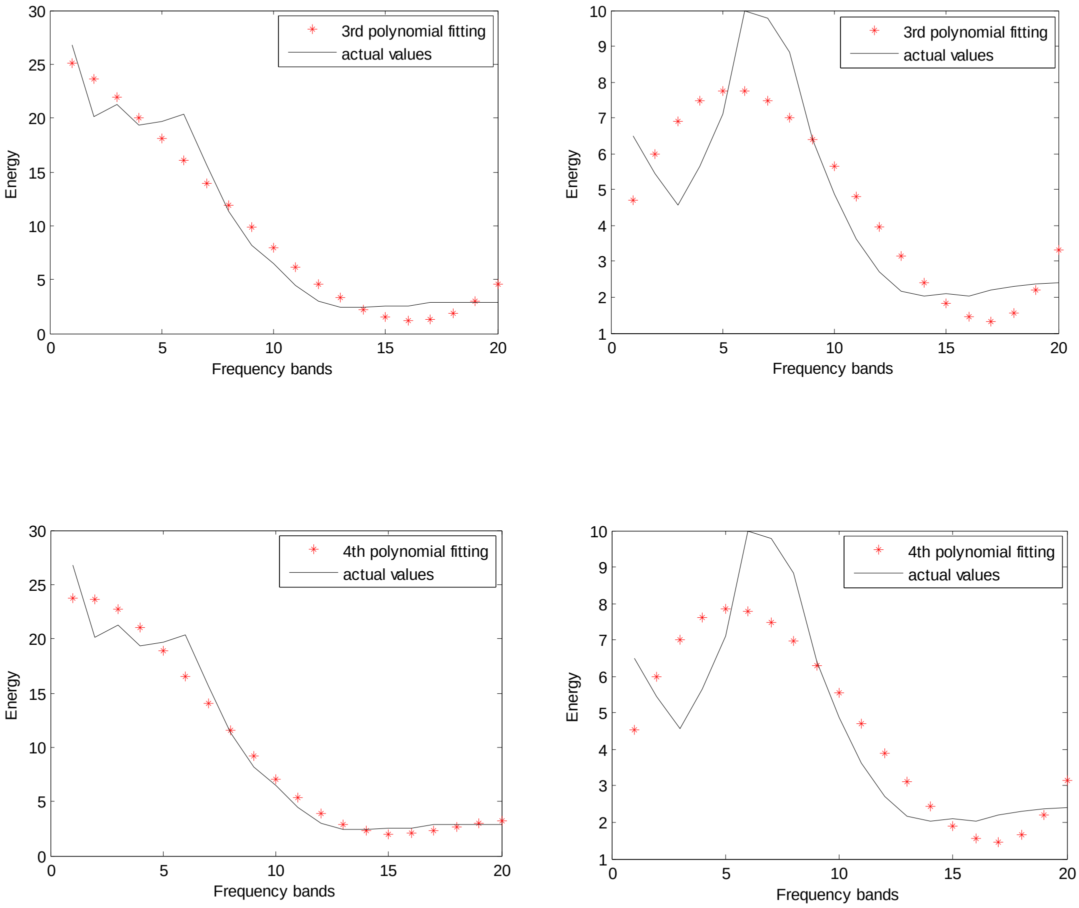


Figure 5.6 Polynomial approximation of order 2, 3 and 4 for the reflected energy from the clean pipe end (left) and the pipe end with a blockage placed before it (right)

It can be seen from the top figure in Figure 5.5 that when the frequency was higher than a certain point, most of the energy value fell in a very narrow area which have no other contribution to the classification system but increase the computational burden and bring uncertainty into the outcome, choosing several more distinguishable frequency bands could not only improve the classification performance but also make the polynomial fitting more effective. To keep the unique character of each pattern, frequency bands from 5 to 12 (200~550Hz) were chosen from which polynomial approximation coefficients will be extracted and used for classification system training. Figure 5.7 gives examples of 3rd polynomial fitting to the entire energy data set and the fitting to the same set of data within the chosen frequency range. The coefficients $a_{3,}a_2, a_1, a_0$ of 3rd polynomial function $f(x) = a_3x^3 + a_2x^2 + a_1x + a_0$ obtained from fitting each set of energy spectrum of all defects will be used for condition classification as input data.

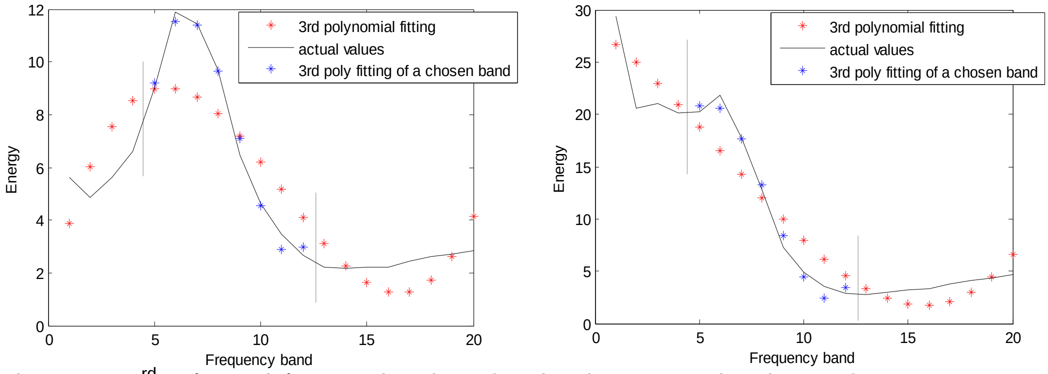


Figure 5.7 3rd polynomial approximation of entire data set and a chosen frequency range of 2 types of defect.

5.3.2 Padé Approximation

Padé approximants are derived by expanding a function as a ratio of two power series and determining both the numerator and denominator coefficients. Its theorem and formulation were given in Chapter 3. It is most frequently used in a control system to approximate the transfer function. In this research, it is expected to simply provide a more accurate approximation to of energy spectrum data because the coefficients in the Pade approximation are able to capture better of the essence of a spectral pattern a polynomial fit. A most recommended Padé approximation for the type of spectra considered in this work is 2nd order approximation in which the polynomial order in numerator is equal to that in the denominator degree [3, 4]:

$$f(x) \square R_2^2(x) = \frac{a_2x^2 + a_1x + a_0}{b_2x^2 + b_1x + 1} \qquad (5.7)$$

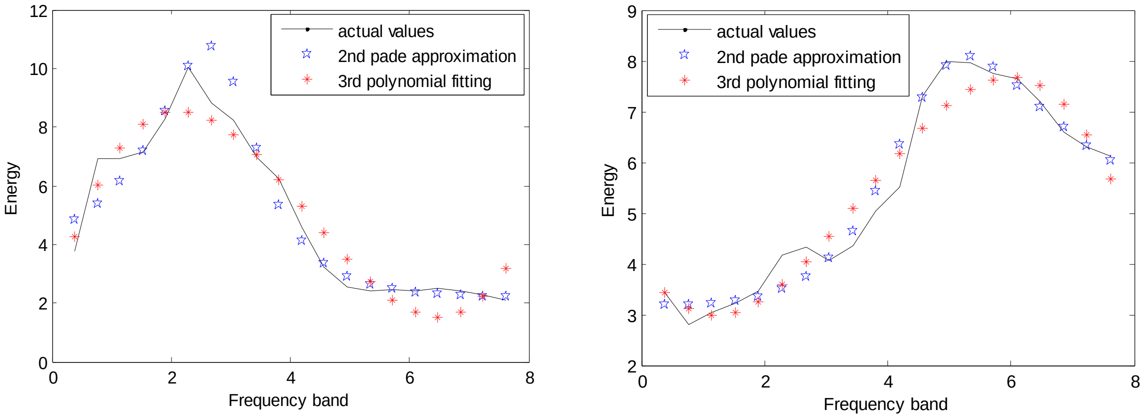


Figure 5.8 2nd Padé approximation and 3rd polynomial approximation of reflected energy spectrum of PE (left) and BK (right)

Comparing with polynomial approximation, Padé approximation has a few advantages: 2nd order Padé approximation can provide a better fit to an entire energy data set without having to choose a certain frequency range; clearly Padé approximation fits better and captures more characters of a hidden pattern in a data set. However, a 2nd order Padé approximation yields 5 coefficients: a_2, a_1, a_0, b_2, b_1 whereas a 3rd order polynomial has 4 coefficients only. As a result, before these features are used in the classification system, an additional selection procedure is necessary. This can help to avoid bringing in the classification process any unwanted features which can cause excessive calculation time and confuse the classifier.

5.3.3 Feature Selection

The assumption here is that features of one particular type of conditions which can exist in a pipe are expected to be characteristic so that the classifier could recognize them and discriminate from those features that are characteristic of some other type of conditions. In order to testify which coefficients are the parameters that a classifier could take advantage of, 4 coefficients of 3rd polynomial and 5 coefficients of Padé approximation derived from 10 sets of energy spectrum data for each of the defect signatures extracted from a range of pipe conditions are plotted as shown in Figure 5.9 and Figure 5.10.

Pipe end (PE), blockage (BK) and lateral connection (LC) are the conditions of interest to us which are to be recognized and classified by the system. As clearly indicated in Figure 5.9 and Figure 5.10, some coefficients can be separated visually. This enables us to train an automatic classifier to recognize each group to which a particular coefficient belongs to a class and test the effectiveness of this classifier using testing samples. In this work, 3rd order polynomial coefficients a_3, a_2, a_1 and 2nd order Padé coefficients a_2, a_1, a_0 are selected as the features for the classification system. For each set of features, a classifier will be trained to generate a cluster for each of the known classes of data and assign a test sample to a particular class that is determined by the kernel function adopted by the classifier.

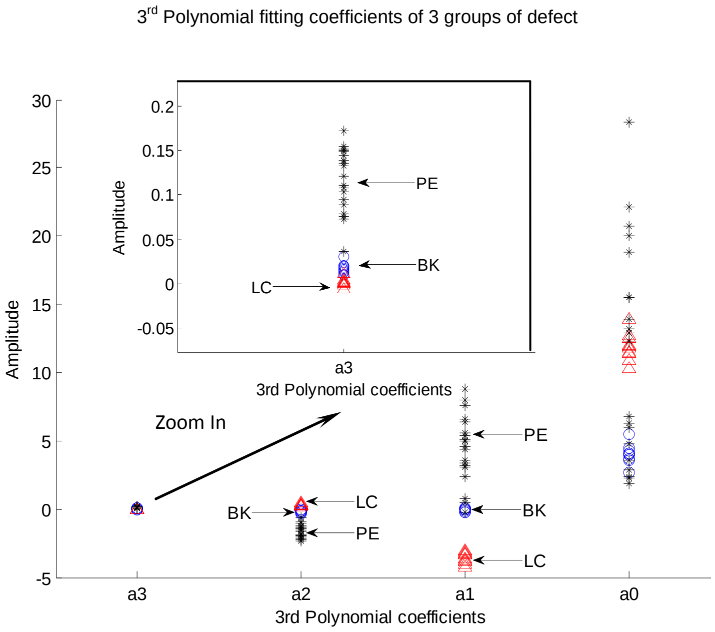


Figure 5.9 The values of the 3rd order polynomial approximation coefficients and their groupings

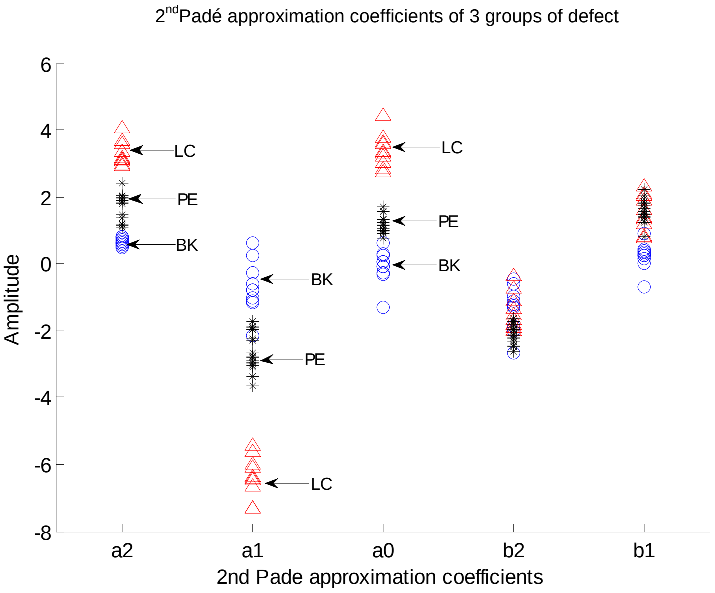


Figure 5.10 The values of the 2nd order Padé approximation coefficients and their groupings.

5.4 Condition Classification Using SVMs Technique

The K-nearest neighbours (KNN) algorithm was used for siphon condition classification as described in Chapter 4. This is a straightforward method which is non-parametric and which works well on data that can be represented as points in a n-dimensional space. In this case, the Euclidean distance can be calculated to measure the similarity or dissimilarity between an object and a labelled group to which it can be called close. The KNN is sensitive to the value of K since the ‘distance’ is not a robust statistical characteristic. If the desired K is not known in advance, one will have to try different values of K and choose a criterion to select one of the results, which will increase the computational burden especially when dealing with a large number of data points. In this chapter, Support Vector Machines (SVMs) was adopted which is considered a state-of-the-art of classification method which is founded strongly on the theoretical foundations developed by Vapnik and Chervonenkis theory [5] (see Chapter 3).

The SVMs is a supervised learning model which means that there are a few parameters which the user needs to define during the training and classification process. The SVMs was developed originally as a binary classifier. However, in this research project there are at least 3 objects which need to be classified and recognized. Therefore, a one-against-all method was introduced here to solve the problem by repeated use one of the classes as a positive class and the rest as a negative class so that two-class classification can be performed. The kernel functions are adopted and used to define a variety of nonlinear relationship between the inputs, if a linear classifier could not separate them by observation. The classification results could be very different due to different choices of kernels, however, there is no single rule to help picking up a suitable kernel prior to the training, linear function and a few basic kernels will be adopted and compared.

5.4.1Defect Classification in the Pipe with Variable Water Levels

As summarized in Table 5.1, acoustic signatures were extracted from a range of pipe conditions. Here each measurement corresponds to one particular water level so that the experiment provides 80 signatures for pipe end, 40 signatures for blockage and 40 signatures for lateral connection condition in the pipe. In feature extraction and selection procedures, 2 sets of coefficients were derived by approximating the reflected energy-frequency spectrum of each signature using two approximation techniques. These coefficients where split into 2 parts: one used for training and the other used for testing. To ensure the balance of the training datasets, we chose to use 32 signatures from each class to train the system. The effect of different size of the training datasets will be discussed in the section 5.4.1.3.

5.4.1.1 Use Polynomial Coefficients as Input Features

The coefficients a_1, a_2, a_3 a_1, a_2, a_3 in the 3rd order polynomial approximation derived from the energy spectrum in the reflection signatures are plotted in Figure 5.11 as a function of the water level for the pipe end, blockage and lateral connection. This approximation was taken over the range from 200 to 550Hz. The range of the value of these coefficients for each defect is rather clear so that these can be separated using linear classifiers and/or other low-order classifiers which provide smooth separation. A linear classifier produces a linear decision boundary that is defined as:

$$\left\{\mathrm{x},\, f(\mathrm{x}) = \left\langle \mathrm{w},\mathrm{x} \right\rangle + b = \gamma \right\} \tag{5.8}$$

A non-linear classifier using a non-linear kernel function

$K(x_i, x_j) = \left\langle \phi(x_i), \phi(x_j) \right\rangle$ is given by:

$$f(\mathrm{x}) = \sum_{i=1}^{n} \alpha_i K(\mathrm{x}, \mathrm{x}_i) + b \tag{5.9}$$

where x_i is a set of support vectors. One popular kernel is a Polynomial kernel of degree- d . This kernel is defined as:

$$K(x_i, x_j) = (\left\langle \phi(x_i), \phi(x_j) \right\rangle + b)^d \tag{5.10}$$

The kernel with $d = 1$ is the linear kernel. The increasing degree results in the increase of the flexibility of the classifier and produces more curvature in the decision boundary. Another widely used kernel is the Gaussian kernel defined by:

$$K(x_i, x_j) = \exp(-\gamma \left\| \phi(x_i) - \phi(x_j) \right\|^2) \tag{5.11}$$

where $\gamma > 0$ is a parameter that controls the width of the Gaussian kernel that is, sometimes parameterized using $\gamma = 1/2\sigma^2$. This parameter plays a similar role as the degree of the polynomial kernel. Normally a Gaussian kernel is referred to as the Gaussian RBF (Radial Basis Function) kernel which is used in support vector classification. Linear kernel, polynomial kernel and

Gaussian RBF kernel are used in the following sections to train the classifier for each set of features.

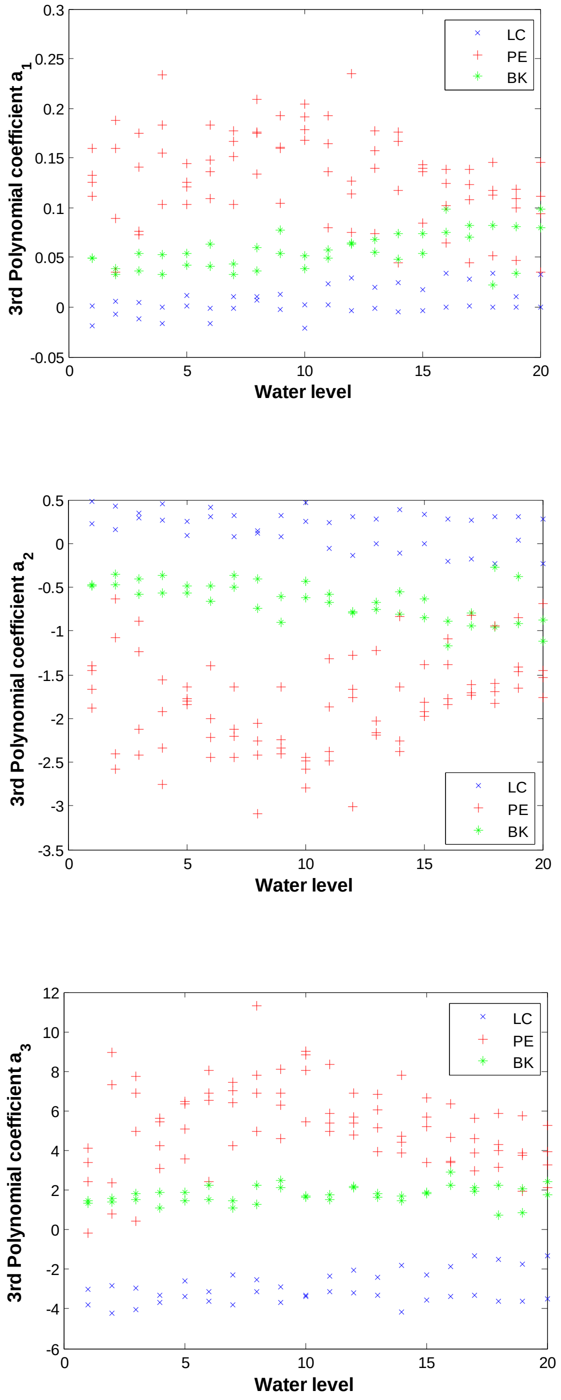
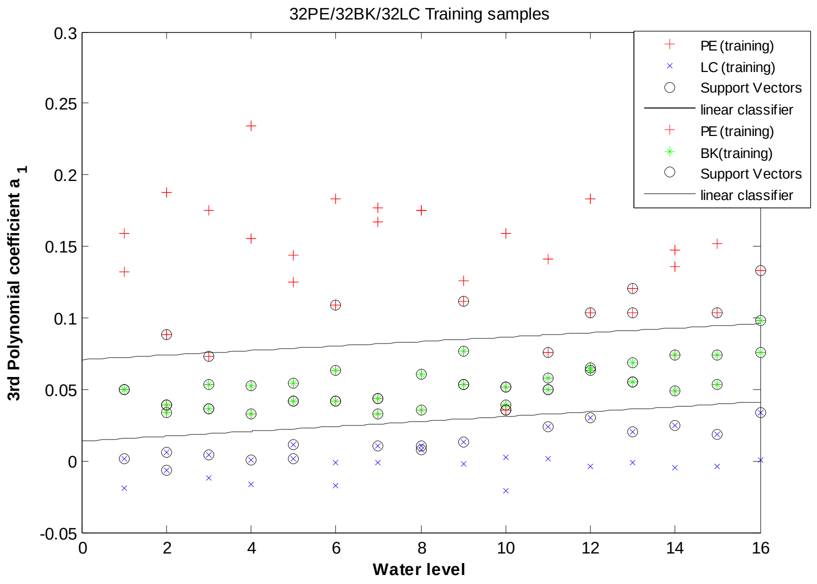
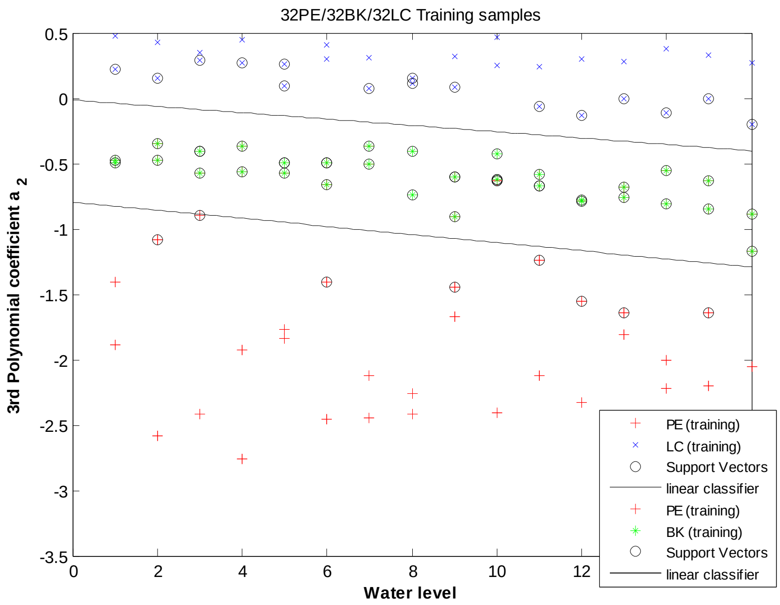


Figure 5.11 The value of the 3rd order polynomial coefficients a_1 (top left), a_2 (top right) and a_3 (left) derived from the energy spectrum of the acoustic signatures for a pipe end (PE), blockage (BK) and lateral connection (LC) and plotted as a function of the water level.

Linear Classifier

A linear classifier corresponds to a separating hyperplane $f(x)$, which is a line in a 2-dimensional space, that passes though the middle of the two classes, separating the two. Once the function is determined, new data x_i can be classified by simple testing the sign of the function $f(x_i)$, so that x_i belongs to the positive class if $f(x_i) > 0$.





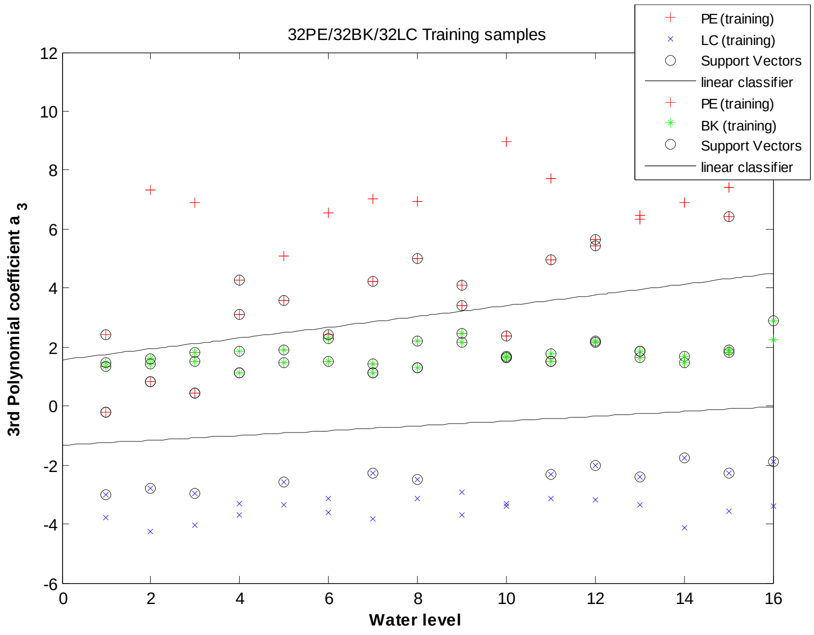


Figure 5.12 An example of the linear SVMs classification obtained by using the 3rd order polynomial coefficients: a_1 (top), a_2 (mid) and a_3 (bottom)

Figure 5.12 gives a linear classification example for 3 in-pipe defects. These results were obtained by using the 3rd order polynomial coefficients as inputs in the SVMs. In total, 96 samples were used to train the classifier, with 32samples of each of the three conditions (PE, LC, BK). In these datasets, one data sample corresponds to one water level for which the original acoustic data was recorded. The application of the one-against-all method enables us to construct 2 binary SVMs classifiers, each of which separates one class from all the rest. At the classification phase, a sample is projected onto the corresponding feature space so that it can be assigned to the class within which its feature appears. The SVMs linear classification results in True-False form using 3 3rd order polynomial coefficients as inputs are given in Table 5.3, the accuracy of all 3 coefficients are given in Table 5.4. Figure

5.13 gives a visual example of how testing samples are classified by the linear classifiers.

Table 5.3 Some Linear classification results in True-False form using polynomial coefficient a_1

Defect/Test	Test1	Test2	Test3	Test4	Test5	Test6	Test7	Test8	Test9	Test10
PE	T	F	T	T	T	T	T	T	F	T
BK	T	F	T	T	T	T	T	T	-	-
LC	T	T	T	T	T	F	T	F	-	-

Figure 5.4 The accuracy of linear SVMs classification using 3rd order polynomial fitting coefficients as input

Defect/ Feature	a_1 (lowest order)	a_2	a_3 (highest order)
Pipe End (PE)	87.5%	92.18%	96.87%
Blockage (BK)	87.5%	87.5%	100%
Lateral Connection (LC)	75%	100%	100%

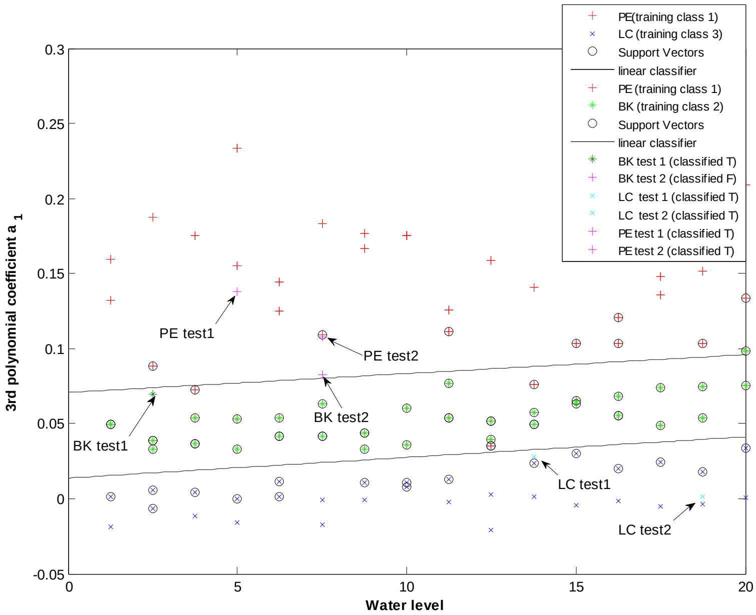


Figure 5.13 Condition recognition of 6 testing samples by linear SVMs using the 3rd order polynomial coefficient a_1 as input features

Non-Linear classifiers

Nonlinear models are used when linear models are just not sufficient to reflect the complexity in the behaviour of the observed feature pattern. Kernels turn linear models into non-linear models by replacing the inner product by a kernel function which represents the data in some higher dimensional feature space in which the data could be linearly separated. The prediction for a test sample x is given by equation (5.12) when g a linear classifier is available:

$$y = \mathrm{sign}(\left\langle \mathrm{w}^{\mathrm{T}}, x \right\rangle + \mathrm{b})$$

$$= \mathrm{sign}(\sum_{i \in SV} \alpha_i y_i x_i{}^{\mathrm{T}} x + b) \qquad (5.12)$$

where SV is a set of support vectors, sign is the signum function which is used to define the decision boundary. Non-linear prediction of x is to simply replace each sample with its feature mapped representation $x \to \phi(x)$:

$$y = \mathrm{sign}(\sum_{i \in SV} \alpha_i y_i \phi(x_i)^{\mathrm{T}} \phi(x))$$

$$= \mathrm{sign}(\sum_{i \in SV} \alpha_i y_i K(x_i, x)) \qquad (5.13)$$

Polynomial kernel and Gaussian kernel are very commonly used to train a classifier for non-linear separations. Figure 5.11 suggested that 3 classes are almost linear separable, therefore low orders of non-linear classifiers would work well to provide smooth decision boundaries. 2nd and 3rd polynomial kernels, Gaussian kernel width parameter σ equals to 0.2 and 2 were tried using the same set of training and testing samples as in the case of linear classification. Figure 5.14 shows classification using the 2nd order polynomial (top) and 3rd order polynomial (bottom) classifiers, from which it can be seen that the 2nd order polynomial classifier generated similar boundaries as the linear classifier therefore the classification results are approximate. The 3rd order polynomial classifiers led to overfitting where misclassified BK test sample 1 (circled in Figure 5.14) and created two unwanted areas by being too sensitive to the samples.

Width parameter σ in the Gaussian RBF kernel determines the area of influence that support vectors have over the feature space. A larger value of σ results in smoother and more regular boundaries while small values of σ add more curvature to the boundaries as indicated in Figure 5.15. A large value of σ can also reduce the number of support vectors required for classification. Similar to the case with the polynomial classifiers, overfitting could occur if the parameter was chosen inappropriately. The accuracy of the class predictions based on the coefficients a_1 as input for all testing samples and linear, polynomial and Gaussian kernels used for separation is quoted in Table 5.5. The results suggest that for linearly separable data, linear classifier is the most suitable machine learning algorithm to design a linear support vector machines system which could provide a fast process and superior performance than other non-linear classifiers.

Table 5.5 Classification accuracy comparison among linear, polynomial and Gaussian kernels using a_1 as input

Defect/Classifier	**Linear**	**Polynomial**		**Gaussian**	
		$d=2$	$d=3$	$\sigma=0.2$	$\sigma=2$
Pipe end	87.5%	82.81%	90.62%	85.94%	92.18%
Blockage	87.5%	75%	37.5%	25%	75%
Lateral connection	75%	75%	62.5%	62.5%	62.5%

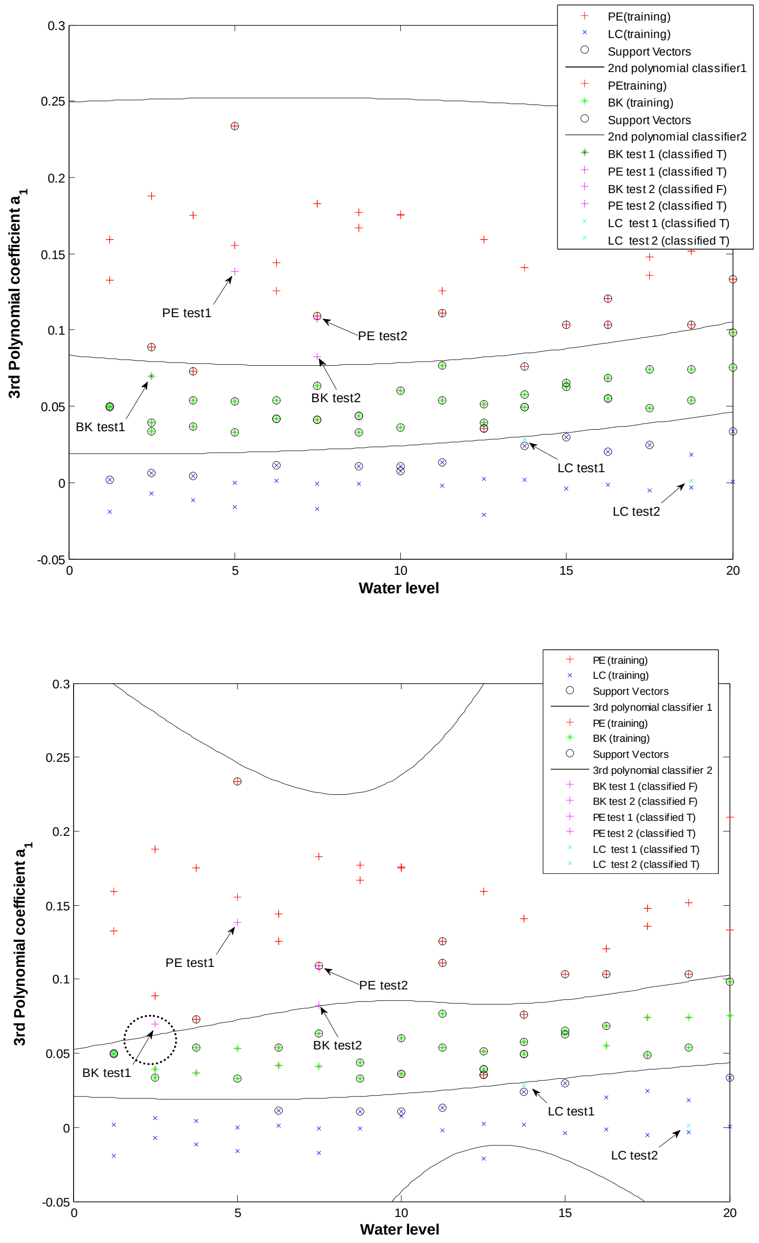


Figure 5.14 Classification examples obtained by 2nd order (top) and 3rd order (bottom) polynomial classifiers

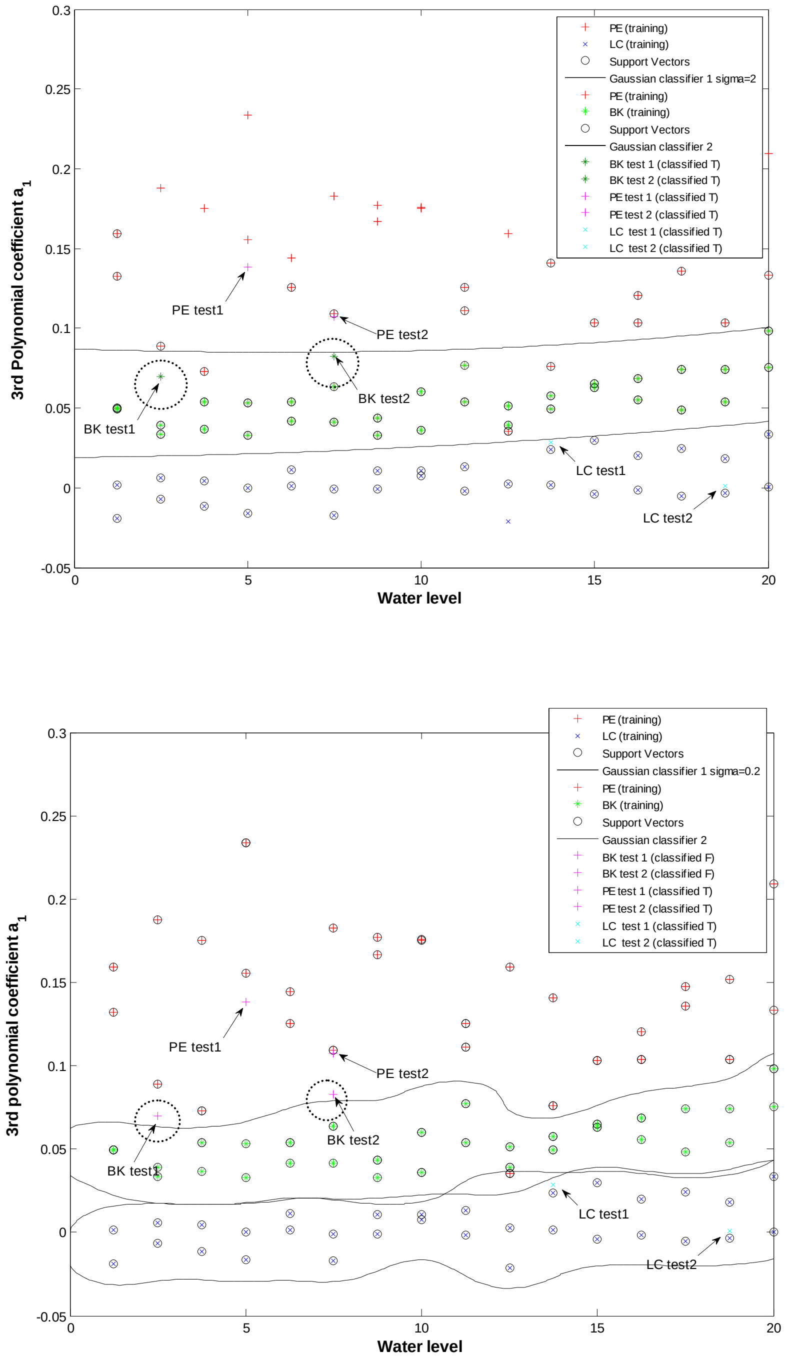


Figure 5.15 Classification examples obtained using Gaussian RBF classifiers with σ =2 (top) and 0.2 (bottom)

5.4.1.2 Use Padé Approximation Coefficients as Input Features

As presented in Chapter 3, Padé approximation as an alternative of polynomial approximation provides more precise fitting to the data in some particular cases. The coefficients a_0, a_1, a_2 in Padé approximation appeared particularly useful and, therefore, were selected for classification as described in the previous section (5.3.3). Feature extraction and selection procedures were repeated as done for using polynomial coefficients as input. The 2nd order Padé approximation coefficients were obtained for the data for all three conditions studied in this work. These coefficients are plotted as a function of water level in Figure 5.16, which shows the pattern that these coefficients following in the feature space. Similar to the patterns observed in the case of polynomial coefficients, Padé coefficients derived for each of the three conditions appear linearly separable so that linear classifiers can be suitable for separating the classes. Figure 5.17 gives examples of using 3 sets of Padé coefficient as input. It is possible to separate the three conditions and test the decision boundaries by introducing testing samples to the system.

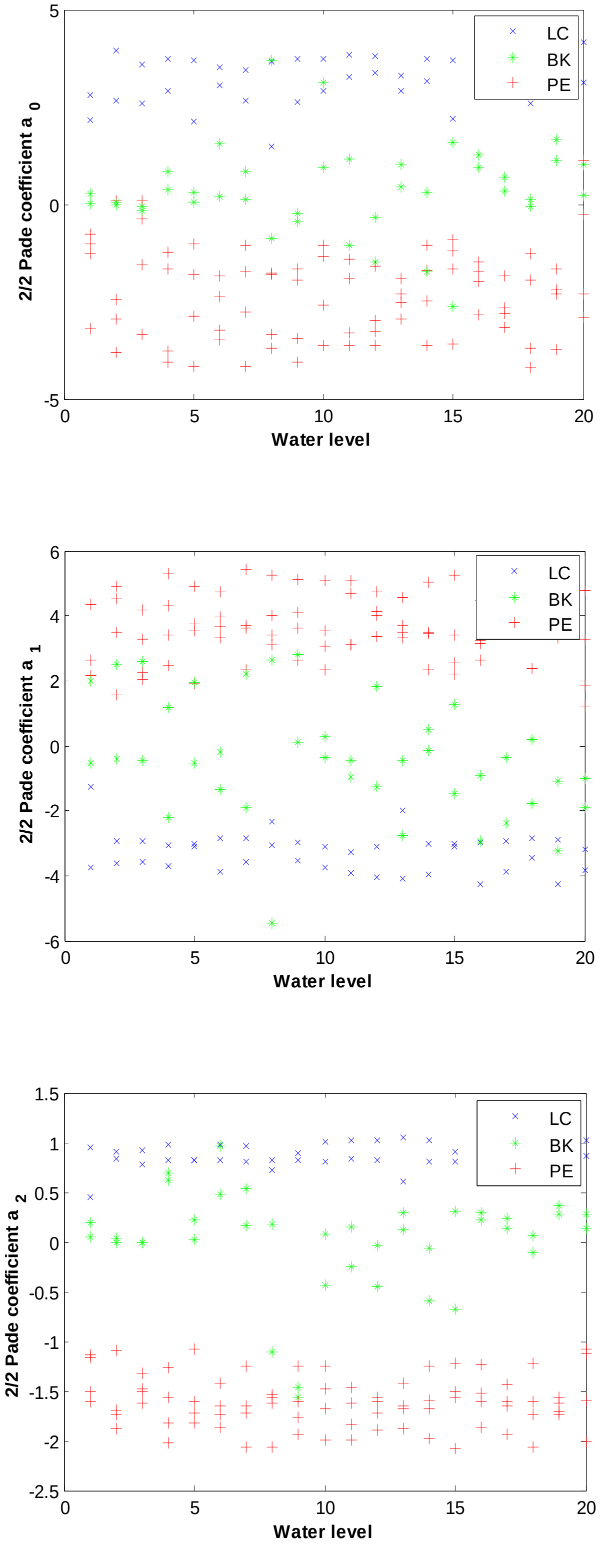


Figure 5.16 The 2nd order Padé coefficients a_0 (top left), a_1 (top right) and a_2 (left) derived from energy spectrum of PE, BK and LC signatures for a range of water level

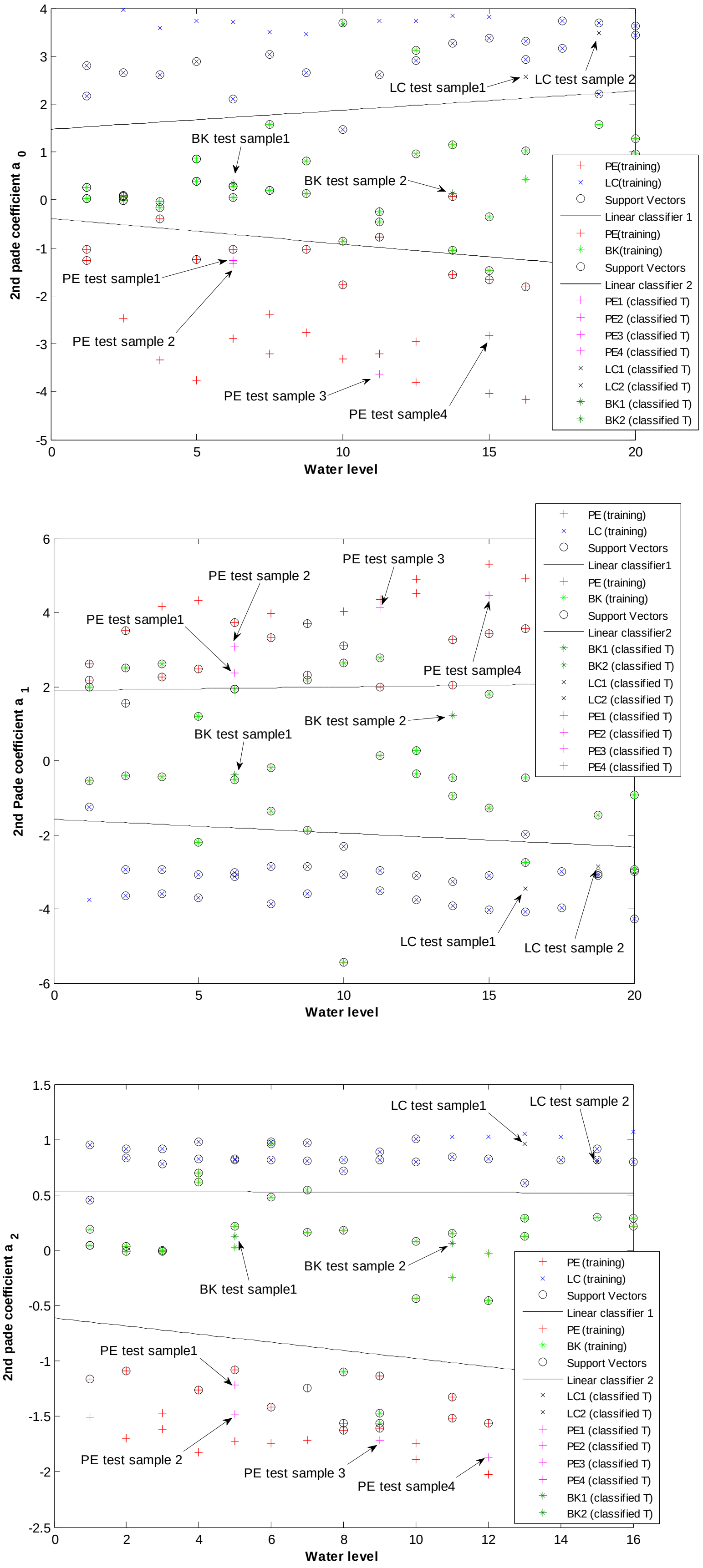


Figure 5.17 Linear classification using the 2nd order Padé coefficients as inputs. Order from low to high: a_0 (top), a_1 (middle) and a_2 (bottom).

Table 5.6 Linear classification accuracy of using 2nd Padé approximation coefficients as input

Defect/ Feature	a_0 (lowest order)	a_1	a_2 (highest order)
Pipe End (PE)	93.75%	96.87%	100%
Blockage (BK)	100%	75%	87.5%
Lateral Connection (LC)	87.5%	100%	100%

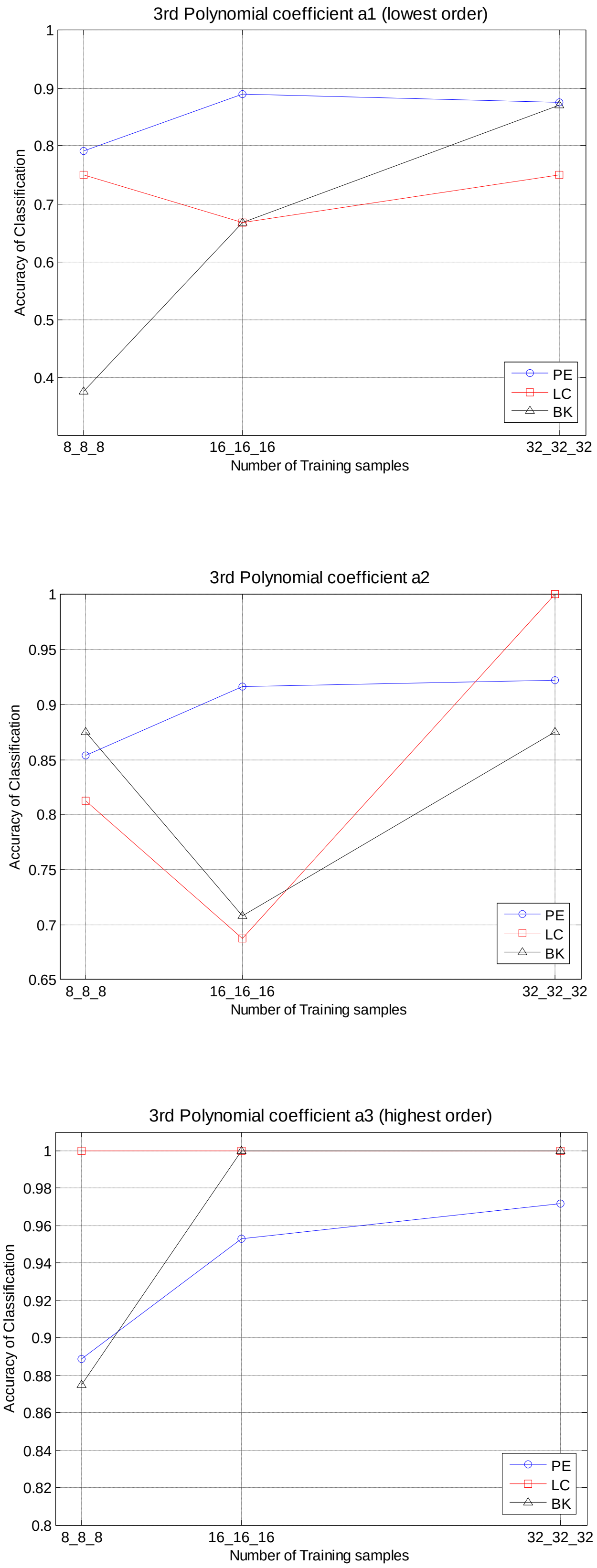
Classification results of using Padé coefficients as input of the system are given in Table 5.6. A comparison of these data with that presented in Table 5.6 suggest that the performance of classification was improved in some occasions by replacing the polynomial coefficients with Padé coefficients, i.e. that Padé approximation provides more accurate fitting when the observations have more complicated structure . A main advantage of polynomials is their simplicity; one of the main disadvantages of polynomial approximation is that high degree polynomials are needed in order to reach a satisfactory accuracy level for polynomial approximations. Padé approximation as a member of ration function families can accommodate a much wider range of shapes than does the polynomial families. It can also be used to model complicated structure with a fairly low degree in both the numerator and denominator, which meansfewer coefficients will be required compared to the polynomial model.

5.4.1.3 Effect of Training Sample Size

There are no well-defined rules of how many training samples are enough to train a system properly. Generally, it is expected that training sets with more samples will be “better” as they will provide better definition of the region of interests, and as a result a model based on a larger sample set should provide more reliable inferences of condition. However, a large number of training samples means expensive computational cost and possibly leads to some confusion in classification and excessive sensitivity of the system. As shown in Table 5.7, 3 sets of training samples containing different size of coefficients will be used to train the system and compare the outcomes. Figure 5.18 and Table 5.8 to 5.10 prove that small size of samples are not enough to train the system to provide a reliable performance, training set 3 contains more than half samples which covered reasonable number of conditions of each defect, it is expected to be rich enough to reflect the complexity of each class. It is also the training datasets used to train the system for classification in the previous sections.

Table 5.7 Number of features and training samples

Features/numbers	Total Amount	Training set 1	Training set 2	Training set 3
	PE/BK/LC	PE/BK/LC	PE/BK/LC	PE/BK/LC
3rd Polynomial coefficients: a_3, a_2, a_1	80/40/40	8/8/8	16/16/16	32/32/32
2nd padé coefficients: a_2, a_1, a_0	80/40/40	8/8/8	16/16/16	32/32/32



3rd Polynomial coefficient a1 (lowest order)

3rd Polynomial coefficient a2

3rd Polynomial coefficient a3 (highest order)

Figure 5.18 Size of training sample against Classification accuracy

Table 5.8 Linear Classification accuracy using different size of training samples of a_1 (lowest)

Defect/No. of samples	8/8/8	16/16/16	32/32/32
PE	79.16%	89%	87.5%
BK	37.5%	66.67%	87.5%
LC	75%	66.67%	75%

Table 5.9 Linear Classification accuracy using different size of training samples of a_2

Defect/No. of samples	8/8/8	16/16/16	32/32/32
PE	85.42%	91.67%	92.18%
BK	87.5%	68.75%	87.5%
LC	81.25%	68.75%	100%

Table 5.10 Linear Classification accuracy using different size of training samples of a_3 (highest)

Defect/No. of samples	8/8/8	16/16/16	32/32/32
PE	88.89%	95.31%	96.87%
BK	87.5%	100%	100%
LC	100%	100%	100%

5.4.2Defect Classification with Multiple Objects Effect

The aim of this study was to test how much the interaction between defects could affect the ability of the classification system to recognize defects and to make the right decision. Table 5.2 lists all the conditions taken into account in the multiple object effect experiments. 5 sets of measurement were taken for each pipe condition, 3 of which were used to collect training features; the other 2 sets were used for testing. Unlike water level effect experiments in which equal numbers of samples were collected for all the defects, datasets used in the multiple objects experiments were imbalanced meaning that each of the classes contained different number of training samples. In total, 60 PE signatures, 60 BK signatures and 30 LC signatures were extracted from the original data.

Many research papers on imbalanced data sets report that the performance of the existing classifiers in this case tends to be biased towards the majority class, i.e. (the class that contains more samples than the others)[5]. A number of solutions to the imbalanced datasets problem were proposed including re-sampling, adjusting the decision threshold and cost-sensitive learning etc [6, 7, 8]. Some of these solutions are tested in this study.

5.4.2.1 Learning from Imbalanced Data

In binary learning from imbalanced datasets, the class with fewer samples is known as the minority class or positive, while the other class with larger samples is called the majority class or negative. A number of solutions to the class-imbalance problem were previously proposed both at the data and algorithmic levels [6, 7]. At the data level, these solutions include different forms of re-sampling such as random oversampling of the minority class, or random under sampling of the majority class. At the algorithmic level, solutions include cost-sensitive learning, over-weighting errors on the minority class, ensemble methods, trained from learning sets with different data distributions, post-processing by tuning the learning classification function to improve performance on minority class, etc.

In this research, the imbalance is due to the fact that different defects simulated in the laboratory and in reality are not always in equal numbers, it is a direct result of the nature of the data space, and the imbalance of this form is referred to as intrinsic. For this type of data, some studies have shown that strategies like some of the re-sampling methods and adjusting the decision threshold provide improved classification performance [7, 8]. The methods adopted in this section include re-size the training samples and introduce a second decision rule.

5.4.2.2 Resample the Training Samples

Among many strategies of learning from imbalanced data, re-sampling methods aid to modify an imbalanced datasets to a balanced distribution and have been proved to be able to improve the accuracy of classifiers [9]. Traditional re-sampling methods like random over-sampling and under-sampling replicate or eliminate samples for the minority class or majority class until it contains as many samples as the other class. These methods have their limitations which related to either a loss of information by removing random samples from the majority class, or lead to overfitting by introducing copies of random existing samples. Informed under-sampling and synthetic sampling are improved alternatives to overcome the limitations in the traditional way [9]. Here, a second classifier, K-nearest neighbours (KNN), is introduced to achieve a more appropriate and specific re-sampling.

K-nearest neighbours (KNN) as a classification technique that was introduced and applied in Chapter 4 in the siphon data analysis. It has shown a great deal of success in various applications including informed under-sampling and synthetic oversampling [10]. KNN under-sampling selects a given number of those majority class samples whose average distance to the other minority classes samples are the smallest. It is then used to refine the majority class by removing these samples until it contains the same size of samples as the other classes. K-nearest neighbours can also be used in oversampling to create artificial data based on the feature space similarities between existing minority examples. In this process, a synthetic sample is created through a random selection of one of the K-nearest neighbours of each of the existing sample x_i in a minority class, which is then multiplied by the corresponding feature vector difference with a random number between [0, 1]. Finally, new data are created and as:

$$x_{new} = x_i + (\hat{x}_i - x_i) \times \delta \tag{5.17}$$

where $\hat{x}_i$ is one of the nearest neighbours of x_i , x_i and $\hat{x}_i$ belong to a same minority class, and $\delta \in [0,1]$ is a random number. The resulting synthetic instance according to (5.17) is a point along the line between x_i and $\hat{x}_i$ .As the minority class is over-sampled by creating "synthetic" examples rather than by over-sampling with replacement, it effectively forces the decision region of the minority class to become more general.We generate synthetic examples in a less application-specific manner, by operating in "feature space" rather than "data space".Replication of the minority class using (5.17) does not cause its decision boundary to spread into the majority class region, but it providesmore related minority class samples to learn from.This enables theuser to definebroader decision regions, leading to more coverage of the minority class.

Feature extraction of multiple objects data is based on the same procedure as that applied in the case of the data obtained for the variable water in the pipe. Here we use 3rd order polynomial and 2nd order Padé approximations to obtain fitting coefficients for the acoustic energy spectrum which corresponds to the reflection of the signals from a pipe defect. Three defect classes contain different sample size: 60 PE signatures, 60 BK signatures and 30 LC signatures. Two options are available to balance these classes: (1) under-sampling the majority classes (PE and BK) to match the size of samples of minority class (LC); (2) oversampling the minority class until it contains the same number of samples as the majority classes.

Under-sampling

The distribution of original 3rd order polynomial coefficient features obtained from multiple defects experiment is given in Figure 5.19. In order to remove 30 samples from PE and BK classes and refine the boundaries, a selection of redundant samples is needed. The redundant samples are those which do not harm the correct classification but increase the classification costs. Selection techniques were studied by the statistical literature of the 60s and 70s and were later investigated by researchers specializing in machine learning. In particular, a recognition-based learning rule of detecting and removing less reliable samples is given by Kubat [11] and it is named One Sided Selection (OSS). This rule is based on the following assumptions:

1. Those samples that either cross or are very close to the borderlines are unreliable as they can increase the sensitivity of the classifier to an exorbitant level. Those samples which are far away from the borderlines are redundant as they could be taken over by other samples and not change the

classification result. The rest of the samples are safe to keep for further classification task.

2. Noisy and borderline samples can be detected using Tomek Links algorithm [12] which is a form of the K-Nearest Neighbours (KNN) algorithm, it was applied as a data cleaning method in order to remove the noisy and borderline instances from the training set. Assume two samples x_i and x_j belonging to different classes, $d(x_i, x_j)$ is the distance between them, a pair of (x_i, x_j) is called a Tomek Link if there is no other sample x_l such that $d(x_i, x_l) < d(x_i, x_j)$ or $d(x_j, x_l) < d(x_i, x_j)$.

3. Redundant samples can be reduced by creating a subset $C \subseteq S$, S is the original training dataset. Initially, C contains all the positive samples (samples from minority class) and one randomly selected negative sample, then 1-NN rule is applied on all the samples in C in an attempt to re-classify S, those training samples that have been misclassified in S are then added to C. Remove all the negative samples participating in Tomek Links from C, all positive samples are retained. The resulting set contains only safe and reliable samples for further classification.

From Figure 5.19 it can be seen that if we use a_0 as a feature, PE and LC are severely overlapped, while the other three polynomial coefficients appear separable. Therefore, polynomial coefficients a_1, a_2, a_3 are more reliable to be used for classification. Figure 5.20 shows the training features after the application of the One Sided Selection rule to original datasets. The new sets contain the same size of samples and are well-balanced. The Padé approximation coefficients were also used for classification for water level data analysis and these have shown better classification results for some conditions. These coefficients will be used for classification and compared with the results from attained with the polynomial fits. Figure 5.21 and Figure 5.22 show the original Padé coefficients distribution and the training sets retained after re-sampling.

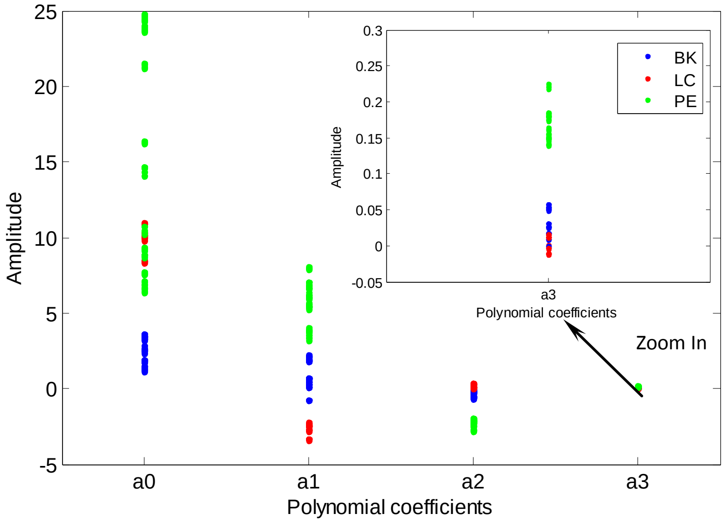


Figure 5.19 Original 3rd order polynomial features distribution

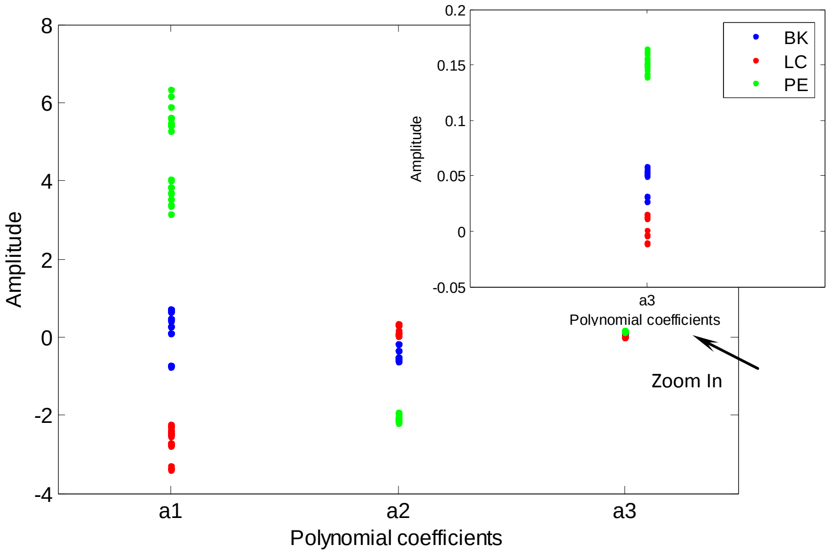
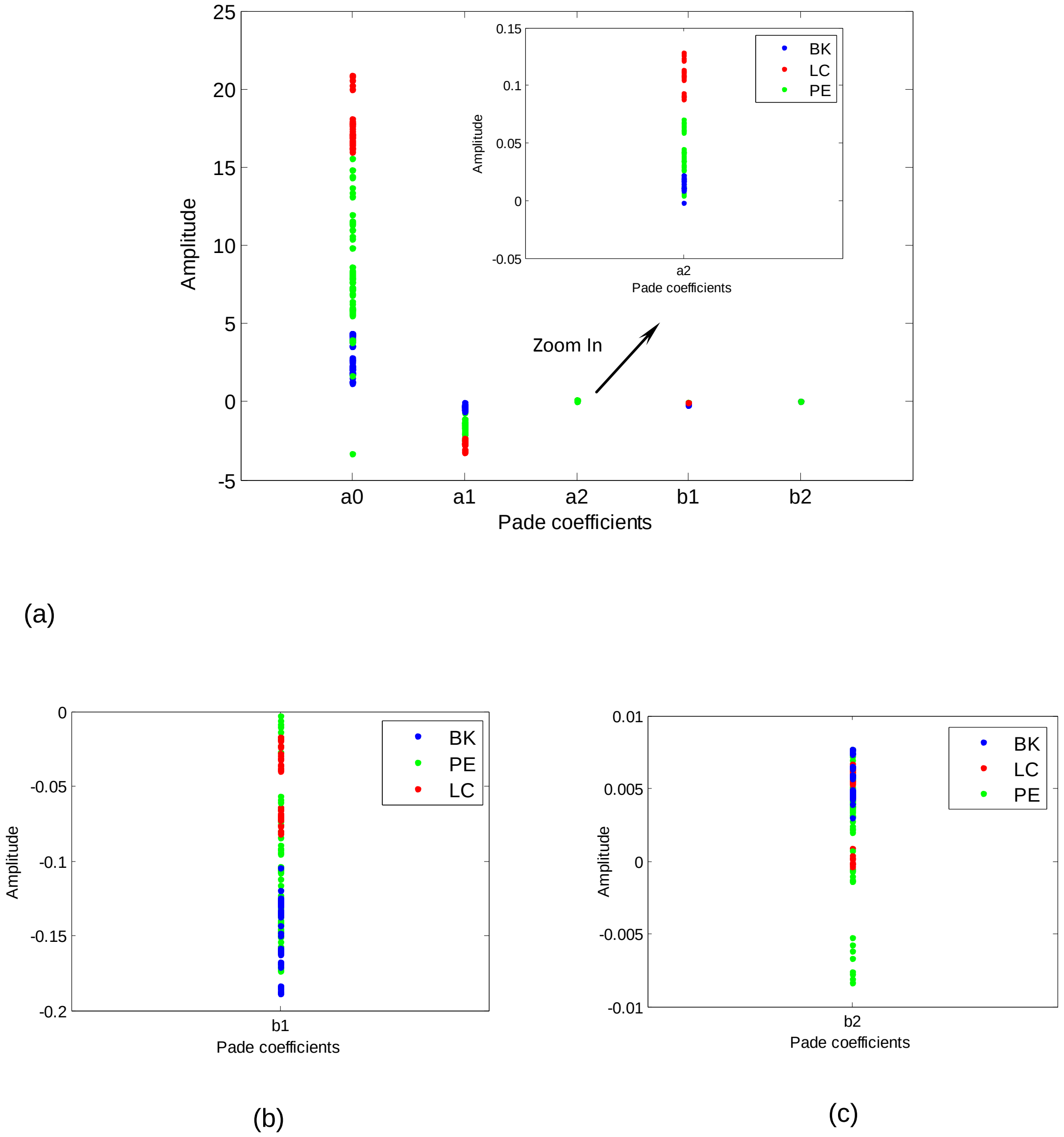


Figure 5.20 Re-sampled 3rd order polynomial features by OSS



(a)

(b)

(c)

Figure 5.21 Original 2nd order Padé features distribution (a) and zoomed-in view of coefficient b_1 (b) and b_2 (c).

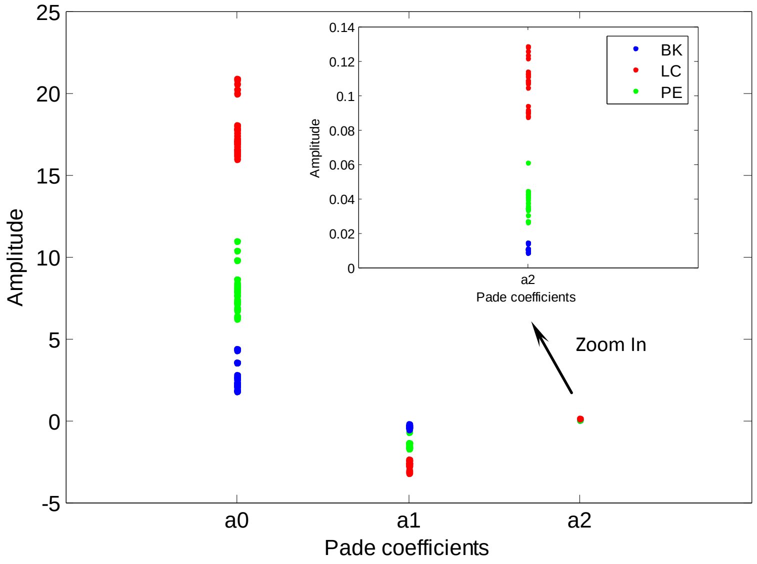


Figure 5.22 Re-sampled 2nd order Padé features by OSS

Oversampling and Data Generation

Synthetic minority oversampling technique (SMOTE) is a powerful method that has shown a great deal of success in various applications [13]. The SMOTE algorithm creates artificial data based on the feature space similarities between existing minority examples. For a subset $S_i \in S$, S is the original training dataset, consider the K-nearest neighbours for each samples $x_i \in S_i$ for some specified integer K, the K-nearest neighbours are defined as the K elements of S_i whose Euclidean distances between itself and x_i are the smallest. A new synthetic sample is then created using equation (5.17). This causes the selection of a random point along the line segment between two specific features. This approach effectively forces the decision region of the minority class to become more general [13].

LC class is in need of 30 new artificial samples to match the size of the other two classes so the data distribution could be balanced for classification. 2 parameters have impact on the new data distribution: the number of nearest neighbours K and weighting vector in equation (5.17) δ . To avoid replicating existing samples, for 6 conditions where all LC data were collected, choose one sample randomly from each condition and these 6 samples will be applied with K-nearest neighbours. New artificial samples are obtained by weighting the 30 nearest neighbours of these 6 samples when K is chosen to be 5, the weighting vector $\delta \in [0,1]$ will be taken randomly.

Figure 5.23 shows the original polynomial coefficient a_0 derived from all the existing LC samples and the zoomed-in view of the 5 nearest neighbours of one feature from LC class. The difference between the feature (sample) and its neighbours can be calculated and multiplied by a random number between 0 and 1 to be added as a feature vector. In this way, a newly created sample should appear along the line segments between the feature and its corresponding neighbour.

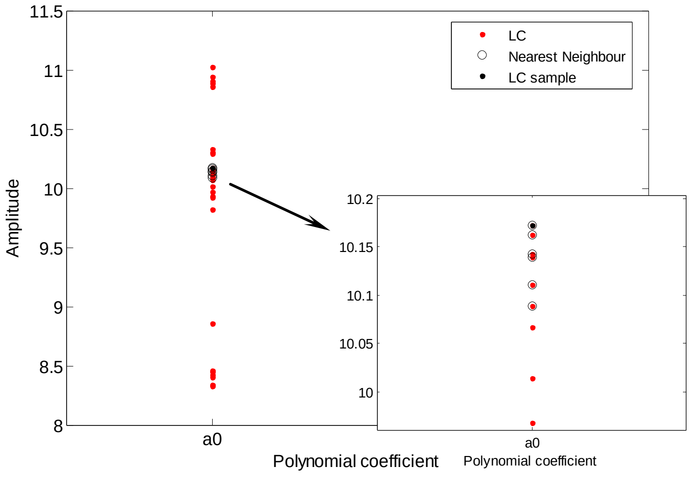


Figure 5.23 Five nearest neighbours of a sample among the class

Figure 5.24 and 5.25 present the polynomial features and Padé features distribution after over-sampling the LC class. The weighting parameter was chosen to be 0.5, each class now contains 60 samples exactly.

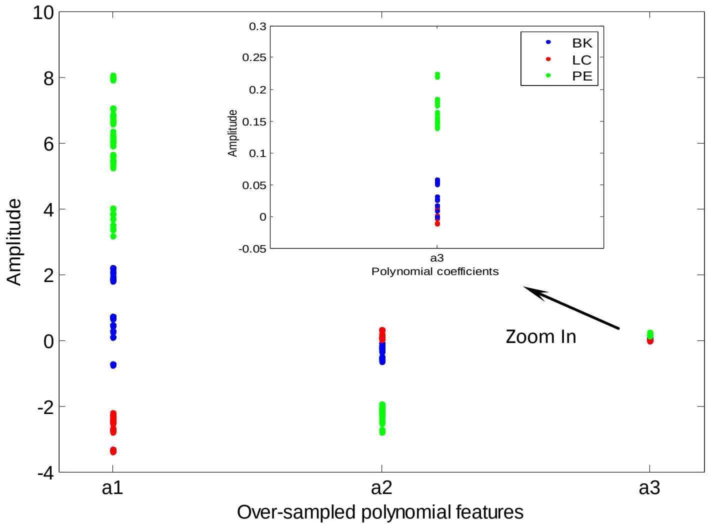


Figure 5.24 Over-sampled 3rd order polynomial features by SMOTE

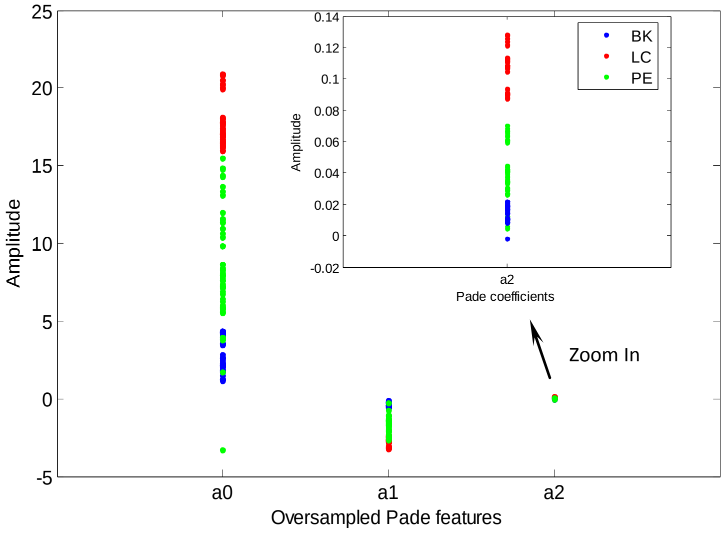


Figure 5.25 Over-sampled 2nd order Padé features by SMOTE

5.4.2.3 Defect Classification using Re-Sampled Features

Two sets of features were obtained by having applied under-sampling to the majority classes and over-sampling to the minority class. 30 samples of each class retained from the former and 60 samples from the latter.

Down-Sampled Features Classification

The number of training samples for a classification depends greatly on the nature of data and experience of the system designer. A small number of training samples can results in the loss of information while a large number can lead to over fitting of the class region. From the 5 sets of experiments conducted for one condition, 3 sets of measurements were picked randomly to train the classification system and the rest 2 sets were used for testing. In order to create balanced training datasets, down-sampling was applied as explained in Section 5.4.2.2 which resulted in 30 samples retained in each class. Linear classifiers were tried first to test if the data were separable. The coefficients (features) obtained with a 3rd order polynomial and 2nd order Padé approximation were used to train the system, as illustrated in Figure 5.26 and Figure 5.27, and the linear classifiers adopted for this purpose were capable of separating condition classes effectively. The linear classification accuracy attained by using the 3rd order polynomial features and the 2nd order Padé approximation features are given in Table 5.11 and Table 5.12, respectively.

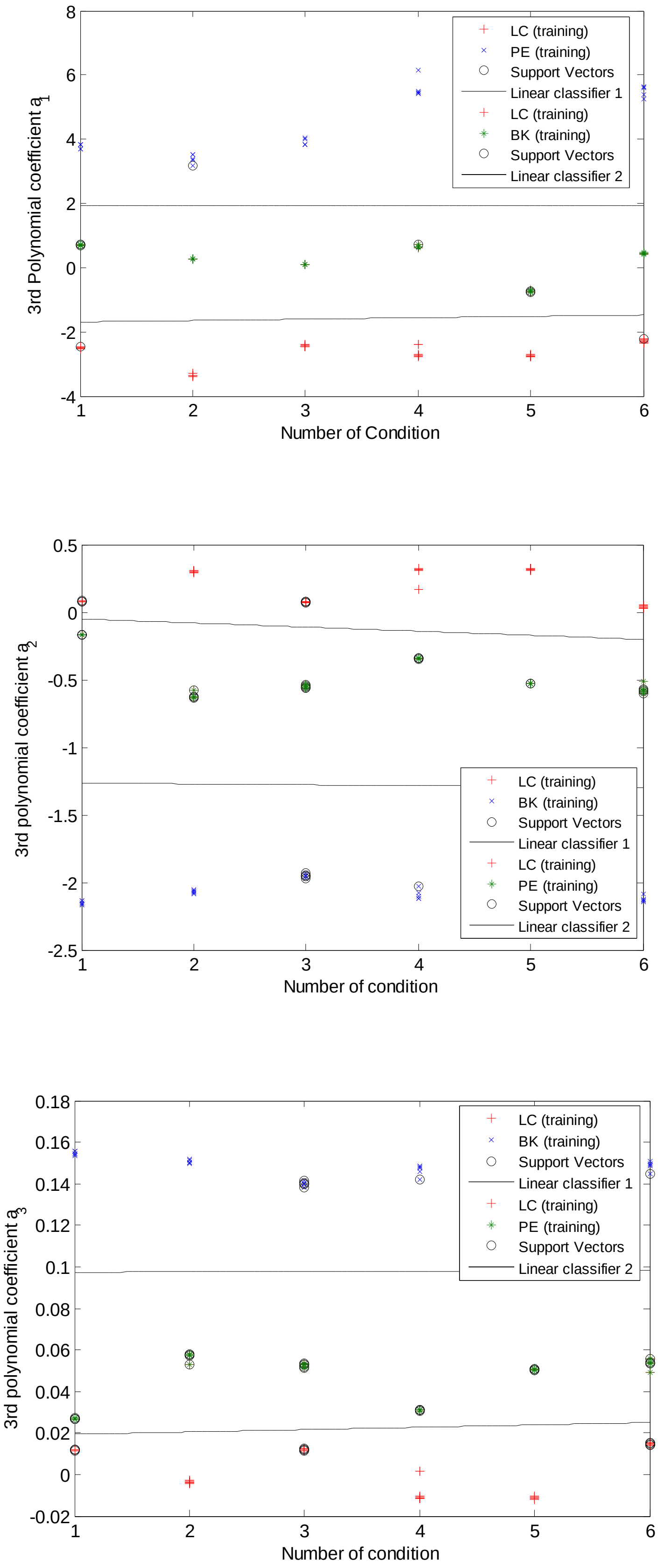
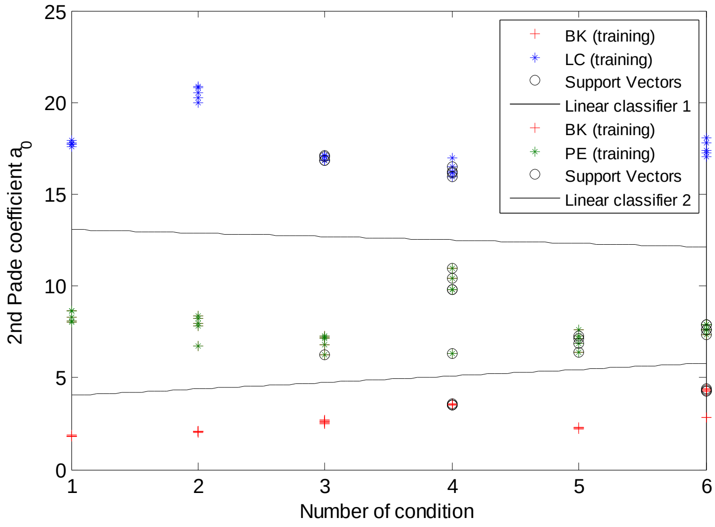
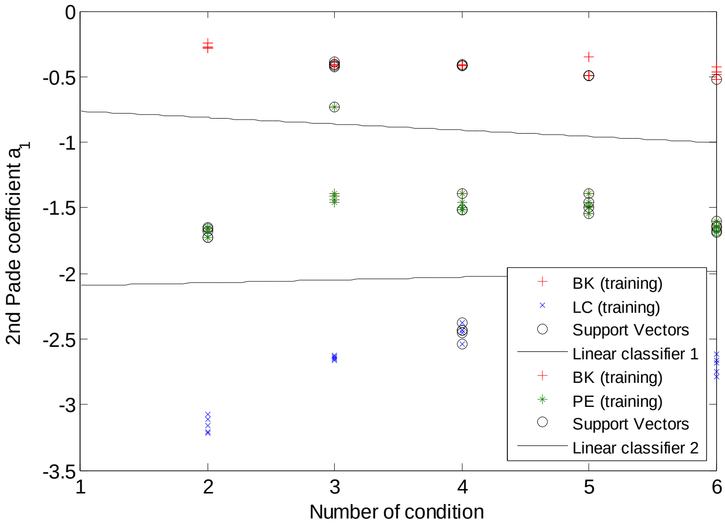


Figure 5.26 Linear classification of under-sampled 3rd order polynomial features

Table 5.11 The accuracy in condition classification attained with linear classifiers and the 3rd order polynomial features

Defect/ Feature	a_1 (lowest order)	a_2	a_3 (highest order)
Pipe End (PE)	100%	100%	100%
Blockage (BK)	83.3%	91.76%	83.3%
Lateral Connection (LC)	100%	100%	100%





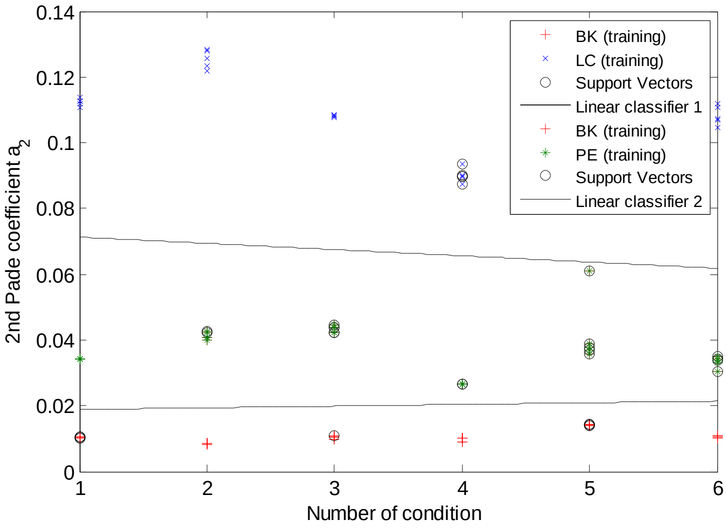


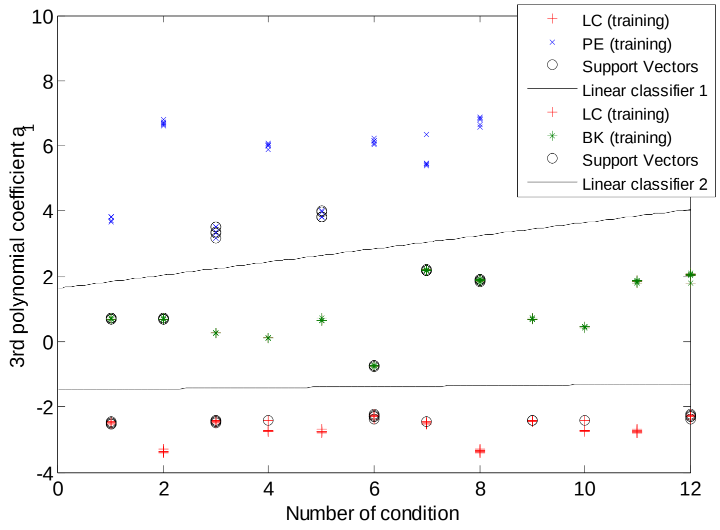
Figure 5.27 Linear classification of under-sampled 2nd order Padé approximation features

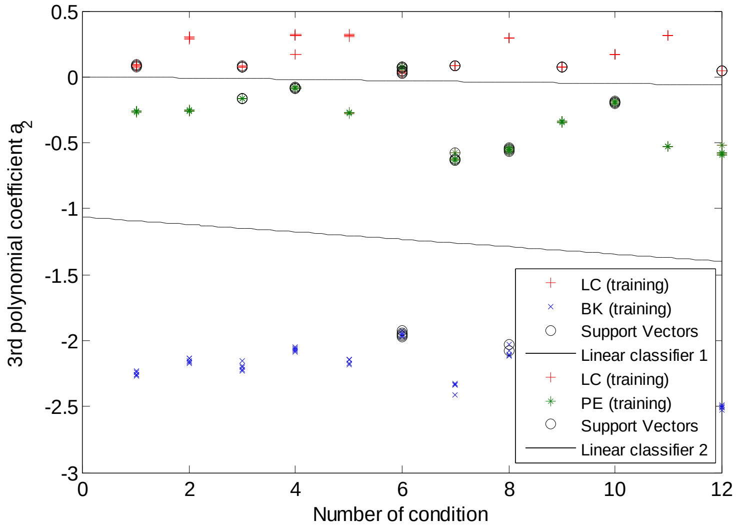
Table 5.12 Linear classification accuracy of using under-sampled 2nd order Padé approximation features

Defect/ Feature	a_0 (lowest order)	a_1	a_2 (highest order)
Pipe End (PE)	83.3%	75%	83.3%
Blockage (BK)	100%	100%	91.67%
Lateral Connection (LC)	100%	100%	100%

Over-Sampled Features Classification

Over-sampling is normally applied to minority class to create artificial samples so that these can be added to the class until it contains the same number of samples as the other classes used in the classification process. The LC class which was used in this work contained only half number of the samples in comparison to that in the other two classes, SMOTE was applied to those samples which were picked out randomly from each of the LC condition signature so that new samples could be generated based on the similarities between the existing samples in feature space. Linear classifiers were trained by using the 3rd order polynomial and 2nd order Padé approximation features which are shown in Figure 5.28 and Figure 5.29. Table 5.13 and Table 5.14 summarize the classification accuracy attained with the features that were used to train and test the system as a part of this classification process.





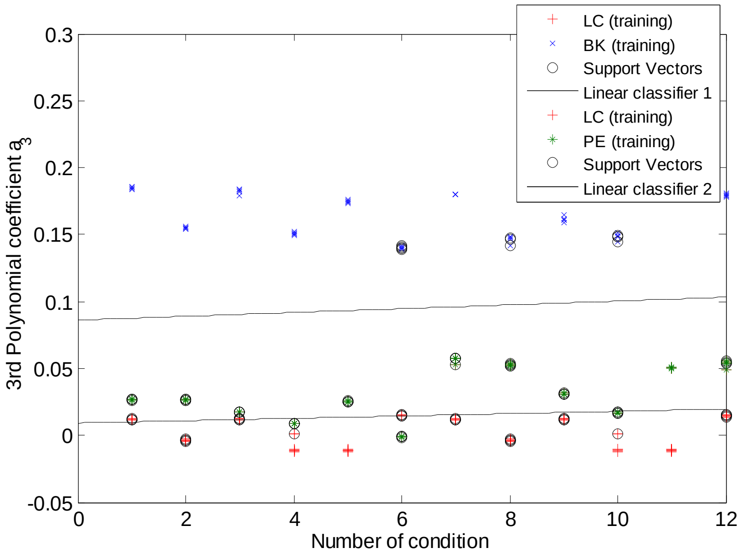


Figure 5.28 Linear classification of over-sampled 3rd order polynomial features

Table 5.13 Linear classification accuracy of using over-sampled 3rd Polynomial features

Defect/ Feature	a_1 (lowest order)	a_2	a_3 (highest order)
Pipe End (PE)	100%	83.3%	75%
Blockage (BK)	91.6%	100%	100%
Lateral Connection (LC)	100%	100%	83.3%

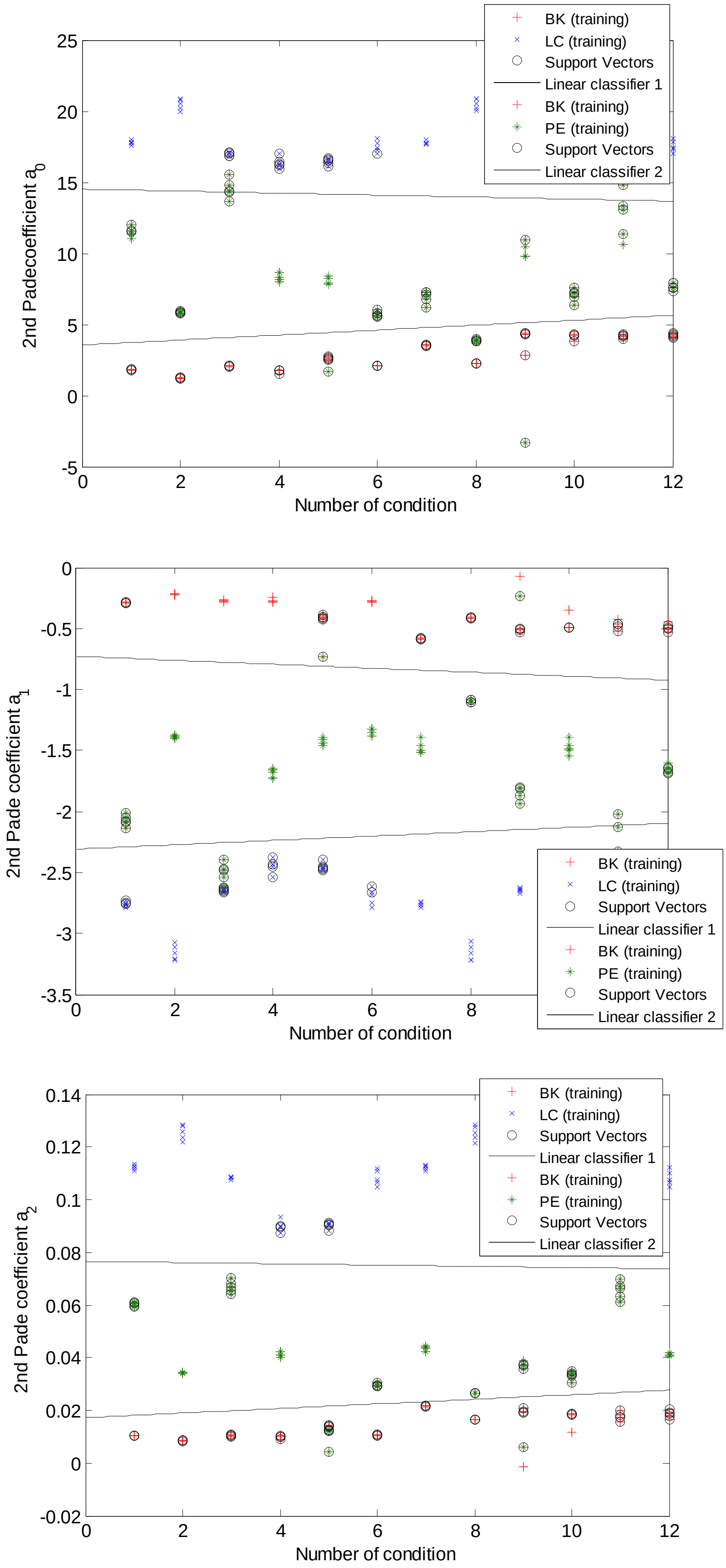


Figure 5.29 Linear classification of over-sampled 2nd order Padé approximation features

Table 5.14 Linear classification accuracy of using over-sampled 2nd Padé features

Defect/ Feature	a_0 (lowest order)	a_1	a_2 (highest order)
Pipe End (PE)	75%	79.1%	83.3%
Blockage (BK)	91.7%	100%	91.7%
Lateral Connection (LC)	100%	100%	100%

The classification results from using down-sampled and over-sampled features above did not show much difference, the classification accuracy of using the 3rd polynomial fitting coefficients and 2nd Padé approximation coefficients both varied between 75% and 100%. The system has the highest success rate of recognizing the lateral connection (LC) defect. The original feature distribution of all three defects were showing linearly separable with rather obvious distance between classes, although the original feature datasets were imbalanced, it did not cause noticeable bias for the linear classifiers, therefore, the performance of advanced re-sampling methods have achieved similar results. However, the problem of imbalanced learning is considered a relatively new challenge that has attracted growing attention from both academia and industry. Due to the inherent complex characteristics of imbalanced data sets, learning from such data requires various principles, algorithms and tools to represent the raw data efficiently so that a classification process can be performed effectively.

5.5 Summary

This chapter is concerned with condition classification analysis of pipe data which were collected under two main conditions: water level effects and multiple object effects. Acoustic signatures were extracted from intensity reflection of defects and used as original input of the classification system. Feature extraction methods and classification techniques were studied and compared.

- The energy spectrum of the acoustic intensity signature show distinguishable pattern of each defect under different pipe condition. Polynomial data fitting and Padé approximation methods were applied to extract pattern coefficients for further classification: less polynomial coefficients are required when the observation has a simple structure; Padé approximation provides better fitting when the observation has a more complicated shape.
- Support vector machines (SVMs) is a powerful state-of-the-art classification method. It belongs to the general category of Kernel methods. A SVMs classification system can be trained by linear or non-linear kernels based on the data distribution. It was shown by the results that the intensity energy spectrum fitting coefficients were linearly separable. Input data preparation, SVM and kernel parameters setting are influential to achieve better classification.

Acoustic properties of a partially filled pipe are proved to be useful in defect classification, appropriate signal pre-processing and feature extraction methods provide relevant features to train a system, support vector machines generate decision boundaries based on the training features and tested by other unlabelled features. Results have suggested that the classification system is capable of recognizing pipe defect under a range of conditions with reasonable accuracy rate.

References

[1]M. T. Bin Ali, *"Development of Acoustic Sensor and Signal Processing Technique", PhD thesis,* University of Bradford. 2010.

[2]Kent L. Gee et al., "Measurement and prediction of nonlinearity in outdoor propagation of periodic signals". *Journal of The Acoustic Society of America* , vol 120, Issue 5, pp 2491-2499. 2006.

[3] C. Brezinski , *"A Bibliography on Continued Fractions, PadeApproximations,Extrapolation*

and Related Subjects". Prensas Universitarias de Zaragosa. Zaragosa. 1991.

[4] M.Vajta, "Some Remarks on Pade-Approximations". *3rd TEMPUS-INTCOM Symposium*, pp 9-14. Veszprém, Hungary. 2000.

[5] V. N. Vapnik, *"The Nature of Statistical Learning Theory",* Springer-Verlag, New York. 1996.

[6] M. Sahare and H. Gupta, "A Review of Multi-Class Classification for Imbalanced Data", *International Journal of Advanced Computer Research* , vol 2, No.2, Issue 5. 2012.

[7] Haibo He and E. A. Garcia, "Learning from Imbalanced Data". *IEEE Transactions on Knowledge and data Engineering* , vol 21, No. 9,. 2009.

[8] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets". *International Journal of Emerging Technology and Advanced Engineering* , vol 2, Issue 4. 2012.

[9] A. Estabrooks et al., "A Multiple Resampling Method for Learning from Imbalanced Data Sets". *Computational Intelligence* , vol 20, number 1. 2004.

[10] J. Zhang and I. Mani, "KNN Approach to Unbalanced Data Distribution: A Case Study involving Inforamtion Extraction". *Workshop on Learning from Imbalanced Datasets II.* Washington DC. 2003.

[11] M.Kubat and S. Matwin, "Addressing the Curse of Imbalanced Training Sets: One-Sided Selection". *14th International Conference on Machine Learning*, pp. 179-186. 1997.

[12]Ian H. Witten et al. *"Data Mining: practical machine learning tools and techniques".* Morgan Kaufmann . 2011.

[13]N. V.Chawla et al., "SMOTE: Synthetic Minority Over-sampling Technique". *Journal of Artificial Intelligence Research* , vol 16, pp 321-357. 2002.

Chapter 6

Field Measurements Analysis

6.1 Introduction

This chapter describes the use of defect classification analysis using acoustic data obtained from different sites in Austria and Australia using a developed acoustic inspection system [1]. Acoustic data were collected from a wide range of pipe defects under field conditions. This data was used to validate the classification system developed as a part of this PhD study (see Chapter 5). The results of this work were used to improve the classification system.

An acoustic signature library was built containing defect signatures extracted using field data from one site in Austria and three sites in Sydney, Australia. Defects signatures include pipe end (PE), lateral connection (LC), displaced joint (DJ) and crack (CR). The number of signatures of cracks collected was much less than the number of other defects, therefore this data will not be added to the classification analysis to avoid biased results. Individual defects were identified and confirmed by CCTV data collected at the same time as the acoustic data.

The combination of conventional CCTV method and acoustic inspection into one device will be an innovative and powerful instrument for sewer inspection. The challenge for that goal is to develop both technologies and improve the recognition capacities for structural and operational sewer conditions. To achieve that aim the research has following objectives: use acoustic data to develop a library containing signatures of defects from the reflected signals; then to identify these signatures with the help of CCTV derived information; apply the data to the classification system by following the feature extraction, selection and classification steps and then finally to quantify the performance of the system using an independent set of field data.

6.2 Field Signatures Library

A live sewer pipe is never perfect. Structural and operational defects such as cracks, poor joints, pipe deformation and blockages are commonly contained in pipes. These defects cause acoustic reflection which can be extracted from the recorded intensity response of the pipe and analysed in terms of the spectral and temporal composition. Table 6.1 presents a list of pipe defects including: pipe end (PE), displaced joint (DJ), crack (CR) and lateral connection (LC). These defects were used to obtain signatures of pipes characteristics. These tests were carried out at (i) Oatlands, Sydney, (ii) Bushland, Sydney, (iii) Carlingford, Sydney and (iv) Anzbach-Laabental, Austria. Figure 6.1 gives some examples of CCTV images that helped to understand the pipe conditions and identify the individual defects. The further analysis used the signatures obtained from this data library and the diameter of all these pipes is 300mm.

Table 6.1 Field Acoustic Data Library

Data file ID	Pipe Material	Pipe length	Defects	Site location
1376358_1376563	Vitrified clay	79.70	CR, PE,DJ	Oatlands, Sydney
1290939_1290935	Vitrified clay	63.93	PE	Bushland, Sydney
1184881_1187589	Cast Iron	26.83	LC	Carlingford, Sydney
1187609_1184877	Vitrified clay	1.96	CR	Carlingford, Sydney
158_157_M6	Concrete	33.1	LC, PE	Anzbach-Laabental, Austria
159_158_M6	Concrete	32.7	LC, PE	Anzbach-Laabental, Austria
160_159_M6	Concrete	37.1	LC, PE	Anzbach-Laabental, Austria
19_18_M7	Concrete	8.3	PE	Anzbach-Laabental, Austria
20_19_M7	Concrete	22.6	LC, PE	Anzbach-Laabental, Austria
23_22_M7	Concrete	34.9	DJ, LC, PE	Anzbach-Laabental, Austria
23_24_M7	Concrete	34.6	DJ, LC	Anzbach-Laabental, Austria
24_25_M7	Concrete	26.7	LC	Anzbach-Laabental, Austria
31_30_M7	Concrete	22.3	LC	Anzbach-Laabental, Austria
32_31_M7	Concrete	34.3	DJ, LC, PE	Anzbach-Laabental, Austria
33_32_M7	Concrete	38.6	LC, PE	Anzbach-Laabental, Austria
41_40_M7	Concrete	36.7	DJ, LC, PE	Anzbach-Laabental, Austria
32_31_M8	Concrete	40.8	LC, PE	Anzbach-Laabental, Austria
36_37_M8	Concrete	34.9	PE	Anzbach-Laabental, Austria
41_42_M8	Concrete	26.6	PE	Anzbach-Laabental, Austria
44_45_M8	Concrete	25.2	LC, PE	Anzbach-Laabental, Austria
46_47_M8	Concrete	35.8	DJ, LC, PE	Anzbach-Laabental, Austria
50_33_M8	Concrete	33.7	LC, PE	Anzbach-Laabental, Austria
109_108_M9	PVC	51.3	DJ, LC, PE	Anzbach-Laabental, Austria
167_166_M9	Concrete	37.5	DJ, LC, PE	Anzbach-Laabental, Austria
177_176_M9	Concrete	35.1	DJ, LC	Anzbach-Laabental, Austria
30_31_M10	Concrete	36.2	DJ, LC	Anzbach-Laabental, Austria
31_32_M10	Concrete	32.7	LC, PE	Anzbach-Laabental, Austria
41_40_M10	Concrete	33.0	LC, PE	Anzbach-Laabental, Austria

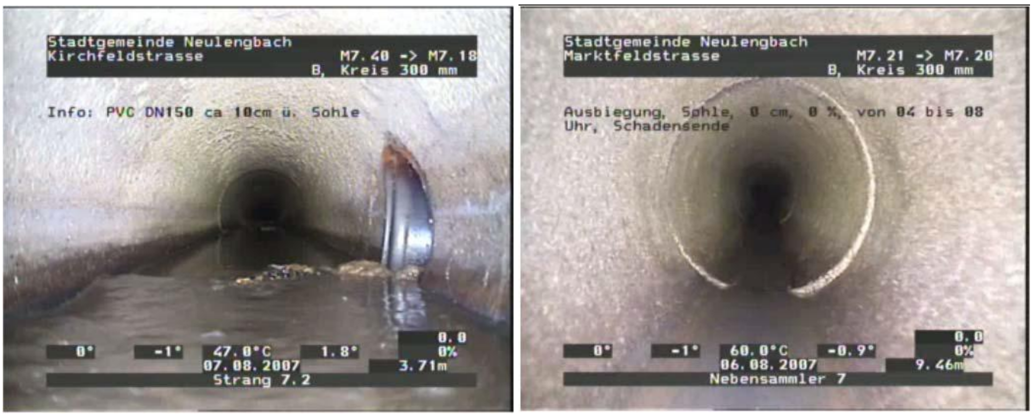


Figure 6.1 Field CCTV images of an open saddle connection (left) and a displaced joint (right).

6.3 Feature Extraction and Preparation

The acoustic intensity signatures of pipe ends, lateral connections and displaced joints were obtained in 20 frequency bands in the range from 100 to 800Hzas listed below:

Frequency band 1	100~150 Hz
Frequency band 2	129~279Hz
Frequency band 3	158~308Hz
Frequency band 4	187~337Hz
Frequency band 5	216~366Hz
Frequency band 6	245~395Hz
Frequency band 7	274~424Hz
Frequency band 8	303~453Hz
Frequency band 9	332-482Hz
Frequency band 10	361-511Hz
Frequency band 11	389-539Hz
Frequency band 12	418-568Hz
Frequency band 13	447-597Hz
Frequency band 14	476-626Hz
Frequency band 15	505-655Hz
Frequency band 16	534-684Hz
Frequency band 17	563-713Hz
Frequency band 18	592-742Hz
Frequency band 19	621-771Hz
Frequency band 20	650-800Hz

The intensity energy (using Equation 5.2) were calculated and plotted as a function of frequency band, Figure 6.2, 6.3 and 6.4 are examples of pipe end, lateral connection and displaced joint signatures, respectively, and their corresponding energy plots. It was noticed that the pattern of reflected energy does not depend significantly on pipe material but on the type of defect.

As discussed in Chapter 5, two data fitting techniques: least squares polynomial fitting and Padé approximation are available to derive the pattern characteristics. Empirically, 3rd order polynomial and 2nd order Padé approximation are able to capture the essence of the pattern with minimum number of coefficients. The numbers of signatures of each defect are not equal which caused imbalanced datasets for condition classification, the concept has been studied in Chapter 5 and re-sampling was proved to be an effective solution to this issue.

The lateral connection class of signatures (LC) contains 44 signatures and it is the largest class. The pipe end class (PE) contains 22 signatures and the displaced joint class (DJ) is the smallest class with 14 signatures. In order to make 3 classes equal in size, there are three options: (i) over-sample PE and DJ classes until they contain the same number of samples as LC class; (ii) down-sample LC and PE classes until their size are the same as DJ class; (iii) a combination of both over-sampling and down-sampling at different rates, set middle size PE class at re-sampling rate 100%.

Table 6.2 gives the size of 3 classes adopting different re-sampling schemes. The detailed analysis of re-sampling will be given in the following section.

Table 6.2 Size of the classes corresponding to different re-sampling scheme

Number/Signatures	Pipe End (PE)	Lateral Connection(LC)	Displaced Joint (DJ)
Original	22	44	14
Over-sampled	44	44	44
Down-sampled	14	14	14
Hybrid	22	22	22

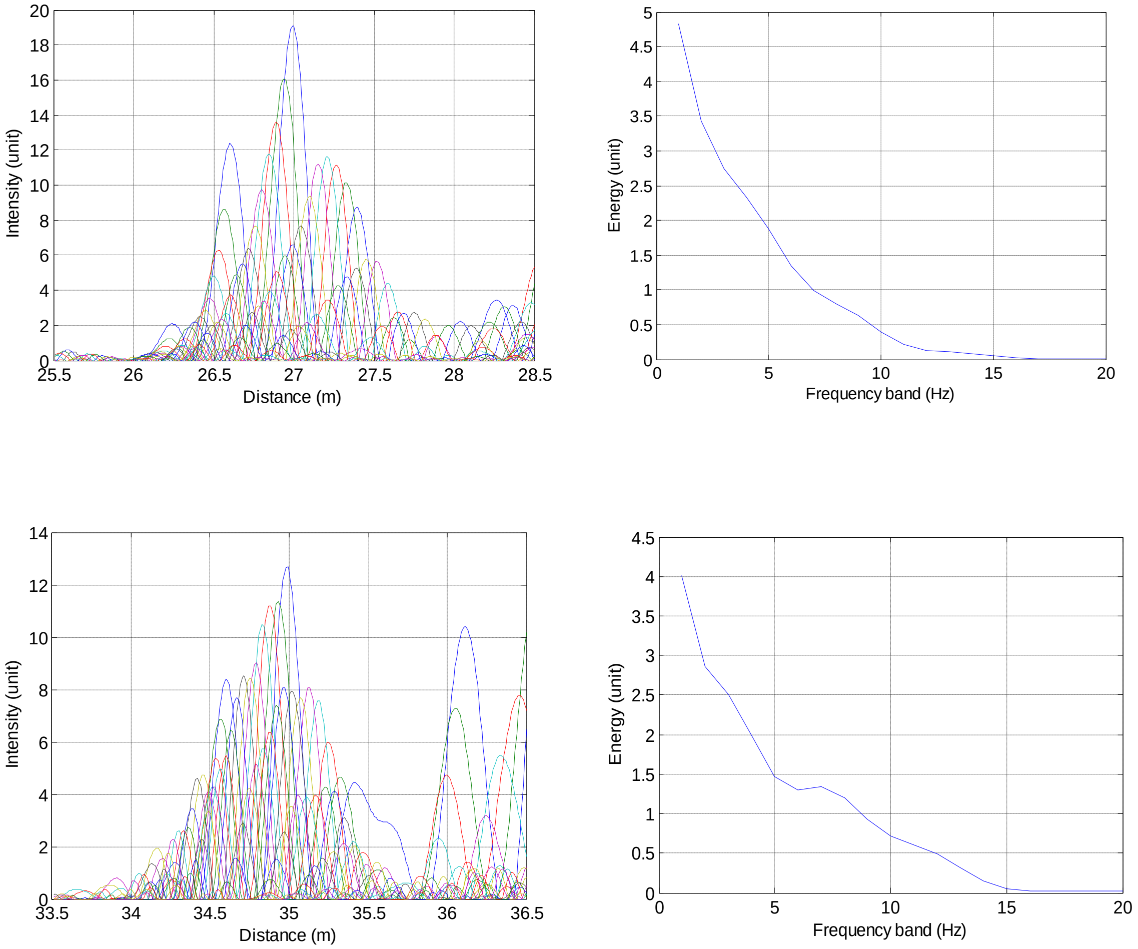


Figure 6.2 Two pipe end signatures obtained from different sites (left) in 20 frequency bands and their corresponding energy spectrum (right)

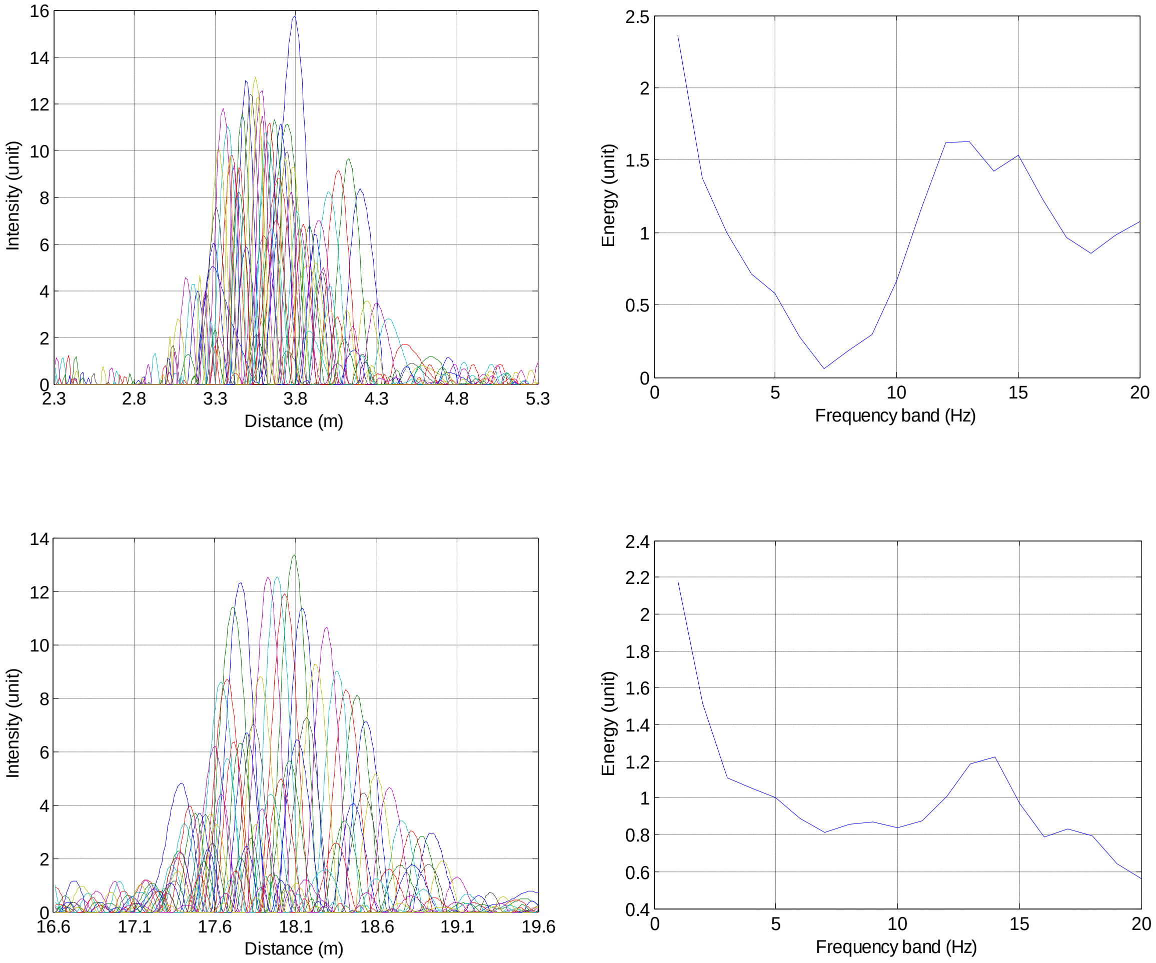


Figure 6.3 Two lateral connection signatures obtained from different sites (left) in 20 frequency bandsand their corresponding energy spectrum (right)

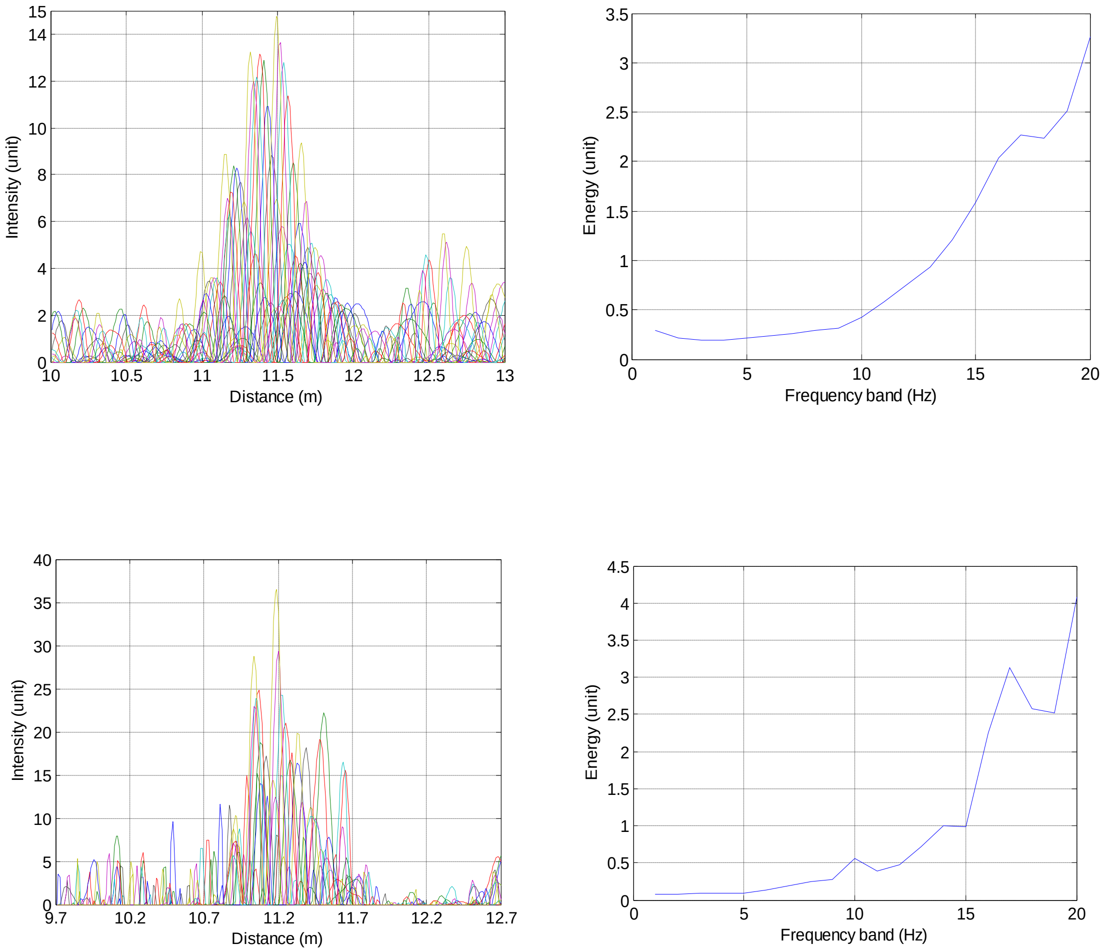


Figure 6.4 Two displaced joint signatures obtained from different sites (left) in 20 frequency bandsand their corresponding energy spectrum (right)

6.3.1 Data Fitting

Polynomial fitting and Padé approximation each has its own advantages on different occasions. Generally, low-order polynomial fitting works better on data curves which have more regular shapes, while Padé approximation can provide more accurate fitting for curves with more complicated structures. Given the fact that the pipe conditions in the field are more complicated and unpredictable, the reflected energy distributions can be in many different shapes, therefore, both fitting techniques will be adopted and their results will be compared.

3rd order polynomial fitting produces 4 coefficients and 2nd order Padé approximation 5 coefficients, in order to only keep the coefficients which are useful for pattern classification, each coefficient's distribution was studied. Figure 6.5 and 6.6 are 3rd order polynomial fitting coefficients and 2nd order Padé coefficients obtained from intensity signatures in the field data library, respectively. Both figures are showing poor clarity among different defects, therefore, the mean and the standard deviation σ of each coefficient set were calculated to measure how spread out the samples are within the class, and how much overlapping between classes caused. Table 6.3 lists the mean and the standard deviation σ values of each set of fitting coefficients obtained by 3rd polynomial and 2nd Padé approximation for all three defects.

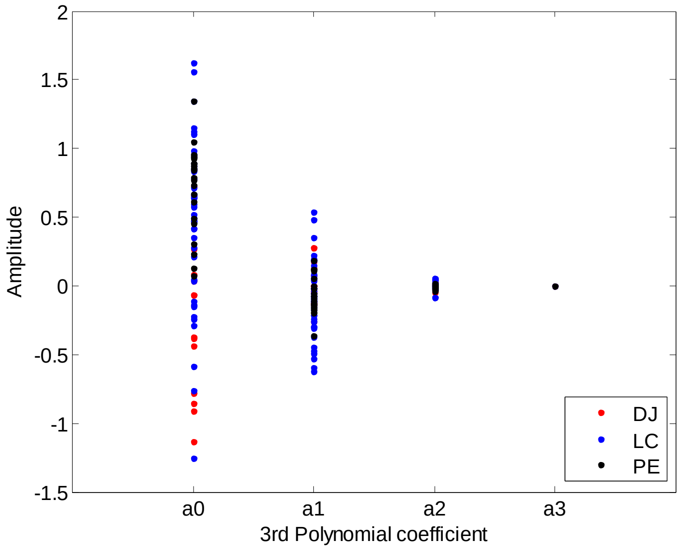


Figure 6.5 All four coefficient sets of 3rd order polynomial fitting obtained from original acoustic intensity signatures.

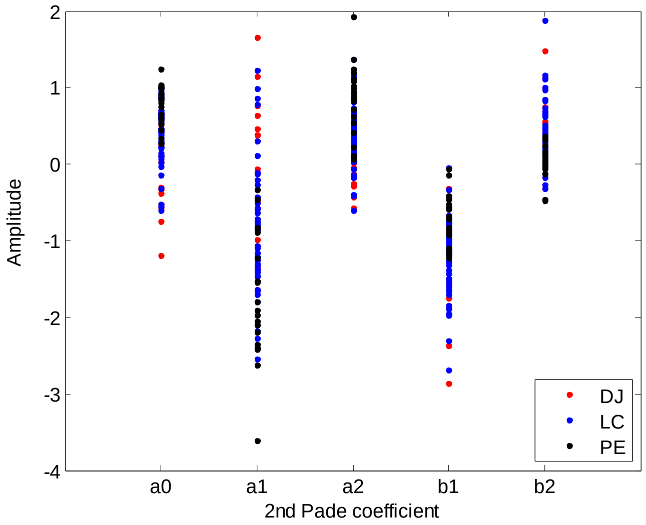


Figure 6.6 All five coefficient sets of 2nd order Padé approximations obtained from the original acoustic intensity signatures

Table 6.3 Standard deviation and mean of each coefficient sets obtained from defect signature energy fitting by 3rd order polynomial and 2nd order Padé approximation

Defect		**PE**		**LC**		**DJ**	
Data fitting coefficient		mean	σ	mean	σ	mean	σ
3rd order polynomial fitting	a_0	0.7168	0.3178	0.4862	0.5932	-0.2309	0.5702
	a_1	-0.0752	0.1297	-0.0958	0.2605	0.0133	0.1579
	a_2	-0.0065	0.0128	0.0007	0.0304	-0.0010	0.0257
	a_3 (×10⁻⁵)	0.1968	0.3972	3.0292	0.9334	-1.8383	0.8838
2nd order Padé approximation	a_0	0.7889	0.2455	0.3890	0.3981	-0.1041	0.4571
	a_1	-1.8141	0.7955	-0.8492	0.8617	0.0555	0.8891
	a_2	0.8301	0.4576	0.3696	0.3995	0.0028	0.3492
	b_1	-0.7271	0.3089	-1.2673	0.5311	-1.4523	0.6277
	b_2	0.0485	0.2180	0.4459	0.4323	0.5291	0.4267

Polynomial coefficients of each class are showing severe overlapping in terms of their mean and standard deviation measurement. Figure 6.7 gives examples of how standard deviation can help to select coefficients for further defect classification. The mean values of the 3rd order polynomial coefficients a_0 for all three defect classes and their standard deviation are shown in the left figure. It can be seen that most samples of PE class were included in the LC class which will make it very difficult to separate these two classes, the other three polynomial coefficients have shown the similar issue. Therefore, polynomial coefficients will not be used for the field data defect analysis. The right figure in Figure 6.7 gives the mean and standard deviation values of the 2nd order Padé approximation coefficient a_0 for all three defects. Although there is some overlapping between classes, it does show the possibility of separating them with suitable data cleaning techniques and classification algorithms. Padé numerator coefficients a_0 , a_1 and a_2 are chosen as input features for defect classification.

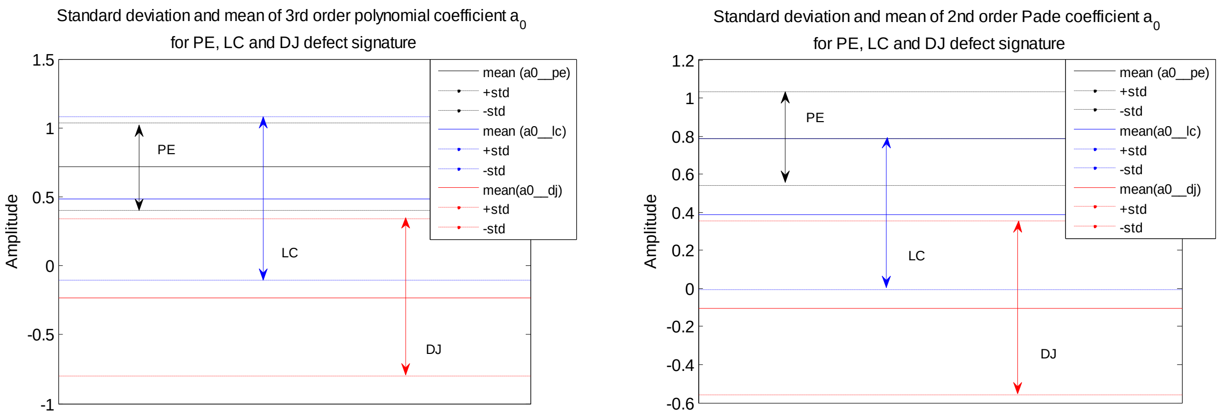


Figure 6.7 Examples of mean and standard deviation obtained from fitting coefficient sets for all threedefects: 3rd polynomial coefficient a_0 (left) and 2nd Padé approximation coefficient a_0 (right).

6.3.2 Re-sampling

The effect of learning from imbalanced datasets has been discussed in Chapter 5 section 5.4.2.1. Down-sampling the majority class or over-sampling the minority class are straight forward solutions for such a binary learning problem. Multiclass imbalanced learning can also adopt solo re-sampling to either remain the biggest class or the smallest class, the combination of the over-sampling and down-sampling strategies can also be useful given the fact that the two approaches are both useful in the presence of imbalanced datasets and appear to learn concepts in different ways. In this section down-sampling, over-sampling and a combination of down and over sampling algorithms will all be applied on the original Padé features extracted using the 2nd order Padé approximation method.

6.3.2.1 Random Over-sampling and Down-sampling

The mechanics of random over-sampling follow naturally from its description: for a set of randomly selected examples from the minority class, augment the original set by replicating the selected examples and adding them to it. Random under-sampling removes samples randomly from the original data set in majority class. The limitations of random re-samplings are also obvious: removing samples randomly from the majority class may cause the classifier to miss important concepts pertaining to the majority class; in regards to over-sampling, multiple instances of certain samples may cause the classifier to become too specific and lead to over-fitting.

Figure 6.8 shows the data distribution of the original LC class and randomly over-sampled DJ and PE classes, each class contains 44 samples; Figure

6.9 gives the distribution of randomly over-sampled DJ class, the original PE class and randomly down-sampled LC class with each class has 22 samples; in Figure 6.10, each class contains 14 samples after PE and LC class have been random down-sampled until they contain the same number of samples as the original DJ class.

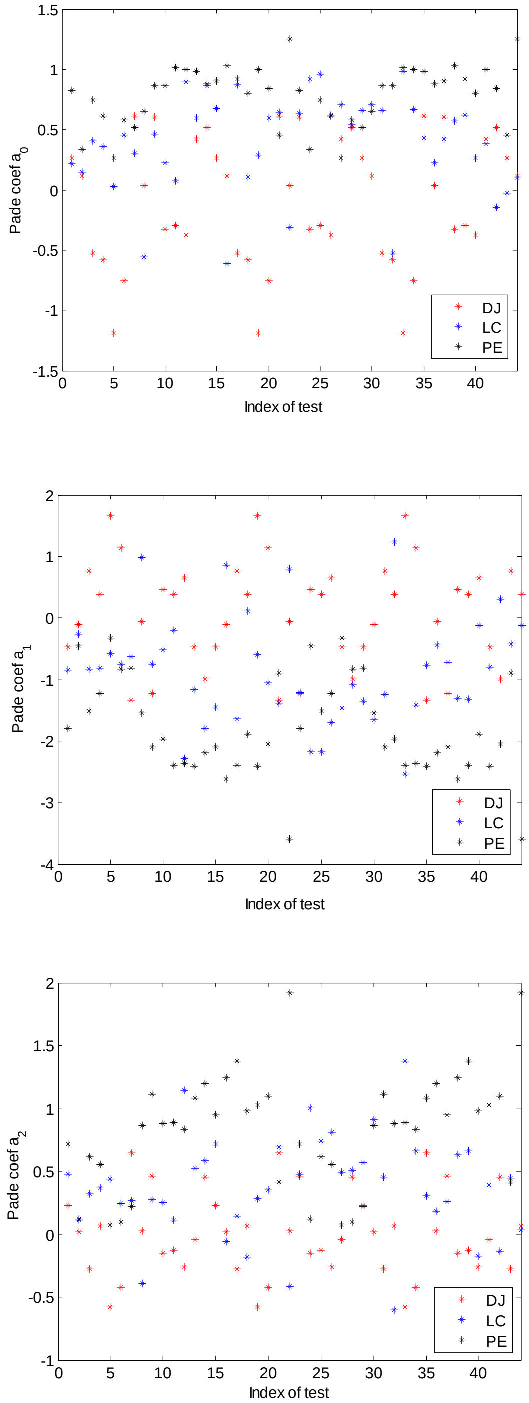


Figure 6.8 Original LC class, random over-sampled DJ and PE class, each classcontains 44 samples

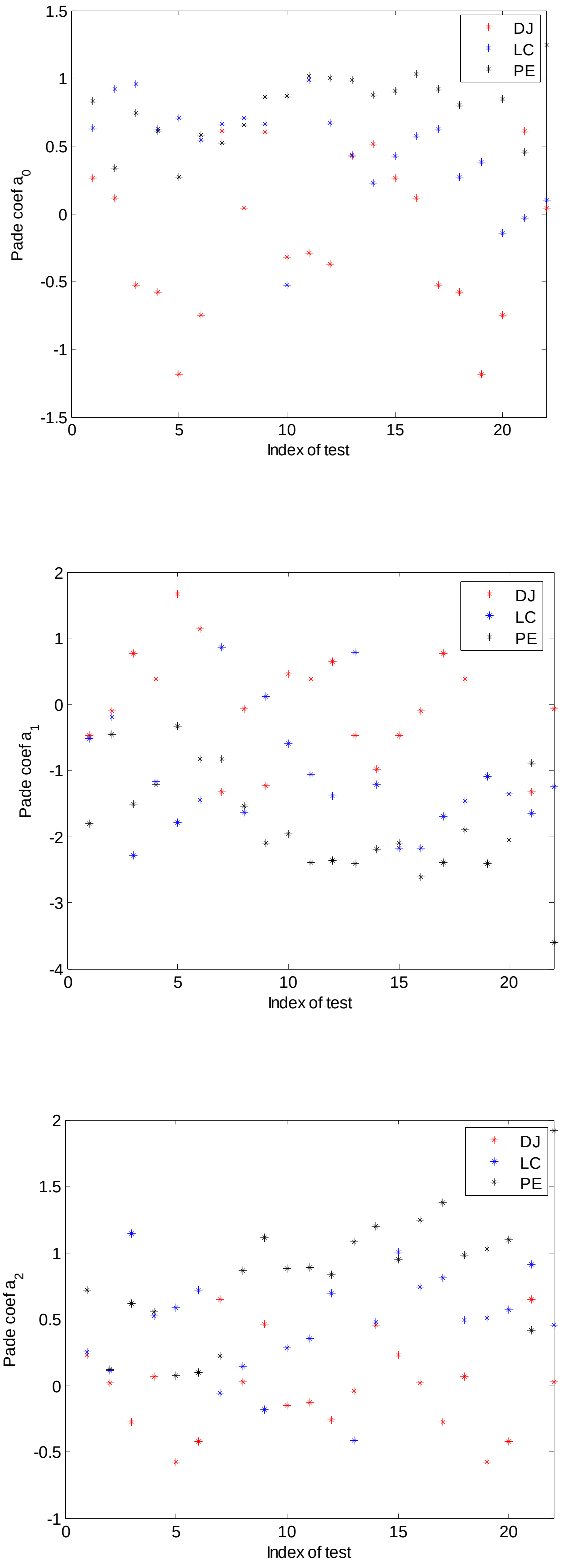


Figure 6.9 Random over-sampled DJ class, the original PE class and random down-sampled LC class, each class contains 22 samples

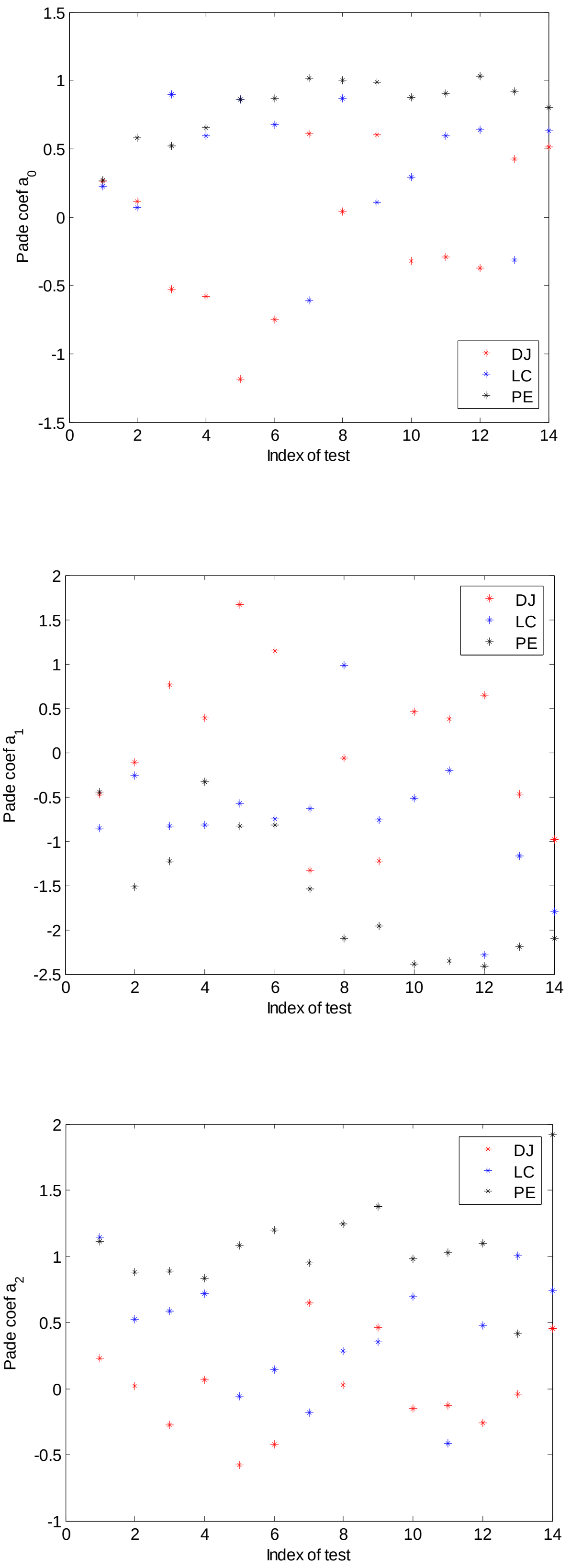


Figure 6.10 Original DJ class, random down-sampled LC and PE class, each class contains 14 samples

6.3.2.2 Nearest-Neighbours Weighted Re-sampling

One example of the many ideas which have been proposed in order to overcome the deficiency of information loss or duplication in the traditional random re-sampling methods is the KNN rule. It can be used to select the majority class samples whose average distances to the class border are the largest, so these samples are “safe” to be removed. It has also been introduced to improve an over-sampling algorithm, so the synthetic samples augment the original dataset in a manner that generally significantly improves learning. Weighted re-sampling algorithms also have their drawbacks including over generalization and variances [2], these algorithms produce more well-defined decision regions which potentially could fail to recognize informative samples brought in by independent datasets (testing datasets).

Based on the characteristics of the original data distribution, a combination of KNN weighted over-sampling and down-sampling scheme is adopted to reshape the original datasets aim to improve the classification performance. . The following procedure is adopted here:

(1) Set the size of PE class as the re-sampling target. It means that in this work the PE class retains its original number and the others are re-sampled until they reach the same size as the PE class;

(2) To reduce to half size the LC class, find Tomek links [3] of class LC & PE and class LC & DJ, and remove the LC samples participated in all Tomek links. Repeat the process until no more Tomek links could be found or the majority class reach the size it desires;

(3) The original DJ class contains 14 samples and it needs to be increased in size by adding artificial samples to match the target size. Find Tomek links in DJ class, apply SMOTE algorithm [4] on the samples except those participated in Tomek links to generate new samples to avoid more overlapping. The integer K depends on the number of Tomek links.

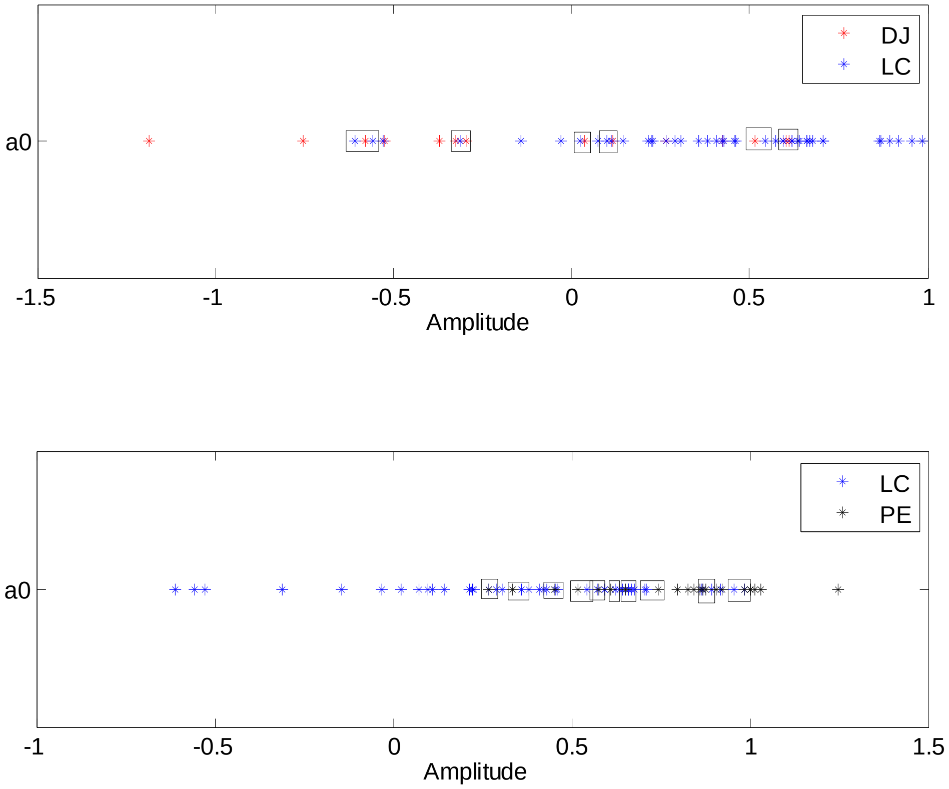


Figure 6.11 Tomek links (the squared sample pairs) found in DJ & LC classes (top) and in LC & PE classes (bottom).

Tomek links removal is a data cleaning technique of attempting to remove as much class label noise as possible, as well as borderline examples that have a higher probability of being incorrect, aiming to reduce overlapping and improve the accuracy of the data classification. The process is a simple modification to the K-NN algorithm, a Tomek link consists of a pair of samples that are each other's nearest neighbour but do not share the same class label.

LC class is the majority class in the learning which needs to remove 22 samples to reach the balance. Using Padé coefficient a_0 as an example as illustrated in Figure 6.11, 7 Tomek links were found within original DJ and LC class and 12 Tomek links within LC and PE class in the first process, LC samples participated in these Tomek links are removed. The second process identified 5 Tomek links within the DJ and LC class and 6 Tomek links within the LC and PE class, only 3 majority samples needed to be moved, the 3 LC samples whose average distance to their nearest neighbours are the smallest are also removed.

The original DJ class has 7 samples participated in Tomek links with samples from the LC class, the other 7 samples in LC class will be fed to the SMOTE algorithm to generate synthetic samples. This was achieved as follows: the four nearest neighbours of each sample were located, the reason of choosing K=4 is to have enough neighbours to generate new samples so that the minority class would be augmented as required. Then multiply the corresponding feature vector difference with a random number $\delta \in$ [0, 1], and the new synthetic sample is obtained by $x_{new} = x_i + (\hat{x}_i - x_i) \times \delta$, where x_i is a random sample in the class, $\hat{x}_i$ is one of the nearest neighbours of x_i in the class, δ is 0.5 here. The resulting new sample is a sample along the line segment joining x_i and one of its selected nearest neighbours. In this way, 17 new samples were generated and 15 of them are picked randomly and added to LC class, the augmented LC class now contains 22 samples as requested. Figure 6.12 illustrates the neighbour relations between the original DJ samples of Padé coefficient a_o and how new samples are located.

Figure 6.13 shows the Padé coefficients a_0, a_1 and a_2 distribution of data down-sampled by Tomek links and over-sampled by the SMOTE algorithm.

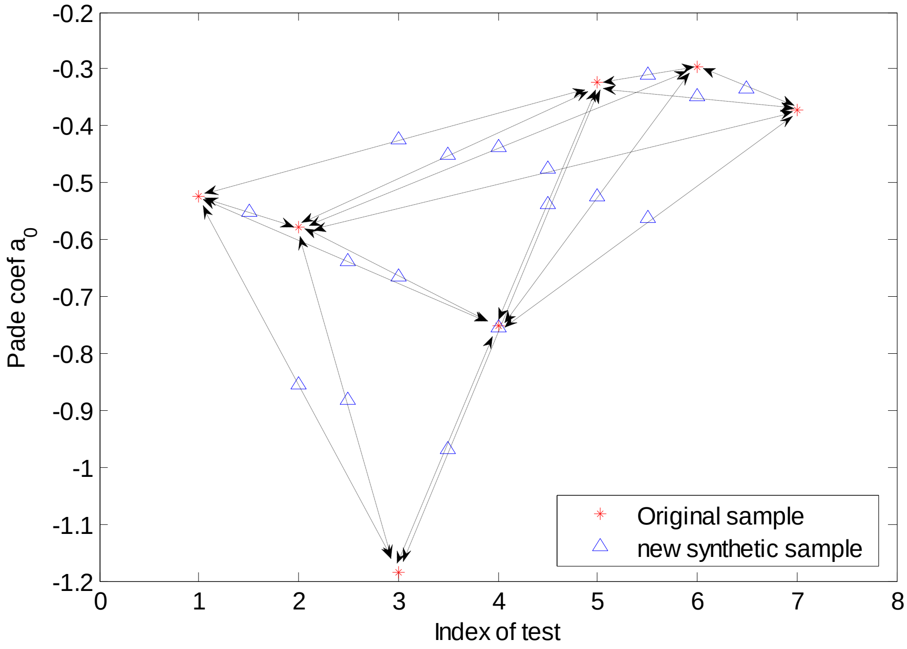


Figure 6.12 Data generation for DJ class using the SMOTE

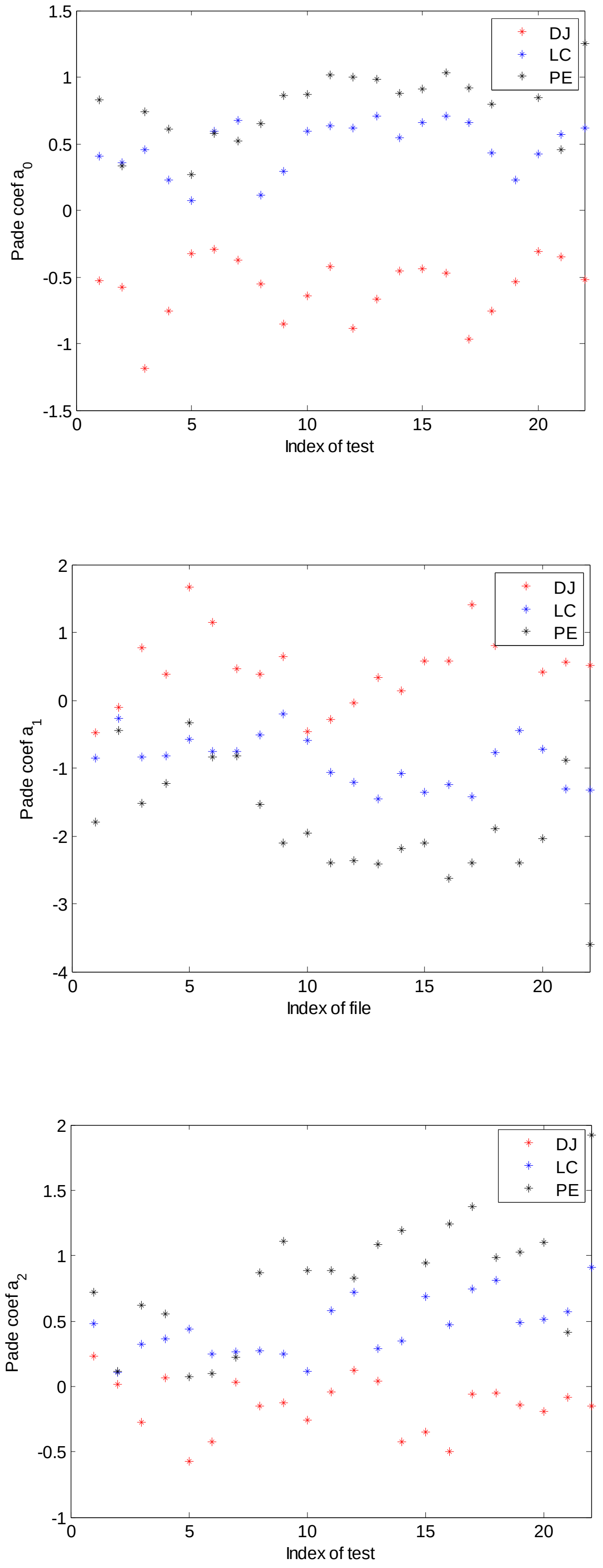


Figure 6.13 Weighted down-sampled LC class, original PE class and over-sampled DJ class by the SMOTE, each class contains 22 samples

6.4 Defect Classification

One dichotomy in statistical pattern recognition is that of supervised learning (labeled training samples) versus unsupervised learning (unlabeled training samples). The distinction is drawn from how the learner classifies data. Unsupervised algorithms seek out similarity between pieces of data in order to determine whether they can be characterized as forming a group, one famous approach is K-nearest neighbour algorithm which has been used in pervious chapters as classifier and re-sampling technique. In supervised learning, the classes are predetermined, each example is a pair consisting of an input object (typically a vector feature) and a desired output value (a supervisory signal), which distinguishes supervised learning from unsupervised learning, the learning designer's task is to search for patterns and construct mathematical models. A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. Among many supervised learning algorithms, Support vector machines (SVMs) is a state-of-the-art technique which seeks to map input vectors to a higher dimensional feature space so the represented or transformed samples can be separated.

Multiclass classification can be achieved by supervised learning methods, also some appropriate unsupervised learning schemes, and it depends on the organization of the input data clusters. Both KNN and SVMs will be applied to the re-sampled balanced datasets obtained in the previous section for defect condition classification, detailed process are given as follows.

6.4.1 Unsupervised Classification and Recognition: KNN

KNN is based on the use of distance measures, a KNN classification system does not require training a model to make a decision, and in other words, imbalanced datasets situation usually will not affect the decision if the number of nearest neighbours is limited not to exceed the size of the minority class. Each Padé coefficient corresponds to a dataset consisting of 14 DJ samples, 44 LC samples and 22 PE samples. Randomly pick 5 samples from each class to test if KNN algorithm is able to assign the correct class label to them with the effect of different value of K. Empirically, $K \cong \sqrt{N}$, N is the number of samples in training [5]. The decision rule of KNN is majority voting which approximates Bayes decision rule on K nearest neighbours of the testing sample awaiting to be assigned. The sample will be assigned to the class which has the highest number of samples in K nearest neighbours. Table 6.4 gives an example of how the recognition accuracy rate is obtained: T indicates correct recognition result, meaning the class which has the highest number of samples in 5 nearest neighbours of the test sample is in fact the right class that the test sample belongs to; F is short for False result and A represents ambiguous result, which suggests that there is tie between two classes as they contain the same number of samples in 5 nearest neighbours, indicating that KNN failed to make a decision.

Table 6.5 to Table 6.9 give the recognition accuracy rates when K was chosen to be close to the square root of the size of training dataset for three Padé coefficient datasets when they are original, random re-sampled and re-sampled using KNN weighted methods.

Table 6.4

Example of KNN recognition results in True-False form of original dataset when K=5

Test ID	DJ1	DJ2	DJ3	DJ4	DJ5	LC1	LC2	LC3	LC4	LC5	PE1	PE2	PE3	PE4	PE5
a_0	A	A	A	F	T	T	F	T	T	F	T	T	T	F	T
a_1	A	A	F	F	T	T	F	F	T	F	T	T	F	F	F
a_2	A	T	A	F	T	T	F	T	T	T	T	T	F	F	T

Table 6.5

Defect recognition using KNN algorithm with the effect of K value

Feature/K value	K=5	K=7	K=9
Padé a_0	53.33%	66.67%	26.67%
Padé a_1	33.33%	53.33%	46.67%
Padé a_2	60%	60%	60%

Table 6.6

KNN defect recognition using random down-sampled data with the effect

Feature/K value	K=3	K=5	K=7
Padé a_0	33.33%	26.67%	53.33%
Padé a_1	33.33%	46.67%	53.33%
Padé a_2	40%	66.67%	66.67%

Table 6.7

KNN defect recognition using random over-sampled data with the effect

Feature/K value	K=7	K=9	K=11
Padé a_0	60%	60%	66.67%
Padé a_1	66.67%	73.33%	60%
Padé a_2	73.33%	80%	73.33%

Table 6.8
KNN defect recognition using hybrid random re-sampled data

Feature/K value	K=5	K=7	K=9
Padé a_0	46.67%	60%	60%
Padé a_1	60%	66.67%	60%
Padé a_2	70%	73.33%	73.33%

Table 6.9
KNN defect recognition using weighted re-sampled data with the effect of

Feature/K value	K=5	K=7	K=9
Padé a_0	53.33%	60%	60%
Padé a_1	60%	60%	53.33%
Padé a_2	66.67%	60%	60%

6.4.2 Supervised Classification and Recognition: SVMs

Support vector machines (SVMs) as introduced in Chapter 3 and 5 is a supervised learning algorithm which has strong theoretical foundations and excellent empirical success in many pattern recognition and data mining applications. However, the performance of SVMs classifier greatly depends on the distribution of training dataset and can be biased if the dataset is imbalanced. Therefore, re-sampling training data to a full balance is necessary for SVMs classification. Re-sampled datasets obtained in section 6.3.2 will be used to train several non-linear SVMs classifiers and their performances will be presented in this section.

In KNN classification, each Pade coefficient’s corresponding training dataset are samples located along the line as shown in Figure 6.6, KNN algorithm used the absolute distance between samples to decide nearest neighbours. That data distribution however is not for SVMs learning as it does not work based on distance measures, therefore, the Padé features are organized as a function of the index of tests as seen in Figure 6.9 to 6.12. What SVMs need to do is to seek a suitable kernel functions to generate separating hyperplanes and class borders for each Padé dataset.

Clearly the data are not linearly separable by inspection, polynomial, RBF kernel and quadratic kernel are employed with the default margin control parameter C (see Chapter 3, section 3.3.2.3) in all runs for each dataset, so the kernels can be compared without the influence of SVM parameter. Each Padé dataset contains 44 samples after random over-sampled, as in KNN classification, 5 samples were picked randomly from each dataset for testing, and the rest of the samples were for training the classifiers. Datasets obtained from other re-sampling methods are being processed in the same way. Table 6.10 and Figure 6.14 display the classification accuracy rates of 3 kernel classifiers trained by 4 sets of re-sampled balanced data.

In general, Tomek links and SMOTE scheme provided better refined training samples for SVMs classifiers than random re-sampling. Classifiers producing smoother decision boundaries achieved more accurate classification due to the inherent characteristics of the original field datasets. Among three Padé coefficients, a_2 is the coefficient of the highest order variable of the defect pattern fitting structure, and its corresponding training dataset received higher classification accuracy rates than the other two coefficients in most runs, which suggests that the highest order coefficient of an approximation function captures more representative information of the pattern in a set of samples.

Table 6.10 SVMs classification accuracy with effect of re-sampling methods and kernel classifiers

Re-sampling	Feature	Polynomial		RBF		Quadratic
		$r=3$	$r=5$	$\sigma=0.8$	$\sigma=2$	
Down sampling	a_0	46.67%	46.67%	40%	46.67%	40%
	a_1	46.67%	33.33%	53.33%	60%	53.33%
	a_2	73.33%	73.33%	80%	80%	80%
Over sampling	a_0	66.67%	46.67%	53.33%	60%	66.67%
	a_1	53.33%	33.33%	60%	66.67%	60%
	a_2	53.33%	60%	60%	60%	60%
Hybrid	a_0	53.33%	40%	53.33%	60%	53.33%
	a_1	53.33%	33.33%	46.67%	53.33%	46.67%
	a_2	46.67%	26.67%	53.33%	60%	53.33%
Tomek links + SMOTE	a_0	60%	60%	66.67%	73.33%	60%
	a_1	73.33%	66.67%	66.67%	73.33%	60%
	a_2	73.33%	73.33%	73.33%	80%	80%

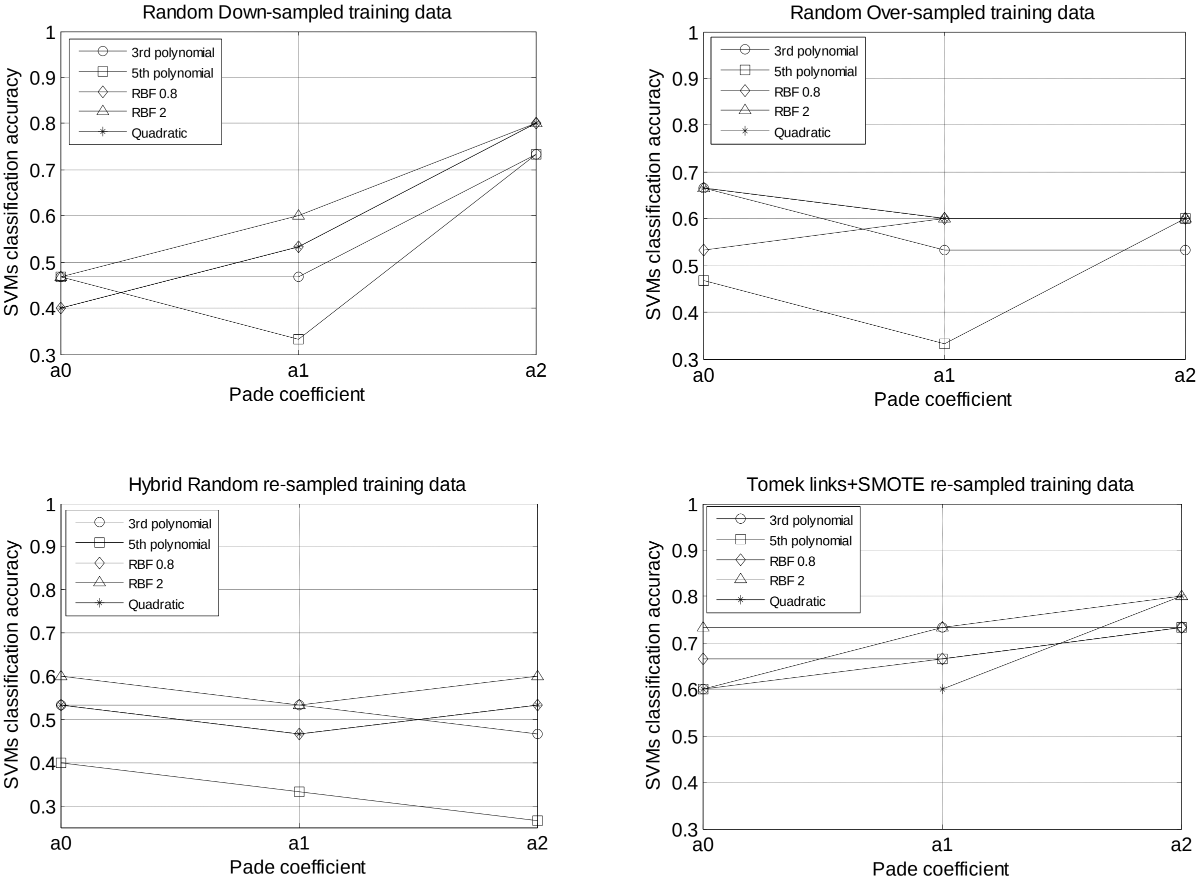


Figure 6.14 SVMs classification accuracy rates:

Random down-sampled training data (top left);

Random over-sampled training data (top right);

Hybrid Random re-sampled training data (bottom left);

Tomek links+SMOTE re-sampled training data (bottom right).

6.4.3 Classification Comparison

(i) Four Re-sampling methods were adopted to re-shape the original imbalanced datasets to reach a full balance: random down-sampling removed samples from majority class may cause the classifier to miss important concepts pertaining to the majority class; random over-sampling replicated data to the original dataset of certain existing minority samples, which narrowed the neighbourhood boundaries and probably cause overfitting. Tomek links and SMOTE were used to re-size the original datasets to a full balance in an informed manner, the training samples would be more well-defined and regularized in this scenario and that generally significantly improves learning. Although the training accuracy will be high in this scenario, the classification performance could be unsatisfied on the independent testing data which have high similarities with the samples were considered 'unwanted' during the re-sampling.

(ii) In K-nearest neighbours classification, the value of K was chosen to be odd numbers close to the square root of the size of training dataset, the classification performance is influenced by the value of K : bigger K results in better classification generally, but it could also leads to bias, a proper choice of K depends on the data and heuristics. KNN is instance-based unsupervised leaning algorithm which doesn't require prior-determined parameter, the imbalanced data distribution would not affect its performance greatly as the other learning methods if the value of K does not exceed the size of the 'most' minority class.

(iii) In Support vector machines learning, a well balanced training data distribution is necessary to avoid biased classification. 3 non-linear classifiers were used to generate separating borders, 4 sets of re-sampled dataset were used to train these classifiers. The classification accuracy rates suggest that: the parameters of kernel function determined the radius of decision boundaries, for more linear alike data distribution in this chapter, parameters produce smoother separating borders define the class regions better hence the classification accuracy rates are higher; various re-sampling methods have been proposed to improve the learning accuracy, Tomek links and SMOTE scheme that used in this chapter is based on the Nearest Neighbours rule and is proved to overcome the inefficiency introduced in the traditional random re-sampling to some extent.

6.5 Summary

In this chapter data collected in the field were used to develop and modify a classification system for pipe conditions and defects. Learning performance of a classifier depends greatly on the distribution of training dataset. Most of learning algorithms assume or expect an equal distribution between classes, but in practice, many applications are facing the problem learning from imbalanced data which is a relatively new challenge for both academia and industry. Various re-sampling methods have been proposed as possible solutions to imbalanced learning, random down-sampling, random over-sampling and a scheme combined Tomek links [3] and SMOTE algorithm [4] have been employed on field data to provide well-balanced training dataset for defect classification.

The choice of classification algorithm depends highly on the nature of the dataset and the type of feature extracted from the data, there is no algorithm that is best for all applications, it is helpful to test multiple algorithms and parameter settings. K-nearest neighbours (KNN) and Support vector machines (SVMs) are used to classify and identify pipe defects based on acoustic characteristics, random down-sampling was found to be the least efficient method to imbalanced learning for both algorithms, random over-sampling was proved to be more effective in KNN learning that SVMs learning because of the unsupervised nature of KNN. SVMs performances well on well-defined training data, informed re-sampling methods such as

Tomek links and SMOTE refine the clusters borders, as a result the SVMs learning is improved.

The first limitation of this system at current stage is that the performance is determined by the quality of the data source, coefficient features are only reliable for classification if extracted from independent acoustic reflection signatures, overlapping with other signatures could make an obvious influence on the classification. The second challenge is the imbalanced learning problem, the numbers of different pipe defects in the real world are most likely to be unequal, although various re-sampling methods with different benefits have been proposed to imbalanced learning, the inherent characteristics of imbalanced data sets can be complex.

References

[1]M.T.Bin Ali, *"Development of Acoustic Sensor and Signal Processing Technique".* PhD Thesis, University of Bradford. 2010.

[2] Haibo He, "Learning from Imbalanced Data". *IEEE Transactions on Knowledge and data Engineering* , vol 21, No. 9. 2009.

[3] I. Tomek, "Two Modifications of CNN". *IEEE Trans. Systems, Man and Cybernetics* , vol 6, no. 11, pp. 769-772. 1976.

[4] N.V. Chawla et al. "SMOTE: Synthetic Minority Over-Sampling Technique". *Journal of Artificial Intelligence Research* , vol. 16, pp. 321-357. 2002.

[5] D. Coomans and D.L. Massart, "Alternative k-nearest neighbour rules in supervised pattern recognition: part 1.k-Nearest neighbour classification by using alternative voting rules". *Analytica Chimica Acta* , vol 136, pp.15-27. 1982.

Chapter 7

Conclusions and Recommendations for Future Work

7.1 Results

One pipe and one siphon structure were used to simulate a range of common conditions which relate to pipe blockage and structural damage.A series of experiments were carried out to collect acoustic data for defect and condition analysis. 5 sets of siphon conditions were studied based on the sound pressure level data. It was found that: (1) the water level inside the siphon had a noticeable effect on the sound pressure particularly in the frequency range below 1000Hz; (2) the effect of air bubbles on the sound propagation in the siphon is relatively small and can be neglected; (3) the sound pressure level data suggest that the effect of different amount of surroundings on the acoustic field in the siphon is progressive, but limited.The greater thehorizontal area covered by dry sand, the lower the sound pressure level resulting in the siphon; and by corollary, higher surrounding water level caused higher interior sound pressure level; (4) the sound pressure level did not depend on the amount of sediment inside the siphon when the frequency was higher than 2500Hz, and the influence of the amount of sediment on the sound pressure level was mostly distinguishablefrom the early time of arrival signals; (5) the sound pressure level was found not sensitive to the wall damages at frequencies lower than 1000Hz and higher than 4000Hz. The time window is also critical to revealing the difference between different types of damage. Acoustic energy derived from the sound pressure level was shown to be useful for the condition analysis of the siphon.

The accuracy of sediment classification using proposed KNN classifier was 60%, which was improved by 20% using wavelet features than filter features on 5 sets of testing samples, the estimation probabilities were also improved between 6% and 14% approximately, therefore added more certainty to the decisions. Damage classification results were 100% for all 5 sets of testing samples using both wavelet and filter features, the estimation probabilities also shown that wavelet features generally led to more definite decisions than filter features.

Sewer pipe condition study focused on the recognition of several pipe defects based on the acoustic intensity signatures. The effect of water levels and multiple defect interaction was studied. The results show that: (1) the acoustic energy spectrum derived from intensity signatures of pipe end, blockage and lateral connection showed somewhat unique patterns which can be used to classify the types of defect; (2) these patterns were relatively unaffected by the various levels of water flow inside the pipe; (3) the presence of multiple defects can have an effect on the energy spectrum pattern of each defect, but this pattern remains distinct and distinguishable by suitable feature extractors and classifiers. The real live sewer conditions are far more complicated especially when the sizes of data collected from different types of defect are not equal, re-sampling as one solution to the problem was studied and applied to the field data. A combination of Tomek links and SMOTE was proven to be effective in re-shaping the original datasets into a better defined form which consequently improved the accuracy of classification.

Lab-simulated sewer defects classification adopting multi-class SVM classifiershave reached approximately 83% and 94% accuracy rate using polynomial approximation and Padé approximation as feature extractor, respectively. Defect samples collected in the field were more complicated and imbalanced between classes, re-sampling methods were applied to re-shape the samples to achieve a better classification performance. The combination of Tomek links and SMOTE as a hybrid re-sampling scheme resulted in approximately 70% accuracy rate of SVM classification and was by roughly 23% higher than random re-samplings. Only Padé features were used for field defects classification as polynomial features have shown too much ambiguity.

7.2 Conclusions

This thesis has been concerned with acoustic methods for classifying and identifying defects in siphon pipes and sewer pipes under various conditions. A pattern recognition system has been developed to recognize condition or defect by finding the relationship between the measured acoustic signal and another signal from a benchmark signature database in which acoustic signatures are assembled to represent a range of common conditions or defects in the pipe. Laboratory data and field data have been used to train the system and evaluate its ability to recognize defects. The signal processing techniques and statistical algorithms that we adopted in the system have been chosen based on the nature of the data distribution in a feature space.

In the case of the siphon structure, the sound pressure level history (SPL) has been determined from acoustic pressure response for a range of siphon conditions. The acoustic energy in the signal has been derived from the SPL data to be used as a main feature to distinguish between siphon conditions through a classification process. Digital filter and Discrete Wavelet Transform (DWT) have been used to filter the sound pressure level data, so that the acoustic energy filtered in a number of frequency bands and wavelet sub-band energy features obtained via the DWT have been the two sets of input characteristics for the classification system. The classifier adopted in the classification phase has been K-nearest neighbours (KNN) algorithm which measures the distance between instances and make the decision using majority vote which is based on the condition probability estimation. The results obtained from this work suggest that the acoustic energy changes noticeably when the siphon condition is changed. The system has been trained to estimate a range of the percentage of the siphon's cross-section which was occupied by the sediment. It has been shown that the use of wavelet energy features improve the classification performance, particularly when the wavelet entropy characteristic is added making effectively the feature space two-dimensional.

In the case of the sewer pipe structure, the classification analysis has been based on the direct and reflected instantaneous acoustic intensity which was generated by a point source inserted in the pipe and measured with a microphone array. A signature library has been built to contain acoustic intensity signatures for a variety of pipe defects which included: pipe end; blockage and lateral connection. The energy spectrum of a defect signature has been found as representative. Classification patterns can be extracted from these signatures using data fitting techniques. Polynomial fitting and Padé approximation methods have been applied to acoustic energy spectra to derive a finite number of coefficients from these patterns. These coefficients have been used as features for the defect classification analysis. It has been shown that the polynomial fitting method is easy to apply and it produces fewer coefficients than the Padé approximation method. However, Padé approximation method can interpolate better a wider range of spectral shapes especially those with complicated spectral characteristics. A state-of-the-art classifier named Support Vector Machines (SVMs) has been applied for machine learning and classification by considering the data distribution in the feature space. It is more suitable than the KNN method for a larger number of data points. SVMs adopt kernel functions to transform the data into a higher dimensional feature space where the represented features are separable.

For supervised classification techniques like SVMs, machine learning performance is likely to be affected by the distribution of the training datasets. If the training datasets are not equally distributed among classes, then the learning will cause misrepresentation of the data and perform poorly against the minority classes. In the real world applications, the defects mostly exhibit an unequal distribution between classes and that provide imbalanced datasets for machine learning and classification. The most popular solution to imbalanced learning is to re-sample the data. Random sampling and advanced sampling techniques have been applied and results have been compared against each other. Tomek links and SMOTE have been adopted as advanced sampling methods to remove or generate samples based on certain criteria to avoid possible information loss or redundancy.

Laboratory generated and field collected data from live pipes have been used to develop and modify a new condition classification system. The classification of field data has been less accurate than that in the case of laboratory data due to the complexity of conditions in real live pipes. Some types of defect can occur less often than the others in the real world and causing small sample size problem. In these circumstances, the combination of imbalanced data and small sample size can be very challenging to achieve robust condition classification. In this respect, multiple re-sampling methods with different rates might be more useful to improve the quality of classification. A combination of KNN weighted down-sampling and over-sampling techniques has been applied to field data to balance the datasets distribution. KNN and SVMs methods have been applied to the re-sampled training datasets and the results suggest that there is no single optimal strategy for all situations. The selection of sampling methods and classification algorithms are problem-dependent so that different techniques or their combinations should be employed.

7.3 Future Work

(1) The proposed classification system in this thesis is focused on using the acoustic energy extracted by different methods. This has been a main feature for condition and defect analysis and classification. It has proved to be possible to classify various conditions and defects accordance with the change in the energy spectrum provided a suitable classification technique is adopted. It can be recommended to study the performance of these classifiers using other acoustic characteristics rather than the energy. For example, these characteristics can include zero-crossing features, temporal and spectral structure of the signal, entropy and dynamism features and cepstrum coefficients. Some of these features have proved to be efficient in speech and music recognition. Signal phase data can be strongly affected by the type of defect and can be used in addition to the energy-based feature extraction methods.

(2) Limited types of pipe defects were studied including: pipe end, blockage, lateral connection and displaced joint due to the limitation of data. There are many other types of defects exist in the real live underground pipes, for example: wall crack, roots and change of wall thickness etc. More of these defects should be studied and their features should be stored to expand the signature library so that more detailed and accurate condition classifications can be expected.

(3) Feature extraction is the most critical phase in a classification system as it decides the form of the features that can be representative. Depending on the feature type desired, the feature extract methods are various. Despite the

digital filter, Discrete Wavelet Transform, Polynomial fitting and Padé approximation methods adopted in this thesis, there are other methods of time-frequency analysis and statistical modelling that have been used successfully in different pattern recognition tasks. A review of these methods has been given in chapter 2.

(4) The field data presented and analyzed in chapter 6 showed obvious overlapping between defect signatures due to the complexity of the environment found in live underground pipes. The SVMs classifier adopted in this thesis works better with well-defined classes of signatures. The classification results suggest that the dimension of the feature space needs to be optimised, so that the performance of different kernel functions could be explored. At the same time, other classifiers used for other purpose should be studied based on a better understanding of the nature of the problem and fundamentals of the classification algorithms.

(5) The imbalanced data learning problem needs to be revisited. Most real-world applications have somewhat imbalanced data problems and learning from imbalanced data can misrepresent characteristics of the data and cause misclassification. Although almost every algorithm presented in the literature claims to be able to improve classification accuracy over certain benchmarks, the fundamental question: *To what extent do imbalanced learning methods help with the learning capabilities, and is there a certain level of desired degree of balance for specific learning algorithms and application domains?* Still remain unanswered.

Appendix A: Matlab Programs

SoundLevelFltData.m	Filter recorded data and calculate sound pressure level of the chosen frequency band
IntensityResponse.m	Calculates acoustic intensity response from pressure impulse response
CalculateIntensity.m	Calculates acoustic intensity from pressure data recorded on a pair of microphones
mywavtree.m	Builds a discrete wavelet decomposition tree and filters input data through a set of wavelet filters
leastsquares_fitting.m	Fits a least-squares polynomial of chosen degree through a set of data, fitting coefficients obtained
padeApp_coef.m	Calculates Pad éapproximation coefficients
knnsearch.m	Finds k nearest neighbours of a chosen sample in a dataset and restores their index numbers in a array
nearestneighbour.com	Finds the nearest neighbour by Euclidean distance
KNNClassifier.m	Classifies a testing sample using a multiple class KNN classifier
svmtrain.m	Trains a support vector machine classifier for binary classification
svmclassify.m	Classifies one sample using a binary support vector machine
testSVMtwomodels.m	Classifies each sample from a dataset using a binary support vector machine classifier
multisvm.m	Trains a support vector machine classifier for multiple-class problem using One-Against-All rule.
testSVMmultimodels.m	Classifies a testing sample using a support vector machine trained by multiple class training datasets
smote.m	Finds k nearest neighbours of a certain sample and generates new artificial samples along the line between the sample and its nearest neighbours.

Appendix B: Publications

Z Feng, K V Horoshenkov and Jim Noras, “Acoustic characterization of theconditions of a water-filled siphon”. CM2011 / MFPT2011. June 2011, Cardiff, UK.

KVHoroshenkov, Z Feng, S J Tait, “Acoustic monitoring of the structural and operational conditions of a live underground siphon via matched field processing”. CM2011 / MFPT2011. June 2011, Cardiff, UK.

Zao Feng, Kirill V. Horoshenkov, M. Tareq Bin Ali, Simon J. Tait, “An acoustic method for condition classification in live sewer networks”. The 18th World Conference on Non Destructive Testing 2012. April 2012, Durban, South Africa.

Zao Feng, M. Tareq Bin Ali, Kirill. V. Horoshenkov, Simon Tait , “Application of KNN classifier for acoustic based pipecondition classification”. The 11th IEEE Sensors conference proceedings. October 2012, Taipei, Taiwan.

Acoustic Characterization of the Conditions of a Water-filled Siphon

Zao Feng, Kirill V. Horoshenkov and Jim Noras

Bradford, West Yorkshire, BD7 1DP, UK

z.feng2@bradford.ac.uk

Abstract

Pattern recognition has been used and developed as a process of advanced analysis of acoustic signal. Designing a robust pattern recognition system involves three fundamental tasks: signal pre-processing, feature extraction and selection, and finally classifier design and optimization. . This paper reports on an application to detect and monitor conditions of a large, water-filled siphon used in underground tunnels. Acoustic signals were collected from 4 hydrophones under various typical siphon conditions and used as input data to study the variation of the acoustic field. The discrete wavelet transform (DWT) was used in feature extraction and k-nearest neighbors (KNN) classification was applied. Subsequently, the system was tested on new unknown data and compared with supervised training samples. Results demonstrated that the acoustic sensors have high reproducibility for collecting signals under operational conditions. The pattern recognition system is also capable of discriminating different pipe conditions but further refinement is needed to improve sensitivity and to compensate for the effect of variable water level and sensor misalignment.

Keywords

Acoustics, siphon, wavelet transform, pattern recognition.

1. Introduction

Acoustics is used widely to determine the conditions of hidden assets, which include pipes, pumping stations and tunnels. It is popular because sound waves provide rapid, effective and non-invasive mean for asset quality control. Historically, Fourier transform-based spectral analysis methods have been used to analyse the collected acoustic data. These are based on time series data processing and calculating global energy-frequency distributions and power spectra. However, the use of Fourier spectral analysis is always limited to linear and stationary systems.In order to overcome these issues, methods of time-frequency analysis, including short-time Fourier transform (STFT), Wigner-Ville Distribution (WVD)(Debnath, 2002) and Wavelet Transform (WT) (Cohen, 1995), have been recently introduced.

In this analysis it is important to be able to determine patterns which are associated with particular system states. For many industrial applications, identifying and classifying patters and extracting features using time-series data constitute an important topic for research. In this research a subset of patterns which represents a range of typical conditions is of a particular interest. Feature extraction and pattern recognition algorithms have been developed and used for analysing signals and for signal classification (Hugo, 1999). These techniques include hidden Markov models (HMM), K-nearest neighbours (KNN)(Richard O.Duda, 2001), decision trees, and neural networks methods(Michael Cowling, 2003). Although these techniques found applications in areas related to voice and speech recognition, image analysis and security, they have not been used extensively for the condition monitoring of civil engineering assets. Therefore, this project concentrates on developing a new methodology for the analysis of acoustic datacollected in a hydraulic siphon.The aim of this project is to develop a robust classification technique to discover a relationship between the acoustic data and a range of classified patterns obtained for a full-scale model of a hydraulic siphon used in London Underground.

This paper is organized as follows: (i) the experimental procedure and data acquisition methods is described in section 2; (ii) section 3 presents a description of the wavelet analysis and K-nearest neighbors algorithm; (iii) the results are reported and discussed in Section 4 (iv) section 5 is the conclusion.

2. Experiments set up and Data collection

Acoustic data were collected in a siphon which was constructed from 450mm diameter concrete pipes in the Hydraulics Laboratory in the University of Bradford. The siphon was 4.2 m long and 2.0 m high. It was installed on a 500mm layer of fine sand in an open top box made of 12mm plywood. The siphon was instrumented with four 25mm hydrophones, 3 of which were installed in the left leg of the siphon. The other hydrophone was installed in the right leg of the siphon 75mm above the speaker and used as a reference receiver. The source was a 50mm diameter, water resistant speaker in a PVCenclosure which able to operate underwater. The hydrophones and the speaker were attached securely to two aluminum tubes which were lowered into the opposite legs of the siphon and kept at the same positions in all of the experiments conducted in the siphon. Figure 1 illustrates the equipment used in this experiment.The siphon was filled with clean water to the level of 900mm below the top of the right vertical pipe (reference water level) in all the experiments except water level test.

The data acquisition and signal processing facilities used in these experiments consisted of: (i) a PC with WinMLS software to control the sound card which generated a sinusoidal sweep in the frequency range of 100 - 6000 Hz; (ii) an 8-channel high-pass hydrophone filter used to remove unwanted low-frequency noise produced by equipment and machinery operated in the laboratory from the signals received on hydrophone H1-H3; (iii) a measuring amplifier and a filter which were used to condition and filter the signal received on the reference hydrophone in the 100 – 4000 Hz range. In addition, a power amplifier was used to drive the underwater speaker. Stereo amplifier and headphones were used to control subjectively the quality of the signal produced by the underwater speaker.

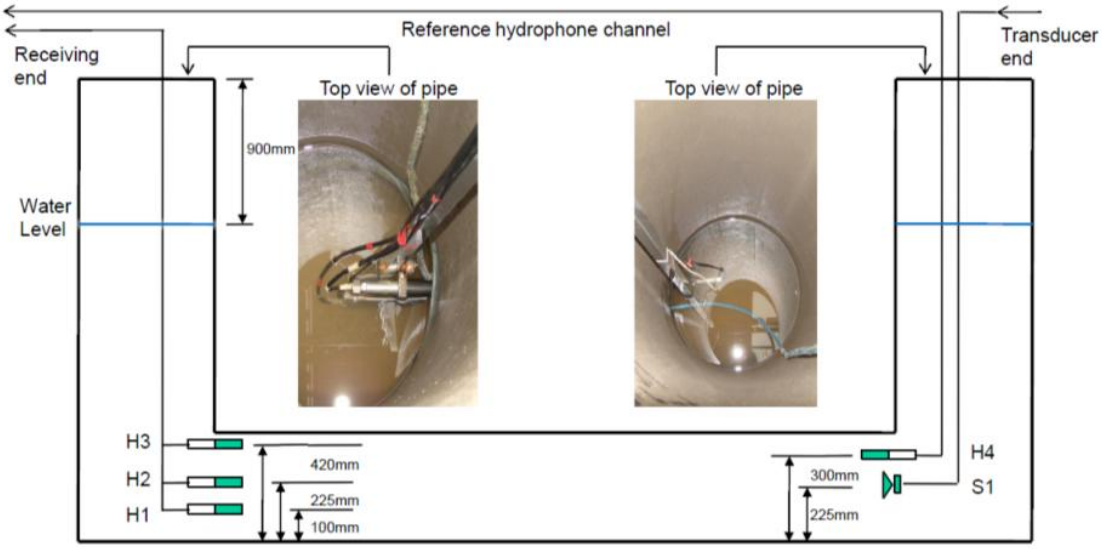


Figure 1 Structure of siphon and sensors

3. Signal processing methodology

For most industrial applications, a classical pattern recognition system consists following components: pre-processing, feature extraction, feature selection and pattern classification (decision making). Feature extraction and recognition methods are very important factors to achieve robustsystem performance. In this work we used the wavelet decomposition and *K*-nearest neighbors method to analyze the collected acoustic data and classify patterns.

3.1. Wavelet Decomposition

The wavelet transform (WT) is an important part of pre-processing and feature extraction phases in a pattern recognition system. It has been designed to analyze the temporal and spectral properties of non-stationary signals and overcomes the shortcomings of Fourier transform by applying adjustable window to achieve the required frequency and temporal resolution. Applications of 1-D discrete wavelet transform are numerous in acoustical signal processing(Christian U. Grosse, 2004)A discrete wavelet transform (DWT) decomposes a signal into mutually orthogonal set of wavelets. The signal to be analyzed is passed through filters constructed by a mother wavelet with different cut-off frequencies and at different scales. A discrete wavelet transform of a discrete time signal *f*(t) with length *N*and finite energy can be written as:

$$DWT(a,b) = \sum_{t=0}^{N-1} f(t) \frac{1}{\sqrt{a}} \psi^*(\frac{t-b}{a}) \qquad (1)$$

where $\frac{1}{\sqrt{a}} \psi(\frac{t-b}{a})$ defines the family of wavelet function, with $a \neq 0$ the scale of the transform and b the spatial (temporal) location, * denotes the complex conjugate.

The process of discrete wavelet transform implemented at each stage can be simplified as low-pass filtering of the signal for the approximations and high-pass filtering of the signal for the details, and then down sampling by half.Filtering a signal corresponds to the convolution of the signal with the impulse response of the filter. The output coefficients can be then expressed mathematically as:

$$y_{high}(k) = \sum_{k=-\infty}^{+\infty} x(n) g(2k-n) \qquad (2a)$$

$$y_{low}(k) = \sum_{k=-\infty}^{+\infty} x(n) h(2k-n) \qquad (2b)$$

where $x(n)$ is the original signal, $y_{high}(k)$ and $y_{low}(k)$ are the outputs of the high-pass filter G and low-pass filter H, respectively, after down sampling by half.

For many signals, it is the low-frequency componentswhich are mostly important. These components define the signal its identity. The wavelet decomposition process can be iterated, with successive approximations being decomposed in turn, so that one signal is broken down into many lower-resolution components.It is called the wavelet decomposition tree(Strang, 1996) as presented in Figure 2.

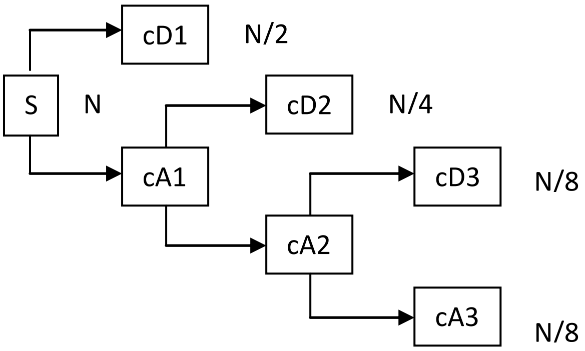


Figure 2 Wavelet decomposition tree.

The coefficients vectors cA and cD can be then used to reconstruct real filtered signals by reversing the decomposition process. The process yields reconstructed approximationsA_i, and details$D_i$which are true constituents of the original signal, so the original signal can be obtained by combining details and approximations, $S = A_i + D_i$.

3.2. K-nearest neighbors (KNN) method

K-nearest neighbors is a common classification technique which is based on the use of distance measures. For a given unlabeled sample x, find the k "closest" labeled samples in the training data set and assign x to the class that appears most frequently within the k-subset. k is the number of considered neighbors. Usually the Euclideandistance is used and it is expressed as:

$$d(x,p) = \sqrt{\sum_{i=1}^{n} |x_i - p_i|^2} \qquad (3)$$

where p is the training data set: $p = \{p_i\}$. A typical procedure for the KNN classification process is:

1) Calculate Euclidean distances of all training data to testingdata.
2) Construct a new matrix with elements are Euclidean distance between testing data and corresponding training data.
3) Pick K number of samples closest to the testing data by choosing K smallest values of Euclidean distance. Larger value of K yields smoother decision regions and, therefore results in a better classification. However, this increases computational burden as further samples are taken into account.
4) Classification: majority vote. K preferably odd to avoid ties.

4. Experimental conditions

The acoustic signals recorded in the siphon at two different conditions were decomposed by applying discrete wavelet transform. These conditions were: (i) clean siphon; (ii) siphon with a controlled amount of blockage. The blockage was simulated with bags of sand. Each of this bags contained approximately 1 kg of fine sand. A maximum of 10 bags were used in these experiments.

Signals with the frequency components higher than 5512Hz were filtered out and low-pass signals were decomposed into 8 frequency bands with each bandwidth equals to $\frac{f_s}{2^n}$, f_s =22050Hz is the sampling frequency, n is the depth of the decomposition.

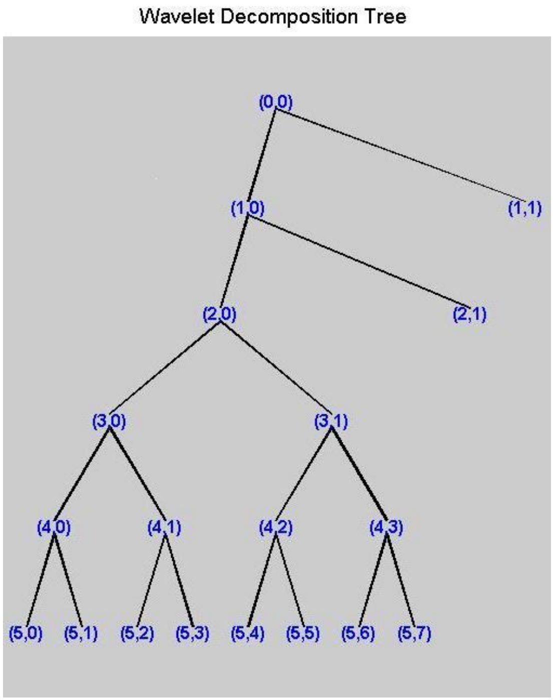


Figure 3 Modified wavelet decomposition tree generated by MATLAB

The frequency bands on the 5thdepth were calculated as following:

$$\frac{n}{2^5} f_s \sim \frac{n+1}{2^5} f_s \qquad (4)$$

Therefore, the frequency bands obtained with this method were:

(1) 0 – 689 Hz;
(2) 689—1378 Hz;
(3) 1378—2067 Hz;
(4) 2067—2756 Hz;
(5) 2756--3445 Hz;
(6)3445—4134 Hz;
(7) 4134—4823 Hz;
(8) 4823—5512 Hz.

This process can be illustrated with a decomposition tree shown in Figure 3.
Note: 8 frequency bands are presented as their index numbers in bracket as displayed above in the following contents.

4.1 Reproducibility test

Figure 4 is an example of the acoustic signal from two blockages in the siphon decomposed into 8 filtered signals by using discrete wavelet transform. This process was repeated on at least 3 signals which were collected under the same siphon condition but at different times so that the reproducibility of this experiment could be determined.

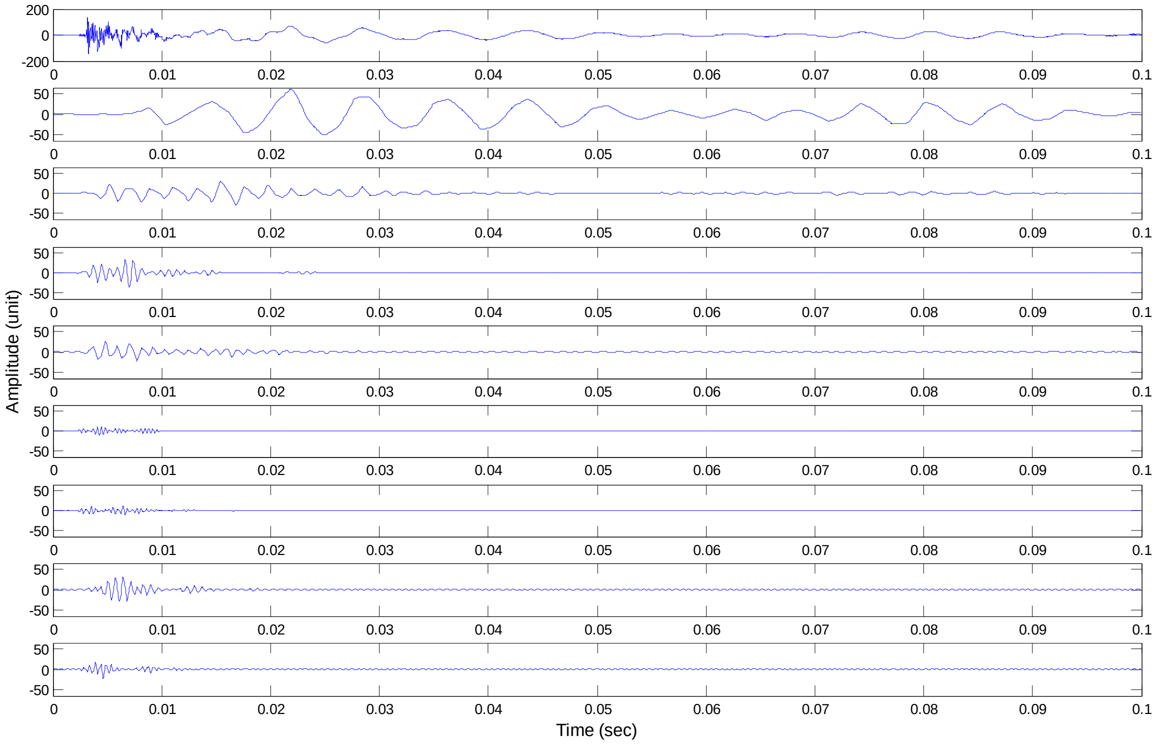


Figure 4 Acoustic impulse response of the siphon with 2 blockages decomposed using *sym4*(Singh & Tiwari, 2006) as mother wavelet. From top to bottom: the original signal plus 8 wavelet outputs with a progressive increase in the frequency band.

Energy and cross-correlation coefficients were calculated to describe the similarity between signals at same frequency range. The energy contained in each signal was calculated according to

$$E = \frac{\int_0^T f^2(t)dt}{f_s} \qquad (5)$$

As the energy of the sound generated by the speaker had varied slightly between individual measurements, the energy percentage in each frequency band was calculated to enable a comparison between these signals

$$Energy\ (\%) = \frac{energy\ \ contained\ \ in\ certain\ \ frequency\ \ range}{total\ \ energy\ \ of\ the\ signal} x100\% \qquad (6)$$

The cross-correlation coefficients were also calculated as

$$matrix\ R(x,y) = \frac{C(x,y)}{\sqrt{C(x,x)C(y,y)}} \qquad (7)$$

where C(x, y) is the covariance of the vector x and y

$$C(x,y) = E(x \cdot y) - E(x) \cdot E(y) \qquad (8)$$

In the above expression E(x) is the expected value of x

$$E(x) = \int_{-\infty}^{+\infty} xf(x)dx \tag{9}$$

where f(x) is the probability function. The maximum deviation (MD) $=max(|x_i - \bar{x}|)$, where x_i represents all samples and $\bar{x}$ is the mean of them.Maximum deviation sensitivity C (%) is calculated as a measure of the reliability of the system, the lower value of C indicates more stable of the system.

$$C = \frac{MD}{\bar{x}} = \frac{\max(\left|x_i - \frac{\sum_{i=1}^{n} x_i}{n}\right|)}{\frac{\sum_{i=1}^{n} x_i}{n}} \tag{10}$$

Table 1 presents the result for the acoustic energy determined from the reproducibility test for the siphon blocked with two sand bags. The number in the brackets in the top row corresponds to the WT band which is defined in the above paragraph. This table also presents the maximum deviation sensitivity C (%) which corresponds to the similarity between the data obtained in reproducibility experiments.

Table 1 Acoustic energy percentage of 2 blockages in the siphon at 8 frequency bands and maximum deviation sensitivity.

Energy (%)	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Test 1	18.61	11.97	2.18	1.20	0.49	1.33	1.55	1.15
Test 2	20.14	11.76	2.30	1.21	0.47	1.35	1.58	1.13
Test 3	20.83	11.29	2.17	1.09	0.50	1.37	1.56	1.10
C(%)	6.29	3.28	3.76	6.57	3.42	1.48	1.07	2.37

Table 2 presents the cross-correlation coefficient obtained in three experiments repeated in the siphon with the same amount of sediment. This table together with the acoustic energy data presents in Table 1 illustrates a very high similarity between the three repeated tests and reproducibility in the experiment.

Table 2 Cross-Correlation coefficients of reproducibility tests of 2 blockages in the siphon

Cross-correlation coefficients	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Test 1 VS Test 2	0.9989	0.9995	0.9992	0.9983	0.9961	0.9988	0.9991	0.9992
Test 1 VS Test 3	0.9993	0.9996	0.9989	0.9991	0.9997	0.9990	0.9993	0.9969
Test 2 VS Test 3	0.9985	0.9991	0.9976	0.9988	0.9979	0.9991	0.9994	0.9980

4.2 Condition classification

The values of the acoustic energy and correlation coefficients calculated for the 8 WT bands were used as features to construct training data matrix. The same process was repeated on the acoustic signals collected from unknown pipe condition and testing data matrix was constructed in the same way, see Table 3 and 4. Both matrices were used with *K*-nearest neighbors algorithm to determine the condition of the siphon from new testing data. The value of *K* was chosen 1 so that only the nearest neighbor from the training data could be found. Example of blockage condition matrices are shown in Table 3.

Table 3 Training data matrix of energy percentage of blockage conditions

Energy (%)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Class1 (clean)	1.2463	7.5603	1.0493	0.5024	0.0368	0.7646	0.0575	1.1635
Class2 (1bag)	0.1836	0.7319	0.3943	0.0366	0.0351	0.0364	0.3502	0.1288
Class3 (2bags)	0.0283	0.0929	0.0496	0.0066	0.0079	0.0131	0.1304	0.0071
Class4(3bags)	0.0235	0.0410	0.0098	0.0040	0.0150	0.0174	0.0194	0.0048
Class5(4bags)	0.0300	0.0286	0.0037	0.0043	0.0017	0.0015	0.0575	0.0020
Class6(5bags)	0.0094	0.0265	0.0118	0.0023	0.0030	0.0036	0.0257	0.0029
Class7(6bags)	0.0054	0.0313	0.0015	0.0027	0.0021	0.0035	0.0115	0.0023
Class8(7bags)	0.0064	0.0392	0.0115	0.0041	0.0185	0.0059	0.0357	0.0057
Class9(8bags)	0.0033	0.0010	0.0009	0.0003	0.0011	0.0029	0.0014	0.0008
Class10(9bags)	0.0015	0.0007	0.0002	0.0001	0.0032	0.0006	0.0009	0.0005
Class11(10bags)	0.0017	0.0006	0.0017	0.0003	0.0018	0.0012	0.0051	0.0016

Table 4 presents the testing data matrix which is composed of the values of the acoustic energy determined for the 8 WT bands. These data correspond to some new conditions against which the proposed method is to be tested. Each element in the testing data matrix is to be compared with the elements in the corresponding column in the training data matrix. In this way the training data closest to the testing data can be found. In this process a new matrix is constructed as shown in Table 5. This matrix lists all the Euclidean distance values which will indicate which training data was the closest to testing data by finding the smallest value of Euclidean distance.

Table 4 Testing data matrix of energy percentage of blockage conditions

Energy(%)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Test data1	1.4616	18.0752	0.4818	0.2592	0.0705	0.1739	1.1908	0.1488
Test data2	0.1999	2.2211	0.0540	0.1568	0.1112	0.0036	0.0964	0.0136
Test data3	0.0955	0.1376	0.0119	0.0046	0.0015	0.0038	0.0168	0.0070
Test data4	0.0057	0.0042	0.0036	0.0021	0.0028	0.0016	0.0144	0.0044

Table5 Euclidean distance matrix of energy percentage of blockage conditions

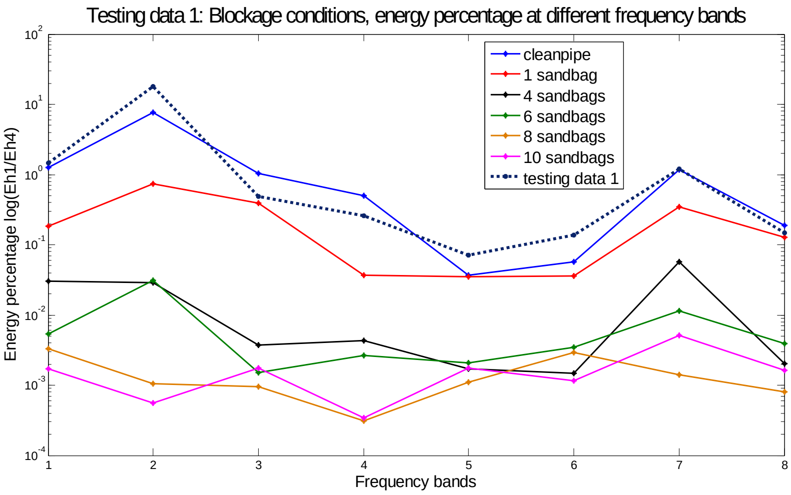
Euclidean distances	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Class1(clean)	0.2152	10.5149	0.5675	0.2432	0.0337	0.0804	0.0273	0.0403
Class2(1bag)	1.2779	17.3433	0.0876	0.2226	0.0355	0.1015	0.8406	0.0200
Class3(2bags)	1.4333	17.9824	0.4322	0.2525	0.0626	0.1247	1.0604	0.1416
Class4(3bags)	1.4381	18.0342	0.4720	0.2552	0.0555	0.1205	1.1714	0.1439
Class5(4bags)	1.4316	18.0466	0.4781	0.2548	0.0688	0.1364	1.1333	0.1467
Class6(5bags)	1.4521	18.0487	0.4701	0.2569	0.0675	0.1343	1.1651	0.1459
Class7(6bags)	1.4561	18.0439	0.4803	0.2565	0.0684	0.1344	1.1793	0.1465
Class8(7bags)	1.4551	18.0360	0.4703	0.2551	0.0520	0.1320	1.1551	0.1430
Class9(8bags)	1.4583	18.0742	0.4809	0.2588	0.0694	0.1349	1.1894	0.1480
Class10(9bags)	1.4600	18.0745	0.4816	0.2590	0.0674	0.1373	1.1899	0.1482
Class11(10bags)	1.4599	18.0747	0.4801	0.2588	0.0688	0.1367	1.1857	0.1471

Table 6 Index of nearest neighbor's class from the training data matrix to testing data matrix

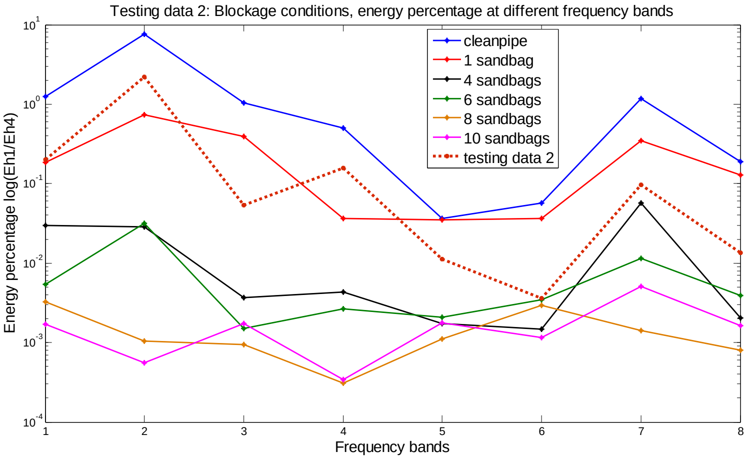
Index No.	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Test data1	1	1	2	2	1	1	1	2
Test data2	2	2	2	2	3	6	3	3
Test data3	5	3	6	5	5	6	4	3
Test data4	7	9	5	6	6	5	7	4

Majority voting was then applied to discover the most common class in the index matrix. In the index matrix Table 6, number 1 appeared 5 times as the most common number of test data 1, number 2 and 5 of test data 2 and test data 3. No obvious majority of any class was found for test data 4 with number 5, 6 and 7 appeared equal times. These results suggest that test data 1, 2 and 3 belong to class 1, 2 and 5, respectively. It is difficult to draw a clear conclusion on test data 4, but it is possible to suggest that its condition was close to any of classes 5, 6 and 7.

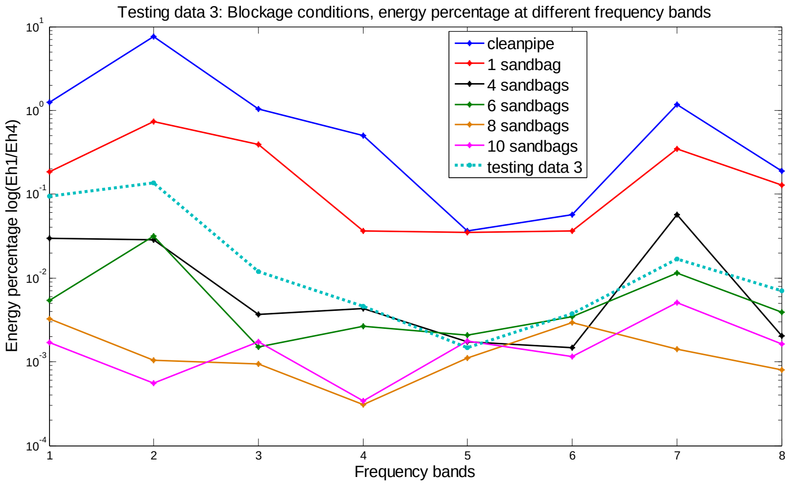
Figures of energy percentage against frequency bands of both testing data and training data support the results derived from *K*-nearest neighbors classification. Figure 5(a) shows the energy percentage against frequency bands of testing data 1 and 6 of training data sets, testing data 1 can be seen is closest to the training data of clean siphon condition which is class 1. It is the result similar to that obtained via the KNN classification method (see Table 6). Figure 5(b), (c) and (d) are testing data 2, 3 and 4 plotted in the same way with same training data sets as in Figure 5(a). All 4 figures illustrate the results consistent with those obtained via the KNN classification method.



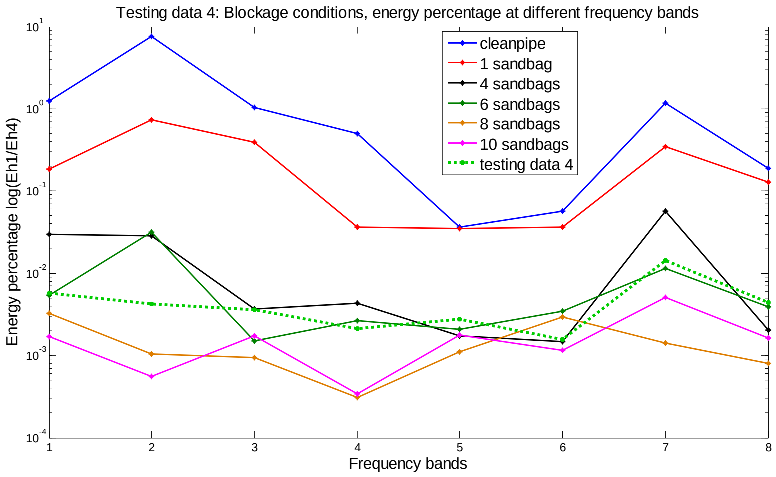
a)



b)



c)



d)

Figure 5 (a-d) Energy percentage against frequency plots of testing and training data

5. Conclusions

Discrete wavelet transform was used as a main signal processing method in reproducibility test, feature extraction and condition classification. Acoustic signals were decomposed into different frequency ranges up to 5512Hz. The energy percentage and cross-correlation coefficients between individual data sets in each frequency band were calculated as characteristic features to describe the degree of similarity between these signals. The reproducibility analysis suggests that the data are reproducible if the condition does not change.

K-nearest neighbor algorithm was used as classification method to classify the condition of the siphon. For this purpose the siphon was blocked with a controlled amount of sand. The results suggest that the acoustic technique and the adopted classification system are capable of discriminating different pipe conditions, but further refinements are needed to tuneits sensitivity and improve its accuracy. Meanwhile, it is also can be seen that the low frequency components of the signal appear to show more accurate results than their high frequency counterparts. Therefore, choosing frequency bands carefully helps to achieve better performance of the adopted classification method and it deserves a further investigation.

References

1. Christian U. Grosse, Florian Finck,'Improvements of AEtechnique using wavelet algorithms, coherence functions and automatic data analysis',Construction and Building Materials 18 , 203-213,2004.

2. Cohen, L. 'Time-Frequency Analysis', Prentice-Hall, New York,1995.

3. Debnath, L, 'Recent developments in the Wigner-Ville distribution and time-frequency signal analysis',PINSA, 68.A.No.1 35-56, 2002.

4. Michael Cowling, R. S.,'Comparison of techniques for environmental sound recognition', Pattern Recognition Letters , 2895-2907, 2003.

5. Richard O.Duda, P. E.,'Pattern Classification',John Wiley & Sons,Inc.,New York ,2001.

6. Singh, B. N., Tiwari, A. K.,'Optimal selection of wavelet basis function applied to ECG signal denoising', Digital Signal Processing 16 , 275-287,2006.

7. Gilbert Strang,'Wavelets and Filter banks', Wellesley-Cambridge Press,Wellesley MA, USA,1996.

8. Hugo T, Grimmelius, Meiler, P. P., Maas, H. L. and Bonnier, B.,'Three State-of-the-Art Methods for Condition Monitoring',IEEE Transactionsonindustrialelectronics, 407-416,1999.

An Acoustic Method for Condition Classification in Live Sewer Networks

Zao FENG, Kirill V. HOROSHENKOV, M. Tareq BIN ALI, Simon J. TAIT
School of Engineering, Design and Technology, University of Bradford, Bradford, BD7 1DP, UK
Phone: +44 1274233867
Z.feng2@bradford.ac.ukK.Horoshenkov@bradford.ac.uk
m.t.binali@Bradford.ac.uks.tait@Bradford.ac.uk

Abstract

Underground pipes are an important part of urban water infrastructure. These pipes are gradually deteriorating due to aging, operational stresses and environmental conditions. In order to be able to manage the underground pipe system efficiently, condition monitoring is needed to provide a clear understanding of the behavior of sewer systems under various hydraulic conditions. This paper reports on the application of a novel acoustic method to study the evolution of blockages and various types of damage in a full scale life sewer pipe which has been installed in the hydraulic laboratory at the University of Bradford. Temporal and frequency characteristics in the behavior of the acoustic intensity are extracted from the acoustic signals recorded on an array of microphones. These characteristics are used for pattern recognition which is based on K-nearest neighbors (KNN) classifier. The obtained results indicate that the pattern recognition system can provide a reliable classification of the pipe condition in the presence and absence of flow.

Keywords: Underground pipe, acoustic intensity, pattern recognition, condition classification

I. Introduction

Internal inspection of pipelines is done by detection systems ranging from simple visual inspection to complex imaging systems. Unlike conventional CCTV system and many other alternatives, acoustic-based methods for inspection of sewers to recognize pipe conditions can be fast, non-invasive and performed on those life pipes which are impassable for a CCTV robot. A laboratory experimental set-up to study the evolution of blockages and effect of damage on the acoustic signal propagation has been installed in Hydraulic Laboratory at the University ofBradford. The results presented in this paper are based on the analysis of acoustic signals which are reflected from various objects deposited in the partly filled pipe. It is shown that these signals carrysufficient information about the conditions of pipe, amount of deposited sedimentsand presence of lateral connections. Sound intensity data are used to extract meaningful features for classification purpose. Sound pressure has been used traditionally to analyze the conditions in pipes. This paper is based on the analysis of sound intensity which is, unlike sound pressure, is vector which direction coincides with the direction in which the acoustic energy propagates. Features extracted from the intensity data from a signaturedatabase for a range of different pipe conditions which are then used as training and testing data in classification procedures.

The classification algorithm applied in this work is K-nearest neighbors (KNN) method. The KNN is a distance-based classifier which is easy to perform.The method doesn't require any knowledge about the system of posterior probabilities. The acoustic intensity data which are used in this work are filtered in several frequency bands so that temporal and frequency features of the reflected acoustic energy could be used as main the features in the KNN trainingand subsequent recognition process.

This paper is organized as follows: Section II presents the experimental set-up and data collection, and signal pre-processing methods. Section III presents a brief introduction of the classification methodology, feature extraction and classification results. Section IV presents the discussion of the accuracy and stability of this method.

.

II. Experimental Testing

A 150mm diameter, 14.4meter long clay pipe was constructed in the Hydraulics Laboratory at the University of Bradford. This type of pipe is representative of small and medium pipes typically found in the UK's underground sewer network. A lateral connected was installed in the middle of the pipe through which simulated blockage can be implanted. The end of the pipe was connected to a water tank which was capable of discharging water at a change of flow rates. The pipe was set on a solid steel beam of the same length. This experimental setup is shown inFigure 1 (a) and (b).

An acoustic sensor which was used in these experiments consisted of four in-line MEMS microphones arranged a PCB board. It was attached to a small loudspeaker which was able to reproduce sound in the audio band. The spacing between the microphones was less than the acoustic wavelength to allow for the intensity measurements. The sensor was connected to a sound card which installed in PC (see Figure 1 (c)). The sensor was attached to the pipe wall of one end of the clay pipe and another end was either open or blocked. Broadband signals of 30 seconds were generated via loudspeaker and its reflections were recorded on four microphones to obtain acoustic intensity.

Three sets of experiments were carried out with water level which was varied from 0 to 20mm inside the pipe to simulate the dry flow conditions typical for real underground life sewers. Acoustic signals were collected for the followingconditions:

(1)empty pipe with closed lateral connection; (2) empty pipe with open lateral connection; (3)pipe with a blockage and closed lateral connection.

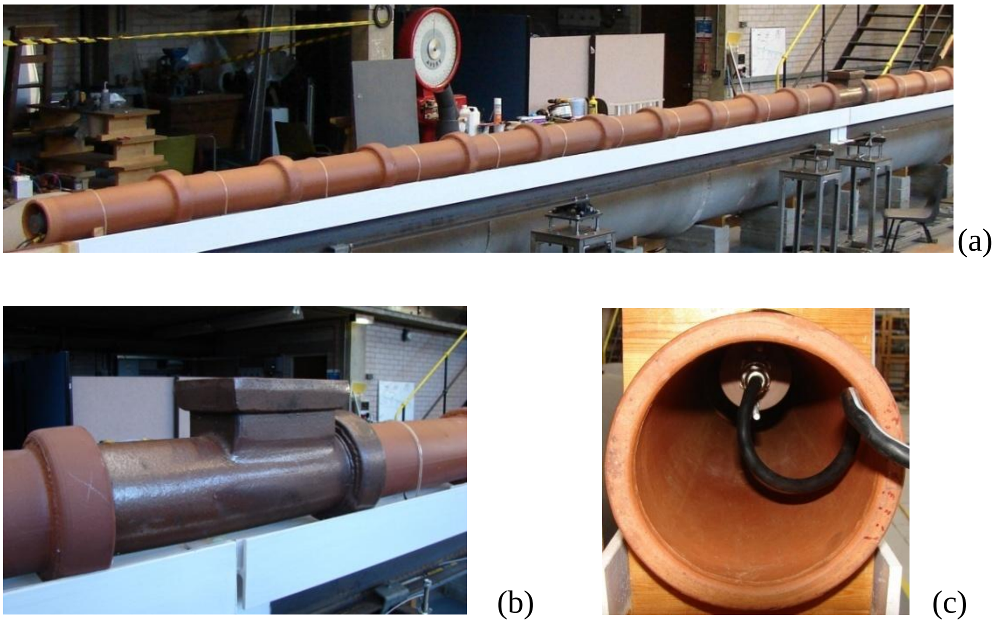


Figure 1.The 150mm clay pipe facility (a); the lateral connection (b) and sensor positioned at the downstream end of the pipe (c).

A 10 second sinusoidal sweep in the frequency range of 50Hz to 15000Hz was used as input signal. This type of time-invariant signal is widely used to measure the transfer function and it is well suited for outdoor measurements. It is less vulnerable to the deleterious effect of time variance [1] and presence of background noise. Recorded reflection signal was deconvolved to obtain the acoustic pressure impulse response which contained information onpipe geometry, sound speed and operational conditions. The broadband impulse response filtered in several narrow bands signal using a digital Butterworth filter. Figure 2 (left) shows an example of the original sinusoidal sweep signal recorded on the four microphones in the clean 150mm pipe. Figure 2 (right) shows the corresponding impulse response filtered in the frequency range of100-1000Hz.

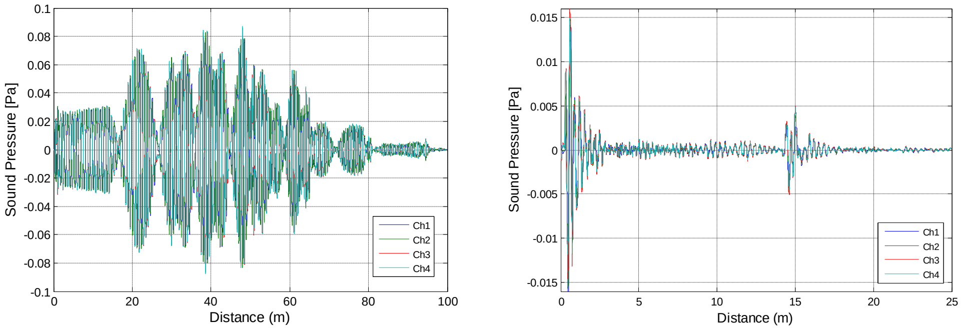


Figure 2 Recorded pressure signal (left) and its filtered impulse response (right) of clean pipe

The acoustic intensity *I*(*t*) can be calculated from data recorded on one pair of microphones using equation (1) and (2), where *p*(*t*) is the acoustic pressure, *u*(*t*) is the acoustic (particle) velocity in the direction of the normal *n* that coincides with the direction of sound wave propagation, Here ρ_0 is the density of air. However, it is difficult to extract exact value of *p* and $\frac{\partial p}{\partial n}$ at the same position, therefore, the approximation (3) and (4) are commonly used, where $p_1(t)$ and $p_2(t)$ are the sound pressures measured on the two microphones which are spaced at the distance Δ□ λ, λ is the acoustic wavelength [1].

Figure 3 present the acoustic intensity in the clean pipe and the clean pipe with an open lateral connection calculated according to this method in the frequency range of 300 – 450 Hz. A strong reflection at approximately 15mand a smaller reflection at 8m (right) can be seen clearly in the intensity plotspresented as a function of distance.

$$\tilde{I}(t) = p(t)u(t) \tag{1}$$

$$u(t) = -\frac{1}{\rho_0}\int \frac{\partial p}{\partial n} d\tau \tag{2}$$

$$p(t) \approx \frac{p_1(t) + p_2(t)}{2} \tag{3}$$

$$u(t) \approx -\frac{1}{\Delta\rho_0}\int_{-\infty}^{t}[p_1(\tau) - p_2(\tau)]d\tau \tag{4}$$

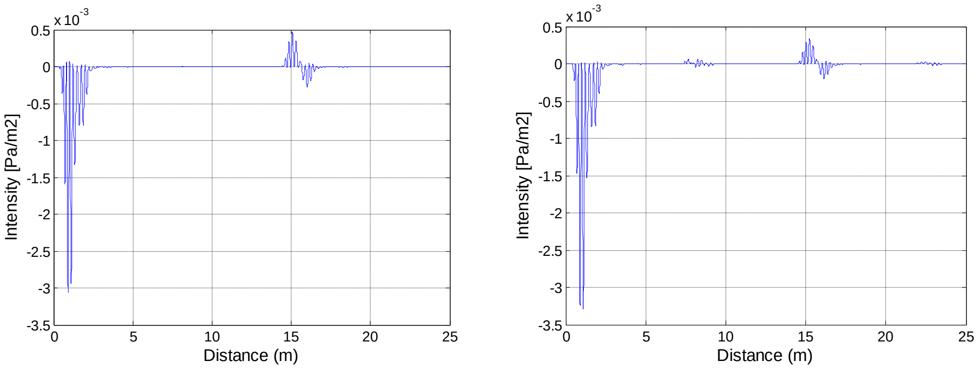


Figure 3 The intensity response of clean 150mm pipe (left) and the pipe with an open lateral connection (right) calculated in the 300 - 450 Hz range

III. Classification and Results

"*Each pattern is rule describing relation between certain context, problem and solution*." is the definition of pattern from *Christopher Alexander*[2]. Pattern is not considered a solution but a description and generalization of the experience which leads to the method how to solve the problem. Pattern classification is the organization of patterns into groupsof those sharing the same set of properties. A typical classification system normally contains four steps: data processing, feature extraction, feature selection and pattern classification [3].

The K-nearest neighbors (KNN) method has proved to be a simple but effective non-parametric classification algorithm. It is based on the use of distance measurement. Given training data $R=\{(x_1,y_1),...,(x_n,y_n)\}$ as a set of labeled samples, KNN classifier assigns a test sample $T(x_i,y_i)$ to the label associates with its *K* number of closest neighbors in *R*. The Euclidean distance is normally used to calculate the distance between test sample and training samples, i.e. $d=\sqrt{\sum_{j=1}^{n}\left|R_i-T_j\right|^2}$. The classification is done by a majority-voting rule, which states that the label assigned to the test sample should be the one which occurs the most among the K nearest neighbors.

Pipe end, blockage and lateral connection were 3 conditions for which the reflected acoustic energy was extracted and used as the signatures in the classification process. For each pipe condition, 20 experiments were carried out with variable water level. A half of these signatures were used to train for the KNN classifier and the other half were used for testing. The acoustic energy was calculated from the intensity signals filtered in the 20 frequency bands. It was used as the main feature in the classification process. Figure 4 shows examples of signatures of recorded from the pipe end and lateral connection. From the intensity plots it can be seen clearly that there are recognizable difference in the sound intensity patterns which can be used for the signature classification.

These intensity data were used to calculate the acoustic energy was obtained for each frequency band according to the following equation:

$$E=\int_{t1}^{t2} I^2(t)dt \tag{5}$$

where [t_1 t_2] is the time window chosen for this integration process. Each signature in the signature database was basically a testing data matrix which contained 10 rows corresponding to different water levels and 20 columns corresponding to different frequency bands. Therefore, each element of this matrix was the energy value in one specific frequency band and for one particular water level.

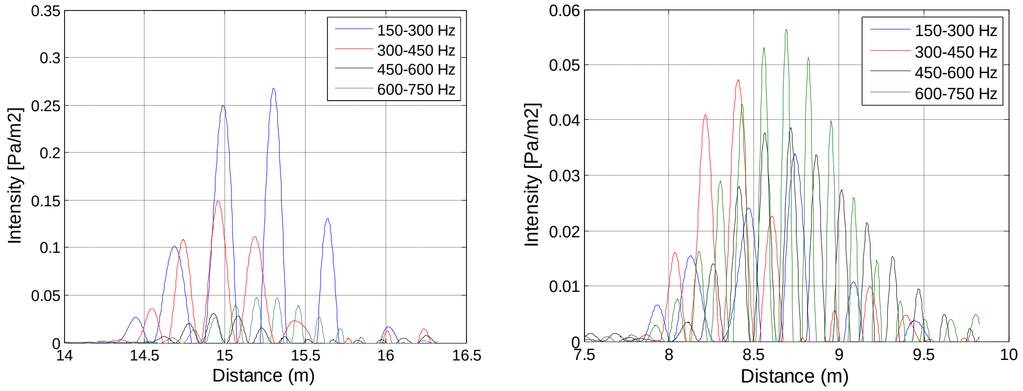


Figure 4 Signature intensity plot of pipe end (left) and lateral connection (right)

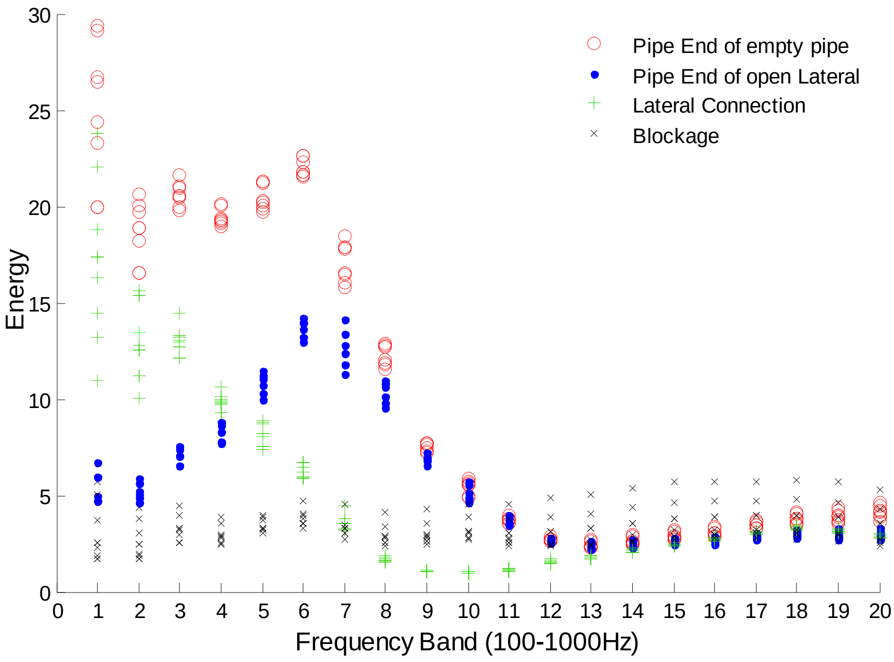


Figure 5. An example of the signature plot

Figure 5 presents the frequency dependence of the energy extracted for the following pipe conditions: pipe end of the clean pipe; pipe end of the clean pipe with an open lateral connection; lateral connection and the blockage. The figure illustrates unique patterns in each of the four signatures and shows that these patterns can be discriminated clearly within the frequency range of 150-450Hz.

The main advantage of the KNN algorithm is that it leads to a very simple approximation of the Bayesian classifier, called the majority voting rule. Assume a training data set $\{X_1, X_2 \cdots X_N\}$ containing total N samples, among which n samples are labeled, $X_n \in \text{class } \omega_i$ $(1 < n < N)$, i is the number of classes. Here comes an unknown sample x which needs to be classified. Draw a hyper-sphere of volume V around x as the estimation range, among which m samples are labeled, $X_m \in \omega_i$ $(1 < m < V)$. Pick K number of samples which are the nearest neighbors of x from $\{X_m\}$ The likelihood function of density estimation using the KNN is: $P(x \mid \omega_i) = \dfrac{K}{nV}$, similarly, the unconditional density is estimated by: $P(x) = \dfrac{m}{NV}$, and the priors are approximated by: $P(\omega_i) = \dfrac{n}{N}$. Therefore, the Bayesian classifier [4] becomes:

$$P(\omega_i \mid x) = \frac{P(x \mid \omega_i) P(\omega_i)}{P(x)} = \frac{\frac{K}{nV} \cdot \frac{n}{N}}{\frac{m}{NV}} = \frac{K}{m} \qquad (6)$$

A large value of K yields smoother decision regions, a smaller value of K improves the classification efficiency. Normally choose $K \leq \sqrt{N}$ for N number of samples using the rule of thumb. It is expected that K should be an odd integer to avoid ties. In this paper, $N=60$ at one frequency band, hence K=7 was chosen. The frequency range of 100 – 450 Hz was found to be the useful range to determine signature types (see Figure 5). This frequency range was split in 5 frequency bands which were used in the classification process. The signature types 'PE', 'BK' and 'LC' stand for pipe end, blockage and lateral connection, respectively.

Table 1 gives majority odds results for 6 signature types of 3 pipe conditions in 5 frequency bands. Table 2 shows the classification results using the adapted KNN algorithm. Estimations of testing data were based on the majority votes and were correct of all signatures. The majority votes were calculated by using equation (6) and the standard deviations were calculated using following equation:

$$s = \sqrt{\frac{1}{N}\sum_{i-1}^{N}(x_i - \overline{x})^2}\ ,\quad \overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{7}$$

where x_i are data samples of one signature, $\overline{x}$ is the mean of these data.

Table 1. Majority odds(%) of 5 frequency bands

Signature types	Freq Hz [100-250]	Freq Hz [200-350]	Freq Hz [300-450]	Freq Hz [400-550]	Freq Hz [500-650]
PE of empty pipe	100%	100%	100%	100%	85.7%
PE with open lateral connection	71.4%	100%	100%	57.1%	28.6%
PE with blockage inside pipe	100%	100%	100%	100%	42.8%
PE with blockage inside and open LC	85.7%	100%	71.4%	57.1%	28.6%
BK inside pipe	14.3%	100%	100%	100%	85.7%
LC of empty pipe	100%	100%	42.8%	85.7%	42.8%

Table 2. Signature classification results using KNN algorithm

Signature types	**Majority votes**	**Estimation**	**Standard Deviation**
PE of empty pipe	97.2%	PE	1.65
PE with open lateral connection	71.4%	PE	0.74
PE with blockage inside pipe	88.6%	PE	1.16
PE with blockage inside and open LC	65.7%	PE	0.62
BK inside pipe	80.0%	BK	0.56
LC of empty pipe	74.3%	LC	0.52

V. Conclusion

A K-nearest neighbours (KNN) algorithm has been developed and used for pipe condition classification. This system is capable of identifying pipe objects of a water-filled pipe using its acoustic signatures. 3 pipe conditions of 20 water levels were studied in this work including: empty pipe, pipe with open lateral connection and blockage inside pipe. Pipe end, lateral connection and blockage signatures were obtained and used in classification. Signatures in frequency 100Hz to 1000Hz were filtered in 20 frequency bands to improve the resolution of classification, a frequency range of 150Hz-450Hz of 5 bands was found to be more useful to discriminate patterns, the amount of data samples in calculation can be reduced using these 5 frequency bands instead of the original 20 bands, as a result, the classification results can be more sensitive to the condition change and the calculations are more efficient. The acoustic energy has been used as the main feature in the classification process. It has been found a useful characteristic which enables discernible difference between signatures to be measured. The proposed system has proved to be reliable to enable to discriminate typical conditions in a partly-filled pipe. Other acoustic parameters will be studied in the future to provide additional dimensions for the classification process and improve its robustness and resolution.

Acknowledgements

The authors would like to acknowledge the work of their technicians at the Mechanical Engineering Workshop and in the Hydraulics Laboratory in the School of Engineering at the University ofBradford.

References

1. BinAli, M. Tareq (2010). 'Development of Acoustic Sensor and Signal Processing Technique', PhD thesis, The University of Bradford.

2. Richard. O.(2000). 'Pattern Classification', NY,USA: Wiley-Interscience.

3. Yella, S. (2006). 'Condition monitoring using pattern recognition techniques on data from acoustic emissions'. Proceedings of the 5th International Conference on Machine Learning and Applications.

4. Bishop, C. M. (1995). 'Bayesian regression and classification'. Computer and Systems Sciences , volume 190.

Application of KNN Classifier for Acoustic Based Pipe Condition Classification

Zao Feng M. Tareq Bin Ali
z.feng2@bradford.ac.ukm.t.binali@Bradford.ac.uk
Kirill V. Horoshenkov Simon Tait
k.horoshenkov@bradford.ac.uks.tait@Bradford.ac.uk
School of Engineering, Design and Technology
University of Bradford
Bradford, UK

***Abstract*—Underground pipeline infrastructure can decay at an accelerating rate due to insufficient quality control, ineffective condition monitoring and maintenance and a general lack of uniformity in design and operation. An intelligent system that is a rapid and reliable decision-making tool to measure the condition of buried water and sewer pipeline infrastructure systems is urgently required by water companies. This paper proposes a novel approach of discriminating and classifying different pipe conditions under various hydraulic conditions. A full scale live pipe was installed in the Hydraulic Laboratory at University of Bradford to study the evolution of blockage and damage effects on acoustic signal propagation. Pattern classification algorithms were studied and applied. In this work, the acoustic intensity and reflected energy were regarded as meaningful signatures that could used for feature extraction and recognition. K-nearest neighbours (KNN) was used as classifier to recognize pipe conditions. The experimental results show that the proposed acoustic based pattern recognition system is a reliable tool that can be used to discriminate between pipe ends, lateral connections and sediment blockages in the presence and absence of flow.**

I. Introduction

The economic and social costs associated with collapse of buried underground infrastructure can be significant. Many pipelines were installed in the first part of the 20th century. The condition of these assets is largely unknown and they continue to deteriorate largely unnoticed [1]. Therefore, regular inspection of these assets is needed. Efficient method for inspection and monitoring of pipelines has been an active area of research for many years and several solutions have been proposed.However, the most common inspection method which is used at present is the close circuit television inspection. It is a detailed method of inspection, but it is slow, expensive and subjective.

Acoustics provides alternative means to inspect pipes more rapidly and objectively. In underground sewer pipes, it is relatively straightforward to inject an acoustic signal into the pipe and listen for the reflections which would inevitably occur when there is a cross-sectional change or change in the acoustic impedance in the pipe wall [2]. The signals reflected from an artefact in the pipe carries information about the nature of this artefact, its extent and severity. It is attractive to have a system in place which can process this information and classify the nature and severity of these defects automatically. This paper presents a novelcondition classification system that has been designed and tested to detect defects of live sewer pipeline using acoustic intensity data. K-nearest neighbours (KNN) is applied in this work and proved to be capable of discriminating pipe segments and defects under dry flow conditions.

I. Experimental Setup

A 150mm diameter, 14.4meter long clay pipe was constructed in the Hydraulics Laboratory at the University of Bradford. A lateral connected was installed in the middle of the pipe through which simulated blockage can be implanted. The end of the pipe was connected to a water tank which was capable of discharging water at a change of flow rates. The pipe was set on a solid steel beam of the same length as shown in Fig. 1(a) and (b). An acoustic sensor which was used in these experiments consisted of an array of four microphones. A small loudspeaker was used to reproduce sound in the audio band. The spacing between the microphones was less than the acoustic wavelength to allow for the intensity measurements. The sensor was connected to a sound card which installed in a PC. The sensor was attached to the top of the clay pipe near a pipe end. The other end of the pipe was either open or blocked as illustrated in Fig.1(c).

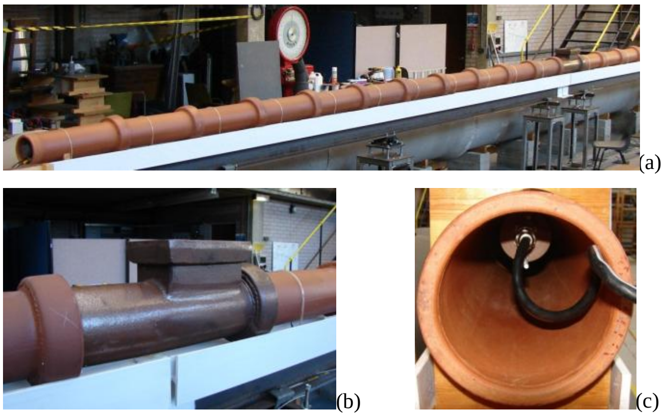


Fig. 1. The 150mm clay pipe facility (a); the lateral connection (b) and sensor position (c).

Four sets of experiments were carried out with water level which was varied from 0 to 20mm inside the pipe to simulate the dry flow conditions typical for real underground life sewers. Acoustic signals were collected for the following conditions: (1) empty pipe with closed lateral connection; (2) empty pipe with an open lateral connection; (3) pipe with a blockage and closed lateralconnection; (4) pipe with blockage and an open lateral connection.

I. Signal Processing

A sine sweep of 10 second long was generated via loudspeaker in the frequency range of 50 – 15000 Hz and its reflections were recorded on four microphones. The signals recorded on the four microphones were deconvolved to obtain the acoustic pressure impulse response which contained information on pipe geometry, sound speed and operational conditions. Fig. 2 shows an example of the acoustic pressure impulse response recorded in the clean, empty pipe which is plotted against the propagated distance ($d = ct$, where c is the speed of sound and t is the time). These calculated impulse responses were filtered using a digital Butterworth filter of 3rd order. The filtered signals recorded on the six microphone pairs were used to calculate the instantaneous acoustic intensity using the method detained in [3] and [4]. The instantaneous intensity data calculated for the six microphone pairs were compensated for the time delay and combined synchronously.

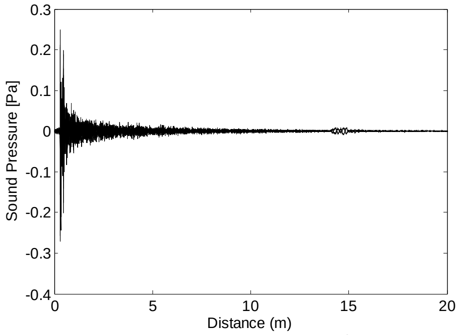
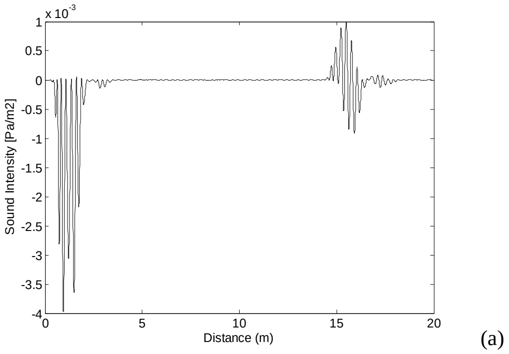


Fig. 2. Recorded pressure impulse response of clean empty pipe

Fig. 3 presents the acoustic intensity in the clean pipe (Fig. 3 (a)) and the clean pipe with an open lateral connection (Fig. 3 (b)) calculated according to this method and filtered in the frequency range of 300 – 450 Hz. A strong reflection at approximately 15m and a smaller reflection at 8m can be seen clearly in the intensity plots presented as a function of distance.



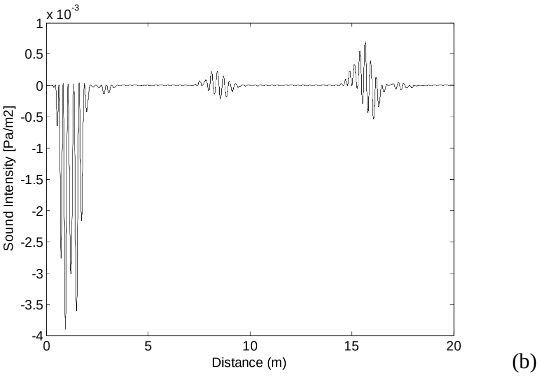


Fig. 3. The intensity response of clean pipe (a) and the pipe with an open lateral connection (b) filtered between 300 - 450 Hz

II. Classification

A. Theory Background

Pattern recognition aims to classify data based either on a priori knowledge or statistical information extracted from the measurements. In our work statistical information was used to carry out the training and automatic classification of pipe defects. For this purpose, a common non-parametric method of the *K*-nearest neighbours (KNN) classifier was adopted [5]. According to this method, *K*-nearest neighbours are computed and majority voting can be applied among the computed samples in the neighbourhood to make classification. This classification is based on the class label of the testing sample which is decided by the majority class of its *k* closest neighbours to which the distance is calculated as Euclidean distance.

B. Feature Extraction

Pipe end, blockage and lateral connection were 3 objects for which the reflected acoustic energy was extracted and used as the signatures in the classification process. For each pipe object, 20 experiments were carried out with variable water level from 0 to 20mm to extract individual acoustic signatures. A half of these signatures were used to train for the KNN classifier and the other half were used for testing. The acoustic energy was calculated from the intensity signals filtered in 20 frequency bands from 100 to 1000Hz with bandwidth equals to 150Hz. The energy of the signals filtered in these frequency bands was then calculated as

$$E = \sum_{t_1}^{t_2} I^2(t) \qquad (1)$$

where $I(t)$ is the instantaneous acoustic intensity and t_1 and t_2 are some time limits which correspond to the time window used for the acoustic signature selection. Figure 4 shows examples of signatures of intensity recorded from the pipe end (Fig. 4(a)) and blockage (Fig. 4(b)).

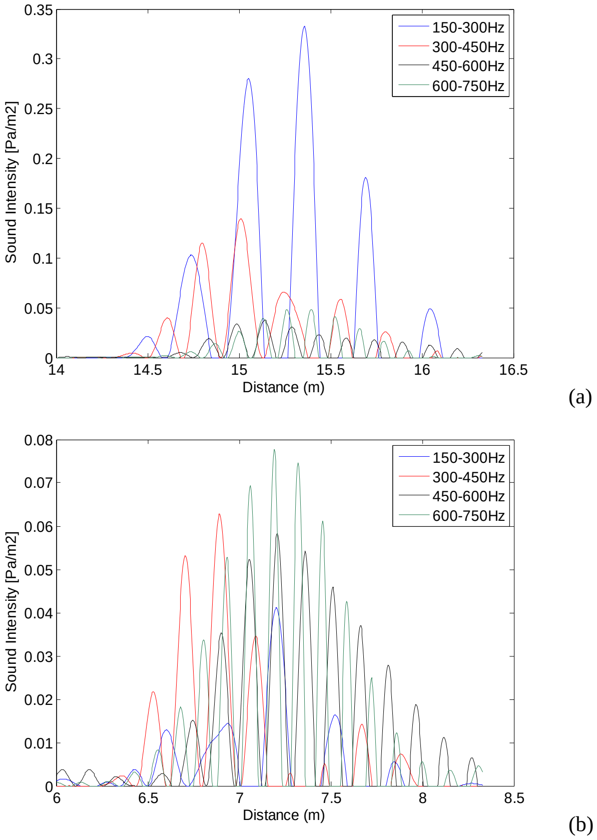


Fig. 4. Intensity signature of pipe end (a) and blockage (b)

An energy data matrix was constructed for each intensity signature for training which contains 10 sets of energy data corresponding to 10 different water levels at 20 different frequency bands from 100 to 1000Hz. A testing data matrix can also be constructed in the same way but only corresponds to 1 water level chosen randomly from the other 10 sets of data of the signature. Therefore, the size of each training data matrix is 10 × 20, each testing data vector is 1 × 20.

In total, 6energy data matrices were constructed: (1) PE1: pipe end of empty pipe; (2) PE2: pipe end of empty pipe with an open lateral connection; (3) PE3: pipe end of pipe with blockage inside; (4) PE4: pipe end of pipe with blockage and an open lateral connection; (5) BK: blockage in the pipe; (6) LC: open lateral connection of empty pipe.

A. KNN Classification

K-Nearest Neighbours (KNN) method is a distance-based classification algorithm, by calculating the distance between the sample data and all labelled data, number *K* of its closest neighbour samples are selected. The sample data will be assigned by majority vote rule to the class most common among its *K* nearest neighbours. The value of *K* depends on the data, normally, a larger value of *K* yields smoother decision ranges but increases the computational burden. Empirically, $K = \left\lfloor \sqrt{N} \right\rfloor$, N is the number of labelled data in one observation.

In this work, there were6 classes of signature with 60 training samples in each class obtained from 10 random water levels data of each acoustic signature in 5 selected frequency bands. Testing samples were chosen from the other water level data in the same frequency range. The original training data matrix and testing data matrix both have 20 rows corresponding to 20 frequency bands, after feature selection only 5 frequency bands remained, therefore, the size of training data and testing data matrices of each signature reduced to 10 × 5 and 1 × 5, respectively. Both training and testing data matrices were in the form of $\mathbf{E} = \left\{ E_{lj} \right\}$ (see equation 1), where l is the index for the water level experiment, j is the frequency band.

Training data in each frequency band from all 6 classes were considered as one observation and KNN was applied to each observation. Each observation contained 60 training data samples, hence *K*=7 was picked for each testing data sample. Fig.5 is the plot of the acoustic energy as a function of the frequency band determined for the training data and for one set of the testing data which consisted of 5 signatures. In this figure the frequency range of 150 -450Hz is marked with the dashed lines in Fig. 5. This is the most useful range for condition classification in which individual signature classes are clearly separated in terms of their spectral energy values.It can be seen that there areconsiderable overlaps of signature classesin the low frequency range below 150 Hz. In the higher frequency range (above 450 Hz) different signatures fell close to each other. These two frequency ranges are deemed unreliable for condition classification. One set of testing data was picked randomly from the testing database, 7 nearest neighbours of the testing data at each frequency band will be computed for decision making of classification.

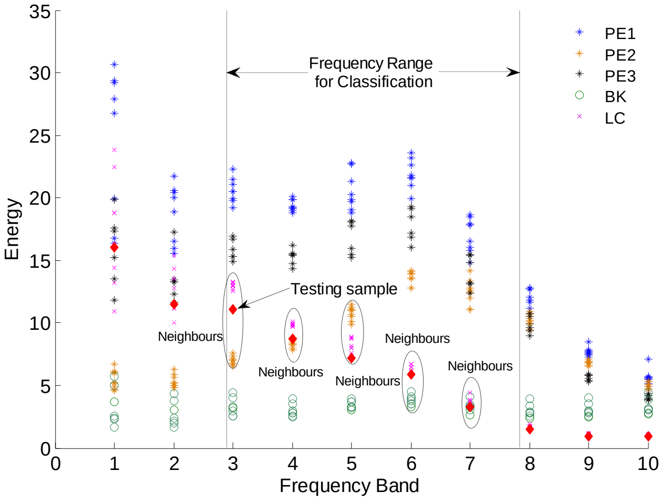


Fig.5. Energy plot of training data and testing data of signatures

Majority vote is the decision rule in KNN classification final process that normally picks the number appearing more than half out of all numbers or simply picks the number has the highest appearance among all. A new class label matrix can be constructed for each testing data set using the class labels of all nearest neighbours:

$$\mathbf{\Lambda} = \left\{ i : \min_{K=7} \left| \mathbf{E}_{training} - \mathbf{E}_{testing} \right|_i \right\} \qquad (2)$$

where N is the number of all signature estimations, μ is the correct class label of the testing data and N_μ is the time of μ shown in the label matrix. The results are shown in the last row of Table I.

Class label matrices of 6 sets of testing data from each signature constructed using equation (2) were given in Table I. Majority vote was applied to each matrix and picked the class number had the highest appearance of all, the proportion of which was calculated using the following:

$$\Pr(X = \mu \mid N) = \frac{N_\mu}{N} \qquad (3)$$

Where N is the number of all signature estimations, is the correct class label of the testing data and is the time of shown in the label matrix. The results are shown in the last row of Table I.

KNN classification was able to discriminate different pipe objects under different conditions when applied carefully to acoustic data within a selected frequency range. The percentages of the correction estimation of all signatures were higher than 70% as shown in Table I. Lateral connection (LC) and pipe ends with open lateral connection (PE2 and PE4) were relatively lower in accuracy of condition classification compared with the others, suggests that lateral connection is the object which could cause more complicated acoustic field in the pipe. Reflected acoustic intensity energy can be used as main feature in KNN classification to discriminate pipe objects but choosing a useful frequency range is crucial.

TABLE I. Example Applications of 7-NN Classification Results

File Group of Testing Data	*PE1*	*PE2*	*PE3*	*PE4*	*BK*	*LC*
Class label matrix $\mathbf{\Lambda}$ K=7	$\begin{bmatrix} 1&1&1&1&1\\1&1&1&1&1\\1&1&1&1&1\\1&1&1&1&1\\1&1&1&1&1\\1&1&1&1&1\\1&1&1&1&2 \end{bmatrix}$	$\begin{bmatrix} 2&2&2&2&3\\2&2&2&3&2\\4&2&2&3&3\\2&2&2&2&1\\2&2&2&2&3\\2&2&2&3&2\\4&2&2&2&3 \end{bmatrix}$	$\begin{bmatrix} 3&3&3&3&2\\3&3&3&3&3\\3&3&3&3&2\\3&3&3&3&3\\3&3&3&3&2\\3&3&3&3&3\\3&3&3&3&2 \end{bmatrix}$	$\begin{bmatrix} 4&4&4&4&4\\4&4&4&7&4\\4&4&4&6&4\\4&4&5&7&4\\4&4&5&4&4\\2&4&4&6&6\\4&4&4&7&4 \end{bmatrix}$	$\begin{bmatrix} 4&5&5&5&6\\4&5&5&5&5\\4&5&5&5&5\\5&5&5&5&5\\4&5&5&5&5\\4&5&5&5&5\\4&5&5&5&5 \end{bmatrix}$	$\begin{bmatrix} 6&6&3&6&6\\6&6&6&6&6\\6&6&3&3&5\\6&6&6&6&5\\6&6&3&6&5\\6&6&6&6&6\\6&6&5&6&5 \end{bmatrix}$
Majority Vote	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 5$
Signature Estimation	PE1	PE2	PE3	PE4	BK	LC
Correct Estimation % of $\mathbf{\Lambda}$	97.14%	71.43%	88.57%	74.28%	80%	74.28%

I. Summary

Pipe end, blockage and lateral connection were studied in this work to develop a classification system which would be able to discriminate these pipe objects under different hydraulic conditions. Acoustic signals were collected from the dry flow conditions realistically simulated in the laboratory. The acoustic intensity and reflected energy were calculated and used as main features in the classification process to recognise automatically pipe conditions based on their unique acoustic signatures. Initially, the acoustic energy data were presented in 20 frequency bands from 100 to 1000Hz to study their dependence on the condition in the pipe. Ultimately, only 5 frequency bands from 150 to 450Hz were selected as it was found that in this frequency range the acoustic energy in the response from different pipe conditions had clearly cognizable patterns.

K-Nearest Neighbours (KNN) method was applied to the acoustic energy data within the selected frequency range, K was chosen to be 7 in this work based on the amount ofdata and class.Classification results suggest that using reflected energy from a suitable frequency range to identify pipe objects under various hydraulic conditions is possible. Acoustic energy distributions against frequency of differentpipe objects have unique patterns and can be useful to take into further study of pipe condition classification.

References

[1]. R. A. Fenner, “Approaches to sewer maintenance: A review”, Elsevier, Urban water 2(2000) pp 343-356.

[2]. Siril Yella, Naren Gupta and Mark Dougherty, “Condition monitoring using pattern recognition techniques on data fromacoustic emissions”, IEEE Computer Science, ICMLA 2006, 0-7695-2735-3.

[3]. M. T. Bin Ali, K.V. Horoshenkov, S. J. Tait, “Rapid detection of sewer defects and blockages using acoustic based instrumentation”, Water science & technology 2010, session 2.4

[4]. M. T. Bin Ali, “Development of acoustic sensor and signal processing technique”, PhD thesis, School of engineering, design and technology, University of BradfordUK, 2010.

[5]. Richard O. Duda, Peter E. Hart amd David G. Stork, “Pattern Classification”, John Wiley & Sons, Inc. New York, 2001.