SOME NEW TECHNIQUES FOR PATTERN RECOGNTION RESEARCH

AND LUNG SOUND SIGNAL ANALYSIS

A Thesis submitted to the Faculty of Engineering

of the University of Glasgow

for the degree of Doctor of Philosophy

by

Roderick B. Urquhart

April 1983

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

"I take the view, and always have done, that if you cannot say what you have to say in twenty minutes, you should go away and write a book about it."

Lord Brabazon

"...that in all things He might have the pre-eminence"

Col 1:18

SUMMARY

This thesis describes the results of a collaborative research programme between the Department of Electronics & Electrical Engineering, University of Glasgow, and the Centre for Respiratory Investigation, Glasgow Royal Infirmary. The research was initially aimed at studying lung sound using signal processing and pattern recognition techniques. The use of pattern recogntion techniques was largely confined to exploratory data analysis, which led to an interest in the methods themselves. A study was carried out to apply recent research in computational geometry to clustering

Two geometric structures, the Gabriel graph and the relative neighbourhood graph, are both defined by a region of influence. A generalization of these graphs is used to find the conditions under which graphs defined by a region of influence are connected and planar. The Gabriel graph may be considered to be just planar and the relative neighbourhood graph to be just connected.

From this two variable regions of influence were defined that were aimed at producing disconnected graphs and hence a partitioning of the data set. A hierarchic clustering based on relative distance may be generated by varying the size of the region of influence. The value of the clustering method is examined in terms of admissibility criteria and by a case study.

An interactive display to complement the graph theoretical clustering was also developed. This display allows a partition in the clustering to be examined. The relationship between clusters in the partition may be studied by using the partition to define a contracted graph which is then displayed. Subgraphs of the original graph may be

used to provide displays of individual clusters. This display should provide additional information about a partition and hence allow the user to understand the data better.

The remainder of the work in this thesis concerns the application of pattern recogntition techniques to the analysis of lung sound signals. Breath sound was analysed using frequency domain methods since it is basically a continuous signal. Initially, a rather ad hoc method was used for feature extraction which was based on a piecewise constant approximation to the amplitude spectrum. While this method provided a useful set of features, it is clear that more systematic methods are required.

These methods were used to study lung sound in four groups of patients:- (1) normal patients, (2) patients with asbestosis, (3) patients with cryptogenic fibrosing alveolitis (CFA) and (4) patients with interstitial pulmonary oedema. The data sets were analysed using principal components analysis and the new graph theroretical clustering method (this data was used as a case study for the clustering method). Three groups of patients could be identified from the data:- (a) normal subjects, (b) patients with fibrosis of the lungs (asbestosis & CFA) and (c) patients with pulmonary oedema. These results suggest that lung sound may be able to make a useful contribution to non-invasive diagnosis. However more extensive studies are required before the real value of lung sound in diagnosis is established.

ACKNOWLEDGEMENTS

I would like to express my thanks to Prof. J.Lamb of the Department of Electronics and Electrical Engineering, Universtity of Glasgow, for the provision of research facilities. I would also like to thank Dr. J.E.S. Macleod for supervising and encouraging my research. Thanks is also due to Mr. D.D. Campbell and Mr. K. Melvin for developing hardware and assembly level programmes for the lung sound work. The assistance of Mrs. L. McCormick, Mrs. A. McVey and Miss A. MacKinnon in programming the PDP-11 and 4070 computers is also appreciated.

This research would not have been possible without the provision of research facilities at Glasgow Royal Infirmary by Dr. F. Moran. The clinical part of the collaborative research was carried out by Dr. S.W. Banham and Dr. C.C.Godley. The work on lung sound was started by J.McGhee, a previous research student.

I am also indebted to Prof. G.T. Toussaint (McGill University), Dr. R.G. Loudon (College of Medicine, University of Cincinnati) and Dr. D.J. Murray-Smith (University of Glasgow) for useful discussions on computational geometry, lung sound and biological signal analysis respectively.

Finally I would like to thank my Ph.D. colleagues and fellow sufferers, especially Ivan Andonovic, Edwin Pun and Paul Siebert from my own year, for some good laughs.

CONTENTS

Chapter 1	Introduction	1
1.1	General Introduction	1
1.2	Aims of Research	5
1.3	Overview of Thesis Contents	6
Chapter 2	Lung Sound	9
2.1	Respiratory System	9
2.2	Lung Sound	13
2.3	Directions in Lung Sound Research	20
Chapter 3	Pattern Recognition & Signal Processing	24
3.1	Introduction	24
3.2	Dimensionality Reduction	26
3.3	Exploratory Data Analysis	29
3.4	Linear Transformations	32
3.5	Spectral Estimation	35
3.6	Segmentation & Feature Extraction Schemes	39
Chapter 4	Geometric Structure in Pattern Recogntion	46
4.1	Introduction	46
4.2	Outline of some Basic Computational Structures	53
4.3	Graphs defined by a Region of Influence	59
4.4	Discussion	67
Chapter 5	A New Graph Theoretical Clustering Method	73
5.1	Introduction	73
5.2	Clustering Methodology	74
5.3	Limited Neighbourhood Sets and Clustering	80
5.4	Interactive Clustering	94
5.5	Overall Discussion	102

Chapter 6 Acquisition and Analysis of Lung Sound Signals 108

6.1 Introduction 108

6.2 Transducers 109

6.3 Lung Sound Signal Analysis 112

6.4 Techniques for Breath Sound Analysis 115

6.5 Discussion 122

Chapter 7 Breath Sound in Respiratory Disease 127

7.1 Introduction 127

7.2 Pattern Analysis 128

7.3 Time Variation of Spectra 131

7.4 Discussion 133

Chapter 8 Conclusions & Suggestions for Further Research 138

8.1 General Conclusions 138

8.2 Suggestions for Further Research 140

Appendix 1 Proposal for a New Nonlinear Mapping Algorithm 147

Appendix 2 List of Published Work 152

INTRODUCTION

1.1 General Introduction

This thesis describes research into new techniques for pattern recognition and for lung sound analysis. The project was originally envisaged as a study of lung sound using pattern recognition and signal processing methods. However as a result of the use of exploratory data analysis techniques, some research was begun into such techniques themselves. Work on applying geometric structures to the clustering problem became a major part of the thesis.

The lung sound work is part of an ongoing collaborative project between the Department of Electronics and Electrical Engineering, University of Glasgow, and the Centre for Respiratory Investigation, Glasgow Royal Infirmary. At the time of writing the project is in its sixth year and has involved development of recording equipment, the recording of patients with different conditions and signal analysis by computer.

During recent years considerable interest has been shown in biological signals [1.1]. Among the most commonly studied signals are the electrocardiogram (ECG), the electroencephalogram (EEG) and electromyogram (EMG). Each of these signals has been studied for some time using signal processing methods. Remarkably, a signal used almost universally in physical examination, **lung sound** [1.2], has received very little attention in terms of signal analysis.

Among the data analysis techniques that are widely available are clustering methods. In view of the widespread and interdisciplinary use of clustering it is disturbing that little attention has been paid

to **clustering methodology** [1.3]. In particular there have been some problems associated with the use of clustering techniques that are based on the visual notion of a cluster.

In recent years considerable advances have been made in the new field of **computational geometry** [1.4] and its application to pattern recognition [1.5]. One area that has not benefited from the progress in computational geometry is cluster analysis. The application of computational geometric techniques to clustering would seem to be an obvious area of research.

An interesting but difficult feature of this thesis is its inter-disciplinary nature. Pattern recognition is itself derived from many fields e.g. electrical engineeering, computer science, psychology, and formal language theory. Physiological and clinical studies on lung sound are relevant to this thesis, and the pattern recognition work here borrows some ideas from geography and geophysics.

1.1.1 Pattern Recognition

Pattern recognition problems are part of everyday life. A human will use and interpret his/her senses and use them to learn about the environment. Seeing objects or hearing a conversation in the presence of noise may be easy to humans but are difficult problems to automate. Conversely machines perform better at tasks such as n-dimensional geometry. Human perception involves many complex processes that are not yet well understood, and so perception by machine should not be direct ed at imitating that of humans.

Historically pattern recognition has closely followed advances in computer technology such as the development of computer graphics and

parallel processing [1.6,1.7]. In the twenty or so years of existence, pattern recognition methods have been applied to many different problems. Among the most widespread applications are character recognition (both of printed and handwritten characters), industrial fault detection, medical imaging and signal analysis. A number of reviews of pattern recognition have been written e.g. Nagy [1.8] and Fu [1.9].

In designing and implementing a pattern recognition system there are many considerations that can be taken into account. One scheme might be:

(1) Problem formulation

(2) Interfacing with the real world

(3) Understanding the nature of input data

(4) Data reduction

(5) Decision taking

The first point is self evident. A digital pattern recognition system will usually be based on a computer and so it is necessary to interface it to the outside world. This stage might involve transducers, signal conditioning and analogue to digital conversion. In order to design a pattern recognition system properly it is necessary to understand the structure of the data. If data structure is understood properly the data properties can be exploited for greater efficiency. Unfortunately it is usually necessary to generate vast quantities of data at the input causing problems with computer resources. Reducing the data to a managable size is often the crux of a pattern recognition problem. Finally the output of a pattern recognition system is usually a decision, whether the decision is made by man or machine.

1.1.2 Information Engineering in Diagnosis - A Personal View

Information engineering techniques have had an impact on a number of areas in medicine. Computers have been used to aid diagnosis, patient care and management. In diagnosis they have allowed a new understanding of some images and signals. Other research has been aimed at providing an automatic computerized diagnostic system [1.10,1.11].

The human body is very complex and so it is not really surprising that at times some systems will not function properly. When such functions are abnormal, disease is said to be present. The problem of diagnosis is therefore one of identifying probable causes of abnormal function which will then allow decisions to be made about patient care and treatment. Unlike simpler systems, such as industrial plant, the information available rarely allows a definite identification of disease to be made. There may be for example a combination of several disease processes and a diagnosis may change as more is learned about a patient. In my view these factors suggest that a radically different approach is required for medical diagnosis when compared with fault finding in industrial systems.

What then should be the objectives of information engineering in diagnosis? In one extreme we might suggest developing some sort of automated diagnostic system and in the other developing sophisticated tools for the clinician. It is rather disturbing that engineers frequently desire to develop systems that replace rather than complement human abilities.

One area in which information engineering may benefit diagnosis is making quantitative measurements on what are normally subjective

physical signs. This allows reproducibility between clinicians and permits a basis for objective comparison. Examples of such signs are auscultation and reading of biological signal traces.

Another very useful area in which information engineering plays a part is the development of **non-invasive techniques.** Some tests and mesurements on patients require placement of probes, or transducers, into the body which can be both hazardous and painful. There are inherent advantages in techniques which avoid this and which cause less stress to a patient. A good example of the use of non-invasive techniques is that of Bache et al [1.12] on cardiac output.

With these considerations in mind, the lung sound project has sought to extend knowledge on lung sound to aid in diagnosis of respiratory disease. This is also the motive behind the use of **pattern analysis** rather than **pattern classification** techniques in this study.

1.2 Aims of Research

This research project originated as a quantitative study of lung sound signals by computer. This arose because of the recent increase in interest in using lung sound; largely as a result of the work of Forgacs in London [1.2]. Forgacs's contribution was rational explanations for each of the categories of respiratory sound which had previously been explained in terms that were less than scientific.

It was hoped that by recording patients from a number of different respiratory diseases, techniques could be developed to detect any differences between the sounds produced by the different groups. The emphasis was on recording and finding significant differences rather than on finding precise physiological information.

Throughout the project it was realized that it was important to consider the nature and appropriateness of the methods used. This led to a major portion of the work being done on geometry and clustering methods. These subjects arose because of an interest in pattern analysis methods that were used extensively in investigating lung sound. The work on clustering methods is probably of general interest in pattern recognition and was not strictly necessary for analysis of lung sound.

To summarize the main aims were twofold:

(1) To investigate new geometric methods in pattern recognition with particular regard to clustering.

(2) To study lung sound signals using pattern recognition techniques with a view to finding a source of clinically useful information.

1.3 Overview of Thesis Contents

Since much of the work will be relevant to clinical as well as biomedical engineering work on lung sound, it is difficult to write in a way that will be useful to both disciplines. A reader who is interested primarily in lung sound research should perhaps read Chapters 1-3 by way of introduction, then skip Chapters 4 & 5 which are concerned with pattern recognition methods, then read Chapters 6-8 which all refer to lung sound analysis. A reader mainly interested in the new pattern recognition methods should concentrate on Chapter 1, then Chapters 3-5 and finally the conclusions in Chapter 8.

The following two chapters (Chapters 2 & 3) are designed to provide background material necessary for this thesis. Some ideas on the lung, the production and properties of lung sounds, pattern

recognition and signal processing are introduced. The presentation of these topics is mainly limited to aspects of each subject that are relavent to later chapters.

The next two chapters (Chapters 4 & 5) are concerned with geometric and clustering techniqes in pattern recognition. In Chapter 4 some basic geometric structures are introduced with later sections describing original work in this area. Chapter 5 is mainly a description of a new clustering method based on a visual idea of the cluster, and of a complementary display to aid the user in interpreting the results. Chapter 5 utilizes some of the geometric structures in Chapter 4.

Chapters 6 & 7 describe work on lung sound. Chapter 6 is mainly about signal acquisition and analysis. The microphone system described was designed by a previous reasearch student - Joseph McGhee. The methods for analysis of lung sound are quite original although they were built on the previous experience of McGhee. Chapter 7 concentrates on the comparison of breath sound between normal and abnormal subjects.

Finally Chapter 8 gives general conclusions on all aspects of the research described in this thesis.

References

1.1 B.McA. Sayers, Exploring biological signals, Biomedical Engineering, 10, 335-341 (1975)

1.2 P.Forgacs, Lung Sounds, Balliere Tindall, London (1978)

1.3 R.Dubes & A.K.Jain, Validity studies in clustering methodology, Pattern Recognition, 11, 235-254 (1979)

1.4 M.I.Shamos, Computational Geometry, Ph.D. Thesis, Yale University (1975)

1.5 G.T.Toussaint, Pattern recognition and geometrical complexity, Proc. 5th Internat. Conf. on Pattern Recognition, Miami, U.S.A., 1324-1346 (1980)

1.6 R.O.Duda & P.E.Hart, Pattern classification and scene analysis, Wiley (1973)

1.7 Y.T.Chien, Interactive Pattern Recognition, Marcel Dekker (1978)

1.8 G.Nagy, State of the art in pattern recognition, Proc. IEEE, 56, 836-862 (1969)

1.9 K.S.Fu, Recent developments in pattern recognition, IEEE Trans. Comput., C-29, 845-854 (1980)

1.10 E.A. Patrick & Shen, Review of pattern recognition in medical diagnosis and consulting relative to a new system model, IEEE Trans. Syst., Man & Cybernet., SMC-4, 1-16 (1974)

1.11 E.A. Patrick & Shen, On the theory of medical diagnosis and consulting, Proc. 1st Internat. Joint Conf. on Pattern Recognition, Washington D.C., U.S.A., 231-234 (1973)

1.12 R.A.Bache, W.M.Gray & D.J.Murray-Smith, Time-domain system identification applied to noninvasive estimation of cardiopulmonary quantities, IEE Proc., 128, Pt.D, 56-64 (1981)

LUNG SOUND

2.1 Respiratory System

2.1.1 Structure

Respiration is a mechanism vital to life. It involves the taking of oxygen into the body and the removal of carbon dioxide. In primates the prinicpal organ responsible for respiration is the lung.

The respiratory system is conventionally divided into the upper and lower respiratory tracts; the upper tract comprising the nose, nasal sinuses and larynx, and the lower tract including the trachea, bronchi and the lungs. The lower tract can be further divided into two functional zones:- the conducting zone and the respiratory zone (or parenchyma).

The trachea extends down from the epiglottis, branching into the left and right main bronchi. The bronchi in turn subdivide into several generations of smaller airways. At some stage these successively smaller airways no longer contain cartilage and become known as **bronchioles.** These also divide and terminate in **alveolar ducts** and **alveoli.** The alveoli are sacs 250μm in size at full inflation and are the sites at which gases diffuse across the membranes between airways and blood vessels. Each lung contains approximately 300 million alveoli.

On the opposite side of the alveolar walls from the airways are the capillaries. These are part of the circulatory system and are responsible for distributing blood within the lungs. Although we are primarily concerned with the airways here, the capillary system in the

lung is vitally important.

Three main processes contribute to the functioning of the lung:

(1) **Ventilation** - the movement of air in and out of the lungs

(2) **Perfusion** - blood flow through the capillaries

(3) **Diffusion** - transfer of gases between alveoli and blood vessels along partial pressure gradients.

Ventilation has two phases, inspiration and expiration. Two movements are responsible for ventilation: (a) the movement of the rib cage by intercostal muscles and (b) the movement of the diaphragm. Clearly energy is expended in ventilating the lungs with contributions coming from the inertia of the rib cage and from resistance to airflow in the airways.

2.1.2 Overview of Respiratory Diseases

Before introducing the ideas of lung sound it is useful to classify respiratory diseases. For the purposes of this thesis we classify these diseases according to functional disturbance: (a) restrictive ventilatory defect and (b) obstructive ventilatory defect.

Before describing the defects it is useful to define two lung function parameters that are measured during a maximal forced expiration that follows a maximal inspiration. The volume of air breathed out during the first second of this manoevre is termed forced expired volume (FEV_1) and the total volume of air expired is termed forced vital capacity (FVC).

Restrictive defect is associated with parenchymal lung disease and is characterised by a marked reduction in lung volume. There is a proportional reduction in both FVC and FEV_1. Fibrosing alveolitis, allergic alveolitis and sarcoidosis are all examples of restrictive

defect.

In contrast obstructive lung disease is characterised by a marked reduction in FEV_1 and a smaller reduction in FVC. Obstructive defect is caused by diseases of the conducting airways such as asthma, bronchitis and emphysema.

2.1.3 Clinical Investigation of Respiratory Disease

Physical examination of a patient is of great importance in diagnosis of respiratory diseases [2.1]. Such procedures include:

1. General examination
2. Examination of the upper respiratory tract
3. Examination of the chest

The general examination will include looking for dysnoea (conscious awareness of breathing), cyanosis (blue colouration of the skin because of reduced haemoglobin), finger clubbing (increased curvature of finger nails leading to distortion of finger tips) and sputum examination.

The examination of the chest will include inspection of the respiratory rate and rhythm, palpitation, percussion and auscultation. In **auscultation** the clinician will listen for the quality and intensity of breath sounds, the presence of adventitious sounds and for vocal resonance.

Other important investigative techniques include chest radiography, blood examination, bronchoscopy, biopsy, pulmonary function tests and skin tests. Of these radiography has been especially valuable in clinical investigation in many diseases of the chest. Bronchoscopy and biopsy are of course invasive techniques.

2.1.4 Auscultation of the Lungs

Auscultation is one of the most widely used examination techniques but often it is considered to be of little clinical value. Lung sounds heard at the chest wall may be divided into a hierarchy. The major division is into **breath sound** which is always present during breathing and **adventitious** or **added** sounds that are not normally present.

Conventionally breath sound is classified into **normal** or **vesicular** breath sound and abnormal **bronchial** breath sounds. An intermediate sound is sometimes referred to as **bronchovesicular.** Normally breath sound is louder in inspiration than in expiration. Bronchial breathing is an abnormal condition caused by the direct transmission of sound through the chest wall resulting in a different quality and intesity of sound.

Adventitious sounds are normally divided into crackles, wheezes and pleural rub. Crackles are **discontinuous** sounds that occur in a number of diseases. These vary in quality and in position within the respiratory cycle. Wheezes are continuous 'musical' sounds with well defined frequency characteristics. Pleural rub is a creaking sound caused by the rubbing together of inflamed pleural surfaces.

The stethoscope will of course be used for listening to voiced sounds and percussion through the chest. However such sounds will not be considered in this thesis.

2.2 Lung Sound

In the last ten years there has been increasing interest in the study of lung sound. During that time there has been the formation of the International Lung Sound Association which aims to promote inter-disciplinary research into lung sound, and has organized an International Conference on Lung Sounds each year since 1976. Unfortunately many of the results on lung sound have appeared only in the lung sound conference abstracts rather in the open literature.

In this section we review the research into lung sound that is useful to this thesis. At the outset it is worth noting that some results are of immediate value to auscultation of the lungs whereas other results involve signal analysis which will require computational facilities.

2.2.1 Historical Background

The value of ausculation of the lungs has been known since Laënnec in the early 19th century. Laënnec in his classic work **L'Auscultation Mediate** [2.2] related sounds heard at the chest wall to anatomical features. He gave the name **rale** to describe the various added sounds heard through the stethoscope. He subdivided rales into four groups which he called moist, mucous, sonorant and sibilant sounds. Since then these terms have become modified in their use and have become more and more ambiguous.

In 1884 Bullar [2.3] carried out some experiments on the site of breath sound generation using an artificial thorax and sheep lungs. He concluded that breath sound was generated in parts of the respiratory tract where air passes from a narrower to a wider space.

In 1925 Cabot & Dodge [2.4] performed what was probably the first frequency domain investigation of lung sound. They played lung sound through a filter bank and were able to distinguish between certain types of crackles.

Very little work was done on lung sound until the work of McKusick et al in 1955 [2.5]. They applied the sound spectrogram to problems of percussion, lung and heart sound. This allowed plots of intensity, time and frequency to be plotted.

During the mid 20th century the importance of auscultation in diagnosis of respiratory diseases diminished with the increased use of the chest X-ray. Furthermore the ideas on lung sound were frequently erroneous. A real breakthrough came with the work of Forgacs in the 1960s and early 1970s which improved clinical knowledge of lung sound and produced far more reasonable explanations of the origins of lung sounds.

The 1970s have seen an increasing interest in lung sound research in the U.K., U.S.A., Japan, Europe and India. This has changed the emphasis from observations that are directly useful to stethoscope auscultation to more advanced signal processing techniques.

2.2.2 Terminology

A glance through relevant sections of some medical textbooks will indicate that there is considerable confusion about lung sound terminology. This has arisen for historical reasons and is discussed by Forgacs [2.6], Cugell [2.7], and Bunin & Loudon [2.8].

Laënnec originally described all added lung sounds as **rales.** This word was then in current use in France for the rattle of sputum heard

in dying patients. Apparently Laënnec used the Latin equivalent **rhonchi** in his casenotes. He used various adjectives to subdivide **rales** but with time the original usages changed and became less precise. Since that time the words rale and rhonchus have been applied to crackle and wheeze respectively.

In 1957 Robertson & Coope [2.9] proposed a new terminology at the end of a rather fanciful essay on Laënnec. The suggested terminology was much less amibiguous that that in use for added sounds. It was

1. Continouous sounds (a) High-pitched wheeze
   (b) Low-pitched wheeze
2. Interrupted sounds (a) Coarse crackling sounds
   (b) Medium crackling sounds
   (c) Fine crackling sounds (crepitations)

The division into continuous and interrupted sounds (wheezes and crackles) is acoustically accurate yet straightforward. Providing the subdivisions of these categories can be defined accurately the adjectives describing them should also be useful. For the purposes of this thesis we follow Forgacs [2.6] in using **crackles** and **wheezes** throughout.

Bunin & Loudon [2.8] give an intersting survey of the use of terminology in case reports up to July 1977. Their work suggests that the term crackle has rarely been adopted; the term **rale** being commonly used in American journals and the term **crepitation** being in frequent use in the British literature. The terms **rhonchi** and **wheeze** seem to enjoy a similar amount of usage. However with the publication of Forgacs' book [2.6] and the work of the International Lung Sound Association it is likely that the newer terminology will be used more frequently.

2.2.3 Breath Sound

Breath sound may be heard at a number of different sites. The term usually refers to the sound always present in breathing heard at the chest wall. However it may also be heard at the trachea or at the mouth. In contrast to continuous adventitious sounds (wheezes) the breath sound does not consist of one or more well defined frequencies but consists of filtered white noise.

Forgacs et al [2.10] drew attention to the use of breath sound at the mouth which is normally barely audible. In chronic bronchitis or asthma it can be heard at some distance away from the patient's mouth. He also showed that there was a linear relationship between the maximum breath sound amplitude and flow rate in both normal and obstructed lungs. He also notes the clinical value of abnormally loud and paradoxically quiet breath sound.

In contrast to breath sound heard at the mouth, breath sound heard at the chest wall is restricted by low pass filtering. There is no definite relationship between the loudness of breath sound heard at the mouth and the loudness of that heard at the chest wall.

The subject of the origin of breath sound has been contraversial from the early 19th century onwards. A number of different sites and mehcanisms have been suggested. Forgacs suggests a central turbulent source whereas Hardin & Paterson [2.11] suggest that it is caused by vortices at junctions between airways. This question is still largely unresolved.

Forgacs carried out some measurements on the attentuation and filtration of breath sound. He concluded that breath sound had an even frequency distribution between 200 and 2000Hz whereas sound recorded

at the chest wall falls at 10-20dB/octave from 200Hz.

Bronchial breathing is a condition associated with consolidation of the lung. Lung tissue may become airless between the chest wall and central airways and transmit sound with much less attenuation and filtration. The bronchial breathing sounds very similar to the tracheal breath sound and has a similar frequency range.

2.2.4 Crackles

Lung crackles have probably received more attention than any other type of lung sound. They are short explosive sounds heard either through the chest wall or through the mouth. Traditionally they have been attributed to the bubbling of secretions in the airways. While this is undoubtedly a reasonable explanation when the main bronchi contain sputum it does not explain crackling in cases of interstitial fibrosis when there is no liquid in the lung

Forgacs [2.12] offered an alternative explanation; he suggested that crackles might be caused by the abrupt opening of small airways in the lung. This is consistent with the observation that in fibrosing alveolis peripheral airways remain shut until late in the inspiration.

Nath & Capel [2.13] made a number of interesting observations on inspiratory crackles. They found that early inspiratory crackles were associated with diseases of airway obstuction including chronic bronchitis, asthma and emphysema. These crackles tend to be scanty, gravity dependent and were usually transmitted to the mouth. In contrast late inspiratory crackles tended to be associated with restrictive lung disorders such as fibrosing alveolitis, pneumonia, pulmonary oedema and asbestosis. In another study Nath & Capel [2.14]

made use of the repetitive nature of crackles were produced at a particular inspired volume rather than at a particular time from the beginning of inspiration.

Subsequent work has been very largely based on waveform analysis of crackles in both time and frequency domains. Two approaches have been prominent both being first used by Murphy's group in Boston: (i) time expanded waveform analysis [2.15] and (ii) spectral analysis based on the discrete Fourier transform (DFT) [2.16]. Time expanded waveform analysis is simply the observation of the nature of lung sound waveforms by chart recorder run with a suitably large time scale. While this is a very simple technique from the point of view of signal processing it has allowed details of waveforms to be used that were not used in older studies e.g. Forgacs and Nath & Capel. Spectral analysis using the DFT has been widely used because of the availability and speed of the fast Fourier transform algorithm.

A number of such studies have suggested that different types of crackle may be associated with different diseases. Murphy and Holford [2.17] were able to differentiate between crackles in asbestosis and cardiac failure using time expanded waveform analysis. Kudoh et al [2.18] noted differences between those in fibrosing alveolitis and bronchitis using sound spectrograms. Mori et al [2.19] studied crackles in tuberculosis using time and frequency domain analysis.

Regrettably there seems to have been very little of clinical value that has arisen from the many studies of crackles. One reason for this may have been inappropriate use of time and frequency domain waveform analysis techniques. Crackles also present a considerable data reduction problem because many of them can be present in one breath cycle. There would also appear to be a basic limit to the value

of crackles in the diagnosis of disease since crackles are usually associated with advanced rather than early stages of diseases such as asbestosis (see Epler et al [2.20]). Murphy however suggests a number of possible applications for analysis of crackles in the future [2.21].

2.2.5 Wheezes

Wheezes are continuous lung sounds that have a 'musical' quality. It is clear from the description 'musical' that these sounds are of well defined pitch. Wheezing is usually associated with obtructive disorders and consequently the sounds are generated in the conducting airways of the lungs. For many years it was assumed that they were generated by an organ pipe type of mechanism. Forgacs however produces convincing evidence to suggest that the wheeze is caused by a mechanism similar to that of a reed in a toy trumpet. He suggests that if the bronchus walls are in contact they will operate as a reed. More recently Grotberg [2.22] has done some theoretical analysis suggesting that flutter in a collapsible channel provides a good model for wheeze behaviour.

Forgacs [2.23] identifies four types of wheezing (a) fixed monophonic wheeze, (b) random monophonic wheeze, (c) sequential inspiratory wheeze and (d) expiratory polyphonic wheeze.

A fixed monophonic wheeze is usually a sign of an incomplete occlusion of a bronchus by a tumour or a foreign body. The random monophonic wheeze is caused by widespread airway obstruction e.g. in asthma. A sequence of short monophonic wheezes may occur in diffuse interstitial pulmonary fibrosis. Polyphonic wheezing is usually associated with widespread airway obstruction.

When compared with crackles relatively little research has gone into the signal processing of wheezes. A number of studies have used spectral analysis techniques to study the time varying frequency content of wheezes. However at the time of writing there would appear to be a lot of fruitful work that could be done on extracting further information from wheezes.

2.3 Directions in Lung Sound Research

In studying lung sound it is felt important to make distinctions in the aims of research projects. Here a distinction is made between a **physiological** and a **clinical** approach. Although these approaches are related they differ in their main objectives and thus may differ in the way experiments are performed.

Very briefly physiological studies of lung sound signals aim to understand the mechanisms of production and transmission of lung sound. This means that accurate measurement and estimation of physiological parameters is of prime importance.

In contrast clinical studies will aim to provide information capable of distingushing normal from abnormal, or disease A from disease B. To this end utility to diagnosis is of prime importance. The recording systems must be designed to operate in normal hospital conditions without elaborate arrangements being made. Some interference is tolerable providing it is of a constant nature over different recordings. The data analysis must aim at **differentiating** different types of sound rather than at accurate parameter estimation.

The investigations described here are primarily of a clinical nature. However it must be stressed that as far as possible attention

has been paid to the physiological factors involved.

References

2.1 M.Schonell, Respiratory Medicine, Churchill Livingstone (1974)

2.2 R.T.H.Laënnec, De l'Auscultation Mediate, ou traite du diagnostic des maladies des poumons et du coeur, fonde principalement sur le nouveau moyen d'exploration, Brosson & Chaude (1819)

2.3 J.F. Bullar, Experiments to determine the origin of respiratory sounds, Proc. R. Soc. London, **37**, 411-423 (1884)

2.4 R.C.Cabot & H.F.Dodge, Frequency characteristics of heart and lung sounds, J.Am.Med.Ass., **84**, 1793-1795 (1925)

2.5 V.A.McKusick, J.T.Jenkins & G.N.Webb, The acoustic basis of chest examination: studies by means of sound spectrography, Am.Rev.Tuberc., **72**, 12-34 (1955)

2.6 P.Forgacs, Lung Sound, Balliere Tindal (1978)

2.7 D.W.Cugell, Sounds of the lungs, Chest, **73**, 311-312 (1978)

2.8 N.J.Bunin & R.G.Loudon, Lung sound terminology in case reports., Chest, **76**, 690-692 (1979)

2.9 A.J.Robertson & R.Coope, Râles, rhonchi and Laënnec, The Lancet, **1**, 417-423 (1957)

2.10 P.Forgacs, A.R.Nathoo & H.D.Richardson, Breath Sounds, Thorax, **26** 288-295 (1971)

2.11 J.C.Hardin & J.L.Paterson, Monitoring the state of the human airways by analysis of respiratory sounds, Acta Astronautica, **6**, 1137-1151 (1979)

2.12 P.Forgacs, Crackles and wheezes, The Lancet, 2, 203-205 (1967)

2.13 A.R.Nath & L.H.Capel, Inspiratory crackles - early and late, Thorax, 29, 223-227 (1974)

2.14 A.R.Nath & L.H.Capel, Inspiratory crackles and the mechanical

events of breathing, Thorax, 29, 695-698 (1974)

2.15 R.L.H.Murphy, S.K.Holford & W.C.Knowler, Visual lung sound characterization by time-expanded waveform analysis, New England J. Med., 296, 968-971 (1977)

2.16 R.L.H.Murphy & K.Sorensen, Chest ausculatation in the diagnosis of pulmonary asbestosis, J.Occup.Med., 15, 272-276 (1973)

2.17 S.K.Holford & R.L.H.Murphy, Differentiation of the rales of pulmonary asbestosis and congestive cardiac failure, 1st Internat. Conf. Lung Sounds, Boston (1976)

2.18 S.Kudoh, A.Shibuya, N.Aisaka, I.Ono, A.Kurashima & R.Mikami, Analysis of rales in patients with fibrosing alveolitis by a new phonopneumographic method using a sound spectrograph, 2nd Internat. Conf. Lung Sounds, (1977)

2.19 M.Mori, K.Kinoshita, H.Morinari, T.Shiraishi, S.Koike & S.Murao, Waveform and spectral analysis of crackles, Thorax, 35, 843-850 (1980)

2.20 G.R.Epler, C.A.Carrington, E.A.Gaensler, Crackles (rales) in the interstitial pulmonary diseases, Chest, 73, 333-339 (1978)

2.21 R.L.H.Murphy, Auscultation of the lung: past lessons, future possibilities, Thorax, 36, 97-107 (1981)

2.22 J.B.Grotberg & S.H.Davis, Fluid-dynamic flapping of a collapsible channel: sound generation and flow limitation, J.Biomechanics, 13, 219-230 (1980)

2.23 P.Forgacs, The functional basis of pulmonary sounds, Chest, 73, 399-405 (1978)

PATTERN RECOGNITION AND SIGNAL PROCESSING

3.1 Introduction

Pattern recognition has been a significant field of study within information science and engineering since the early 1960s. It has grown with the increasing sophistication and availability of computer technology, and has drawn on many disciplines e.g. mathematics, statistics, control theory, psychology and formal language theory. Unfortunately there have been very few research journals devoted exclusively to this field.

Probably the most widely studied source of data in pattern recognition is pictorial data. Time varying signals of one or more channels are widely used in medicine and geophysics and constitute another important class of data. As with image data, a systematic approach to the pattern recognition of such signals is of great value. In this chapter an attempt is made to introduce the basic notions of pattern recognition and signal processing relavent to succeeding chapters with particular emphasis on the interface between and techniques common to these fields.

The pattern recognition process is frequently divided into a number of interacting components. In practice these components may not be easily distinguished, but give a convenient representation of the process. Chien [3.1] divides pattern recogniton into three basic stages:- data acquisition, pattern analysis and pattern classification. Data acquisition involves interfacing the pattern recogniton system to the real world situation under study. In biological signal analysis this stage would involve transducing,

conditioning and sampling the signals. Pattern analysis might involve finding suitable dimensionality reduction methods and exploratory data analysis. Pattern classification involves the design and implementation of decision logic for classifying input data items.

Another division of the pattern recognition field is by the approach made to the problem. In the **statistical** or **geometrical** approach [3.2] to pattern recognition data is explored or classified by representing it as an n-dimensional vector. The **structural** approach [3.3] in contrast presupposes that the data can be described recursively by simpler patterns. It was found that formal language theory was particularly appropriate to this approach which is why the terms **syntactic** and **linguistic** pattern recognition are used. Both these approaches have proved useful in pattern recognition of signals.

Signal processing or conditioning has always been an important part of the electrical engineering field. With the development of the computer, **digital signal processing** has added an important new dimension to this area. For example the ability to sample and store signals removes the usual time restraints and allows the time scale to be altered or even reversed. Additionally the discrete versions of the Fourier transform and correlation functions have changed what were primarily theoretical tools to practical ones. Operations to condition and analyse signals may be performed in either the time or frequency domains making the transitions between these domains especially important. In fact the study of spectral estimates has become an important field in its own right.

In applying pattern recognition techniques to signals the approach will depend on how much knowledge is available about the processes which generate the signals. If sufficient knowledge is

available, problems may be solved by modelling the process itself. Since this ideal situation will not always occur, some interfacing between a signal and a pattern recognition system may be possible by fitting a model to the signal. Finally at worst a purely pattern recognition approach may be used which assumes nothing about the incoming signals.

In this chapter some basic techniques of pattern analysis and signal processing are outlined. The emphasis here is on pattern analysis rather than pattern classification because of the almost exclusive use of pattern analysis techniques in this thesis. The presentation of material is intended to emphasize techniques common to the two areas of interest notably some linear transformations that may be applied to both feature extraction and spectral analysis. Finally segmentation and feature extraction techniques for signals are briefly reviewed.

3.2 Dimensionality Reduction

In statistical pattern recognition, data input usually takes place with a data matrix. Suppose we have a data set of N patterns $[X_1,X_2,\ldots,X_N]^T$ where each pattern is characterized by n **measurements** or **features** $[X_1,\ldots,X_n]$. Each pattern X_i corresponds to an n-dimensional vector i.e. $X_i = [x_{i1},x_{i2},\ldots,x_{in}]$. The data set may be represented by an N x n data matrix. We may view this data set in one of two ways:

(1) The set $P = \{p_1,p_2,\ldots,p_N\}$ of N points in an n-dimensional **primal space** [3.4] (or feature space) each p_i corresponding to a pattern.

(2) The set $Q = \{q_1,q_2,\ldots,q_n\}$ of n points in an N-dimensional **dual**

space [3.4] (or pattern space) each q_i corresponding to a feature.

In general pattern recognition algorithms have operated in the primal space. However Daly [3.4] points out that a problem posed in the dual space is sometimes easier to solve than the corresponding one in the primal space. Also primal-dual algorithms operating in both spaces may be of value. In the statisical literature the distinction is also apparent with principal coordinates analysis being computed in the primal space and principal components analysis being computed in the dual space.

In either space computation can take place by representing points by their cartesian coordinates. However geometric constructions offer a much more compact description of spatial ideas. Hence it may be convenient to transfer our data representation from the coordinates to geometric structures.

Frequently input data in pattern recognition is of a very high dimensionality and will have a certain amount of redundancy. This "curse of dimensionality" impairs an understanding of the nature of the data and can worsen the complexity of the algorithms involved. Two methods of dimensionality reduction may be distinguished, namely **feature selection** and **feature extraction.** Given input data of n-dimensions and a reduced dimensionality of m, we may either **select** a subset of m of the n input variables or **extract** m variables by choosing a subset of m variables in a transformed space.

The problem of dimensionality reduction is of course vital to exploring data. In this case mapping algorithms are frequently used to reduce the dimensionality of a data set to two or three dimensions and therefore allow a person to view the point set. This particular approach will be considered in the next section whereas more general

considerations will be dealt with here.

The aim of feature selection is to choose a subset having m of the input variables so that there is no need to use redundant or less useful ones. If we assign a cost to taking a particular set of measurements, feature selection will minimize the cost by removing the necessity of making all the measurements. In contrast, feature extraction utilizes all the input measurements and obtains a lower dimensional vector by some transformation. A number of reviews of dimensionality reduction techniques are available e.g. Levine [3.5], Kittler [3.6] and Toussaint [3.2].

Feature selection is frequently achieved by optimizing a criterion that is ~~based on~~ related to classification error. Among such criteria are those based on probabalistic distance measures (e.g. Mahalanobis distance), dependence measures and Euclidean distance. However the main drawback of these techniques is that the criterion must be evaluated $\binom{n}{m}$ times. Clearly this can lead to considerable computation as n grows larger, rapidly becoming impossible. Various suboptimal "top down" and "bottom up" approaches are available [3.2,3.7] which involve much less computation.

Feature extraction methods could be based on linear or nonlinear transformations of the input data. Among the best known methods are those based on the Karhunen-Loève and other rotational transformations, those based on finding discriminant vectors, and those based on separability measures. In contrast to feature selection techniques, feature extraction involves finding an optimal transformation matrix and is computationally less demanding.

3.3 Exploratory Data Analysis

In the pattern recogintion field problems of exploratory data analysis frequently arise. Given a multivariate data set what is its underlying structure? Do any 'natural' groupings exist in the data? These problems may be tackled by a number of techniques including clustering methods and mapping algorithms. This field is closely related to dimensionality reduction particularly with regard to some of the mapping techniques. Two main areas are briefly reviewed here - cluster analysis and mapping algorithms.

3.3.1 Cluster Analysis

Cluster analysis has proved to be a useful tool in biological studies using multivariate data. A number of methods have been developed for this purpose and have found ready acceptance in pattern recognition. Many contributions have been made to cluster analysis by pattern recognition researchers themselves. The fact that there are many different ways of interconnecting and grouping points has led to a proliferation of new techniques in different part of the literature. Suitability of a particular clustering technique depends a great deal on the application. For example in numerical taxonomy very specific requirements on cluster formation are made, whereas in pattern recognition it may be quite acceptable to cluster points in a way analagous to human visual perception. Also the appropriateness of a particular technique to the data is very important. To a greater or lesser degree a technique imposes a structure on the data and so the utility of a particular technique will depend on how appropriate the imposed structure is.

In general, approaches to cluster analysis have been heuristic rather than theoretical. Jardine, Jardine & Sibson [3.8], Jardine & Sibson [3.9], Sibson [3.10] and Wright [3.11,3.12] have developed formal approaches to cluster analysis. The approaches of these authors have led directly to methods based on their propositions [3.13,3.14]. In contrast most new techniques in the pattern recognition field have been based on applying techniques such as optimization or graph theory to the clustering problem.

With a very diverse literature, there is unfortunately a confusing range of terminology in current use. For example in the taxonomy literature the word **classification** is used synonymously with **clustering**. Here we restrict the use of the word classification to denote the assignment of unknown patterns into pre-specified classes.

Recently there has been mounting concern over the use of cluster analysis in pattern recognition. These methodological questions, raised largely through the work of Dubes and Jain, involve the interpretation of results from a clustering method and will be considered in Chapter 5. There are a number of reviews on cluster analysis in the literature e.g. Cormack [3.15], Everitt [3.16] and Scoltock [3.17].

3.3.2 Mapping Algorithms

Another approach to exploratory data analysis is to find a low dimensional representation of high dimensional data. Clearly if data can be adequately represented in two or three dimensions, the user can conceptualize geometry of the data set which may allow better judgements to be made on the problem. It is fairly obvious that

obtaining low dimensional displays of multivariate data is akin to dimensionality reduction problems such as feature selection.

Mapping techniques may be subdivided into iterative and non-iterative methods. A noniterative mapping can be calculated by a precise formula and will therefore be unique. Iterative techniques however involve optimization of some objective function which compares high and low dimensional representations. We will briefly review some linear non-iterative mappings and some nonlinear iterative mappings.

Probably the two best known types of linear transformation are the principal components (or Karhunen-Loève) transformation and discriminant vectors transformation. Each of these techniques is based on finding eigenvalues and eigenvectors of a matrix.

In the case of principal components an attempt is made to find a transformation that preserves the structure of the data in the least mean square error sense. If labels are available discriminant vectors may be used to find a transformation that gives the best representation according to a discrimination criterion.

Iterative mapping techniques have evolved in two similar ways. In the statistics literature such techniques have been known as **multidimensional scaling** [3.18-3.21] whereas in pattern recognition they have been called **nonlinear mappings** [3.22,3.23]. However the approaches may be distinguished. A major practical difference is that, in general, multidimensional scaling techniques are capable of using non-metric dissimilarities whereas nonlinear mappings in pattern recogniton tend to be based on the Euclidean metric. Let d^*_{ij} denote the distance between entities i and j in the higher dimensional space and let d_{ij} denote the distance in the lower dimensional space. Multidimensional scaling techniques attempt to satisfy a monotonicity

constraint i.e. one where the rank order of the d^*_{ij}s is the same as the d_{ij}s; in practice this is impossible to achieve and so an objective function is used to get as close to this as possible. Nonlinear mappings tend to be based on optimizing an error function between the two representations.

Mapping algorithms are reviewed by Chien [3.1], Everitt [3.24] and Terekhina [3.25].

3.4 Linear Transformations

Among the most commonly used techniques for feature extraction and mappings to lower dimensions are linear transformations. Among these techniques that are widely used are the principal components/Karhunen-Loève transformation and the discrete Fourier transform (DFT). Both these techniques may be approached from the point of view of either time series analysis or multidimensional feature rotations. In practical terms a multidimensional feature vector may be processed in the same way as a time series. In view of the uses to which these techniques are put in later chapters we explain the principal components transformation as a feature rotation and the DFT from the point of view of time series.

3.4.1 Principal Components/Karhunen-Loève Transformation

The title of this section suggests the all too frequent disparity between the statistics literature and the engineering literature when describing identical or similar techniques. An attempt is made to introduce the ideas behind this technique drawing from both literatures.

Frequently we encounter a data set of n correlated variables, and it would be useful to transform them to a new set of uncorrelated variables. Such a set of uncorrelated variables is termed the **principal components** of the data and will be a linear combination of the original variables.

Since our objective will be dimensionality reduction it is important that the first few principal components account for most of the variation in the data set.

Let $\underline{X}^T = [X_1,...,X_n]$ be an n-dimensional random vector variable having mean $\underline{\mu}$ and covariance Σ. We seek a vector $\underline{Y}$ of new variables $Y_1,..,Y_n$ which are uncorrelated and whose variances decrease from first to last. Each Y_j will be a linear combination of all the X_is

i.e. $Y_j = a_{1j}X_1 + a_{2j}X_2 + ... + a_{nj}X_n$

$= \underline{a}_j^T\underline{X}$

where $\underline{a}_j^T = [a_{1j},...a_{nj}]$ is a vector of constants subject to the condition $\underline{a}_j^T\underline{a}_j = 1 \quad \underline{a}_k^T\underline{a}_j = 0 \quad k \neq j$

The first principal component Y_1 is found by choosing $\underline{a}_1$ so that the variance of $\underline{a}_1^T\underline{X}$ is maximized subject to $\underline{a}_1^T\underline{a}_1 = 1$. Similarly the second principal component Y_2 is found by choosing $\underline{a}_2$ so that Y_2 has the maximum possible variance while being uncorrelated with Y_1. All the other Y_js are derived in the same way.

It can be shown (see e.g. Chatfield & Collins [3.26]) that the vectors $\underline{a}_j^T$ are the eigenvectors of the covariance matrix Σ where a particular vector $\underline{a}_j^T$ is the eigenvector corresponding to the jth largest eigenvalue. Furthermore the eigenvalues can be interpreted as the variances of the different components.

The Karhunen-Loève transformation in the engineering literature is defined in the same way. A number of variations in forming the

covariance matrix are possible depending on the purpose of the transformation and on whether a priori labels are available [3.27-3.29].

3.4.2 The Discrete Fourier Transform

The discrete Fourier transfom (DFT) has proved to be a useful tool in many aspects of electrical engineering. For the purposes of this chapter it is especially useful as a tool for feature extraction and time series analysis.

As in other linear transformations, each output variable Y_k is found by a linear combination of the input variables X_i

$$Y_k = \sum_{i=0}^{n-1} X_i \exp(-2j\pi ki/n)$$

where $j = \sqrt{-1}$. In general the X_is may be complex numbers and the Y_ks are always complex. This expression is usually written as

$$Y_k = \sum_{i=0}^{n-1} X_i \, W^{ki} \qquad \text{where } W = \exp(-2j\pi/n)$$

The DFT is then given by Y_k for k=0,...,n-1

Clearly if we consider the X_is as samples of a time series we obtain the familiar use of the DFT to transform the series to a sampled frequency domain representation Y_k.

The DFT has come into widespread use through the availability of the fast Fourier transform (FFT) algorithm [3.30] which has allowed rapid computation of the DFT. Computation of the DFT requires computation proportional to n^2 whereas the FFT algorithm requires only $O(n\log_2 n)$ time. With such an efficient algorithm the DFT has become a popular tool in spectral analysis and feature extraction. The properties of the DFT and the development of the FFT algorithm are described by many authors [3.31-3.33].

3.5 Spectral Estimation

This section briefly reviews techniques for estimating power spectra. Although a variety of techniques are mentioned, we concentrate mainly on those based on the discrete Fourier transform since they are computationally efficient. The discrete Fourier transform is also used in biological signal analysis to obtain amplitude and phase spectra [3.55]; however this section is restricted to power spectral estimation. A comprehensive review of spectral analysis techniques is given in a paper by Kay & Marple [3.34] and in a reprint series edited by Childers [3.35].

One of the first pioneering steps in spectral analysis was the "periodogram" approach of Schuster [3.36] who studied variation in sun spot numbers. Wiener discovered the relationship between autocorrelation and power spectral density which was implemented in the moving average (MA) approach of Blackman & Tukey [3.37]. This became the most popular method of spectral estimation until the development of the FFT algorithm, which gave a fast route to the periodogram estimate.

During the late 1960s several modelling approaches to spectral estimation were developed. The maximum entropy method (MEM) [3.38,3.39] and autoregressive method (AR) [3.40] were shown to be equivalent for one dimensional data [3.41]. These methods offered an improvement over methods based on the DFT, especially for short data records, but at a cost of additional computation.

We concentrate now on the properties of direct spectral estimation based on the DFT. Firstly we must consider some preliminary

results. Consider a deterministic analogue signal x(t). Assuming that the signal energy is finite the continuous Fourier transform (CFT) X(f) exists and is given by

$$X(f) = \int_{-\infty}^{\infty} x(t)\exp(-j2\pi ft)dt$$

The term "spectrum" S(f) of x(t) is frequently used to describe the squared modulus of the CFT.

$$S(f) = |\ X(f)\ |^2$$

S(f) is an energy spectral density (ESD).

If instead of a continuous function x(t) we have a data sequence x_n sampled at equally spaced intervals Δt over a finite time window (n=0 to n=N-1) we may develop the discrete Fourier transform (DFT) consisting of N equally spaced frequency values (frequency spacing $\Delta f = 1/N\Delta t$).

$$X_m = \Delta t \sum_{n=0}^{N-1} x_n \exp(-2j\pi m\Delta f n\Delta t)$$
$$= \Delta t \sum_{n=0}^{N-1} x_n \exp(-2j\pi mn/N)$$

We may now define the periodogram ESD estimate by

$$S_m = |\ X_m\ |^2 \text{ for } m=0,\ldots N-1$$

If we consider the signal to be given by a wide sense stationary process the derivation is somewhat different. For a stationary random process the autocorrelation function is given by

$$R_{xx}(\tau) = E[x(t+\tau)x^*(t)]$$

This is related to the power spectral density (PSD) P(f) by

$$P(f) = \int_{-\infty}^{\infty} R_{xx}(\tau)\exp(-2\pi f\tau)d\tau$$

If we additionally assume that the process is ergodic in first and second moments, we may substitute time averages for ensemble averages and give the autocorrelation function by

$$R_{xx}(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t+\tau)x^*(t)dt$$

The power spectral density may then be given by

$$P(f) = \lim_{T\to\infty} \frac{1}{2T} \left| \int_{-T}^{T} x(t)\exp(-j2\pi ft)dt \right|^2$$

The (raw) periodogram estimate of the PSD is a sampled data version of the above expression.

$$\hat{P}_{per}(f) = \frac{1}{N\Delta t} \left| \Delta t \sum_{n=0}^{N-1} x_n \exp(-j2\pi fn\Delta t) \right|^2$$

$$\hat{P}_{per}(f_m) = \hat{P}_m = \frac{1}{N\Delta t} |X_m|^2$$

where the X_m are the values of the DFT of the sequence x_n.

A blind use of the periodogram spectral estimate may be misleading. This is mainly because the variance of the estimate is poor [3.39]. The expression for variance is given by

$$\text{var}[P_{per}(f)] = P(f)^2 + O(N^{-1})$$

and if r and s are integers and $g=2\pi r/N$ and $h=2\pi s/N$ then the covariance is given by

$$\text{cov}[P_{per}(g),P_{per}(h)] = O(N^{-1})$$

These results are crucial since the variance of the periodogram estimate will not change no matter how large N is. Also individual ordinates in the PSD estimate have small covariances compared with their variances giving rise to the irregular appearance of the periodogram. In fact it can be shown that the ordinates are asymptotically independent chi-squared variables with two degrees of freedom.

Two approaches have been suggested for obtaining more consistent spectral estimates from the periodogram. The first is to smooth the periodogram and the second is to average several periodograms.

Consider an ordinate X_m in the periodogram. From the previous arguments it is clear that the p ordinates before and p ordinates after X_m may be considered approximately independent chi-squared variates. This suggests that a better estimate would be an average of the periodogram ordinates in the neighbourhood of X_m.

* For details of proof see A. PAPOULIS, PROBABILITY, RANDOM VARIABLES AND STOCHASTIC PROCESSES McGRAW-HILL 1965

i.e.

$$P_{smooth}(f_m) = 1/(2p+1) \sum_{j=-p}^{p} P_{per}(f_{m+j}) \quad m \neq 0$$

it can be shown that

$$var[P_{smooth}(f_m)] = (1/(2p+1)).P_m(f_m)^2 + O(N^{-1}) \quad m \neq 0$$

and hence that the averaging of $(2p+1)$ periodogram ordinates has resulted in a reduction in variance by a factor of $(2p+1)$. The smoothed ordinates are now chi-squared variates with $(4p+2)$ degrees of freedom. However the bias of the spectral estimate is likely to increase with m resulting in a tradeoff between the two parameters [3.42].

This approach was suggested by Daniell [3.43] in 1946. Subsequently other more complex filters have been suggested for spectral smoothing. Bartlett [3.44] developed an alternative approach which was based on dividing the time series into segments and averaging the periodograms of individual segments. This also leads to a reduction in the variance of the estimate. Welch [3.45] suggested an FFT-based procedure for doing this following the work of Bingham et al [3.46] on modified periodograms.

The frequency resolution of the periodogram estimates are limited by the (explicit or implicit) windowing applied to the input time series. If the data is not tapered (implying a rectangular window) the DFT contains significant sidelobes. This is because the multiplication of the input time series implies convolution of the desired transform with the transform of the window in the frequency domain. The effect of the sidelobes is known as spectral leakage. Spectral leakage can be reduced by a choice of a suitable data window. Harris [3.47] provides a detailed comparison of various windows, some of his results being corrected by Nuttall [3.48].

In contrast to the problems of the periodogram some of the more recent modelling approaches provide both good estimates and good frequency resolution. Windowing is not required and smooth spectral estimates may be obtained directly. However there are additional problems that are associated with fitting a model to the signal. A big problem is to select the correct model order; if the order is too low the estimate will be smooth, and if the model order is too high the estimate will contain spurious detail.

3.6 Segmentation & Feature Extraction Schemes

Although not actually implemented systematically during the course of this research it is worth briefly considering the problem of feature extraction from a signal. This involves some of the common ground between pattern recognition and signal processing alluded to already.

In the consideration of spectral estimation techniques above, it was noted that the time series was assumed to be stationary. Many signals are locally stationary but exhibit longer term changes. It may then be justifiable to compute local spectral estimates for each stationary interval. Signal analysis may then be considered in two stages:- (1) the segmentation of the signal into quasi-stationary intervals and (2) finding a representation for the signal over that interval. This type of approach is outlined by Sanderson & Segen [3.49] who applied this approach to EEG analysis.

The first stage depends on detecting change in a time series. Several approaches have been suggested, Segen & Sanderson [3.50] have developed an approach based on a transformed sequence, and Büdenstein & Praetorius [3.51] use a spectral error measure based on linear

prediction [3.49]. The ideal way of representing a segment involves some sort of model. Autoregressive models have been frequently suggested as a basis for specifying signal segments [3.47,3.48,3.50]. It is to be hoped that the model parameters provide suitable features for a pattern recognition system.

A number of advantages arise from a piecewise stationary approach to signal analysis. Appropriate segmentation will allow preservation of local detail such as transients which are often significant in biological signal analysis. Clustering of segment parameters has been suggested as a means of data reduction. If the segments cluster well the signal may be represented by a sequence of symbols. This then allows a linguistic [3.54] as opposed to statistical approach to be used.

References

3.1 Y.T.Chien, Interactive Pattern Recognition, Marcel Dekker (1978)

3.2 G.T.Toussaint, Recent progress in statistical methods applied to pattern recognition, Proc. 2nd Int. Joint Conf. on Pattern Recognition, 479-488, Copenhagen, Denmark (1974)

3.3 T.Pavlidis, Structural Pattern Recognition, Springer-Verlag (1977)

3.4 J.A.Daly, Some dual problems in pattern recognition, Pattern Recognition, 3, 73-84 (1971)

3.5 M.D.Levine, Feature extraction: a survey, Proc. IEEE, 57, 1391-1407 (1969)

3.6 J.Kittler, Mathematical methods of feature selection in pattern recognition, Int. J Man-Machine Studies, 7, 609-637 (1975)

3.7 J.Kittler, Feature set search algorithms, NATO Advanced Study Institute on Pattern Recognition and Signal Processing, Paris, France, Sijhoff & Noordhoof, 41-60 (1978)

3.8 C.Jardine, N.Jardine & R.Sibson, The structure and construction of taxonomical hiearchies, Math. Biosciences, 1, 173-179 (1967)

3.9 N.Jardine and R.Sibson, A model for taxonomy, Math. Biosciences, 2, 465-482 (1968)

3.10 R.Sibson, A model for taxonomy II, Math. Biosciences, 6, 405-430 (1970)

3.11 W.E.Wright, A formalization of cluster analysis, Pattern Recognition, 5, 273-282 (1973)

3.12 W.E.Wright, An axiomatic specification of Euclidean analysis, Computer J., 17, 355-364 (1974)

3.13 N.Jardine & R.Sibson, The construction of hierarchic and non-

hierarchic classifications, Computer J., 11, 177-184 (1968)

3.14 W.E.Wright, Gravitational clustering, Pattern Recognition, 9, 151-166 (1977)

3.15 R.M.Cormack, A review of classification, Jl.R.Statist.Soc.Ser.A, 134, 321-367 (1971)

3.16 B.Everitt, Cluster analysis, Heinemann Educational (1974)

3.17 J.Scoltock, A survey of the literature of cluster analysis, Computer J., 25, 130-134 (1982)

3.18 R.N.Shepard, The analysis of proximities: multidimensional scaling with an unknown distance function I, Psychometrika, 27, 125-140 (1962)

3.19 R.N.Shepard, The analysis of proximities: multidimensional scaling with an unknown distance function II, Psychometrika, 27, 219-246 (1962)

3.20 J.B.Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, Psychometrika, 29, 1-27 (1964)

3.21 J.B.Kruskal, Nonmetric multidimensional scaling: a numerical method, Psychometrika, 29, 115-129 (1964)

3.22 J.W.Sammon, A nonlinear mapping for data structure analysis, IEEE Trans. Comput., C-18, 401-409 (1969)

3.23 C.L.Chang & R.C.T.Lee, A heuristic relaxation method for nonlinear mapping in cluster analysis, IEEE Trans. Systems, Man & Cybernet., SMC-3, 197-200 (1973)

3.24 B.Everitt, Graphical techniques for multivariate data, Heinemann Educational (1978)

3.25 A.Y.Terekhina, Methods of multidimensional data scaling and visualization - a survey, Automation and Remote Control, 34, 1109-1121 (1973)

3.26 C.Chatfield & A.J.Collins, Introduction to Multivariate Analysis, Chapman & Hall (1980)

3.27 S.Watanabe, Proc 4th Prague Conf on Information Theory (1965)

3.28 Y.T.Chien & K.S.Fu, On the generalized Karhunen-Loeve expansion, IEEE Trans. Information Theory, IT-13, 518-520 (1968)

3.29 J.Kittler & P.C.Young, A new approach to feature selection based on the Karhunen-Loeve expansion, Pattern Recognition, 5, 335-352 (1973)

3.30 J.W.Tukey & J.W.Tukey, An algorithm for machine computation of complex Fourier series, Math. Comput, 19, 297-301 (1965)

3.31 W.T.Cochran et al, What is the fast Fourier transform?, IEEE Trans. Audio & Electroacoustics, AU-15, 45-55 (1967)

3.32 J.W.Cooley, P.A.W.Lewis & P.D.Welch, Application of the fast Fourier transform to computation of Fourier integrals, Fourier series and convolution integrals, IEEE Trans. Audio & Electroacoustics, AU-15, 85-90 (1967)

3.33 E.O.Brigham, The fast Fourier transform, Prentice-Hall (1974)

3.34 S.M.Kay & S.L.Marple, Spectrum analysis - a modern perspective, Proc. IEEE, 69, 1380-1419 (1981)

3.35 D.Childers (Ed.), Modern spectrum analysis, IEEE Press (1978)

3.36 A.Schuster, The periodogram of magnetic declination as obtained from the records of the Greenwich Observatory during the years 1871-1895, Trans. Camb.Phil.Soc., 18, 107-135 (1899)

3.37 R.B.Blackman & J.W.Tukey, The measurement of power spectra from the point of view of communication engineering, Dover (1959)

3.38 J.P.Burg, Maximum entropy spectral analysis, Proc. 37th Meeting Society Exploration Geophysicists, Oklahoma City, USA (1967)

3.39 J.P.Burg, A new analysis technique for time series data, NATO

Advanced Study Institute with emphasis on underwater acoustics, Enschede, Neth. (1968)

3.40 E.Parzen, Statistical spectral analysis (single channel case) in 1968, Tech.Rept.12, Stanford University (1968)

3.41 A.van den Bos, An alternative interpretation of maximum entropy spectral analysis, IEEE Trans. Information Theory, IT-17, 493-494 (1971)

3.42 D.R.Brillinger, Time series data analysis and theory, Holt, Rinehart & Winston (1975)

3.43 P.J.Daniell, Discussion of paper by M.S.Bartlett, J.Roy.Statist.Soc., Suppl.8, 27 (1946)

3.44 M.S.Bartlett & J.Medhi, On the efficiency of procedures for smoothing periodograms from time series with continuous spectra, Biometrika, 42, 143-150 (1955)

3.45 P.D.Welch, The use of fast Fourier transform for the estimation of power spectra: a method based on averaging over short, modified periodograms, IEEE Trans. Audio Electroacoustics, AU-15, 70-73 (1967)

3.46 C.Bingham, M.D.Godfrey & J.W.Tukey, Modern techniques of power spectrum estimation, IEEE Trans. Audio & Electoacoustics, AU-15, 56-66 (1967)

3.47 F.J.Harris, On the use of windows for harmonic analysis with the discrete Fourier transform, Proc.IEEE, 66, 51-83 (1978)

3.48 A.H.Nuttall, Some windows with very good sidelobe behaviour, IEEE Trans. Acoustics, Speech and Signal Processing, ASSP-29, 84-89 (1981)

3.49 A.C.Sanderson & J.Segen, A pattern-directed approach to signal analysis, Proc. 5th Internat. Conf. on Pattern Recognition, Miami, USA, 613-617 (1980)

3.50 J.Segen & A.C.Sanderson, Detecting change in a time series, IEEE

Trans. Information Theory, IT-26, 249-255 (1980)

3.51 G.Bödenstein & H.M.Praetorius, Feature extraction from the electroencephalogram by adaptive segmentation, Proc.IEEE, 65, 642-652 (1977)

3.52 J.Makhoul, Linear prediction: a tutorial review, Proc. IEEE, 63, 561-580 (1975)

3.53 A.C.Sanderson, J.Segen & E.Richey, Hierarchical modelling of EEG signals, IEEE Trans. Patt. Anal. & Mach. Intelleg., PAMI-2, 405-415 (1980)

3.54 V.Mottl′ & I.Muchnik, Linguistic analysis of experimental curves, Proc. IEEE, 67, 714-736 (1978)

3.55 B.McA.Sayers, Exploring biological signals, Biomedical engineering, 10, 335-341 (1975)

GEOMETRIC STRUCTURE IN PATTERN RECOGNITION

4.1 Introduction

In analysis of multivariate data, we are confronted with a set of N n-dimensional measurements which might be represented spatially by a cartesian coordinate system. In pattern classification we are concerned with deciding whether a given vector lies in a particular class. In pattern analysis or exploratory data analysis we seek to find structure in the data set. Clearly the analysis of multivariate point sets suggests a geometric approach; yet until recently such problems were not tackled in an explicitly geometric way. With the growth of computational geometry, new approaches have been suggested based on formal geometric structure.

Probably the first suggestion of the power of computational geometry was in Shamos and Hoey's use of the Voronoi diagram to solve a wide range of geometric problems [4.1]. This showed that by developing computationally efficient algorithms for fundamental structures such as the Voronoi diagram, efficient algortihms were also available for a wider range of problems.

More recently computational geometry has been applied to a number of problems in pattern recognition. Examples include nearest neighbour decision rules [4.2], the shape of a set of points [4.3], decomposition of polygons [4.4], and cluster analysis [4.5]. For comprehensive reviews of work in this area, the reader is referred to Toussaint [4.6,4.7].

In this chapter we first consider the basic geometry of a multivariate point set, then define some fundamental geometric

structures and list some of their properties. For this it is necessary to introduce the ideas of computational complexity and graph theory. Finally we consider other structures based on the idea of the region of influence [4.5].

4.1.1 Simple Geometry in n-Dimensions

Before discussing geometric structures in pattern recognition it is worth considering some simple terms in n-dimensional geometry; a full treatment is given in Kendall [4.8].

We are concerned with a space S_n typefied by the variables $(X_1, X_2, \ldots, X_n)$ where the X_is can take any real value. An equation in the variables X_i defines a subspace of S_n which may be termed a **variety**. If the degree of the equation defining the variety is r, the variety is said to have **order** r and is denoted by V^r_{n-1}. A point in the subspace will be specified by n-1 coordinates.

Linear spaces are of particular importance. These are varieties of order 1 and are known as **flats** or **hyperplanes**. In S_n p linear equations define an (n-p)-flat. Clearly a p-flat is a flat space of p-dimensions.

A **polygon** is a very useful structure in 2-dimensions and is being increasingly used in pattern recognition [4.4]. The generalization of a polygon in n-dimensions is termed a **polytope**. A polytope is a figure bounded by a set of (n-1)-flats. Clearly two (n-1)-flats will intersect meeting at an (n-2)-flat.

The minimal number of (n-1)-flats that can enclose a space to form a polytope is (n+1). Hence in two dimensions such a structure is a triangle; in three dimensions it is a tetrahedron and in n-dimensions a **simplex**.

Taking Kendall's example it is useful to consider the simple properties of the simplex.

A triangle has 3 sides and 3 vertices.

A tetrahedron has 4 faces, 6 sides and 4 vertices.

A 4-dimensional simplex has $\binom{5}{1}$ 3-flats meeting in $\binom{5}{2}$ = 10 ways to form 2-flats. These meet in triplets in $\binom{5}{3}$ ways to form 1-flats and in sets of 4 in $\binom{5}{4}$ ways to form vertices.

The case for any number of dimensions can be worked out in a similar fashion. These properties of the simplex are essential to an understanding of say the Delaunay structure in higher dimensions.

4.1.2 Computational Complexity

Whenever a computational method is being applied it is obviously an advantage if the computer resources available are used most efficiently. Whether an algorithm is for serial or parallel processing, a vital consideration is the computation time and in particular the effect on time if the problem grows in size. One of the most spectacular examples of this is the discrete Fourier transform which was computationally infeasible until the development of the fast Fourier transform algorithm [4.9].

In order to consider the effect of size of input N on running time we assume a random access machine with infinite precision and asymptotic time complexity [4.6].

Firstly we define the commonly used O- and Ω-notation.

Let $f(N) = O(g(N))$ iff there exists a positive constant c such that $|f(N)| < c.|g(N)|$ for all N above some finite value.

Let $f(N) = \Omega(g(N))$ iff there exists a positive constant c such

that $|f(N)| > c\ |g(N)|$ for all N above some finite value.

If the two conditions exist for a particular algorithm then $O(f(N))$ and $\Omega(f(N))$ are the upper and lower bounds for the time complexity of the problem.

An algorithm is said to be **optimal** iff the upper and lower bounds are equal to within a positive constant $f(N) = O(g(N)) = \Omega(g(N))$. For a particular size of input N the **worst case complexity** is defined to be the maximum complexity over all possible inputs. Hence an 'O(g(N)) algorithm' is an algorithm with worst case complexity O(g(N)). If we assume a distribution of inputs, we may define the **expected complexity** to be the average complexity over all inputs of size N from that distribution.

The distinction between worst case and expected complexity is important [4.10]. It is possible for an algorithm to have a lower bound on worst case complexity that is greater than its average case complexity. An example of this is the 2-dimensional convex hull problem where many optimal algorithms have been developed (i.e. O(NlogN) and Ω(NlogN) worst case performance) but for which Bentley and Shamos found an O(N) expected time algorithm for some inputs [4.10].

Many of the results in geometric complexity are only for sets of points in the plane, but apart from some image processing problems, data sets used in pattern recognition are generally of higher dimensionality. Bentley and Shamos [4.11] and Bentley [4.12] derive some results in geometric complexity in higher dimensions but a significant problem is that an algorithm may be efficient in terms of N the number of points but may increase exponentially with dimensionality n. Thus some algorithms that are inefficient in terms

of N may run faster than those efficient in N in higher dimensions.

4.1.3 Graph Theory

Graph theory has emerged as a tool in many fields. Because of the intuitive appeal of representing data by a set of points and a set of interconnections, graph theory has received wide use in pattern recognition. In this section we define a graph and describe some simple properties.

A **simple graph** G (Fig 4.1(a)) is the pair (V(G),E(G)) where V(G) is a non-empty finite set of **vertices** (or **nodes** or **points**) and E(G) is a finite set of unordered pairs of elements of V(G) called **edges**; alternatively V(G) is the **vertex set** and E(G) is the **edge set** of G.

A **general graph** G (Fig 4.1(b)) is defined as above but allowing the existence of **loops** (edges joining vertices to themselves) and allowing **multiple edges** between a pair of vertices.

Two vertices of a graph G are said to be **adjacent** iff there is an edge joining them. The **degree** of a vertex G is the number of edges incident on that vertex.

Two graphs G_1 and G_2 are said to be **isomorphic** iff there is a one to one correspondance between the vertices of G_1 and G_2 such that the number of edges joining a pair of vertices in G_1 is equal to the number joining the corresponding pair in G_2.

Two graphs G_1 and G_2 are said to be **homeomorphic** iff they can both be obtained from the same graph G_3 by adding new vertices of degree two into the edges of G_3.

A **subgraph** S of G is merely a graph all of whose vertices lie in V(G) and all of whose edges lie in E(G) i.e. $V(S) \subseteq V(G)$ and $E(S) \subseteq E(G)$. A **supergraph** of G is a graph of which G is a subgraph.

A simple graph which has every pair of vertices adjacent is called a **complete graph.** A complete graph on n vertices is denoted by K_n (Fig 4.1(c)). A **bipartite graph** is a graph whose vertex set V can be divided into two disjoint subsets V_1 and V_2 such that every edge of G joins a vertex of V_1 to a vertex of V_2. A **complete bipartite graph** $K_{r,s}$ (Fig 4.1(d)) is a graph with r and s members of V_1 and V_2 respectively and where any given member of V_1 is connected to every member of V_2.

A **connected graph** is a graph that cannot be expressed as the union of two graphs. Clearly any **disconnected graph** can be expressed as the union of a number of connected subgraphs. More informally in a connected graph it is possible to travel from a given vertex to any other vertex by travelling along graph edges.

An **elementary contraction** of a graph G is obtained by identifiying two adjacent points u and v so that u and v are replaced by a new point w which is adjacent to the neighbours of both u and v. A **contraction** of a graph G (Fig 4.2) is obtained by applying one or more elementary contractions.

A **disconnecting set** e of a graph G is a set of edges of G whose removal disconnects G. If additionally no subset of e is a disconnecting set then e is a **cutset.** If a cutset contains only one edge it is called a **bridge.** A **separating set** v of a graph G is a set of vertices of G whose removal disconnects G. If this set has one member the vertex is called a **cut-vertex** (or cutnode or cutpoint).

A **nonseparable graph** is connected and has no cutpoints (e.g. the graph of Fig 4.2(a)). A **block** of a graph (Fig 4.3) is a maximal nonseparable subgraph. An alternative name for a block is **biconnected component.** An alternative definition for block is that a subgraph is a

block iff for every distinct triple of vertices p_i, p_j, p_k there exists a path between p_i and p_j not including p_k for sets of points with three or more vertices.

A **plane graph** is a graph drawn in the plane in such a way that no two edges intersect except at a vertex to which they are both incident. A **planar graph** is any graph isomorphic to a plane graph. There are a number of special properties of planar graphs with Kuratowski's theorem and Euler's formula being particularly useful.

Theorem 4.1 (Kuratowski) A graph is planar iff it contains no subgraph homeomorphic to K_5 or $K_{3,3}$.

Clearly this result is useful in proving whether or not a graph is planar.

If we consider a plane graph, it is clear that it defines regions known as **faces.** There will always be one region that is unbounded known as the **exterior face.** The number of faces may be found using **Euler's formula.**

Theorem 4.2 (Euler's formula) Let G be a plane graph with l vertices, m edges and n faces. Then l-m+n=2.

From a consideration of Euler's formula we can obtain the useful upper bound for the number of edges in a planar graph having 3 or more vertices. Since every face is bounded by at least 3 edges and since each edge divides two faces $3n < 2m$. Substituting into Euler's formula we obtain Corollary 4.2A.

Corollary 4.2A If G is a simple connected planar graph with N vertices, the number of edges m in G satisfies $m < 3N-6$

A **forest** is a simple graph containing no circuits and a connected forest is called a **tree.** A **spanning tree** is a connected graph consisting of one tree.

4.2 Outline of some Basic Geometric Structures

Six basic geometric structures are introduced and briefly reviewed, each being applicable to a wide range of problems in pattern recognition. These are (1) the convex hull (CH), (2) the Voronoi diagram, (3) the Delaunay triangulation (DT), (4) the Gabriel graph (GG), (5) the relative neighbourhood graph (RNG), and (6) the minimal spanning tree (MST). The convex hull is included for completeness rather than for detailed discussion later. For simplicity we first consider these structures in the plane (Fig 4.4) but we note that each generalizes to higher dimensions. In the following subsections let $P = \{p_1, p_2, .., p_N\}$ denote a set of N distinct points in the plane.

For the same set of points P these six structures are closely related. The convex hull is a subgraph of the Delaunay triangulation and the vertices on the convex hull are the same vertices as those having tiles of infinite size in the Voronoi diagram. Four structures, the DT, the GG, the RNG and the MST are very closely related i.e.

$$DT \supseteq GG \supseteq RNG \supseteq MST$$

4.2.1 The Convex Hull (CH)

The convex hull (CH) of the set P (Fig 4.4(a)) is the minimum area convex set of vertices of P. The convex hull problem may be subdivided into two problems depending on the input data - the convex hull of a set of points and the convex hull of a polygon [4.6]. The convex hull has a wide range of applications in pattern recognition including tests for linear separability, cluster admissibility, and

concavity, describing the shape of a set of points and image processing. More recently Edelsbrunner et al [4.3] have considered a generalization of the convex hull, the α-hull, and have shown its importance in finding the shape of a set of points. There is a considerable literature describing convex hull algorithms, but they will not be discussed in any detail here.

4.2.2 The Voronoi Diagram

The construct known as the Voronoi diagram [4.13] in computational geometry, the Dirichlet tesselation [4.14] in mathematics and Thiessen polygons [4.15] in geography is very widely used. Applications range from interpolation and finite element analysis [4.16] to nearest neighbour decision rule editing [4.2].

The Voronoi diagram (Fig 4.5(b)) consists of N disjoint regions or tiles, each tile enclosing one point p_i. A tile T_i is defined by

$$T_i = \{x: d(x,p_i) < d(x,p_j) \text{ for all } i \neq j\}$$

In general tiles meet in threes at points known as **Voronoi points** - unless there are four or more cocircular points. The straight line segments at the boundaries of adjacent tiles are called **Voronoi edges.** Each tile consists of a finite convex polygon except for those points lying on the convex hull, whose tiles extend to infinity.

Shamos and Hoey [4.1] describe an O(NlogN) optimal algorithm for finding the Euclidean planar Voronoi diagram based on a divide-and-conquer approach. A number of algorithms have been proposed for finding the n-dimensional Voronoi diagram including those of Brown [4.17] and Bowyer [4.18].

Some theoretical results are available on the properties of the Voronoi diagram including those of Sibson [4.19] and Miles

[4.20,4.21]. Lee and Wong [4.40] consider computation of the Voronoi diagram in L_1 and L_∞ metrics.

4.2.3 The Delaunay Triangulation (DT)

Considering the Voronoi diagram of a point set P, we can form a triangulation by joining a pair of points p_i,p_j iff they share a common Voronoi edge. The resulting triangulation is called the Delaunay [4.22] or locally equiangular triangulation [4.23]. This triangulation has a number of useful properties including the circle criterion [4.1].

Lemma 4.1 (Circle criterion). Any edge p_i,p_j is an edge of the Delaunay triangulation iff there exists a point x such that the circle centred at x passing through p_i and p_j contains no other points from P (Fig 4.5(a)).

Corollary 4.1A Any edge (p_i,p_j) on the convex hull of P is an edge of the Delaunay triangulation.

Lemma 4.2 The triangle p_i,p_j,p_k is a Delaunay triangle iff its circumcircle contains no other points of P.

In addition to the circle criterion, the Delaunay triangulation possesses the locally equiangular property. Sibson [4.23] showed that the circle criterion and locally equiangular property were equivalent. Consider a set of four points a,b,c,d forming a convex quadrilateral adbc in the plane. This quadrilateral may be triangulated by using either ab or cd as diagonals. We now state the **max-min angle criterion** which is used to make the triangualtion as nearly equiangular as possible.

Lemma 4.3 (Max-min angle criterion) If two triangles in a

triangulation have a common edge, they form a quadrilateral with that edge as a diagonal. If that quadrilateral is strictly convex, then the replacement of the existing diagonal by the alternative one must not increase the minimum of the six angles in the two triangles forming the quadrilateral.

Either criterion will yield a triangulation which is a dual of the Voronoi diagram unless degeneracies exist. Suppose we have a cyclic pentagon abcde, then from the Voronoi diagram we see that there are no edges which will triangulate the pentagon (Fig 4.5(b)). Instead we have a Delaunay pentagon or more generally a Delaunay polygon. Sibson [4.23] calls the construct with Delaunay polygons the **Delaunay pretriangulation** and the triangulation formed by the arbitrary triangulation of all Delaunay polygons the **completion** of the Delaunay pretriangulation. A misconception was held that the Delaunay triangulation was also the minimum weight triangulation [4.1], however Lloyd showed that this was incorrect by counter-example [4.24].

Since the Voronoi diagram in the plane is a planar graph it has a maximum of 3N-6 Voronoi edges (Corollary 4.2A), and since each of these edges corresponds to an edge of the DT, the planar DT can be computed from the Voronoi diagram in O(N) time. Hence the DT can be computed via the Voronoi diagram in O(NlogN) time [4.1]. More recently Lee and Schachter [4.25] describe two algorithms for computing the planar DT directly running in O(NlogN) and $O(N^2)$ time.

The ideas of the circle criterion are easily extended to higher dimensional spaces. Instead of using triangles we have n-simplices whose points lie on the surfaces of n-dimensional hyperspheres. Watson [4.26] describes an algorithm for finding Delaunay simplices by updating the structure one point at a time.

4.2.4 The Gabriel Graph (GG)

Given the set of N points in P we may define the disk of **influence** (Fig 4.6(a)) of points p_i,p_j by

$$DISK(p_i,p_j) = \{x:[d^2(x,p_i)+d^2(x,p_j)] \leq d^2(p_i,p_j)\ i \neq j\}$$

Clearly p_i and p_j lie on the edge of the disk of influence which has diameter $d(p_i,p_j)$. We may use this to define the Gabriel graph (Fig 4.4(d)).

$$(p_i,p_j) \in GG \text{ iff } p_k \notin DISK(p_i,p_j) \text{ for all } p_k \in P\ i\neq j\neq k$$

The terms **least squares adjacency criterion** and **least squares adjacency graph** [4.27] are alternative names for the above definition and the Gabriel graph respectively. Gabriel and Sokal [4.28] originally defined the graph for use in geographical variation analysis. Since then Matula and Sokal have derived a number of properties of the Gabriel graph in the plane.

Howe [4.29] showed that the GG was a subgraph of the DT, and Matula and Sokal [4.27] showed that the minimal spanning tree was a subgraph of the GG. Howe gave a lemma defining the relationship between the GG and DT.

Lemma 4.4 [4.29] The Gabriel graph of P is a subgraph of the Delaunay triangulation of P. Also any edge of the DT, (p_i,p_j) is also an edge of the GG iff the line segment joining p_i to p_j intersects the Voronoi edge common to tiles T_i and T_j at a point other than the endpoints of the Voronoi edge.

This result is used in Matula and Sokal's O(NlogN) algorithm for the GG of a planar set [4.27]. This algorithm finds both the Voronoi diagram and the DT, then obtains the GG from the DT using Lemma 4.4. From Lemma 4.4 it is clear that Miles's definition of **full neighbours**

[4.21] corresponds to the definition of Gabriel neighbours. A number of related results are available for the expected number of Gabriel neighbours of a point $E(N_G)$ as $N \to \infty$. In two dimensions Matula and Sokal show that for points uniformly distributed in the unit square, $E(N_G)$ is 4. In the corresponding result for three dimensions Miles [4.21] gives $E(N_G)$ as 8, and finally Devroye [4.30] generalizes these results to give $E(N_G)$ as 2^n for any underlying density in n-dimensions.

4.2.5 The Relative Neighbourhood Graph (RNG)

Another way of defining neighbourhood has been suggested by Lankford [4.31]. He defined points p_i,p_j to be **relatively close** iff

$$d(p_i,p_j) < \max[d(p_i,p_k),d(p_j,p_k)] \text{ for all } k=1,\ldots,N \; i\neq j\neq k$$

Toussaint [4.32] gave a more convenient definition using $\leq$ instead of $<$ in the above expression. Only Toussaint's definition will be considered below.

An equivalent definition is to define the RNG of P (Fig 4.4(e)) using a **lune of influence** (Fig 4.6(b)) for each p_i,p_j where

$$\text{LUNE}(p_i,p_j) = \{x: d(p_i,p_j) < \max[d(x,p_i),d(x,p_j)] \; i\neq j\}$$

As in the GG definition, the RNG may be defined by linking points p_i,p_j iff $\text{LUNE}(p_i,p_j)$ is empty.

Toussaint [4.32] proved that the RNG is a subgraph of the DT and that the MST was a subgraph of the RNG. He stressed that the RNG does not impose a particular structure such as a tree or triangulation, and also extracts a perceptually meaningful structure from a set of points.

O'Rourke [4.41] describes properties of and gives algorithms for

the RNG in L_1 and L_∞ metrics.

4.2.6 The Minimal Spanning Tree (MST)

The minimal spanning tree of a set P (Fig 4.4(f)) is formed by connecting points in P so that the sum of edge lengths is the minimum over all spanning trees of P. Shamos and Hoey [4.1] show that the MST can be computed by finding the DT first then obtaining the MST from the DT. In higher dimensions Bentley and Friedman [4.33] give some algorithms with fast expected time.

Zahn [4.34] describes some of the properties of the MST and considers its perceptual relevance. Because of this he suggests that it is a good structure for cluster analysis. Gower and Ross [4.35] show the link between the MST and nearest neighbour (single linkage) cluster analysis.

4.3 Graphs defined by a Region of Influence

4.3.1 Definitions

In the above section it was noted that the Gabriel graph and relative neighbourhood graph have similar definitions - in each case a pair of points are connected iff a specified region is empty - the disk of influence and lune of influence respectively. We may generalize these ideas and consider the set S of graphs defined by a **region of influence.**

Let $S = \{S_1, S_2, \ldots\}$ denote the set of graphs which have vertex set P and whose edge sets are defined with respect to a region of influence; let $R = \{R_1, R_2, \ldots\}$ denote the corresponding set of

regions of influence; and let $R_1(p_i,p_j)$ denote the region of influence formed by applying the definition R_1 to the pair of points p_i,p_j. Any graph S_1 of P is defined by

$$(p_i,p_j) \in S_1 \text{ iff } p_k \notin R_1(p_i,p_j) \text{ for all } k=1,..,N \; i\neq j\neq k$$

$$d(p_i,p_k)>0,\; d(p_j,p_k)>0$$

where the region R_1 defining S_1 is given by the set

$$R_1(p_i,p_j) = \{x:\; f[d(x,p_i),d(x,p_j)] < d(p_i,p_j)\; i\neq j\}$$

where

$$R_1(p_i,p_j) = \emptyset \text{ when } d(p_i,p_j) = 0$$

and

$$R_1(p_i,p_j) = R_1(p_j,p_i)$$

i.e. f is well behaved in the sense of yielding a finite non-empty region for $d(p_i,p_j) > 0$.

Clearly by using this generalization $R_{GG}(p_i,p_j)$=DISK(p_i,p_j) and $R_{RNG}(p_i,p_j)$=LUNE(p_i,p_j) we obtain the GG and RNG. It might be expected that the region of influence will determine some of the basic properties of these graphs - this is considered in a later subsection.

4.3.2 Algorithms for finding graphs $S_1 \in S$

It is possible to consider general algorithms for any graph $S_1 \in S$. Two general algorithms are given, both being based on applying 'region tests' to candidate graph edges. The first, GEN-1 is simply a generalization of Toussaint's RNG-1 algorithm [4.32], and the second, GEN-2 is a generalization of Urquhart's RNG algorithm [4.36].

Consider the definition of a graph defined by a region of influence S_1 given above. A logically equivalent expression is

$$(p_i,p_j) \notin S_1 \text{ iff } p_k \in R_1(p_i,p_j) \text{ for any } k=1,..,N \; i\neq j\neq k$$

Toussaint's RNG-1 algorithm considers every one of the N(N-1)/2 possible edges and tests each possible edge using all the other points in P. This algorithm easily generalizes for any graph $S_1 \in S$ giving GEN-1. In Urquhart's RNG algorithm [4.36] it is noted that the basic operation of testing a candidate edge (p_i,p_j) with another point p_k in Toussaint's algorithm involves three inter-point distances. If we replace this basic operation by one that tests each of the possible edges in the triple of points we will be able to reject some edges and throw them away. The elimination of possible edges substantially reduces the number of edges that have to be tested. This throw away scheme is easily generalized to any graph $S_1 \in S$.

Algorithm GEN-1

(1) Find the interpoint distances $d(p_i,p_j)$ for all i,j=1,...,n, i>j

(2) For each pair of points (p_i,p_j) apply the appropriate region test using each p_k for all k=1,...,n $i \neq j \neq k$

(3) Iff the given pair (p_i,p_j) satisfies the region test for all p_k, then $(p_i,p_j) \in S_1$

Algorithm GEN-2

(1) Find the interpoint distances $d(p_i,p_j)$ for all i,j=1,...,n, i>j

(2) For each pair of points (p_i,p_j) that has not previously been eliminated, do step 3 using p_k for all k=1,...,n $i \neq j \neq k$ (unless (p_i,p_j) becomes eliminated by step 3 for some k). Iff (p_i,p_j) satisfies the region test for all p_k then $(p_i,p_j) \in S_1$.

(3)(i) Apply region test to (p_i,p_j) using p_k

(ii) Apply region test to (p_i,p_k) using p_j

(iii) Apply region test to (p_j,p_k) using p_i

(iv) Eliminate any of these pairs that fails the region test

Algorithm GEN-1 runs in $O(n^3)$ time and requires the storage of $n(n-1)/2$ distances. Algorithm GEN-2 runs in less than $O(n^3)$, but more than $O(n^2)$ time, and requires the storage of $n(n-1)/2$ distances and $n(n-1)/2$ binary indicators to record eliminated edges. Although there is no formal worst case complexity analysis for GEN-2 it is clear in practice that there is a substantial saving in computation when compared with GEN-1.

4.3.3 Obtaining Subgraphs of the Delaunay Triangulation

The Delaunay triangulation of a set $P = \{p_1, p_2, \ldots, p_n\}$ of n points in the plane can be found in $O(n \log n)$ worst case running time [4.25]. Clearly it is important to know whether subgraphs of the DT can be computed with similar efficiency. Suppose we have a subgraph $S_m \in S$ of the DT defined by a region of influence $R_m \in S$. If R_m is defined such that if any points lie within $R_m(p_i,p_j)$, at least one will be a Delaunay neighbour of both p_i and p_j. An $O(n \log n)$ algorithm may be constructed for the Haltbury graph S_m, since algorithms in this class will be able to test each edge of the Delaunay triangulation for inclusion in S_m by simply testing the set of common Delaunay neighbours. If one of the Delaunay neighbours lies within $R_m(p_i,p_j)$ then $(p_i,p_j) \notin S_m$, otherwise $(p_i,p_j) \in S_m$. Such a region of influence is said to have the **Delaunay neighbour property.** Urquhart proposed computing the planar RNG using this type of algorithm [4.36], but Toussaint [4.37] showed by counter-example that the algorithm would not always yield the RNG; hence the question of the Delaunay neighbour property of a region is important.

Consider a set of four points a,b,c,d forming a convex quadrilateral adbc in the plane (Fig 4.7). This quadrilateral may be triangulated by using either ab or cd as diagonals. The Delaunay

triangulation, by the circle criterion, will have ab as a diameter iff d lies strictly outside the circumcircle of abc; cd as a diameter iff d lies strictly inside the circumcircle of abc; and either ab or cd as a diameter in the degenerate case of adbc being concyclic. In the following discussion, degenerate cases are considered in the context of the completion of the Delaunay pretriangulation [4.23].

Lemma 4.5: Suppose two points a and b are connected by an edge of the Delaunay triangulation. If one or more points lie within a circle having ab as a chord, exactly one will be a Delaunay neighbour of both a and b (degenerate cases are considered below).

Proof: A third point c will not be a Delaunay neighbour of ab, if any point lies within the circumcircle of abc [4.1]. If we consider the circumcircle of abc (Fig 4.7), we can denote the region within the circumcircle on the opposite side of ab from c by D_1 and the region within the circumcircle on the same side of ab as c by D_2.

Case (1) If a point d lies in D_1, from the locally equiangular property ab cannot be chosen as a diagonal of adbc and hence cannot be an edge of the Delaunay triangulation. However, since we know a priori that ab ϵ DT, there is a contradiction and so d may not exist within D_1.

Case (2) If a point e lies within D_2, then we can replace the circumcircle abc by the circumcircle abe and repeat the argument of case(1). If a point still lies within the current circumcircle, a new one is found until no points lie in the current D_2. At this point the circumcircle abz, where z denotes the last point to be found within D_2, is empty and so z is a Delaunay neighbour of both a and b.

Also if no points lie within a circle having ab as a chord, but one or more points lie on the boundary of the circle, then one (if only

one point lies on the boundary) or two of these will be Delaunay neighbours of both a and b.

Corollary 4.5A: Suppose two points a and b are connected by an edge of the Delaunay triangulation. If one or more points lie within the circular region R_c having ab as a diameter, then one point will be a Delaunay neighbour of a and b, and so the region of influence R_c possesses the Delaunay neighbour property.

Proof: This is obvious, and the same degenerate case considerations apply.

Corollary 4.5B: The only regions of influence having the Delaunay neighbourhood property are those whose boundaries consist of two major arcs of circles of equal radius, having ab as a chord, the arcs lying on opposite sides of ab; such regions include the circular region having ab as a diameter (R_c).

Proof: We first show that a necessary condition for a region to have the Delaunay neighbour property, is that it is bounded by an arc of a circle passing through ab. From Lemma 4.5, a circular region having ab as a chord, and containing at least one point, is guaranteed to contain a Delaunay neighbour of ab. Consider a region D, lying on one side of ab and containing a single point c. The point c will be a Delaunay neighbour of ab iff no points lie within the circle abc [4.1]. Suppose that D has a boundary that does not include an arc of a circle having ab as a chord, then if c is just inside the boundary of D, we may have c such that $\vec{D} \cap C \neq \emptyset$, where C denotes the region bounded by the circle abc. Thus it is possible for a point lying outside D to lie within the circle abc, and c is not guaranteed to be a Delaunay neighbour of ab. Hence the region D does not possess the Delaunay neighbour property. If however, D was bounded by an arc of a circle

passing through ab, from Lemma 4.5 c would be a Delaunay neighbour of ab. Thus any region $R_m \in R$ having the Delaunay neighbour property, must be bounded by arcs of circles having ab as a chord. By definition, R_m (if drawn in Euclidean space) possesses bilateral symmetry about ab; hence R_m must be defined by arcs of equal radius, each having ab as a chord. Since a region $R_1 \in R$ defined by minor arcs of circles passing through ab defines a graph $S_1 \notin DT$, the only regions having the Delaunay neighbour property are those bounded by major arcs or the circle having ab as a diameter.

Since it is preferable that regions of influence be defined by simple functions f the practical range of those having the Delaunay neighbour property is restricted to the circular region with ab as a diameter, i.e. $R_c = R_{GG}$. Thus, in practice, the only graph that is conveniently computed from the DT by testing the set of Delaunay neighbours is the GG.

From the above discussion, it is clear that the planar RNG algorithm of Urquhart [4.36] is only approximate ($R\hat{N}G_U$) since the lune does not possess the Delaunay neighbour property. Toussaint and Menard [4.38] describe a better approximation ($R\hat{N}G_T$) where the approximations are related by

$$RNG \subseteq R\hat{N}G_T \subseteq R\hat{N}G_U \subseteq GG$$

Matula and Sokal [4.27] describe an optimal algorithm for computing the Gabriel graph using both the Voronoi diagram and the DT. Clearly since R_{GG} possesses the Delaunay neighbour property, we may compute just the DT using the algorithm of Lee and Schachter [4.23] then apply region tests to each DT edge using the neighbours of that edge - this algorithm being optimal also.

Recently Bowyer [4.18] and Watson [4.26] have produced efficient

algorithms for computing the Delaunay triangulation in more than two dimensions. If P is defined for a k-dimensional Euclidean space, Bowyer's algorithm runs in $O(a_k n^{(1+1/k)} + b_k n)$ time, and Watson's in $O(n^{(2k-1)/k})$ time. However the factor b_k in Bowyer's algorithm increases significantly with increasing dimensionality. With large dimensionalities (e.g. k>15 for a mainframe computer) the algorithm is therefore liable to become inoperable [4.39]. The definition of the Delaunay triangulation, Lemma 4.5 and its Corollaries may be generalized to k-dimensional spaces, and so the Gabriel graph may be computed from the Delaunay simplex by merely applying the region test to each edge using its set of Delaunay neighbours.

4.3.4 Connectivity and Planarity

It is to be expected that the definition of R_1 will determine whether or not S_1 is connected. Consider Lemma 4.6.

Lemma 4.6: R_{RNG} is the maximal region of influence $R_1 \in R$ that is guaranteed to give a connected graph S_1.

Proof: From the definition of R_1, a necessary condition for $S_1 \in S$ to be connected is that it will always join a point to its nearest neighbour. Any region $R_1 \not\in R_{RNG}$ (Fig 4.8) is not guaranteed to connect a point to its nearest neighbour, and so S_1 will not necessarily be connected. This Lemma may easily be illustrated for a region $R_1 \not\in R_{RNG}$ by constructing a three point example. (See also Toussaint [4.32])

Lemma 4.7 R_{GG} is the minimal region of influence that is guaranteed to define a planar graph of a planar set P.

Proof Consider a region of influence that is just smaller than R_{GG} i.e.

$$R_{*} = \{x: d^2(x,p_i)+d^2(x,p_j) < d(p_i,p_j)\ i \neq j\}$$

note that the R_{GG} is defined using $\leq$ rather than $<$. Consider the set P of six cocircular points, then the graph S_{*} of P is given in Fig 4.9(a). Clearly this graph is isomorphic with $K_{3,3}$ (Fig 4.9(b)) and hence S_{*} is not planar by Kuratowski's theorem (Theorem 4.1). Thus any graph defined by a region of influence $R_1 \not\equiv R_{GG}$ is not guaranteed to be planar.

4.4 Discussion

This chapter examines a number of structures that may be used in statistical pattern recognition. Such structures offer computationally attractive solutions to problems in pattern recognition. It is now widely known that the DT, GG, RNG, and MST are closely related graphs. However the differences between the GG and RNG are not clear.

The difference between these structures becomes significant when constructing algorithms. The planar GG may be computed from the planar DT in O(N) time, however the planar RNG may not. An investigation of the differences between these structures was based on generalizing both structures as graphs defined by a region of influence. This showed that efficient computation of such graphs depended heavily on the shape of the region of influence.

Two further important properties arise from allowing the region of influence to vary. The RNG may be considered to be 'just connected' and the GG to be 'just planar'. The use of a region just larger than the lune used in defining the RNG may result in a disconnected graph, and a region just smaller than the disk used for the GG may yield a non-planar graph. The connectivity result is not limited to 2-

dimensions and is vital to the work of Chapter 5

References

4.1 M.I.Shamos & D.Hoey, Closest point problems, Sixteenth Annual IEEE Symposium on Foundations of Computer Science, 151-162 (1975)

4.2 G.T.Toussaint & R.S.Poulsen, Some new algorithms and software implementation methods for pattern recognition research, IEEE Computer Society's Third International Computer Software and Applications Conference, 55-63 (1979)

4.3 H.Edelsbrunner, D.G.Kirkpatrick & R.Seidel, On the shape of a set of points in the plane, Tech Rept F71, Technishe Universitat Graz (1981)

4.4 G.T.Toussaint, Decomposing a simple polygon using the relative neighbourhood graph, Proceedings of the Allerton Conference, Urbana, U.S.A (1980)

4.5 R.B.Urquhart, Graph theoretical clustering based on limited neighbourhood sets, Pattern Recognition, 15, 173-188 (1982)

4.6 G.T.Toussaint, Pattern recognition and geometrical complexity, Proc. 5th International Conference on Pattern Recogniton, 1324-1347, Miami, U.S.A. (1980)

4.7 G.T.Toussaint, Computational geomtric problems in pattern recognition, in Pattern Recognition Theory and Applications, J.Kittler (Ed.), NATO Advanced Study Institute, Oxford, England (1981)

4.8 M.G.Kendall, A course on the geometry of n dimensions, Griffin (1961)

4.9 J.W.Cooley & J.W.Tukey, An algorithm for the machine calculation of complex Fourier series, Math. Comput., 19, 297-301 (1965)

4.10 J.L.Bentley & M.I.Shamos, Divide and conquer for linear expected time, Inf. Proc. Lett., 7, 87-91 (1978)

4.11 J.L.Bentley & M.I.Shamos, Divide and conquer in multidimensional space, Proc. 8th Annual ACM Symposium on Theory of Computing, 220-230, Hershey (1976)

4.12 J.L.Bentley, Multidimensional divide-and-conquer, CACM, 23, 214-229 (1980)

4.13 G.Voronoi, Nouvelles applications des parametres continus a la theorie des formes quadratiques. Deuxieme memoire: recherches sur les parallelloedres primitifs, J. Reine Agnew Math., 134, 198-287 (1908)

4.14 P.J.Green & R.Sibson, Computing Dirichlet tesselations in the plane, Computer J., 21, 168-178 (1978)

4.15 A.H.Thiessen & J.C.Alter, Precipitation averages for large areas, Monthly Weather Review, 39, 1082-1084 (1911)

4.16 R.Sibson, The Dirichlet tesselation as an aid in data analysis, Scand. J. Statist., 7, 14-20 (1980)

4.17 K.Q.Brown, Voronoi diagrams from convex hulls, Inf. Proc. Lett, 9, 223-228 (1979)

4.18 A.Bowyer, Computing Dirichlet tesselations, Computer J., 24, 162-166 (1981)

4.19 R.Sibson, A vector identity for the Dirichlet tesselation, Math. Proc. Camb. Phil. Soc., 87, 151-155 (1980)

4.20 R.E.Miles, On the homogenous planar Poisson point process, Math. Biosciences, 6, 85-127 (1970)

4.21 R.E.Miles, The random division of space, Suppl. Adv. Appl. Prob., 243-266 (1972)

4.22 B.Delaunay, Sur la sphere vide, Bull. Acad. Science USSR VII Class Sci. Mat. Nat., 793-800 (1934)

4.23 R.Sibson, Locally equiangular triangulations, Computer J., 21, 243-245 (1978)

4.24 E.L.Lloyd, On the triangulation of a set of points in the plane, Tech. Rept. MIT/LCS/TM-88, MIT, Cambridge, U.S.A. (1977)

4.25 D.T.Lee & B.J.Schachter, Two algorithms for constructing a Delaunay triangulation, Int. J. Comput. Inf. Sci., 9, 219-242 (1980)

4.26 D.F.Watson, Computing the n-dimensional Delaunay tesselation with application to Voronoi polytopes, Computer J., 24, 167-172 (1981)

4.27 D.W.Matula & R.R.Sokal, Properties of Gabriel graphs relavent to geographical variation analysis and the clustering of points in the plane, Geogr. Anal., 12, 205-222 (1980)

4.28 K.R.Gabriel & R.R.Sokal, A new statistical approach to geographic variation analysis, System.Zool., 18, 259-278 (1968)

4.29 S.E.Howe, Estimating regions and clustering spatial data: analysis and implementation of methods using the Voronoi diagram, Ph.D. thesis, Brown University, Providence, U.S.A. (1978)

4.30 L.Devroye, Personal communication (1981)

4.31 P.M.Lankford, Regionalization: theory and alternative algorithms, Geogr. Anal., 1, 196-212 (1969)

4.32 G.T.Toussaint, The relative neighbourhood graph of a finite planar set, Pattern Recognition, 12, 261-268 (1980)

4.33 J.L.Bentley & J.P.Friedman, Fast algorithms for constructing minimal spanning trees in coordinate spaces, IEEE Trans. Computers, C-27, 97-105 (1978)

4.34 C.T.Zahn, Graph-theoretical methods for detecting and describing Gestalt clusters, IEEE Trans. Computers, C-20, 68-86 (1971)

4.35 J.C.Gower & G.J.S.Ross, Minimum spanning trees and single-linkage cluster analysis, Applied Statistics, 18, 54-64 (1969)

4.36 R.B.Urquhart, Algorithms for computation of relative neighbourhood graph, Electron.Lett., 14, 556-557 (1980)

4.37 G.T.Toussaint, Comment on 'Algorithms for computing relative neighbourhood graph', Electron.Lett., 14, 860-861 (1980)

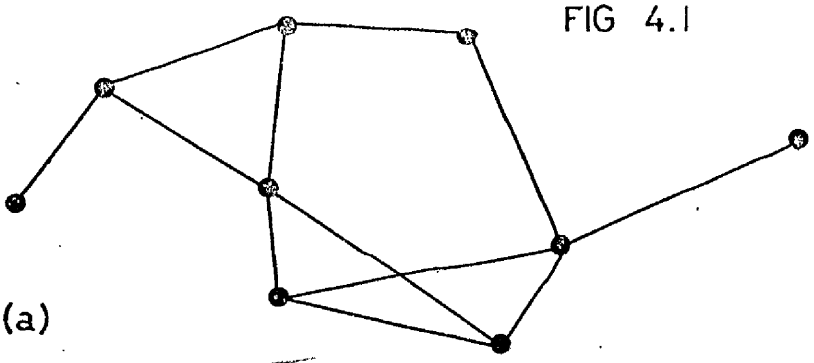
4.38 G.T.Toussaint & R.Menard, Fast algorithms for computing the planar relative neighbourhood graph, in Methods of Operations Research, Koln, W.Germany (1980)

4.39 A.Bowyer, Personal Communication

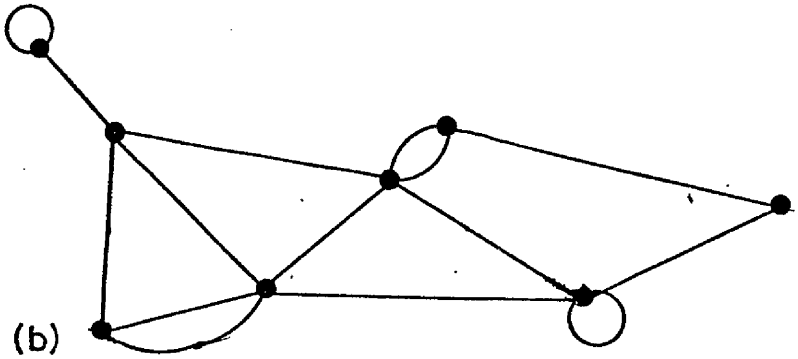
4.40 D.T.Lee & C.K.Wong, Voronoi diagrams in L_1 (L_{∞}) metrics with 2-dimensional storage applications, SIAM J. Comput., 9, 200-211 (1980)

4.41 J. O'Rourke, Computing the relative neighbourhood graph in the L_1 and L_{∞} metrics, Pattern Recognition, 15, 189-192 (1980)

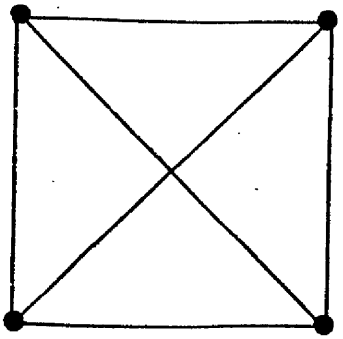
FIG 4.1



(a)

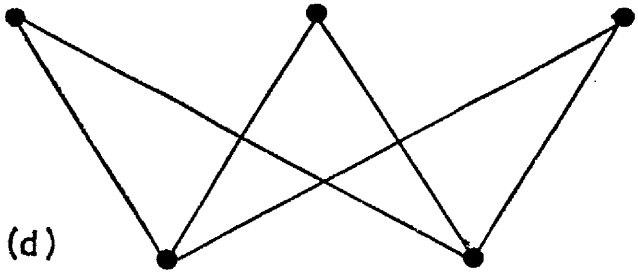


(b)



(c)

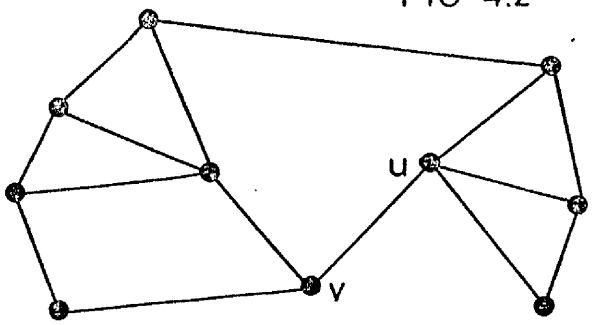
K_4



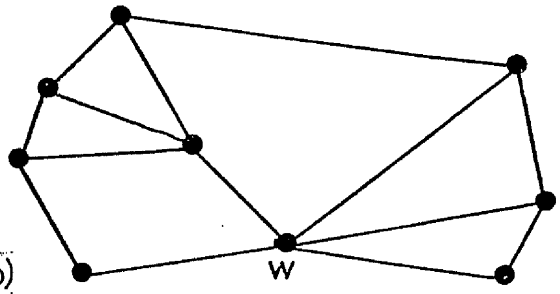
(d)

$K_{2,3}$

FIG 4.2

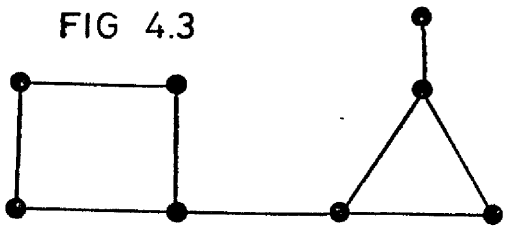


(a) u&v are contracted to w

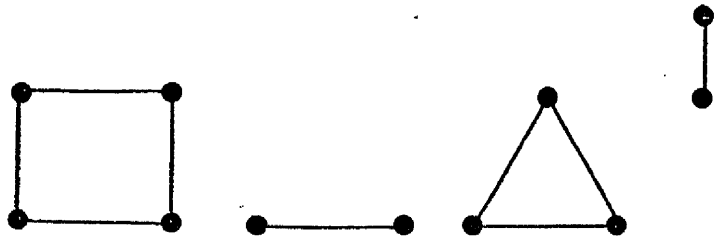


(b)

FIG 4.3

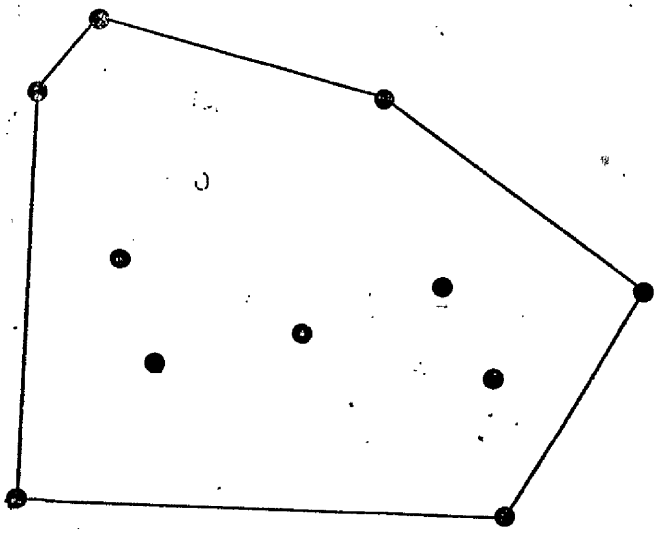


(a) a graph G

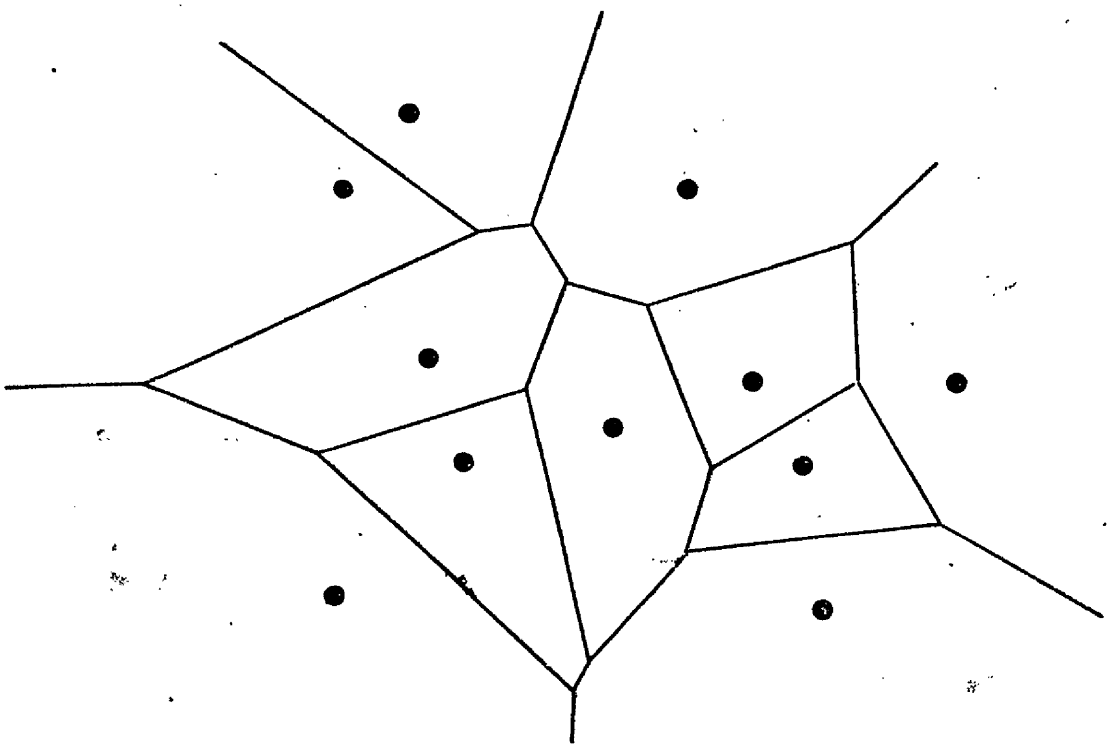


(b) blocks of G

FIG 4.4

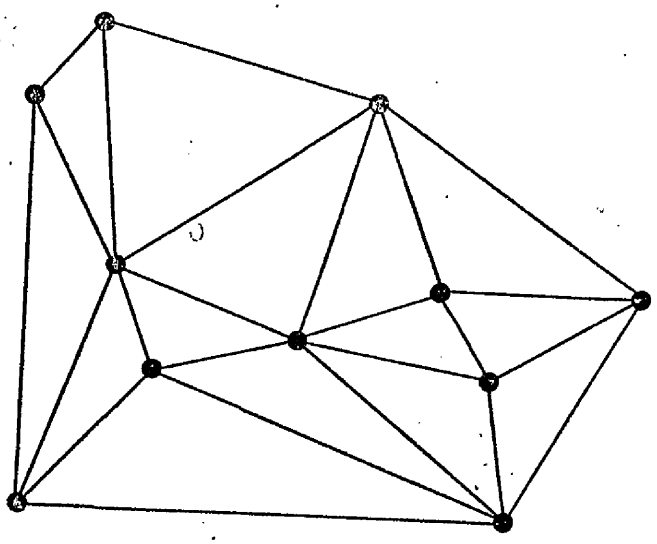


(a)

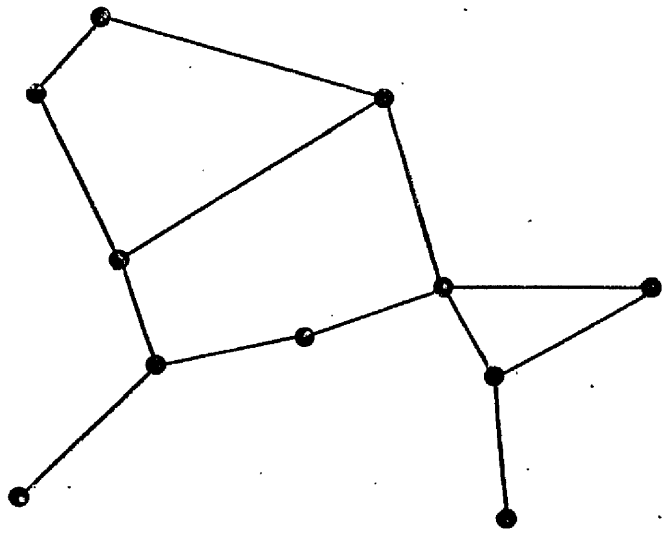


(b)

FIG 44

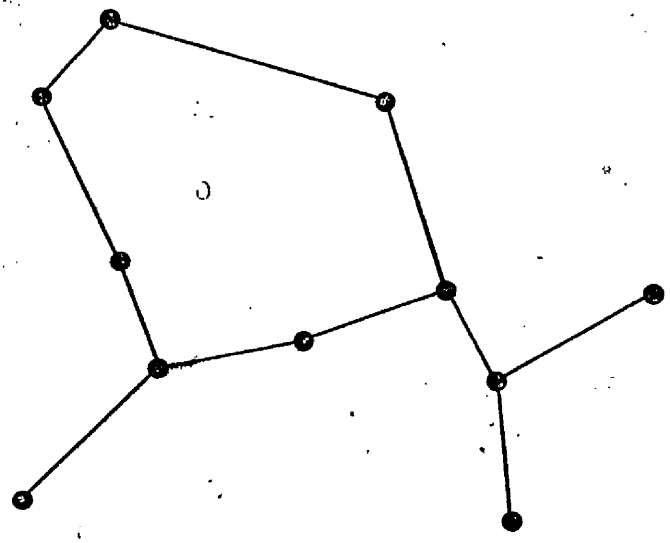


(c)

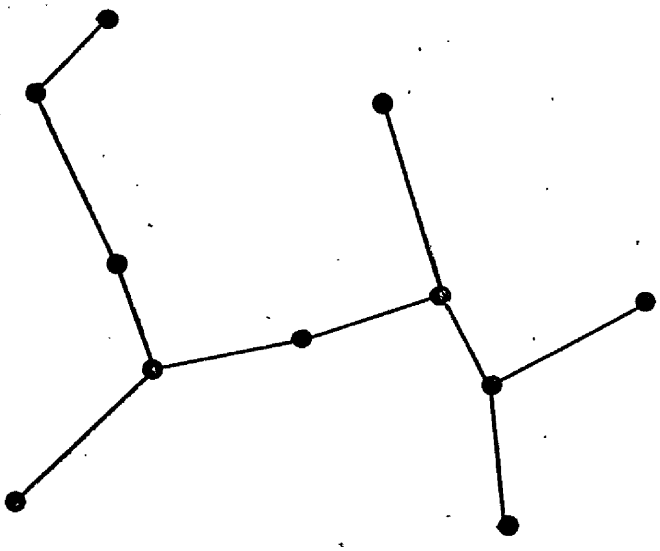


(d)

FIG 4.4



(e)



(f)

FIG 4.5

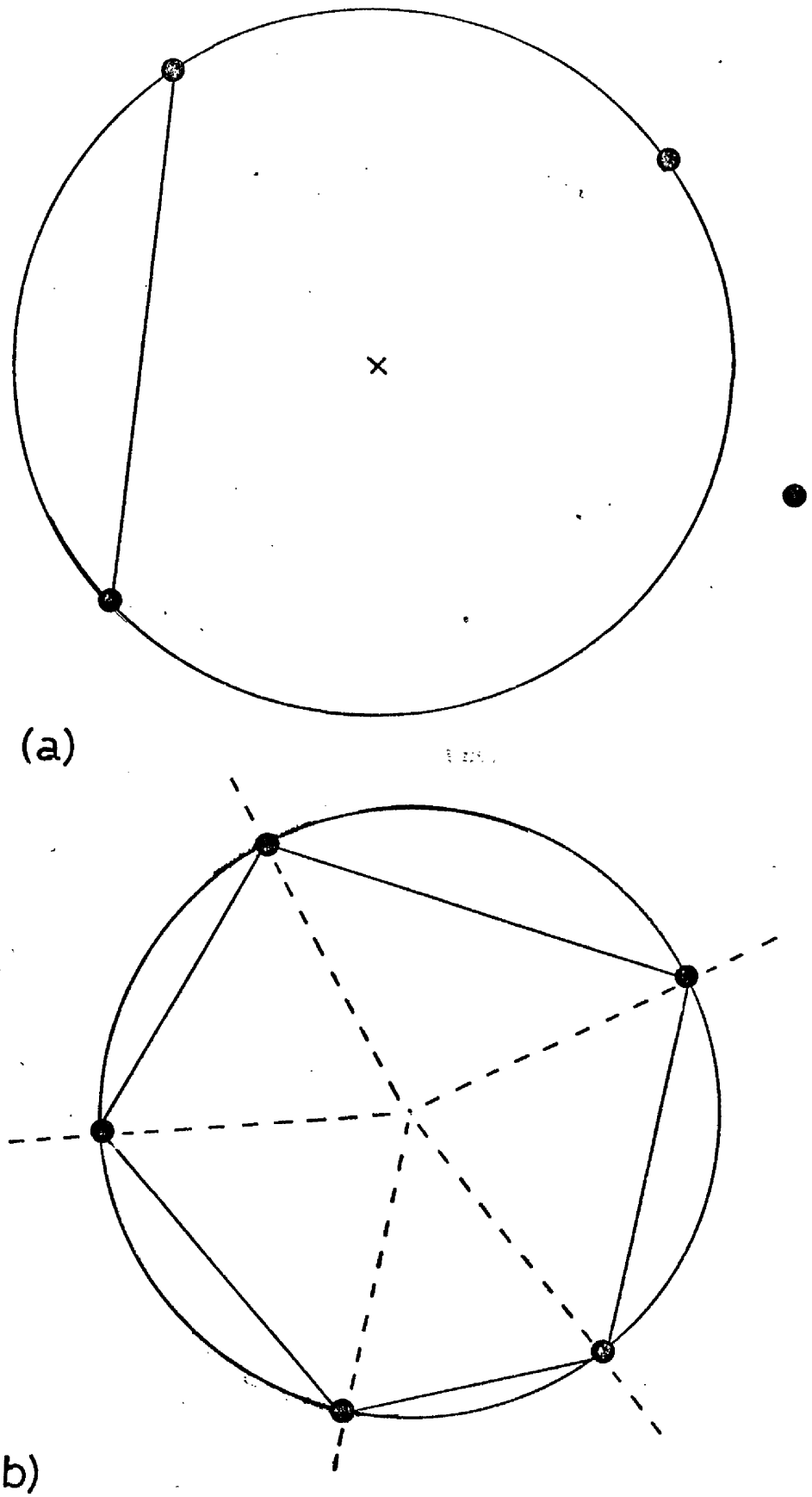
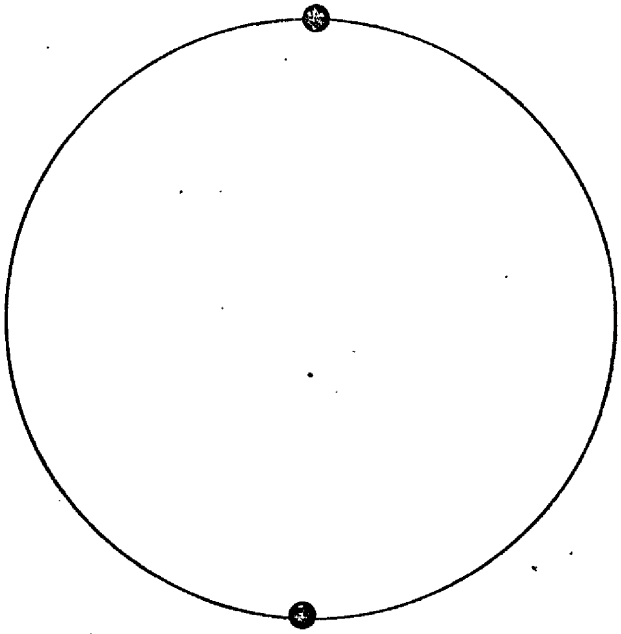
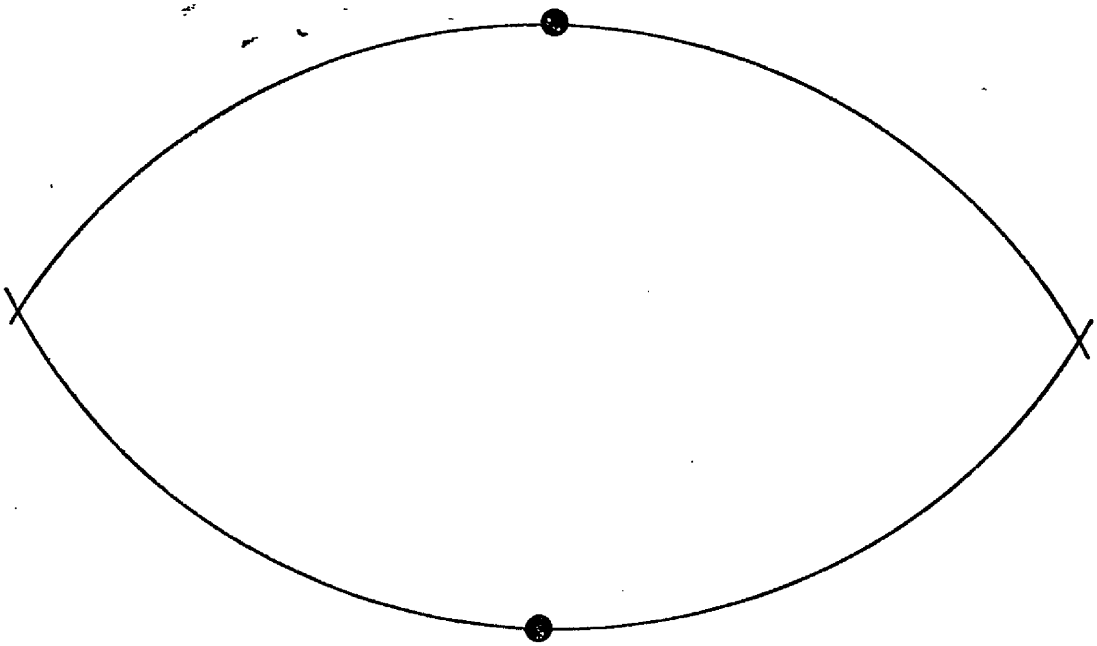


FIG 4.6



(a)



(b)

FIG 4.7

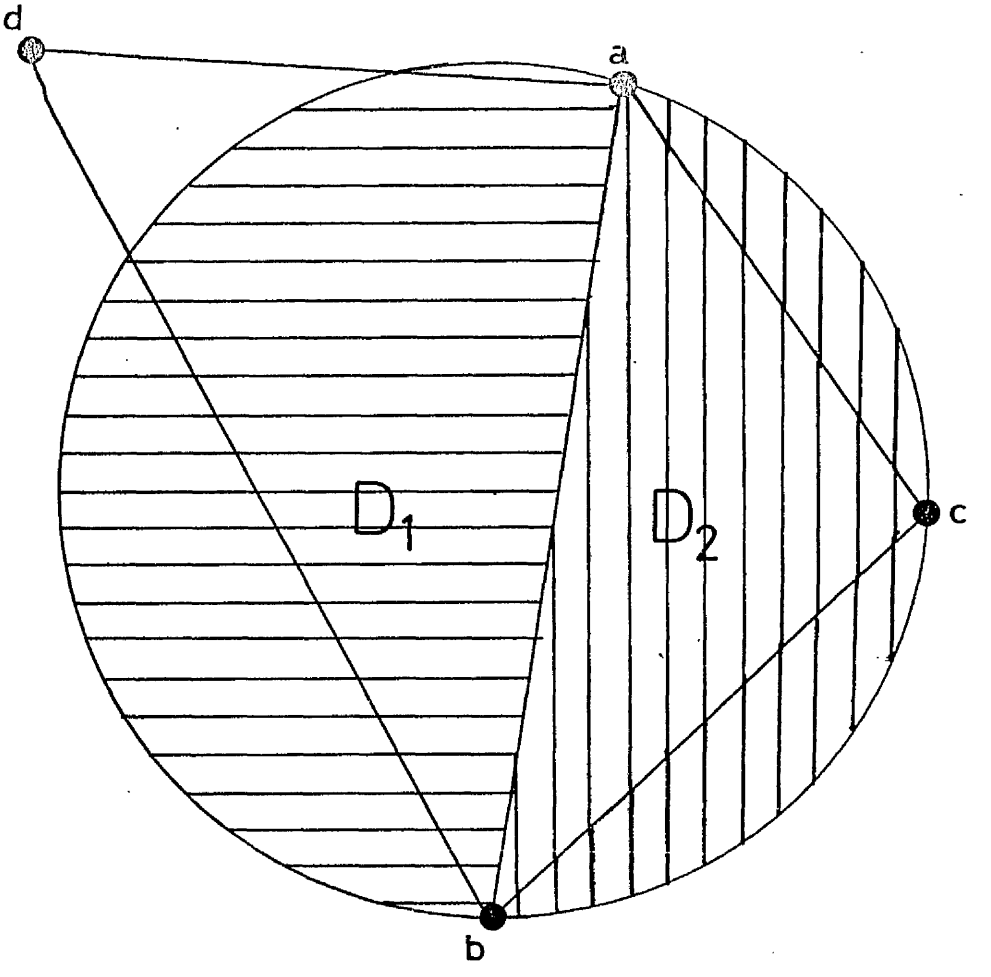


FIG 4.8

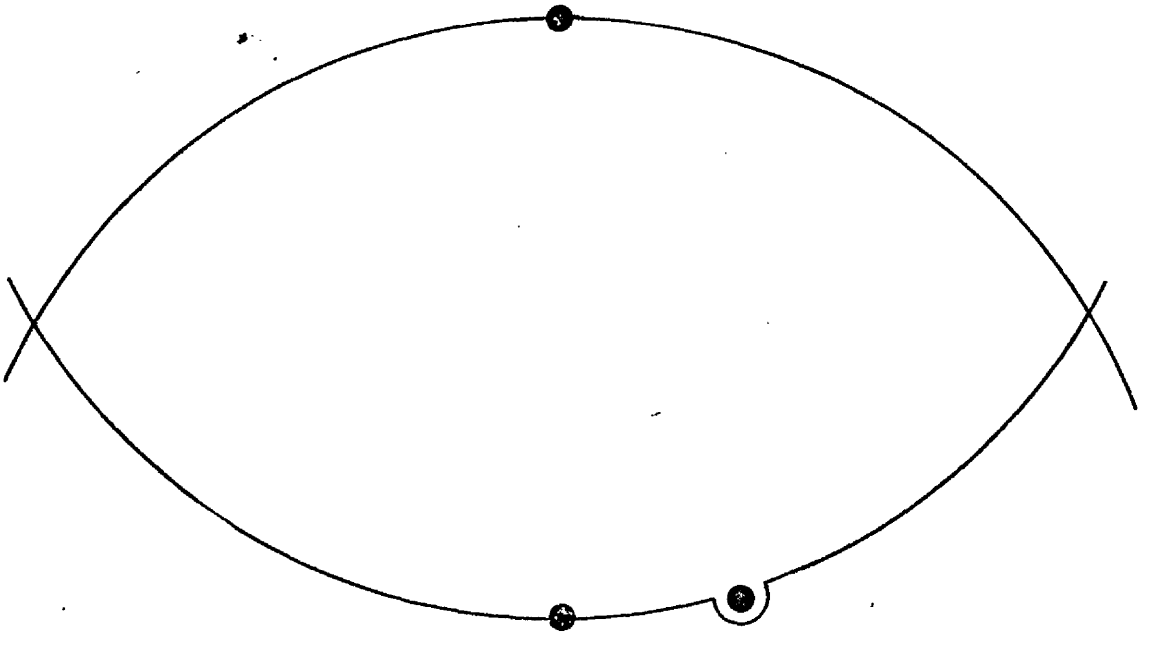
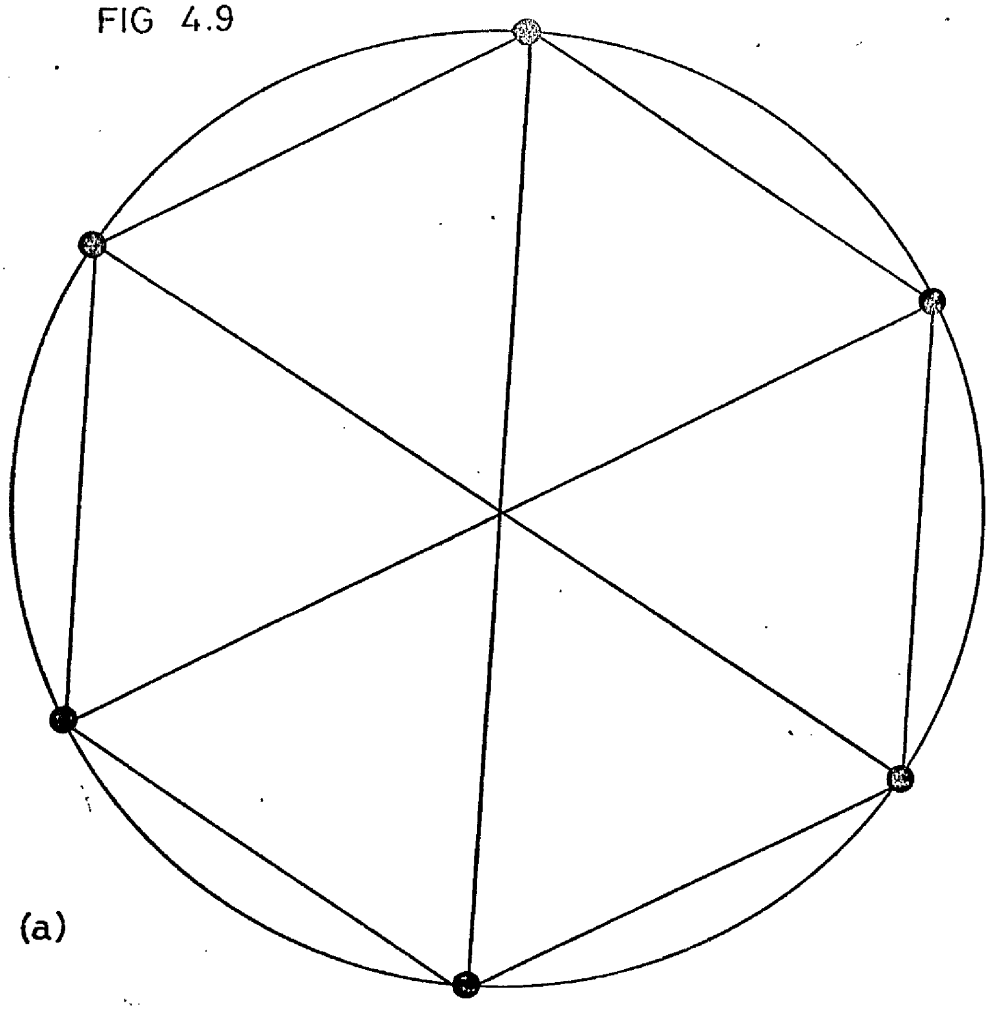
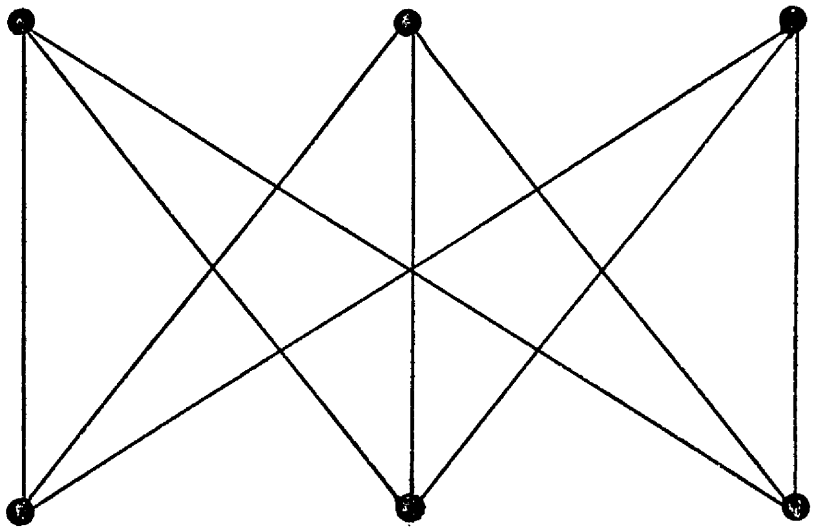


FIG 4.9



(a)



(b)

A NEW GRAPH-THEORETICAL CLUSTERING METHOD

5.1 Introduction

Clustering is an approach to the analysis of data common to pattern recognition, and to the biological and social sciences. Clustering methods partition the data into subsets so that two data points within a cluster are considered similar in some sense whereas two data points from different clusters are considered dissimilar. Over a number of years, a variety of approaches have been devised; the idea of a cluster does not have a precise universal definition, and so a given clustering method will reflect a particular interpretation of the clustering problem.

The clustering methods may be grouped into a number of classes e.g. (a) parametric methods, which assume underlying Gaussian distributions in the data; (b) methods which partition data according to a dissimilarity matrix e.g. single link and complete link; (c) methods which are based on a visual perceptual model of clusters [5.1-5.3]. Graph theoretical techniques have been used extensively in clustering methods. Such techniques include (i) breaking of the minimal spanning tree (MST) of the data set e.g. single link clustering and Zahn's edge inconsistency methods [5.3]; (ii) the use of directed trees [5.2,5.4]; (iii) nearest neighbour techniques [5.3,5.5-5.7] (although these are not necessarily graph theoretical); and (iv) graph colouration [5.8]. Some of these techniques have proved useful in analysis of data containing non-globular clusters.

In the following study, an approach is considered based on a visual model of clustering. This work was motivated by a consideration

of Zahn's classic paper on MST methods [5.3]. A hierarchic clustering method is described which is based on limited neighbourhood sets [5.1]. The performance of this technique is considered using a number of two-dimensional problems, and is shown to avoid some of the problems associated with Zahn's original methods. Two of the geometrical structures (the relative neighbourhood graph (RNG) and Gabriel graph (GG)) recently suggested for use in pattern recognition by Toussaint [5.9] are exploited. These graphs may be generalized giving a family of graphs defined with repect to a **region of influence**. Such graphs connect two data points whenever all the remaining points in the data set lie outside the region of influence of the two data points (Chapter 4). In this way we may choose a definition that may be used to partition data directly.

Some methodological problems are discussed with particular emphasis on the problem of **cluster identification**. An extension of the graph theoretical clustering method which gives an interactive approach to this problem is described. The operation of a display for interactive clustering is illustrated by taking the lung sound data set of Chapter 7 as a simple case study.

5.2 Clustering Methodology

5.2.1 Problems in Clustering Methodology

There are a number of problems associated with using clustering methods. A clustering algorithm will be guaranteed to partition a data set regardless of whether there is any significant cluster structure. Thus it is important to ask a number of questions about the data and

to establish whether the clustering results are due to genuine structure in the data or are merely artifacts of the clustering algorithm. Some important questions are listed below (those referring exclusively to hierarchic methods are marked* and those referring to a visually based method are marked+):

1. Does the input data show a clustering tendency or is it random?

2. Is the method appropriate to the data?

3* Does the hierarchy reflect the data structure? (Cluster stability)

4* Is any partition a good summary of the data?

5. Are any of the clusters real? (Cluster validity)

6+ What type of clusters have been produced? (Cluster identification).

Clustering algorithms are data dependent and so results will only be meaningful if an appropriate clustering method is chosen [5.11]. Having obtained a partition or a hierarchic clustering, it is important to establish which clusters if any are valid. This problem of **cluster validity** involves qualifying the results by means of suitable statistical tests increasing the value of the clustering to the user [5.12]. Unfortunately there are few straightforward validity methods short of computationally expensive Monte Carlo simulations. Many of these techniques are only appropriate to absolute (as opposed to relative) distance methods as they are based on the random graph hypothesis. However recently Backer & Jain [5.13] and Bailey & Dubes [5.14] have suggested new approaches.

In hierarchic clustering, the data set is 'fitted' to a sequence of nested partitions, conventionally displayed as a dendrogram, but

the imposed structure obscures the more complex relationships between the clusters obtained, unless the data is ultrametric or near ultrametric. The problem of **cluster stability** [5.15] involves (a) whether the hierarchy is representative of the overall data structure, (b) whether any individual partition is a good summary of the data and (c) whether any individual clusters in the hierarchy are real rather than being an artifact of the clustering algorithm. Smith & Dubes [5.15] consider various statistics to test cluster stability.

Problems occur with the use of the visually oriented non-parametric methods frequently used in patttern recognition. Such methods were developed following the realization that conventional parametric methods would fail to take account of unusually shaped clusters and impose a particular structure instead. But with the resulting variety of acceptable cluster types, the user has additional problems in interpreting the partitions produced. For example the user might draw different conclusions according to whether a given cluster was globular, chained or bridged. Possibly a user would benefit more from solving this problem of **cluster identification** than from being able to detect a very wide range of cluster types. Surprisingly few aids are available for this problem.

In pattern recognition, cluster analysis is considered to be a tool for exploring data and suggesting hypotheses. Advances in interactive graphics are particularly amenable to exploring data. Usually interactive pattern recognition techniques involve linear, non-linear or functional mappings of the multivariate data set onto a 2-dimensional display (see Chien [5.16]), and some of these have been described as 'interactive clustering' methods [5.17,5.18]. However there is little scope for interaction with the results of conventional

clustering methods apart from the interactive entry of parameters before or during computation. This led to the development of a simple display to complement the above graph theoretical clustering method.

5.2.2 Using a Visual Clustering Methodology

A visual model of the clustering problem is one in which clusters are defined (in two-dimensions) in a way that relates to human visual perception. A method based on this model should produce clusterings in 2 or more dimensions similar to those perceived in two-dimensional scatter diagrams.

It is clear that local, global and contextual factors have a part to play in human visual perception of dot patterns. Consider a few dot patterns:- Fig 5.1(a) and Fig 5.1(b) would probably be considered to have a number of well separated clusters; in Fig 5.1(c) and Fig 5.1(d) the clusters may be distinguished by local changes in point density; in Fig 5.1(d) and Fig 5.1(e) there is a 'bridge' connecting the two subclusters; and in Fig 5.1(f) the data might be divided into three subsets around the local maxima in point density. Context undoubtedly plays a part in the visual identification of clusters [5.19,5.20].

It is useful to contrast the objectives of a visual clustering methodology with for example clustering in numerical taxonomy. A taxonomist is interested in classifiying populations of organisms into a hierarchy of species. A clustering technique will therefore be the means of obtaining such a hierarchy of organisms. Because of this very specific requirement there may be very strict restrictions on which algorithms are suitable [5.21]. The basic data unit is the **operational taxonomic unit** (OTU) which may be the smallest representative of a

biological population.

The visual model of clustering allows the separation of populations with unusually shaped or differently spaced distributions, and allows a chain of points to be a reasonable cluster (Fig 5.1(b)). The visual model is more suited to pattern recognition than to the construction of taxonomical hierarchies since the OTU is itself representative of some population [5.21], whereas generally in pattern recognition, all members of (undefined) populations are likely to be present. Thus in constructing taxonomical hierarchies, the clustering problem is a question of clustering representatives of distributions, whereas often in pattern recognition clustering is a technique to explore data and detect the distributions, making use of the wide range of cluster types that are acceptable in the visual model. Implicit in visually oriented clustering methods such as that of Zahn [5.3], is the notion that relative distance is important in the pattern space, in contrast to the single link method which is based on absolute distance.

Zahn's method of cluster analysis involved finding the MST of the data set, then removing the MST edges that were found to be inconsistent. In view of the fact that the MST is a tree, the removal of one link will partition the data set. This technique was generally successful in detecting disjoint clusters (e.g. Fig 5.1(a) & Fig 5.1(b)), but did not always work directly for changes in point density (e.g. Fig 5.1(c) & Fig 5.1(d)). Heuristic solutions were offered to problems of bridged and touching Gaussian clusters (e.g. Fig 5.1(e) & Fig 5.1(f)) but these were specific to each particular problem, and would require a priori knowledge of the nature of the data set to be effective.

Jarvis [5.6] has pointed out that with some arrangments of clusters, Zahn's methods may fail to break obviously inconsistent edges (Fig 5.2(a)). Also the MST does not necessarily contain every consistent edge and so the removal of a single edge may lead to an inappropriate clustering. Another problem of using the MST is that, at a low level, it is very sensitive to changes in the position of a particular point [5.19]. Jarvis [5.6] tried to overcome some of the problems of using the MST by combining MST methods with the shared nearest neighbour clustering method [5.7]. However, it is questionable whether the k-NNs give the 'best' set of neighbours for representing data structure [5.5,5.22].

Interestingly Jarvis suggested that the MST has a limitation on its ability to represent data structure in terms of relative distance. The alternative to combining the MST with shared nearest neighbours is to ask whether another graph would give a better representation of data, in the relative distance sense, than the MST. Recently, in an important study, Toussaint [5.23] has suggested that the relative neighbourhood graph (RNG) is better at extracting a perceptually meaningful structure from a data set than the MST. The RNG does not impose a particular structure (e.g. tree or triangulation) on the data, and thus is better able to represent the data; the RNG is also much less sensitive to changing the position of a point. Toussaint showed that the MST is a subgraph of the RNG, and gave a number of examples where the RNG extracted a meaningful structure from the data.

In view of these properties, it is suggested that the RNG or graphs with similarly flexible properties e.g. the Gabriel graph [5.24] might provide a better framework for relative distance clustering than the MST.

5.3 Limited Neighbourhood Sets and Clustering

Having outlined some of the problems of using a visual clustering methodology we now outline a new clustering method. It is very easy to fall into the trap of 'selling' a new clustering technique and overlook the deficiencies. We will attempt to avoid this by considering objective measures such as the admissibility criteria, but the evaluation of the new technique is far from exhaustive.

5.3.1 The Limited Neighbourhood Concept

Lankford [5.1] describes criteria for clustering algorithms that were defined by Neely. These criteria are based on a visual model of clustering, and are collectively known as the limited neighbourhood concept.

To introduce this topic, a number of clustering definitions are presented. Let $P = \{p_1, p_2, \ldots p_n\}$ denote n points in space, and let $d(p_i, p_j)$ denote the distance between points p_i and p_j according to the given metric. A clustering C(P) is a partition of P into m non-empty subsets of P denoted by $C(P) = \{c_1, c_2, \ldots, c_m\}$. The clustering should in some way reflect the structure of the data, without making any a priori assumptions.

Clustering methods may be either hierarchic or non-hierarchic. A hierarchic clustering is a sequence of nested clusterings, whereas a non-hierarchic clustering seeks a clustering that is optimal according to some criterion. A strictly hierarchic clustering C(P) is a sequence of n clusterings $C(P) = C_1, C_2, \ldots C_n$, where C_1 contains n clusters each

consisting of a single data point $C_1 = \{c_{11}, c_{12}, \ldots c_{1n}\}$ and C_n contains a single cluster with n points $C_n = \{c_{n1}\}$. Each clustering C_j is associated with a dissimilarity value d^*_j where $d^*_{j-1} < d^*_j$ for $j=2,\ldots,n$; and where C_j is formed by merging two clusters of C_{j-1}. The dissimilarity value d^* must satisfy the ultrametric inequality for all points in P [5.25,5.26] i.e.

$$d^*(x,y) \leq \max \; [\, d^*(x,z), d^*(y,z) \,] \quad x \neq y \neq z$$

in addition to the normal requirements for a metric. Consideration of a dendrogram which is a graphical representation of a hierarchic clustering, shows that the hierarchy is meaningless unless the dissimilarity values are ultrametric.

A dendrogram is very simply a representation of a hierarchy. One axis represents dissimilarity, and at each partition a horizonatal line is drawn at the dissimilarity corresponding to the partition. vertical lines are used to denote the two subclusters formed (see Fig 5.3). A number of different styles have been used to draw dendrograms but they all contain essentially the same information.

The basic approaches to hierarchic clustering algorithms (as opposed to methods) are the agglomerative and the divisive ones. The agglomerative algorithm starts with clustering C_1 and then progressively merges clusters until the single cluster C_n is reached. In contrast, the divisive algorithms initially have clustering C_n which is divided at each level until individual points at level C_1 are reached.

The limited neighbourhood set criteria [5.1] are (in modified form):

(1) Connectivity: all points within a cluster $c_i \in C(P)$ should be connected. (In the original statement of these criteria Neely used

'strongly connected' rather than 'connected'.)

(2) Consistency: for adding a new point t with a set of neighbours N(t), if $c_i \in C(P)$ and $N(t) \subset c_i$ then there exists a cluster $c_i' \in C(P \cup t)$ such that c_i' is identical to $c_i \cup t$ for some partition of P.

(3) Local stability: suppose $c_i \in C(P)$ and a point t is added such that $N(t) \cap c_i = \emptyset$ then there is a clustering $c_i' \in C(P \cup t)$ such that $c_i = c_i'$ for some partition of P.

These criteria do not assume any particular distribution and so are suitable for use with a visual model of clustering. In Lankford's paper, the concept of relatively close neighbours and hence the RNG is introduced. A general purpose neighbourhood set algorithm is used to compute 'association' between points which then clusters the data. However the use of the RNG to form neighbourhood sets was not considered to be very good, probably because of the clustering algorithm rather than because of a property of the graph itself.

Returning to the visual concept of clustering, it is interesting to note that both Zahn [5.3] and Toussaint [5.23] use connected graphs to describe the data set. While these provide good descriptions of dot patterns, it would be interesting to find a graph that is disconnected when obvious clusterings occur (each connected subgraph corresponding to a cluster). If data were partitioned directly in this way, the graph would immediately fulfil the above connectivity criterion; clearly, and if it also defined a reasonable set of neighbours it would meet the consistency and local stability requirements. Obviously such graphs would be of immediate value in the clustering problem.

5.3.2 Graphs defined by a variable Region of Influence

Regions of influence that are larger will cause fewer pairs of points to be joined, and whereas the RNG and GG are connected graphs, regions larger than the lune may lead to disconnected graphs (Lemma 4.7). In this section three variable region definitions are proposed, namely those suggested by Figs 5.4(a) and 5.4(b), as the basis of a clustering method based on the GG and RNG respectively.

These definitions are now considered formally. It should be noted that although the Delaunay triangulation (DT), like the GG and the RNG is capable of producing limited neighbourhood sets, the DT cannot be defined by a region of influence in the same way.

Clearly, any graph $S_1 \in \mathcal{S}$ will be defined according to some notion of neighbourhood, and will define a limited neighbourhood set. If f is based on a simple operator, or a simple combination of operators, the resulting graph will be reasonably easy to interpret and compute: the regions of Fig 5.4(a) and 5.4(b) were chosen because they correspond to simple combinations of functions.

From Lemma 4.7 any graph $S_1 \in \mathcal{S}$ where $R_1 \cap \bar{R}_{RNG} \neq \emptyset$ may be disconnected. Therefore such graphs may be usable in detecting clusters. Consider the following examples of such regions of influence (corresponding to the regions of Figs 5.4a and 5.4b) which incorporate explicitly an idea of relative distance and a parameter σ :

$$R_1(p_i,p_j,\sigma) = R_{GG}(p_i,p_j) \cup \{x:\min[d(x,p_i),d(x,p_j)] < \sigma.d(p_i,p_j) i\neq j\}$$

$$R_2(p_i,p_j,\sigma) = R_{RNG}(p_i,p_j) \cup \{x:\min[d(x,p_i),d(x,p_j)] < \sigma.d(p_i,p_j) i\neq j\}$$

where σ is a factor of **relative edge consistency.** Obviously, $S_1(\sigma) \subseteq$ GG and $S_2(\sigma) \subseteq$ RNG. Thus $S_1(\sigma)$ is obtained from the GG by removing edges (p_i,p_j) if the ratio of $d(p_i,p_j)$ to $\min[d(p_i,p_a),d(p_j,p_b)]$ is

greater than σ, where p_a and p_b denote the nearest Gabriel neighbours to p_i and p_j respectively ($p_a \neq p_j$, $p_i \neq p_b$).

Variation of σ controls the fragmentation of the data set and hence it might be expected that varying σ would give a sequence of nested clusterings. (Differences in clusterings produced by $S_1(\sigma)$ and $S_2(\sigma)$ may occur at a low level of dissimilarity, depending on the relative distances of nearest relative or Gabriel neighbours to the edge in question). We may therefore associate with each edge of the GG or RNG, a similarity σ' (or a dissimilarity $d'=1/\sigma'$) at which the edge is broken. As the graph edges are broken it is clear that the breaking of some edges will partition the data set. We may therefore associate the similarity (or dissimilarity) value required to break that edge with the partition formed. We denote the value of d' corresponding to the edge that partitions the data by d^*. Since σ is defined as a ratio of distances it is easily shown that the ultrametric dissimilarity coefficient d^* is continuous, and hence $S_1(\sigma)$ and $S_2(\sigma)$ may be used in hierarchic clustering.

The objective of hierarchic clustering is to produce a set of nested clusterings. The effect of increasing the measure of relative edge consistency σ, is that of progressively breaking the data set into a greater number of smaller clusters. Clearly not every link that is broken will partition the data set, this being desirable since it avoids spurious partitions (c.f. Zahn's method).

It is clear that this method will allow 'chained clusters' which is consistent with a visual model of clustering. This is often criticized as being a defect in a clustering method, but Jardine & Sibson [5.21] stress that continuity in the ultrametric dissimilarity coefficient is more important.

5.3.3 Examples

A number of two-dimensional dot patterns are given in order to see the effectiveness of the clustering method in terms of examples that can be interpreted visually. These problems are similar to those given by Zahn [5.3], and a value of $\sigma = 0.5$ has been used frequently for comparison.

The first examples (Figs 5.5-5.7) show a number of clusters which may be easily separated visually, either because of being spatially separate or because they have different point densities. The method successfully distinguishes between these rather obvious clusters.

The second set of problems are examples of touching clusters (Figs 5.8-5.10). Evidently the question of touching clusters is not well defined [5.2], however each of these examples would probably be regarded as touching. Fig 5.8 shows two clusters joined by a few obvious strays, the clusters being distinguished successfully. In Fig 5.9 there is a homogeneous bridge linking the clusters, which is not split. Fig 5.10 problem shows touching clusters with obviously different point densities in each cluster; these are clearly distinguished.

In order to see the operation of the hierarchic clustering algorithm on a real data set, the two class Iris data was used. The classes used (Iris Setosa and Iris Versicolor) are well known to be disjoint, so a partition might be expected at a relatively high d^*. The projection of the GG of the Iris data (Fig 5.11(a)) was obtained using the first two Karhunen-Loeve axes. The dendrogram (Fig 5.11(b))

shows a definite partition into the two Iris classes ($d^* = 6.199$), and partitioning within the classes at a lower level.

5.3.4 A New Hierarchic Clustering Algorithm

The clustering methods based on $S_1(\sigma)$ and $S_2(\sigma)$ may be used in two ways. The first is the obvious 'direct' approach, partitioning the data set for a given value of σ (e.g. $\sigma = 0.5$). The second is to produce a hierarchic clustering from one of these graphs. The 'direct' clustering is fairly obvious from a consideration of algorithms for computing any graph $S_1 \in S$. The clustering methods based on $S_1(\sigma)$ and $S_2(\sigma)$ could be implemented by either divisive or agglomerative algorithms. Here we consider an agglomerative algorithm.

The clustering algorithm has been implemented on a GEC 4070 computer in four programs. The aim is to produce a dendrogram to represent the hierarchic clustering. We require to find the ordering of the n points along the bottom of the dendrogram (which we store in an n-element 'dendroram position vector') as well as the partition levels.

Program 1

(a) Obtain the GG or RNG in the required metric

(b) For each edge $(p_i,p_j) \in$ GG (or RNG) find the value of d' required to break that edge ($d' = 1/\sigma'$).

Program 2

Sort graph edges in ascending order of dissimilarity

Program 3

(a) Form n clusters of one member (i.e. C_1)

(b) For j=1,n-1

(i) Take new graph edge (p_i,p_k) until p_i and p_k belong to different clusters.

(ii) Form partition at level d^* C_{j+1} by merging $c_{j.m}$ and $c_{j.l}$ $(d'=d^*)$

(iii) Store label and number of points in cluster 'lost' through merging.

(c)(i) Initialize dendrogram position vector to C_n

(ii) For j=n-1,2

If $c_{j.q} \subset c_{j+1.r}$ replace first h entries of $c_{j+1.r}$ in dendrogram position vector by $c_{j.q}$, where h is number of points in $c_{j.q}$ (already found in (b)(iii))

Output value of d^*

Output position of partition for drawing dendrogram

Program 4

Draw dendrogram in the style of Rohlf [5.27].

Program 1 may use one of a number of algorithms (see Section 4.3.2). This requires the storage of n(n-1)/2 real locations and n(n-1)/2 binary locations for GEN-2. Program 2 requires the storage of e graph edges and workspace. Program 3 requires the storage of 6 vectors of n integers for merging and forming the dendrogram.

Initially the dendrograms were plotted with the partitions arranged as they appeared in the program. However the plotting time was significantly reduced by plotting the smaller side of the partition closer to the stem. This style of plotting also seems to make the dendrogram easier to interpret.

An interesting feature of this algorithm is that if at the start of program 2 we input the distance d corresponding to each graph edge rather than d', the single link dendrogram will be computed. This takes advantage of the fact that both the GG and RNG are supergraphs

of the MST, which has been shown to contain all edges required for single link clustering [5.28]. Thus the method can do relative distance and single link clustering in the one operation, to yield complementary descriptions of the data set.

Additionally we might use a hybrid of the relative and absolute distance as the basis of a clustering method. This can be done by multiplying the relative and absolute distances for each edges and using the hybrid edge dissimilarity d" in the clustering algorithm (where $d''(p_i,p_j) = d'(p_i,p_j).d(p_i,p_j)$).

5.3.5 Admissibility Criteria

The clustering techniques described here have had a measure of success in clustering data using an approach based on relative distance. Clearly there will be some disadvantages associated with the method and an objective comparison with other methods is required. We now consider the admissibility criteria, defined by Fisher & Van Ness [5.29], and subsequently used by Dubes & Jain [5.11]. We firstly define the criteria then list how they apply to the new clustering method. The properties are fairly easily understood and will not be proved formally.

Consider a set $P = \{p_1,p_2,\ldots,p_M\}$ of M points clustered into k clusters $c_1,c_2,\ldots,c_k$ where

$c_1 = \{p_1,\ldots,p_j\},\ c_2 = \{p_{j+1},\ldots\},\ldots,\ c_k = \{\ldots,p_M\}$

Let $Q = \{q_1,q_2,\ldots,q_M\}$ denote any reordering of the points and let the set of k clusters $c'_1,c'_2,\ldots,c'_k$ be the **image** of $c_1,\ldots,c_k$ where

$c'_1 = \{q_1,\ldots,q_j\},\ c'_2 = \{q_{j+1},\ldots\},\ldots,\ C'_k = \{\ldots,q_M\}$

Then a clustering $C(P) = \{c_1,\ldots,c_k\}$ is said to be **image admissible** if

it does not have an image that is uniformly better in the sense that (1) $d^*(p_i,p_j) > d^*(q_i,q_j)$ where the ith and jth points belong to the same cluster and

(2) $d^*(p_i,p_j) < d^*(q_i,q_j)$ where the ith and jth points belong to different clusters.

A clustering $C(P) = \{c_1,\ldots,c_k\}$ is said to be **convex admissible** if the convex hulls of each of the k clusters do not intersect. Clearly in some cases it is desirable to have non-linearly separable clusters whereas in other situations it is not. Hence this criterion may be helpful in choosing an appropriate technique.

Given any set of points P we find the 'linkage' or MST of each cluster $L_1,L_2,\ldots,L_k$. The clustering is **linkage admissible** if the linkages are pairwise disjoint (an infringement of this criterion is shown in Fig 5.12). It is worth noting that this definition is only applicable to two dimensional data.

A data set is **well structured k-group** if there exists a clustering $C(P) = \{c_1,\ldots,c_k\}$ such that all within cluster distances are shorter than between cluster distances

A clustering is **exact tree admissible** if it generates the same dendrogram as the single linkage (nearest neighbour method) given ultrametric data).

Some criteria refer to specific situations. Proportion admissibility refers to the duplication of points in clusters. A clustering method is **point proportion admissible** if points can be duplicated without altering the resulting clustering. A clustering method is **cluster proportion admissible** if it gives the same clustering for duplication of points over an entire cluster. A procedure is **cluster omission admissible** if the omission of an entire

cluster in the input data does not result in a change in the remaining clusters.

In some situations the data is only of ordinal significance. In such cases only the rank order of the dissimilarities is significant. A clustering technique is **monotone admissible** if applying a monotone transformation to the dissimilarity matrix results in no change to the clustering.

We now give some admissibility results.

Lemma 5.1 Clustering based on the graphs S_1 & S_2 is image admissible.

This follows from the fact that the clustering is not determined by the order in which data is presented to the algorithm. The derived dissimilarity d^* is ultrametric ensuring that the method is image admissible.

Lemma 5.2 Clustering based on the graphs S_1 & S_2 is not convex admissible.

The graphs S_1 & S_2 allow nonlinearly separable clusters and hence do not meet this criterion.

Lemma 5.3 Clustering based on the graphs S_1 & S_2 is connected admissible.

This criterion is infringed if a point from one cluster lies on an MST edge of another cluster. Let us assume that the clustering based on S_1 or S_2 is not connected admissible. Then we have a situation similar to Fig 5.12 with the point c lying on or over the edge of the MST edge (a,b). Using S_1 or S_2 c lies within DISK(a,b) or within LUNE(a,b) and hence a and b cannot be connected by S_1 or S_2. Thus the position of c requires that a and b belong to different clusters giving a contradiction. Hence S_1 and S_2 are connected

admissible.

Lemma 5.4 Clustering based on the graphs S_1 & S_2 is not k-group admissible.

Lemma 5.5 Zahn's MST clustering method is not k-group admissible.

Both Zahn's MST clustering method and the S_1/S_2 clustering method are based on **relative** rather than **absolute** distance. They will therefore permit some intra-cluster absolute distances to be greater than some inter-cluster absolute distances subject to intra-cluster relative distances being less than inter-cluster relative distances.

Lemma 5.6 Clustering based on the graphs S_1 & S_2 are cluster proportion admissible.

The clustering based on S_1 and S_2 obtains its edge dissimilarity $d'(p_i,p_j)$ by considering neighbours that are not coincident with either p_i or p_j. The method is therefore insensitive to the duplication of either points or entire clusters.

Lemma 5.8 Clustering based on the graphs S_1 & S_2 is cluster omission admissible.

Lemma 5.9 Zahn's MST clustering method is not cluster omission admissible.

Fig 5.2(a) illustrates the fact that Zahn's method is not cluster omission admissible. The removal of the outlying cluster could result in a change in the remaining clustering. It is also evident that the omission of the same cluster will not alter the clustering if S_1 or S_2 are used.

Lemma 5.10 Clustering based on the graphs S_1 & S_2 is not monotone admissible.

This follows from the fact that the clustering is based on the ratio rather than the rank order of inter-point distances. If it

relied only on the rank order of distances it would be monotone admissible.

5.3.6 Appraisal

Since the techniques described are analagous in some ways to the MST methods of Zahn, it is worth returning to some of the problems encountered with that method. The problem of Fig 5.2(a) does not arise, and the obviously disjoint clusters are separated easily (Fig 5.2(b)). The problem of the graph not containing all consistent links is less likely to occur in methods based on the GG or RNG than in MST methods.

Obviously disjoint clusters are characterized by the clusters lasting over a reasonably wide range of d^* (Fig 5.11(b)), homogeneous clusters having different point densities (e.g. Fig 5.10(a)) tend to fragment over a narrow range of d^* (Fig 5.10(c)), but the dendrogram of random data from a uniform distribution, however, has neither of these properties.

It is worth observing that the clustering results above resemble those of Gowda & Krishna's mutual nearest neighbour (MNN) method [5.5]. In fact there is a tenuous link between the methods. We may define a **mutual nearest neighbourhood graph** of value 2, $S_{MNN}(2) \in S$ which links points having a mutual neighbourhood value of 2 and whose region definition is

$$R_{MNN}(p_i,p_j,2) = \{x:\min[d(x,p_i),d(x,p_j)] < d(p_i,p_j)\ i\neq j\}$$

Clearly $R_{MNN}(p_i,p_j,2) = R_1(p_i,p_j,1.0) = R_2(p_i,p_j,1.0)$ (see Fig 5.4(c))

Zahn [5.3] also gave a number of heuristic techniques for

clustering, involving separate tactics for dealing with clusters having necks, touching Gaussian clusters and point density problems. There may be problems in using these heuristics, but in view of their extensive use, it is worth considering the use of such heuristics in the framework of the GG and the RNG.

Zahn's heuristics were based on obtaining information along a diameter, or near diameter of the MST. The heuristic used in the touching cluster problem (joined by a neck) involved finding the depth of branching from the diameter. A partition was made at the best local minimum in the diameter histogram. The touching Gaussian cluster problem was approached by finding the minimum of a point density histogram along a cluster diameter.

It is clear that Zahn's touching cluster method cannot be directly embedded within the GG or RNG since branching depth implies a tree. However a heuristic might be constructed using the degree of a point on the diameter rather than branching depth. Fig 5.9 shows an example where any of these heuristics may produce useful results, however in Fig 5.13 none of the heuristics were likely to be very effective (see histograms, Figs 5.9(b) and 5.13(b)). The difference between the two results is probably attributable to the MST diameter passing through the centres of the clusters in Fig 5.9 whereas it passed close to the cluster boundary in Fig 5.13. However, the fact that the graphs $S_1(\sigma)$ and $S_2(\sigma)$ leave a neck may be exploited interactively [5.30] as described later.

An alternative to Zahn's Gaussian cluster heuristic might be to compute point density along the diameter using the neighbours of that point rather than just neighbours along the diameter. Such an approach would be less sensitive to actual details of the diameter.

Each of the above suggestions still relies on the MST diameter to partition the data and may be criticized on the same basis as Zahn's heuristics. So the above heuristics will yield less meaningful results than the clustering method described in section 5.3.4.

5.4 Interactive Clustering

We now return to the problem of cluster identification. In this section we propose a way in which information contained in the GG or RNG may be exploited to help in this problem. The result is an interactive display that complements the clustering method described in section 5.3.

5.4.1 An Interactive Clustering Display

The fundamental idea behind the interactive display is that the user selects a partition of interest from the dendrogram, then inter- and intra-cluster relationships are represented graphically. A hierarchic clustering method based on any of the above ideas will define a spanning tree in the data set, the spanning tree being a subgraph of the GG (or RNG) but not necessarily the minimal spanning tree (MST), leaving much of the geometric information contained in the GG (or RNG) unused. A partition of the data set may be used to define subgraphs and a contraction of the GG (or RNG), where each subgraph corresponds to a cluster in that partition and the vertices of the contracted graph correspond to individual clusters, enabling all the geometric information in the GG (or RNG) to be accessed and used interactively. The geometric ideas used for connecting points may be

extended to describing interconnection of clusters; it would be interesting for example to know the neighbourhood set of a given cluster.

Consider a hierarchic clustering $C(P) = C_1, C_2, \ldots, C_N$ where C_1 contains N clusters of 1 point (i.e. the set P), and C_N is a single cluster of N points. For a particular partition $C_k = \{c_{k1}, \ldots, c_{kl}, \ldots, c_{kp}\}$ where $p=N-k+1$ the **subgraph** G_{kl} of the GG (or RNG) corresponding to cluster c_{kl} has a vertex set $V(G_{kl})=c_{kl}$ and has an edge set $E(G_{kl})$ given by $\{(p_i,p_j)$ for all $(p_i,p_j) \in GG$, $p_i \in c_{kl}$, $p_j \in c_{kl}\}$. The **contraction** G_c of the GG (or RNG) corresponding to a partition C_k is obtained by contracting each edge of the GG (or RNG) within a cluster i.e. G_c has a vertex set $V(G_c)=C_k$ and edge set $E(G_c) = GG \setminus E(G_{kl})$ $l = 1,2,\ldots,p$

Having defined inter- and intra-cluster relationships in terms of subgraphs and a contraction of the GG, the crucial question is how these graphs are best displayed. A simple solution was to arrange the vertices of the graph around a circle in a meaningful way and then to mark in the graph edges. It was decided to position points using graph properties rather than use multidimensional scaling favoured by Jardine and Sibson [5.31] on their display for B_k clusters.

Cutpoints and cutedges are of considerable importance in identifying simple cluster types. This is especially so if the clustering method relies on point density changes because 'necks' in a cluster will remain unbroken unless they correspond to a change in point density. Thus if the cutpoints can be identified interactively this limitation of the clustering can be overcome.

Points were arranged around the display in such a way that biconnected components appeared on the same sector of the display.

Aho, Hopcroft and Ullman [5.32] describe an efficient algorithm to detect biconnected components of a graph of e edges in O(e) time. Each biconnected component is arranged in depth first order.

The display may be used interactively once the partition is selected and either the contraction or subgraph is chosen. Graph edges may then be broken interactively by varying dissimilarity to give an understanding of the geometry of the clusters in the partition or to give clues to the type of cluster generated.

5.4.2 Algorithm for finding the Display

Having described the display informally we now consider how to form such a display. Firstly we give some mathematical preliminaries, these results being stated without proof, the proofs being given in Aho, Hopcroft & Ullman [5.32].

The algorithm used for finding the display is based on a depth first search of a connected graph. Such a search partitions the edge set E of G into two subsets T and B. The set T is the set of **tree edges** and B is the set of **back edges** (Fig 5.14). An edge (p_i,p_j) is placed in the set T if, when we are at the vertex p_i examining the edges (p_i,p_j), the vertex p_j has not been visited; otherwise (p_i,p_j) is placed in B. The subgraph (V,T) of G is called the **depth first spanning tree**, and the vertex at which the tree was started is called the **root.**

Aho et al [5.32] describe an efficient algorithm for the depth first search of a graph. The algorithm runs in O(max(n,e)) time for a graph with n vertices and e edges. A tree may be ordered from its root by considering the root to be the senior ancestor and the points

connected to it to be descendents. This idea is embodied in the following lemma.

Lemma 5.11 If (p_i,p_j) is a back edge, then in the spanning tree p_i is an ancestor of p_j or vice versa.

The depth first search imposes a natural order on the vertices of the graph. This approach may then be modified to find blocks or biconnected components in the graph.

We can define a relation on the edge set E by saying that the edges e_x and e_y are related if there is a cycle containing both e_x and e_y $(x \neq y)$. This equivalence relation partitions the edge into classes so that two edges are in the same class if they lie on a common cycle. For a particular class of edges E_i having a vertex set V_i, we say that the graph $G_i = (V_i,E_i)$ is a biconnected component of G.

The following lemmas provide information on biconnectivity.

Lemma 5.12 Let $G_i = (V_i,E_i)$ be a biconnected component of a connected graph $G = (V,E)$ for $1 \leqslant i \leqslant k$ then

(1) G_i is biconnected for each i, $1 \leqslant i \leqslant k$

(2) For all $i \neq j$ $V_i \cap V_j$ contains at most one vertex.

(3) a is a cutvertex of G iff $a \in V_i \cap V_j$ for some $i \neq j$.

Lemma 5.13 Let $G = (V,E)$ be a connected graph, and let $S = (V,T)$ be a depth first spanning tree for G. Vertex a is a cutvertex of G iff

(1) a is the root and has more than one descendent or

(2) a is not the root and for some descendent s of a there is no back edge between any descendent of s and an ancestor of a.

Aho et al gave an algorithm for finding the biconnected components of a graph in O(e) time for a graph with e edges.

The display was formed using this information. The idea was to place points around the display so that the biconnected components

were displayed together. The algorithm would output points after using information from the depth first spanning tree and the biconnected components.

First all the points in the first biconneceted component are positioned in depth first order. Then the next biconnected component in the display is positioned in the same way, and the process is continued until all points are positioned around the display.

5.4.3 A Clustering Case Study

In order to understand better the problems of the new graph theoretical clustering methods and interactive display, we describe a case study using real data. Since this thesis is concerned with both theory an a particular application of pattern recognition methods, it is appropriate that we use the lung sound data sets of Chapter 7. As described later (Chapter 7) if one class (pulmonary oedema) is removed, the remaining data forms two obvious clusters (Figs 7.2(b) & 7.3(b)). Two versions of the data were available:- (a) the original 20-dimensional data set and (b) a set based on the first six features.

We therefore consider the application of the relative distance methods $S_1(\sigma)$ and $S_2(\sigma)$ to the more difficult 6-dimensional data set. Clusterings of this data based on absolute and hybrid distance are given in Chapter 7. The presentation here is therefore aimed at examining problems associated with the new methods.

The difference in difficulty in clustering the 20- and the 6-dimensional data sets by $S_1(\sigma)$ is apparent in Figs 5.15 and 5.16(a). Allowing for a few outliers, the 20-dimensional data partitions into two major clusters over a dissimilarity range of 1.25-1.55; however

the corresponding partition that occurs in the 6-dimensional data (Fig 5.16(a)) is less than obvious, even allowing for outliers. At a dissimilarity of 1.75 the plotting of the graph edges on the first two principal components of the data (Fig 5.16(b)) shows that no partitioning of the data has occured although many of the GG edges have been removed. At dissimilarities of 1.5 & 1.45 (Fig 5.16(c) & (d)) some of the points corresponding to normal patients (marked *) have broken away from the main cluster whereas the rest of the normals are linked to the asbestosis and CFA points (⊡ and △ respectively).

This situation was investigated using the interactive display (Fig 5.17). The inter-cluster connected graph (Fig 5.17(a) shows that all the clusters are linked to the main one (cluster 31) which spans most of the data set (Fig 5.17(d)). As the display is reduced to a dissimilarity of 1.45, the outlier points break away from cluster 31 without revealing any structure.

Fig 5.18 shows the results of using $S_2(\delta)$ to cluster the same data set. Allowing for outliers, the data forms a partition of two obvious clusters for dissimilarities between 1.6 and 1.7. Figs 5.18(b)-(d) show the fragmentation of the data set down to a dissimilarity of 1.5. Note that one CFA point (△) is clustered with the normals.

The inter-cluster relationships at a dissimilarity level of 1.5 are shown in Fig 5.19; the clusters with numbers greater than 140 being normals. The normals are connected to the other points by two edges (31,160) and (66,160). The edge (31,160) breaks without yielding a partition whereas (66,160) causes a partition between the normals and other points.

Displays of individual clusters are given in Fig 5.20. Fig

5.20(a) shows cluster 31 (the main asbestosis/CFA cluster in Fig 5.18(a)) is very well connected even at a low level reflecting that cluster's dense nature. In Fig 5.20(c) cluster 66 appears to be rather straggly with a few tails, and cluster 160 is simply a pair of points (Fig 5.20(d)).

Comparison of Figs 5.16(a) and 5.18(a) suggests that the partitioning produced by $S_2(\sigma)$ is better than that given by $S_1(\sigma)$. While it is pointless basing conclusions on one data set it is possible that the fact that the RNG is sparser than the GG means that $S_2(\sigma)$ gives a better partition than $S_1(\sigma)$. It is worth noting that the poor results obtained by using $S_1(\sigma)$ are improved by the use of hybrid dissimilarities (Fig 7.4(d)).

5.4.4 Discussion

The case study should serve to illustrate how additional information on a clustering may be recovered interactively. This display of course does have a number of limitations. Firstly the method depends on the ability of the GG (or RNG) and their associated edge dissimilarities to reveal the geometry of the data. Secondly the attempt to provide cluster identification is dependent on finding cutpoints in the graph displayed.

There is also the effect of dimensionality, which will directly affect réponse times on the display. The arrangement of points around the display is found in O(e) time where e is the number of graph edges. Devroye [5.33] has shown that the expected number of Gabriel neighbours of a vertex is 2^n for any underlying density as $N \to \infty$. (At present the expected degree of a vertex in the RNG is not known).

Therefore this computation may increase significantly with dimensionality, deteriorating response times and so degrading user performance [5.34].

There is also a limit to the number of points that may be usefully displayed around a circular display. As the number of points increases to 40 and more it becomes difficult to see the structure of the graph displayed. However this effect could be minimized by marking the cutnodes and cutedges in a different colour, or by replacing individual points in a large cluster by a low level clustering.

Toussaint and Poulsen [5.35] describe a heuristic dual space feature selection method based on Zahn's MST clustering methods [5.3]. This involved clustering the n features in N-dimensional space then selecting say one feature per cluster. Wismath, Soong and Akl [5.17] obtained encouraging results by using this heuristic in conjunction with a non-linear mapping algorithm. Clearly the interactive cluster display could be used instead of the nonlinear mapping. However Roberts, Henderson and Hanka [5.36] point out that proximity in the dual space - implicit in the heuristic - is not guaranteed to give a good feature set.

The idea of using a circular display to show geometry in a clustering need not be confined to the above methods. In a partitional clustering it would be interesting to study inter-cluster relationships perhaps by finding the Gabriel graph of cluster centroids, then displaying the graph interactively. In some data sets the points may simply form a cloud with no spatial cluster structure. If a **priori** labels are available it is possible that points with a particular label occupy a paticular part of the space. In such a case if the 'cloud' of points can be broken up by a clustering algorithm

the display could be used to explore the inter-cluster geometry.

5.5 Overall Discussion

A nonparametric hierarchic clustering method has been developed based on the concept of limited neighbourhood sets [5.37]. By considering a family of graphs (which includes the GG and RNG), two types of graph have been defined that are capable of partitioning the data set according to relative distance. The method is capable of distinguishing disjoint clusters and homogeneous clusters separable by changes of point density. It must be stressed that whether or not the clusters produced by the method are reasonable to the user, the dendrogram gives a meaningful summary of local data structure in terms of relative distance.

Some of the limitations of the clustering method are considered. Although the treatment is far from exhaustive the admissibility criteria do provide a basis for comparison with some other methods. The fact that the technique is locally sensitive is worth considering when using the technique. The ability of the method to partition data is limited by the extent to which the structure is shown at a local level. This presents no problem in well structured point sets, but in many situations global considerations should be taken into account. In such situations it is probable that a global clustering algorithm would be appropriate although it is possible that some sort of preprocessing of the input data could allow the cluster structure to be better reflected at a local level [5.38].

A comparison with Zahn's heuristics for clustering using the MST

is given. The illustrations suggest that some of Zahn's heuristics, notably those using the MST diameter will not always work well. It might be wondered how useful results will be when based solely on some heuristic tactic as opposed to some consistent criterion. Other heuristics e.g. that of Lee [5.39] are designed for efficient computation of methods at the expense of suboptimal results. Although Dubes & Jain [5.11] report useful results from Zahn's heuristics it is felt that they must be used with care.

There are a number of problems associated with using clustering algorithms. Two of these - the problem of cluster identification and the lack of information on inter-cluster relationships - may be tackled interactively. The display was successful on some simple problems by providing the user with information that would otherwise be lacking. Despite limitations to the display, this facility should prove useful to the user. Clearly the display used is by no means the only one that is possible, and other information could be accessed using different means. Further research into interactive aids for the cluster user is urged.

Finally it should be stressed that there is no such thing as a universal clustering method. Since the value of a particular technique is data dependent, we should ideally have a range of techniques available offering different types of clustering [5.11]. In this respect the above clustering method should provide a reasonable clustering based on the visual idea of the cluster.

References

5.1 P.M.Lankford, Regionalization: Theory and alternative algorithms, Geogrl. Anal., 1, 196-212 (1969)

5.2 R.Mizoguchi and M.Shimura, A non-parametric algorithm for detecting clusters using hierarchical structure, IEEE Trans. Patt. Anal. Mach. Intelleg., PAMI-2, 292-300 (1980)

5.3 C.T.Zahn, Graph-theoretical methods for detecting and describing Gestalt clusters, IEEE Trans. Comput., C-20, 68-86 (1971)

5.4 W.L.G.Koontz, P.M.Narendra and K.Fukunaga, A graph-theoretical approach to non-parametric cluster analysis, IEEE Trans. Comput., C-25, 936-944 (1976)

5.5 K.C.Gowda and G.Krishna, Agglomerative clustering using the concept of mutual nearest neighbourhood, Pattern Recognition, 10, 105-112 (1978)

5.6 R.A.Jarvis, Shared nearest neighbour maximal spanning trees for cluster analysis, Proc. 4th Joint Conf. on Pattern Recognition, 308-313, Kyoto, Japan (1978)

5.7 R.A.Jarvis and E.A.Patrick, Clustering using a similarity measure based on shared nearest neighbors, IEEE Trans. Comput., C-22, 1025-1034 (1973)

5.8 M.Delattre and P.Hansen, Bicriterion Cluster Analysis, IEEE Trans. Patt. Anal. & Mach. Intelleg., PAMI-2, 277-291 (1980)

5.9 G.T.Toussaint, Pattern recognition and geometrical complexity, Proc. 5th International Conference on Pattern Recognition, 1324-1347, Miami U.S.A. (1980)

5.10 R.A.Fisher, The use of multiple measurements in taxonomic problems, Ann.Eugenics, 7, 178-188 (1936)

5.11 R.Dubes and A.K.Jain, Clustering techniques: the user's dilemma, Pattern Recognition, 8, 247-260 (1976)

5.12 R.Dubes and A.K.Jain, Validity studies in clustering methodologies, Pattern Recognition, 11 (1979)

5.13 E.Backer and A.K.Jain, A clustering performance measure based on fuzzy set decomposition, IEEE Trans. Patt. Anal. & Mach. Intelleg., PAMI-3, 66-74 (1981)

5.14 T.A.Bailey and R.Dubes, Cluster validity profiles, Pattern Recognition, 15, 61-83 (1982)

5.15 S.P.Smith and R.Dubes, Stability of a hierarchical clustering, Pattern Recognition, 12, 177-187 (1980)

5.16 Y.T.Chien, Interactive Pattern Recognition, Marcel Dekker (1978)

5.17 S.K.Wismath, N.P.Soong and S.G.Akl, Feature selection by interactive clustering, Pattern Recognition, 14, 75-80 (1981)

5.18 H.Niemann, Interactive clustering of patterns, Proc. 4th International Joint Conference on Pattern Recognition, 301-304, Kyoto Japan (1978)

5.19 B.Rosenberg and D.J.Langridge, A computational view of perception, Perception, 2, 415-424, (1973)

5.20 J.F.O'Callaghan, Human perception of homogeneous dot patterns, Perception, 3, 33-45 (1974)

5.21 N.Jardine and R.Sibson, Mathematical taxonomy, Wiley (1971)

5.22 J.F.O'Callaghan, An alternative definition for neighbourhood of a point, IEEE Trans. Comput., C-24, 1121-1125 (1975)

5.23 G.T.Toussaint, The relative neighbourhood graph of a finite planar set, Pattern Recognition, 12, 261-268 (1980)

5.24 D.W.Matula and R.R.Sokal, Properties of Gabriel graphs relavent to geographic variation analysis and the clustering of points in the

plane, Geogrl. Anal., 12, 205-222 (1980)

5.25 C.Jardine, N.Jardine and R.Sibson, The structure and construction of taxonomical hierarchies, Math. Biosciences, 1, 173-179 (1967)

5.26 S.C.Johnson, Hierarchical Clustering Schemes, Psychometrika, 32, 241-254 (1967)

5.27 F.J.Rohlf, Hierarchical clustering using the minimum spanning tree, Comput.J., 16, 93-95 (1973)

5.28 J.C.Gower and G.J.S.Ross, Minimum spanning trees and single-linkage cluster analysis, Applied Statistics, 18, 54-64 (1969)

5.29 L.Fisher and J.W.Van Ness, Admissable clustering procedures, Biometrika, 58, 91-104 (1971)

5.30 R.B.Urquhart and J.E.S.Macleod, Interactive hierarchic clustering: a display based on graph theoretical methods, Proc. International Conference on Cybernetics and Society, IEEE, 11-15, Atlanta U.S.A. (1981)

5.31 N.Jardine and R.Sibson, The constrction of hierarchic and non-hierarchic classifications, Computer J., 11, 177-184 (1968)

5.32 A.V.Aho, J.E.Hopcroft and J.D.Ullman, The design and analysis of computer algorithms, Addison-Wesley (1974)

5.33 L.Devroye, Personal communication

5.34 L.Leiker, Human factors relevant to the computer user, Proc. Southeast Region ACM Conf., Atlanta, U.S.A. (1981)

5.35 G.T.Toussaint and R.S.Poulsen, Some new algorithms and software implementation methods for pattern recognition research, IEEE Computer Society's Third International Computer Software and Applications Conference, 55-63 (1979)

5.36 S.J.Roberts, L.P.Henderson and R.Hanka, The dual space and its use in feature selection, Proc. 5th International Conference on

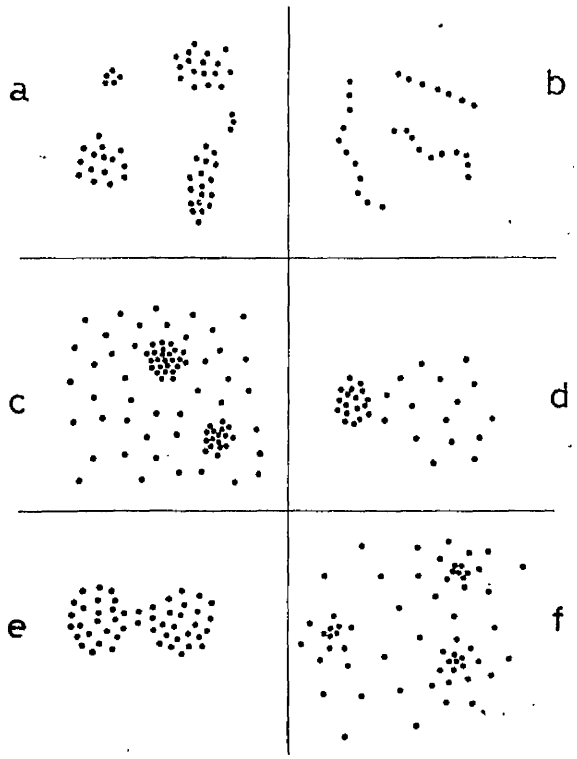
Pattern Recogntion, 1188-1190, Miami U.S.A. (1980)

5.37 R.B.Urquhart, Graph theoretical clustering based on limited neighbourhood sets, Pattern Recognition, 15, 173-188 (1982)

5.38 F.R.Dias Velasco and A.Rosenfeld, Some methods for the analysis of sharply bounded clusters, IEEE Trans. Syst., Man & Cybernet., 10, 511-518 (1980)

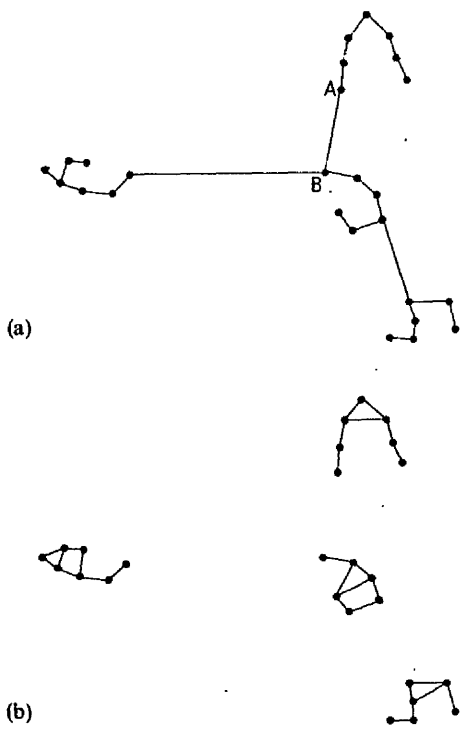
5.39 R.C.T.Lee, A sub-minimal spanning tree approach for large data clustering, Proc. 2nd Internat. Joint Conf. on Pattern Recognition, Copenhagen, 22-26 (1974)

FIG 5.1



. Examples of visually identifiable clusters.

FIG 5.2



(a) Inconsistency in Zahn's method (using depth of 3). Link AB will be broken at inconsistency of 1.95, but if the outlying cluster is removed AB breaks at inconsistency of 3.23; (b) method based on $S_1(\sigma)$. Link always breaks at $d^* = 3.33$ ($\sigma = 0.3$).

FIG 5.3

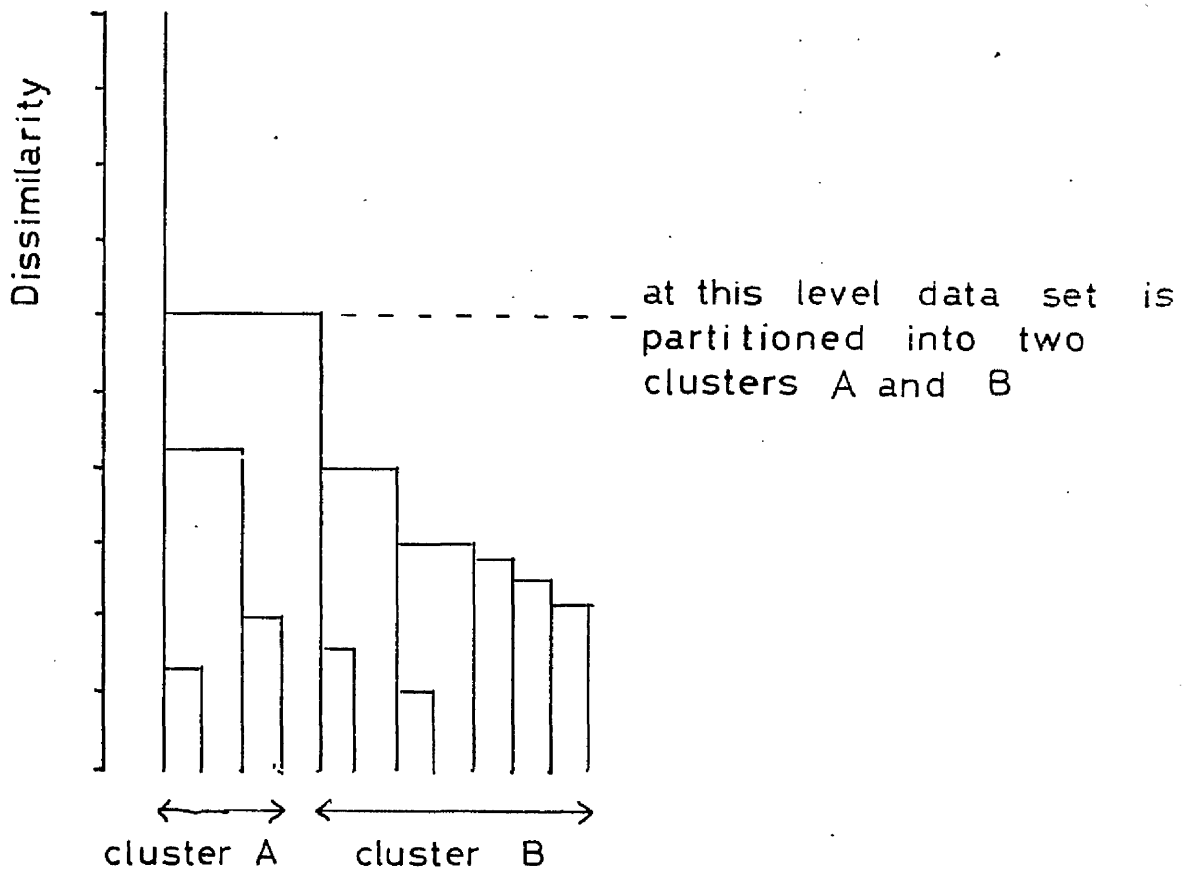
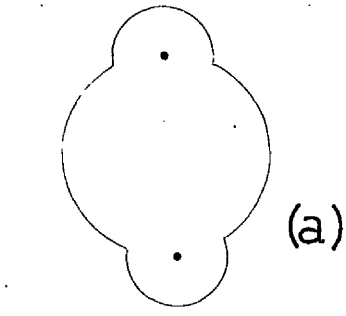
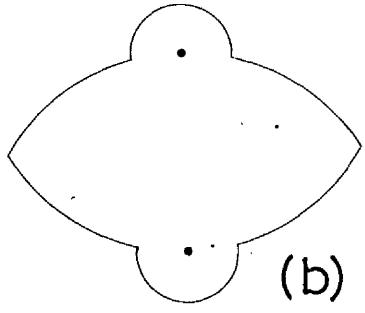

Dissimilarity
at this level data set is partitioned into two clusters A and B
cluster A
cluster B

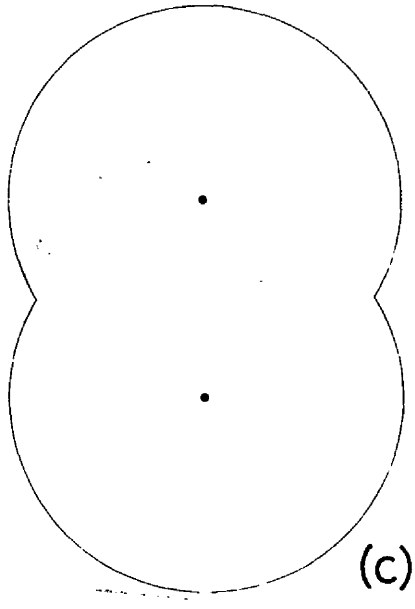
FIG 5.4



(a)

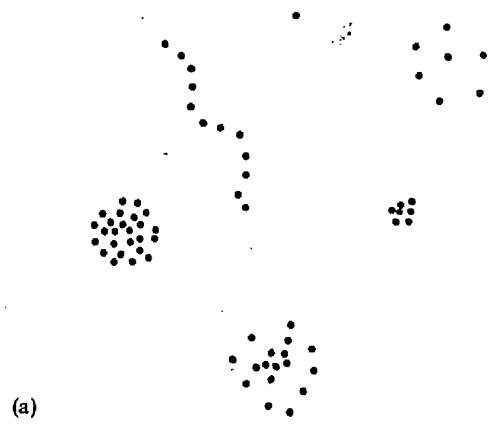


(b)

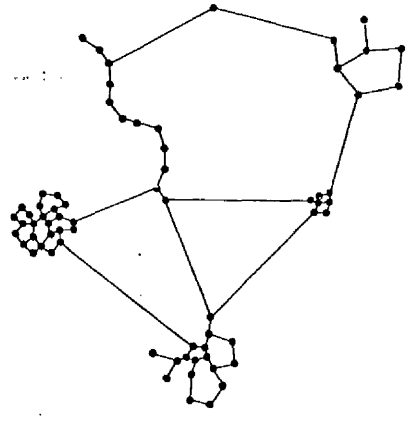


(c)

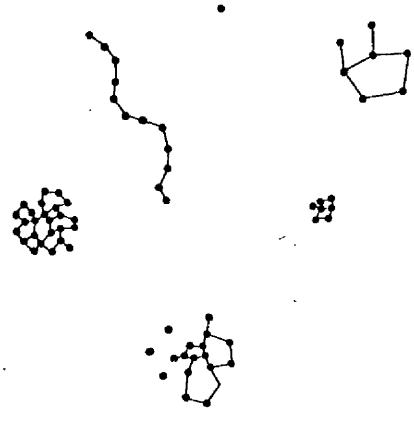
FIG 5.5



(a)

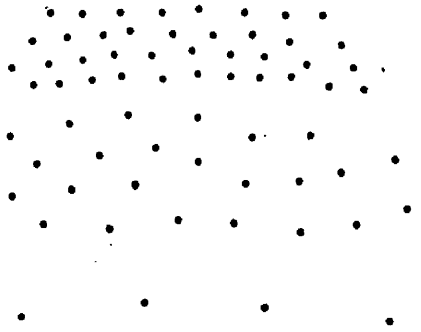


(b)



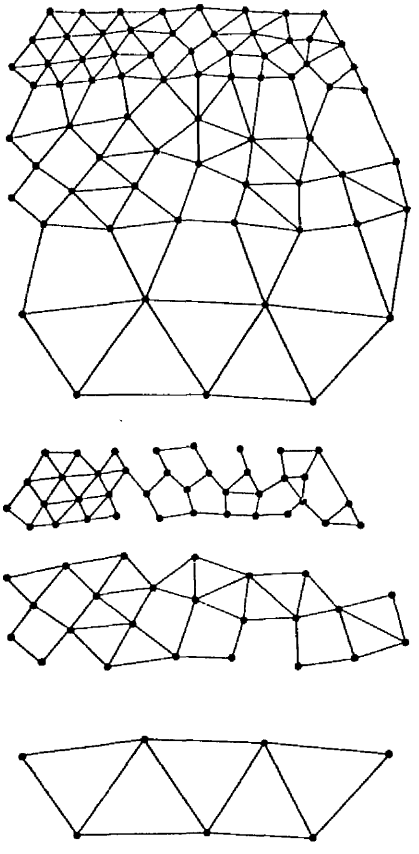
(c)

(a) A set of points showing several obvious clusters; (b) relative neighbourhood graph of (a); (c) $S_2(0.5)$ of (a).



(a)

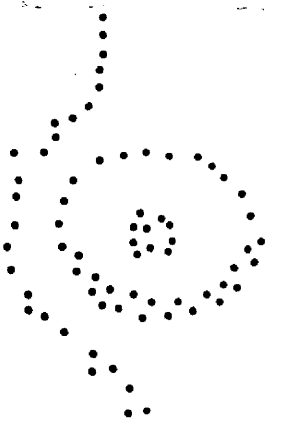
FIG 5.6



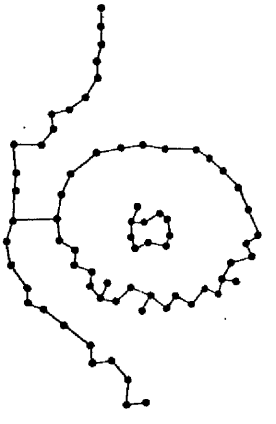
(b)

(c)

(a) A set of points showing three homogeneous regions of different point densities; (b) GG of (a); (c) $S_1(0.5)$ of (a).



(a)

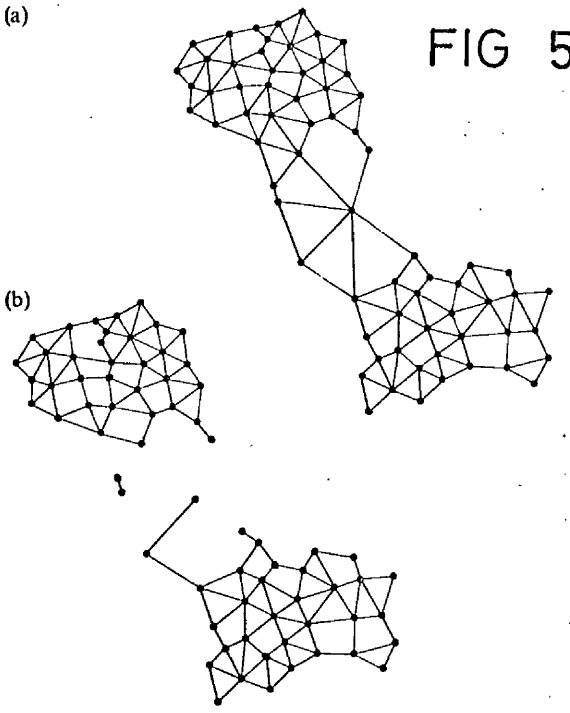


(b)

FIG 5.7

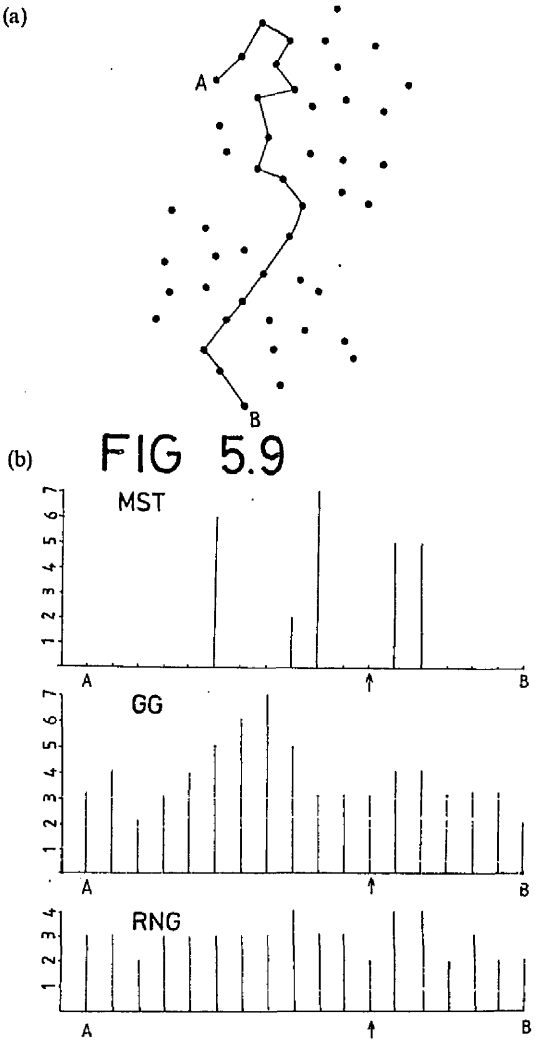
(a) A set of points; (b) $S_2(0.5)$ of (a). Note that $S_2(0.55)$ separates the chained cluster from the annulus, fragmenting the chain slightly.

FIG 5.8

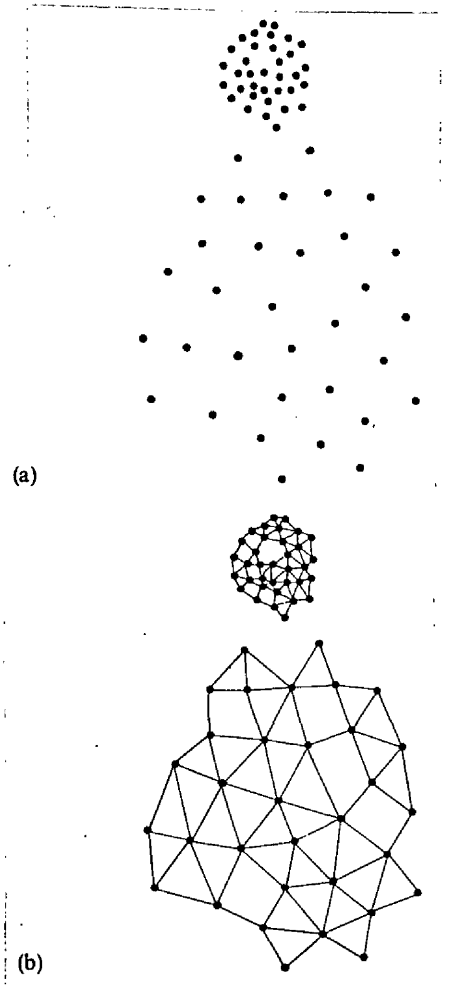


(a) Gabriel graph of a set of points forming two clusters with intermediate strays; (b) $S_1(0.5)$ of (a).

FIG 5.9



(a) A set of points having a neck between subclusters [no partition produced by either $S_1(0.5)$ or $S_2(0.5)$]; (b) histograms of MST diameter using MST, GG and RNG [$S_1(0.5)$ and $S_2(0.5)$ have histograms almost identical to those of the GG and RNG respectively]. The MST histogram uses branching depth, whereas the GG and RNG ones use degree of each point. The arrow indicates the obvious cutpoint.



For caption see over.

FIG 5.10

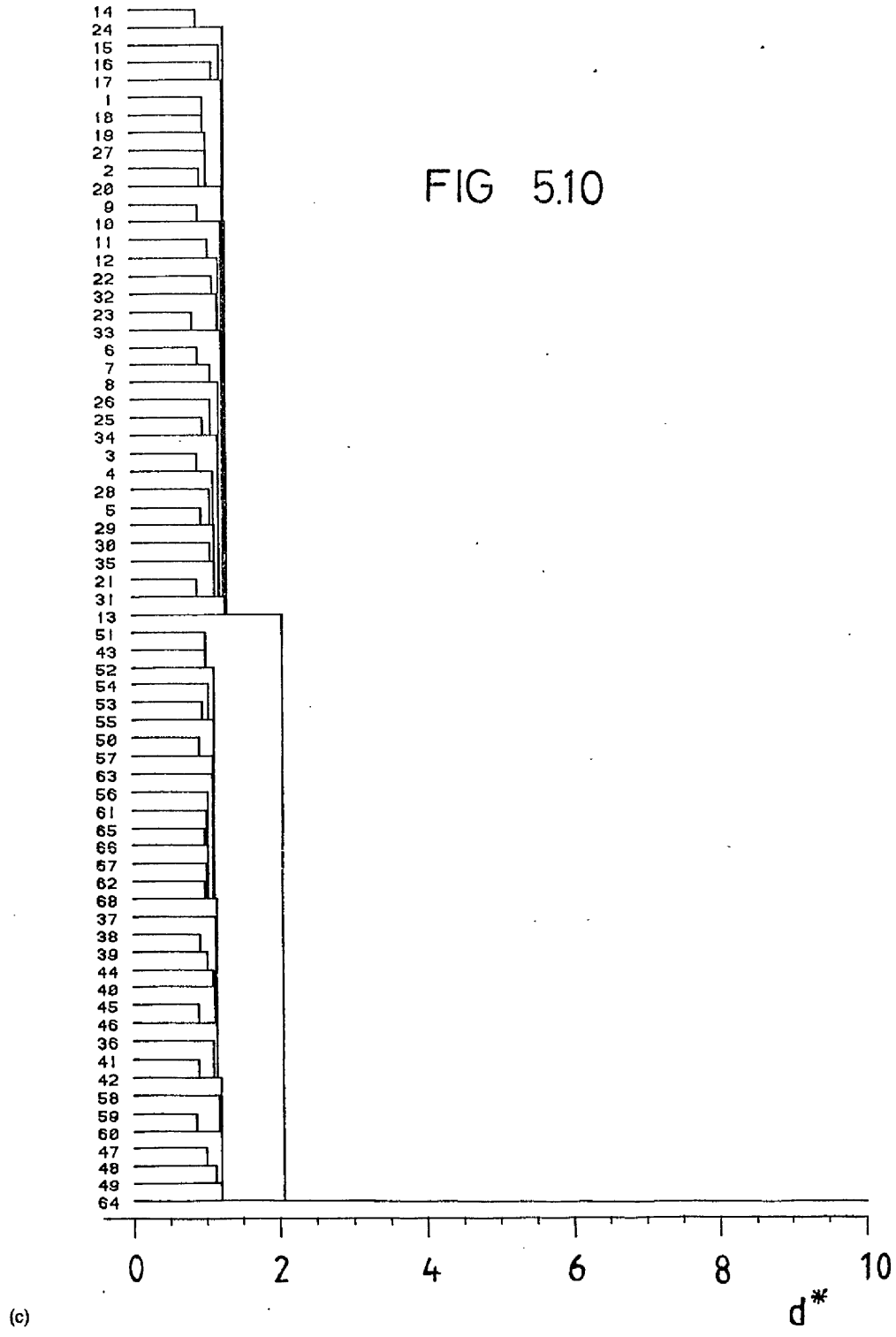
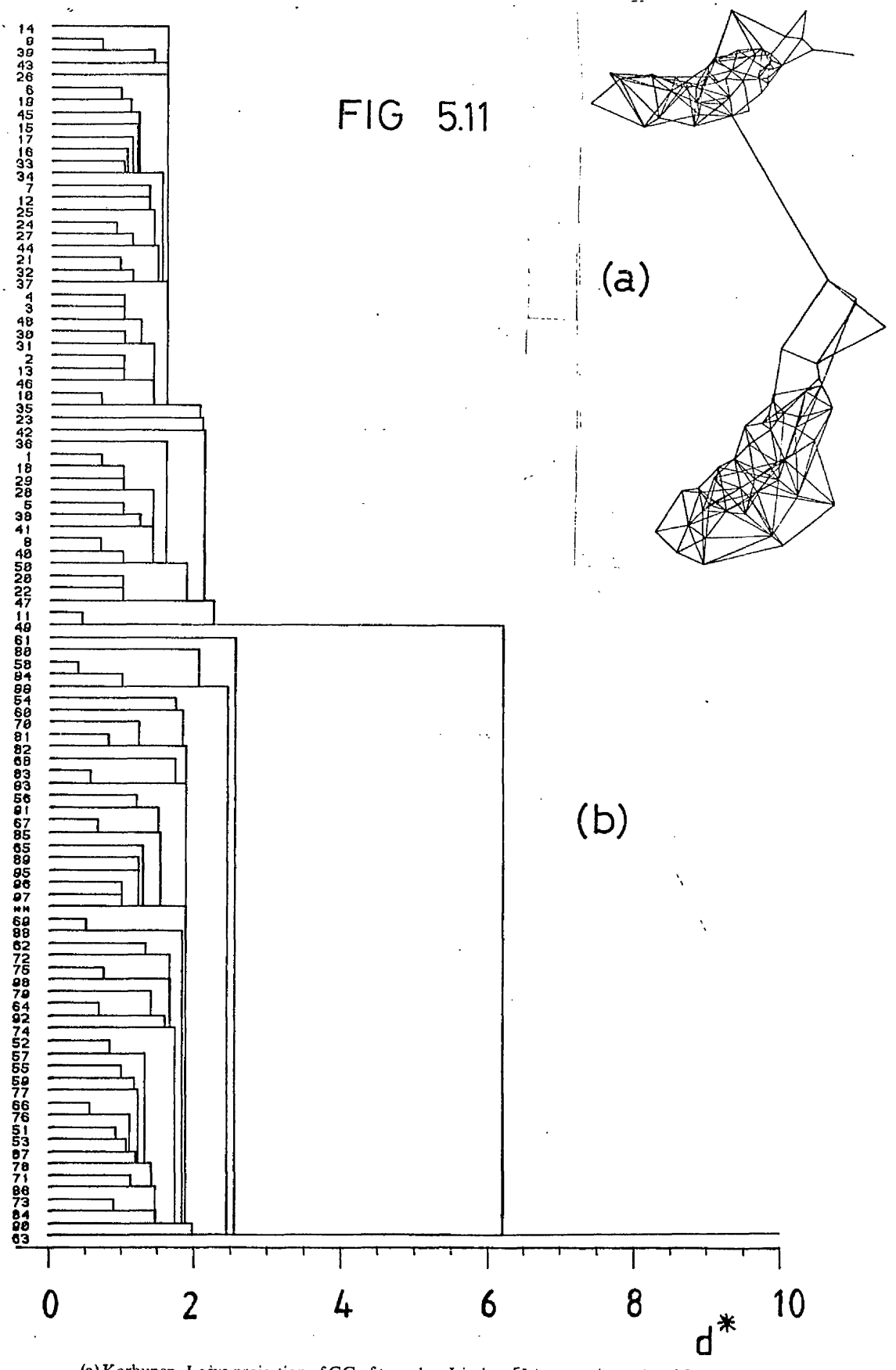


FIG 5.10

(c)

(a) A set of points showing two touching clusters of different point densities; (b) $S_1(0.5)$ of (a); (c) dendrogram of (a).



(a) Karhunen–Loève projection of GG of two-class Iris data [*Iris setosa* (upper) and *Iris versicolor* (lower)]; (b) dendrogram of two class Iris data with *I. setosa* 1–50 upper partition and *I. versicolor* 51–100 lower partition.

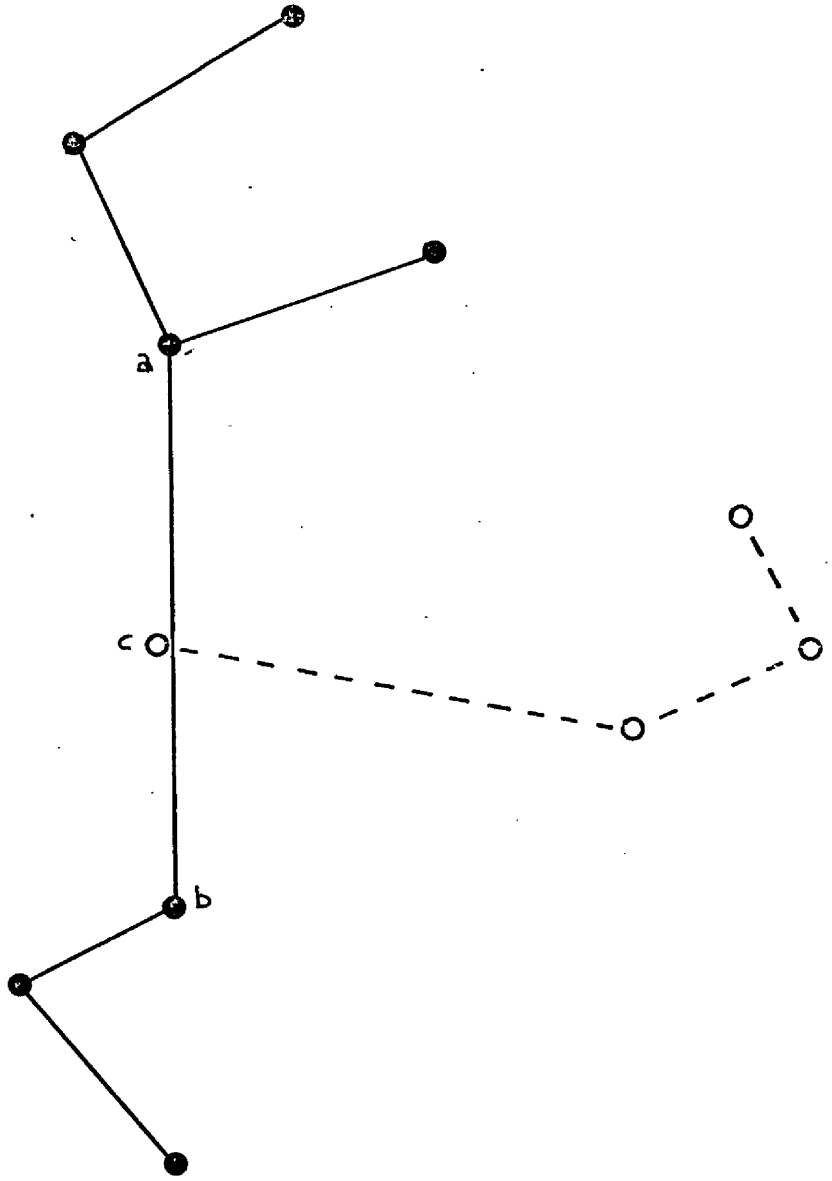


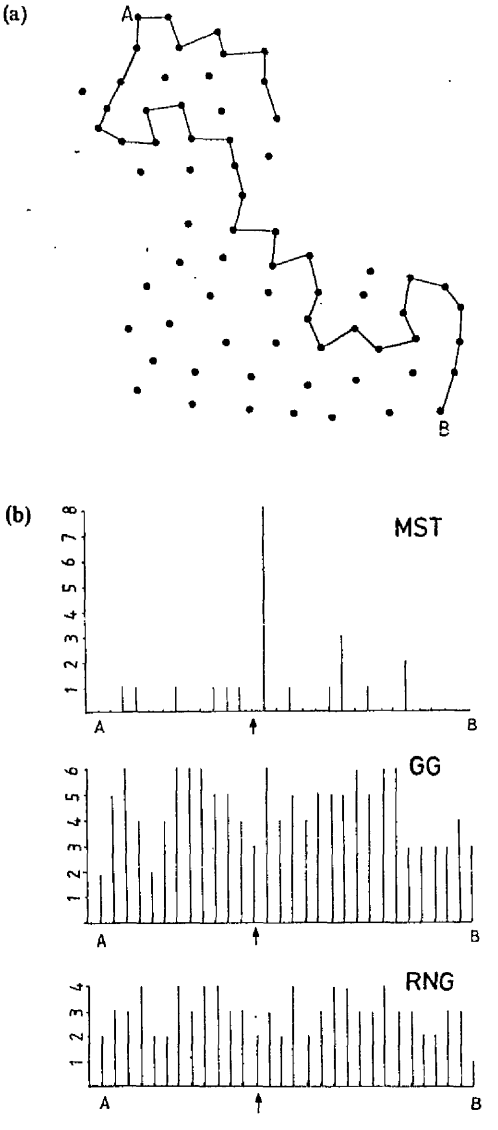
FIG 5.12
Infringement of linkage admissibility

Table 1

Criterion	S_1/S_2	Zahn MST	Single link
Image	yes	yes	yes
Convex	no	no	no
Connected	yes	yes	yes
Exact tree	*	*	yes
k-group	no	no†	yes
Point proportion	yes	yes	yes
Cluster proportion	yes	yes	yes
Monotone	no	no	yes
Cluster omission	yes	no†	yes

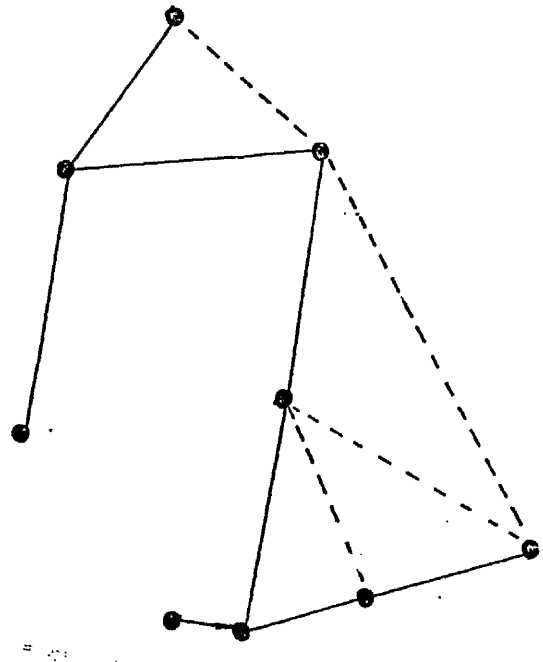
* Not applicable since method based on relative distance.
† See Jarvis.[6]

FIG 5.13



(a) Set of points showing a neck between two subclusters; (b) histograms along the MST diameter. Note the difficulty in drawing conclusions from the histograms – except, perhaps, that of the GG. Arrow indicates obvious cutpoint.

FIG 5.14



Depth first search:
Tree edges are solid
Back edges are dashed

FIG 5.15

S_1 clustering of the 20-D lung sound data

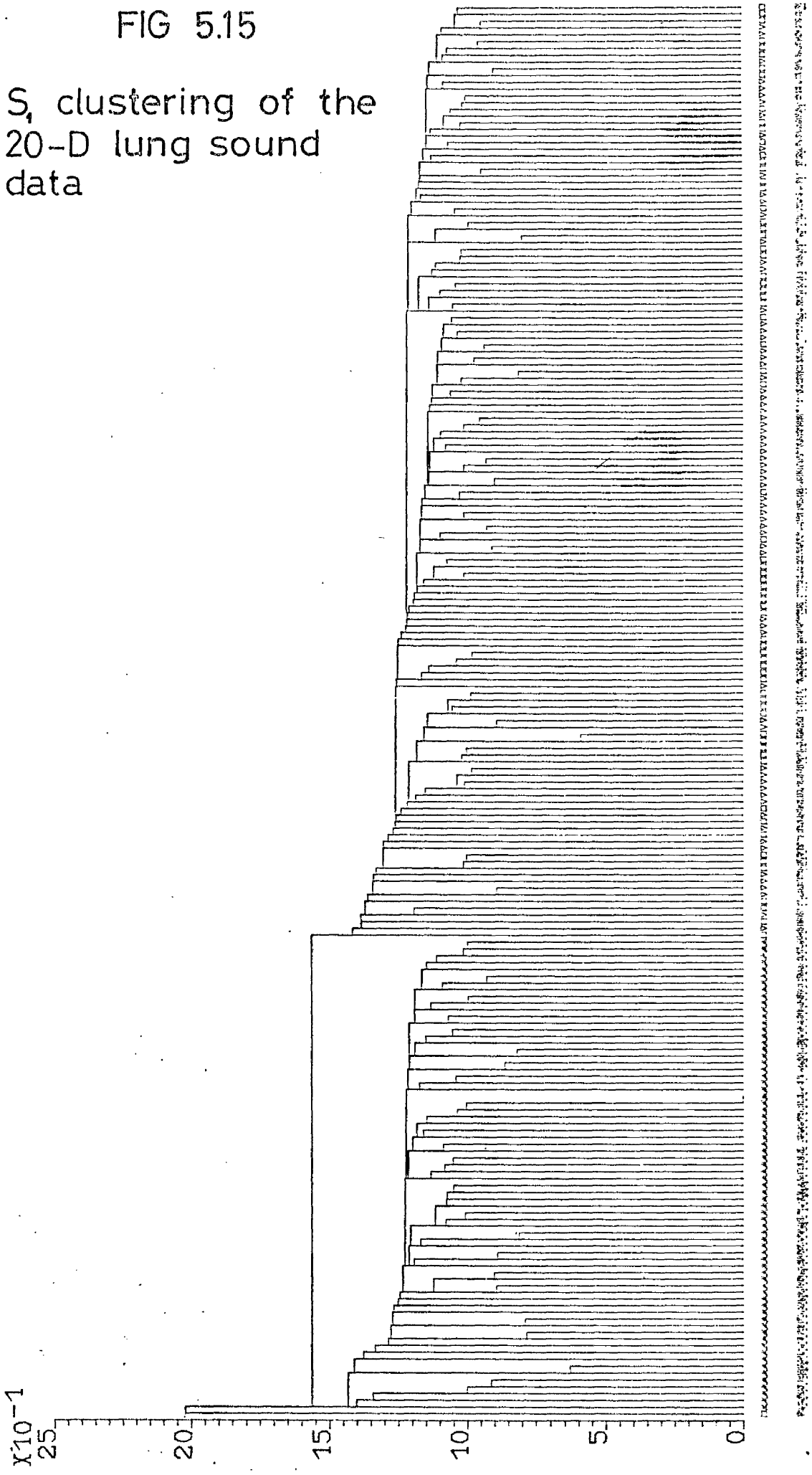


FIG 5.16(a)

S_1 clustering of the 6-D lung sound data

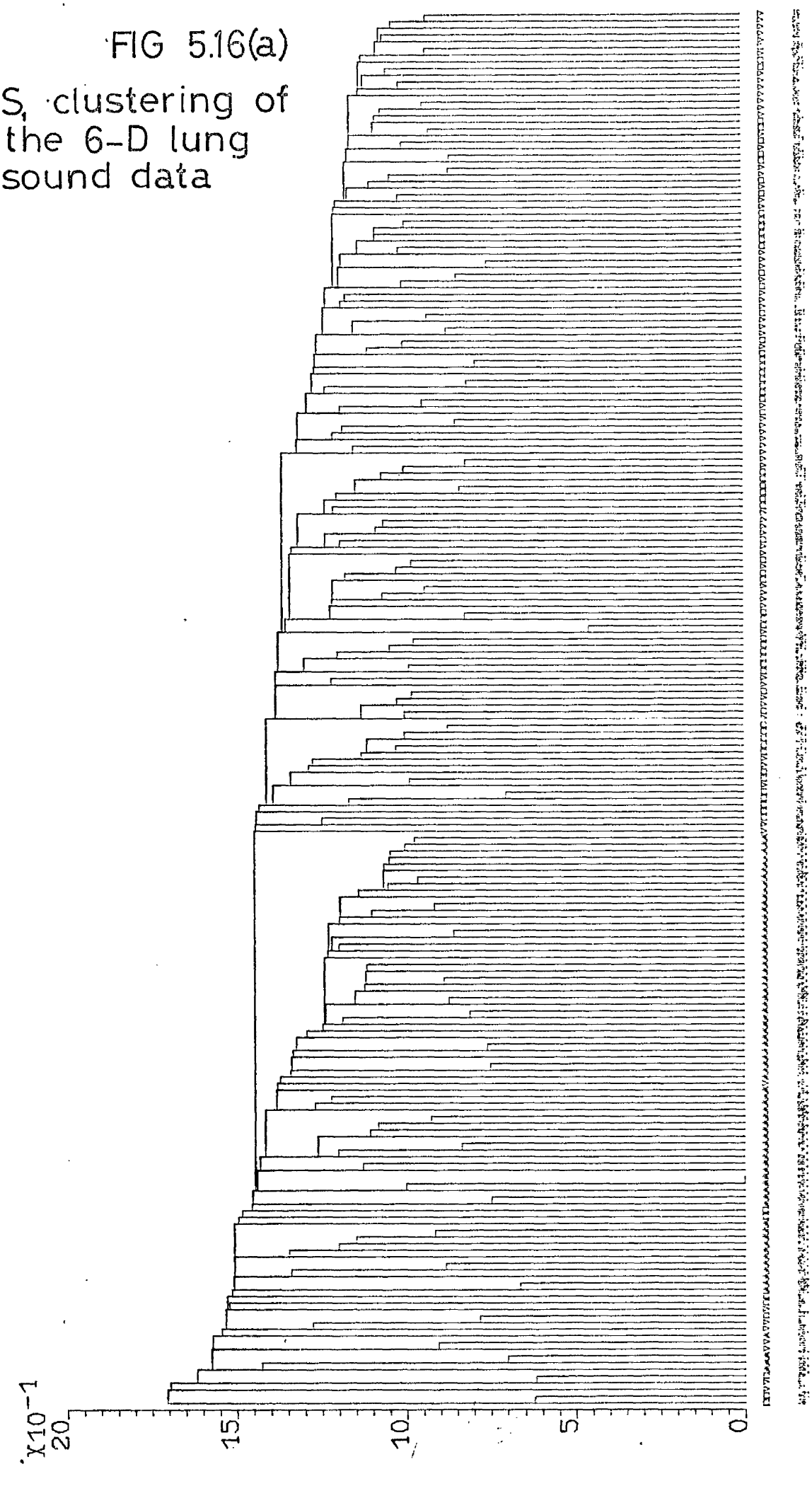


FIG 5.16(b)

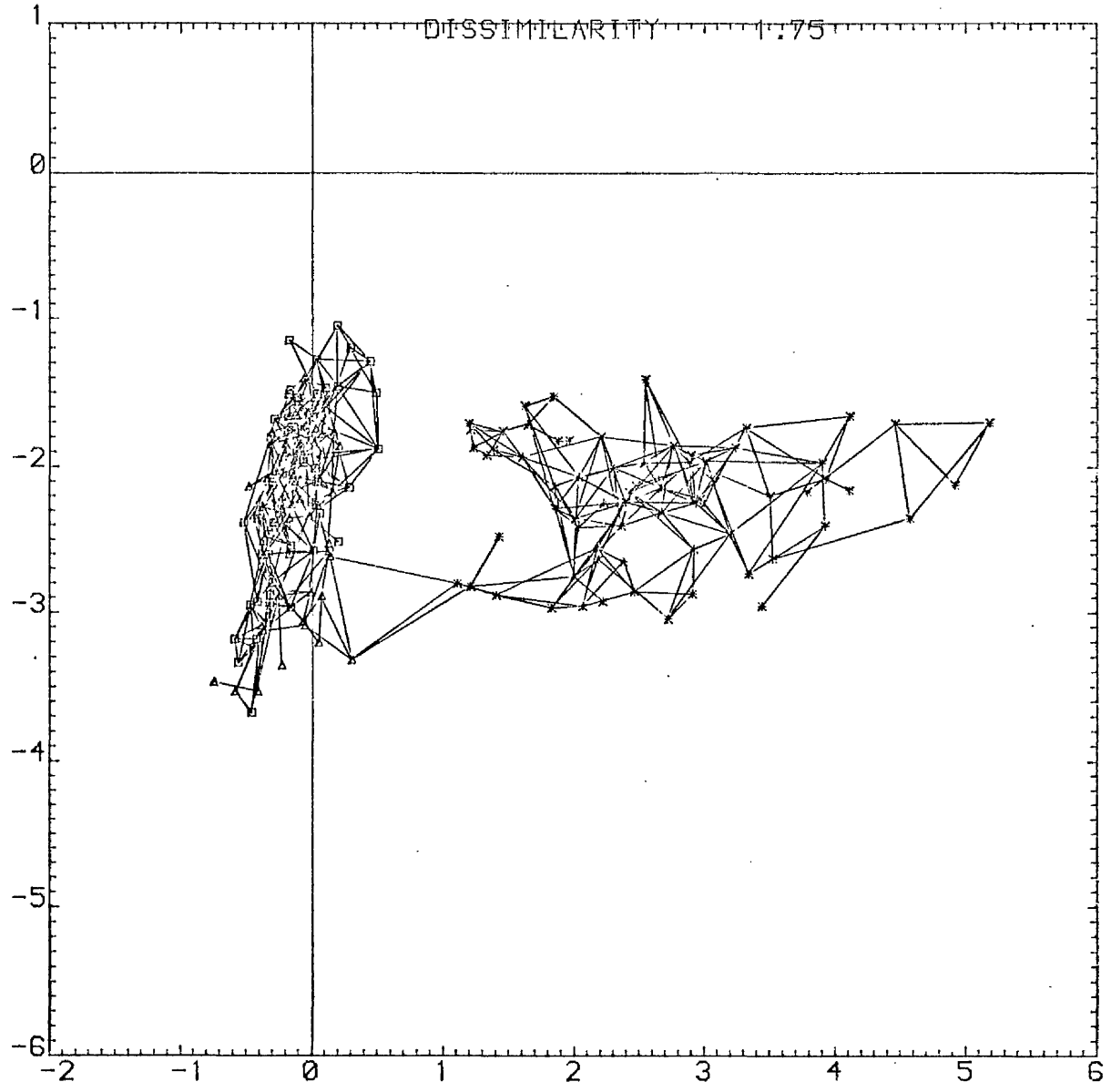


FIG 5.16(c)

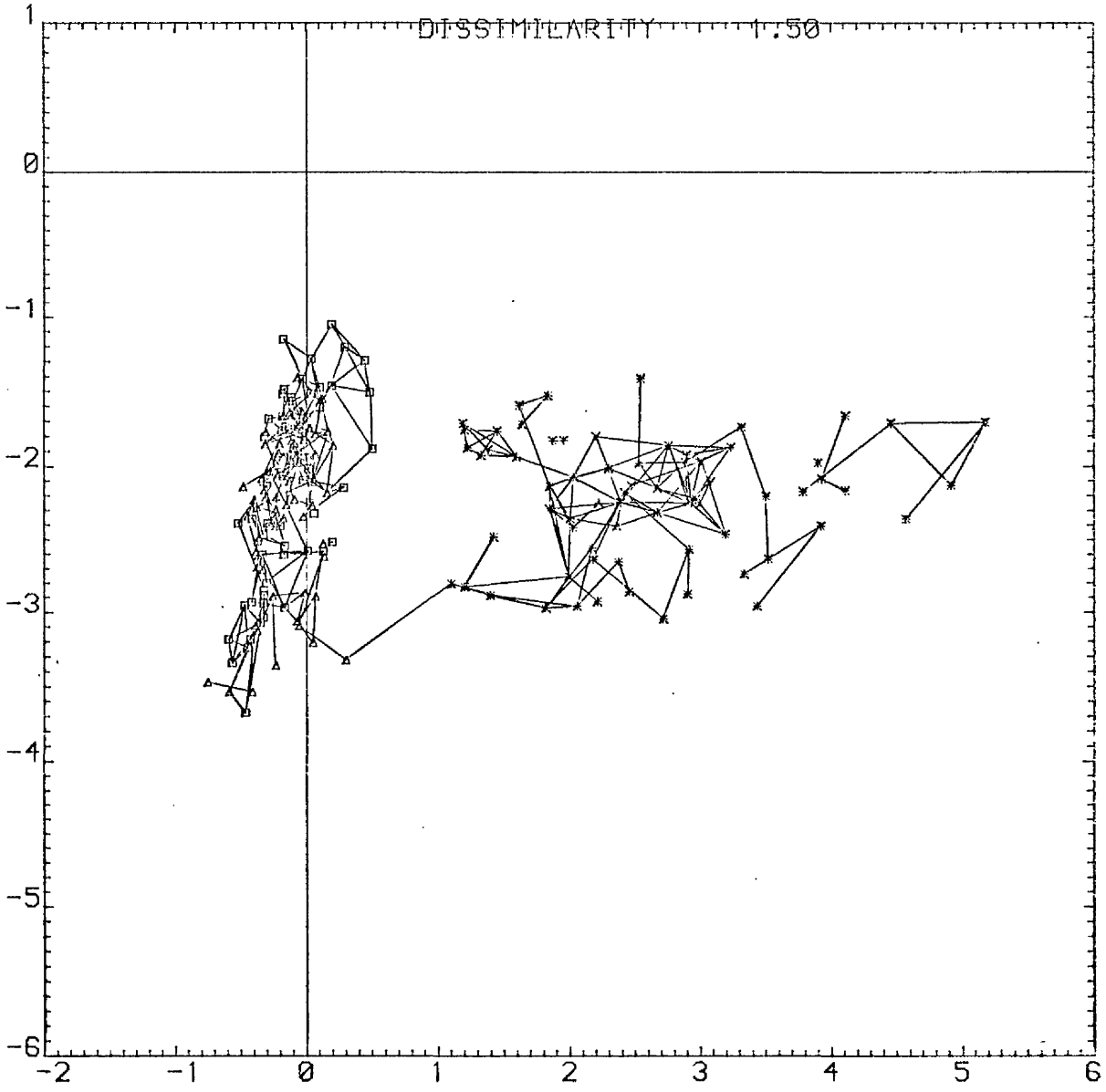


FIG 5.16(d)

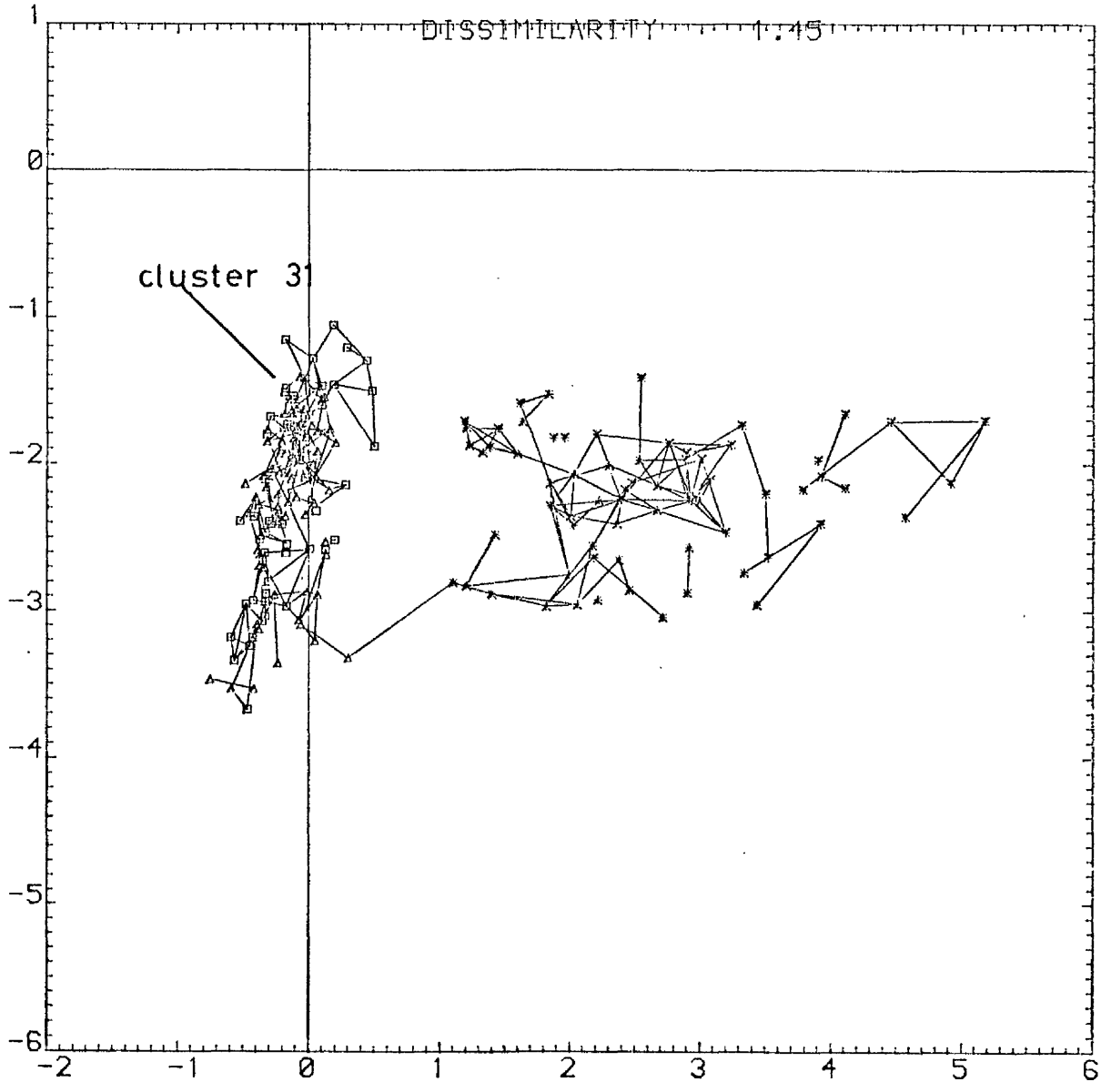
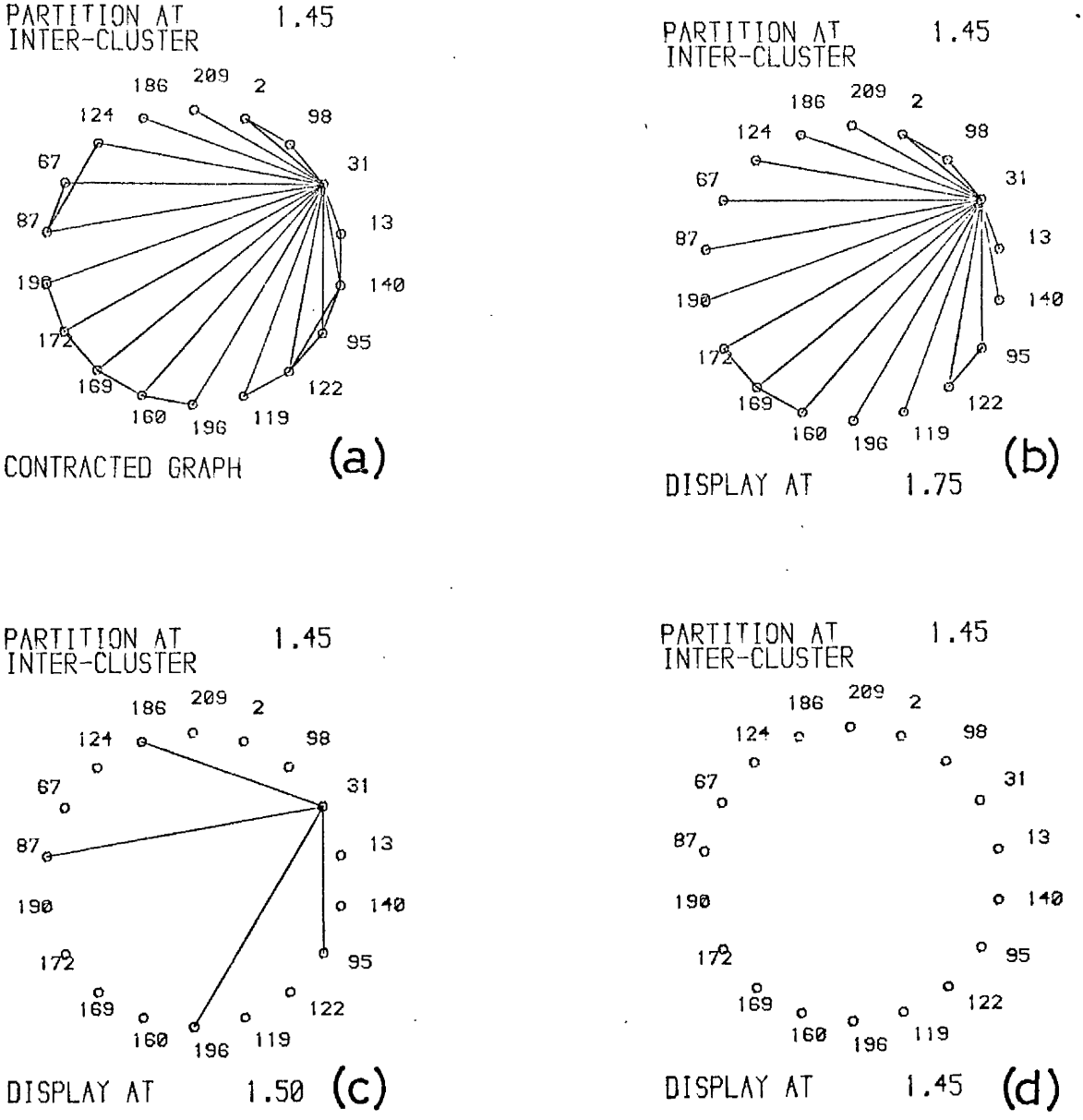


FIG 5.17



Inter-cluster display corresponding to FIG 5.16

FIG 5.18(a)

S_2 clustering of the 6-D lung sound data

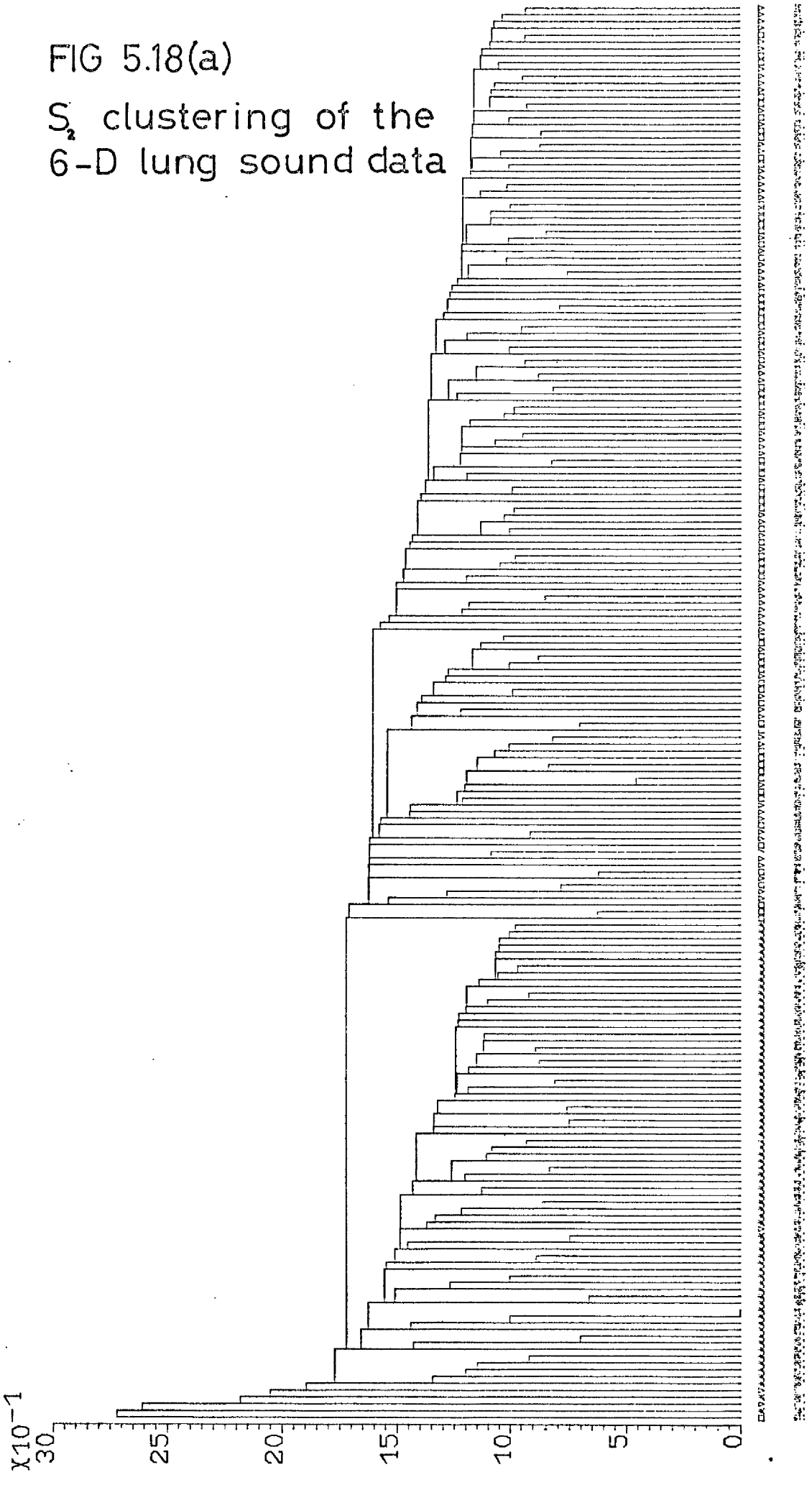


FIG 5.18(b)

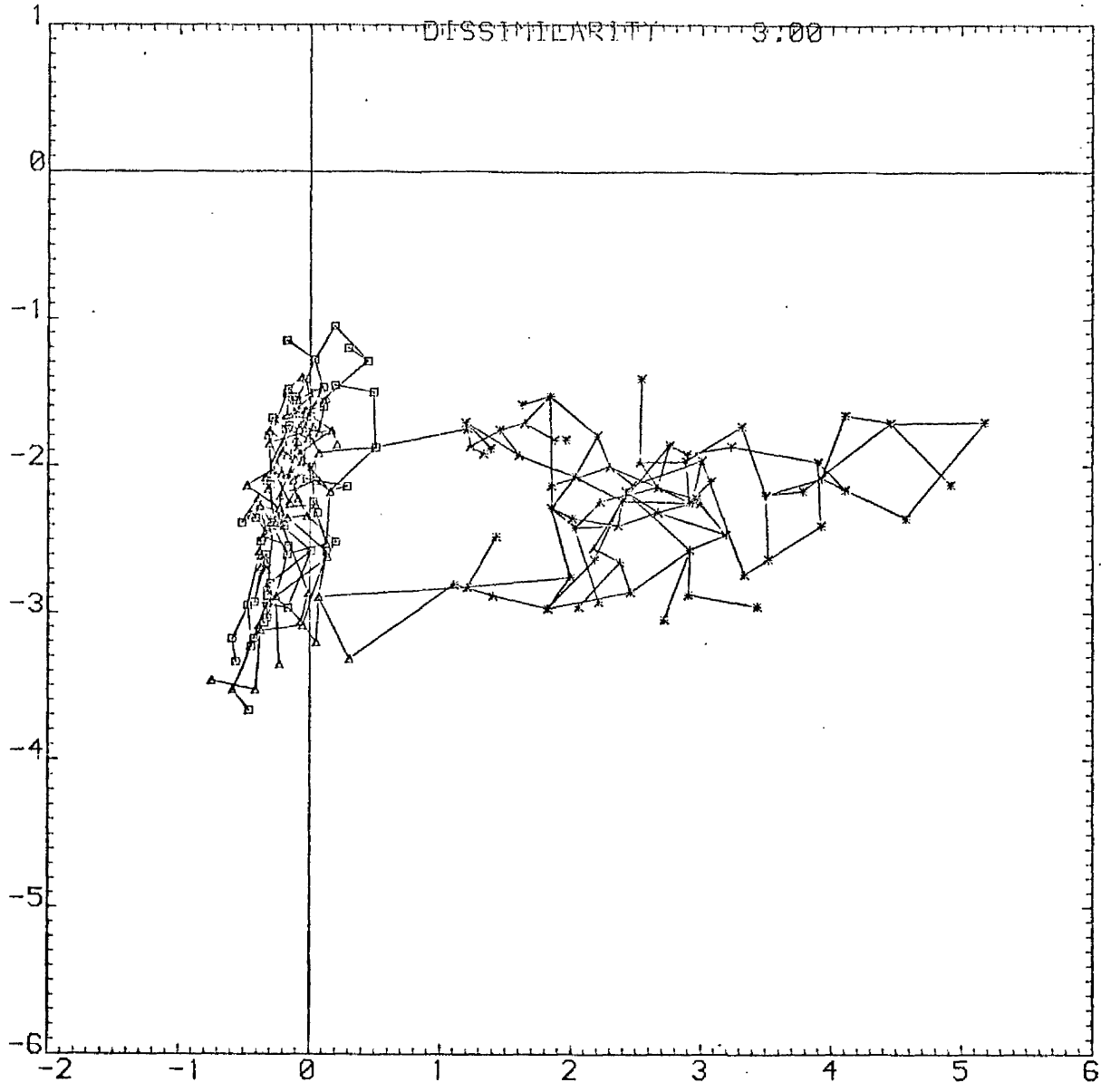


FIG 5.18(c)

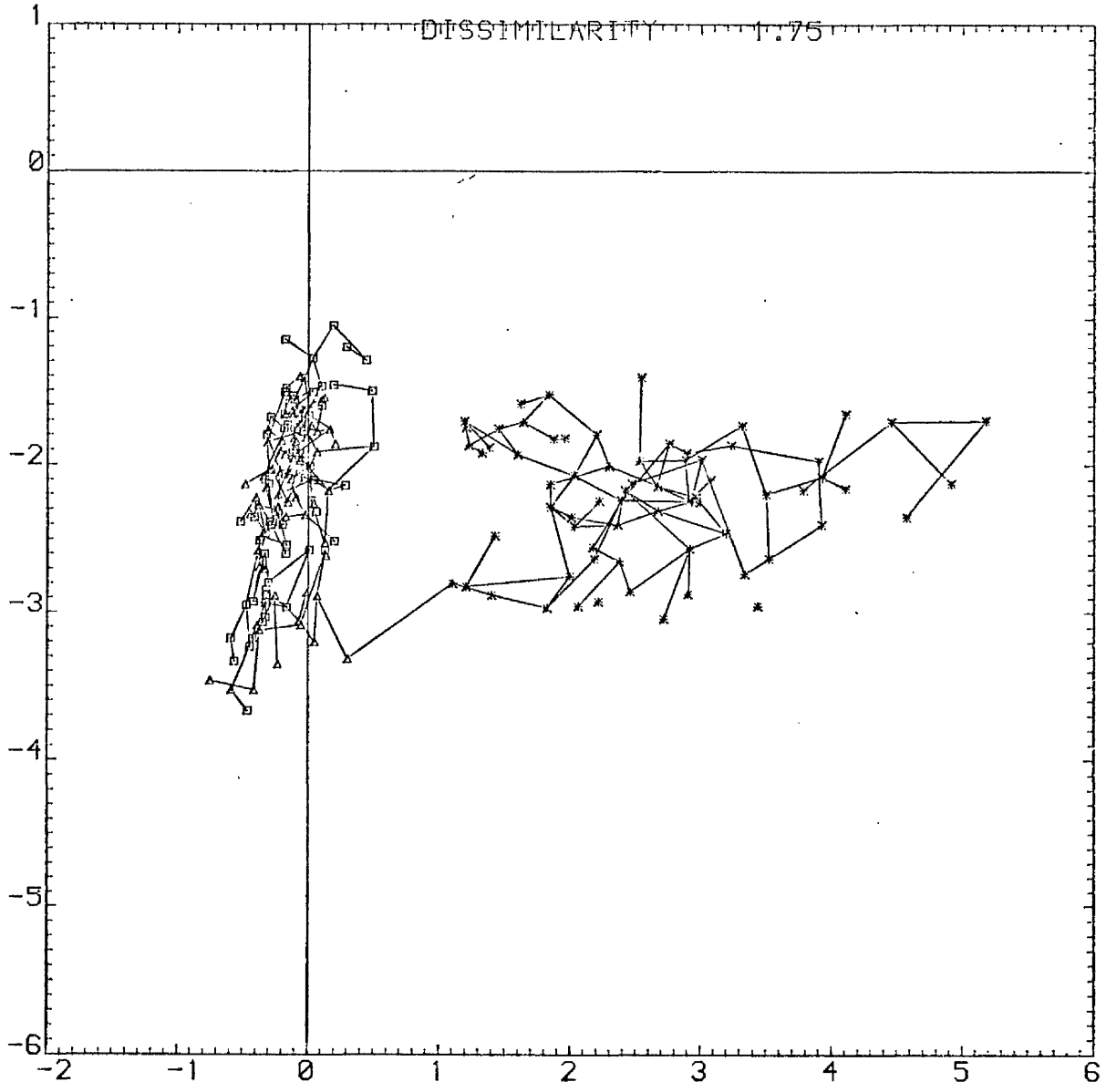


FIG 5.18(d)

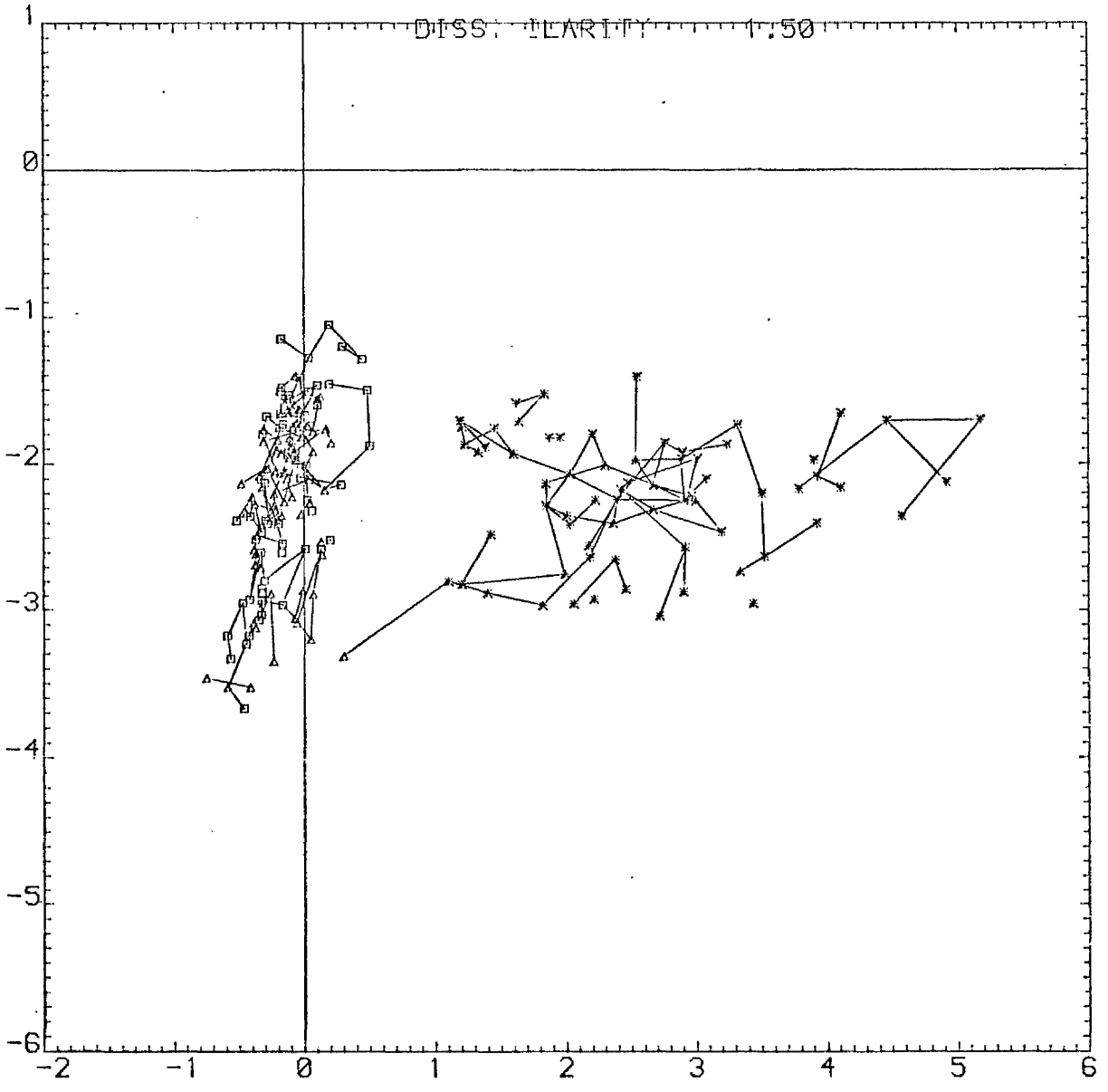
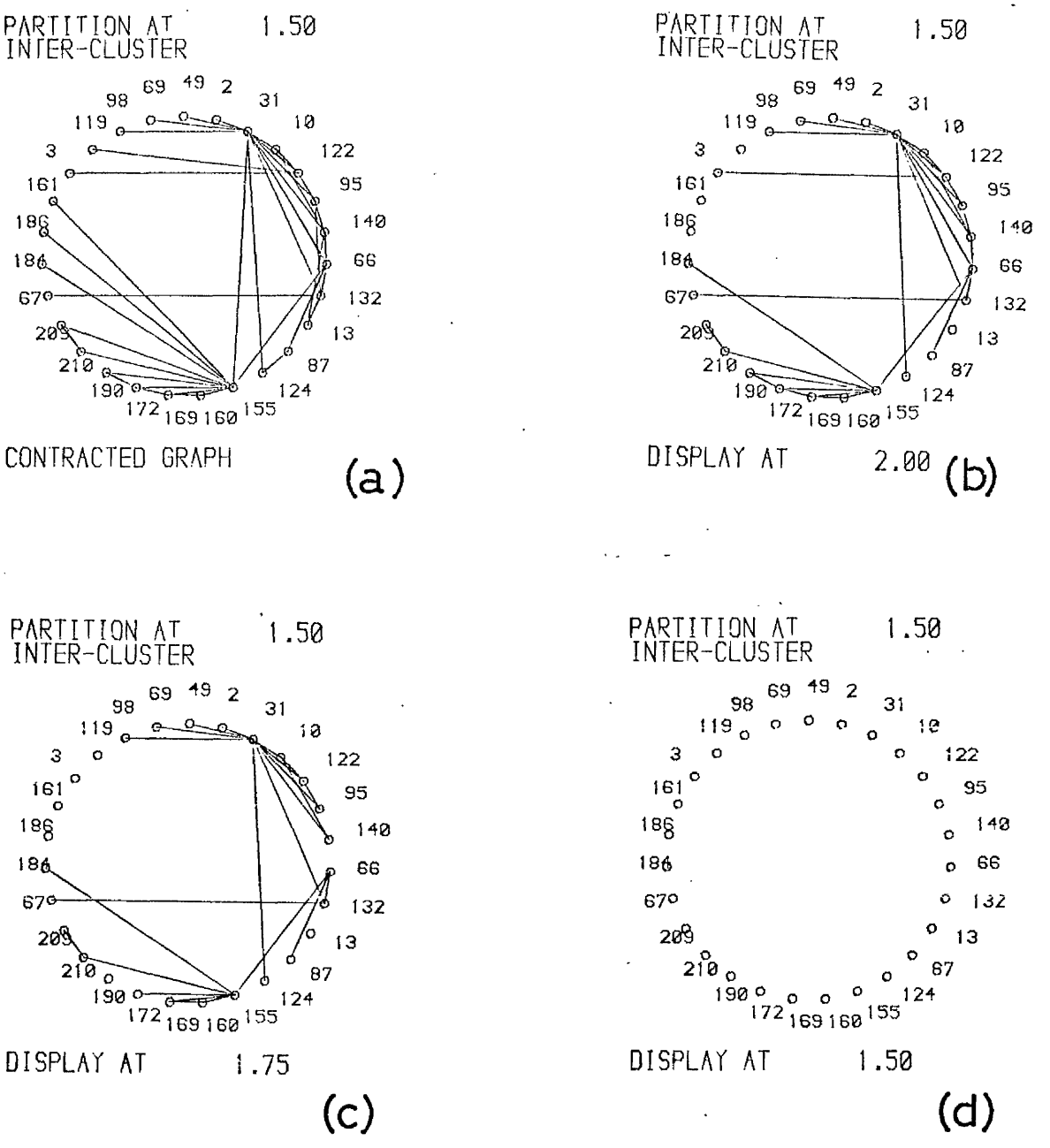
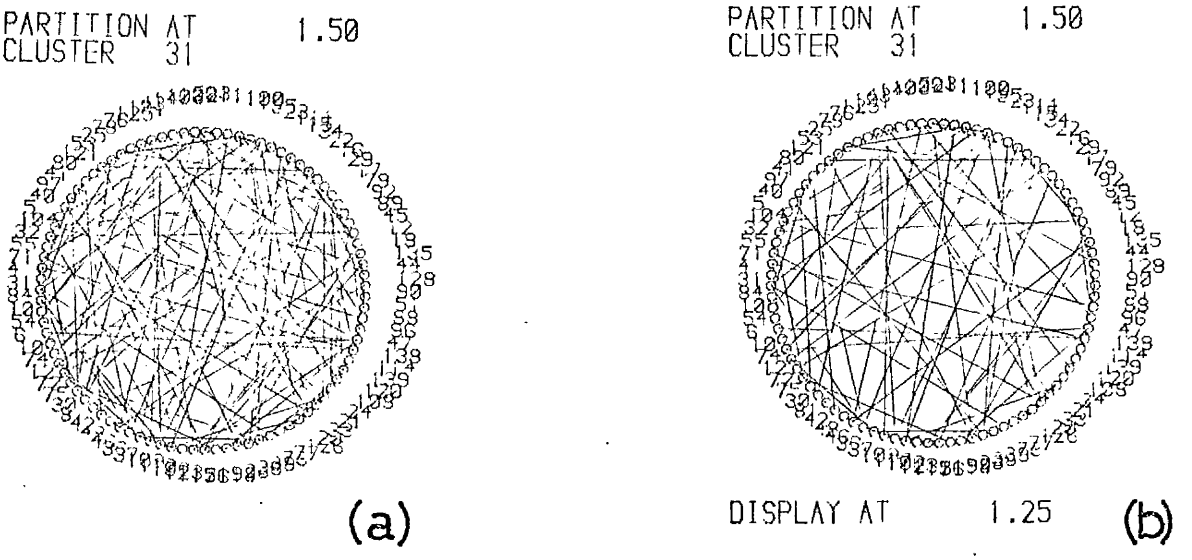


FIG 5.19



Inter-cluster display corresponding to FIG 5.18

FIG 5.20



Intra-cluster display corresponding to Fig 5.18

ACQUISITION & ANALYSIS OF LUNG SOUND SIGNALS

6.1 Introduction

Analysis and interpretation of lung sound requires several stages. The first is the acquisition of the signal by means of a suitable transducer. Then various conditioning and analytical techniques may be used to analyse the signals. This chapter briefly discusses the techniques used during lung sound studies and describes the basic properties of the sounds studied.

Throughout the theoretical sections of the thesis it has been mentioned that the appropriateness of analysis techniques is data dependent. This is certainly true of lung sound and an attempt is made to discuss the relevance of the different approaches that are used.

It is unfortunate that relatively little work has been done on the design and testing of transducers for lung sound. Frequently work has been published giving details of parameters derived from lung sound with no regard for the limitations of the transducers. Similarly there has been relatively little systematic work done on techniques for lung sound signal analysis. The availability of the FFT algorithm has enabled a large number of workers to use spectral analysis, although apparently paying little regard to the assumptions inherent in the technique.

It was decided from the outset to aim for clinical rather than a physiological investigations of lung sound. Recordings were therefore made in relatively normal hospital conditions. It was felt that if any of the techniques that were developed were to be useful, it would be necessary for them to be used at the bedside rather than in a special

soundproofed environment. Therefore a major consideration has been reduction of ambient noise during recording. Also it is felt that the recording/analysis system must tolerate some noise.

6.2 Transducers

In earlier times a physician performed auscultation by placing his ear against the patient's chest. This was unsatisfactory since it was not always advisable to be so close to a patient's body. Laënnec made a considerable advance by inventing the stethoscope. This removed the need for physical contact between the doctor and patient and doubtless improved the reception. Progress has been made since then and the binaural stethoscope has become a familiar feature of today's doctor. Ertel et al [6.1-6.3] carried out a number of studies on stethoscope acoustics.

Design of a suitable lung sound transducer will depend on the particular site chosen for recording. We might divide the sites into three contrasting areas - (1) the mouth, (2) the trachea and (3) the chest wall. At the mouth the lung sound transmission is simply through air in the mouth. Tracheal sound is transmitted through the cartilage of the trachea and a thin layer of tissue below the skin. At the chest wall the sound will be transmitted through the lung tissue, the pleural membrane, the intercostal muscles and the ribs before reaching the surface. The acoustic impedances at different sites will of course influence the ideal design for a transducer. A transducer for picking up sound at the mouth will be designed for detecting sound in air whereas one designed to pick up sound from the chest wall must be designed to match the impedance of the chest wall.

Regrettably the fact that the chest has a much larger impedance than air has rarely been taken into account. A notable exception is the work of Burch & Stock [6.4] who designed an improved stethoscope for listening to heart sounds. In view of the complex structure of the chest wall it is difficult to estimate the impedance and produce a design that couples well. Also the problem of designing suitable tests for lung sound transducers is largely unresolved with many authors quoting results of testing the microphone in air. Gavriely et al [6.5] tested their transducer using a vibrating rubber membrane which took account of skin vibration but not of the transmission of sound through the chest.

The transducer used for the experiments described here was designed by McGhee [6.6] and based on earlier work by Guard [6.7]. It consists of a General Radio ¼-inch electret microphone (Type 1962-9602) with a matching preamplifier, both of which fit into a tubular aluminium enclosure and is linked via a fixed gain amplifier to an FM tape recorder. The microphone and preamplifier have a -2dB bandwidth of 5Hz-20kHz (manufacturer's data) although this response will be modified by the enclosure.

The enclosure is basically a thick aluminium tube surrounding the microphone and preamplifier. This is to act as a mechanical mass element which attentuates ambient sound reaching the microphone. The enclosure is fitted with a rigid Tufnol diaphragm 0.64mm thick which is designed to crudely match the microphone to the chest. This improves the ability of the microphone to detect sound from the chest and simultaeneously reduces the ambient sound recorded.

A certain amount of work has been done on testing the microphone and enclosure system by Howie [6.8] but results are still of a

preliminary nature. The aim of his experiments was to investigate whether the microphone system was effective at picking up sound from within the body. He studied sound transmitted through air and through gelatin (which was used to simulate the mechanical properties of human flesh), and showed that the microphone response in air is quite different from that obtained when sound is passed through gelatin (Fig 6.1). The reasonably flat response obtained through gelatin, over the frequency range of interest, encourages the belief that the transducer design is effective at coupling into the chest, however it is realized that the transducer design is still far from ideal. Once useable signals were being recorded this research concentrated on signal analysis rather than acquisition

Initially two FM tape recorders were used (an Ampex FR1300 and a Racal Store-4D). The recordings were carried out at a tape speed of $60in.s^{-1}$ ($1524mm.s^{-1}$), ensuring that all frequencies from 0 to 20kHz could be recorded. Tests were carried out to determine any differences between recording levels in the two recorders and compensation was carried out in software. Following the preliminary study, recordings were made exclusively on the Racal Store-4D at $30in.s^{-1}$, which allowed the recording of frequencies between 0 and 10kHz.

All recordings made were of sounds from the postero-basal segments of the lower lobes, since there may be some variation in amplitude and spectra when sounds are recorded at widely separated sites over the chest. The subjects were seated and the microphone assembly was hand-held against the chest wall over the ninth or tenth intercostal space posterially. Auscultation through a stethoscope was performed immediately before recording, and the sound was monitored by headphones during recording.

At first the recorded signals were digitized and stored directly by a PDP11/45 computer. Due to software overheads the tape speed was reduced by a factor of 32. For the analysis of the asbestosis recordings software was developed by a technician to log data using an SBC-100 microcomputer. This permits logging in real time and allows much larger sampled data files. Data files are then transferred to the PDP11/45 for analysis. A schematic of the recording and analysis system is given in Fig 6.2.

Pneumotachygraphs are frequently used to monitor the transitions between inspiration and expiration during lung sound recordings. In early recordings the transitions were identified aurally but in all the later ones a thermistor probe was used. A bead thermistor was mounted on a probe that was attached to a headset worn by the patient. The change in temperature associated with direction of breathing gave a signal which was differentiated, amplified and recorded as a second channel. This provided a useful alternative to the pneumotachygraph which is cumbersome and may slightly affect the recording (see Section 7.3).

6.3 Lung Sound Signal Analysis

A survey of lung sound research was given in Chapter 2 and one on signal processing in Chapter 3. This section is specifically on the use of signal processing techniques in lung sound analysis. A subsection on adventitious sounds is included but it is not intended to be comprehensive.

6.3.1 Breath Sound

Breath sound is a continuous signal but has no harmonic structure. It is a random signal in contrast to the adventitious sounds which have regular waveforms. Since the signal is continuous it might be expected that frequency domain techniques would prove useful. The cyclic nature of breathing suggests that the envelope of the breath sound is of interest. Weiss & Carlson [6.9] show some interesting traces but apparently there is no quantitative work using this method.

Probably the earliest work on spectral analysis of breath sound was that of Cabot & Dodge in 1925 [6.10]. With the invention of the sound spectrograph, McKusick et al [6.11] investigated the spectral properties of breath sound. Subsequently Banaszak et al [6.12] have used filter banks to study breath sound at different flow rates.

The next technological influence was the availability of FFT algorithms which offered an efficient route to spectal analysis. Unfortunately the published work in this area shows little regard for the different ways in which the technique may be used.

Mori et al [6.13] showed spectra from a number of different diseases. However the spectra shown are of low resolution and on a very limited number of patients. In probably the most thorough investigation of the spectral properties of breath sound, Gavriely et al [6.5] used smoothed averaged spectra. While this is reliable statistically, any time variation in the spectra will become obscured by averaging. An unfortunate feature of their paper is that they termed ampitude spectra 'power spectra' which is confusing. From the averaged spectra they computed the slope of the log of the curve and a 'maximal frequency' which is the maximum frequency at which lung sound

is observable. However their parameter of maximal frequency will probably depend on the dynamic range of A/D converters. Nevertheless their results are important in providing a thorough analysis over a variety of chest locations.

In Gavriely's paper it was suggested that spectral properties of breath sound would be expected to change with disease. Confirmation of this came independently with the work of Urquhart et al [6.14] which is described and dicussed in Chapter 7. Since then Chowdhury & Majumder [6.15] have found differences in spectra in patients with tuberculosis. Unfortunately the different spectra described here are not easily compared because of differing recording equipment and estimation methods.

6.3.2 Adventitious Sounds

In recent years crackles have been studied by a relatively large number of groups in the U.S.A., Japan and Europe. Despite the intensive effort the results obtained so far, using both time and frequency domain methods, have been rather disappointing.

In the time domain Murphy et al [6.15] have suggested time expanded waveform analysis for investigating crackles, which simply investigates crackles by plotting them on a large time scale. This approach yielded some interesting results and has been used by a number of workers (e.g.[6.16-6.18]). Mori et al [6.18] examined the time domain properties of crackles in tuberculosis using a method based on zero-crossing.

Murphy & Sorenson [6.19] also investigated the spectral properties of crackles using the FFT algorithm and again this idea has

been adopted by other workers [6.18,6.20-6.22]. However few of the workers have noted that the DFT assumes periodicity in the time domain. Thus the spectral results are really based on a train of equally spaced crackles rather than on individual crackles. Perhaps better results could be obtained using the chirp z-transform (CZT) algorithm which would take acount of decays in the signal.

Forgacs [6.23] discusses the clinical significance of wheezing at different frequencies. However the only reports of digital spectral analysis are those of Baughman & Loudon [6.24] and Maeda et al [6.25] who used the DFT to study the variation of the frequency content of wheezes during the course of breathing.

6.4 Techniques for Breath Sound Analysis

At this stage it is worth stressing that the spectra used in the initially were amplitude spectra, as in Gavriely et al [6.5]. However for later work it was decided to use power spectra since it would allow a direct comparison between DFT-based and other estimation techniques.

6.4.1 Spectral Analysis

Since breath sound is basically a continuous signal, a frequency domain approach was adopted from the outset. This was initially carried out by McGhee [6.6] and continued in the work described here. A number of conflicting requirements influence the approach to studying the signals.

Firstly the lung sound signal is cyclic in nature. Different

mechanisms are involved in inspiration and expiration and so these phases must be segmented and considered separately. Secondly, and although this has not yet been shown in a systematic study, it is to be expected that the signal properties may change during the course of an inspiration or expiration. Initially therefore it was decided to concentrate on the latter portion of the inspiration partly because of the fact that in some diseases crackles appear during this part of the cycle.

It was also felt desirable to use a DFT-based approach to spectral analysis for efficiency in computation. The computation of the DFT using the FFT algorithm immediately leads to a number of problems. The readily available FFT algorithms require the number of input points to be a power of 2. This means that with the duration of the inspiration varying from individual to individual, the differences had to be accounted for by either fixing the length of the time series used or by adjusting the sampling rate so that the the number of input points corresponded to the number of samples required to digitize the late inspiratory segment. For the preliminary study McGhee [6.6] selected the latter approach.

The selection of this approach causes some immediate problems. Firstly a variation in the time domain sampling rate causes a corresponding variation in the frequency resolution of the DFT. This must then be taken into account in any further processing especially when comparing spectra.

The use of the DFT naturally leads to a periodogram estimate of the power spectrum. As discussed previously (Chapter 3) the raw periodogram estimate has a very high variance and so it is desirable to reduce the variance by either smoothing individual periodograms or

averaging periodograms from different time segments.

In view of the fact that little is known about the breath to breath variation of the signals it was felt desirable to smooth rather than average periodograms. The use of smoothing is also useful in patients who are difficult to record, since it means that the recording of only one uncontaminated breath cycle is necessary for analysis. However in a preliminary study the variation in sampling rate prevented the use of standard periodogram smoothing methods. Instead a rather ad hoc method was used to simultaneously smooth the spectrum and extract features. This method is described in detail in the next subsection.

Despite the encouraging result obtained in the preliminary study (Chapter 7) it was felt unwise to continue with such an ad hoc spectral estimate in any systematic study of lung sound. In view of this, software was developed to allow a more rigorous approach to spectral estimation. Firstly it was decided to have a constant sampling rate which would then allow conventional smoothing techniques to be used, although it hoped that the differing lengths of breath cycle will not adversely affect the results. Secondly it was decided to investigate the maximum entropy method (MEM) of spectral analysis which allows statistically reliable spectral estimates to be obtained using time series of different lengths.

The use of the various smoothing techniques is illustrated in Figs 6.3-6.5. Fig 6.3 shows the smoothing of a raw periodogram spectral estimate, the removal of the spurious spikiness allows the spectrum to be more easily understood. In Fig 6.4 the same data is smoothed by (1) a Daniell window [6.23], (2) a cosine window and (c) an autoregressive model. The Daniell window gives a more spiky

estimate than the cosine window. The autoregressive smoothing [6.24], or maximum entropy spectral analysis [6.25], does not require computation of the DFT and is shown for comparison. The use of the autoregressive method requires the selection of a suitable model order. Fig 6.5 shows the effect of varying the model order with lower order models giving much smoother estimates at the expense of losing detail.

Since a good deal of work on selection of model order was required before maximum entropy method spectra could be used routinely, the later work on spectral analysis (e.g. Section 7.3) used cosine window smoothing of the periodogram. This cosine window is essentially the same as the sine window used by Gavriely et al [6.5].

6.4.2 Normalization and Feature Extraction

In order to carry out exploratory data analysis it is crucial to provide a suitable data reduction stage. In the preliminary study it was considered of great importance to compare a given frequency interval in one spectrum with the corresponding range in another. Visual inspection suggested that spectral shape at lower frequencies was significant. The frequency range 0-400Hz was divided into 20 unequal frequency intervals (Table 1), the closer spacing at the low frequency end being intended to take account of the relative significance of these frequencies. Furthermore it was felt that that the shape of the spectral envelope was significant.

The time interval digitized was always the latter half of the inspiration. As indicated earlier the sampling rate was allowed to vary to take account of differing lengths of breath cycle. The time

interval was always in the 0.6-1.4s range but with the anti-aliasing precautions being based on the lowest sampling rate used (730 Hz). The logging was always 2048 points for use with a 2048-point FFT algorithm.

A heuristic piecewise constant approximation was devised to extract data from the amplitude spectrum over the chosen frequency intervals. The approximation took place in two stages: firstly peak detection over a given interval, and secondly averaging of peaks within each interval. For a particular spectral amplitude a_n at frequency f_n to be recognised as a peak:

(i) a_n must satisfy the conditions $a_n > a_{n-1}$ and $a_n > a_{n+1}$

(ii) f_n must lie outside the range over which mains frequency (at 50Hz) and its 3rd and 5th harmonics were likely to vary and

(iii) f_n must lie no closer to another selected peak thatn 4Hz in order to take account of the differences in spectral resolution. The value 4Hz was derived from the lowest resolution spectra in which the minimum peak width $f_{n+1}-f_{n-1}$ was 3.54Hz. This value was rounded up to 4Hz.

The amplitude value corresponding to each frequency interval was then taken to be the average of the selected peaks in the interval (Fig 6.6). Unnormalized results obtained directly from this feature extractor proved to be disappointing.

Since the magnitude of the spectra is related to signal magnitude in the time domain, it is influenced by extraneous factors such as chest wall thickness. It was therefore considered necessary to normalize the spectral values, which is also consistent with the observation that the spectral shape rather than the actual magnitude is significant in biological signals [6.26].

The values in a particular interval were normalized by dividing the area under the approximation within the interval by the total area under the approximation. The resulting values are then dimensionless and hence are in arbitrary units. This normalization proved crucial in obtaining the results of Chapter 7.

The software developed for analysing the asbestosis recordings aimed to avoid the complications of the above method. In retrospect it was felt that a suitably smoothed spectra would yield a better indication of spectral shape than the envelope of the spectral peaks (each of which would be subject to high variance). While no feature extractor has been tested a suitable normalization for the spectral estimate is proposed below.

It is suggested that in order to extract data from a smoothed spectral estimate we should either divide the frequency intervals into a set similar to Table 1 or select a set of frequency values. Data could then be extracted from the power spectrum by averaging the smoothed periodogram over each interval (giving a rather better piecewise constant approximation) or by taking ordinates corresponding to the chosen frequency values.

The values would then be normalized by

(1) Power = $\Sigma \hat{P}_m \Delta f$ for the piecewise constant approximation

(2) $\Sigma \hat{P}_m$ for ordinates

where $\hat{P}_m$ is the power spectral estimate at frequency f_m and Δf is the interval between two frequencies in the DFT i.e. $\Delta f = f_{m+1} - f_m$

6.4.3 Influence of Interference on Spectra

Before exploring lung sounds using spectral analysis techniques it is useful to establish the extent of interference from various sources. If the interference is likely to be constant, and the objectives of the experiment are purely comparative, then the interference constitutes a systematic error that does not affect any comparisons made.

Three sources of interference must be considered:- (a) ambient sound, (b) internal biological sound and (c) handholding. The ambient sound is reduced as far as possible by the transducer design. Internal biological sounds include the heart, the muscles and the gut. Bad contact with the microphone is easily noted during recording (and hence data rejected) providing headphones are used, but there could additionally be some low frequency interference due to unsteadiness in handholding.

The combined effects of ambient sound and handholding were investigated by handholding the transducer against a block of gelatin under normal ambient conditions. This type of interference appears to be small compared with late inspiratory breath sound (Fig 6.7).

Yoganathan et al [6.27,6.28] have studied heart sounds and have observed significant components in their signals below 100 Hz. However their results are not directly relevant to a consideration of interference to lung sound recordings because of their analysis method. Since heart sounds are transient, they remove the segments of the signal between beats and thus allow the signals to be analysed as if they were continuous. Therefore their spectra do not represent interference possible in lung sound recordings. Also there was no

evidence of the characteristic heart sound waveforms appearing in recorded data indicating that the precaution of recording from the right lung is adequate.

6.5 Discussion

The results above suggest that although the matiching of the microphone to the chest was not ideal, the means of recording lung sound are sufficiently good for experimental work. The levels of interference are low enough to permit useful extraction of data from the lung sound. In any case any systematic errors would be relatively unimportant in comparative studies between disease groups.

The problem of analysing a segment of signal from varying lengths of breath was tackled in two different ways. The earlier ad hoc techniques did produce useful results (Chapter 7) yet would not be advisable for more than an exploratory study. A more thorough study of lung sound requires a more rigorous approach to spectral estimation, based either on smoothing the periodogram or on MEM. For MEM to be used as a tool a thorough investigation of appropriate model order, e.g. using the Akaike information criterion [6.29], would need to be undertaken using a wide variety of recordings. Additionally an investigation of the stationarity of the signal would give useful information on the best way to approach spectral estimation.

The feature extraction problem was solved by comparing the spectra over a number of frequency intervals. However for this information to be meaningful the spectra had to be normalized to emphasize spectral shape rather than signal intensity. It is likely that a similar feature extraction method with normalization could be

employed on future DFT-based spectral estimation methods. Alternatively features might be obtained from coefficients obtained during maximum entropy spectral analysis.

References

6.1 P.Y.Ertel, M.Lawrence, R.K.Brown & A.M.Stern, Stethoscope acoustics: I the doctor and his stethoscope, Circulation, 34, 889-898 (1966)

6.2 P.Y.Ertel, M.Lawrence, R.K.Brown & A.M.Stern, Stethoscope acoustics: II transmission and filtration patterns, Circulation, 34, 899-909 (1966)

6.3 P.Y.Ertel, M.Lawrence & W.Song, How to test stethoscopes, Medical Research Engineering, 8, 7 (1969)

6.4 C.R.Burch & J.P.P.Stock, A new diaphragmatic stethoscope, Brit.Heart J., 4, 447-454 (1961)

6.5 N.Gavriely, Y.Palti & G.Alroy, Spectral characteristics of normal breath sounds, J.Appl.Physiol: Respirat. Environ. Exercise Physiol., 50, 307-314 (1981)

6.6 J.McGhee, Unpublished results

6.7 D.R.Guard, The generation of breath sounds and their transmission through the chest wall, Ph.D. Thesis, University of Southapton (1976)

6.8 K.Howie, The frequency response of a microphone for measuring human lung sounds, Final year project report, University of Glasgow (1981)

6.9 E.B.Weiss & C.J.Carlson, Recording of breath sounds, Am. Rev. Respirat. Dis., 105, 835-839 (1972)

6.10 R.C.Cabot & H.F.Dodge, Frequency characteristics of heart and lung sounds, J.Am.med.Ass., 84, 1793-1795 (1925)

6.11 V.A.McKusick, J.T.Jenkins & G.N.Webb, The acoustic basis of chest examination: studies by means of sound spectrography, Am.Rev.Tuberc., 72, 12-34 (1955)

6.12 E.F.Banaszak, R.C.Korry & G.C.Snider, Phonopneumography, Am. Rev.

Respirat. Dis., 107, 449-455 (1973)

6.13 M.Mori, N.Kinoshita, H.Morinari, T.Shiraishi, S.Koike & S.Murao, Spectral analysis of breath sounds, Nippon Kyobu Shikk. Gakk. Zasshi, 16, 503-512 (1978)

6.14 S.K.Chowdhury & A.K.Majumber, Digital spectral analysis of respiratory sound, IEEE Trans. Biomed. Eng., BME-28, 784-788 (1981)

6.15 R.L.H.Murphy, S.K.Holford & W.C.Knowler, Visual lung-sound characterization by time-expanded waveform analysis, N.Eng.J.Med., 296, 968-971 (1977)

6.16 S.K.Holford & R.L.H.Murphy, Differentiation of the rales of pulmonary asbestosis & congestive heart failure, 1st Internat. Conf. Lung Sounds (1976)

6.17 S.Kudoh, A.Shibuya, N.Aisaka, I.Ono, A.Kurashima & R.Mikami, Analysis of rales in patients with fibrosing alveolitis by a new pnonopneumographic method using a sound spectrograph, 2nd Internat. Conf. Lung Sounds (1977)

6.18 M.Mori, K.Kinoshita, H.Morinari, T.Shiraishi, S.Koike & S.Murao, Waveform and spectral analysis of crackles, Thorax, 35, 843-850 (1980)

6.19 R.L.H.Murphy & K.Sorensen, Chest auscultation in the diagnosis of pulmonary asbestosis, J.Occup.Med., 15, 272-276 (1973)

6.20 Y.Fujuara & T.Okayasu, Recording and analysis of the Velcro rale, Trans. 9th Pulmonary Fibrosis Res. Conf. (1974)

6.21 Y.Homma, Y.Minami, Y.Ohsaki & M.Murao, Velcro rale - physiological analysis and simulation of the rale, Clin. Physiol., 7, 157, (1977)

6.22 S.Kudoh, K.Ichikawa, S.Kitamura, K.Kosaka, A.Shibuya, N.Aisaka & I.Ono, Acoustic characteristics of 'crackle' analysised by sound spectrograph - comparison with time-expanded waveform and power

spectrum by FFT, Nippon Kyobu Shikk. Gakk. Zasshi, 16, 711-720 (1978)

6.23 P.J.Daniell, Discussion of a paper by M.S.Bartlett, J.Roy.Statist.Soc., Suppl. 8, 27 (1946)

6.24 J.Makhoul, Linear prediction: a tutorial review, Proc. IEEE, 63, 561-580 (1975)

6.25 J.P.Burg, Maximum entropy spectral analysis, Proc 37th Meeting Exploration Geophysicists, Oklahoma City, U.S.A. (1967)

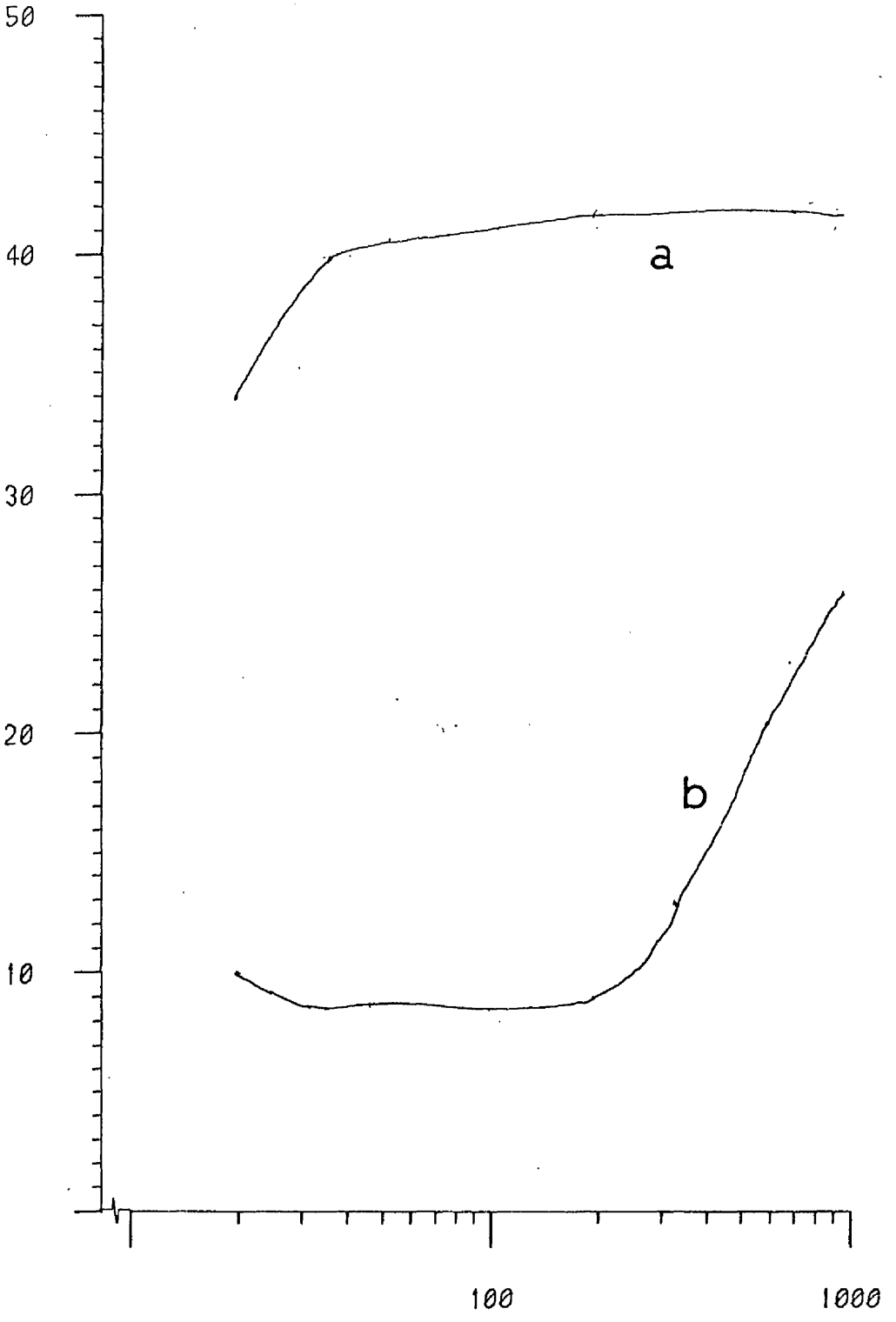
6.26 B.McA.Sayers, Exploring biological signals, Biomedical Engineering, 10, 335-341 (1975)

6.27 A.P.Yoganathan, R.Gupta, F.E.Udwadia, J.W.Miller, W.H.Concoran, R.Sarma, J.L.Johnson, R.J.Bing, Use of the fast Fourier transform for frequency analysis of the first heart sound in normal man, Med. & Biol. Eng., 14, 69-73 (1976)

6.28 A.P.Yoganathan, R.Gupta, F.E.Udwadia, R.Sarma, R.J.Bing, Use of the fast Fourier transform for frequency analysis of the second heart sound in normal man, Med. & Biol. Eng., 14, 455-460 (1976)

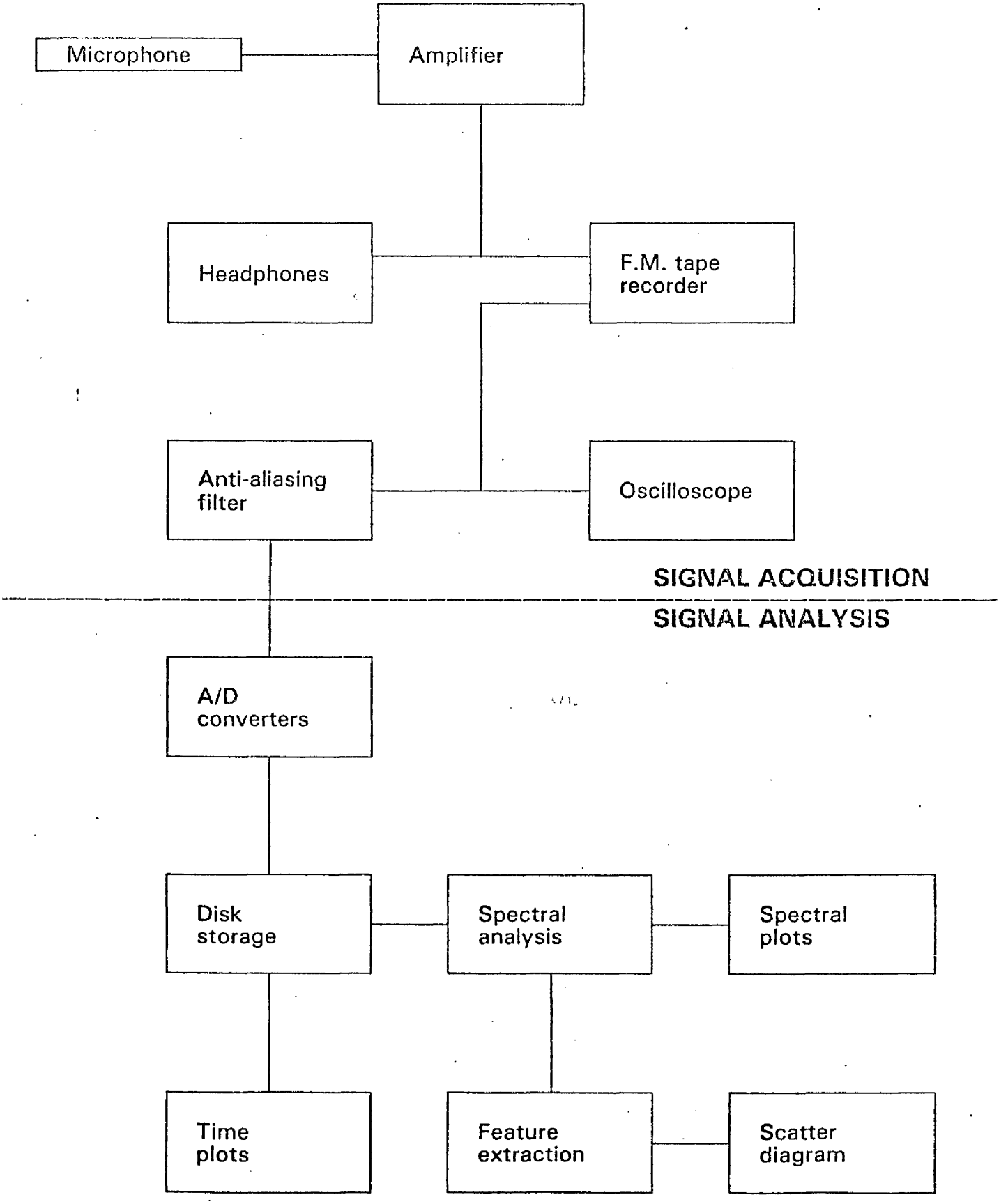
6.29 H.Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control, AC-19, 716-723 (1974)

FIG 6.1



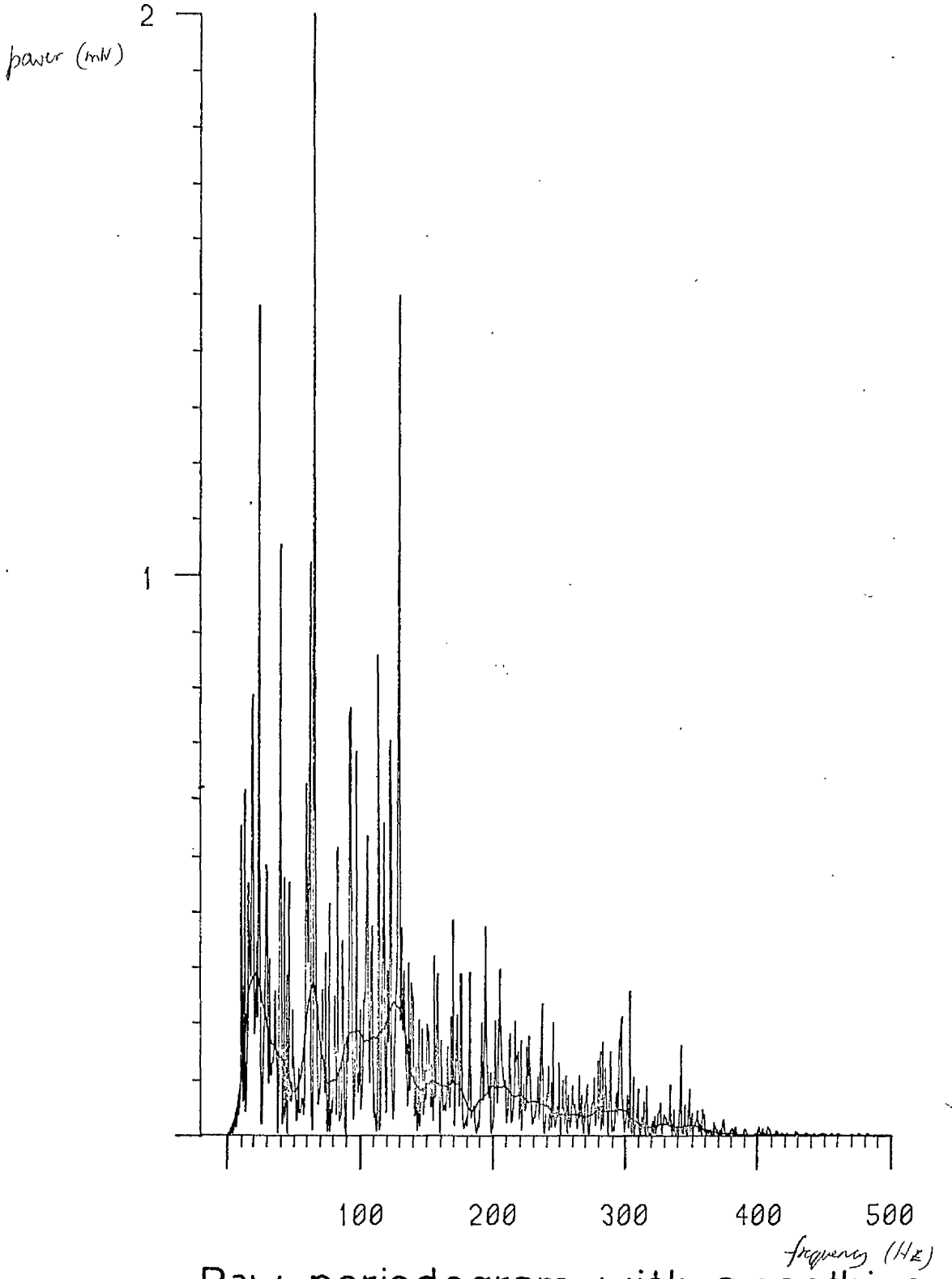
microphone response
through a. gelatin
b. air

FIG 6.2



SCHEMATIC OF ACQUISITION AND ANALYSIS OF LONG SOUND SIGNALS

FIG 6.3



Raw periodogram with smoothing

FIG 6.4

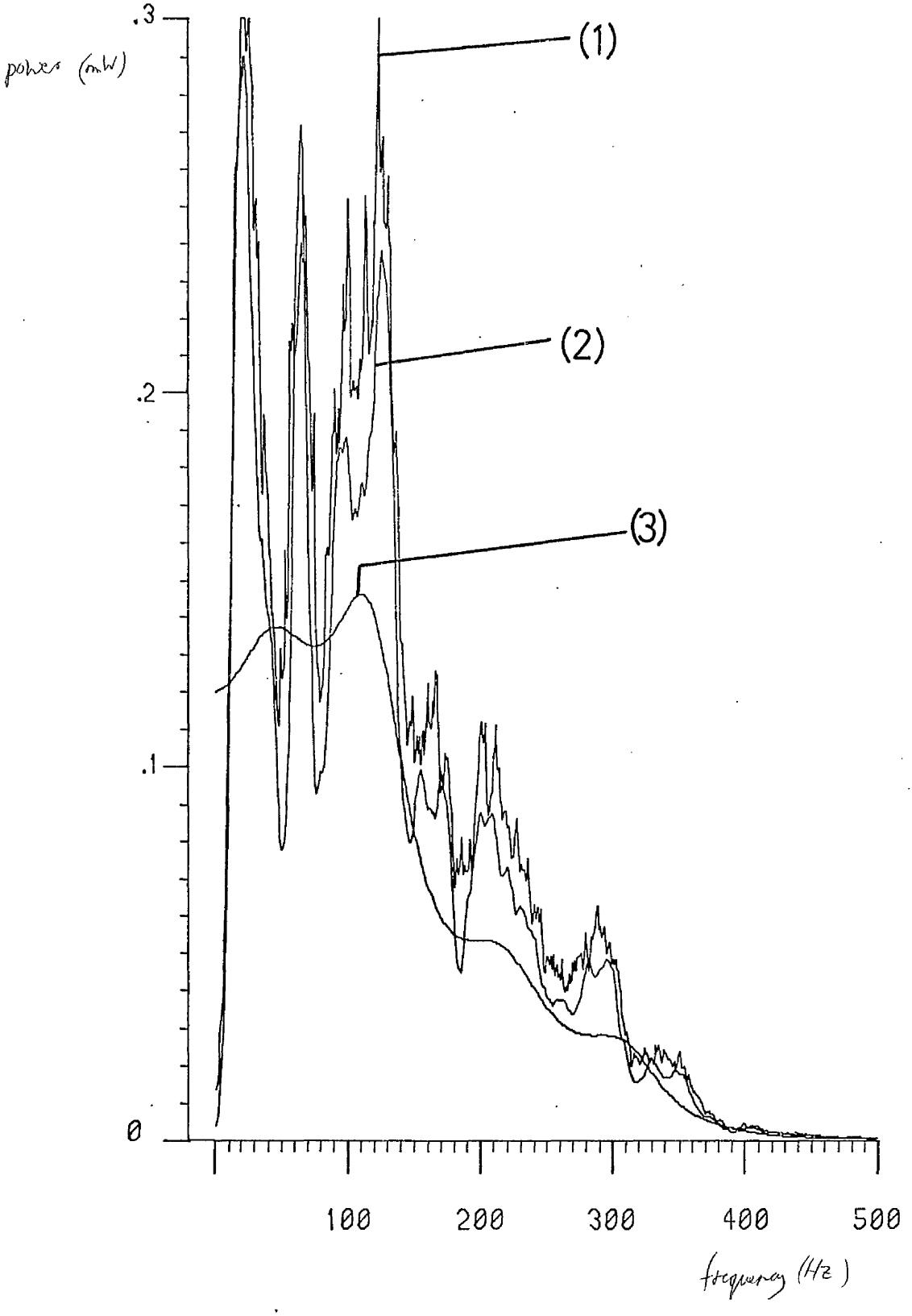
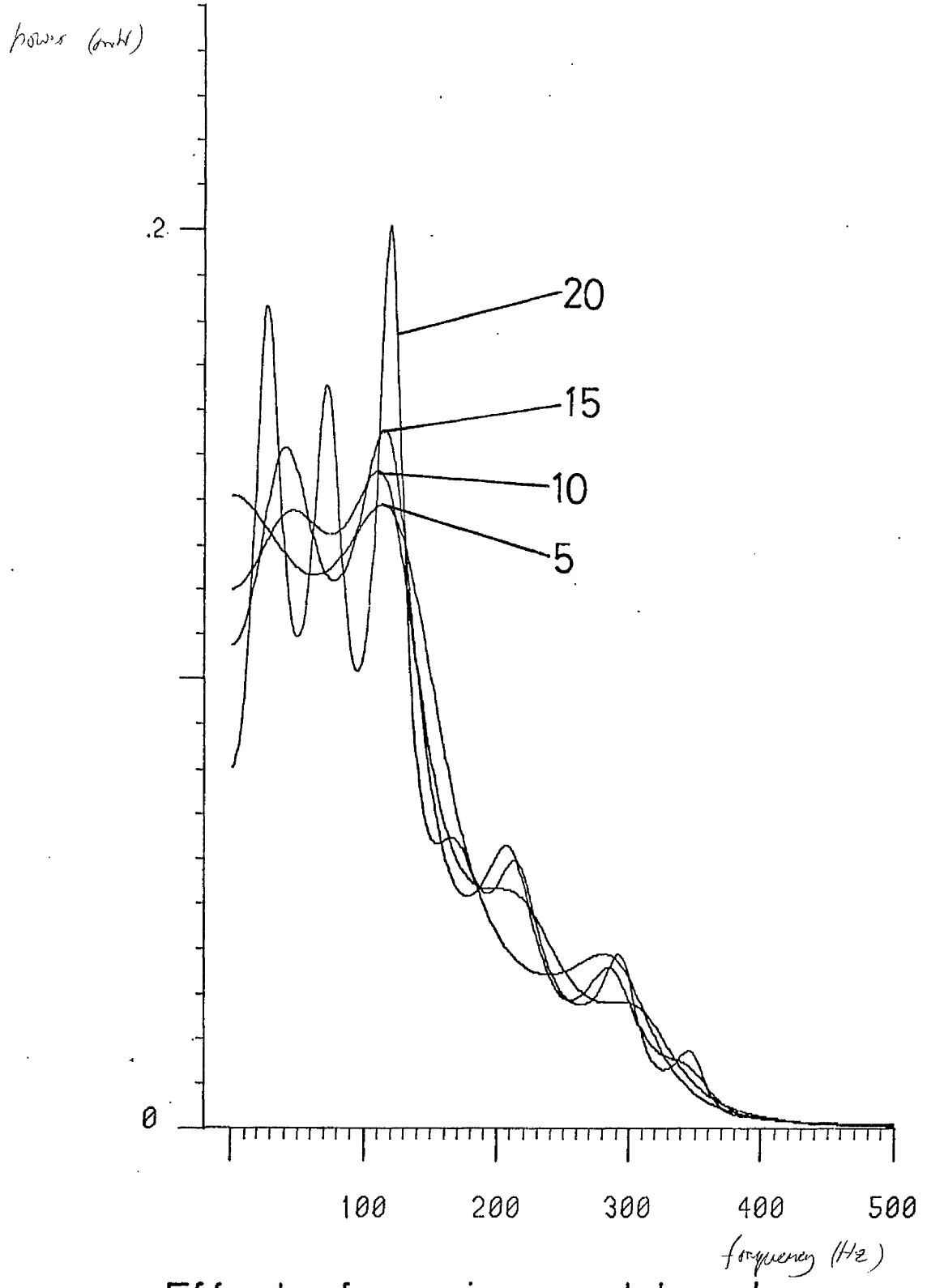


FIG 6.5



Effect of varying model order

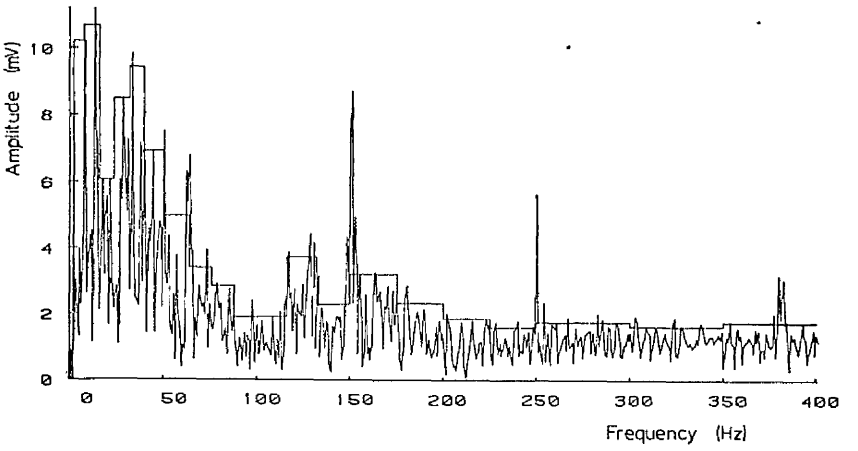


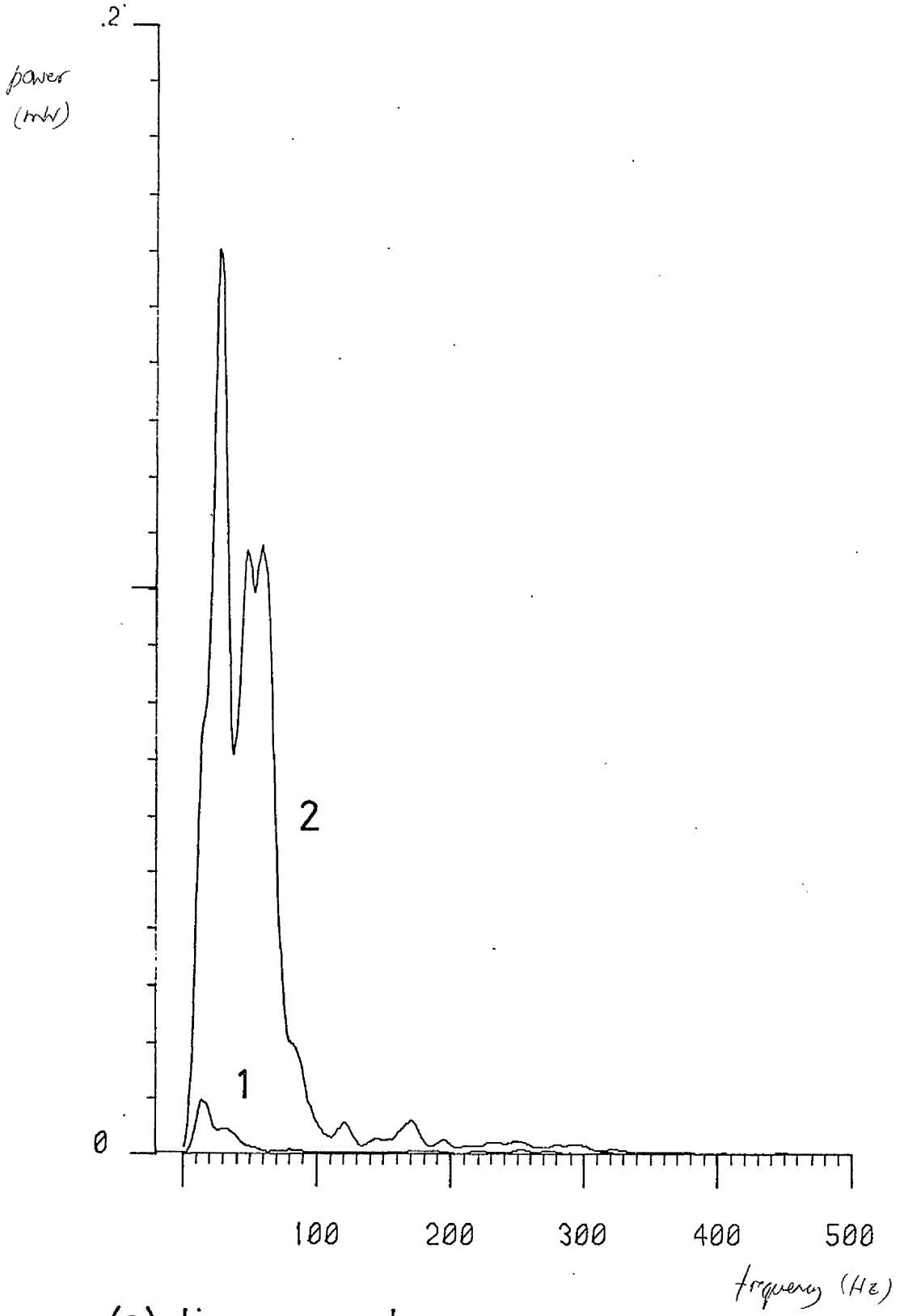
FIG 6.6

Table 1

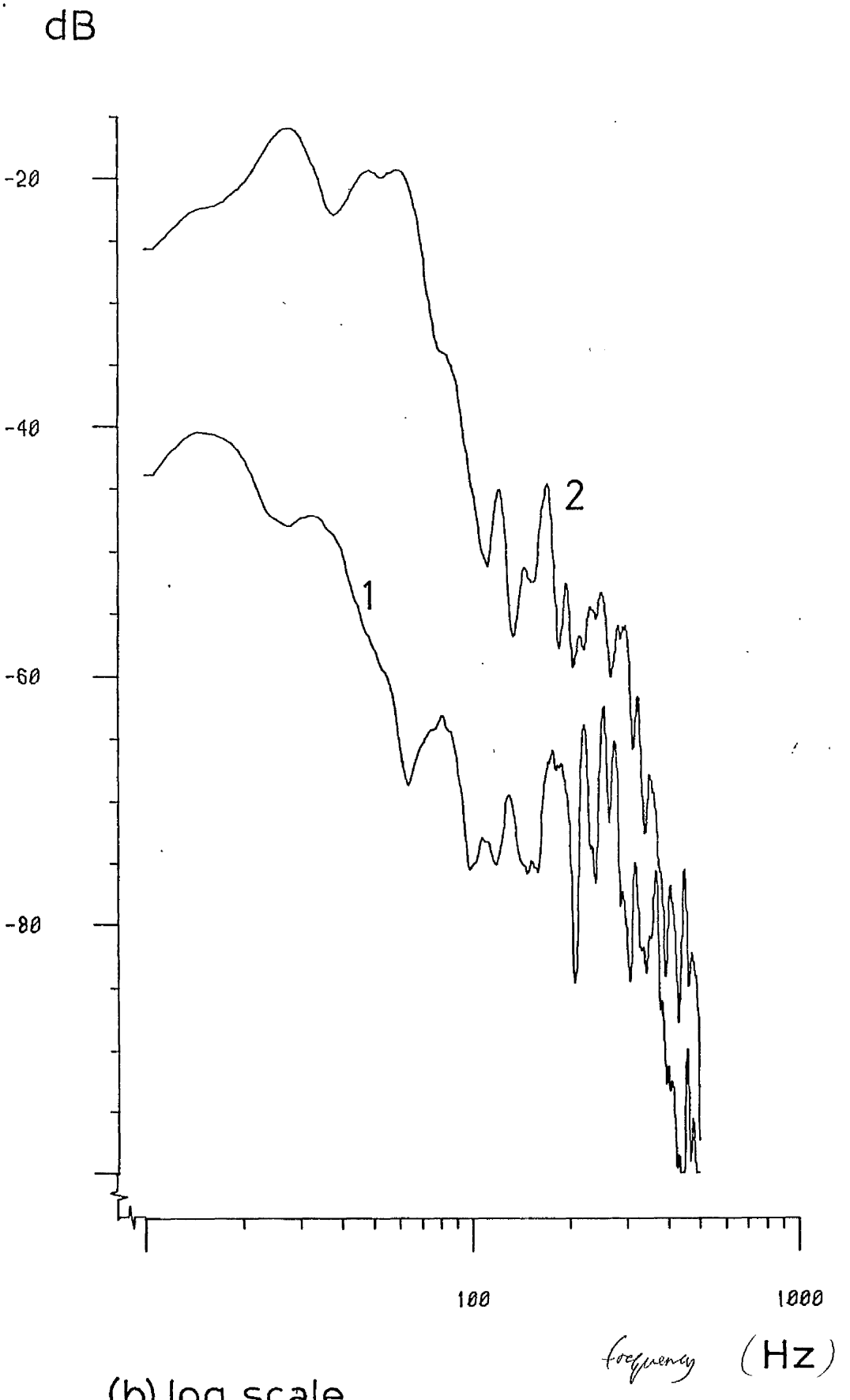
Interval	Lower frequency (Hz)	Upper frequency (Hz)
1	2	8
2	8	16
3	16	24
4	24	32
5	32	40
6	40	50
7	50	64
8	64	76
9	76	88
10	88	100
11	100	116
12	116	132
13	132	150
14	150	175
15	175	200
16	200	225
17	225	250
18	250	300
19	300	350
20	350	400

FIG 6.7

1. interference
2. late inspiratory breath sound



(a) linear scale



(b) log scale

BREATH SOUND IN RESPIRATORY DISEASE

7.1 Introduction

If any biological signal is to be of value in clinical medicine it must first be shown to provide information suitable for diagnosis and assessment of disease. In this context an experiment was devised to compare, by pattern analysis techniques, the spectra of lung sounds in patients with various respiratory diseases and in normal subjects.

Despite the advances in respiratory medicine in recent years, some diseases are difficult to diagnose. An example of such diseases are the pneumoconioses which are caused by damage to the parenchyma by inhaled dust. Asbestosis is an example of a pneumoconiosis with widespread risk in a number of industries. The effects of asbestos fibres on the lung are long term and may occur after withdrawal from exposure [7.1]. There is no method capable of the early detection of asbestosis and any more sensitive technique than those already available would be valuable.

With these considerations in mind our preliminary study included four groups: five patients with asbestosis, a disease associated with progressive fibrosis of the lungs; five patients with cryptogenic fibrosing alveolitis (CFA), which also is a fibrotic process but has a more marked inflammatory response; five hospital inpatients with radiological evidence of pulmonary oedema; and five healthy male non-smokers as controls. All the recordings were performed using the methods described in Chapter 6.

Investigation of the amplitude spectra of late inspiratory sound suggested that they tended to differ slightly from group to group (Fig

7.1). This then led to an analysis of the data set by extracting features from the amplitude spectrum. For each inspiration analysed, a 20-dimensional feature vector was extracted using the piecewise constant approximation described in Section 6.4. Seventy inspirations were analysed from each group. The 20-dimensional data set was then passed on for pattern analysis.

Two complementary pattern analysis techniques were used:- linear mappings and cluster analysis. The linear mapping used was the principal components or Karhunen-Loeve transformation [7.2], and the clustering method [7.3] used is the new graph theoretical method described in Chapter 5 of this thesis. The data set described here is used in Chapter 5 as a case study where a more technical consideration of the results is given.

Since the results given in this section 7.2 are based on late inspiratory sound, and since the spectra differ somewhat from those published elsewhere, the time variation of spectra is considered briefly in Section 7.3. The effect of recording using a pneumotachygraph is also shown.

7.2 Pattern Analysis

7.2.1 Principal Components Analysis

Principal components analysis was used throughout the development of the feature extractor. It provided a useful 2-dimensional representation of the data and gave a measure of feedback on the quality of the features produced. Information gleaned using principal components analysis led to the use of a normalization in feature

extraction.

The normalized features were first processed by computing the principal components of the 20-dimensional data set. Most of the variance is accounted for in the first two principal components (46% and 27% respectively). A plot of the first two principal components is given in Fig 7.2(a), each point representing one inspiration. There are 70 points in each of the 4 groups of 5 subjects i.e. an average of 14 inspirations per subject.

Several observations can be made. It is quite clear that the CFA and asbestosis groups are completely interpenetrating in the first two principal components. In view of the similarity between these two diseases, the overlap is not surprising. However the region containing these two groups together is completely separated from the region containing the normal group (Fig 7.2(b)) and is easily discriminated from the pulmonary oedema region. There is some overlap between the pulmonary oedema and normal groups but the group means are well separated.

A consideration of the spectra in Fig 7.1 suggests that at higher frequencies the features contain mainly system noise. It was therefore decided to examine the effect of removing the higher frequency features. Useful results were obtained by this feature selection and we show the results of using 6 features corresponding to 0-50Hz.

The principal components analysis of the six dimensional data set (Fig 7.3(a)) suggests that most of the information is contained in the 1st component whose eigenvalue is dominant (the first two principal components account for 65% and 16% of the data variance). Again there is a good separation of the fibrotic diseases from the normals (Fig 7.3(b)).

7.2.2 Cluster Analysis

Despite the fact that good results were obtained using principal components analysis, it is useful to study the data by means of a technique that works in the higher dimensional space. It was decided to use the graph theoretical clustering methods described in Chapter 5 to see if the groups clustered as well in 2Ø- or 6-dimensions as they appear to in Figs 7.2 & 7.3. The matter of greatest interest in diagnosis is how well separated the normals are from those with fibrotic disease. (Pulmonary oedema is easily diagnosed by other means than lung sound, whereas fibrotic dieseases are harder to detect). It was therefore decided to concentrate on asbestosis, CFA and normals.

The clustering programs described in Chapter 5 allow clustering based on relative distance, absolute distance or a combination of the two. Also it may be based on either the relative neighbourhood graph or the Gabriel graph. In either case the absolute distance clustering is identical to nearest neighbour or single linkage clustering. All three types of dissimilarity were tried and a strikingly similar picture emerged. Full details of the relative distance clustering in 6-dimensions are given in Section 5.4.3 as a case study of the clustering technique, but the single link and hybrid dissimilarity clusterings are shown in Fig 7.4. The treatment given in this chapter concentrates on results rather than methods.

The general picture is that the normals are separable from the fibrotic diseases both in terms of absolute and relative distance. This confirms the picture suggested by principal components analysis. In Section 5.4.3 the display of the main cluster containing asbestosis

and CFA confirms that the groups are highly interpenetrating and that that cluster has no obvious subdivisions.

7.3 Time Variation of Spectra

The results obtained in Section 7.2 were obtained by analysis of late inspiratory breath sound. While a number of authors have proposed time varying spectral analysis based on the DFT (e.g. [7.4]), very little systematic analysis has been undertaken. Frequently papers have been published stating that the maximum power in the breath sound spectrum is in the 100-300Hz range [7.5]. In view of the fact that the late inspiratory spectra, e.g. those in Fig 7.1, are different it is important to see how dependent the spectra are on the phase of the respiratory cycle. Another point of interest is the effect on sound recording of breathing through a pneumotachygraph which has been employed in a number of studies.

A number of recordings were processed by sampling at 997.4 Hz using the SBC100 microcomputer. This permits loggings of up to 15k samples thus allowing the selection of segments for further processing from the sampled data. The signals were analysed using 1024 point FFT algorithms followed by spectral smoothing. A higher sampling rate and larger size of transform could have been tried, but this exploratory study aimed only to cover the frequencies used in earlier studies.

Fig 7.5 shows the time plot of a normal subject. The early part of the inspiration obviously contains higher frequency components than the late part. The late inspiration is also easily separated from the expiration by its lower frequencies. These simple observations were confirmed by dividing the time series up into 1 second segments with

50% overlap. The overlap is useful to avoid detail lost by use of a 4-term Blackman-Harris window [7.6]. Fig 7.6(a) shows that in early inspiration there is considerable power above 100Hz, but this diminishes rapidly during the first half of the inspiration. The lower frequency sound is then comparable to that of Fig 7.1 during the second half inspiration. The expiration (Fig 7.6(b)) apparently does not change so dramatically.

Fig 7.7 shows an inspiration from the same normal subject as Figs 7.5-7.6. This time the subject is using a pneumotachygraph and the signal is attenuated. Fig 7.8 shows that the spectral pattern during the course of the inspiration is basically the same although at a lower amplitude. A further comparison was obtained by averaging three late inspiratory segments (Fig 7.9).

Finally we show time varying spectra in a case of asbestosis (Fig 7.11). The length of breath cycle shown in Fig 7.10 is much shorter in this case but it may give an interesting comparison. The early and mid sections of the inspiration have a greater proportion of high frequency sound than in normals (Fig 7.4) and adventitious sounds are present. The late inspiratory segment (segment 4) also has more higher frequency components than the normal despite the fact that there is virtually no evidence of crackles in that segment.

These results suggest a clear distinction between the spectra in early and late inspiratory sound. The spectra of early inspiratory sound correspond well with spectra reported in the literature whereas the late inspiratory spectra correspond well with those of Fig 7.1.

7.4 Discussion

The lung sounds detected at the chest wall are the summation of breath sounds and any additional adventitious sounds relating to disease of the airways or lung parenchyma. This preliminary study was confined to conditions in which the only adventitious sounds present were crackles. Although differences between crackles can be detected using the stethoscope, the features identified have not proved reliable in clinical practice where inter-observer variability is marked [7.7]. Electronic recordings offer a more precise tool for the characterization of crackles, and one study employing time-expanded waveform analysis has differentiated between the crackles of asbestosis and those of cardiac failure [7.8].

In contrast, the breath sound component of the lung sounds in the respiratory diseases in which crackling occurs have received little attention. Several factors may have contributed to this situation. Firstly auscultation using the stethoscope is less sensitive to the low frequency sounds and so, in clinical practice, it is seldom possible to comment on changes in breath sound unless there is a gross alteration. Secondly, recording equipment designed for studying crackling may selectively filter out low frequencies [7.9,7.10].

In this study differences in the lower frequency sounds were found to distinguish the recordings into three groups. In view of the fact that crackles in the conditions studied have spectral peaks above 400Hz it is most unlikely that they contributed to the low frequencies analysed. Also the energy contained in the crackles is much smaller than that of the underlying breath sound. (However it might be possible to remove the effect of crackles by using the recent robust

resistant spectral estimation methods [7.15].) The breath sound may therefore contain diagnostically significant information of a type hitherto unsuspected in the diseases studied.

A number of mechanisms might account for change in the low frequency lung sounds. The normal breath sound is determined by both the physical events within the airways producing an acoustic signal, i.e. generation, and the modifications by the structure through which the sound passes to reach the chest wall (transmission). While the site of generation is still uncertain, it is widely accepted that transmission through the lung is the major factor resulting in the predominantly low frequency of the sound heard at the chest wall. Many factor including flow rates, regional ventilation, lung volumes and posture [7.11,7.12] influence the breath sounds in normal subjects. Therefore it is likely that the structural and physiological abnormalities associated with lung disease account for the alteration in the lower frequency sounds irrespective of any adventitious sounds produced.

It is realized that in recording a number of individuals there will be differences in flow rate. However existing research suggests that flow rate affects sound intensity rather than spectral shape. Forgacs [7.13] describes the variation of sound intensity with flow rate and Banaszak et al [7.14] give evidence to suggest that spectral shape does not change with flow rate. However the time varying spectra suggest that the spectral shape during the course of a breath cycle is related to the mechanics of breathing. The higher frequency components of the sound are probably related to lower lung volumes and so an explanation of the spectral differences in asbestosis in terms of reduced lung volume cannot be ruled out.

The fact that the recordings were obtained from recordings made in a general ward rather than in a special soundproofed environment is also encouraging. If analysis of lung sound signals can be improved, a useful contribution might be made to the repetoire of non-invasive diagnostic techniques in general.

Finally it is worth noting that the above results were based on breath sound of relatively low frequency. These frequencies have been ignored since 1929 when Hannon & Lyman [7.16] reported tracings very similar to Fig 7.5.

Refererences

7.1 M.R.Becklake, Editorial: Asbestos-related diseases of the lungs and pleura - current clinical issues, Am.Rev.Respir.Dis, 126, 187-194 (1982)

7.2 C.Chatfield & A.J.Collins, Introduction to Multivariate Analysis, Chapman & Hall (1980)

7.3 R.B.Urquhart, Graph theoretical clustering based on limited neighbourhood sets, Pattern Recogntion, 15, 173-188 (1982)

7.4 F.T.Wooten, W.W.Waring, M.J.Wegman, W.F.Anderson, J.D.Conley, Med.Instrum., 12, 254-257 (1978)

7.5 R.L.H.Murphy & S.K.Holford, Lung Sounds, American Thoracic Society (1980)

7.6 F.J.Harris, On the use of windows for harmonic analysis with the discrete Fourier transform, Proc. IEEE, 66, 51-83 (1978)

7.7 L.D.Hudson, R.D.Conn, R.S.Matsubara & A.H.Pribble, Non-specificity of qualitative descriptions of crackles, 1st Internat. Conf on Lung Sound, Boston (1976)

7.8 S.K.Holford & R.L.H.Murphy, Differentiation of the rales of pulmonary asbestosis and congestive heart failure, 1st Internat. Conf. on Lung Sounds, Boston (1976)

7.9 M.Mori, K.Kinoshita, H.Morinari, T.Shiraishi, S.Koike & S.Murao, Waveform and spectral analysis of crackles, Thorax, 35, 843-850 (1980)

7.10 A.R.Nath & L.H.Capel, Lung crackles in bronchiectasis, Thorax, 35, 696-699 (1982)

7.11 P.Leblanc, P.T.Macklem & W.R.D.Ross, Breath sounds and distibution of pulmonary ventilation, Am. Rev. Resepir. Dis., 102, 10-16 (1970)

7.12 Y.Ploy-Song-Sang, R.R.Martin, W.R.D.Ross, R.G.Loudon & P.T.Macklem, Breath sound and regional ventilation, Am. Rev. Respir. Dis., 116, 187-199 (1975)

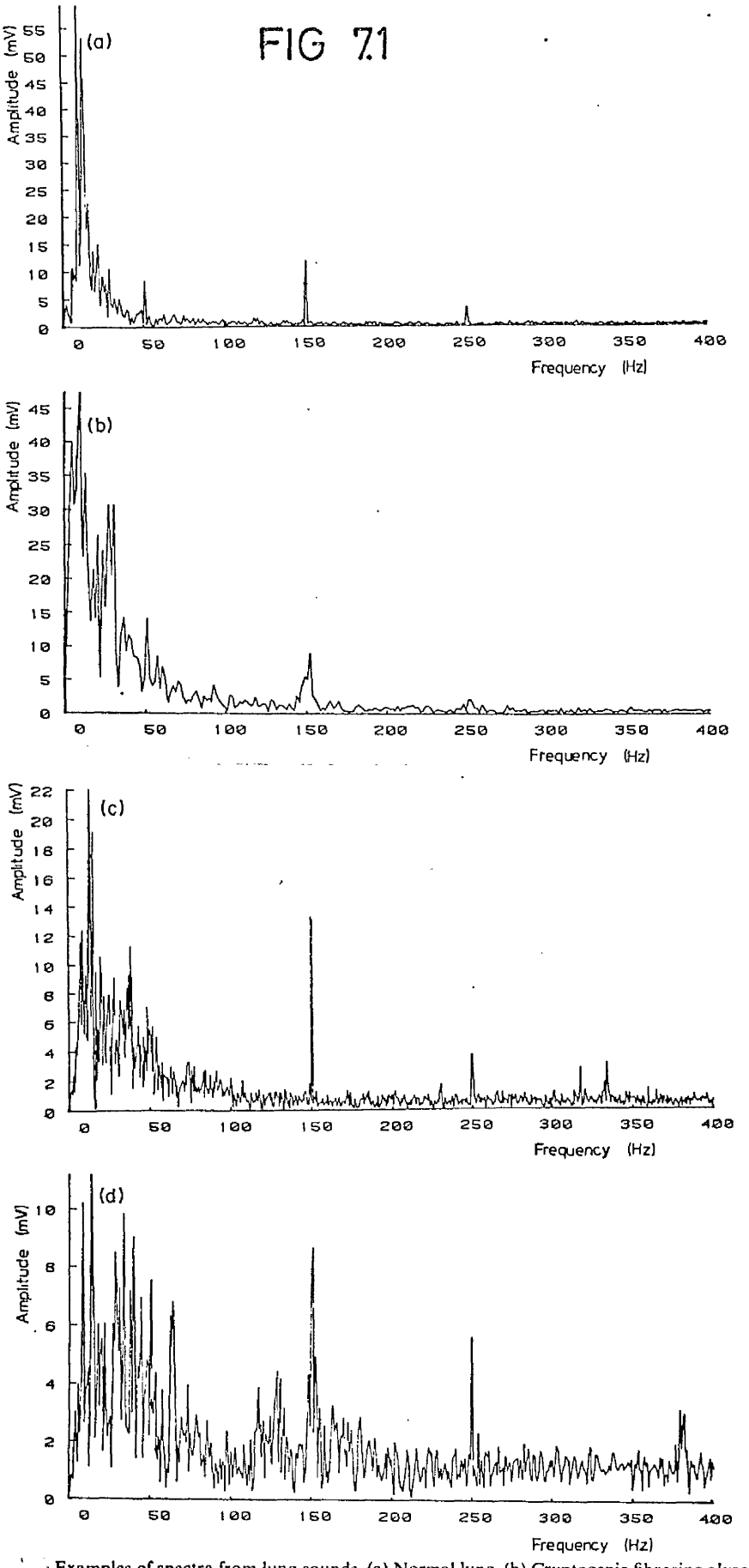
7.13 P.Forgacs, Lung Sounds, Balliere-Tindall, London (1978)

7.14 E.F.Banaszak, R.C.Kory, G.L.Snider, Phonopneumography, Am. Rev. Respir. Dis., 107, 449-455 (1978)

7.15 R.D.Martin & D.J.Thomson, Robust-resistant spectral estimation, Proc. IEEE, 70, 1097-1115 (1982)

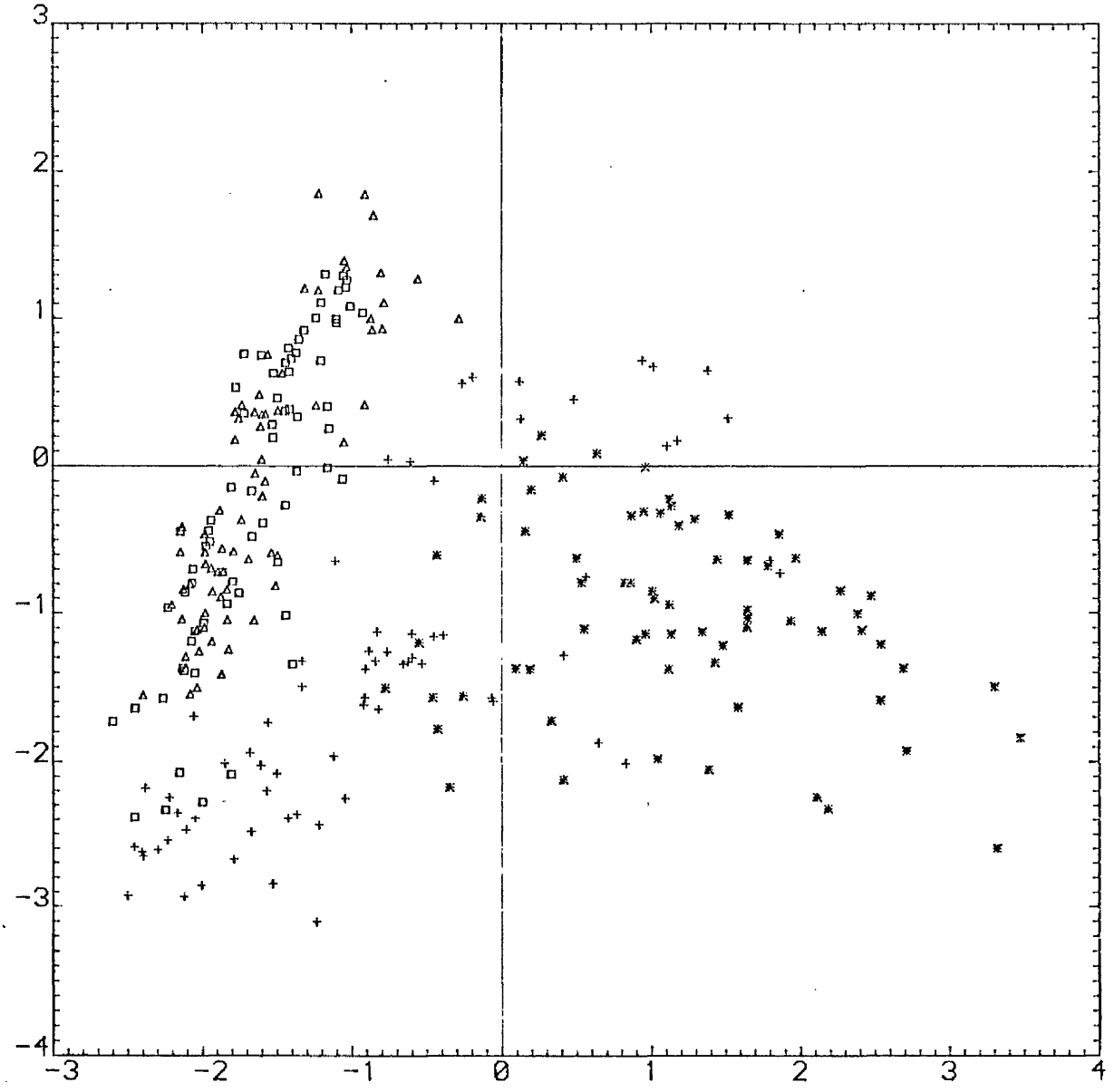
7.16 R.R.Hannon & R.S.Lyman, Studies in pulmonary acoustics II, The transmission of tracheal sounds through freshly extenterated sheep lungs, Am.Rev.Tuberc., 19, 360-375 (1929)

FIG 7.1



Examples of spectra from lung sounds. (a) Normal lung. (b) Cryptogenic fibrosing alveolitis (CFA). (c) Asbestosis. (d) Interstitial pulmonary oedema. Note mains peaks at 50, 150, and 250 Hz. The absolute magnitudes of the spectra varied but within each group the spectral shape tended to remain the same.

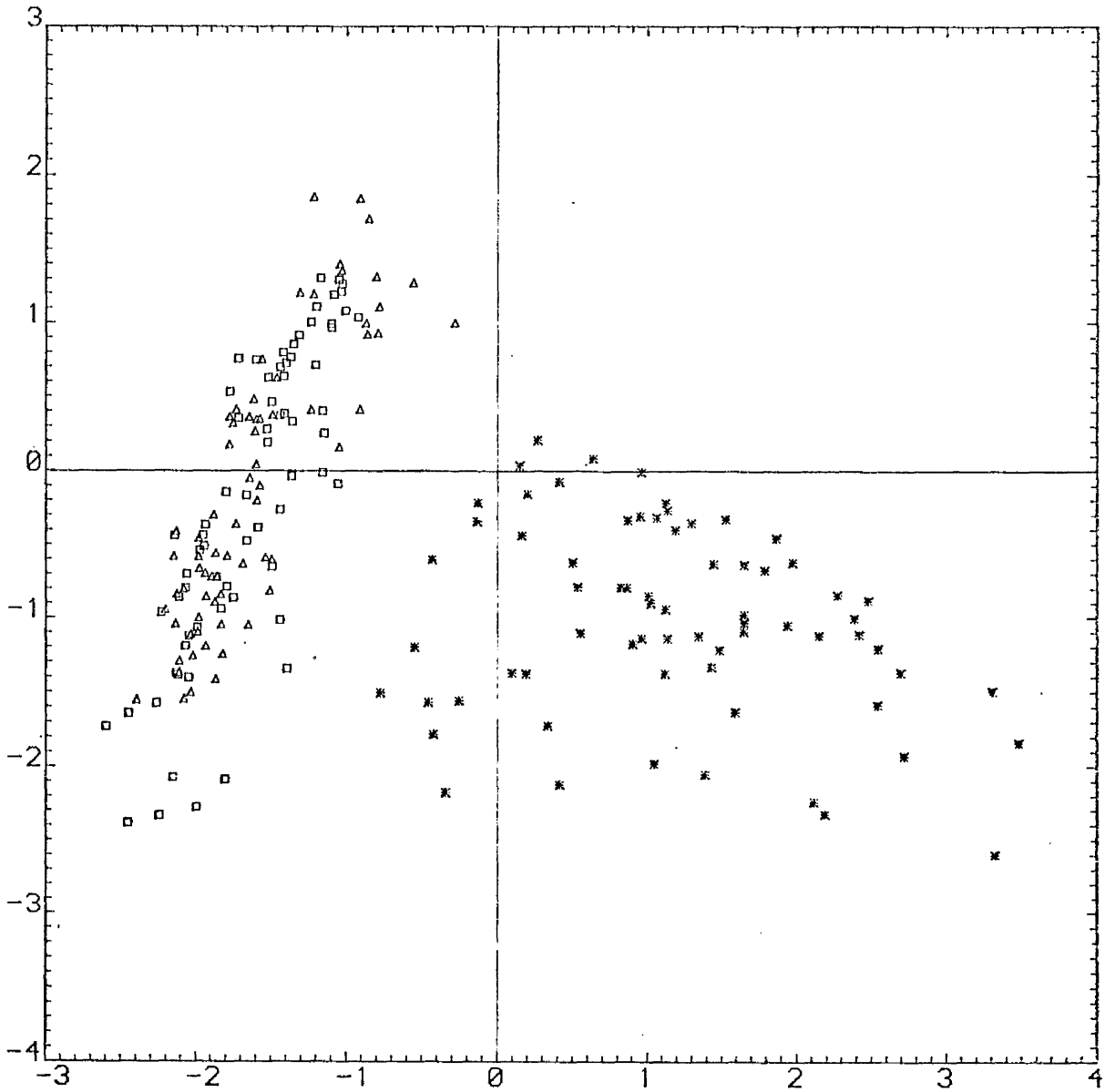
FIG 7.2



(a) Plot of the first two principal components of the 20-dimensional lung sound data

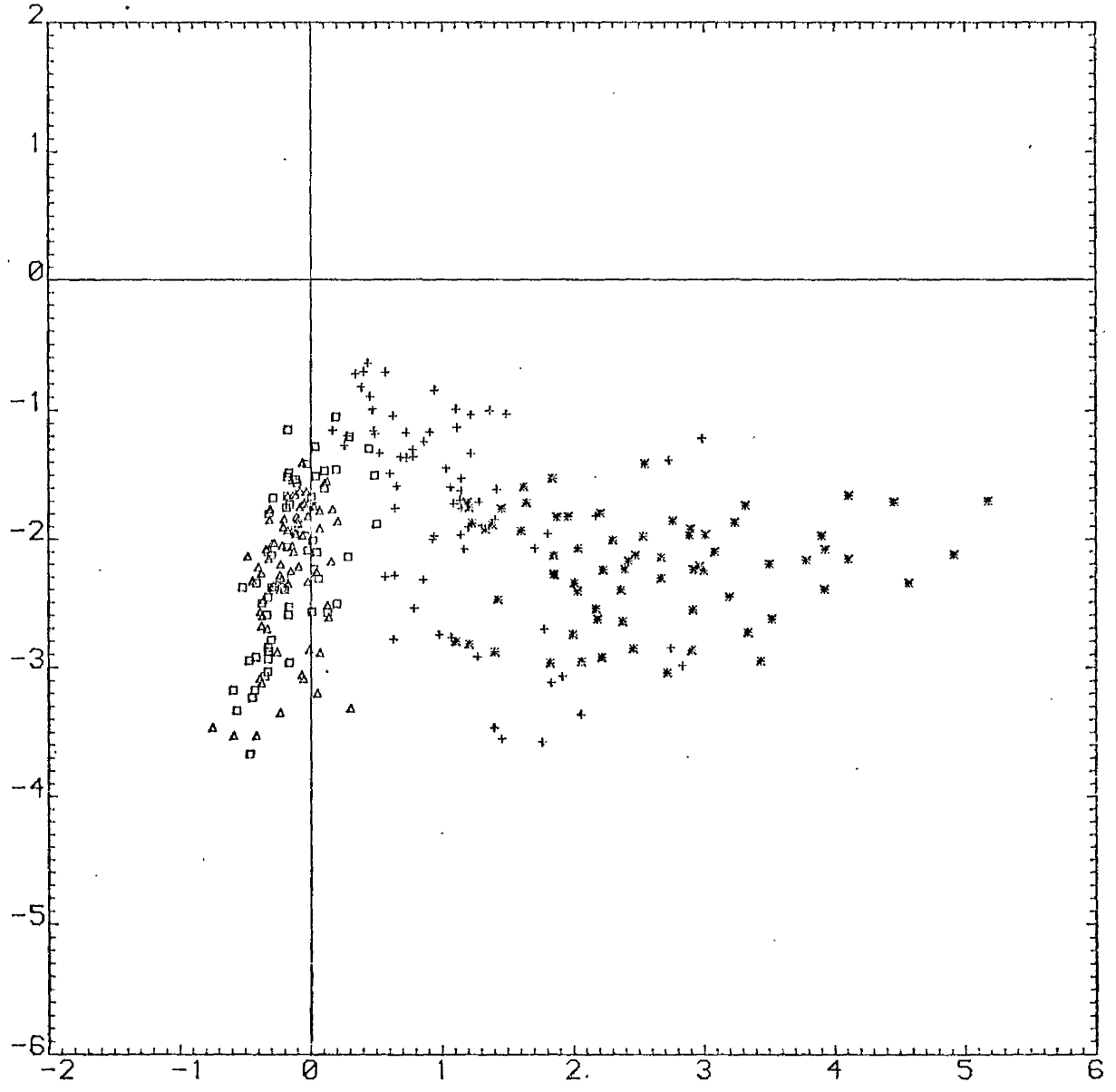
Key:
□ asbestosis
△ CFA
+ pulmonary oedema
* controls

FIG 7.2



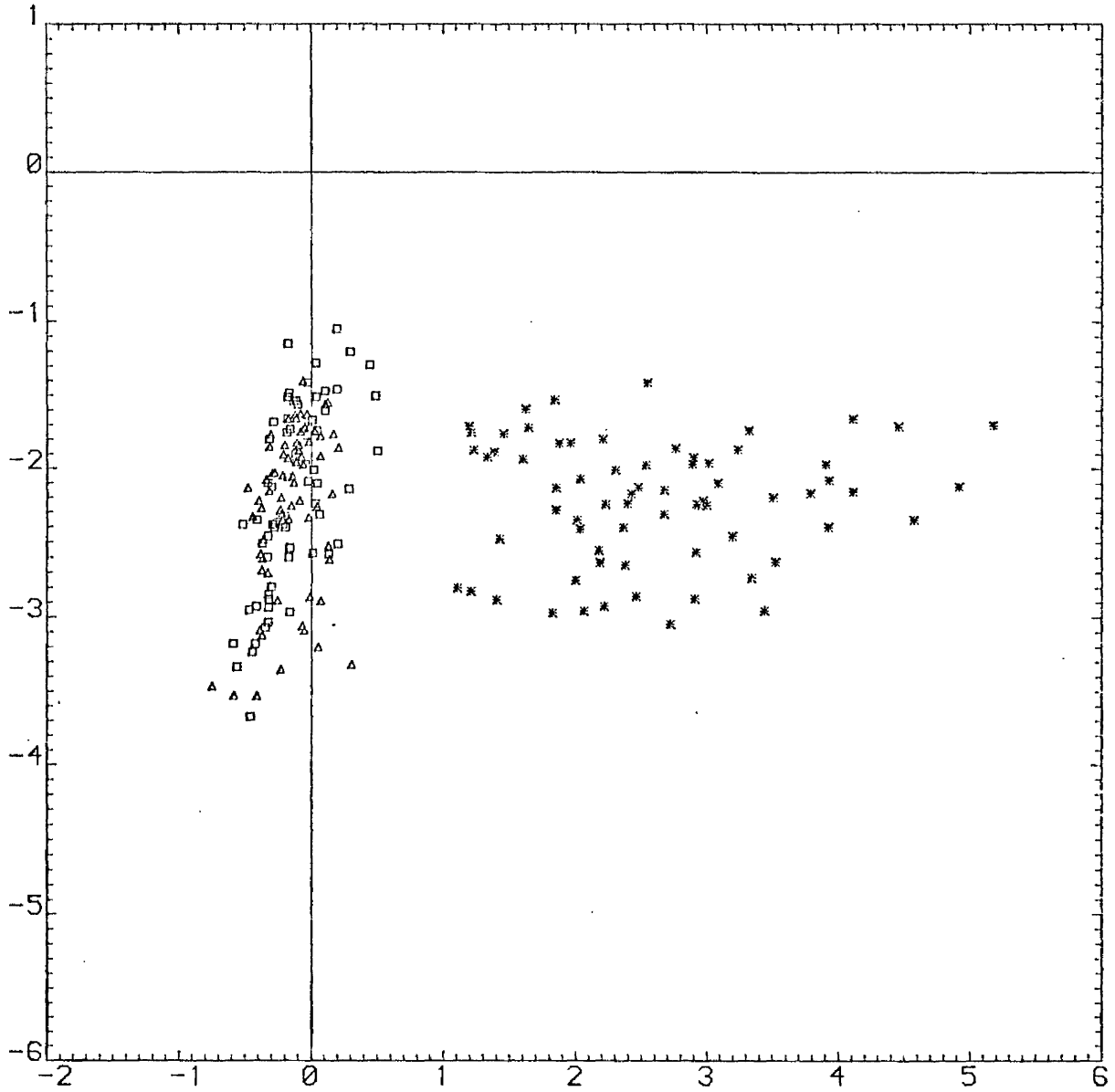
(b) Same as (a) but omitting pulmonary oedema

FIG 7.3



(a) Plot of the first two principal components of the 6-dimensional lung sound data

FIG 7.3



(b) Same as (a) but omitting pumonary oedema

FIG 7.4(a)

Single link clustering of the 20-D data set

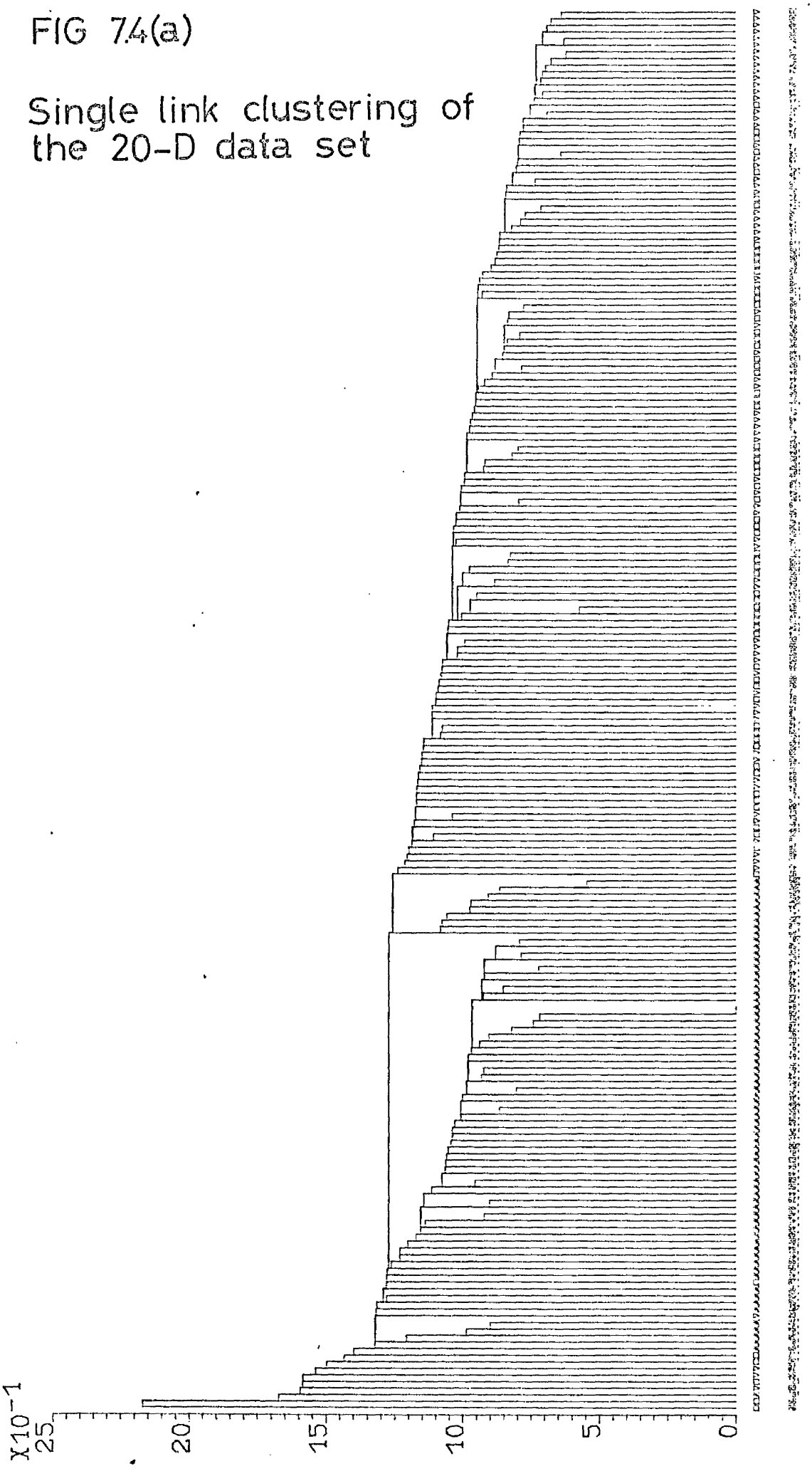


FIG 7.4(b)

Single link clustering of the 6-D data set

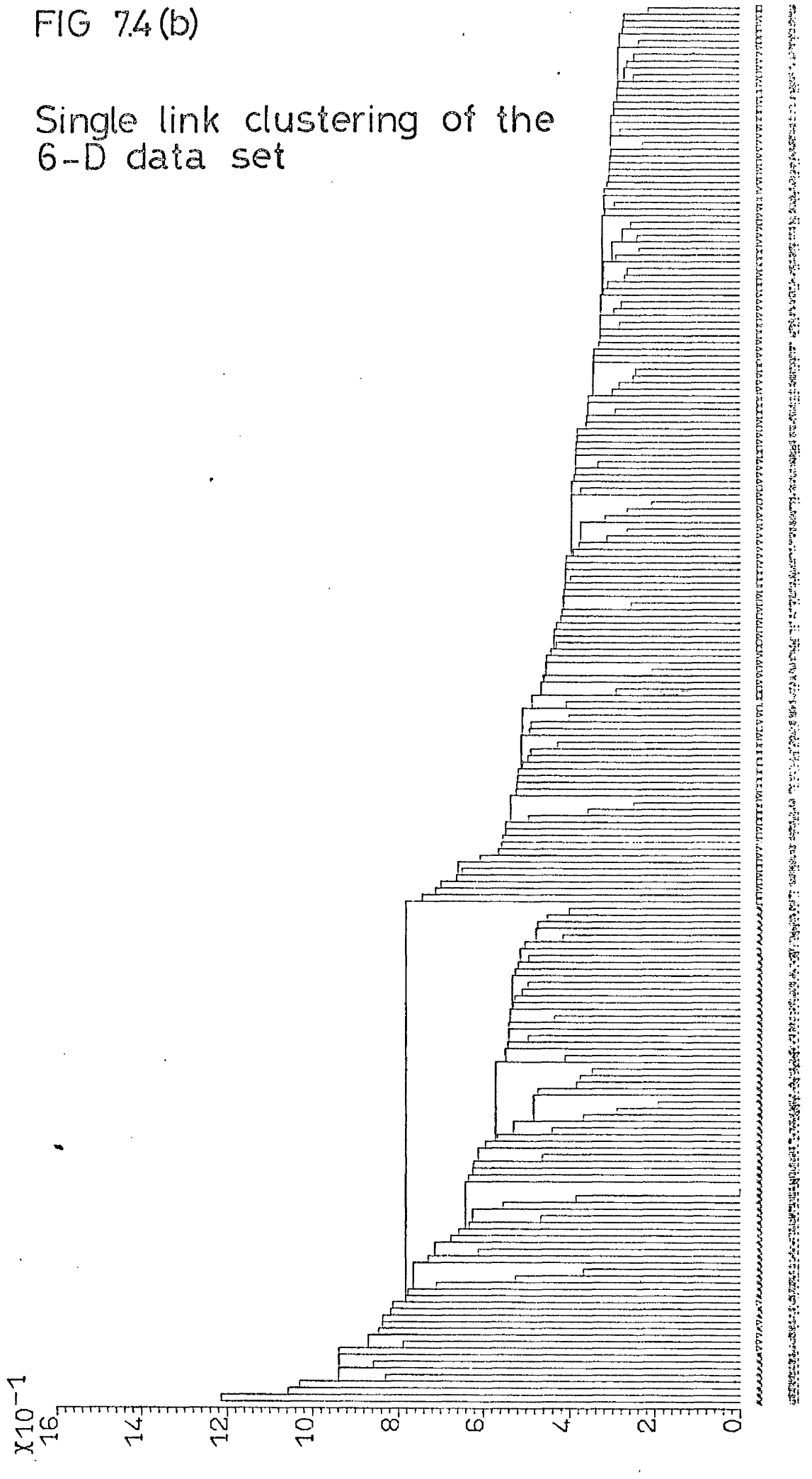


FIG 7.4(c)

Hybrid clustering of the 20-D data set using the G.G.

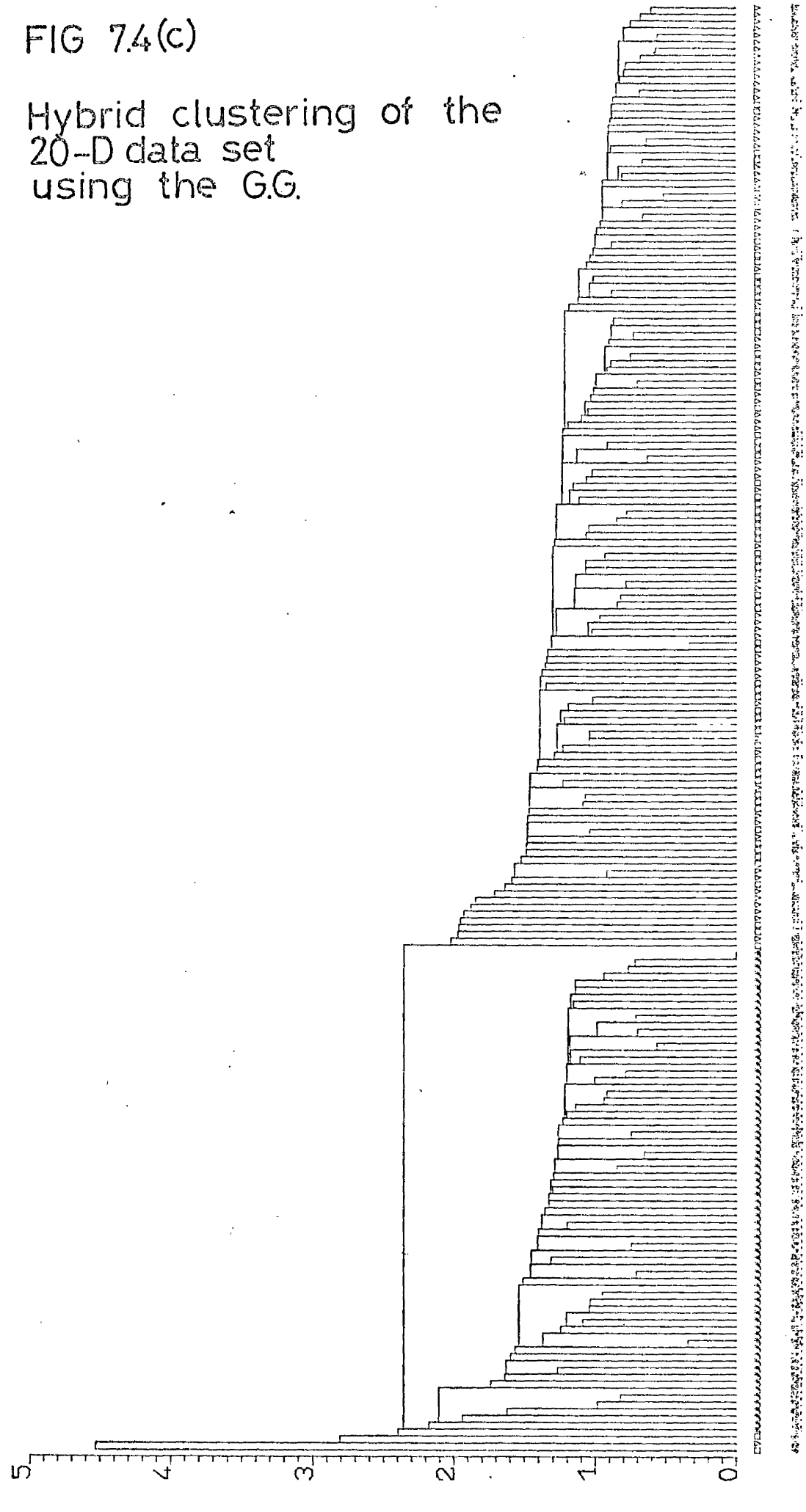


FIG 7.4(d)

Hybrid clustering of the 6-D data set using the G.G.

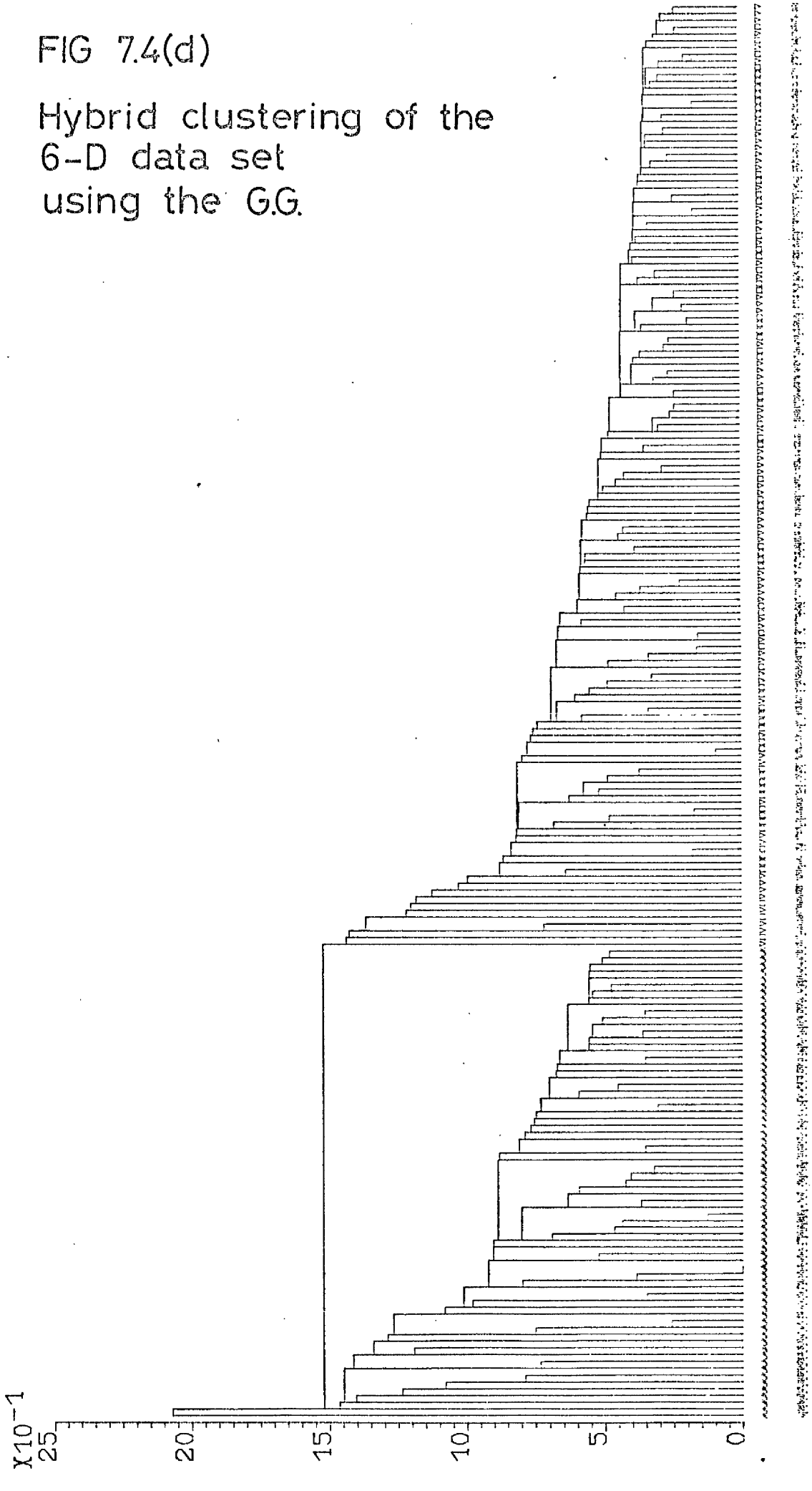


FIG 7.5

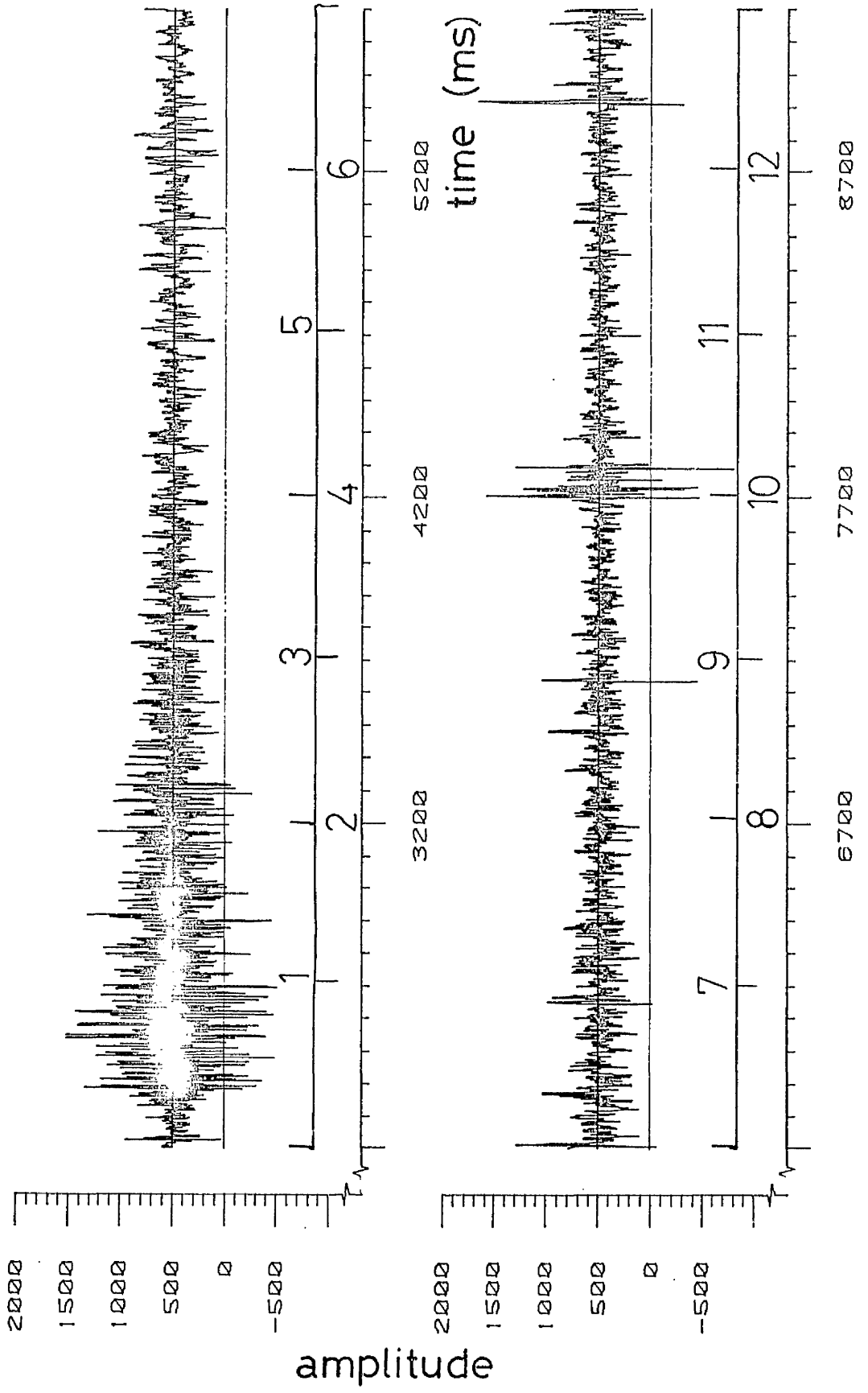
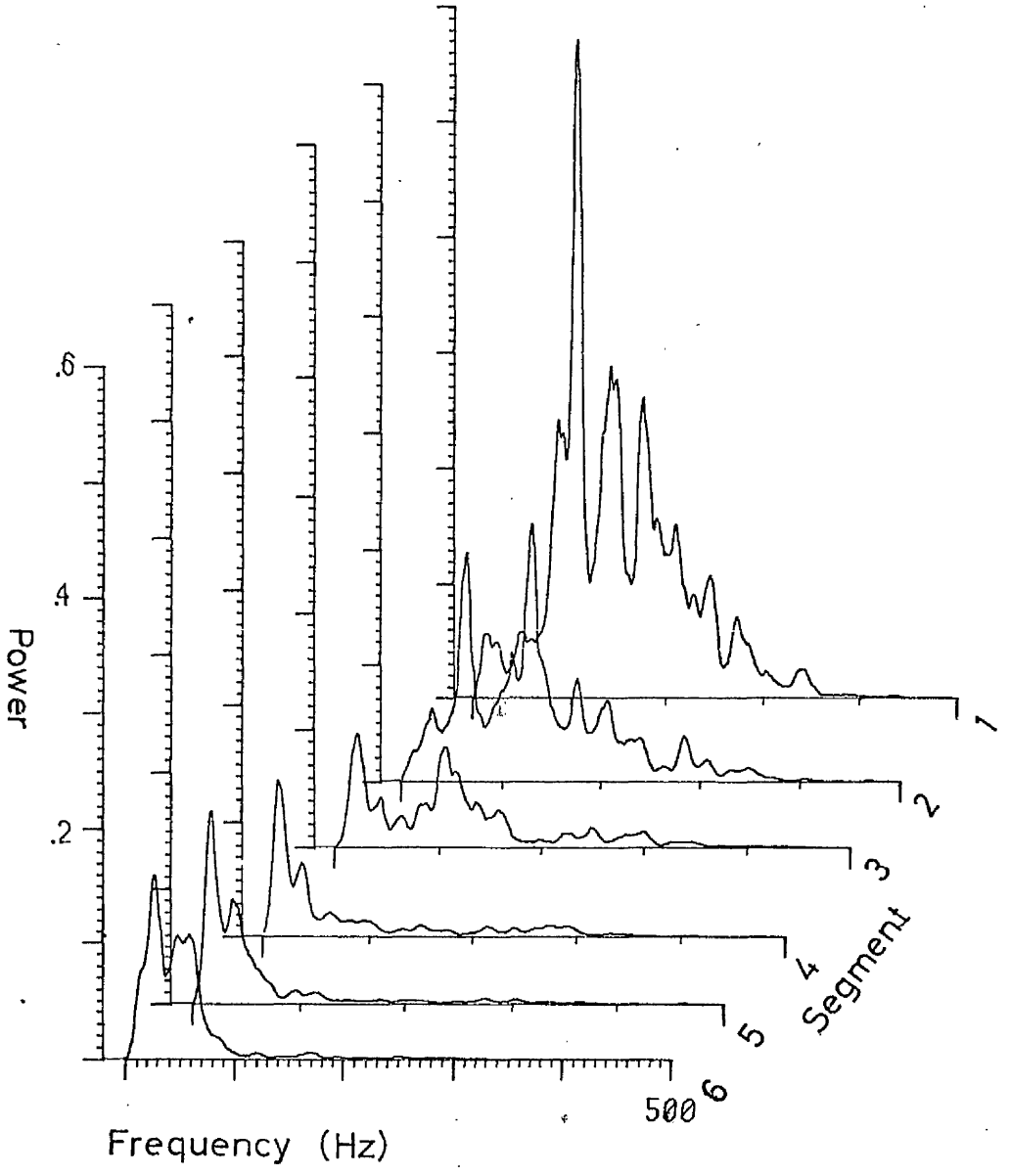

time (ms)
3200
4200
5200
6700
7700
8700
1
2
3
4
5
6
7
8
9
10
11
12
2000
1500
1000
500
0
-500
amplitude

FIG 7.6(a) Inspiration

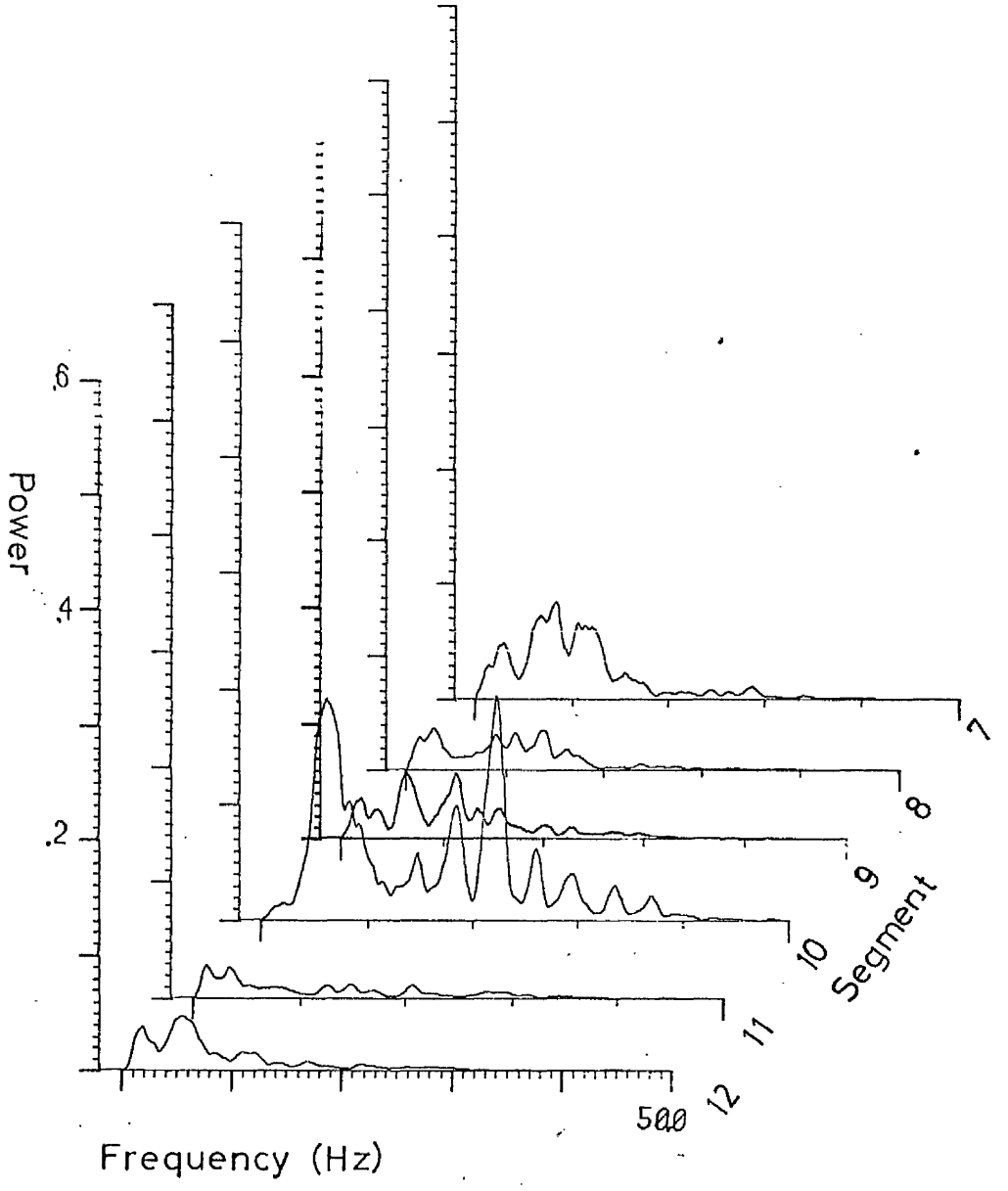
Segment numbers correspond to those in Fig 7.5



Note the appreciable power above 100 Hz which diminishes during the course of inspiration

FIG 7.6(b) Expiration

Segment numbers correspond to those in Fig 7.5



Note the effect of microphone movement in segment 9

FIG 7.7

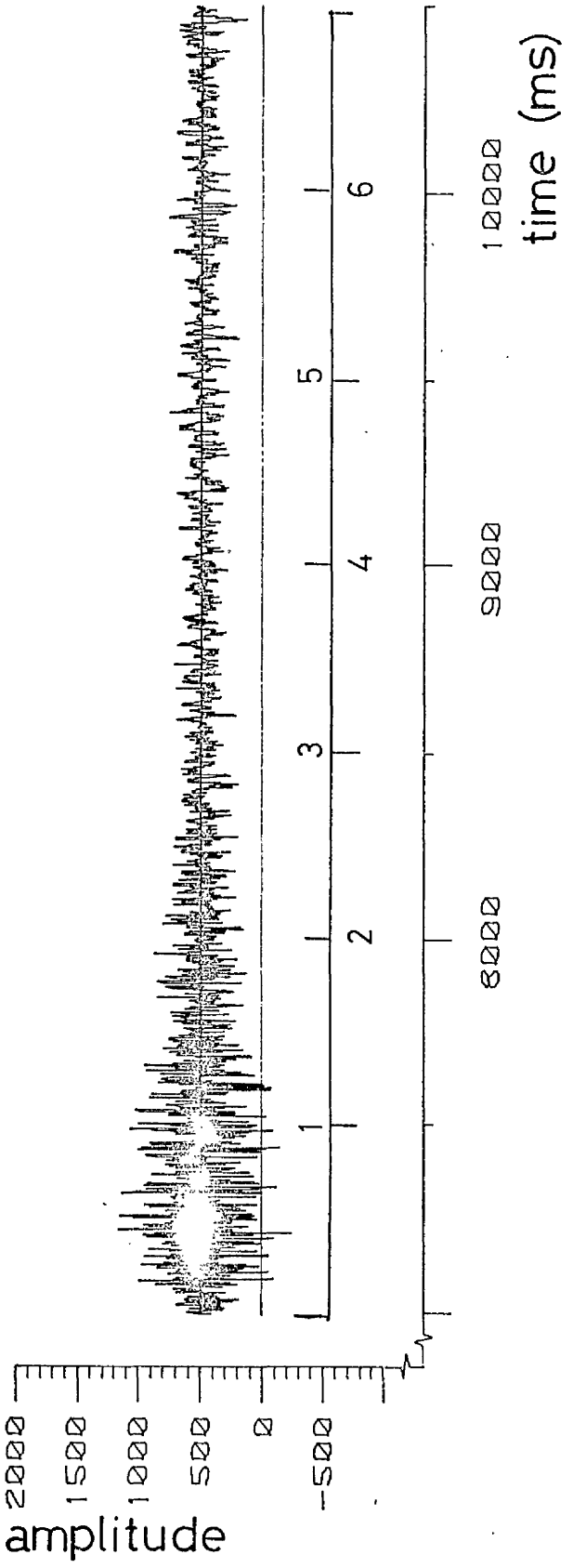

2000
1500
1000
500
0
-500
amplitude
1
2
3
4
5
6
8000
9000
10000
time (ms)

FIG 7.8 Inspiration

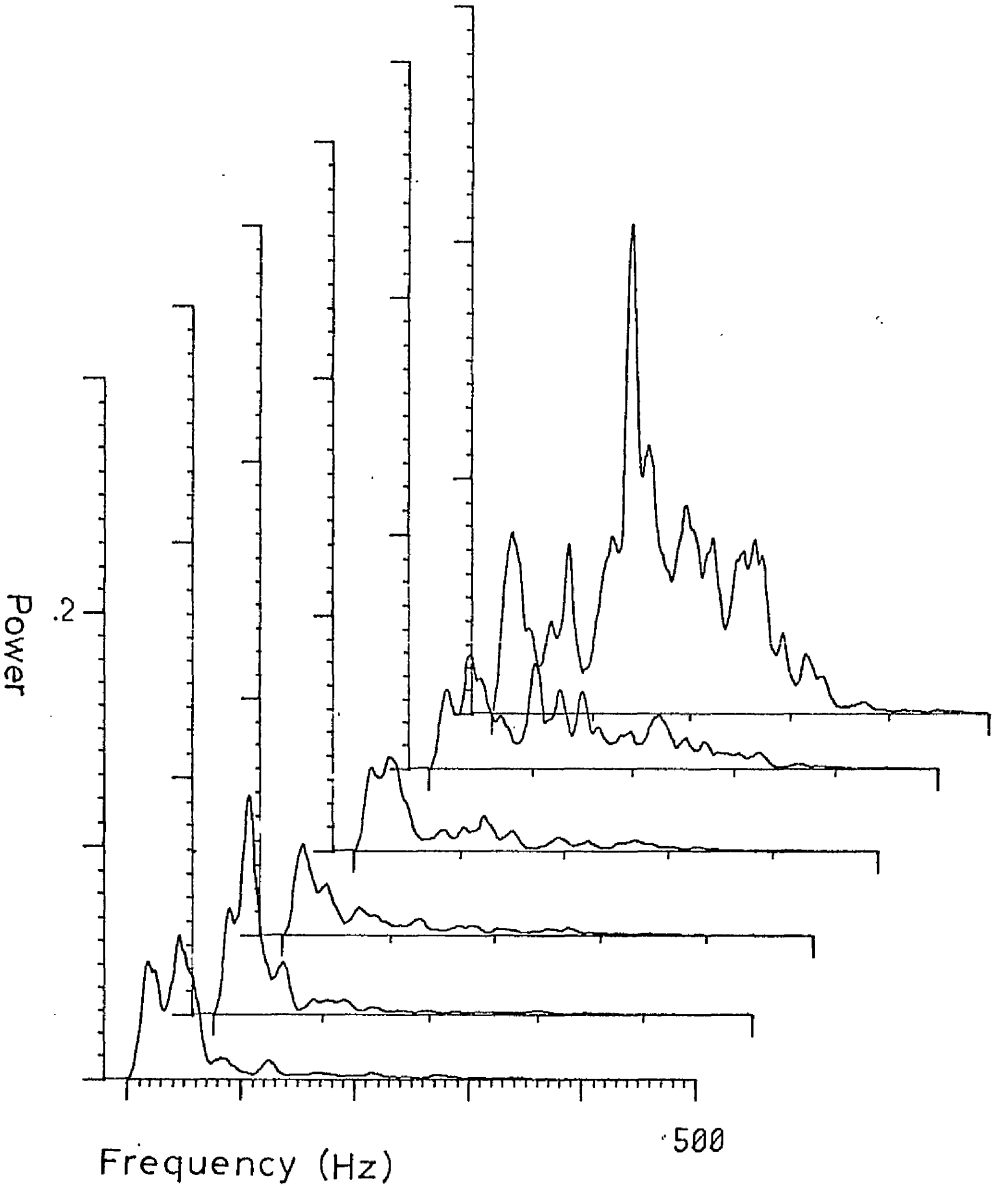


FIG 7.9(a)

Averaged spectra plotted on a linear scale

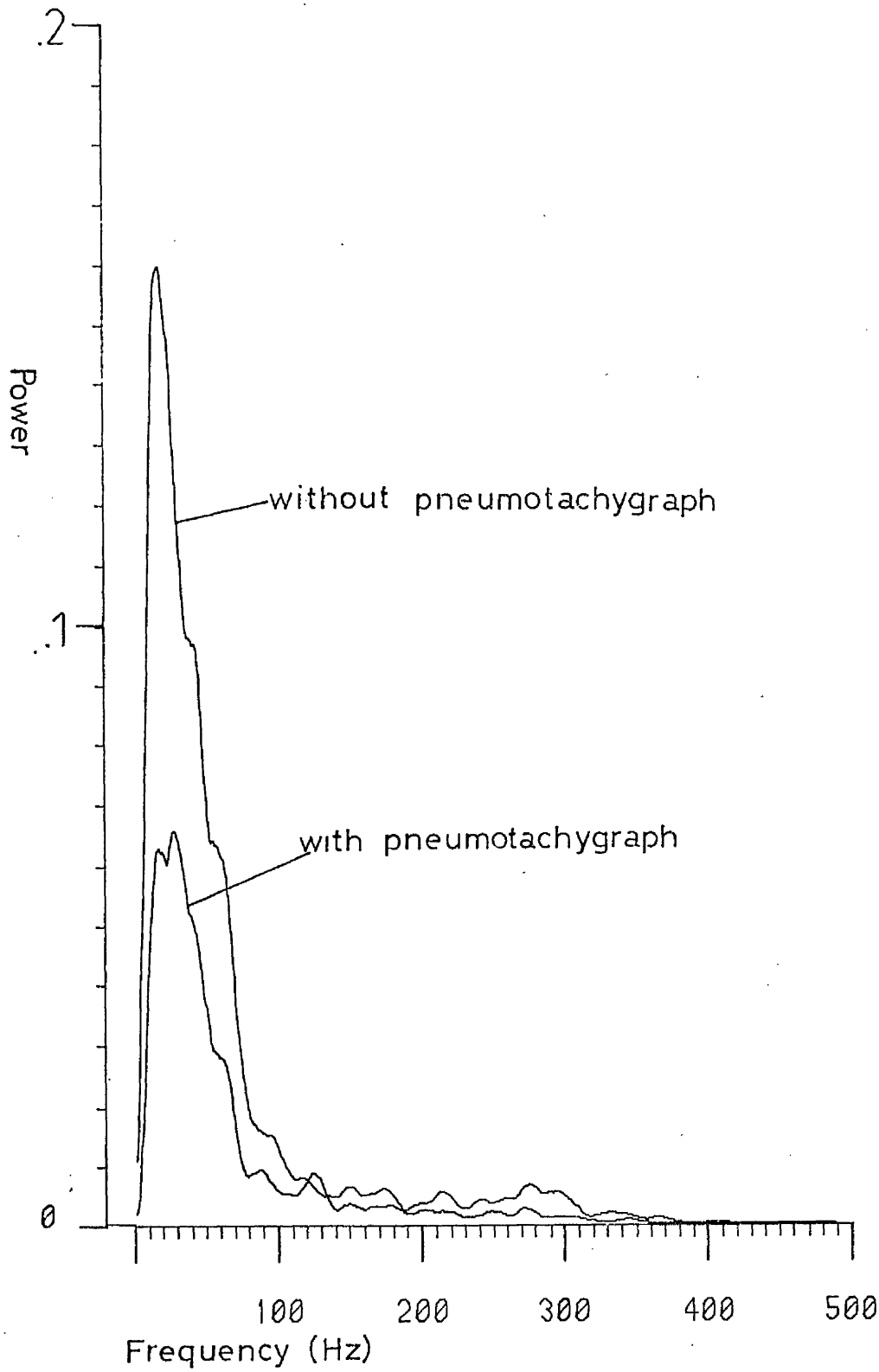


FIG 7.9(b)

Averaged spectra plotted on a log scale

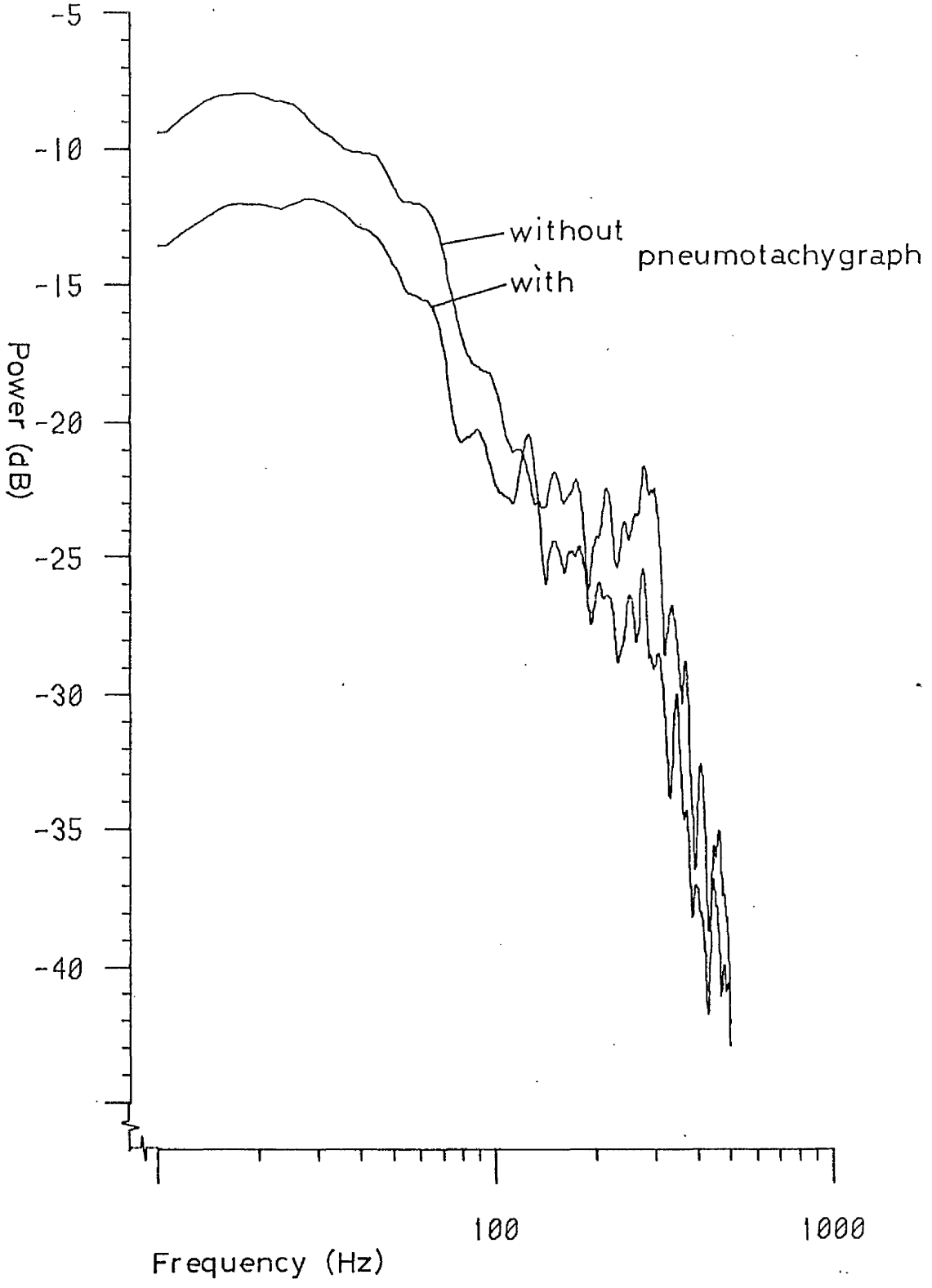


FIG 7.10

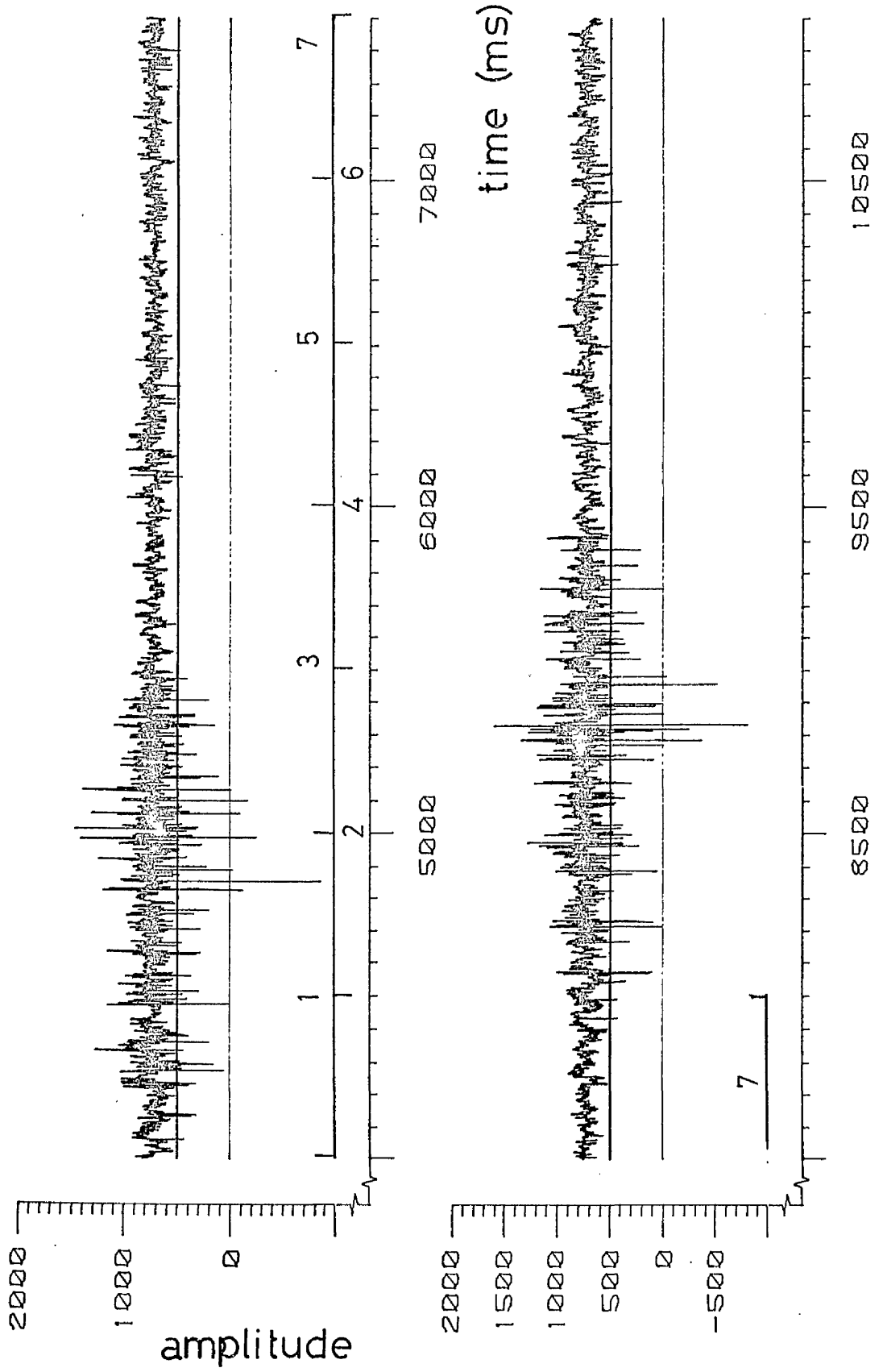
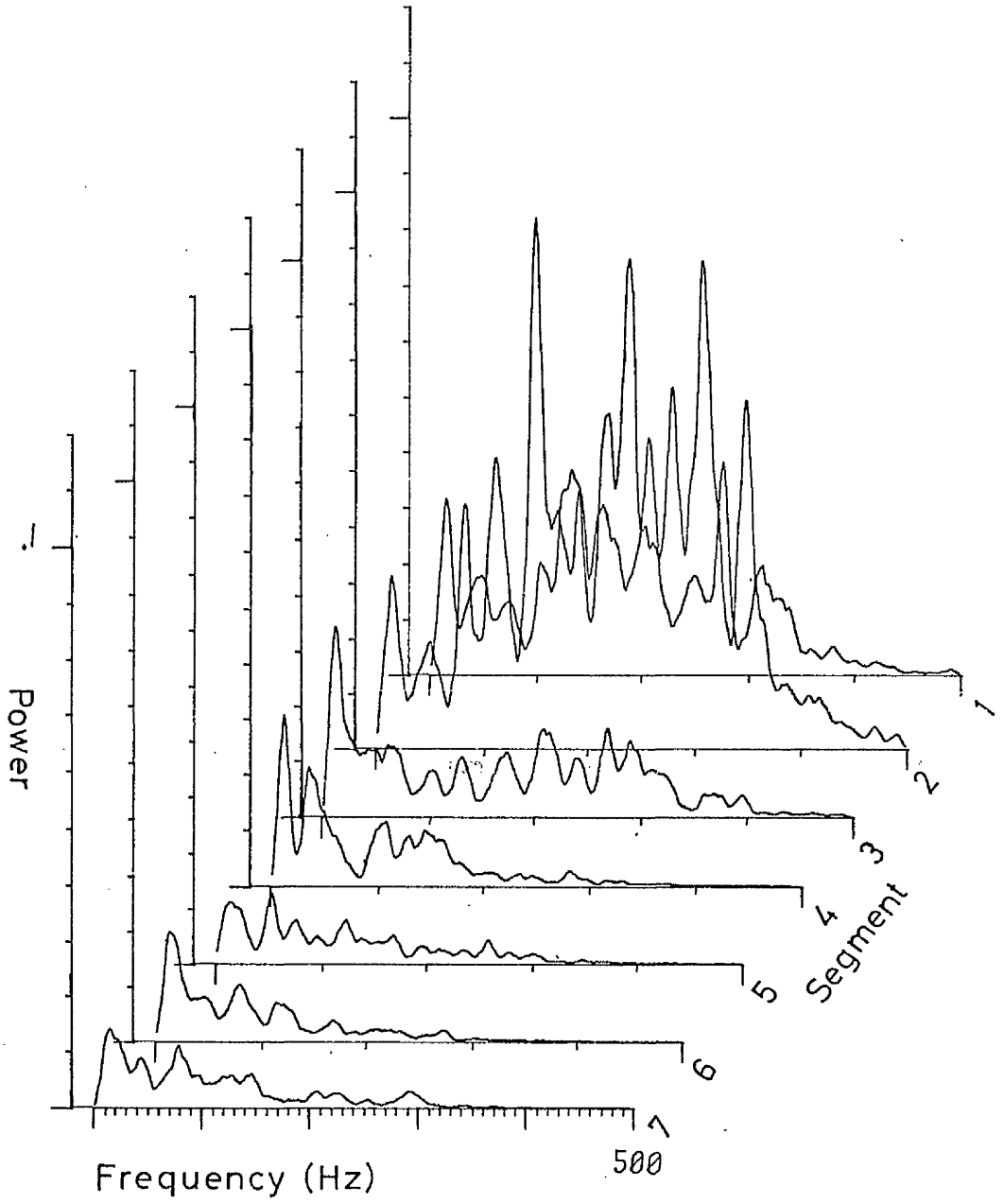

amplitude
2000
1000
0
1
2
3
4
5
6
7
5000
6000
7000
time (ms)
2000
1500
1000
500
0
-500
7
8500
9500
10500

FIG 7.11

Segment numbers correspond to those in Fig 7.10



Inspiration segments (1-4)

Expiration segments (5-7)

CONCLUSIONS & SUGGESTIONS FOR FUTURE RESEARCH

8.1 General Conclusions

Although it is hoped that the research described in this thesis reads as a whole, the conclusions are best divided into two sections. The new geometrically-based pattern recognition techniques should be considered as general research tools whereas the methods developed for lung sound signal analysis are for a particular application.

8.1.1 Pattern Recognition Techniques

This work has extended the application of computational geometric techniques to pattern recognition. Some properties of the Gabriel graph (GG) and relative neighbourhood graph (RNG) have been derived. These have been based on generalizing the definitions of the GG and RNG and considering graphs defined by a region of influence. The effect of the region of influence on the connectivity and planarity of the resulting graph are particularly inportant and the connectivity property has allowed the development of a new non-parametric clustering technique.

The clustering work is similar in its aims to the earlier work of Zahn [8.1] in that it attempts to form clusters according to a visual perceptual model for the cluster. However there are some deficiencies in the use of the minimal spanning tree (MST) for clustering. The new clustering method is capable of detecting clusters that are either disjoint or separable in terms of point density while overcoming some of the limitations of the MST. An important property of the method is

that the dissimilarity coefficient implied by the technique is continuous. The principal conclusion of this part of the work is that the RNG and GG are better structures on which to base a visually-based clustering method than the MST.

The use of such structures has also allowed the development of an interactive tool for studying the inter-relationships between clusters and for identifiying the types of cluster found. This aid should give the user a greater amount of information on a partition in the clustering and hence gain a greater insight into the results. It may also be viewed as a method for recovering information from the GG (or RNG) that is not used in the clustering.

Results obtained using these methods have been encouraging although it must be realized that any clustering technique is limited by the appropriateness of its underlying assumptions to the data being studied. Hence this clustering technique might find a useful place in a suite of clustering methods each of which would generate clusters according to a different set of assumptions [8.2].

8.1.2 Lung Sound Signal Analysis

This work has shed further light on whether lung sound signals could be used in the diagnosis of respiratory disease. In a preliminary study comparing three diseases with normals, using spectral analysis and pattern recognition techniques, the following groups were distinguished: (a) normal subjects, (b) patients with fibrosis of the lungs (asbestosis and cryptogenic fibrosing alveolitis) and (c) patients with interstitial pulmonary oedema. Evidence was found to suggest that these differences were attributable to changes in breath sound rather than added sounds. This conclusion

does have encouraging implications for diagnosis since alteration in breath sound had not previously been recognised as important in the diseases studied.

Additionally some work was done to investigate ways of improving the analysis of lung sound signals. This was to identify sources of systematic error in recording and to avoid the use of ad hoc methods. The main conclusion of this work was to use better spectral estimates by smoothing the periodogram and to normalize the spectral features. New software was developed to process the recordings systematically, based on the improved methods of analysis, but time did not permit the completion of this work.

The ultimate aim of this work would be to develop some sort of diagnostic aid and/or screening test for asbestosis. The results so far do not exclude this possibility although a considerable widening of the research is required to prove that this is more than speculative.

8.2 Suggestions for Further Research

8.2.1 Pattern Recognition Techniques

One area in which there is a need for further research is the properties of the GG and RNG in higher dimensions and in non-Euclidean geometries. Matula & Sokal [8.3] derived a number of useful properties of the planar GG. Knowledge of some of these properties for the RNG e.g. the number of edges in a maximal graph would be useful especially in higher dimensions.

Another more speculative line of research is to apply the graph

theoretical techniques to optimization of non-linear mapping algorithms. This is proposed in some detail in Appendix 1.

8.2.2 Lung Sound Signal Analysis

One of the long term technical objectives of this work must be to provide suitable instrumentation to make acquisition and analysis of lung sound a routine matter. When this is done the bulk of the research can become clinical.

A major limitation of the work on lung sound has been the difficulty of data acquisition without dedicated equipment. Frequently during the asbestosis recording programme suitable patients were visiting the Royal Infirmary but equipment could not be brought together quickly enough to organize a recording. Additionally the long time taken to analyse tapes prevented any feedback of results into the clinical environment. When compared with direct data logging the use of the tape recorder is very cumbersome and time consuming.

A major improvement would be to invest in a microcomputer-based data acquisition and analysis unit. Such a unit, if designed with suitably simple controls, could allow doctors to make recordings at short notice and allow much more rapid data collection. The provision of standard analyses e.g. breath sound spectral estimates would permit the use of quantitative lung sound analysis in the clinical situation. This would probably allow far more insight into lung sound from the clinical angle than is possible at present. Sampled data could regularly be transferred to a computer at the University for any further processing without time being wasted on data logging. Such a unit could either be built from scratch or by adapting a standard microcomputer. Another difficulty was in persuading outpatients to

volunteer to be subjects in the experiments.

A number of specific projects naturally arise from the work described int this thesis. These can be divided into (A) improving existing analysis techniques and (B) furthering clinically orientated research into lung sound. In a sense both types of project could work hand in hand since progress in improving analysis techniques is best judged by its suitability to clinically based research.

These suggestions are listed below:-

A1 Transducers

Very little is currently known about the best way in which to design and test lung sound trasducers. One of the major problems is the estimation of the acoustic parameters of the chest wall and to incorporate this information into future transducer designs. An obvious approach would be to study the transmission of sound across a sheep's thorax.

A very interesting line of research would be to investigate hydrophones. The acoustic properties of the thorax are broadly similar to those of water and so hydrophones might be more suitable transducers than the microphones used at present.

A2 Signal Stationarity

Spectral estimation techniques frequently assume stationarity of the signal. While inspiration and expiration have been carefully segmented, it would be useful to investigate the stationarity of the signal using techniques discussed in Section 3.6. Ideally this would lead to a splitting of the lung sound signal into quasi-stationary segments.

Clearly segmentation would have to be considered in some sort of hierarchy [8.4]. At the top level there is segmentation into

inspiration and expiration. Below this we require to investigate whether an inspiration (or expiration) should be be subdivided for spectral analysis. At a low level we require segmentation into breath sound and adventitious sounds.

A3 Spectral Estimation

The problem of variation in the length of the breath cycle led initially to an ad hoc approach to spectral estimation. A more systematic approach avoided the ad hoc solution but still did not solve the problem of having input data (samples from a breath cycle) of differing lengths. This problem could largely be overcome by using either the chirp z-transform (CZT) algorithm [8.5] or maximum entropy spectral analysis [8.6].

Unlike the usual FFT algorithms, the CZT algorithm may have a number of output points that is unrelated to the number of input data points. Therefore the CZT algorithm could be used to compute DFTs of a fixed resolution while being given input sequences of varying lengths.

Maximum entropy spectral analysis is equivalent to fitting an autoregressive model to the data. Since the data length is not determined by the order of the model it is clear that this method is less impaired by variation in the length of the breath cycle.

Yet another alternative is to seek higher resolution at the low frequency end of the spectrum. Oppenheim et al [8.7] describe an FFT based method for doing this, however care would need to be taken in smoothing a spectral estimate based on this technique.

So far no accurate estimate of the influence of crackles on breath sound spectra has been obtained. Recently robust spectral estimation methods have been described for excluding impulse-like interference from signals [8.8], and such a method would solve this problem

In section 6.4.1 the problem of differing lengths of breath cycle were discussed. A rather different approach to pattern recogntion of lung sound would be to match patterns of different length using dynamic time warping [8.9] which is widely used in speech recognition where problems arise because of words being spoken at different rates. Itakura [8.10] has shown that linear predictive coding of data matches well with dynamic time warping and Montpetit [8.11] has already demonstrated linear predictive coding of lung sound.

B1 Asbestosis Study

A larger scale recording programme for studying breath sound in asbestosis has already been completed. This study involves the recordings of over 30 individuals with a history of asbestos exposure. The aim is to extend work done in the preliminary study and gain further insight into the potential use of breath sound in the diagnosis of asbestosis.

B2 Studies of Other Diseases

It would obviously be interesting to know whether and how breath sound changes in a number of diseases. For example it would be interesting to compare asbestosis with other occupational lung diseases. Sarcoidosis while having a broadly similar radiological pattern to asbestosis is usually associated with fewer crackles, hence it would be interesting to see if there were any differences between the diseases in terms of breath sound as well.

B3 Possible Relationship with Lung Volume

It has been suggested that alteration in breath sound in asbestosis is caused by a change in filtration of the sound passing through the lung. Since asbestosis is a restrictive defect, late inspiratory breath sound is likely to be associated with smaller lung

volumes than with normals. Therefore it would be interesting to study spectra at different stages in the breath cycle and see if there is any correlation with lung volume. If such a relationship existed it might in part account for the spectral differences observed in asbestosis.

Although this experiment appears to be physiological in its aims, it is felt that it would be a useful test of the diagnostic value of breath sound in asbestosis. If there is a strong correlation between breath sound spectra and lung volume or transpulmoary pressure, there is little point in measuring breath sound since the same information could be obtained from pulmonary function tests. The experiment would probably have to be based on breathing manoevres similar to those used by Nath & Capel [8.12].

B4 Bronchial Breathing

Until recently bronchial breathing was the only recognised case of breath sound changing with respiratory disease. However virtually no work has been done to quantify these changes. It would be interesting to examine changes in breath sound during the development and disappearance of bronchial breathing.

References

8.1 C.T.Zahn, Graph-theoretical methods for detecting and describing Gestalt clusters, IEEE Trans. Comput., C-20, 68-86 (1971)

8.2 R.Dubes & A.K.Jain, Clustering techniques - the user's dilemma, Pattern Recognition, 8, 247-260 (1976)

8.3 D.W.Matula & R.R.Sokal, Properties of the Gabriel graph relevant to geographical variation analysis and the clustering of points in the plane, Geogrl. Anal., 12, 205-222 (1980)

8.4 A.C.Sanderson, J.Segen & E.Richey, Hierarchical modelling on EEG signals, IEEE Trans. Patt. Anal. & Machine Intelleg., PAMI-2, 405-415 (1980)

8.5 L.R.Rabiner, R.W.Schafer & CM.Rader, The chirp z-transform algorithm, IEEE Trans Audio & Electroacoust., AU-17, 86-92 (1969)

8.6 J.P.Burg, Maximum entropy spectral analysis, Proc. 37th Meeting Society of Exploration Geophysicists, Oklahoma City (1967)

8.7 A.Oppenheim & D.Johnson, Computation of spectra with unequal resolution using the fast Fourier transform, Proc. IEEE, 59, 299-301 (1971)

8.8 R.D.Martin & D.J.Thomson, Robust-resistant spectral estimation, Proc. IEEE, 70, 1097-1115 (1982)

8.9 H.Sakoe & S.Chiba, Dynamic programming algorithm optimization for spoken word recognition, IEEE Trans., ASSP-26, 43-49 (1978)

8.10 F.Itakura, Minimum prediction residual principle applied to speech recogntition, IEEE Trans., ASSP-23, 67-72 (1975)

8.11 J.-M.Montpetit, Personal communication

8.12 A.R.Nath & L.H.Capel, Inspiratory crackles and the mechanical events of breathing, Thorax, 29, 695-698 (1974)

APPENDIX 1

PROPOSAL FOR A NEW NONLINEAR MAPPING ALGORITHM

Mapping algorithms are commonly used in exploratory data analysis. Such methods may be either linear or nonlinear. In turn the nonlinear methods may be either iterative or noniterative. Here we propose a new method of optimizing iterative nonlinear mappings. The term 'nonlinear mapping' (NLM) is normally reserved for techniques similar to that of Sammon [A1.1], whereas the term 'multidimensional scaling' (MDS) is usually restricted to methods following Shepard [A1.2,A1.3] and Kruskal [A1.4,A1.5].

Let us consider a point set P in h-dimensional space ($h > 2$), we require to map the point set into a 2-dimensional configuration which in some sense preserves the structure of the data. Let d^*_{ij} denote the distance between points p_i and p_j in the h-dimensional space and let d_{ij} denote the corresponding 2-dimensional distance.

In designing a nonlinear mapping algorithm four questions are important [A1.6]:

(1) Which distance or dissimilarity measure should be used ?

(2) How do we choose an error function E ?

(3) How do we obtain an initial configuration prior to optimization ?

(4) When do we stop the optimization ?

Here we are only concerned with the choice of E since that determines the approach to optimization. Chien [A1.6] has shown that with some data sets the choice of a good initial configuration can improve the convergence rate of the mapping.

The 2-dimensional representation is obtained by minimizing the error function E. Chien [A1.6] lists a number of different error functions which emphasize different aspects of data structure.

These functions are all in the form

$E = f(e_{ij})$ for all $i,j = 1,\ldots,N$ $i \neq j$ and where $e_{ij} = (d^{*}_{ij} - d_{ij})$

E is therefore a function of 2N variables (the coordinates of the 2-dimensional representation) and contains N(N-1)/2 terms.

A major problem in optimizing E is computation time. Chang & Lee [A1.7] describe a heuristic method for reducing computation time known as the 'frame' algorithm. This chooses a frame of M points where $M \ll N$ then maps the data in two stages. Initially a 2-dimensional representation is found for the frame, then the remaining (N-M) points are positioned using the frame as a reference. They also propose optimizing the error function one term at a time by the 'relaxation method'. This results in a considerable reduction in computation time but at a cost of less accurate representation.

The disadvantage of the frame algorithms is that the reduced set of terms that are optimized is not chosen systematically. This results in a loss of structural information if the choice of frame is poor.

Consider a generalization of the error function

$E = f(e_{ij})$ for all i,j where $(p_i,p_j) \in G$

where G is a set of point pairs. Each of the two stages of the frame algorithm fit into this framework where only the inter-relationships determined by the frame are listed in G for each stage. In fact we may regard G as the edge set of a graph. The expression of E used by Sammon [A1.1] is then based on the complete graph K_N. The description of the error function in terms of a graph immediately suggests an alternative approach to error functions which is outlined below.

It is to be assumed that to some extent users are interested in local data. In other words the configuration of a particular point with respect to its neighbours is more important that its relationship with individual points in distant clusters. For example in Fig A1.1 we are more likely to be interested in the relationship of b to d than that of b to e. So let us assume that we require a point to be well represented in relation to its immediate neighbours at a low level. Following Terekhina [A1.8], Chien [A1.6] lists a number of error functions including ones that emphasize local structure. However local structure is interpreted simply as short distances, and all interpoint distances are included in the error function. This interpretation neglects important inter-cluster relationships (e.g. in Fig A1.1 the relationship of c to e is an important 'local' relationship even although d_{ce} is relatively large).

Thus a good approach to optimizing an NLM would seem to be to combine the local and graph approaches by choosing a suitable graph. Two graphs are obvious candidates – the Gabriel graph and the relative neighbourhood graph. Both these graphs extract a local structure from an h-dimensional data set, but the edges are not resticted to just short distances. If we choose a suitable graph the relaxation method can then be used to find a 2-dimensional configuartion.

We could simply use the RNG or GG directly. The expected number of edges in an h-dimensional GG is 2^h. The expected number of edges in an h-dimensional RNG is not known but will be substantially less. A useful constraint is that a point in h-dimensional space may be uniquely specified in terms of its distance to (h+1) neighbours.

We should then seek an 'augmented' RNG, RNG^+ that satisfied that constraint for each point. Therefore the problem of finding a new

mapping will partly be that of suitably augmenting the RNG. It is fairly easy to top up the RNG with edges so that each point has (h+1) or more neighbours, however if there is a cluster structure to the data a cluster might be connected to the rest of the data by only one or two RNG edges. Hence we might also want to find the major clusters in the data first (just a few!) and then sure that each cluster is connected to (h+1) neighbouring clusters. Another approach to finding a 2-dimensional representation is to ensure that the augmented graph connects a point to at least 3 neighbours to yield a unique 2-dimensional configuration.

Should this approach succeed there might also be an application of the RNG to the MDS of a point set. Furthermore the graph might be the key to 'recursive applicability' of the mapping [A1.6].

References

A1.1 J.W.Sammon, A nonlinear mapping for data structure analysis, IEEE Trans. Comput., C-18, 401-409 (1969)

A1.2 R.N.Shepard, The analysis of proximities: multi-dimensional scaling with an unknown distance function I, Psychometrika, 27, 125-140 (1962)

A1.3 R.N.Shepard, The analysis of proximities: multi-dimensional scaling with an unknown distance function II, Psychometrika, 27, 219-246 (1962)

A1.4 J.B.Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, Psychometrika, 29, 1-27 (1964)

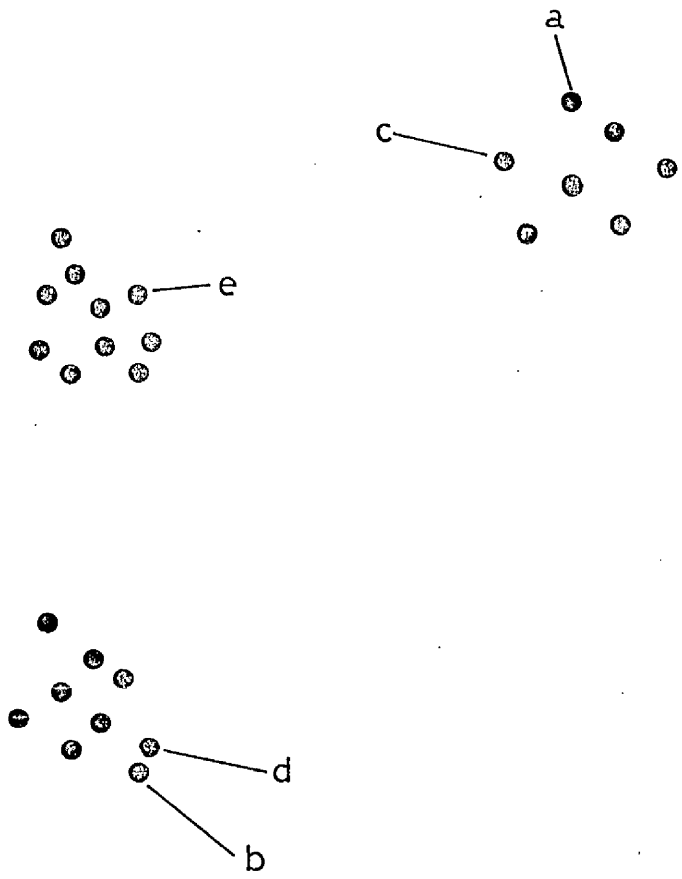
A1.5 J.B.Kruskal, Nonmetric multidimensional scaling - a numerical method, Psychometrika, 29, 115-129 (1964)

A1.6 Y.T.Chien, Interactive Pattern Recogntion, Marcel Dekker (1978)

A1.7 C.L.Chang & R.C.T.Lee, A heuristic relaxation method for nonlinear mapping in cluster analysis, IEEE Trans. Systems, Man & Cybernet., SMC-3, 197-200 (1973)

A1.8 A.Y.Terekhina, Methods of multidimensional scaling and visualization - a survey, Automation and Remote Control, 34, 1109-1121 (1973)

FIG A1.1



APPENDIX 2

LIST OF PUBLISHED WORK

1. R.B.Urquhart, Algorithms for computation of relative neighbourhood graph, Electronics Letters, 14, 556-557 (1980)

2. R.B.Urquhart, J.McGhee, J.E.S.Macleod, S.W.Banham & F.Moran, The diagnostic value of pulmonary sounds: a preliminary study by computer-aided analysis, Computers in Biology & Medicine, 11, 129-139 (1981)

3. R.B.Urquhart & J.E.S.Macleod, Interactive hierarchic clustering: a display based on graph theoretical methods, Proc. International Conference on Cybernetics & Society, Atlanta, USA, 11-15 (1981)

4. R.B.Urquhart, Graph theoretical clustering based on limited neighbourhood sets, Pattern Recognition, 15, 173-187 (1982)

5. S.W.Banham, R.B.Urquhart, J.E.S.Macleod & F.Moran, Alteration in the low frequency lung sounds in respiratory disorders associated with crackles, in press European Journal of Respiratory Disease (1983)