ACOUSTIC TUBE SHAPE RECOVERY WITH SPECIFIC APPLICATION TO SPEECH ANALYSIS

GREGORY JOHN BIELBY, B.E. (HONS.), B.Sc.

Being a thesis submitted

for the

DEGREE OF DOCTOR OF PHILOSOPHY

in the

DEPARTMENT OF ELECTRICAL AND

ELECTRONIC ENGINEERING

THE UNIVERSITY OF ADELAIDE

AUGUST 1983

THIS THESIS IS DEDICATED TO MY PARENTS
FOR THEIR UNDYING LOVE, ENCOURAGEMENT,
PATIENCE AND SUPPORT.

TABLE OF CONTENTS

CHAPTER 5 (cont.)

ABSTRACT

A lossless acoustic tube model and linear predictive analyses are reviewed, which represent the conventional model and analysis procedure for vocal tract shape recovery from the speech waveform. The effects of a non-white excitation on acoustic tube shapes recovered by linear prediction are investigated, and result in the definition of special acoustic tube shapes which can be recovered for non-white excitations of certain durations.

Conventional pre-emphasis techniques for removing glottal pulse excitation effects from the acoustic tube shape recovered by linear prediction are evaluated and, in general, are shown to perform poorly. A new adaptive pre-emphasis filter is defined, and shown to produce improved acoustic tube/vocal tract shape recovery in comparison with conventional pre-emphases for both synthetic and real speech waveforms of five vowel sounds.

A lossy termination of the acoustic tube model is investigated, and two new analysis procedures are presented, one based on autocorrelation functions, and another based on a transfer function of the acoustic tube model. The lossy termination analysis based on autocorrelations is shown to be applicable to real time applications of speech analysis, and both analyses produce an improvement in acoustic tube/vocal tract shape recovery in comparison with conventional analyses, when a lossy termination is present.

The new adaptive pre-emphasis filter and the lossy termination analysis based on autocorrelations are combined into a single speech analysis procedure. This new speech analysis procedure is evaluated with synthetic and real speech for five vowel sounds, and is shown to produce improved acoustic tube/vocal tract shape recovery in comparison with existing techniques.

SUMMARY

Acoustic tube models are widely used as models of the human vocal tract, and permit the acoustic waveforms in the vocal tract to be specified from the geometry and properties of the vocal tract. However, the inverse problem of determining the geometry or properties of a set of acoustic tubes, which model the vocal tract, is complex, and not completely resolved today. This thesis investigates the inverse problem to improve the accuracy of recovering acoustic tube geometries from the radiated acoustic waveform.

A lossless acoustic tube model and linear predictive analyses are reviewed, and the relationship between the results of a linear predictive analysis and the lossless acoustic tube model is defined. The conditions that the acoustic tube model must satisfy for a predictive analysis of its output or radiated acoustic waveform to recover its shape are presented. Of these conditions, many are not satisfied in the human vocal tract or during the production of speech; therefore, a linear prediction of a speech waveform does not identify an acoustic tube model the shape of which is the same as the vocal tract. An area distance measure is defined to provide a quantitative measure of the similarity between an acoustic tube shape recovered by an analysis of the output waveform from a set of acoustic tubes and the shape of those acoustic tubes.

For linear prediction to recover the shape of a set of acoustic tubes, those acoustic tubes must be excited by a white excitation. The effects of a non-white excitation of acoustic tubes on the acoustic tube shape recovered by linear prediction are investigated, and show that poor acoustic tube shape recovery occurs. A class of acoustic tubes is determined which can be recovered after a linear predictive analysis for non-white excitations of restricted duration. The necessary post analysis to recover the acoustic tube shape is presented, and the areas of application are discussed.

A new adaptive pre-emphasis filter is presented, the form of which is defined from measurements of the required pre-emphasis to whiten a range of glottal pulse waveforms, which are representative of the glottal pulse waveforms used to excite the vocal tract during the production of voiced speech. Evaluation of the new pre-emphasis filter with real and synthetic speech waveforms shows that improved acoustic tube shape recovery is achieved in comparison with the acoustic tube shapes recovered by previously used pre-emphasis techniques. The evaluations consider a wide range of sampling frequencies, and show a consistent improvement in acoustic tube shape recovery.

Conventional acoustic tube shape recovery by linear prediction requires a lossless termination of the acoustic tubes, which is not the case for the vocal tract. The effects on the acoustic tube shape recovered by linear prediction for a lossy termination of acoustic tubes are presented, and are shown to be significant. A general model of the loss at the termination of acoustic tubes due to radiation is reviewed, and simplified by the known conditions which exist at the lips, i.e. the termination of the vocal tract.

A number of autocorrelation analysis procedures for a lossy termination of acoustic tubes are defined, and evaluated with synthetic and real speech waveforms. An analysis procedure is also presented which is derived from a transfer function of the acoustic tube model and uses constraints, based on physical restrictions in the vocal tract, to overcome an ambiguity problem. The autocorrelation analyses are shown to be applicable to a wider range of situations than the transfer function analysis, especially for real time applications. These new analysis methods are shown to produce improved acoustic tube/vocal tract shape recovery in comparison with conventional analyses, when a lossy termination is present.

The new adaptive pre-emphasis filter and a lossy termination analysis are combined into a single speech analysis procedure. This new speech analysis procedure is evaluated with real and synthetic speech waveforms of five vowel sounds, and is shown to produce improved vocal tract/acoustic tube shape recovery in comparison with existing analysis techniques.

STATEMENT OF ORIGINALITY

This thesis contains no material which has been accepted for the award of any other degree or diploma in any University, and to the best of the author's knowledge and belief contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

G.J. BIELBY

ACKNOWLEDGEMENTS

Over the five years it has taken to perform the research presented in this thesis, I have been helped by many different people at an academic, technical and personal level, and I am in debt to them all.

I am greatly indebted to Professor R.E. BOGNER for his guidance and academic help in my first years of research, and his then allowing me to develop my research skills to a point of self-sufficiency. I have indeed been fortunate to have been supervised by a person of the highest calibre, academically and personally, who is able to open up new paths, from generalizing aquired scientific facts to disclosing truths which others have not perceived. I would like to thank Dr. B.R. DAVIS for his help and guidance during the year that Professor R.E. BOGNER was on study leave.

There have been many fellow post graduates who have helped me over the years, but none more than Mr. D.S. FENSOM. He has always been interested and willing to help with difficulties and, where possible, find a solution or a line of attach which has been fruitful. His honest opinions of many of my ideas and methods are also greatly appreciated, and I am especially indebted to DAVID FENSOM for supplying the real speech data. I acknowledge Dr. N.I. SMITH, Dr. T.K. KO, Dr. K. ESHRAGHIAN, Mr. P. BUDIMIR, Mr. M. LIEBELT, Mr. N. WOTTON, Mr. M. RAYMOND, Dr. R.J. CLARKE, Dr. A. ROY, Mr. R.C. BRYANT, Mr. J. NOONAN, Mr. B. COWLEY, Mr. N. MARTIN, and Mr. A. DAY for their moral support and their making life interesting and bearable as a post graduate student.

Special thanks must go to Mr. R.C. NASH, who managed to keep a troublesome computing system running long enough for me to finish my research work. Thanks are also directed to the staff and ancillary staff of the Department of Electrical and Electronic Engineering for all the help they have given.

I am also much indebted to ETTA CLARK for typing the thesis manuscript, and for using her expert typing skills to make this task relatively simple for me. The help of ELIZABETH HENDERSON with the production of some of the figures and tables is also acknowledged.

Finally, but most importantly, the research work presented in this thesis would never have been started or completed without the love, encouragement, support and patience of my parents, brothers and their families. Although they did not understand what I was doing and why, they have provided the greatest contribution to this thesis.

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

The mathematical modelling of acoustic waveforms in general acoustic tubes is described by NEWTON's second law of motion, the second law of thermodynamics, and the equation of continuity or conservation of mass. Such a description enables, from the properties of a set of acoustic tubes and the source excitation, a complete model of the resulting acoustic waveforms and their characteristics. The inverse problem of determing the geometry or properties of a set of acoustic tubes from their radiated or output acoustic waveform is more complex, and is not completely resolved today. This thesis investigates the inverse problem to improve the accuracy of recovering the acoustic tube geometry.

A major application of acoustic tube models is to the modelling of the human vocal tract. This has proved to be very successful, due to the human vocal tract being an acoustic tube which is excited by acoustic waveforms to produce speech, a radiated acoustic waveform. Acoustic tube models were first used to model speech production by CHIBA and KAJIYAMA in 1941, and the research up until the early sixties [DUNN 1950; STEVENS, KASOWSKI and FANT 1953; FANT 1960; KELLY and LOCHBAUM 1962] concentrated on acoustic tube modelling of speech production. The major result of this research was to show that the shape of an acoustic cavity determined the frequency charateristics of the output or radiated waveform, under idealized conditions.

As the understanding of the speech production mechanism matured with the use of acoustic tube modelling, the inverse problem was investigated. The initial studies of the inverse problem can be traced to MERMELSTEIN and SCHROEDER [1965] and MERMELSTEIN [1967], who used the relationships developed between the frequency characteristics of the output acoustic waveform and the acoustic cavity shape. More specifically, the cavity shape was determined by perturbation of formant frequencies of nearly uniform acoustic tubes. This work was extended by HEINZ [1967] and SCHROEDER [1967], but had many inherent disadvantages. The main disadvantages were that convergence of the perturbation process could not be guaranteed, and the resultant acoustic tube shape depended heavily on the boundary conditions at the source and the length of the acoustic cavity.

Another approach to the inverse problem which has excellent potential to recover the human vocal tract shape is the measurement of the impulse response of the vocal tract at the lips. The studies in this area were initiated by SCHROEDER [1967], and a major paper was presented in 1971 by SONDHI and GOPINATH [1971]. The impulse response of the vocal tract was measured as the pressure developed at the lips to a unit impulse of volume velocity applied to the lips. Measurements were taken with the subject in a phonating position with an impedance measuring tube at the lips, where the excitation was developed and pressure measurements taken.

Recovering the vocal tract shape by measuring the impulse response at the lips overcomes the speech analysis problems of an unknown excitation of the vocal tract and complex radiation of sound at the lips. The major disadvantage of this technique is that the measuring apparatus does not allow a natural speaking environment.

It is not applicable to analysis of speech sounds, and so its application is basically limited to an alternative for determining vocal tract shapes via X-ray photography. The disadvantages outweigh the advantages, and so little work has been completed in this area, although SONDHI [1977 and 1979] has argued that this is the only procedure for accurately determining the vocal tract shape from acoustic signals.

Up until the late sixties, the speech analysis problem was performed via frequency or spectral analysis, but a number of restrictions were encountered. One of these restrictions was the necessity for relatively long speech segments to provide the required spectral resolution, and hence rapidly changing speech events could not be followed. The periodic nature of voiced speech masks the spectral information between pitch harmonics, and causes unsatifactory results for high pitched voices, which are commonly encountered in women and children. Many solutions to these and other problems have been presented, with the most successful being a direct analysis of the speech waveforem.

Of the early techniques used for direct analysis of the speech waveform, the most significant were those of SAITO and ITAKURA [1968], and ATAL and SCHROEDER [1967, 1968a, 1968b]. The papers by ATAL and SCHROEDER were concerned with predictive coding of the speech waveform, while a maximum likelihood approach was presented by ITAKURA and SAITO. These papers can be recognised as the beginnings of the linear prediction analysis procedure which is now widely used for speech analysis. The term "linear prediction" was first used by ATAL [1970a] in 1970, and his paper of the same year

[ATAL 1970b] presented the first attempt at directly computing an acoustic tube model of the vocal tract from the speech waveforem.

The results of ATAL [1970b] and, soon after, ATAL and HANAUER [1971] proved to be very important in cementing the bond between linear prediction and the acoustic tube model. They showed that a transfer function of M poles is always realizable as the transfer function of a set of M commensurated cylindrical acoustic tubes. Thus, a unique discrete acoustic tube shape can be constructed from a given order transfer function polynomial. The important relationship, that the inverse filter obtained by linear prediction of speech waveforms is an equivalent representation of the acoustic tube model, was presented by WAKITA [1971] soon after. Other significant studies around that time which lead to different formulations of linear prediction were presented by WAKITA [1973], WAKITA and GRAY [1974], MAKHOUL and WOLF [1972], ITAKURA and SAITO [1972], ITAKURA et al [1972], GRAY and MARKEL [1974], and NARASIMHA [1974].

The object of this thesis is to produce improved acoustic tube shape recovery from the radiated or output acoustic waveform of a set of acoustic tubes. The excitation, radiation conditions, and general acoustic tube parameters considered are similar to those in acoustic tube models of the human vocal tract. Therefore, a direct application of the improved acoustic tube shape recovery procedures developed in this thesis to speech analysis is possible, to provide improved area function recovery of the vocal tract.

The linear predictive analysis of speech has been highly successful, mainly due to the simplicity and ease with which acoustic tube parameters are recovered directly from the output acoustic waveform. For this reason, the majority of the work presented in this thesis uses linear predictive analysis techniques directly, or as a basis for developing new acoustic waveform analysis techniques.

1.2 THE PHYSICAL VOCAL TRACT AND SPEECH PRODUCTION

If an accurate model of the vocal tract is to be produced, then it is important that its structure be well defined and understood. This section provides a brief description of the vocal tract and the speech production process, sufficient to define the speech terms used throughout this thesis. A complete description of the vocal tract and the speech production process is found in the texts of FLANAGAN [1972] and FANT [1960].

A lateral mid-sagittal section of the human vocal tract, showing the positions of the major features, is presented in Figure 1.1. The vocal tract is defined to start at the opening of the vocal cords, i.e. the glottis, and extends up to the lips. The region between the glottis and the back of the mouth is called the pharynx, and contains the epiglottis, whose major function is to close off the air passage to the lungs during swallowing. The major changes in cross-sectional area of the human vocal tract occur in and around the mouth cavity, with the major articulators in this region being the lips, jaw, tongue and teeth. The amount of coupling of the nasal cavity or tract to the vocal tract is controlled by the velum. Speech sounds which are produced with the nasal tract acoustically coupled (i.e. the velum is open) to the vocal tract are called nasalized sounds.

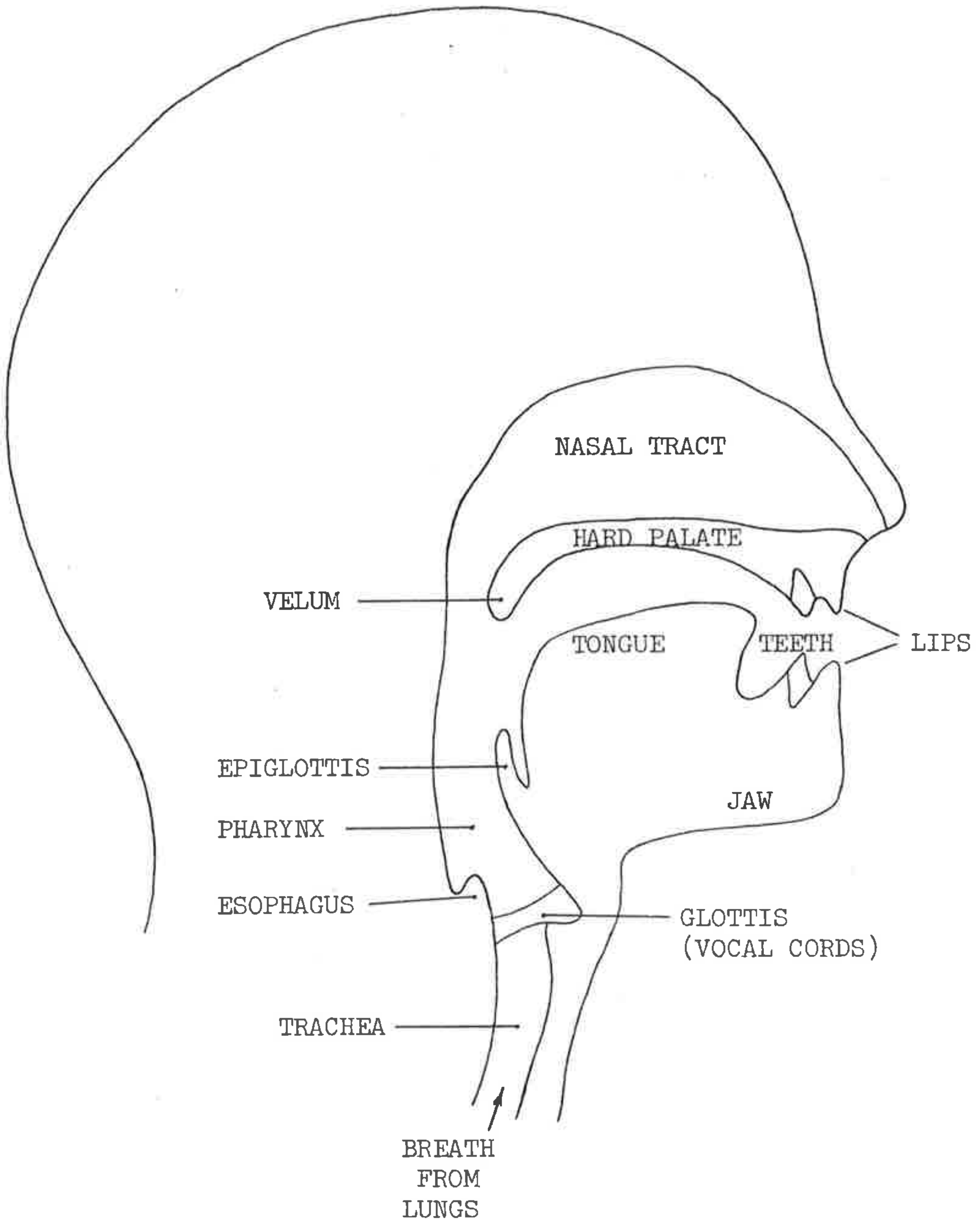


FIGURE 1.1: Lateral mid-sagittal section of the human vocal tract.

Speech sounds are produced by forcing air from the lungs past the glottis and through the vocal tract to the lips, where an acoustic pressure waveform is radiated. The sound produced depends upon the physical conditions at the glottis, the position of the articulators in the vocal tract, and the degree of nasal coupling. Speech sounds are classified into three basic groups, i.e. voiced, unvoiced and plosive sounds, by the manner in which the acoustic waveforms in the vocal tract are excited.

If the vocal folds are held tight, therefore closing the glottis, then pressure builds up in the sub-glottal regions during the exhalation process. Eventually, the pressure becomes sufficient to force open the vocal folds, opening the glottis, and allowing air flow into the vocal tract. This air flow causes a decrease in pressure below the glottis and in the glottal orifice. The tension on the vocal folds and this decrease in pressure causes the vocal folds to snap shut and stop the air flow into the vocal tract. This process is then repeated to provide a nearly periodic pulse excitation of the vocal tract. Sounds produced by this nearly periodic pulse excitation are called voiced sounds.

Excitation of the vocal tract for sounds called unvoiced is produced by turbulent air flow. For these sounds, the vocal folds are held open, and air is allowed to pass the glottis freely while a constriction is made in the vocal tract. This constriction causes a turbulent air flow around the constriction, giving rise to a noise like excitation of the vocal tract. In the case of unvoiced sounds, excitation therefore does not occur at the glottis, but at the point where the constriction of the vocal tract occurs.

Plosive sounds are generated by a complete closure of the vocal tract, causing a build-up of pressure, which is suddenly released. The sounds in this category can be produced with or without vocal fold vibration, and so voiced and unvoiced plosive sounds are possible. In general, the closure of the vocal tract occurs in the mouth region, e.g. at the lips, teeth or tongue.

A very important group of sounds is the vowels, due to their relatively large usage in spoken English. The vowels are non-nasalized voiced phonemes, and the study of speech sounds in this thesis is restricted to vowels. Therefore, the glottal excitations of the vocal tract investigated later in this thesis are of the near periodic glottal pulse type.

1.3 THE NEED FOR ACCURATE VOCAL TRACT SHAPE RECOVERY

One of the goals of speech research has been to accurately recover the shape of the human vocal tract from the speech sound. Many of the early speech waveform analysis techniques (ITAKURA and SAITO [1968], ATAL and SCHROEDER [1967, 1968a, 1968b], ATAL [1970a, 1970b], ATAL and HANAUER [1971], and WAKITA [1972]) produced results in the form of estimated vocal tract shapes. However, as speech research continued, it became obvious that most of the applications for speech research do not require recovery of the vocal tract shape, or even the use of a vocal tract model. Despite this, there still exists a need to recover vocal tract shapes accurately, and these needs can be grouped into the following areas:

a) The need to improve our understanding of speech production.

b) Diagnostic applications in medicine and aids for the speech handicapped.

c) Speech coding and synthesis.

Of the above three groups, the first is the most important, as it leads to a large number of applications, and has the potential to improve existing analysis/synthesis methods of speech. At present, there still exists a lack of knowledge about the position of the articulators when uttering sounds, and their movement when a string of phonemes is pronounced. The position of articulators was first determined by X-ray photography, FANT [1960], CHIBA and KAJIYAMA [1941], PERKEL [1965], and FUJIMURA et al [1968], but limitations on the amount of X-ray exposure that is safe for humans, and transparency of some vocal tract features, causes this procedure to be inadequate for determining vocal tract shapes over a wide range of speech sounds. Other methods, such as placing detectors in the mouth, have also been found to be inadequate. Hence, only elementary information about articulator positions during speech production is known.

Knowledge of articulator position would be helpful in speech recognition. This is due to the fact that articulator positions are largely speaker invariant, and have little redundancy, which is why phonetic descriptions of speech are based on articulator position. The area of speech synthesis could also benefit, since the knowledge of the constraints on the movements of articulators from one position to another may help intelligibility of the synthesized speech.

Increased understanding of speech production through accurate vocal tract shape recovery should permit improvements in speaker identification and verification. In the indentification process, verification of the true position of certain or all of the articulators increases the probability of a correct identification. Accurate vocal tract shape recovery should also permit determination of which articulators are important in an identification situation.

One medical application of accurate vocal tract shape recovery could be for assessment of incorrect articulator movement and deformations of the vocal tract. The latter would require a simple and portable system for it to be cost-effective. Incorrect articulator movement could be determined by display of vocal tract shapes while phonemes are spoken. If the articulator movement were restricted by incorrect growth of the articulator or absence of muscle etc., then medical procedures could be implemented to correct the situation.

Application of accurate vocal tract shape recovery to the speech handicapped is far greater than the medical application. The major group of speech handicapped are those who are deaf, and so cannot use the normal feedback path of hearing to learn and maintain speech quality. For these people, the visual feedback of vocal tract shapes could provide a means by which quality speech can be produced. Those people with speech impediments may also be helped by vocal tract shape display to show how and where certain articulators need to be placed to produce a certain sound.

In a typical system used for speech handicapped people, a target shape is displayed, showing the correct position of the articulators and, as the person speaks, the position of his articulators is displayed for comparison. The articulators in the correct position can be identified, and the appropriate corretion made to those articulators in incorrect positions. CRICHTON and FALLSIDE [1974] described such a system, and the need for an accurate vocal tract shape recovery procedure is obvious, not only for display of the person's vocal tract shape, but also for the production of the target shape.

Speech coding and synthesis, which are required in vocoders, is another application area where improvements may be made if accurate vocal tract shape recovery is available. The requirements of vocoders for a small number of slowly varying parameters could be satisfied by articulator position found from accurate vocal tract shape recovery. At the receiving end of the vocoder, if parameters modelling the glottal waveform and vocal tract shape parameters were available, then using a model of speech production for speech synthesis may provide improved synthesized speech quality.

The above discussions do not exhaust the applications which would benefit from accurate vocal tract shape recovery from a direct analysis of the speech waveform. However, those applications presented above are sufficient to show that there should be significant benefits to be gained in several areas if accurate vocal tract recovery were possible.

1.4 THE NEED FOR FURTHER INVESTIGATION

The previous section established the need for accurate models of the vocal tract so that accurate shapes can be recovered. This section briefly considers the available techniques for vocal tract shape recovery, and comments on the accuracy and limitations of these techniques, and hence establishes a need for further investigation.

The first method of finding vocal tract shapes and articulator positions was by X-rays. While this procedure is initially attractive, many disadvantages, such as the limited X-ray dosage a person can be given, restrict its usage. Problems are experienced with the articulators, which are made of fleshy materials and so are relatively transparent to X-rays, as well as articulators being masked by bones, skin, etc., making accurate shape determination difficult. Some of these problems can be overcome by coating the tract with materials to highlight the articulators, but difficulties still persist with their application and discomfort to the subject.

Tracing the movement of articulators with X-rays requires films to be made while the sound is phonated. Problems exist with synchronization of the film with the sound, noise due to the camera, and the lack of definition found in individual frames of film. Thus, studies of articulator movement using this procedure can be inaccurate, and are difficult to implement.

Since X-ray photographs only provide a one dimensional view of the vocal tract, accurate vocal tract cross-sectional area measurements still require additional procedures. Some of the methods used to obtain three dimensional shapes include mouth cavity measurements with palatograms (false plates similar to those used by dentists), and film recordings of the lip opening.

Despite all of these problems FANT [1960] managed to obtain a series of vocal tract area functions for twenty-three phonemes from a single speaker. This is the only large group of vocal tract cross-sectional area shapes available and, as a consequence, they are used as benchmarks for verifying new analysis techniques. In fact, they are used later in this thesis as comparisons for the recovered vocal tract shapes from spoken vowels. Other researchers using X-ray techniques to obtain vocal tract shapes have been CHIBA and KAJIYAMA [1941], PERKELL [1965], and FUJIMURA et al [1968], and a summary of their work is presented by FLANAGAN [1972].

A procedure which overcomes the problems of X-ray photographs is one based on the measurement of the driving point response at the lips. This method recovers the area function exactly under the plane wave assumption for a lossless vocal tract. Original studies in this area were performed by SCHROEDER [1967], who describes the apparatus used to measure the impedance at the lips in his paper (SCHROEDER [1967]). It consists of an impedance tube with a flexible coupler, which the subject places against the lips. Measurements are then made while the subject moves the vocal tract as in normal speech, but without phonating.

Theoretical advances on the problems of computing the area function from the poles and residues of the driving point impedance at the lips were made by GOPINATH and SONDHI [1970]. These advances overcame glottal source excitation, and the non-uniqueness problem of the recovered vocal tract shape. SONDHI and GOPINATH later defined a procedure which produces more accurate vocal tract shape recovery from the impulse response at the lips. The same apparatus is used, but the analysis takes place in the time domain rather than the frequency domain.

The apparatus used to measure lip impulse response removes any lip radiation, and the remaining losses in the vocal tract (i.e. due to wall motion, viscosity and heat conduction) cause only small perturbations in the recovered area function. SONDHI and GOPINATH [1971] showed that lip impulse response measurements cannot discover the presence of losses in the vocal tract, or correct for them. However, SONDHI and GOPINATH [1971] showed that, if the loss distribution is a series loss proportional to the reciprocal of the area and a shunt loss proportional to the area, then a correction can be made. SONDHI [1974] showed that this loss distribution approximates vocal tract losses, and so lip impulse response procedures can also provide a first order correction of vocal tract losses.

Hence, accurate vocal tract shape recovery can be obtained from measurements of the impulse response at the lips, with less accurate results for lip impedance measurements due to its inability to account for vocal tract losses. The problems with X-ray photography are eliminated, as well as the problems that time and frequency domain analysis of speech waveforms have with glottal excitation and lip radiation. Recent work with lip impedance and impulse response techniques has been performed by DESCOUT, TOUSIGNANT and LECOURS [1975] and SONDHI [1977, 1979].

There are significant disadvantages with the lip impulse response and lip impedance techniques, which prevent their application to general situations. These disadvantages revolve around the requirement of the measuring apparatus being at the lips of the speaker. Thus, an unnatural non-phonating environment is required for measurements, and the techniques cannot be applied to speech waveforms. Therefore, the application of these techniques is limited to special research investigations.

Recent advances in ultrasonic techniques provide another alternative to the harmful X-ray techniques. However, this method, and all those considered in this section, require controlled conditions and direct access to the speaker, and necessitate complex and time-consuming computations. Hence, none of these methods is applicable to the most common situation, where only the speech waveform is available. Thus, there exists a need for alternative methods which can analyse the speech waveform directly and obtain accurate area function recovery.

Direct analysis of the speech waveform requires a model of the speech production system, and the acoustic tube model mentioned earlier has been very successful. An accurate model of the vocal tract must take into account

a) time variation of the vocal tract shape,

b) excitation of the acoustic waveforms in the vocal tract,

c) radiation of the speech waveform from the lips,

d) losses produced by heat conduction and viscous friction at the vocal tract walls,

e) softness of the vocal tract walls,

f) branches of the vocal tract, e.g. the nasal cavity.

The assumptions of linear prediction, which is used to recover vocal tract shapes from the speech waveform, ignore all of the above properties of the vocal tract. Even so, it has been suggested that the acoustic tube shape that is recovered by linear prediction of the speech waveform is a reasonable approximation of the vocal tract shape. MARKEL and GRAY [1976] have concluded that, if speech is properly pre-emphasized, and if boundary conditions of the acoustic tube are properly chosen, then very reasonable vocal tract shapes can be estimated from the speech waveform by linear predictive procedures. However, SONDHI [1977, 1979] has disputed this, and claims that it is not possible to obtain accurate vocal tract shapes from the speech waveform.

The proper pre-emphasis referred to in the above conclusion by MARKEL and GRAY [1976] is to remove the effects of excitation of the vocal tract and radiation of speech at the lips. Pre-emphasis was originated by WAKITA [1973], who applied a constant +12 dB per octave spectral boost to the waveform to account for glottal excitation of the vocal tract. A constant -6 dB per octave spectral correction is also applied simultaneously, to correct for the radiation impedance at the lips, so that the nett pre-emphasis is a +6 dB per octave spectral correction.

Application of pre-emphasis to the speech waveform before analysis alters the recovered acoustic tube shape, as indicated by Figure 4.11 of MARKEL and GRAY [1976] (originally from WAKITA [1972]). However, the conclusion by MARKEL and GRAY that the figure shows that reasonable area function recovery results from pre-emphasis is disputed by SONDHI [1977, 1979]. SONDHI concludes that the figure shows that the area function recovered is very strongly dependent on the assumed source and radiation characteristics. The experimental work of ENGEBRETSON and VEMULA [1974] supports this conclusion, by showing that the glottal waveform changes markedly from one sound to another. Hence, inaccurate area function recovery still occurs, even after the constant +6 dB per octave pre-emphasis is applied.

Another problem which causes significant inaccuracies in the recovered vocal tract shape is the correct choice of boundary conditions for the acoustic tube model. The boundary conditions are the terminating conditions at the glottis, or source, and lip, or output, ends of the acoustic tube model. If linear predictive analysis is used, then the boundary conditions must be chosen firstly to make the acoustic tube and linear prediction models equivalent, and secondly to guarantee a unique vocal tract shape from the speech waveform. SONDHI [1977, 1979] and other researchers have shown that satisfying these requirements does not lead to accurate area function recovery.

Additional problem areas which also cause significant errors when accurate area function recovery is required are: losses within the vocal tract, uniqueness problems, and a lack of high frequency information. It is claimed by SONDHI [1977, 1979] that these and the above problems are only overcome by measurement of the impulse response at the lips. His conclusion is that, if all the poles and zeros of the input impedance at the lips are known for the closed glottis condition, then the vocal tract area function can be uniquely recovered for lossless plane wave propagation in the tract. However, it has already been argued that the lip impedance procedure is not applicable to analysis of the speech waveform.

Acceptable vocal tract shape recovery has been shown [PINSON 1967, STEIGLITZ et al 1977, ROGERS 1974, and MARKEL and GRAY 1976] for analysis of voiced sounds during the interval when the glottis is closed, i.e. when the vocal tract is not being excited. Closed glottis interval analysis does, however, have many disadvantages, which makes it an unreliable method for vocal tract shape recovery. A major disadvantage is the restriction of the analysis procedure to speech sounds where the glottis is closed, and research by ROSENBERG [1971] on glottal excitation has shown that glottal closure does not occur as a general rule, even for voiced sounds.

Closed glottis interval analysis requires special analysis procedures using covariances instead of autocorrelations, due to the small duration of the closed glottal interval, and these procedures do not guarantee a solution. Another problem with the analysis process is that, at present, there does not exist a method which accurately and reliably determines the closed glottis interval directly from the speech waveform [ANANTHAPADMANABHA and YEGNANARAYANA 1979]. Therefore, even though closed glottis interval analysis shows a potential for accurate vocal tract shape recovery, a number of fundamental and unsolved analysis problems prevent it from being an accurate and reliable process for all speech sounds.

The discussion presented in this sections leads to the conclusion that a procedure for accurate and reliable vocal tract shape recovery from the speech waveform did not exist, and so there is a need for further investigation. This thesis performs further investigations towards improved vocal tract shape recovery from the speech waveform.

1.5 THESIS ORGANIZATION

This chapter has provided a basic introduction to acoustic tube models and their application to human vocal tract modelling. A brief history has been presented of the analysis techniques which, along with the acoustic tube model, have been used to obtain vocal tract shapes directly from the speech waveform. A description of the human vocal tract and the speech production mechanism was presented, to define the speech terms that are used throughout this thesis.

The second half of this chapter investigated the need for accurate models of the vocal tract, and the necessity for further investigations to be performed. A detailed discussion of the application areas which require accurate vocal tract shape recovery from the speech waveform was presented, to establish the need for accurate models of the vocal tract. Techniques for vocal tract shape recovery both indirectly and directly from the speech waveform were discussed, and their inadequacies presented. This discussion showed that there did not exist a procedure which could accurately and reliably recover vocal tract shapes directly from the speech waveform, hence establishing the need for the further investigations performed in this thesis.

Chapter 2 reviews acoustic tube modelling and acoustic waveform analysis by linear prediction. A detailed description and definition of acoustic tubes results in the definition of a mathematical waveform model of single acoustic tubes, junctions between acoustic tubes, and a concatenation of commensurate acoustic tubes. Boundary conditions for a concatenation of commensurate acoustic tubes are discussed and defined for an acoustic tube model of the vocal tract.

A review of linear prediction analysis procedures and their application to analysis of acoustic waveforms generated from commensurate acoustic concatenations is presented in Chapter 2. After an introduction of the basic principles of linear prediction, the covariance, autocorrelation and lattice formulations are presented in detail. This is followed by the definition of the necessary conditions of an acoustic tube concatenation for linear prediction to recover its shape from the acoustic waveform at its termination. The relationship that exists between linear prediction and the analysis of acoustic tubes is then presented in both the theoretical and the practical sense. The last section of Chapter 2 defines an area distance measure which is used throughout this thesis to measure the difference or similarity between the recovered acoustic tube shape and the correct one.

Chapter 3 examines the effects of a non-white excitation of commensurate acoustic tubes on the acoustic tube shape recovered by linear predictive procedures. Investigations of these effects permit a detailed understanding of the manner in which an autocorrelation linear prediction procedure performs an analysis in the autocorrelation domain. This results in an understanding of the effects of non-white excitation on acoustic tube shape recovery. The latter half of Chapter 3 defines and investigates special acoustic tube shapes which can be recovered by a linear prediction analysis, and some post-analysis processing when certain types of non-white excitations have been used. The areas of application for these special acoustic tube shapes are also presented.

The non-white excitations investigated in Chapter 3 have, in general, an arbitrary shape, which contrasts with voiced speech sounds where the glottal excitation of the vocal tract has a pulse shape. Chapter 4 utilizes this basic knowledge of the form of the glottal excitation of voiced sounds to remove the excitation pulse from the speech waveform and allow linear prediction to achieve improved acoustic tube shape recovery. Initially, Chapter 4 investigates the effects of glottal pulse excitation on acoustic tube shape recovery by linear prediction. This is followed by an evaluation of the effectiveness of conventional methods to remove glottal pulse excitation from the speech wave. A new adaptive pre-emphasis filter is then defined, to include the advantages and successful features of conventional pre-emphasis filters, but overcome some of their disadvantages and deficiencies.

The new adaptive pre-emphasis filter defined in Chapter 4 is evaluated in Chapter 5 with glottal pulse waveforms, and real and synthetic speech waveforms. The evaluation procedure compares the recovered acoustic tube shape after using the new adaptive pre-emphasis and conventional pre-emphasis techniques with the original acoustic tube shape. Therefore, not only is the performance of the new adaptive pre-emphasis filter presented, but also the performance of conventional pre-emphasis filters to account for glottal pulse excitation. The evaluations presented in Chapter 5 consider a wide range of different sampling frequencies.

Another necessary condition for acoustic tube shapes to be recovered by linear prediction is that the set of acoustic tubes must have a lossless termination. In reality, a loss occurs at the termination of a set of acoustic tubes, and Chapter 6 investigates the effects of a lossy termination on acoustic tube shape recovery, and develops new analysis procedures for improved acoustic tube shape recovery. The first part of Chapter 6 defines a general model for realistic termination conditions which apply for the speech application. Using this lossy termination model, an evaluation of the effectiveness of conventional linear prediction to recover acoustic tube shapes is performed.

After stating the necessary requirements for an analysis procedure to include a lossy termination, Chapter 6 presents a number of autocorrelation analysis procedures which take into account a lossy termination of the acoustic tubes. Each of the autocorrelation methods is evaluated with synthetic data from an acoustic tube model with a lossy termination, and the autocorrelation method which provides the best improvement in acoustic tube shape recovery is evaluated with real speech data.

An analysis procedure which takes into account a lossy termination of acoustic tubes via a transfer function of the acoustic tube model is presented in Chapter 6. This analysis procedure forces physical vocal tract shape constraints onto the recovered acoustic tube shape, therefore reducing the problem of non-uniqueness and ensuring a realistic vocal tract shape is recovered. To recover an acoustic tube shape, the analysis procedure must solve a set of non-linear simultaneous equations, and the problems that this creates are discussed.

Chapter 7 combines the new adaptive pre-emphasis filter defined in Chapter 4, to account for glottal pulse excitation, with the autocorrelation analysis procedure defined in Chapter 6, to account for a lossy termination of acoustic tubes, into a single analysis procedure. The performance of the analysis procedure to provide improved acoustic tube/vocal tract shape recovery is evaluated using synthetic and real speech waveforms.

A set of conclusions for the various topics considered in this thesis is presented in Chapter 8.

CHAPTER 2

ACOUSTIC TUBE MODEL AND LINEAR PREDICTION

2.1 INTRODUCTION

The purpose of this chapter is to review and define the previously known acoustic tube model and linear predictive analysis procedures which form the basis of the work presented in this thesis. The acoustic tube model defined in this chapter is designed to be a valid model of the vocal tract shape and suitable for analysis by linear predictive procedures. Initially, a single acoustic tube is defined physically and mathematically in terms of its acoustic time waveforms. This is followed by a mathematical description of the acoustic time waveforms at the junction of two adjacent acoustic tubes, which leads to a complete description of a set of acoustic tubes.

A number of restrictions are placed on the generality of the acoustic tubes, an important one being on the length of each acoustic tube. Since all the analysis procedures in this thesis assume sampled time waveforms, then the length of each acoustic tube is fixed to a multiple of the sampling period. Hence, the sets of acoustic tubes considered in this thesis are commensurate. Boundary conditions for the acoustic tubes are defined by consideration of the requirements of linear prediction to recover the acoustic tube shape from the output or radiated acoustic waveform.

Numerous analysis procedures have been developed over the years to recover the characteristics, e.g. the shape, of acoustic tubes. Since, in the majority of practical applications, e.g. speech analysis, only the radiated waveform is available, those analysis procedures which recover the acoustic tube characteristics directly from the radiated waveform are the most widely used. One analysis procedure fitting into this category is LINEAR PREDICTION, which precisely recovers idealized acoustic tube shapes from the radiated waveform.

The origins of linear prediction have been traced by SORENSON [1970] to GAUSS in 1795 and LEGENDRE in 1806, who independently invented the linear least squares estimation, or prediction. Since that time, many formulations of linear prediction have been produced, with the term "linear prediction" being introduced by WEINER [1949]. Recent interest in linear prediction can be attributed to SAITO and ITAKURA [1966], with their maximum likelihood formulation, as applied to speech analysis. Since then, the development of linear prediction in association with speech analysis has led to linear prediction becoming a very powerful general analysis tool.

The strong association between speech analysis and linear prediction has led to linear prediction being used as a benchmark for evaluating new speech analysis procedures, in many situations. One of these situations is vocal tract shape recovery, and linear predictive procedures are used to evaluate the effectiveness of the new vocal tract shape recovery procedures that are developed in this thesis. Also, some of the new analysis procedures use linear prediction as a starting point for their development.

Hence, the definitions and discussions of linear predictive techniques presented in this chapter are important to the remainder of this thesis.

After defining linear prediction, this chapter presents a general formulation from basic principles. During the development of linear prediction and its applications to various areas, a number of simplifications of the general formulation lead to numerous forms of the linear prediction process. All these different forms can be grouped into three basic formualtions, and the details of these are also presented in this chapter. A general comparison of the three basic formulations is presented, to show the advantages and disadvantages of each, and which formulations are best suited to certain situations. This is followed by a definition of the excitation requirements that must be satisfied by an all pole system if linear prediction is to identify it from its output waveform.

Following the definition and presentation of the linear prediction procedures and acoustic tube models, the relationship between the two is presented. This shows the manner in which linear prediction recovers the acoustic tube parameters from its output, or radiated acoustic time waveform. In practice, not all the assumptions and requirements of linear prediction and the acoustic tube model are satisfied, and so incorrect acoustic tube model parameter recovery, e.g. its shape, often occurs. The last section of this chapter defines and investigates the performance of area distance measures, to provide an indication of the difference between the recovered and original acoustic tube shapes.

2.2 THE ACOUSTIC TUBE MODEL

The theory of acoustics which permits a description of time waveforms within a single acoustic tube is well-known, and is found in the many classical acoustics texts, e.g. RAYLEIGH [1926], STEWART and LINDSAY [1930], MORSE [1948], KINSLER and FREY [1950], BERANEK [1954], and MORSE and INGARD [1968]. Using the basic equations of acoustics produces the one dimensional wave equation, or the Webster horn equation (WEBSTER 1919), which are equivalent if the cross-sectional area of the acoustic tube is constant. From the one dimensional wave equation, a model of a set of acoustic tubes can be derived.

This section initially defines the properties of the acoustic tubes that the basic acoustic equations assume, as well as those that provide tractable mathematics. Using these properties, a complete acoustic waveform description of a single acoustic tube is derived from the one dimensional wave equation. A waveform description at the junction of two acoustic tubes is then produced by considering waveform continuity requirements. Following this, the necessary boundary conditions are defined for a set of commensurate acoustic tubes. Finally, this section combines the waveform descriptions of a single acoustic tube and the junction between acoustic tubes with the boundary conditions, to produce an acoustic tube model. In this combination step, a number of simplifications are made to the waveform description, to provide a relatively simple mathematical description of the set of commensurate acoustic tubes.

2.2.1 DEFINING THE ACOUSTIC TUBES

The acoustic tubes used in this chapter have basic restrictions placed on their physical and acoustic properties, in order to provide a tractable mathematical analysis, and to fit the analysis conditions under which the acoustic tube model is used. This section presents and explains the required physical and acoustic properties of a set of acoustic tubes.

The acoustic tubes used in this thesis have a physical length, ℓ, and a constant cross-sectional area over that length. An important quantity of the acoustic tube is the round trip propagation delay, i.e. the time required for an acoustic signal to propagate from one end of the acoustic tube to the other and back again, which is equal to the division of 2ℓ by the velocity of the acoustic waveform in the acoustic tube. All the analysis procedures used in this thesis assume that the acoustic waveforms are sampled at equal intervals of time, which is equal to the round trip propagation delay of a single acoustic tube.

The necessity for tractable mathematical analysis of the acoustic waveforms within an acoustic tube requires that only plane wave propagation occurs. Hence, the transverse dimensions of any acoustic tube must be considerably less than the smallest wavelength of the acoustic waveforms propagating in the acoustic tube. In practical situations, the plane wave propagation requirement is satisfied by low pass filtering the acoustic waveform before analysis.

Completely lossless acoustic tubes are required so that there is no heat conduction through the walls of the acoustic tube, no viscous friction between the medium in the acoustic tube and the acoustic tube walls, and the acoustic tube walls are rigid so that there is no loss due to acoustic tube wall vibration. The medium within the acoustic tube must also be lossless. In practice, the lossless acoustic tube assumptions are violated in varying degrees.

All the acoustic tube dimensions and properties are assumed to be time invariant, or at least so over the time interval during which a waveform analysis is performed. In terms of the acoustic waveforms, this time invariance translates to the requirements of waveform stationarity, i.e. the waveform properties, such as autocorrelation, are independent of the time origin. For the majority of practical applications, waveform stationarity is ensured by choosing a suitable time period in which an analysis is to be performed.

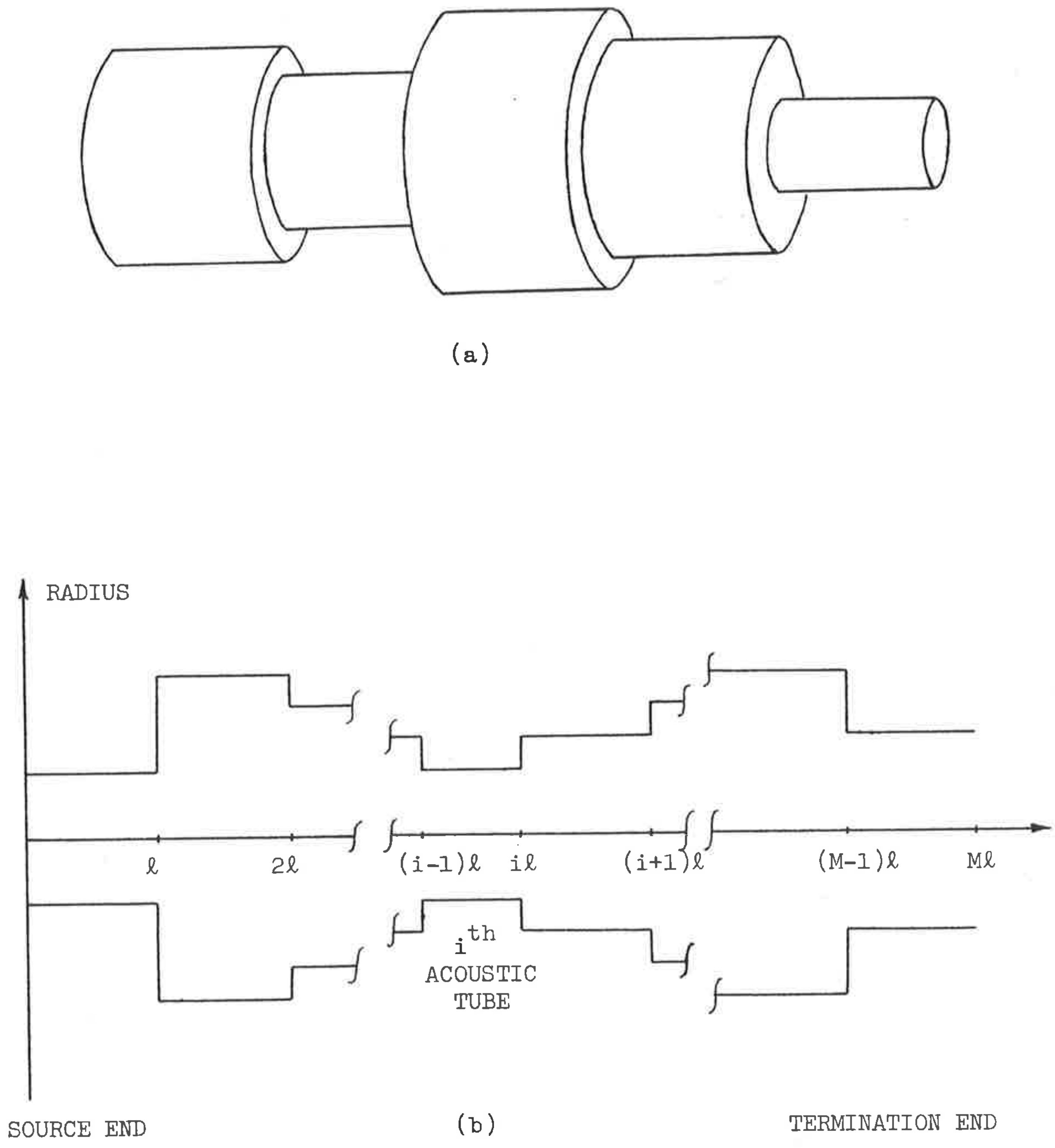
The set of commensurate tubes does not have any branches along its length, and so has only two ends. One end is denoted as the input, or source, end, and the other the output, or termination, end. External excitation of the acoustic tubes occurs at the source end, with the radiated, or output, acoustic waveform radiating from the termination end. The necessity for no branches to the acoustic tubes does restrict the areas of application in which the acoustic tube model developed in this section can be applied.

Boundary conditions are applied to the commensurate acoustic tubes to reduce the complexity of the mathematical waveform analysis needed to produce a model of the acoustic tubes. In simple terms (a more exact description is given in Section 2.2.4), the boundary conditions at the source end have the source acoustically matched to the acoustic tube at the source end. The boundary condition at the terminating end is an acoustic short circuit which prevents any loss of acoustic energy from the terminating end of the set of acoustic tubes. This latter boundary condition does not allow any radiated waveform, a condition which is clearly violated in any practical situation.

Figure 2.1 presents a perspective and cross-section of a typical set of commensurate acoustic tubes satisfying the assumptions and requirements detailed above. For reference purposes, each acoustic tube is numbered from the source end, so the ith acoustic tube starts at a distance $(i-1)\ell$ and finishes at a distance $i\ell$ from the source end. There are a total of M acoustic tubes in each set of commensurate acoustic tubes.

2.2.2 WAVEFORM DESCRIPTION OF A SINGLE ACOUSTIC TUBE

This section provides a waveform description of the pressure and volume velocities in an acoustic tube which satisfies the assumptions and requirements stated in the previous section. The pressure waveform a distance x from the input or source end of the ith acoustic tube in a set of commensurate acoustic tubes, at a time t, is denoted as $p_i(x,t)$. At the same distance x from the source end of the ith acoustic tube, and at the same time t, the volume velocity is denoted as $W_i(x,t)$.



(a)

(b)

FIGURE 2.1: A typical set of commensurate acoustic tubes in (a) perspective view and (b) cross sectional view.

The relationships between pressure and volume velocity within a single acoustic tube have been known for some time, and are derived from the momentum and conservation of mass equations. Using the same assumptions of the properties of an acoustic tube as detailed in the previous section, MORSE and INGARD [1968] have shown that for the *i*th acoustic tube the momentum equation can be written as

$$\frac{\partial p_i(x,t)}{\partial x} = -\frac{\rho}{A_i}\frac{\partial W_i(x,t)}{\partial t} \qquad (2.1)$$

and the continuity of mass equation can be written as

$$\frac{\partial W_i(x,t)}{\partial x} = -\frac{A_i}{\rho c^2}\frac{\partial p_i(x,t)}{\partial t} \qquad (2.2)$$

where ρ defines the density of the medium in the set of acoustic tubes, c the velocity of sound in that medium and A_i the constant cross-sectional area* of the *i*th acoustic tube.

Equations 2.1 and 2.2 are the one dimensional equations of motion within the *i*th acoustic tube. Differentiating Equation 2.1 by x and Equation 2.2 by *t*, and eliminating the volume velocity, produces the one dimensional wave equation for pressure as

$$\frac{\partial^2 p_i(x,t)}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 p_i(x,t)}{\partial t^2} \qquad (2.3)$$

*for simplicity, cross-sectional area is often shortened to area.

Similar differentiation and elimination of the pressure waveform produces the one dimensional wave equation for volume velocity as

$$\frac{\partial^2 w_i(x,t)}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 w_i(x,t)}{\partial t^2} \tag{2.4}$$

If a forward (i.e. towards the termination end of the set of acoustic tubes) travelling waveform is denoted by a superscript +, and a backward (i.e. towards the source end of the set of acoustic tubes) travelling waveform is denoted by a superscript −, then the solution of Equation 2.3 is written as

$$p_i(x,t) = p_i^+\left(t - \frac{x}{c}\right) + p_i^-\left(t + \frac{x}{c}\right) \tag{2.5}$$

Similarly, the solution of Equation 2.4 is written as

$$w_i(x,t) = w_i^+\left(t - \frac{x}{c}\right) - w_i^-\left(t + \frac{x}{c}\right) \tag{2.6}$$

The forward and backward travelling pressure and volume velocity waveforms within the ith acoustic tube are depicted in Figure 2.2.

It has been shown by MARKEL and GRAY [1976] that the forward and backward travelling pressure waveform in the ith acoustic tube are found in terms of $w_i^+\left(t - \frac{x}{c}\right)$ and $w_i^-\left(t + \frac{x}{c}\right)$ as

$$p_i^+\left(t - \frac{x}{c}\right) = \frac{\rho c}{A_i} w_i^+\left(t - \frac{x}{c}\right) + C_1 \tag{2.7}$$

and

$$p_i^-\left(t + \frac{x}{c}\right) = \frac{\rho c}{A_i} w_i^-\left(t + \frac{x}{c}\right) + C_2 \tag{2.8}$$

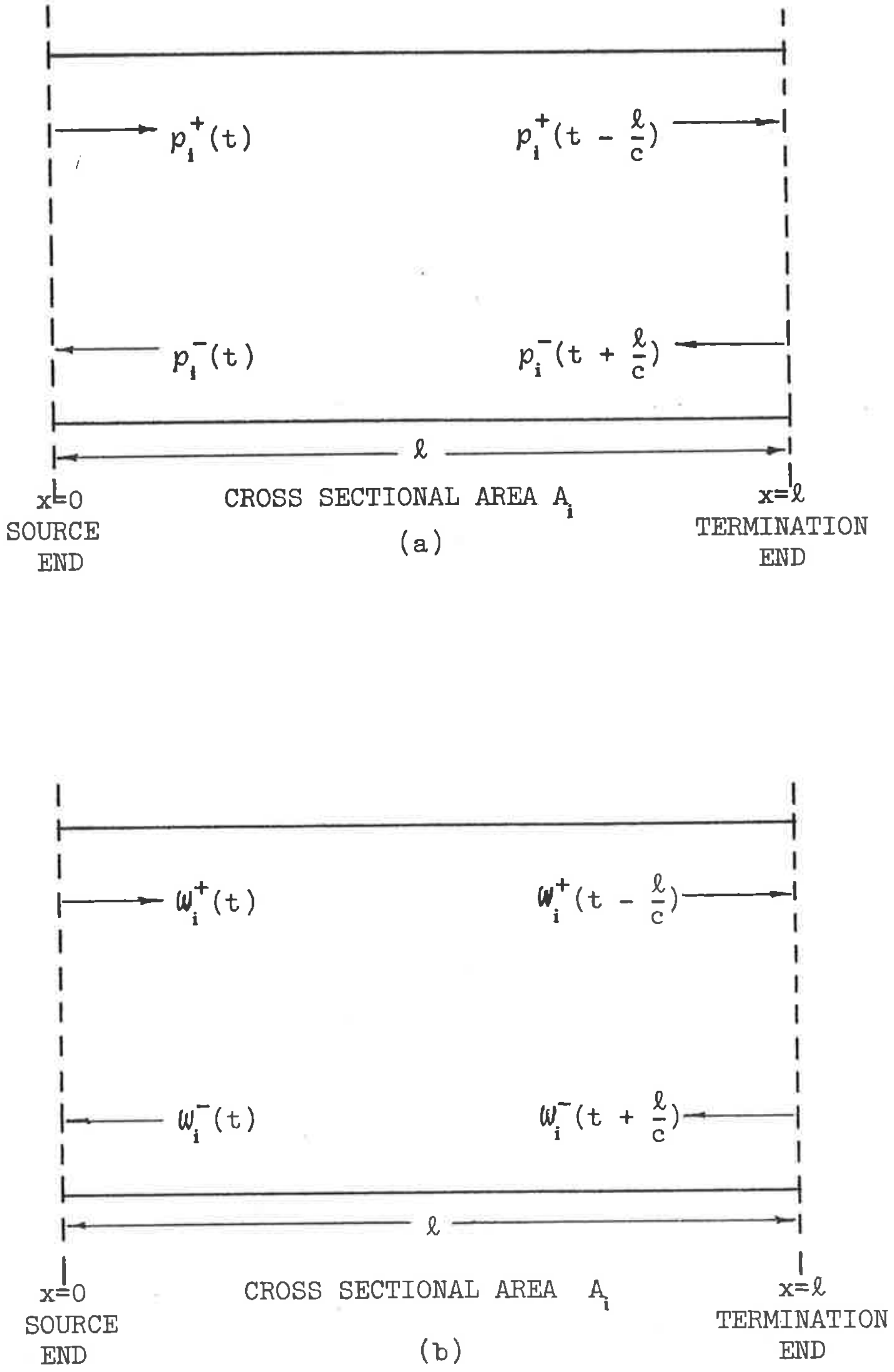

$p_i^+(t)$
$p_i^+(t - \frac{\ell}{c})$
$p_i^-(t)$
$p_i^-(t + \frac{\ell}{c})$
ℓ
x=0
SOURCE END
CROSS SECTIONAL AREA A_i
(a)
x=ℓ
TERMINATION END
$W_i^+(t)$
$W_i^+(t - \frac{\ell}{c})$
$W_i^-(t)$
$W_i^-(t + \frac{\ell}{c})$
ℓ
x=0
SOURCE END
CROSS SECTIONAL AREA A_i
(b)
x=ℓ
TERMINATION END

FIGURE 2.2: Forward and backward travelling waveforms in the i^{th} acoustic tube for (a) pressure and (b) volume velocity.

where C_1 and C_2 are constants. In practice, it is the pressure variation and not the absolute pressure value which is important, and so the constants C_1 and C_2 are neglected. Equations 2.7 and 2.8, along with Equation 2.5, enable the pressure waveform in the ith acoustic tube to be expressed in terms of $w^+\left(t-\frac{x}{c}\right)$ and $w^-\left(t+\frac{x}{c}\right)$ as

$$p_i(x,t) = \frac{\rho c}{A_i}\left[w_i^+\left(t-\frac{x}{c}\right) + w_i^-\left(t+\frac{x}{c}\right)\right] \qquad (2.9)$$

Therefore, all the waveforms in the ith acoustic tube are found from just the knowledge of the forward and backward travelling volume velocity waves.

2.2.3 WAVEFORM DESCRIPTION AT THE JUNCTION OF ACOUSTIC TUBES

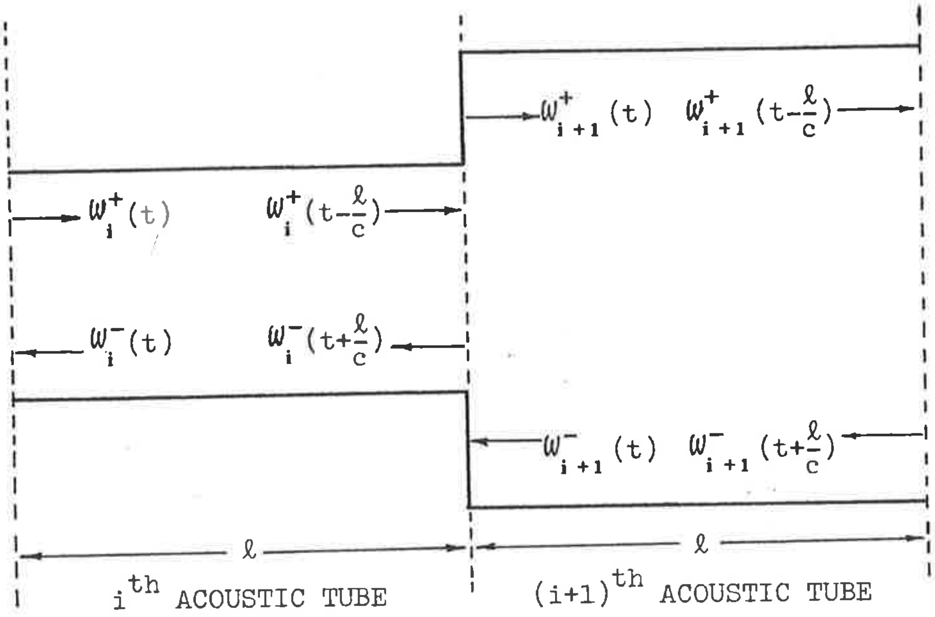
The relationships for volume velocity and pressure waveforms at the junction of two acoustic tubes are defined by requiring continuity of volume velocity and pressure across the junction. This section produces these relationships at the junction of the ith and $(i+1)$th acoustic tubes in a commensurate set of acoustic tubes, as depicted in Figure 2.3a.

Continuity of both pressure and volume velocity waveforms at the junction of the ith and $(i+1)$th acoustic tubes requires

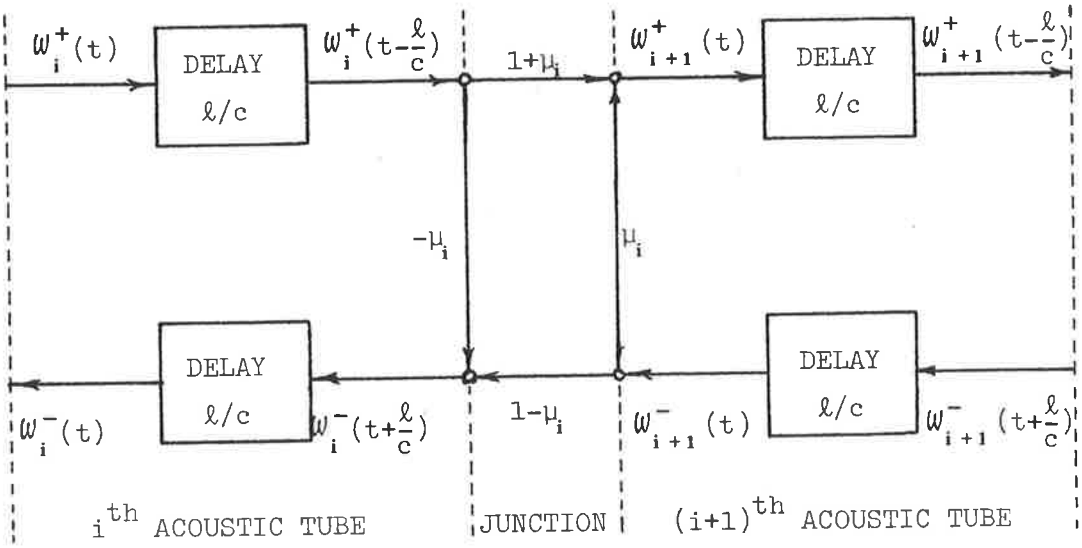
$$p_i(\ell,t) = p_{i+1}(0,t) \qquad (2.10)$$

and

$$w_i(\ell,t) = w_{i+1}(0,t) \qquad (2.11)$$


$w_i^+(t)$
$w_i^+(t-\frac{\ell}{c})$
$w_i^-(t)$
$w_i^-(t+\frac{\ell}{c})$
$w_{i+1}^+(t)$
$w_{i+1}^+(t-\frac{\ell}{c})$
$w_{i+1}^-(t)$
$w_{i+1}^-(t+\frac{\ell}{c})$
ℓ
ℓ
i^{th} ACOUSTIC TUBE
$(i+1)^{th}$ ACOUSTIC TUBE

(a)



(b)

FIGURE 2.3: The junction of the i^{th} and $(i+1)^{th}$ acoustic tubes depicted in (a) and shown in flow diagram form in (b).

Using Equations 2.7 and 2.8 (neglecting constants C_1 and C_2) the continuity equation for pressure, i.e. Equation 2.10, is rewritten as

$$\frac{\rho c}{A_i}\left[w_i^+\left(t-\frac{\ell}{c}\right) + w_i^-\left(t+\frac{\ell}{c}\right)\right] = \frac{\rho c}{A_{i+1}}\left[w_{i+1}^+(t) + w_{i+1}^-(t)\right] \qquad (2.12)$$

The continuity of volume velocity equation, i.e. Equation 2.11, is rewritten as

$$w_i^+\left(t-\frac{\ell}{c}\right) - w_i^-\left(t+\frac{\ell}{c}\right) = w_{i+1}^+(t) - w_{i+1}^-(t) \qquad (2.13)$$

by use of Equation 2.6. Rearrangement of Equations 2.12 and 2.13 produces

$$w_{i+1}^+(t) = \frac{2A_{i+1}}{A_{i+1}+A_i} w_i^+\left(t-\frac{\ell}{c}\right) + \frac{A_{i+1}-A_i}{A_{i+1}+A_i} w_{i+1}^-(t) \qquad (2.14)$$

and

$$w_i^-\left(t+\frac{\ell}{c}\right) = -\frac{A_{i+1}-A_i}{A_{i+1}+A_i} w_i^+\left(t-\frac{\ell}{c}\right) + \frac{2A_i}{A_{i+1}+A_i} w_{i+1}^-(t) \qquad (2.15)$$

Equations 2.14 and 2.15 completely define the volume velocity waveforms at the junction of the ith and $(i+1)$th acoustic tubes, and are referred to as the JUNCTION EQUATIONS.

The waveform $w_{i+1}^+\left(t-\frac{\ell}{c}\right)$ which is incident on the junction of the ith and $(i+1)$th acoustic tubes, from the source side of the junction, is transmitted into the $(i+1)$th acoustic tube as $w_{i+1}^+(t)$ and reflected back into the ith acoustic tube as $w_i^-\left(t+\frac{\ell}{c}\right)$. Therefore, from Equation 2.14, the transmission coefficient from the ith to $(i+1)$th acoustic tubes is $2A_{i+1}/(A_{i+1}+A_i)$. The reflection coefficient for a waveform incidient on the junction from the source side is then $-(A_{i+1}-A_i)/(A_{i+1}+A_i)$ from Equation 2.15. Similar consideration of the waveform $w^-_{i+1}(t)$ incident on the junction from the termination side produces the transmission and reflection coefficients from the $(i+1)$th to the ith acoustic tube as $2A_i/(A_{i+1}+A_i)$ and $(A_{i+1}-A_i)/(A_{i+1}+A_i)$, respectively.

The symbol μ_i is used to denote the reflection coefficient for a volume velocity incident on the junction of the ith and $(i+1)$th acoustic tubes, from the termination side, i.e.

$$\mu_i = \frac{A_{i+1}-A_i}{A_{i+1}+A_i} \tag{2.16}$$

An important property of μ_i is that its magnitude is always less than or equal to unity (i.e. $-1 \leqslant \mu_i \leqslant 1$) since the cross-sectional areas A_i and A_{i+1} are always positive for real acoustic tubes. Using μ_i the JUNCTION EQUATIONS, i.e. Equations 2.14 and 2.15, are written as

$$w^+_{i+1}(t) = (1+\mu_i)w^+_i\left(t-\frac{\ell}{c}\right) + \mu_i w^-_{i+1}(t) \tag{2.17}$$

and

$$w^-_i\left(t+\frac{\ell}{c}\right) = -\mu_i w^+_i\left(t-\frac{\ell}{c}\right) + (1-\mu_i)w^-_{i+1}(t) \tag{2.18}$$

Figure 2.3(b) presents these JUNCTION EQUATIONS as a signal flow diagram, using the conventions defined by OPPENHEIM and SCHAFER [1975].

2.2.4 BOUNDARY CONDITIONS

A brief description of the boundary conditions required for a set of commensurate acoustic tubes was given in Section 2.2.1. These boundary conditions are necessary to reduce the complexity of the acoustic tubes so that the model of the acoustic tubes is described by tractable mathematics. This section details the necessary boundary conditions at the source and termination ends of the acoustic tubes.

At the source end of the acoustic tubes, the excitation source is required to be acoustically matched to the first acoustic tube. Therefore, the transmission coefficient from the source into the first acoustic tube, and vice versa, is unity. The reflection coefficient for waveforms incident on the junction of the first acoustic tube and the source is zero. Thus, the source boundary condition requires the loss of all the backward travelling volume velocity waveform at the source end of the first acoustic tube.

A pictorial and signal flow diagram description of the source boundary condition is presented in Figure 2.4, with the source or excitation volume velocity denoted by $W_0^+(t)$. The volume velocity waveform absorbed at the junction of the source and the first acoustic tube is denoted by $W_0^-(t)$, and is not used in analysis or synthesis procedures, and so is only included for completeness. The source excitation is assumed to occur at exactly the source end of the first acoustic tube, so that there is no time delay between the source waveform $W_0^+(t)$ and $W_1^+(t)$.

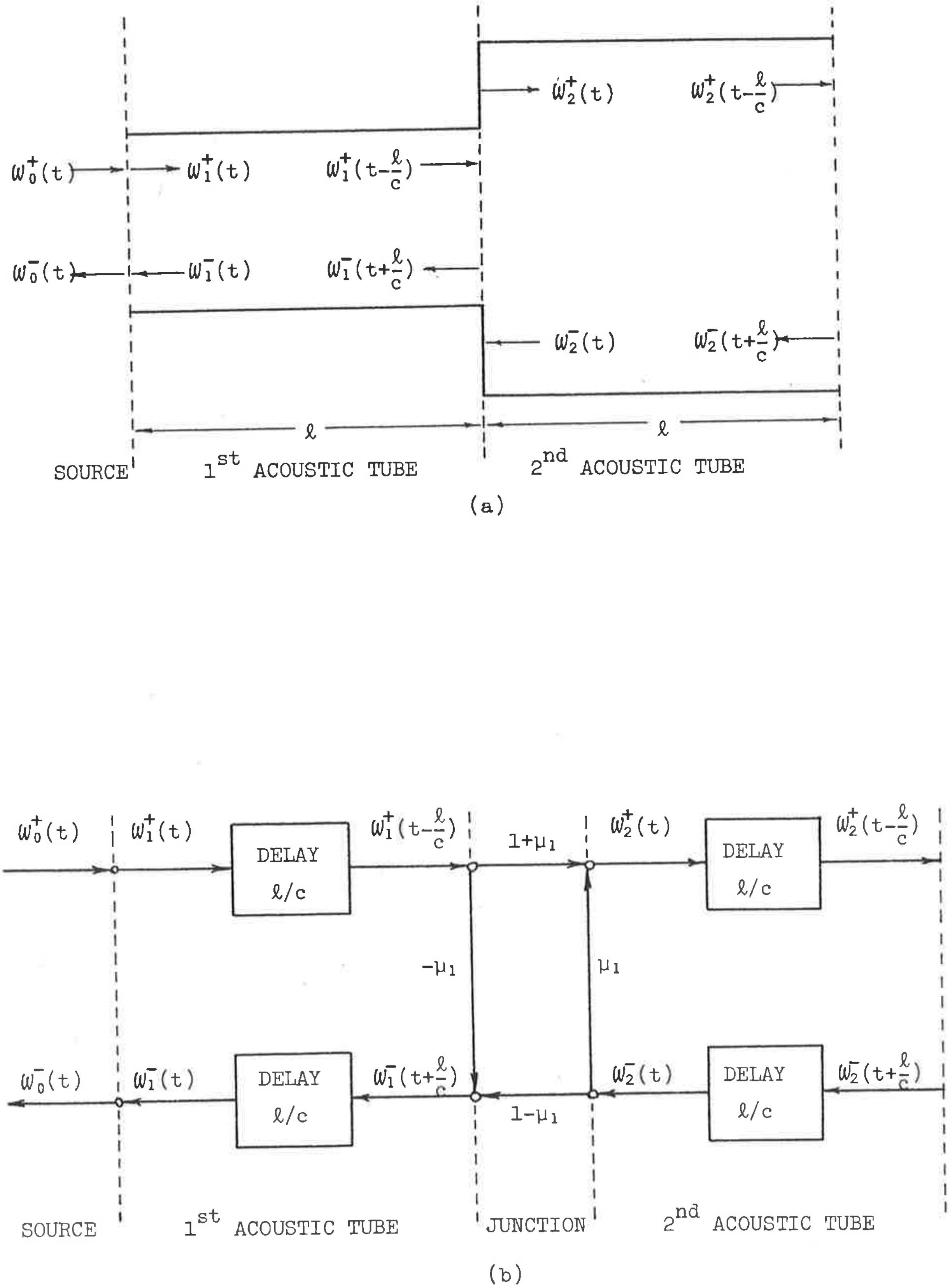


FIGURE 2.4: Source boundary condition depicted in (a) with the signal flow diagram presented in (b).

The termination boundary condition requires an acoustic short circuit at the termination end of the Mth or termination acoustic tube of the set of commensurate acoustic tubes. Thus, zero acoustic pressure but finite acoustic volume velocity occurs at the termination. Defining the acoustic impedance of the termination by Z_{M+1}, then the termination boundary condition requires Z_{M+1} to be zero. The effective termination cross-sectional area, i.e. A_{M+1}, is found from the equation

$$A_{M+1} = \frac{\rho c}{Z_{M+1}} \tag{2.19}$$

as infinity. Equation 2.16 defines the termination reflection coefficient as

$$\mu_M = \frac{A_{M+1} - A_M}{A_{M+1} + A_M} \tag{2.20}$$

which on substitution of A_{M+1} equal to infinity defines the termination reflection coefficient to have a value of unity, i.e.

$$\mu_M = 1.0 \tag{2.21}$$

A pictorial and signal flow diagram description of the termination boundary condition is presented in Figure 2.5. The reflection coefficient for the volume velocity in the Mth acoustic tube incident on the termination, i.e. $w_M^+(t - \ell/c)$, is $-\mu_M$ which, from Equation 2.21, has a numerical value of -1. Therefore, all of $w_M^+(t - \ell/c)$ is reflected back into the Mth acoustic tube with a phase reversal and none is lost into the termination. The zero termination impedance required by the termination boundary condi-

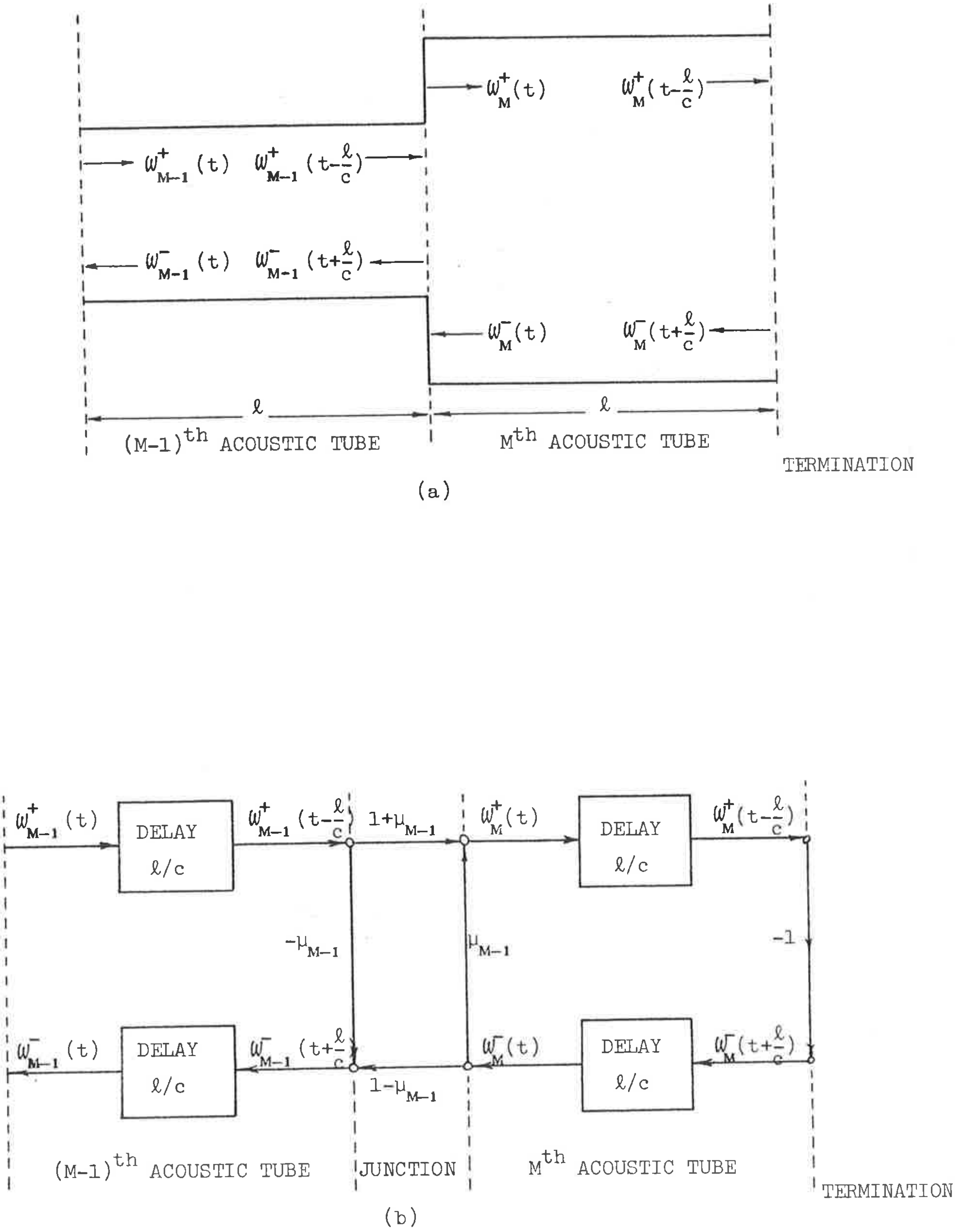


FIGURE 2.5: Termination boundary condition depicted in (a) with the signal flow diagram presented in (b).

tion prevents any pressure waveform from being radiated from the commensurate acoustic tubes.

2.2.5 COMPLETE ACOUSTIC TUBE MODEL

A complete description of the acoustic waveforms in a single acoustic tube, at the junction of acoustic tubes, and the necessary boundary conditions for a set of commensurate acoustic tubes have been presented in the previous sections. This section combines these descriptions to provide a complete definition and waveform description of the model used for the commensurate acoustic tubes. This model is referred to as the acoustic tube model.

Firstly, a simplification of the mathematical description of the acoustic tubes is performed. The signal flow diagram for two commensurate acoustic tubes satisfying the ideal assumptions and requirements set out in Section 2.2.1 is presented in Figure 2.6. The impulse response of this two acoustic tube system is found in APPENDIX A as

$$W_3^+(t) = 2(1+\mu_1)\sum_{k=0}^{\infty}(-\mu_1)^k\delta\left(t - \frac{2(k+1)\ell}{c}\right) \qquad (2.22)$$

where $\delta(t)$ is the impulse function. Equation 2.22 shows that the soonest an impulse can reach the termination is $2\ell/c$. Then succussive impulses due to reflections at the junction of the acoustic tubes reach the termination at multiples of $2\ell/c$ later. The time interval $2\ell/c$ corresponds to the round trip propagation delay within a single acoustic tube.

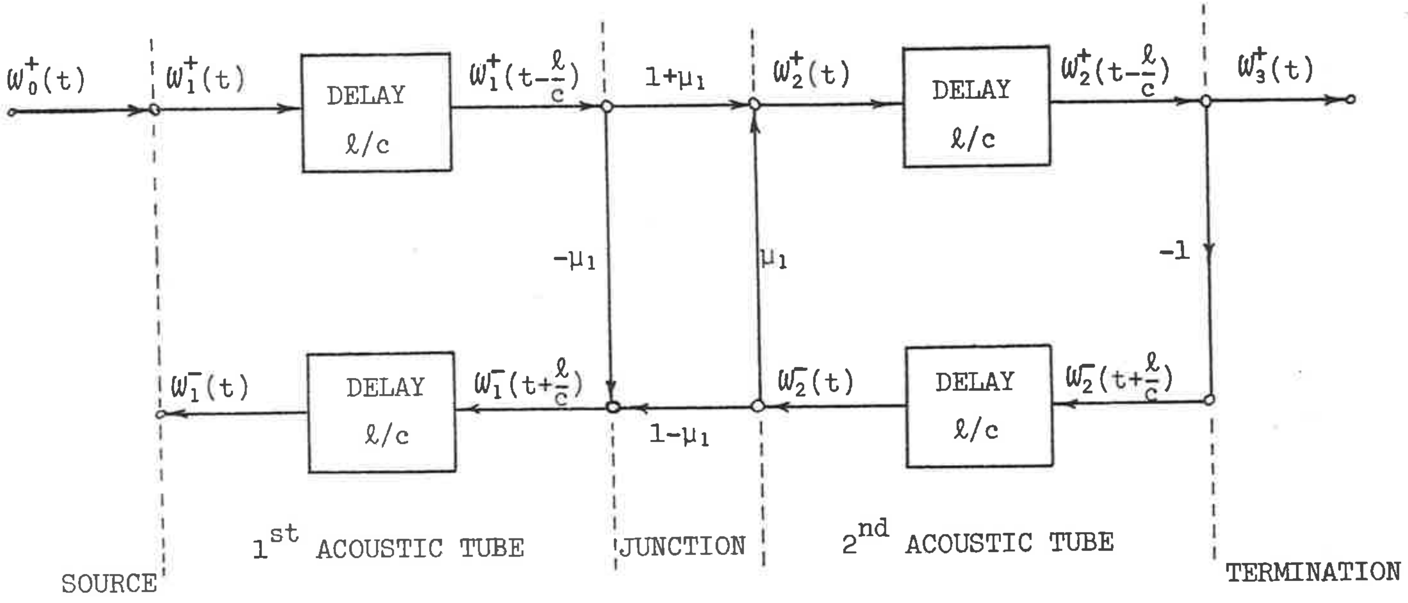


FIGURE 2.6: Signal flow diagram for a two acoustic tube concatenation.

For the general situation where there are M commensurate acoustic tubes in a concatenation, the quickest an impulse can reach the termination from the source end is $2M\ell/c$. Due to the constant cross-sectional area along each acoustic tube then successive impulses, due to reflections at the junction of acoustic tubes, reach the termination at multiples of $2\ell/c$ later. These observations can be generalised to any excitation function which is bandlimited to frequencies below $\pi/(2\ell/c)$ [RABINER and SCHAFER 1975]. This enables the output waveform of the acoustic tubes to be sampled with a period of $T = 2\ell/c$ without any loss of information.

Defining all the time quantities in terms of the sampling period T creates some difficulties with the equivalent discrete time model of the analog acoustic tubes. Figure 2.7(a) shows the correct discrete time model of an acoustic tube. The half sample delay, i.e. $T/2$, implies that an interpolation half-way between sample values is necessary. This interpolation is undesirable, and is overcome by placing all the waveform propagation delay in the reverse path (or in the forward path), as depicted in Figure 2.7(b).

Using the discrete time model of an acoustic tube depicted in Figure 2.7(b) does not alter any of the waveform delays around any closed path within the acoustic tubes. However, the overall waveform delay from the input to output is incorrect. From both the practical and theoretical point of view, this is of minor significance, and can be easily conpensated for by appropriately advancing or delaying the input or output waveforms respectively. The model shown in Figure 2.7(b) not only eliminates interpolation bewteen waveform sample values but also allows the set of acoustic tubes

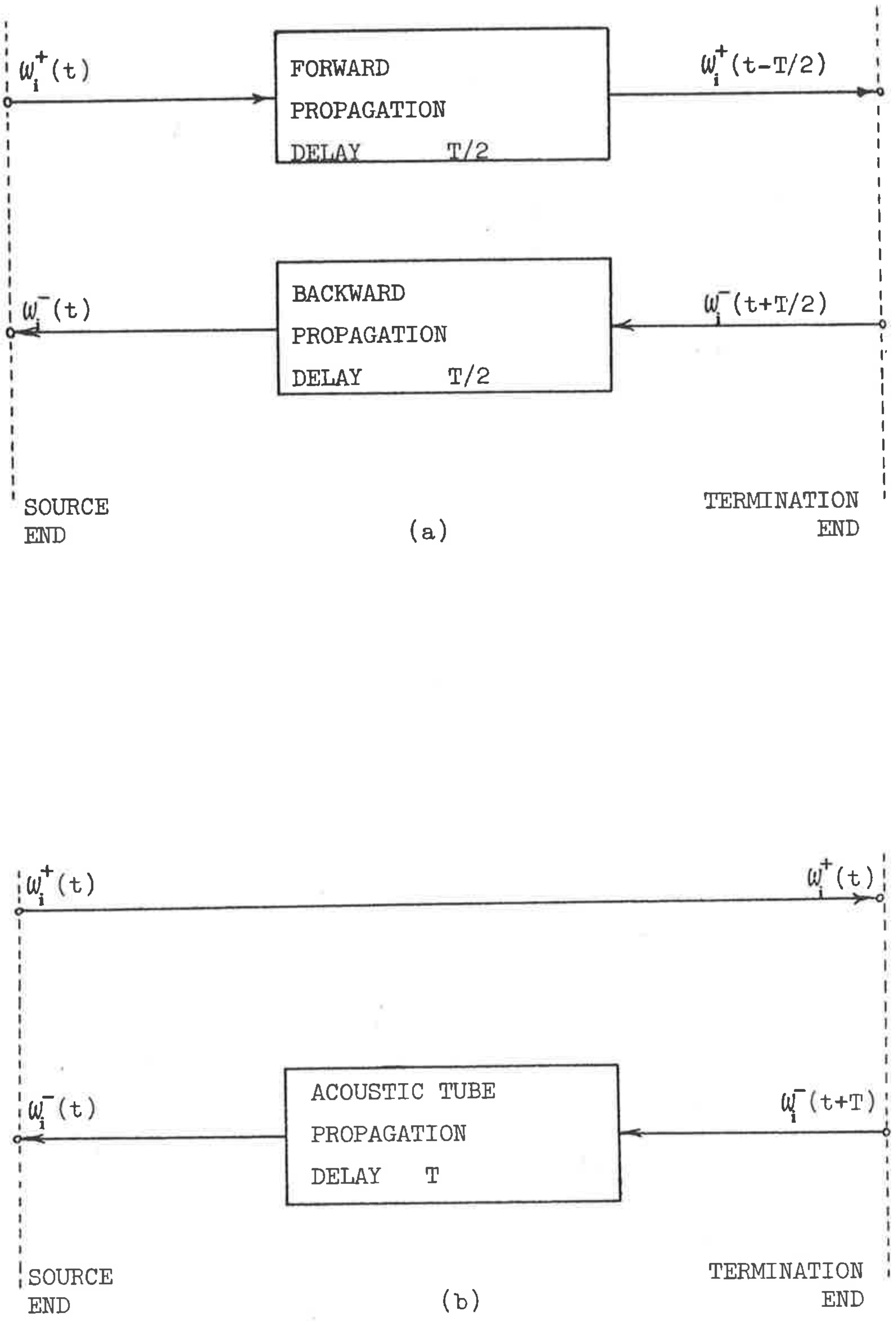


FIGURE 2.7: The i^{th} acoustic tube shown in (a) the correct discrete time model and (b) a discrete time model with all the propagation delay in the backward path.

to be defined by a set of difference equations, hence permitting iterative calculation of waveform sample values.

The discrete time model of a single acoustic tube presented in Figure 2.7(b) permits a simpler set of symbols to be used to define the waveform quantities. To such an end, the symbol $U_i(n)$ is used to define the forward travelling volume velocity in the ith acoustic tube which is leaving the junction between the (i-1)th and ith acoustic tubes, at time instant nT, i.e.

$$U_i(n) = W_i^+(nT) \tag{2.23}$$

The symbol $V_i(n)$ defines the backward travelling volume velocity in the ith acoustic tube which is leaving the junction of the ith and (i+1)th acoustic tubes, at time instant nT, i.e.

$$V_i(n) = W^-(nT) \tag{2.24}$$

Using Equations 2.23 and 2.24, the junction equations for the waveform description at the junction of the ith and (i+1)th acoustic tubes, i.e. Equations 2.17 and 2.18, are rewritten as

$$U_{i+1}(n) = (1+\mu_i)U_i(n) + \mu_i V_{i+1}(n-1) \tag{2.25}$$

and

$$V_i(n) = -\mu_i U_i(n) + (1-\mu_i)V_{i+1}(n-1) \tag{2.26}$$

The signal flow diagram equivalent for these JUNCTION EQUATIONS, i.e. Equations 2.25 and 2.26, is presented in Figure 2.8(a). These JUNCTION EQUATIONS are the required difference equations which allow iterative computation of any discretely sampled

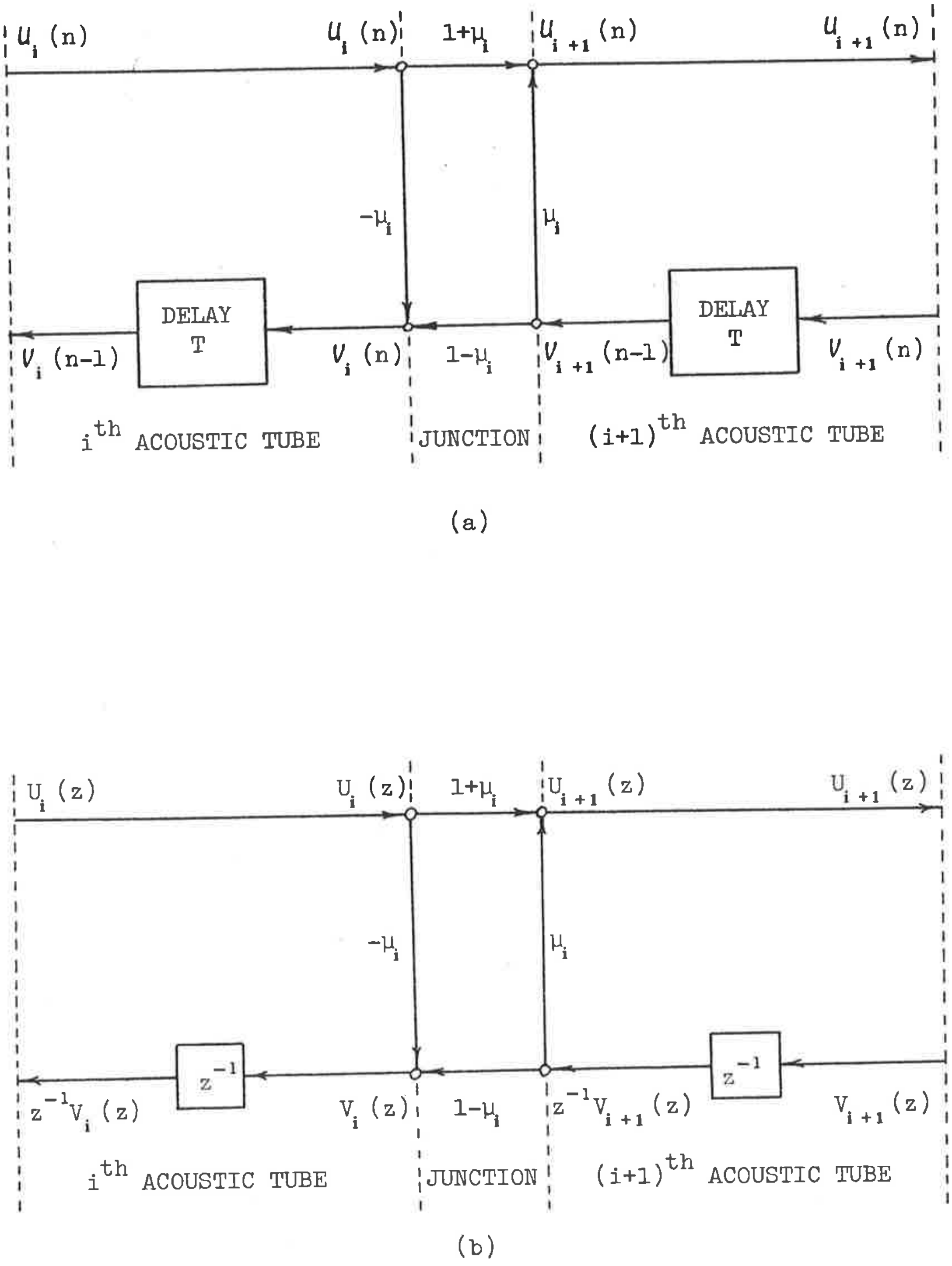


FIGURE 2.8: Signal flow diagrams for two acoustic tubes in (a) the time domain and (b) the z domain.

acoustic volume velocity waveform within an acoustic tube or set of commensurate acoustic tubes.

In many situations, the z transform equivalent of the JUNCTION EQUATIONS is required. With $U_i(z)$ as the z transform of $u_i(n)$, i.e.

$$U_i(z) = \sum_{n=-\infty}^{\infty} u_i(n) z^{-n} \qquad (2.27)$$

and $V_i(z)$ as the z transform of $v_i(n)$, i.e.

$$V_i(z) = \sum_{n=-\infty}^{\infty} v_i(n) z^{-n} \qquad (2.28)$$

then the z transform equivalents of the JUNCTION EQUATIONS, i.e. 2.25 and 2.26, are

$$U_{i+1}(z) = (1+\mu_i)U_i(z) + \mu_i z^{-1} V_{i+1}(z) \qquad (2.29)$$

and

$$V_i(z) = -\mu_i U_i(z) + (1-\mu_i) z^{-1} V_{i+1}(z) \qquad (2.30)$$

The signal flow diagram equivalent of the above z transform JUNCTION EQUATIONS is presented in Figure 2.8(b).

The signal flow diagrams presented in Figure 2.8 are often referred to as the four multiplier configuration. This is due to the necessity to perform four multiplications for each iteration with this configuration. In most real time applications, multiplications are time consuming, and so modifications to the JUNCTION EQUATIONS have been performed (e.g. OPPENHEIM and SCHAFER 1975, ITAKURA and SAITO 1971) to reduce the number of multiplications. The results of such modifications lead to the two and one-multiplier configurations. These configurations, however, do not have the one

one to one correspondence between the model and physical waveforms at the junction, and so are not considered here.

Using the signal flow diagrams of a single acoustic tube, i.e. Figure 2.7(b), and the junction of two acoustic tubes, i.e. Figure 2.8(a), the time domain signal flow diagram for a set of commensurate acoustic tubes is presented in Figure 2.9. The corresponding z domain equivalent of Figure 2.9 is presented in Figure 2.10. These signal flow diagrams define all the volume velocity time and z domain waveforms at any point and discrete time instant within a set of commensurate acoustic tubes. The restrictions, assumptions and boundary conditions defined in Section 2.2.1 are satisfied by these signal flow diagrams, and so represent the acoustic tube model of M commensurate acoustic tubes. Both analysis and synthesis of volume velocity waveforms are possible with the signal flow diagrams presented in Figures 2.9 and 2.10.

Signal flow diagrams similar to those of Figures 2.9 and 2.10 were first used by KELLY and LOCHBAUM [1962] for speech synthesis in 1962. Since then, researchers (e.g. ITAKURA and SAITO [1971] and GRAY and MARKEL [1973, 1975]) have developed digital signal processing techniques from these types of signal flow diagrams, which are often referred to as DIGITAL LATTICES, LADDER FILTERS or LATTICE FILTERS.

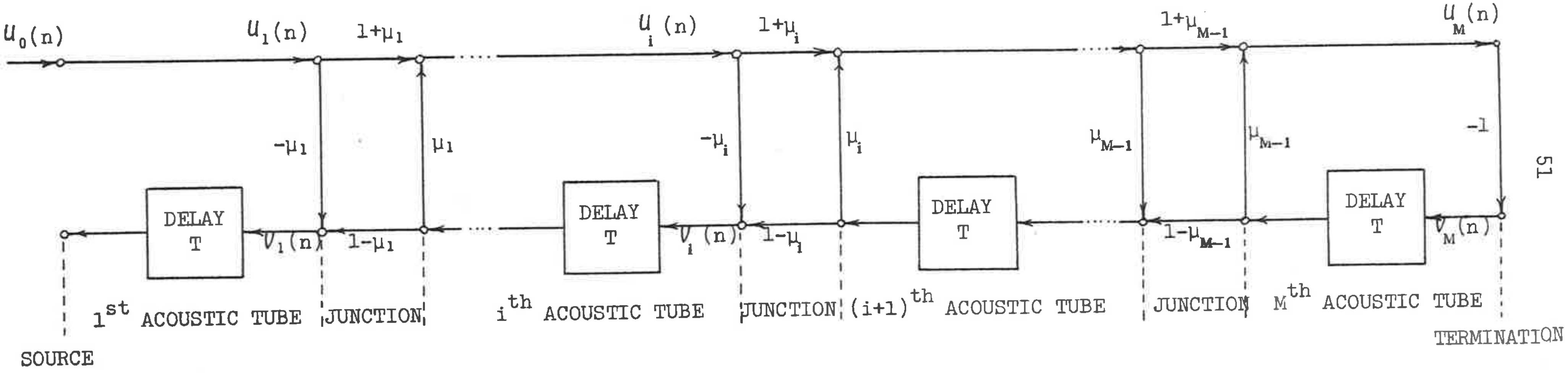


FIGURE 2.9: Flow diagram of acoustic tube model in time domain.

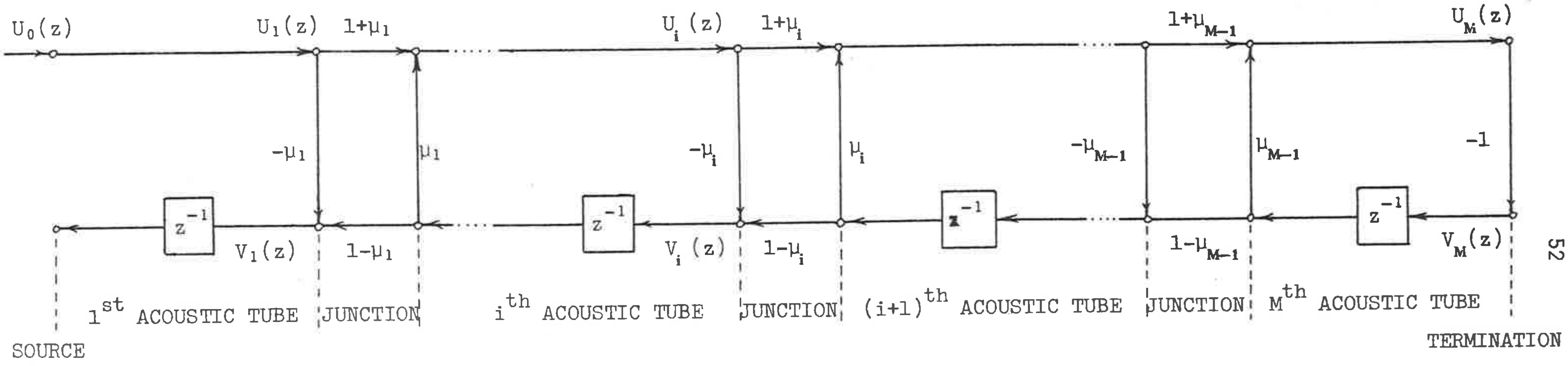


FIGURE 2.10: Flow diagram of acoustic tube model in z domain.

2.3 LINEAR PREDICTION

Linear prediction has developed into a very useful analysis technique for a wide range of applications. The reasons for the success of linear prediction lie in its ability to provide accurate system identification for certain classes of signal with a relatively small amount of mathematical computation. This section initially defines and presents the basic principles of linear prediction analysis. For each application of linear prediction, a specific formulation is necessary, and this section contains a detailed description of three basic formulations. A discussion of the advantages and disadvantages, and the most likely areas of application for each of the basic formulations, is also presented.

Linear predictive analysis techniques have been used for a long time (e.g. the origins of linear prediction have been attributed to GAUSS in 1795 and LEGENDRE in 1806 [SORENSON 1970]). In the engineering area, linear prediction has been used in the areas of control and information theory, under the names of "system estimation" and "system identification." The term "system identification" aptly describes linear prediction because, once a linear prediction has been performed on a data sequence, an all pole system model is uniquely identified.

Linear prediction is defined as a procedure which estimates or predicts the next sampled time waveform output of a system by a linear combination of a number of the system's past outputs. The goal of any linear prediction process is to produce a minimum error between its predicted output values and the actual output values of the system. The manner in which linear prediction achieves this minimum error in a particular situation determines the type of formulation of linear prediction used in that situation.

The first application of linear prediction to speech processing was by SAITO and ITAKURA [1966] with a maximum likelihood formulation, with a condensed English version of their paper appearing in 1968 [ITAKURA and SAITO 1968] and a detailed paper in 1970 [ITAKURA and SAITO 1970]. During this period, a number of papers by ATAL and SCHROEDER [1967, 1968a, 1968b] investigated predictive coding of speech. In 1970, the term LINEAR PREDICTION was introduced in connection with speech processing by ATAL [1970a]. The inverse filter formulation of linear prediction was presented by MARKEL [1971b] and detailed later by MARKEL and GRAY [1976]. The covariance formulation of linear prediction was introduced by ATAL and HANAUER [1971].

During 1972, the PARCOR formulation of linear prediction was presented in English by ITAKURA and SAITO [1972, et al 1972], and WAKITA [1972] in the same year showed that it could be interpreted as a detailed structure of an inverse filter. Also during 1972, MAKHOUL and WOLF [1972] presented a unified presentation of the covariance and autocorrelation methods. Many other formulations than those mentioned so far have been produced, and the works of KAILATH [1974], MAKHOUL [1975] and MARKEL and GRAY [1976] provide an excellent description and summary of the most useful formulations.

The differences between many formulations of linear prediction is generally small, and those differences often only occur because of the different way in which a problem is viewed or approached. In the application of linear prediction to speech processing, three basic formulations can be found, to which all other formulations are equivalent. These basic formulations are called the covariance [ATAL and HANAUER 1971], autocorrelation [MARKEL and GRAY 1976, MAKHOUL 1975, MARKEL and GRAY 1973], and the lattice [MAKHOUL 1977] formulations. A full presentation of each of these formulations is given in the following sections. The difference between these three basic formulations of linear prediction is the manner in which various computations are performed.

The following section presents and defines the basic principles of linear prediction, from which the various formulations are derived.

2.3.1 BASIC PRINCIPLES

A linear predictor estimates or predicts a value of a signal by a linear combination of the system's present and past output values. If the discrete sampled output from a system at a time nT (n is an integer and T the sampling period) is denoted by $s(n)$, then a pth order linear prediction estimates $s(n)$ as $\hat{s}(n)$, where

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \qquad (2.31)$$

The constants a_k are referred to as the linear predictor coefficients, and these are to be determined such that the error between $s(n)$ and $\hat{s}(n)$ is minimized in some manner. The error between $s(n)$ and $\hat{s}(n)$ is denoted as $e(n)$, defined by

$$e(n) = s(n) - \hat{s}(n) \qquad (2.32)$$

and referred to as the residual or predictor error.

Using the definition of a pth order linear predictor, i.e. Equation 2.32, the predictor error is written as

$$e(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \qquad (2.33)$$

In the z domain, the predictor error, $e(n)$, is denoted as $E(z)$, and defined by

$$E(z) = \sum_{n=-\infty}^{\infty} e(n) z^{-n} \qquad (2.34)$$

and the signal, $s(n)$, in the z domain is denoted as $S(z)$ and defined by

$$S(z) = \sum_{n=-\infty}^{\infty} s(n) z^{-n} \qquad (2.35)$$

The z domain equivalent of Equation 2.33 is then written as

$$E(z) = S(z)\ (1 - \sum_{k=1}^{p} a_k z^{-k}) \qquad (2.36)$$

From the linear predictor point of view, its input signal is $s(n)$ ($S(z)$ in the z domain) and its output is $e(n)$($E(z)$ in the z domain), and so the transfer function of the linear predictor, denoted as $A(z)$, is found from Equation 2.36 as

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \qquad (2.37)$$

From Equation 2.37 it is seen that the transfer function of linear prediction, $A(z)$, is all zero. In terms of using linear prediction as a system identification procedure, then $E(z)$ is the input signal and $A(z)$ is the output signal, and so the system is identified as $1/A(z)$. Therefore, it is easily seen why $A(z)$ is referred to as the INVERSE FILTER of the system, and why linear prediction is an ALL POLE identification process.

The linear predictor coefficients, a_k, are chosen such that the predictor error, $e(n)$, is minimized, in order that the best possible estimate of future signal samples is obtained. Historically, the method of determining the predictor coefficients, a_k, has been to minimize the sum of the mean square predictor errors over the period of analysis. A justification for minimizing this sum is that the prediction coefficients obtained from an analysis of a time invariant all pole system excited by white noise or an impulse are the same as those of the system, i.e. a correct identification is made. This approach also results in a simple set of linear simultaneous equations, which can be efficiently solved to obtain the predictor coefficients, a_k.

The sum of the square predictor error is defined as

$$E = \sum_{n=n_0}^{n_1} e^2(n) \qquad (2.38a)$$

$$= \sum_{n=n_0}^{n_1} \left(s(n) - \hat{s}(n)\right)^2 \qquad (2.38b)$$

$$= \sum_{n=n_0}^{n_1} \left(s(n) - \sum_{k=1}^{p} a_k s(n-k)\right)^2 \qquad (2.38c)$$

The summation range, i.e. from n_0 to n_1, determines the type of formulation of the linear prediction process, and is discussed in detail later in this chapter. Strictly speaking, a division of Equation 2.38 by the number of samples in the summation, i.e. $n_1 - n_0 + 1$, should be performed to obtain the mean. However, such a division does not affect the derivation of the linear predictor, and so is omitted.

Minimization of the mean square predictor error, E, with respect to the predictor coefficients requires the solution of the set of equations obtained from

$$\frac{\partial E}{\partial a_j} = 0 \qquad (2.39)$$

for $j = 1,2, \ . \ . \ . \ , p$. Substituting Equation 2.38c into Equation 2.39 produces

$$2 \sum_{n=n_0}^{n_1} \left(s(n) - \sum_{k=1}^{p} a_k s(n-k)\right)\left(-s(n-i)\right) = 0 \qquad 1 \leqslant j \leqslant p \qquad (2.40)$$

which is rearranged to

$$\sum_{n=n_0}^{n_1} s(n-j)s(n) = \sum_{k=1}^{p} a_k \sum_{n=n_0}^{n_1} s(n-j)s(n-k) \qquad 1 \leqslant j \leqslant p \qquad (2.41)$$

If the quantity $\phi(j,k)$ is defined as

$$\phi(j,k) = \sum_{n=n_0}^{n_1} s(n-j)s(n-k) \qquad (2.42)$$

then Equation 2.41 simplifies to

$$\phi(j,0) = \sum_{k=1}^{p} a_k \phi(j,k) \qquad 1 \leqslant j \leqslant p \qquad (2.43)$$

Equation 2.43 represents a linear set of p equations in p unknowns which when solved provides a set of predictor coefficients which minimize the mean square error between the signal being analysed and that produced by the linear predictor. Therefore, the basic steps required for a linear predictive analysis are to firstly calculate $\phi(j,k)$ for $1 \leqslant j \leqslant p$ and $0 \leqslant k \leqslant p$, and secondly solve Equation 2.43 for the predictor coefficients, a_k. Calculation of $\phi(j,k)$ can be performed in many ways, and results in many different formulations of linear prediction. However, all these different formulations can be equated to two basic formulations, namely the autocorrelation and covariance formulations, which are presented in the following sections.

2.3.2 COVARIANCE FORMULATION

Assuming that the signal, $s(n)$, is only available from the finite interval of n from zero to N-1, i.e. N samples, then the covariance* formulation minimizes the mean square predictor error over this finite interval. Equation 2.38c shows that the range of $s(n)$ to determine E is from n equal to n_0-p to n_1. Therefore, for the covariance method $n_0 = p$ and $n_1 = N$-1, i.e.

$$E = \sum_{n=p}^{N-1} e^2(n) \tag{2.44a}$$

$$= \sum_{n=p}^{N-1} \left(s(n) - \sum_{k=1}^{p} a_k s(n-k)\right)^2 \tag{2.44b}$$

Hence, the quantity $\phi(j,k)$ is evaluated as

$$\phi(j,k) = \sum_{n=p}^{N-1} s(n-j)s(n-k) \qquad \begin{matrix} 1 \leqslant j \leqslant p \\ 0 \leqslant k \leqslant p \end{matrix} \tag{2.45}$$

for the covariance method of linear prediction.

*The term covariance used here does not have the normal meaning of of a signal correlation with its mean removed.

To reduce confusion, the covariances calculated by Equation 2.45 are denoted as $c(j,k)$ where

$$c(j,k) = \phi(j,k)\Big|_{\substack{n_0=p\\ n_1=N-1}} \qquad \begin{matrix}1\leqslant j\leqslant p\\ 0\leqslant k\leqslant p\end{matrix} \qquad (2.46a)$$

$$= \sum_{n=p}^{N-1} s(n-j)s(n-k) \qquad \begin{matrix}1\leqslant j\leqslant p\\ 0\leqslant k\leqslant p\end{matrix} \qquad (2.46b)$$

This allows the system of linear equations that are used to determine the predictor coefficients, a_k, (i.e. Equation 2.43) to be written as

$$c(j,0) = \sum_{k=1}^{p} a_k c(j,k) \qquad 1\leqslant j\leqslant p \qquad (2.47)$$

The basic steps of the covariance formulation of linear prediction are therefore to, firstly, calculate the covariances defined by Equation 2.46b and, secondly, to solve Equation 2.47 for the predictor coefficients, a_k.

The major difficulty with the covariance formulation is the necessary step of solving the linear set of equations, i.e. Equation 2.47. Equation 2.47 is presented in Figure 2.11 as a matrix equation. The $p{\times}p$ matrix $\mathbb{C}$ is called the covariance matrix, and is positive definite and symmetric. The symmetry of the covariance matrix, i.e.

$$c(j,k) = c(k,j) \qquad (2.48)$$

is easily shown from Equation 2.46b. However, since different data is used to calculate different covariances along any diagonal, then the covariance matrix is not Toeplitz [GRENANDER and SZEGO 1958].

$$\begin{array}{ccccc}
\begin{bmatrix}
C(1,1) & C(1.2) & \cdots & C(1,p) \\
C(2,1) & C(2,2) & \cdots & C(2,p) \\
C(3,1) & C(3,2) & \cdots & C(3,p) \\
\vdots & \vdots & & \vdots \\
C(p,1) & C(p,2) & \cdots & C(p,p)
\end{bmatrix} & \times &
\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{bmatrix} & = &
\begin{bmatrix} C(1,0) \\ C(2,0) \\ C(3,0) \\ \vdots \\ C(p,0) \end{bmatrix} \\
\begin{array}{l}\text{p × p COVARIANCE} \\ \text{MATRIX, } \mathbf{C}\end{array} & &
\begin{array}{c}\text{p × 1} \\ \text{PREDICTOR} \\ \text{COEFFICIENTS} \\ \underset{\sim}{a}\end{array} & &
\begin{array}{c}\text{p × 1} \\ \text{COVARIANCE} \\ \text{VECTOR} \\ \underset{\sim}{c}\end{array}
\end{array}$$

(a)

$$\begin{array}{ccccc}
\begin{bmatrix}
C(1,1) & C(1,2) & \cdots & C(1,p) \\
C(1,2) & C(2,2) & \cdots & C(2,p) \\
C(1,3) & C(2,3) & \cdots & C(3,p) \\
\vdots & \vdots & & \vdots \\
C(1,p) & C(2,p) & & C(p,p)
\end{bmatrix} & \times &
\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{bmatrix} & = &
\begin{bmatrix} C(1,0) \\ C(2,0) \\ C(3,0) \\ \vdots \\ C(p,0) \end{bmatrix} \\
\begin{array}{l}\text{p × p COVARIANCE} \\ \text{MATRIX, } \mathbf{C}\end{array} & &
\begin{array}{c}\text{p × 1} \\ \text{PREDICTOR} \\ \text{COEFFICIENTS} \\ \underset{\sim}{a}\end{array} & &
\begin{array}{c}\text{p × 1} \\ \text{COVARIANCE} \\ \text{VECTOR} \\ \underset{\sim}{c}\end{array}
\end{array}$$

(b)

FIGURE 2.11: Matrix equations for covariance formulation in (a) the general form and (b) the simplified form.

One popular method of solving the linear set of Equations 2.47 is by CHOLESKY decomposition [CLASEN 1966, ATAL and HANAUER 1971, RABINER and SCHAFER 1978]. This decomposition procedure decomposes the covariance matrix into a lower, V, and upper V^t (where t denotes transpose) triangular and diagonal, D, matrices, i.e.

$$C = V D V^t \tag{2.49}$$

There are numerous processes by which the decomposition is performed [FADDEEV and FADDEEVA 1963, ATAL and HANAUER 1971, ATAL and SCHROEDER 1970a, RABINER and SCHAFER 1978] with the classical GRAM-SCHMIDT orthogonalization procedure [MARKEL and GRAY 1976] being popular. Once the covariance matrix decomposition is performed, a simple recursive process determines the predictor coefficients, a_k. The details of the recursive process are found in FADDEEV and FADDEEVA [1963] with a simple version being presented in RABINER and SCHAFER [1978].

The covariance matrix can be singular, which implies that the stability of the covariance formulation of linear prediction cannot be guaranteed. To a certain degree the stability of the covariance formulation is dependent on the number of signal samples, N, used to calculate the covariances. In practice, the stability of the covariance formulation increases as the value of N increases [RABINER and SCHAFER 1978]. In the limiting case where N is very large, then the covariance formulation approximates the autocorrelation formulation for which stability is guaranteed.

The major advantage of the covariance formulation is its ability to identify an all pole filter from a finite portion of its impulse response. Theoretically, the autocorrelation formulation requires the impulse response out to infinity, which in practice is not available, to perform an exact identification. However, in practical applications, the advantage of the covariance formulation to provide accurate identification from finite length signals must be weighed up against the possibility of singular covariance matrices and, hence, unstable solutions.

2.3.3 AUTOCORRELATION FORMULATION

In the autocorrelation formulation of linear prediction, knowledge of the signal $s(n)$ is required for all time. In practice, this knowledge is not possible, and so it is necessary to assume the signal $s(n)$ is zero outside the time interval over which the signal has been measured. As for the covariance formulation, the number of signal samples available to the linear prediction process is N, i.e. $s(n)$ is known for $0 \leqslant n \leqslant N-1$. An alternative to assuming the signal is zero outside a certain range of n is to window the infinite length signal. This windowing process multiplies the signal $s(n)$ by a window function, $W(n)$, which is zero outside the region $0 \leqslant n \leqslant N-1$, to produce the windowed signal $s'(n)$, i.e.

$$s'(n) = s(n)\,W(n) \qquad -\infty \leqslant n \leqslant \infty \qquad (2.50)$$

where

$$W(n) = \begin{array}{ll} \text{non zero value} & 0 \leqslant n \leqslant N-1 \\ \text{zero value} & n < 0 \text{ and } n > N-1 \end{array} \qquad (2.51)$$

The type of window function, $W(n)$, used does affect the results obtained by an autocorrelation formulation of linear predictive analysis on $s'(n)$. A discussion of the types of windows and their effects is presented by BLACKMAN and TUKEY 1958, MARKEL 1971a, MAKHOUL and WOLF 1972, BINGHAM et al 1967, EBERHARD 1973, WANG 1971, and WELCH 1961. If the signal $s(n)$ is stationary, then as N increases the effects of the window function, $W(n)$, decrease until in the limit, when N is very large, the window function is ignored completely. For simplicity, all the signals used in this thesis, in connection with an autocorrelation formulation of linear prediction, are windowed with a rectangular window, i.e.

$$W(n) = \begin{array}{ll} 1 & 0 \leqslant n \leqslant N-1 \\ 0 & n < 0 \text{ and } n > N-1 \end{array} \qquad (2.52)$$

unless otherwise stated. This implicit windowing allows the dash to be dropped from $s'(n)$ without producing ambiguities.

For the autocorrelation formulation of linear prediction, the summation limits over which the predictor error, $e(n)$, is minimized are from n equal to zero to $N+p-1$, i.e. $n_0 = 0$ and $n_1 = N+p-1$, so that E is written as

$$E = \sum_{n=0}^{N+p-1} e^2(n) \qquad (2.53)$$

Therefore, the quantity $\phi(j,k)$ is evaluated as

$$\phi(j,k) = \sum_{n=0}^{N+p-1} s(n-j)s(n-k) \qquad \begin{array}{l} 1 \leqslant j \leqslant p \\ 0 \leqslant k \leqslant p \end{array} \qquad (2.54)$$

From the limits of n, j and k it is easily seen that signal values $s(n)$ outside the window (where they are zero) are used to evaluate

$\phi(j,k)$. Elimination of these signal values from the evaluation of $\phi(j,k)$ allows Equation 2.54 to be written as

$$\phi(j,k) = \sum_{n=0}^{N-1-(j-k)} s(n)s(n+j-k) \qquad \begin{array}{l} 1 \leqslant j \leqslant p \\ 0 \leqslant k \leqslant p \end{array} \qquad (2.55)$$

Further simplification of Equation 2.55 is achieved by noting that the values of $\phi(j,k)$ do not depend on the absolute values of j and k, but only the difference between them, i.e. $(j-k)$. This is easily shown by changing the absolute values of j and k, but not the difference between them, e.g.

$$\phi(j+q,k+q) = \sum_{n=0}^{N-1-((j+q)-(k+q))} s(n)s(n+(j+q)-(k+q)$$

$$= \sum_{n=0}^{N-1-(j-k)} s(n)s(n+j-k)$$

$$= \phi(j,k) \qquad (2.56)$$

If i is used to denote the difference between j and k, i.e. $i = j-k$, then the symbol $R(i)$ is used to denote $\phi(j,k)$ for the autocorrelation formulation of linear prediction, i.e.

$$R(i) = \phi(j,k)\Big|_{i=j-k}$$

$$= \sum_{n=0}^{N-i-1} s(n)s(n+i) \qquad (2.57)$$

For N small, $R(i)$ as defined by Equation 2.57 represents the short time autocorrelation function [RABINER and SCHAFER 1978], and when N is very large the autocorrelation function of the signal $s(n)$. For simplicity $R(i)$ is referred to as the autocorrelation function at the ith (signal) lag, irrespective of the value of N.

In terms of the autocorrelation function, $R(i)$, the set of simultaneous equations which are used to obtain the predictor coefficients, a_k, i.e. Equation 2.43, is written as

$$R(i) = \sum_{i=1}^{p} a_k R(i-k) \qquad 1 \leqslant i \leqslant p \qquad (2.58)$$

This equation is generally referred to as the NORMAL EQUATION, and in matrix notation is written as

$$\underset{\sim}{R} = \underset{\sim}{a}\,\mathrm{R} \qquad (2.59)$$

where R is called the autocorrelation matrix, $\underset{\sim}{a}$ is the predictor coefficient vector, and $\underset{\sim}{R}$ is the autocorrelation vector. The Normal Equations in matrix form are presented in Figure 2.12.

The autocorrelation function is symmetric, i.e.

$$R(i) = R(-i) \qquad (2.60)$$

for all i, and so the $p \times p$ qutocorrelation matrix, R, is Toeplitz [GRENANDER and SZEGO 1958]. This Toeplitz property allows the solution of the Normal Equations to obtain the predictor vector, $\underset{\sim}{a}$, to be performed in a more efficient manner than for the covariance method. Numerous solution procedures exist, with the most popular and well-known being those of LEVINSON [1947], which was adapted by ROBINSON [1967], and DURBIN [1960]. All these solution procedures are recursive, which allows quick solutions with the same number of computations each time. At present, the most computationally efficient (i.e. requiring fewest computations) recursive solution procedure for the Normal Equations is that of LEROUX and GUEGUEN [1977].

$$\begin{array}{ccccc}
\begin{bmatrix} R(0) & R(1) & R(2) & \cdots & R(p-1) \\ R(1) & R(0) & R(1) & \cdots & R(p-2) \\ R(2) & R(1) & R(0) & \cdots & R(p-3) \\ \vdots & \vdots & \vdots & & \vdots \\ R(p-1) & R(p-2) & R(p-3) & \cdots & R(0) \end{bmatrix} & \times & \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{bmatrix} & = & \begin{bmatrix} R(1) \\ R(2) \\ R(3) \\ \vdots \\ R(p) \end{bmatrix} \\
\begin{gathered} p \times p \\ \text{AUTOCORRELATION} \\ \text{MATRIX, } R \end{gathered} & & \begin{gathered} p \times p \\ \text{PREDICTOR} \\ \text{COEFFICIENTS} \\ \underset{\sim}{a} \end{gathered} & & \begin{gathered} p \times 1 \\ \text{AUTOCORRELATION} \\ \text{VECTOR} \\ \underset{\sim}{R} \end{gathered}
\end{array}$$

FIGURE 2.12: The general form of the normal equations.

The simple form of the autocorrelation matrix and the efficient solution algorithms of the Normal Equations, i.e. Equation 2.58, implies that the autocorrleation formulation of linear prediction is generally faster than the covariance formulation. A Toeplitz matrix is guaranteed to be nonsingular, and it has been shown [MARKEL and GRAY 1976, MARKHOUL 1975, and MARKEL and GRAY 1973] that in the absence of computational inaccuracies, this Toeplitz property ensures the autocorrelation formulation identifies a stable all pole filter. This is a significant advantage over the covariance formulation for which stability cannot be guaranteed.

A disadvantage of the autocorrelation formulation is that it cannot perform a correct system identification unless all the output signal of this system is available. A truncation or windowing of the signal as proposed above does not allow the exact autocorrelation of the system's output signal to be calculated correctly. Hence, incorrect system identification results from incorrect autocorrelation values.

2.3.4 LATTICE FORMULATIONS

Another formulation of linear prediction, called the lattice formulation, is obtained by considering the internal structure of the inverse filter. Section 2.3.1 showed the equivalence between linear predictive analysis and inverse all pole filtering analysis. In relation to the autocorrelation and covariance formulations, the lattice formulation combines the steps of calculating covariances or autocorrelation, and then solving a set of linear simultaneous equations, into a single recursive algorithm.

The lattice formulation of linear prediction has the advantages of guaranteed filter identificiation stability without the need for windowing of the signal sequences. The sensitivity of filter parameter identification is small, and stability is also preserved when finite word length computations are used [MARKEL and GRAY 1976]. The number of computations required by the lattice formulation to perform a system identification is similar to that required by the covariance and autocorrelation formulations, although the early lattice formulations [ITAKURA and SAITO 1968, 1970 and 1971] required four times as many computations. A summary of various efficient lattice formulations has been presented by MAKHOUL [1977].

In the following section, a basic derivation of the lattice formulation of linear prediction from the inverse all pole filter is presented. The process of calculating various quantities within the lattice formulation is performed in many ways, and a description of traditional and recently formulated methods is presented in Section 2.3.4.2. The computationally efficient lattice methods, described by MAKHOUL [1977], are presented later in this thesis.

2.3.4.1 Basic Principles

The lattice formulation of linear prediction is derived from the internal structure of the inverse filter defined by Equation 2.37, i.e.

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \qquad (2.61)$$

At the ith recursive stage of the lattice formulation, a lattice or inverse filter is defined which is the linear predictor of order i for the analysed signal, $s(n)$. The inverse or lattice filter at this ith recursive step is denoted as $A^{(i)}(z)$, and its corresponding predictor coefficients are denoted as $A_k^{(i)}$. Using Equation 2.61 $A^{(i)}(z)$ is written as

$$A^{(i)}(z) = 1 - \sum_{k=1}^{i} a_k^{(i)} z^{-k} \tag{2.62}$$

An analysis of $s(n)$ is performed by having the input of the inverse filter, $A^{(i)}(z)$, as the signal $s(n)$ and the output of the inverse filter is then the predictor error denoted as $e^{(i)}(n)$. If $E^{(i)}(z)$ and $S(z)$ are used to denote the z transform of $e^{(i)}(n)$ and $s(n)$ respectively, then

$$E^{(i)}(z) = A^{(i)}(z)S(z) \tag{2.63}$$

The iterative procedures used to solve the Normal Equations, i.e. Equation 2.58, use a recursive updating procedure to determine the next predictor coefficient, a_k. In terms of $a_k^{(i)}$ this updating procedure has the form [DURBIN 1960, MAKHOUL 1975, WAKITA 1973, MARKEL and GRAY 1976, RABINER and SCHAFER 1978]

$$a_k^{(i)} = a_k^{(i-1)} - k_i a_{i-k}^{(i-1)} \qquad 1 \leqslant k \leqslant i-1 \tag{2.64}$$

where k_i is a constant whose relationship with the inverse or lattice filter is developed later in this section. Substitution of Equation 2.64 into the defintion of $A^{(i)}(z)$, i.e. Equation 2.62, enables the ith inverse filter to be written in terms of the $(i-1)$th inverse filter as

$$A^{(i)}(z) = A^{(i-1)}(z) - k_i z^{-i} A^{(i-1)}(z^{-1}) \qquad (2.65)$$

Using Equation 2.65 allows Equation 2.63 to be written as

$$E^{(i)}(z) = A^{(i-1)}(z)S(z) - k_i z^{-i} A^{(i-1)}(z^{-1})\, S(z) \qquad (2.66)$$

The first term of Equation 2.66 is recognised as the z transform of the predictor error for the (i-1)th inverse filter. The second term is simplifed by defining $B^{(i-1)}(z)$ as

$$B^{(i-1)}(z) = z^{-i+1} A^{(i-1)}(z^{-1})\, S(z) \qquad (2.67)$$

Therefore, Equation 2.66 reduces to

$$E^{(i)}(z) = E^{(i-1)}(z) - k_i z^{-1}\, B^{(i-1)}(z) \qquad (2.68)$$

By defining the z transform $B^{(i-1)}(z)$ as

$$B^{(i-1)}(z) = \sum_{n=-\infty}^{\infty} b^{(i-1)}(n) z^{-n} \qquad (2.69)$$

then the time domain form of Equation 2.68 is

$$e^{(i)}(n) = e^{(i-1)}(n) - k_i b^{(i-1)}(n-1) \qquad (2.70)$$

From the definition of $B^{(i-1)}(z)$, i.e. Equation 2.67

$$B^{(i)}(z) = z^{-i} A^{(i)}(z^{-1})\, S(z) \qquad (2.71)$$

which on substitution of Equation 2.65 becomes

$$B^{(i)}(z) = z^{-i} A^{(i-1)}(z^{-1})\, S(z) - k_i A^{(i-1)}(z)S(z) \qquad (2.72)$$

Using the definitions of $B^{(i-1)}(z)$ and $E^{(i-1)}(z)$, i.e. Equations 2.63 and 2.67 respectively, Equation 2.72 reduces to

$$B^{(i)}(z) = z^{-1} B^{(i-1)}(z) - k_i E^{(i-1)}(z) \qquad (2.73)$$

which in the time domain is

$$b^{(i)}(n) = b^{(i-1)}(n) - k_i e^{(i-1)}(n) \qquad (2.74)$$

The z domain Equation 2.68 and 2.73, or the equivalent time domain equation, i.e. Equations 2.70 and 2.74, form a set of recursive equations which enable continuous updating of the lattice or inverse filter, provided the constants k_i are known. The set of Equations 2.68 and 2.73 or Equations 2.70 and 2.74 is called the LATTICE EQUATIONS. Initial conditions for the lattice equations are obtained from boundary conditions. Since the linear prediction procedures developed here are to be applied to the acoustic tube model developed earlier in this chapter, the initial conditions of the lattice equations are generated from the boundary conditions of the acoustic tube model.

If the zeroth recursive stage corresponds to no prediction of $s(n)$, then the terminating condition of the acoustic tube model requires

$$E^{(0)}(z) = -B^{(0)}(z) = S(z) \qquad (2.75a)$$

or

$$e^{(0)}(n) = -b^{(0)}(n) = s(n) \qquad (2.75b)$$

to be the initial conditions of the lattice equations, in the z and time domains respectively. The signal flow diagram representation of the lattice equations and the above initial conditions is presented in Figure 2.13 for both the z and time domains. From Figure 2.13 it is easily seen why this formulation of linear prediction is called the lattice formulation.

An interpretation of the quantity $B^{(i)}(z)$ is obtained by considering the time domain equivalent of Equation 2.71, i.e.

$$b^{(i)}(n) = s(n-i) - \sum_{k=1}^{i} a_k^{(i)} s(n+k-i) \qquad (2.76)$$

Comparison of Equations 2.76 and 2.33 reveals that $b^{(i)}(n)$ is the predictor error for a linear prediction of the signal sample $s(n-i)$ from the signal sequence $s(n+k-i)$ for $1 \leqslant k \leqslant i$. A close examination of this sequence shows that it contains the same signal values that are used to predict $s(n)$ via Equation 2.31 (if $p = i$), but in the reverse or backward order. Therefore, $b^{(i)}(n)$ is referred to as the backward predictor error sequence for the ith stage of the lattice formulation. In line with this definition, $e^{(i)}(n)$ is referred to as the forward predictor error sequence for the ith stage of the lattice formulation.

To use the lattice formulation of linear prediction described above, it is necessary to know the value of k_i at the ith stage before proceeding onto the $(i+1)$th stage. There exists a large number of different ways for determining or obtaining a value of k_i, and each leads to a different procedure for the lattice formulation. The following section provides a brief description of the most popular procedures for the determination of k_i.

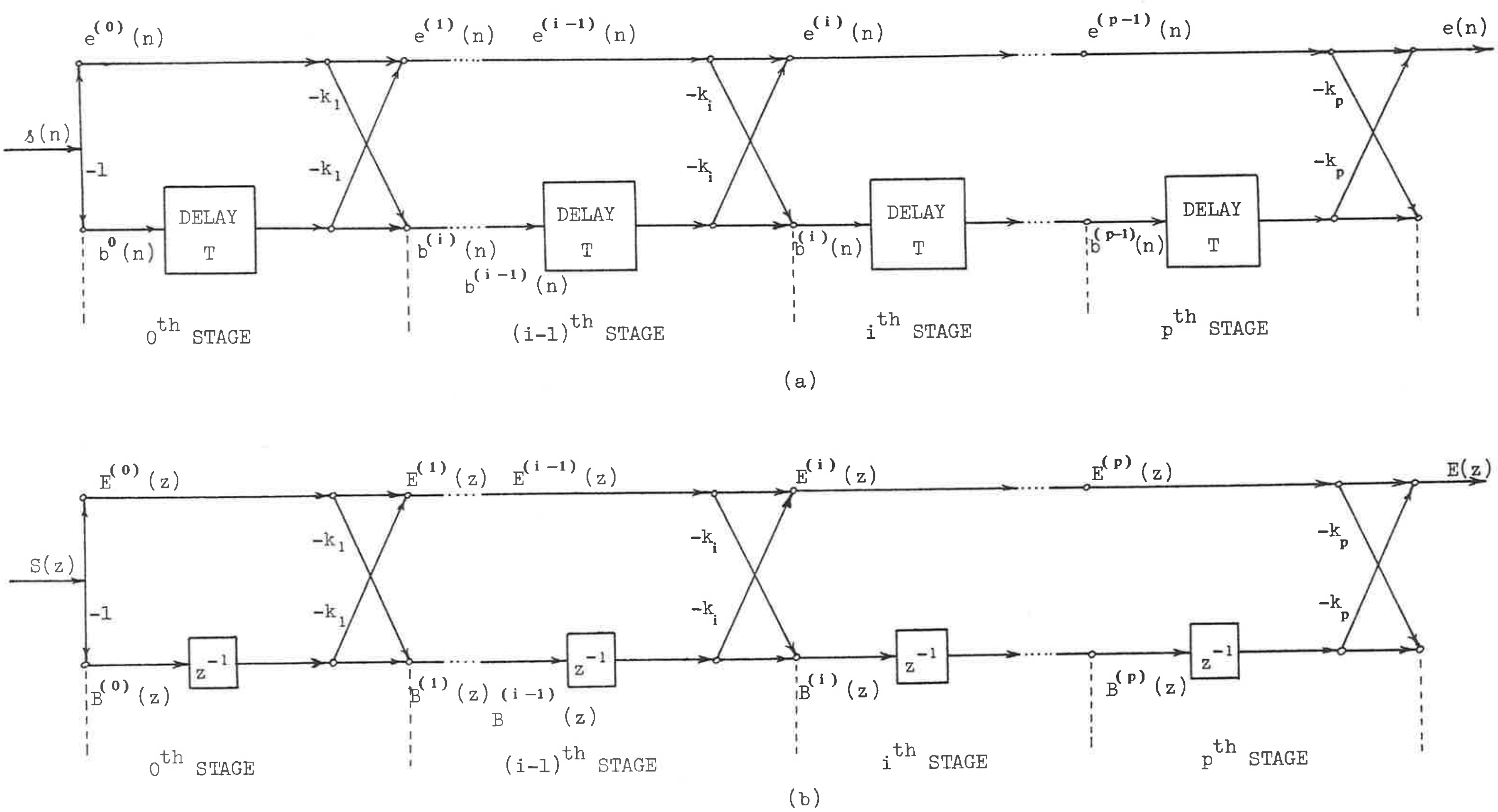


FIGURE 2.13: Signal flow diagrams for lattice equations in (a) the time domain and (b) the z domain.

2.3.4.2 Calculation of k_i

In 1969, ITAKURA and SAITO [1969] presented a technique for determining k_i from the forward and backward predictor error sequences of the (i-1)th stage of the lattice formulation of linear prediction. The criterion used to determine k_i is to simultaneously minimize the total squared forward and backward prediction errors, which results in a least mean square formulation of linear prediction. It has been shown [MARKEL and GRAY 1973] that a necessary and sufficient condition for simultaneous minimization of the forward and backward predictor errors is for the predictor error sequences at the ith stage to be orthogonal to all the predictor error sequences at preceding stages. This orthogonality requirement can be stated as

$$\sum_{n=-\infty}^{\infty} e^{(i)}(n)e^{(j)}(n) = 0 \qquad (2.77a)$$

and

$$\sum_{n=-\infty}^{\infty} b^{(i)}(n)b^{(j)}(n) = 0 \qquad (2.77b)$$

for $1 \leqslant j \leqslant i-1$.

The above orthogonality requirements, and hence the simultaneous minimization of the total squared forward and backward predictor errors, are satisfied by choosing k_i as a partial correlation coefficient, i.e.

$$k_i = \frac{\sum_{n=-\infty}^{\infty} e^{(i-1)}(n)b^{(i-1)}(n-1)}{\left[\sum_{n=-\infty}^{\infty}\left(e^{(i-1)}(n)\right)^2 \sum_{n=-\infty}^{\infty}\left(b^{(i-1)}(n-1)\right)^2\right]^{\frac{1}{2}}} \qquad (2.78)$$

The analysis procedure produced by ITAKURA and SAITO [1969] using Equation 2.78 in the lattice formulation is termed the PARCOR formulation of linear prediction, where the word PARCOR is derived from PARTIAL CORRELATION. The Parcor formulation is detailed in the flow chart of Figure 2.14 and the signal flow diagram of Figure 2.15, which is often referred to as the PARCOR LATTICE.

Another method for determining k_i was proposed by BURG [1968, 1975, RABINER and SCHAFER 1978], based upon minimizing the sum of the mean square forward and backward predictor errors. The resulting expression for k_i which minimizes this sum of predictor errors is

$$k_i = \frac{2\sum_{n=-\infty}^{\infty} e^{(i-1)}(n)b^{(i-1)}(n-1)}{\sum_{n=-\infty}^{\infty}\left(e^{(i-1)}(n)\right)^2 + \sum_{n=-\infty}^{\infty}\left(b^{(i-1)}(n)\right)^2} \tag{2.79}$$

A lattice formulation of linear prediction based on Equation 2.79 is performed in the same manner as the Parcor formulation. Therefore, the same flow chart and signal flow diagrams which describe the Parcor formulation also describe the Burg formulation if any reference to Equation 2.78 is replaced by a reference to Equation 2.79.

Minimizing the mean square error of the forward and backward prediction errors independently produces two more lattice formulations of linear prediction. Minimizing either prediction errors ensures the orthogonality of that prediction error as described by Equation 2.77. Minimizing the mean square error of the forward prediction error results in [MARKEL and GRAY 1973, 1976]

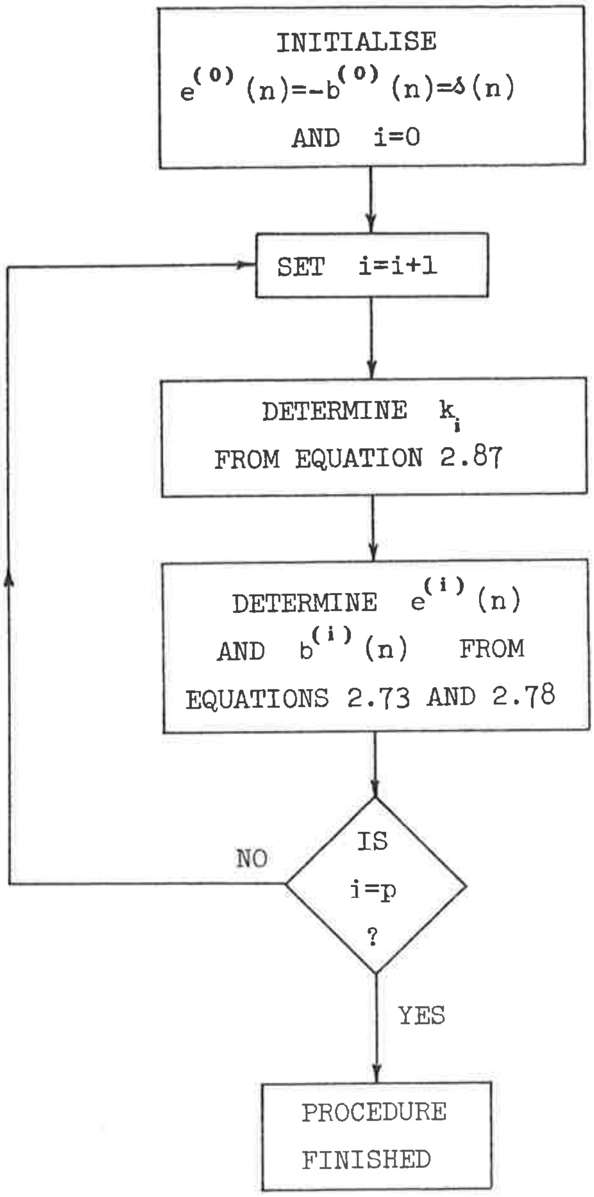


FIGURE 2.14: Flow chart for parcor formulation of a linear predictive analysis.

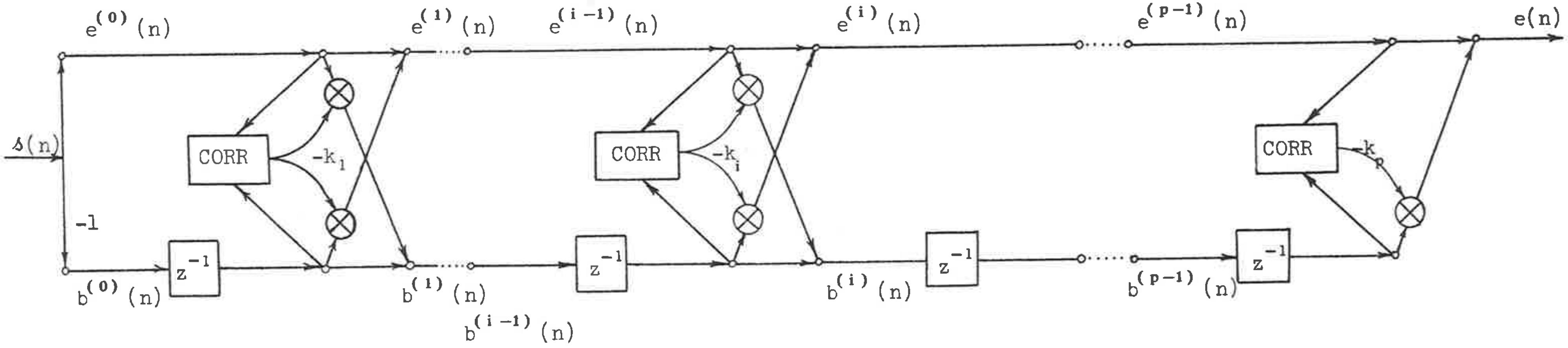


FIGURE 2.15: The parcor lattice.

$$k_i = \frac{\sum_{n=-\infty}^{\infty} e^{(i-1)}(n)b^{(i-1)}(n-1)}{\sum_{n=-\infty}^{\infty} \left(b^{(i-1)}(n-1)\right)^2} \tag{2.80}$$

The procedure based on k_i being determined by Equation 2.80 is the FORWARD lattice formulation of linear prediction, and is described by Figures 2.14 and 2.15 with all references to Equation 2.78 replaced by a reference to Equation 2.80.

Minimizing the mean square error of the backward prediction error yields [MARKEL and GRAY 1973, 1976]

$$k_i = \frac{\sum_{n=-\infty}^{\infty} e^{(i-1)}(n)b^{(i-1)}(n-1)}{\sum_{n=-\infty}^{\infty} \left(e^{(i-1)}(n)\right)^2} \tag{2.81}$$

The procedure based on k_i being determined by Equation 2.81 is called the BACKWARD lattice formulation of linear prediction. The flow chart of Figure 2.14 and the signal flow diagram of Figure 2.15 describe the BACKWARD lattice formulation if the references to Equation 2.78 are replaced by a reference to Equation 2.81.

Numerous other methods for calculating k_i exist which are all described by the general method defined by MAKHOUL [1977]. However, the lattice formulations presented above represent the most commonly used and popular lattice formulations. All the lattice formulations presented above guarantee the stability of the all pole filter they identify [MARKEL and GRAY 1976, MAKHOUL 1977, and RABINER and SCHAFER 1978], i.e.

$$-1 \leqslant k_i \leqslant 1 \tag{2.82}$$

for all i. For stationary signals, the k_i, as determined by the above lattice formulations with the boundary condition of Equation 2.75, are the same, i.e. all formulations are equivalent. This property is easily shown [MAKHOUL 1977] since under these conditions

$$\sum_{n=-\infty}^{\infty} \left(e^{(i-1)}(n)\right)^2 = \sum_{n=-\infty}^{\infty} \left(b^{(i-1)}(n-1)\right)^2 \tag{2.83}$$

for all i. It is also possible to show that not only are the various lattice formulations equivalent but the covariance, autocorrelation and lattice formulations are the same under these conditions [MAKHOUL 1977].

2.3.5 EXCITATION REQUIREMENTS OF LINEAR PREDICTION

For linear prediction to accurately identify a system from its output waveform, $s(n)$, that system must satisfy certain conditions. In this chapter it has been shown that linear prediction identifies an all pole filter; therefore, for a system to be identified correctly it must only be described by poles. This section defines the necessary excitation requirements of a system so that it is identified correctly by linear prediction.

The production of a signal $s(n)$ by a system and its subsequent analysis by linear prediction is depicted in Figure 2.16. The system transfer function, denoted as $H(z)$, has a gain constant G and when excited by a function $U(z)$ produces an output signal $S(z)$. Since $H(z)$ must be all pole for a correct identification, then

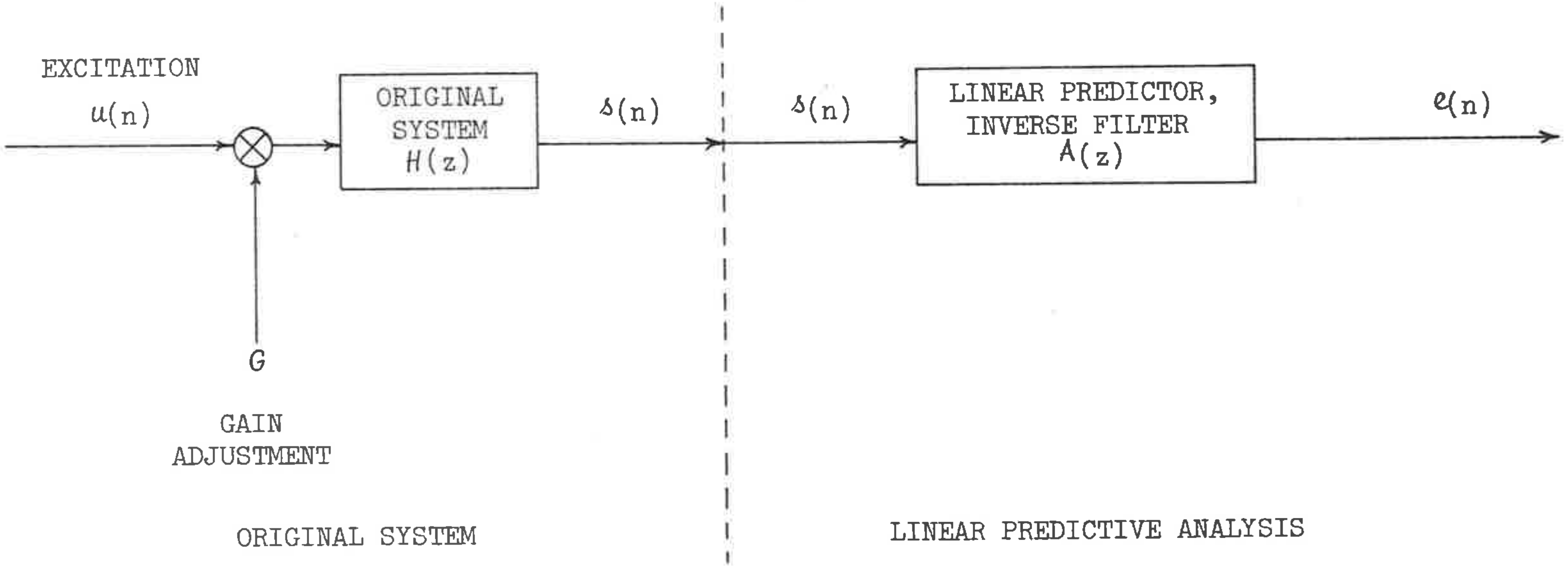


FIGURE 2.16: Generation of a signal $s(n)$ and its subsequent analysis by linear prediction.

$$H(z) = \frac{S(z)}{U(z)}$$

$$= \frac{G}{1 - \sum_{k=1}^{q} \alpha_k z^{-k}} \qquad (2.84)$$

where α_k are the system coefficients which define the poles of $H(z)$. In the time domain Equation 2.84 becomes

$$s(n) = \sum_{k=1}^{q} \alpha_k s(n-k) + Gu(n) \qquad (2.85)$$

A linear predictive analysis of the system output signal $s(n)$ defines an inverse filter $1/A(z)$ which filters $s(n)$ to produce the error sequence $e(n)$. In the z domain this filtering process is written as

$$\frac{1}{A(z)} = \frac{S(z)}{E(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \qquad (2.86)$$

and in the time domain as

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + e(n) \qquad (2.87)$$

An identification of the system $H(z)$ by linear prediction requires Equations 2.85 and 2.87 to be equivalent. Hence, the order of the linear predictor, p, must be equal to q so that the predictor coefficients identify the system coefficients correctly, i.e. $a_k = \alpha_k$. Also, the product of the gain and excitation of the system must be equal to the predictor error, i.e.

$$e(n) = Gu(n) \tag{2.88}$$

Therefore, the excitation of the system being identified must have the same properties as the predictor error of the linear prediction process.

It has been shown (MARKEL and GRAY 1974, pp 139-143) that the criterion for minimizing the mean square predictor error of the linear predictor is equivalent to choosing the inverse filter to maximize the spectral flatness of the predictor error. Hence, the linear predictor acts as a spectral whitening filter so that for a large enough predictor order, p, the predictor error signal $e(n)$ is white. In other words, the linear predictor removes all the predictable (or non-random) signal values of $s(n)$, and the non-predictable (or random) signal values appear as the error signal, $e(n)$.

Since the error signal has a white or flat spectrum, then by Equation 2.88 the excitation of the system, $u(n)$, must also have have a white or flat spectrum. Therefore, the ideal system excitation for linear prediction is white noise or an impulse. Any other system excitation does not allow Equation 2.85 to be equivalent to Equation 2.87 and so $a_k \neq \alpha_k$, i.e. the system is incorrectly identified.

2.4 RELATIONSHIP BETWEEN LINEAR PREDICTION AND THE ACOUSTIC TUBE MODEL

So far in this chapter the acoustic tube model and linear predictive analysis techniques have been defined and presented independently of one another. This section presents the relationship between linear prediction and the acoustic tube model, and in particular the manner in which the results of a linear predictive analysis may be used to identify an acoustic tube model.

The volume velocity waveforms within a set of commensurate acoustic tubes have been shown to satisfy Equations 2.25 and 2.26. Rearrangement of these equations so that the volume velocities in the ith acoustic tube are expressed in terms of those in the $(i+1)$th acoustic tube produces

$$U_i(n) = \frac{1}{1+\mu_i}\big(U_{i+1}(n) - \mu_i V_{i+1}(n-1)\big) \qquad (2.89)$$

and

$$V_i(n) = \frac{1}{1+\mu_i}\big(-\mu_i U_{i+1}(n) + V_{i+1}(n-1)\big) \qquad (2.90)$$

If a normalized forward travelling volume velocity is defined as

$$\bar{U}_i(n) = \prod_{j=1}^{i-1} (1+\mu_j) U_i(n) \qquad (2.91)$$

and a normalized reverse travelling volume velocity is defined as

$$\bar{V}_i(n) = \prod_{j=1}^{i-1} (1+\mu_j) V_i(n) \qquad (2.92)$$

then Equations 2.89 and 2.90 can be written as

$$\bar{U}_i(n) = \bar{U}_{i+1}(n) - \mu_i \bar{V}_{i+1}(n-1) \qquad (2.93)$$

and

$$V_i(n) = -\mu_i \bar{U}_{i+1}(n) + \bar{V}_{i+1}(n-1) \qquad (2.94)$$

respectively.

Comparison of the equations defining the lattice formulation of linear prediction, i.e. Equations 2.70 and 2.74, with Equations 2.93 and 2.94 above reveals a basis for an equivalence between the acoustic tube model and linear prediction. The acoustic tube model was derived as a synthesis model, whereas linear prediction is an analysis process, and so the source end of one is the termination end of the other. Hence, the jth acoustic tube of the acoustic tube model corresponds to the $(M-j)$th stage of the linear prediction process. Taking this into account allows

$$e^{(j)}(n) \equiv \bar{U}_{M-j}(n) \qquad 1 \leqslant j \leqslant M \qquad (2.95a)$$

and

$$b^{(j)}(n) \equiv \bar{V}_{M-j}(n) \qquad 1 \leqslant j \leqslant M \qquad (2.95b)$$

and hence

$$\mu_j = k_{M-j} \qquad 1 \leqslant j \leqslant M \qquad (2.95c)$$

The consequence of Equation 2.75 is that the acoustic tube model and linear prediction are equivalent. Equation 2.75c shows that the reflection coefficients of the acoustic tube model are determined directly from a linear predictive analysis of the output acoustic volume velocity of the acoustic tube model. The equivalence of the acoustic tube model and linear prediction was first presented by ATAL and HANAUER [1971] and later by WAKITA [1972, 1973] through an inverse filter description. This equivalence is only valid if all the assumptions of the acoustic tube model satisfy the assumptions of linear prediction and vice versa.

Equation 2.95c shows that the lattice formulations of linear prediction identify the reflection coefficients of a set of acoustic tubes. Conversion of reflection coefficients to an acoustic tube shape is found to be simpler (i.e. via Equation 2.16) than converting a_k's (which are recovered by the covariance and autocorrelation formulations of linear prediction) into an acoustic tube shape [MARKEL and GRAY 1976]. Therefore, in applications where acoustic tube shape recovery is the goal, lattice formulations of linear prediction are favoured.

2.5 THE CONCEPT OF AREA DISTANCE

Throughout this thesis a number of new and existing analysis procedures are used to recover the shape of acoustic tubes. To evaluate the accuracy of a recovered acoustic tube shape, synthetic data is used, which is generated from a known (referred to as original) acoustic tube shape via the acoustic tube model equations, detailed in Section 2.2. This knowledge of original and recovered acoustic tube shapes then allows a comparison to be made and, hence, an evaluation of the particular analysis process to recover acoustic tube shapes. The manner in which a comparison of acoustic tube shapes is made in this thesis is presented in this section.

During the development of speech recognition systems, a large number of distance measures have been defined [ITAKURA 1975, GRAY and MARKEL 1976, RABINER and SCHAFER 1978]. However, these distance measures are not suitable here, as they are designed to find a distance between a single data set and a large group of reference data sets. Also, those distance measures are designed for distances between predictor coefficients, a_k, and not cross-sectional areas of acoustic tubes, A_i, i.e. the requirements of the distance measures for the A_i's are different from those for the a_k's.

The set of original cross-sectional areas, A_i $(1 \leqslant i \leqslant M)$, is denoted by the vector $\underset{\sim}{A}$ and the recovered cross-sectional areas, A'_i $(1 \leqslant i \leqslant M)$, are denoted by the vector $\underset{\sim}{A}'$. Therefore, the distance measure required must determine the similarity between the vectors $\underset{\sim}{A}$ and $\underset{\sim}{A}'$. Many distance measures between vectors have been defined [FU 1980, DUDA and HART 1973] with the most well-known and used being a Euclidean distance, d_E, defined as

$$d_E = \left[\sum_{i=1}^{M} (A_i - A'_i)^2 \right]^{\frac{1}{2}} \qquad (2.96)$$

Of the many other distance measures, the most appropriate for the comparison of $\underset{\sim}{A}$ and $\underset{\sim}{A}'$ are

MINKOWSKY DISTANCE
$$d_M = \left[\sum_{i=1}^{M} |A_i - A'_i|^p \right]^{1/p} \qquad (2.97)$$

CAMBERRA DISTANCE
$$d_{CA} = \sum_{i=1}^{M} \frac{|A_i - A'_i|}{|A_i + A'_i|} \qquad (2.98)$$

CHEBYCHEV DISTANCE
$$d_{CH} = \max_{i} |A_i - A'_i| \qquad (2.99)$$

ABSOLUTE VALUE DISTANCE $$d_A = \sum_{i=1}^{M} |A_i - A'_i| \qquad (2.100)$$

CITY BLOCK DISTANCE $$d_{CB} = \sum_{i=1}^{M} w_i |A_i - A'_i| \qquad (2.101)$$

where w_i is a weighting function. The Absolute value and Euclidean distances are special cases of the Minkowsky distance, i.e. for $p = 1$ and $p = 2$ respectively.

Of the above distance measures, the Chebychev distance can be eliminated immediately as inappropriate because it provides the same distance for a recovered acoustic tube shape which is correct for all but one, where $A_i - A'_i \approx 5$ say, and an acoustic tube shape where $A_i - A'_i \approx 5$ for all i. In determining the similarity of recovered and original cross-sectional areas, it is important that the absolute value of A_i or A'_i be considered. This is shown by considering the case where $A_1 = 1$, $A'_1 = 1.3$, $A_2 = 5$ and $A'_2 = 5.4$, where A_2 is obviously recovered better than A_1 despite the fact that $|A_1 - A'_1| < |A_2 - A'_2|$. Therefore, the Minkowsky, Euclidean and Absolute value distance measures, which apply equal weighting to $|A_i - A'_i|$ regardless of the absolute values of A_i or A'_i, are not satisfactory in their present form.

The Camberra distance measure takes into account the absolute values of A_i and A'_i, with a division by $|A_i + A'_i|$. However, the Camberra distance measure does not produce the same distance for equidistant values of A'_i each side of A_i. If $A'_i = A_i + x$ then the Camberra distance is $\frac{x}{2A_i+x}$, whereas for $A'_i = A_i - x$ the Camberra distance is $\frac{x}{2A_i-x}$. If x is comparable with A_i then a significant bias of the distance occurs, which is undesirable. The City Block distance has the provision to apply a weight to the difference

$|A_i - A'_i|$ and so make an unbiased distance measure which takes into account the absolute values of A_i and A'_i. However, it is not clear how the weighting function could be chosen to satisfy such requirements.

Turning back to the Minkowsky, Euclidean and Absolute Value distances, a normalization of $|A_i - A'_i|$ before summation would remove the above stated deficiency of those distance measures. This normalization can be performed in many ways, and experiments show that a normalization by A_i is most appropriate. Experiments have also shown that values of p greater than 2 for the Minkowsky distance are not cost effective in terms of extra computational complexity. Therefore, the distance measure chosen is

$$d = \left[\sum_{i=1}^{M} \left(\frac{A_i - A'_i}{A_i} \right)^2 \right]^{\frac{1}{2}} \qquad (2.102)$$

which is a normalized Euclidean distance. It is easily verified that the same distance is achieved for equidistant A'_i around A_i, and that for $A_1 = 1$, $A'_1 \doteq 1.3$, $A_2 = 5$ and $A'_2 \doteq 5.4$ (the example used earlier) the distances obtained show A_2 is recovered better than A_1.

The distance measure defined by Equation 2.102 is referred to as the AREA DISTANCE, even though it is not a true distance measure. A true distance measure must satisfy the three axioms of reflexivity $(d(i,i) = 0)$, symmetry $(d(i,j) = d(j,i))$ and triangle inequality $(d(i,q) \leqslant d(i,p) + d(p,q))$, and the area distance does not satisfy symmetry. This is due to normalization by A_i only, which is necessary for two or more different tube shapes being recovered from the same synthetic data to have consistent area distances.

2.6 SUMMARY

This chapter has reviewed and defined the acoustic tube model and linear prediction procedures which are used throughout this thesis. Definition of the acoustic tubes within the acoustic tube model was presented from basic physical restrictions, the requirements of tractable mathematics and waveform sampling. A complete definition of an acoustic tube model of a set of commensurate acoustic tubes was described from a waveform description of a single acoustic tube and the junction of adjacent acoustic tubes. The acoustic tube model was presented in both a mathematical equation and signal flow diagram formulation.

After defining the process of linear prediction, a general mathematical description of the linear predictive analysis procedure was presented. Following this, the details of the three basic formulations of linear prediction, namely Covariance, Autocorrelation and Lattice, were presented. A comparison between each formulation was made and the applications in which each proves to be more effective than the others. The Lattice formulation of linear prediction can be implemented in many ways, and the most popular implementations were defined in this chapter.

Linear prediction has been used successfully as an analysis process for acoustic waveforms derived from sets of commensurate acoustic tubes, and the relationship between linear prediction and the acoustic tube model was presented in this chapter. This relationship was derived by the equivalence of the Lattice formulation of linear prediction with the equations of the signal flow diagram of the acoustic tube model. Hence, the manner in which a linear predictive analysis of an acoustic waveform identifies a unique set of commensurate acoustic tubes was presented.

Finally, a discussion of the various distance measures that can be employed to determine the similarity or dissimilarity of two acoustic tube shapes was presented. This discussion included the suitability of the various distance measures to the applications in this thesis, and resulted in the determination of a single distance measure that is used in the remainder of this thesis.

CHAPTER 3

NON-WHITE EXCITATION OF ACOUSTIC TUBES

3.1 INTRODUCTION

In both Chapters 1 and 2, linear prediction has been shown to be a powerful analysis tool, mainly due to the ease and simplicity with which linear prediction performs system identification. In Chapter 2 it was shown that, for linear prediction to recover an acoustic tube model from its output waveform, certain assumptions of the acoustic tube model and requirements of linear prediction must be satisfied. In most situations, many of these assumptions and requirements are not satisfied, especially in the speech application considered in this thesis and, as a result, inaccurate acoustic tube model recovery occurs.

A goal of the work presented in this thesis is to produce an analysis procedure which has more realistic assumptions and requirements than those of linear prediction and the acoustic tube model. To produce such a new analysis procedure, the effects of relaxing some of the assumptions of linear prediction are considered in turn. This chapter considers the effects of relaxing the white excitation requirement of linear prediction on the acoustic tube shape recovered by linear predictive procedures.

The research reported in this chapter was performed so that the effects of non-white excitation on acoustic tube shape recovery could be understood and be applied to improving vocal tract shape recovery from the speech waveform. The investigations resulted in a better understanding of the autocorrelation formulation of linear prediction and how non-white excitation modifies the recovered acoustic tube shape. A mathematically exact description of the effects of non-white excitation on acoustic tube shape recovery is very complex and, since there is no direct application to vocal tract shape recovery, this mathematical description is not presented. Instead, a number of typical examples are presented which illustrate each of the important investigation results. Excitation of the vocal tract is by positive only waveforms, and so the non-white excitations considered in this chapter are positive only, although in most cases the results and effects presented are also valid for general non-white excitations.

A white analog waveform has its frequency spectrum flat, i.e. has the same value, at all frequencies. This is slightly modified in the case of sampled waveforms which are used throughout this thesis. For samples at a frequency of f_s to accurately represent the analog waveform, that waveform must be bandlimited to exclude all positive frequencies above $f_s/2$ and negative frequencies below $-f_s/2$ (Nyquist Theorem). Therefore, a sampled waveform is white if its frequency spectrum is flat between the frequencies of $-f_s/2$ and $f_s/2$. If T is the sampling period, i.e. $T = 1/f_s$, a restriction of the frequency range to within $\pm f_s/2$ implies the longest duration for a white excitation waveform is $T/2$. Hence, the non-white excitations considered in this chapter have durations which are longer than $T/2$.

Section 3.2 investigates the type and magnitude of errors that occur for acoustic tube shapes recovered by linear prediction of the output waveform from acoustic tubes satisfying the assumptions of the acoustic tube model and excited by non-white excitations. A wide range of acoustic tube shapes and positive excitation waveforms is used so that a general conclusion about the effects of positive non-white excitations can be made. Section 3.2.1 provides a complete description of the autocorrelation formulation of linear prediction via a detailed description of the manner in which an acoustic tube model is identified. As a result of the understanding of the linear prediction process, an explanation of the acoustic tube shape recovered by linear prediction for certain non-white excitations is presented in Section 3.2.2.

The investigations reported in Sections 3.2.1 and 3.2.2 permit a prediction of the effects on acoustic tube shapes recovered by linear prediction for certain types of acoustic tube shapes and positive non-white excitations. Once the effects of a positive non-white excitation on acoustic tube shape recovery are known or predictable, then they can be removed to allow accurate acoustic tube shape recovery. The definition of acoustic tube shapes which permit the removal of positive non-white excitation effects, termed special acoustic tube shapes, is presented in Section 3.3. Two procedures by which special acoustic tube shapes can be recovered are presented in Section 3.4.

Relaxation of the white excitation assumption is possible if a constraint is placed on the acoustic tube shape, i.e. if it has to be one of the special acoustic tubes described in Section 3.4. The application areas for special acoustic tubes are considered in Section 3.5, with an important application being to non-commensurate sets of acoustic tubes which frequently occur in nature.

3.2 EFFECT OF NON-WHITE EXCITATION

For acoustic tube shape recovered by linear prediction to be the same as the original acoustic tube shape, certain constraints and assumptions, as detailed in Chapter 2, must be satisfied. This section investigates the effects of violating the white excitation requirement on the acoustic tube shape recovered by linear predictive procedures.

To study the effects of non-white excitation, it is necessary that no other effects such as radiation, losses, etc., cloud the results. Therefore, all the data used in this section is synthetic, with all the assumptions of linear prediction and the acoustic tube model satisfied, except for the white excitation assumption. Using synthetic data also has the advantage of being able to control the exact form of the non-white excitation and the original acoustic tube shape, hence permitting accurate comparisons of original and recovered acoustic tube shapes and, therefore, the determination of the effects of non-white excitation.

A detailed procedure for generating the synthetic data from the acoustic tube model is presented in Appendix B. Since no radiation effects are considered, the termination reflection coefficient is unity, i.e. $\mu_M = 1.0$. Using synthetic data ensures that the waveforms being analysed are stationary and are available for long time intervals. In Chapter 2 it was concluded that waveforms with the above properties are efficiently analysed by the autocorrelation formulation of linear prediction. Hence, all linear predictive analyses performed in this section are via the autocorrelation formulation, and in particular via the Parcor procedure (see Sections 2.3.3 and 2.3.4.2).

The first non-white excitation used is of rectangular shape, i.e. the excitation waveform $U_0(n)$ is defined as

$$U_0(n) = \begin{matrix} 1 & \quad 0 \leqslant n \leqslant L \\ 0 & \quad n < 0,\ n < L \end{matrix} \qquad (3.1)$$

where L is the duration of the excitation. This rectangular shaped excitation is a crude approximation of the typical excitations which often occur in reality. To use realistic acoustic tube shapes and ones which satisfy the speech applications considered later, the acoustic tube shapes used to generate synthetic data are derived from real vocal tract shapes. Five vocal tract shapes as measured by FANT [1970] for the Russian vowels $|a|$, $|e|$, $|i|$, $|o|$ and $|u|$ are used. A detailed description of acoustic tube shapes used to approximate these five vocal tract shapes is presented in Appendix C.

Synthetic data is generated, as shown in Appendix B, for a ten commensurate acoustic tube approximation of the vowels $|a|$, $|e|$, $|i|$, $|o|$ and $|u|$ (see Appendix C) with non-white excitation as defined by Equation 3.1 and L varying from one to four. The area distances between the original acoustic tube shapes and those recovered by a Parcor analysis of the synthetic data (see above) are presented in Table 3.1. A general trend of increasing area distance as L increases is observed in Table 3.1, which implies that larger errors in acoustic tube shape recovery occur as the excitation duration increases. A wide variation of area distances from one vowel to another is observed in Table 3.1, which suggests that the magnitude of the errors in acoustic tube shape recovery has a strong dependence on the original acoustic tube shape.

	VALUE OF L			
VOWEL	1	2	3	4
/a/	1.85	2.44	2.33	5.00
/e/	1.53	2.38	3.50	6.98
/i/	4.56	10.81	28.49	16.32
/o/	2.12	2.77	3.71	6.55
/u/	3.55	5.80	4.90	11.44

TABLE 3.1: Values of area distances for five vowel shapes and four excitation waveforms.

To show the actual errors that occur in acoustic tube shape recovery for the situations in Table 3.1 a plot of the original and recovered acoustic tube shapes for the vowels $|a|$ and $|i|$ are presented in Figure 3.1. The vowels $|a|$ and $|i|$ are chosen from the five in Table 3.1, as they represent the extremes, i.e. largest and smallest of area distances in Table 3.1. Figure 3.1 shows that large errors in acoustic tube shape recovery occur for the type of non-white excitation considered. As the duration of the excitation, i.e. L, increases, the recovered acoustic tube shape has little resemblance to the original acoustic tube shape. Figure 3.1 and Table 3.1 show that an area distance of greater than 2.0 implies gross errors in acoustic tube shape recovery.

To investigate the effects of non-white excitation further, a non-white excitation described by

$$u_0(n) = \begin{cases} 1 & n=0 \\ h & n=1 \\ 0 & n\neq 0 \text{ or } 1 \end{cases} \qquad (3.2)$$

is used instead of that described by Equation 3.1. The area distances between the original and recovered (by Parcor) acoustic tube shapes, when the non-white excitation described by Equation 3.1 is used with h varying from 0.02 to 0.8, are presented in Table 3.2. The results in Table 3.2 show an almost linear relationship between area distance and the value of h, i.e. when h doubles, the area distance approximately doubles.

A plot of the original and recovered acoustic tube shapes for the vowels $|e|$ and $|i|$, using the synthetic data generated for Table 3.2 is presented in Figures 3.2 and 3.3. Only the acoustic tube shapes for the vowels $|e|$ and $|i|$ are presented, as they represent the extremes, i.e. smallest and largest area distances, of the re-

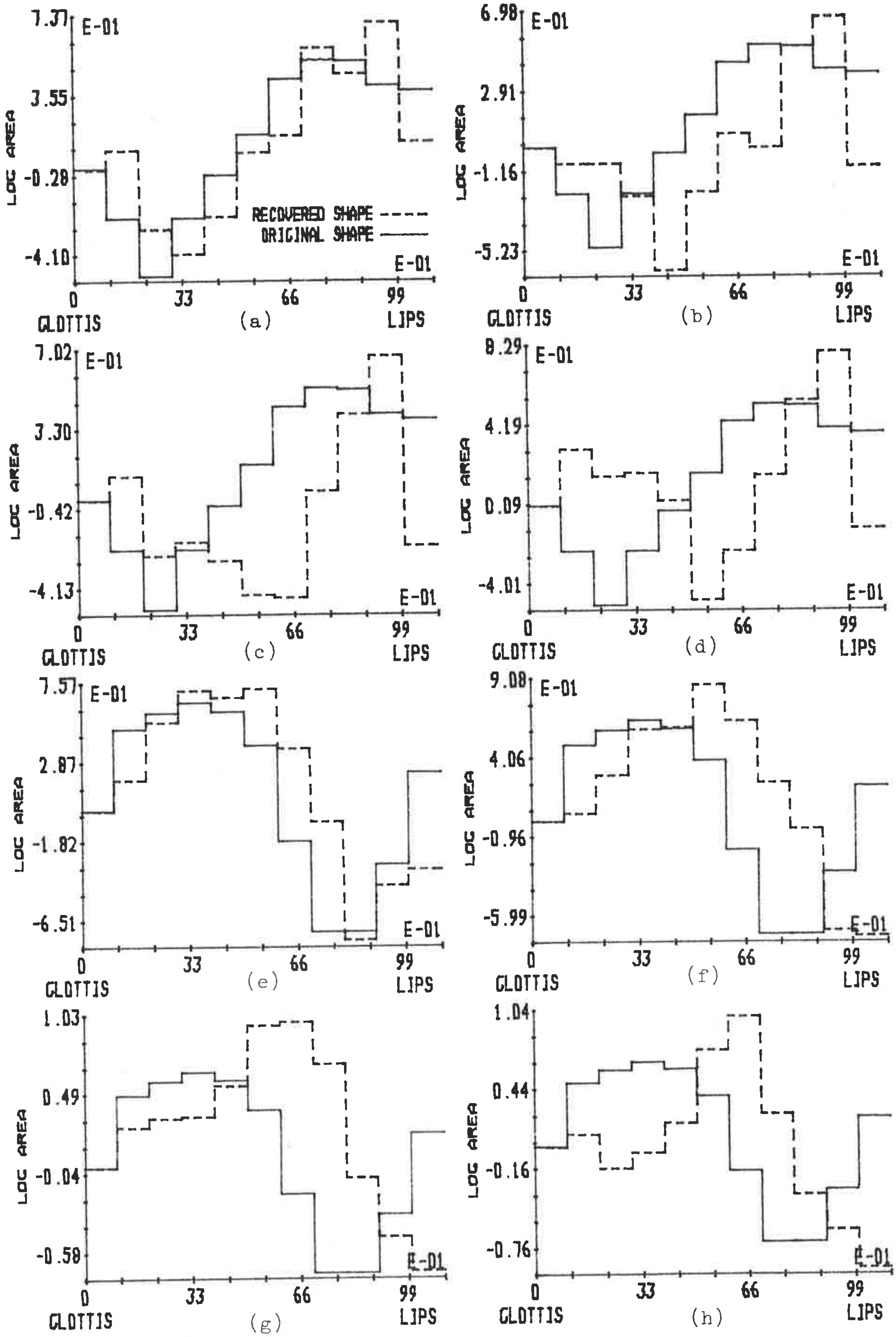


FIGURE 3.1: Comparison of recovered and original acoustic tube shapes for the vowel /a/ with excitation parameter L equal to (a) 1, (b) 2, (c) 3 and (d) 4, and for the vowel /i/ with L equal to (e) 1, (f) 2, (g) 3, and (h) 4.

	VALUE OF h						
VOWEL	0.02	0.05	0.10	0.20	0.40	0.60	0.80
/a/	0.08	0.18	0.33	0.56	0.91	1.26	1.63
/e/	0.07	0.17	0.33	0.60	1.00	1.24	1.42
/i/	0.11	0.28	0.57	1.15	2.33	3.40	4.19
/o/	0.08	0.19	0.35	0.60	0.98	1.33	1.75
/u/	0.09	0.21	0.39	0.72	1.42	2.25	3.10

TABLE 3.2: Values of area distances for five vowel shapes and various excitation waveforms.

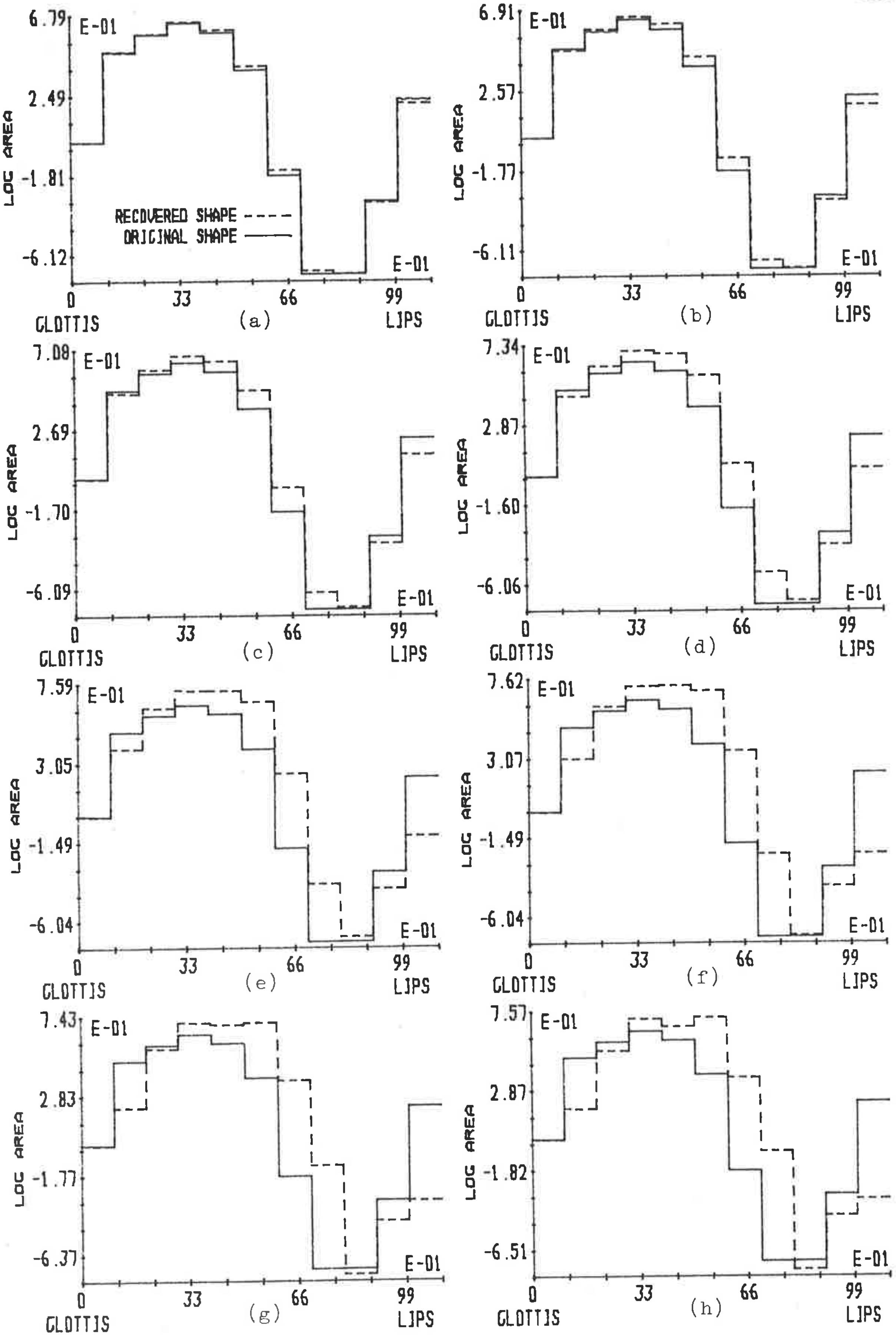


FIGURE 3.2: Comparison of recovered and original acoustic tube shapes for the vowel /i/ with excitation parameter h equal to (a) 0.02, (b) 0.05, (c) 0.1, (d) 0.2, (e) 0.4, (f) 0.6, (g) 0.8 and (h) 1.0.

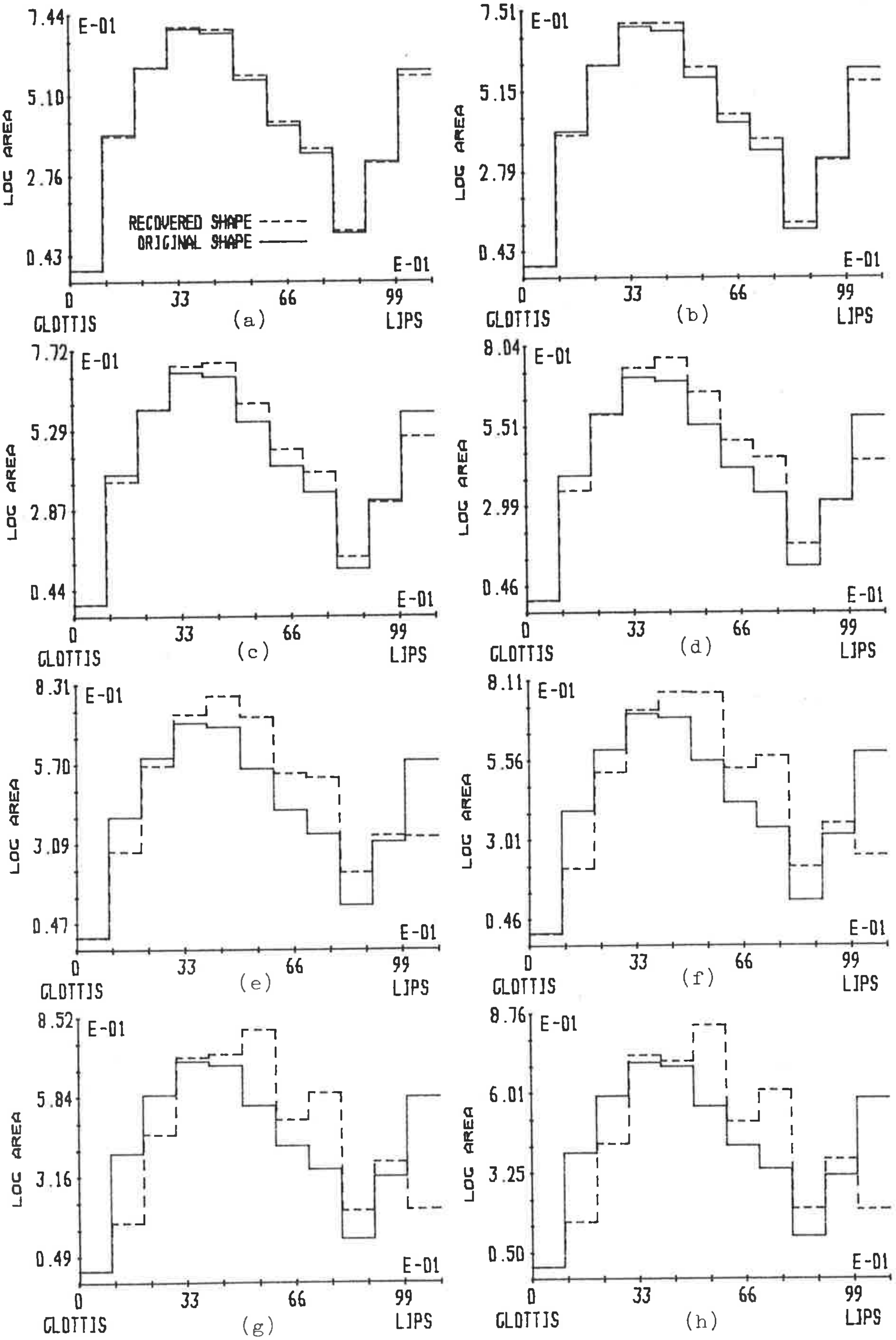


FIGURE 3.3: Comparison of recovered and original acoustic tube shapes for the vowel /e/ with excitation parameter h equal to (a) 0.02, (b) 0.05, (c) 0.1, (d) 0.2, (e) 0.4, (f) 0.6, (g) 0.8 and (h) 1.0.

sults presented in Table 3.2. Figures 3.2 and 3.3 show that significiant departures in the recoverd acoustic tube shape from the original acoustic tube shape only occur as h approaches unity. Hence, depending on the application, acceptable acoustic tube shapes are recoverd if only a relatively small departure from a white excitation occurs.

A comparison of area distances in Table 3.2 and the acoustic tube shapes in Figures 3.2 and 3.3 shows that reasonable acoustic tube shape recovery only occurs when h is small. Combining this result with the earlier one, where acceptable acoustic tube shape recovery occurred if the duration of the excitation was small, leads to the conclusion that small departures from the white excitation assumption can cause gross errors in acoustic tube shape recovery. Hence, in general, non-white excitation causes errors in acoustic tube shape recovery sufficiently large that little resemblance may occur between the original and recovered acoustic tube shapes.

To obtain an understanding of how non-white excitation causes errors in acoustic tube shape recovery, a much simpler set of acoustic tubes (i.e. having fewer cross-sectional area changes) is considered. The original and recovered acoustic tube shapes of a simple set of acoustic tubes is presented in Figure 3.4. The non-white excitation used for this example is defined by Equation 3.1 with $L = 1$. The recovered acoustic tube shape in Figure 3.4 resembles the original acoustic tube shape if the decaying oscillations are removed. Therefore, in this example the effect of the non-white excitation is to superimpose a decay-

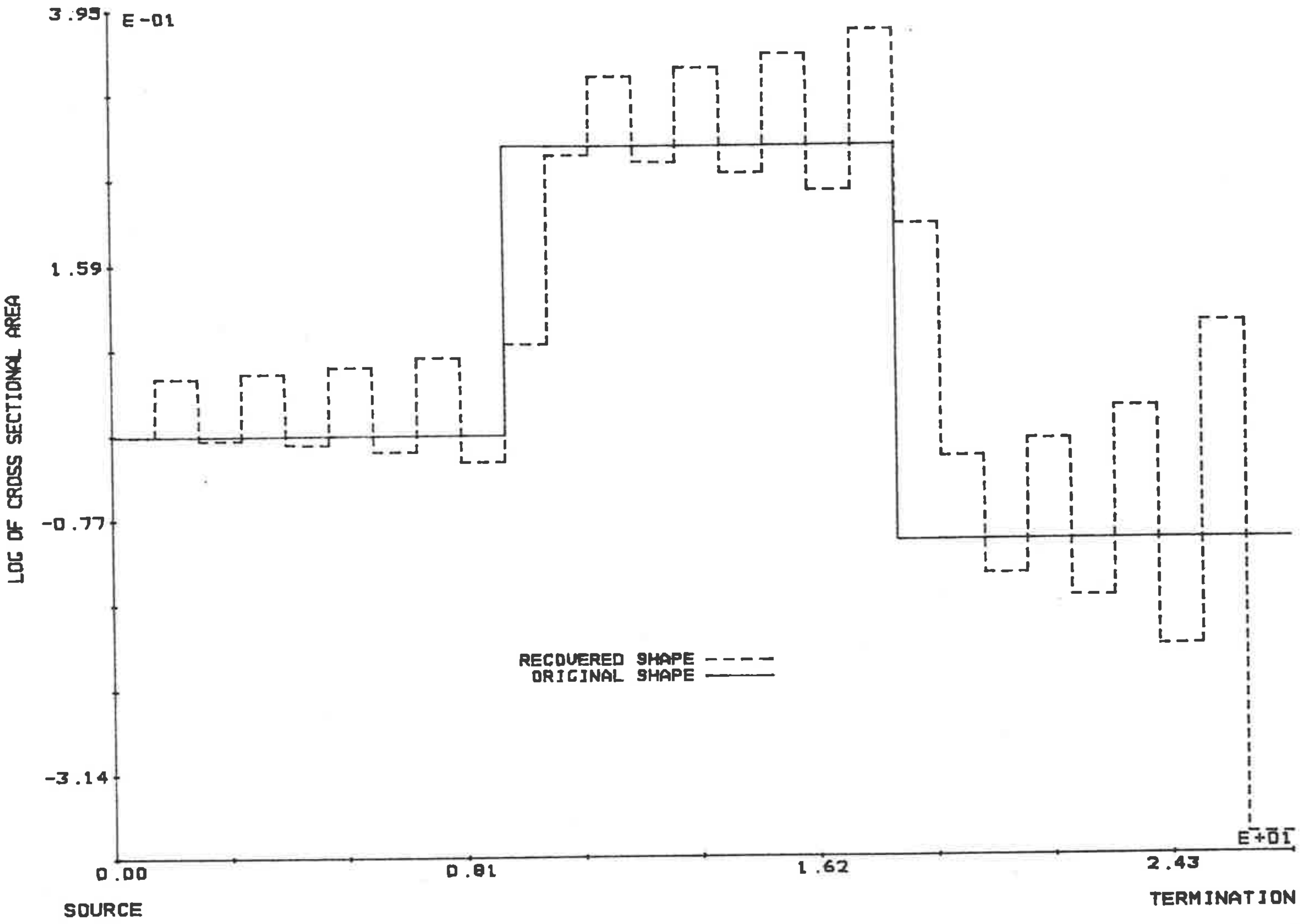


FIGURE 3.4: Comparison of recovered and original acoustic tube shapes for an excitation of two consecutive equal height impulses.

ing oscillation onto an acoustic tube shape which resembles the original.

If the mechanisms producing the observed decaying oscillations in the recovered acoustic tube shape of Figure 3.4 were understood, then the oscillations might be removed to provide a correct acoustic tube shape recovery. An understanding of the mechanisms producing the decaying oscillations described above requires a complete understanding of the manner in which the linear prediction process is performed. This understanding is developed in the following section. Section 3.2.2 then explains some of the effects of non-white excitation on acoustic tube shape recovery.

3.2.1 THE LINEAR PREDICTION PROCESS

Many researchers [MARKEL and GRAY 1973 and 1976, ATAL and HANAUER 1971, and MAKHOUL 1973 and 1975] have shown that the autocorrelation formulation of linear prediction performs autocorrelation matching. A linear prediction of order M produces an all pole filter such that the first $M+1$ autocorrelation values of its impulse response are a scaled match to the first $M+1$ autocorrelation values of the signal $s(n)$ being analysed. If the excitation of the system producing $s(n)$ has the same energy as the predictor error of the linear prediction process, then the autocorrelation matching is exact, i.e. the scale factor is unity.

In the following discussion the energies of the linear predictor error and the excitation of the system producing $s(n)$ are assumed to be the same so that an exact matching of autocorrelation functions occurs. This assumption is not restrictive in any way, since it has been shown [MAKHOUL 1975] that an arbitrary scaling of the autocorrelation function of $s(n)$ does not affect the results of a linear predictive analysis. This is verified later in this section, where it is shown that reflection coefficients are chosen from a ratio of autocorrelation values and not the absolute values of any autocorrelations.

The signal $s(n)$ being analysed is known over the infinite interval $0 \leqslant n \leqslant \infty$ and its autocorrelation function is denoted as $R(j)$ and defined by

$$R(j) = \sum_{n=0}^{\infty} s(n)s(n-j) \tag{3.3}$$

The impulse response, $\hat{s}(n)$, of the all pole filter identified by an ith order linear prediction of $s(n)$ is known over the same interval as $s(n)$, i.e. $0 \leqslant n \leqslant \infty$, and its autocorrelation function is denoted as $A^{(i)}(j)$ and defined by

$$A^{(i)}(j) = \sum_{n=0}^{\infty} \hat{s}(n)\hat{s}(n-j) \tag{3.4}$$

An Mth order linear prediction therefore produces

$$A^{(M)}(j) = R(j) \tag{3.5}$$

for $0 \leqslant j \leqslant M$.

A basic understanding of the linear prediction process is obtained by considering the autocorrelation matching for an ith and $(i+1)$th order linear prediction. An ith order linear prediction produces

$$A^{(i)}(j) = R(j) \tag{3.6}$$

for $0 \leqslant j \leqslant i$, and an $(i+1)$th order linear prediction produces

$$A^{(i+1)}(j) = R(j) \tag{3.7}$$

for $0 \leqslant j \leqslant i+1$. Therefore, in terms of an acoustic tube model (see Figure 3.5) linear prediction identifies the $(i+1)$th acoustic tube such that

$$A^{(i+1)}(i+1) = R(i+1) \tag{3.8}$$

The identification of the ith acoustic tube does not change the matching of $A^{(i)}(j)$ to $R(j)$ for $0 \leqslant j \leqslant i$ performed by previous acoustic tube identifications.

The reasons why linear prediction identifies the ith acoustic in the manner described above is explained by consideration of the acoustic tube model shown in Figure 3.5. If the acoustic tube model containing i acoustic tubes has been identified, then all the signal paths from which $\hat{s}(n)$ is produced have also been identified. When the $(i+1)$th acoustic tube is being identified by linear prediction, only one extra signal path is introduced and the paramenter k_{i+1} associated with this extra signal path is being identified by the linear prediction procedure. The extra signal path is from the $(i+1)$th acoustic tube to the termination and back to the $(i+1)$th acoustic tube. Since the time delay in this signal path is $(i+1)T$ then the contribution to $A^{(i+1)}(j)$ can only be for $j = (i+1)$, $2(i+1)$, $3(i+1)$, . . . etc. Hence, if linear prediction matches autocorrelations of the signals $s(n)$ and $\hat{s}(n)$ then, in identifying the $(i+1)$th acoustic tube, linear prediction must ensure

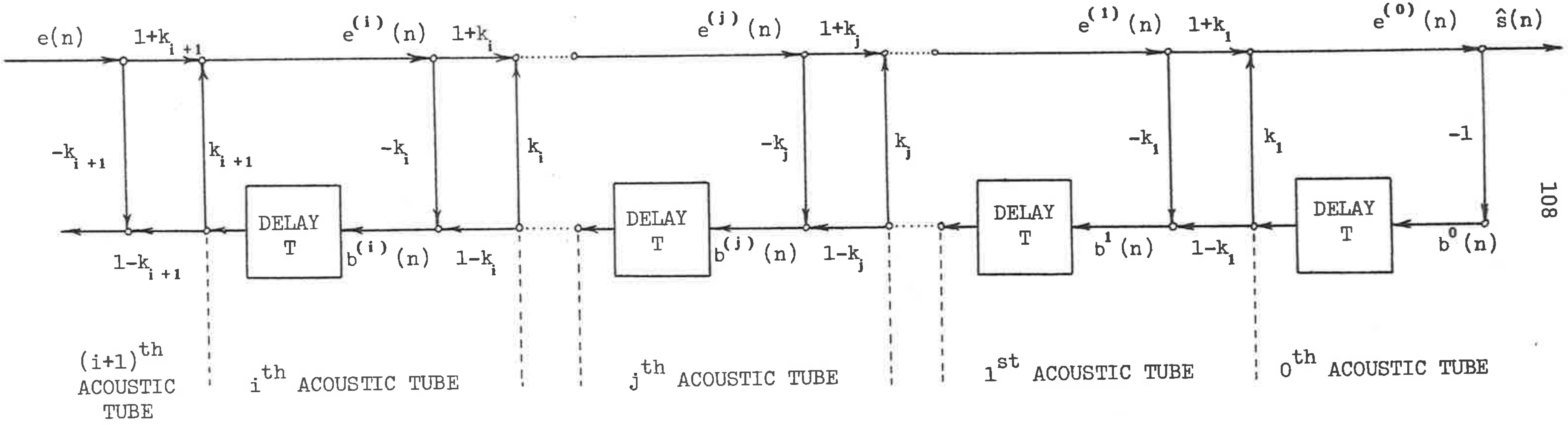


FIGURE 3.5: Flow diagram for the acoustic tube model of (i+1) commensurate acoustic tubes.

$A^{(i+1)}(i+1) = R(i+1)$ because the $(i+1)$th acoustic tube cannot change $A^{(k)}(i+1)$ for $0 \leqslant k \leqslant i$.

The next step in understanding the linear prediction process from an autocorrelation point of view is to establish the manner in which the ith acoustic tube is identified, i.e. the identification of k_i. Chapter 2 detailed some of the many methods for calculating k_i and showed for a termination reflection coefficient of unity and $s(n)$ being stationary that all the methods are equivalent. In terms of the acoustic tube variables defined in Figure 3.5, k_i is found as

$$k_i = \frac{\sum_{n=0}^{\infty} e^{(i-1)}(n) b^{(i-1)}(n-1)}{\sum_{n=0}^{\infty} \left(e^{(i-1)}(n)\right)^2} \tag{3.9}$$

(see Chapter 2). Equation 3.9 does not permit a general understanding of the manner in which linear prediction matches $A^{(i)}(i)$ to $R(i)$ when identifying the ith acoustic tube.

For the acoustic tube model defined in Figure 3.5, the signals $e^{(0)}(n)$ and $b^{(0)}(n)$ are determined directly from $\hat{s}(n)$ as

$$e^{(0)}(n) = \hat{s}(n) \tag{3.10a}$$

and

$$b^{(0)}(n) = -\hat{s}(n) \tag{3.10b}$$

Hence, by using Equations 3.10 and 3.9, k_1 is determined as

$$k_1 = -\frac{\sum_{n=0}^{\infty} \hat{s}(n)\hat{s}(n-1)}{\sum_{n=0}^{\infty} \left(\hat{s}(n)\right)^2} \tag{3.11}$$

Using the definition of $A^{(i)}(j)$, i.e. Equation 3.4, Equation 3.11 is rewritten as

$$k_1 = -\frac{A^{(1)}(1)}{A^{(1)}(0)} \qquad (3.12)$$

Since linear prediction matches $A^{(i)}(j)$ to $R(j)$ for $0 \leqslant j \leqslant i$, then the identification of k_1 matches $A^{(1)}(j)$ to $R(j)$ for $0 \leqslant j \leqslant 1$ and, hence, Equation 3.12 may be rewritten as

$$k_1 = -\frac{R(1)}{R(0)} \qquad (3.13)$$

Therefore, the first acoustic tube is identified from the negative ratio of $R(1)$ to $R(0)$.

From Equation 3.9 k_2 is defined by linear prediction as

$$k_2 = \frac{\sum_{n=0}^{\infty} e^{(1)}(n) b^{(1)}(n-1)}{\sum_{n=0}^{\infty} (e^{(1)}(n))^2} \qquad (3.14)$$

The acoustic tube model signal flow diagram of Figure 3.5 enables $e^{(1)}(n)$ and $b^{(1)}(n)$ to be expressed in terms of $\hat{s}(n)$ as

$$e^{(1)}(n) = \left(\frac{1}{1+k_1}\right)\left(\hat{s}(n) + k_1\hat{s}(n-1)\right) \qquad (3.15a)$$

and

$$b^{(1)}(n) = \left(\frac{-1}{1+k_1}\right)\left(k_1\hat{s}(n) + \hat{s}(n-1)\right) \qquad (3.15b)$$

Substitution of Equation 3.15 into the expression for k_2, i.e. Equation 3.14, using the definition of $A^{(i)}(j)$, i.e. Equation 3.4, and Equation 3.6 with $i = 2$, permits k_2 to be written as

$$k_2 = \frac{-\frac{R(2)}{R(0)} + k_1^2}{(1 - k_1^2)} \tag{3.16}$$

The impulse response for a single acoustic tube is obtained by setting the excitation, i.e. $e^{(1)}(n)$, to an impulse. Hence, using the signal flow diagram of Figure 3.5, the impulse response of a single acoustic tube is

$$s(n) = \begin{cases} 0 & n < 0 \\ 1 + k_1 & n = 0 \\ -k_1 s(n-1) & n > 0 \end{cases} \tag{3.17}$$

Substitution of Equation 3.17 into the definition of $A^{(i)}(j)$, i.e. Equation 3.4, with $i = 1$, defines the autocorrelation function $A^{(1)}(j)$ of a single acoustic tube as

$$A^{(1)}(j) = \frac{(1+k_1)}{(1-k_1)}(k_1)^j \tag{3.18}$$

Hence,

$$\frac{A^{(1)}(2)}{A^{(1)}(0)} = k_1^2 \tag{3.19}$$

and Equation 3.16 is then rewritten as

$$k_2 = \frac{-\frac{R(2)}{R(0)} + \frac{A^{(1)}(2)}{A^{(1)}(0)}}{(1-k_1^2)} \tag{3.20}$$

The normalized autocorrelation of $s(n)$ is denoted by $\bar{R}(j)$ and defined as

$$\bar{R}(j) = \frac{R(j)}{R(0)} \qquad j \geqslant 0 \qquad (3.21)$$

and the normalized autocorrelation function of $\hat{s}(n)$, generated from i commensurate acoustic tubes, is denoted by $\bar{A}^{(i)}(j)$ and defined as

$$\bar{A}^{(i)}(j) = \frac{A^{(i)}(j)}{A^{(i)}(0)} \qquad j \geqslant 0 \qquad (3.22)$$

Therefore, Equation 3.13 is rewritten as

$$k_1 = - \bar{R}(1) \qquad (3.23)$$

and Equation 3.20 as

$$k_2 = \frac{-\bar{R}(2) + \bar{A}^{(1)}(2)}{(1-k_1{}^2)} \qquad (3.24)$$

by using Equation 3.21 and 3.22.

Equation 3.24 shows that linear prediction calculates k_2 by adding $-\bar{R}(2)$ and $\bar{A}^{(1)}(2)$ and scaling the result by $(1-k_1{}^2)$. Therefore, the autocorrelation matching of $A^{(2)}(2)$ to $R(2)$ takes into account the normalized autocorrelation value $\bar{R}(2)$ of the signal $s(n)$ being analysed and the contribution of a single acoustic tube to the autocorrelation function of $\hat{s}(n)$ of two commensurate acoustic tubes, i.e. $A^{(1)}(2)$. The scale factor $1/(1-k_1{}^2)$ accounts for the scaling a signal from the second acoustic tube obtains in travelling to the termination (i.e. scaled by $1+k_1$) and then from the termination back to the second acoustic tube (i.e. scaled by $1-k_1$) (see Figure 3.5).

In general, the identification of k_i of an acoustic tube by linear prediction is performed via Equation 3.9, which is rewritten in terms of $\bar{R}(i)$ and $\bar{A}^{(i-1)}(i)$ as

$$k_i = \frac{-\bar{R}(i) + \bar{A}^{(i-1)}(i)}{\prod_{j=1}^{i-1}(1-k_j^2)} \qquad (3.25)$$

Hence, in general, the matching of $A^{(i)}(i)$ to $R(i)$ by linear prediction $\left(A^{(i)}(j)\right.$ already matched previously to $R(i)$ for $\left.0 \leqslant j \leqslant i-1\right)$ is performed by identifying k_i as the sum of the negative value of $\bar{R}(i)$ and the normalized autocorrelation contribution $\bar{A}^{(i-1)}(i)$ from the previous acoustic tubes that have been identified. A scaling by $1/\prod_{j=1}^{i-1}(1-k_j^2)$ is necessary to account for the scaling of a signal travelling from the ith acoustic tube to the termination $\left(\text{a scaling of } \prod_{j=1}^{i-1}(1+k_i)\right)$ and the scaling for that signal travelling back to the ith acoustic tube $\left(\text{a scaling of } \prod_{j=1}^{i-1}(1-k_i)\right)$ (see Figure 3.5).

The general expression for k_i, i.e. Equation 3.25, shows that only normalized autocorrelations are used to calculate k_i and, hence, a scaling of the autocorrelation functions does not affect the value of k_i. Therefore, relative rather than absolute autocorrelation function values are important for a linear predictive analysis of a signal $s(n)$.

Equation 3.25 shows that linear prediction takes into account the previous acoustic tubes that have been identified. This can be considered as a built-in memory which allows linear prediction to correct for incorrect acoustic tube identifications that may have occurred previously. Any corrections are subject to the constraint of matching $A^{(i)}(i)$ to $R(i)$ at the ith acoustic tube identification which may cause later acoustic tube identifications to progressively correct for one incorrect acoustic tube identification.

For non-white excitations, the observed acoustic tube shape recovered by linear prediction differs from the original acoustic tube shape firstly because of incorrect acoustic tube identifications by linear prediction and secondly due to the inbuilt correction process of linear prediction. The following section investigates, in detail, both the initial errors caused by non-white excitation and the correction process and their effects on acoustic tube shape recovery.

3.2.2 UNDERSTANDING SOME EFFECTS OF NON-WHITE EXCITATION

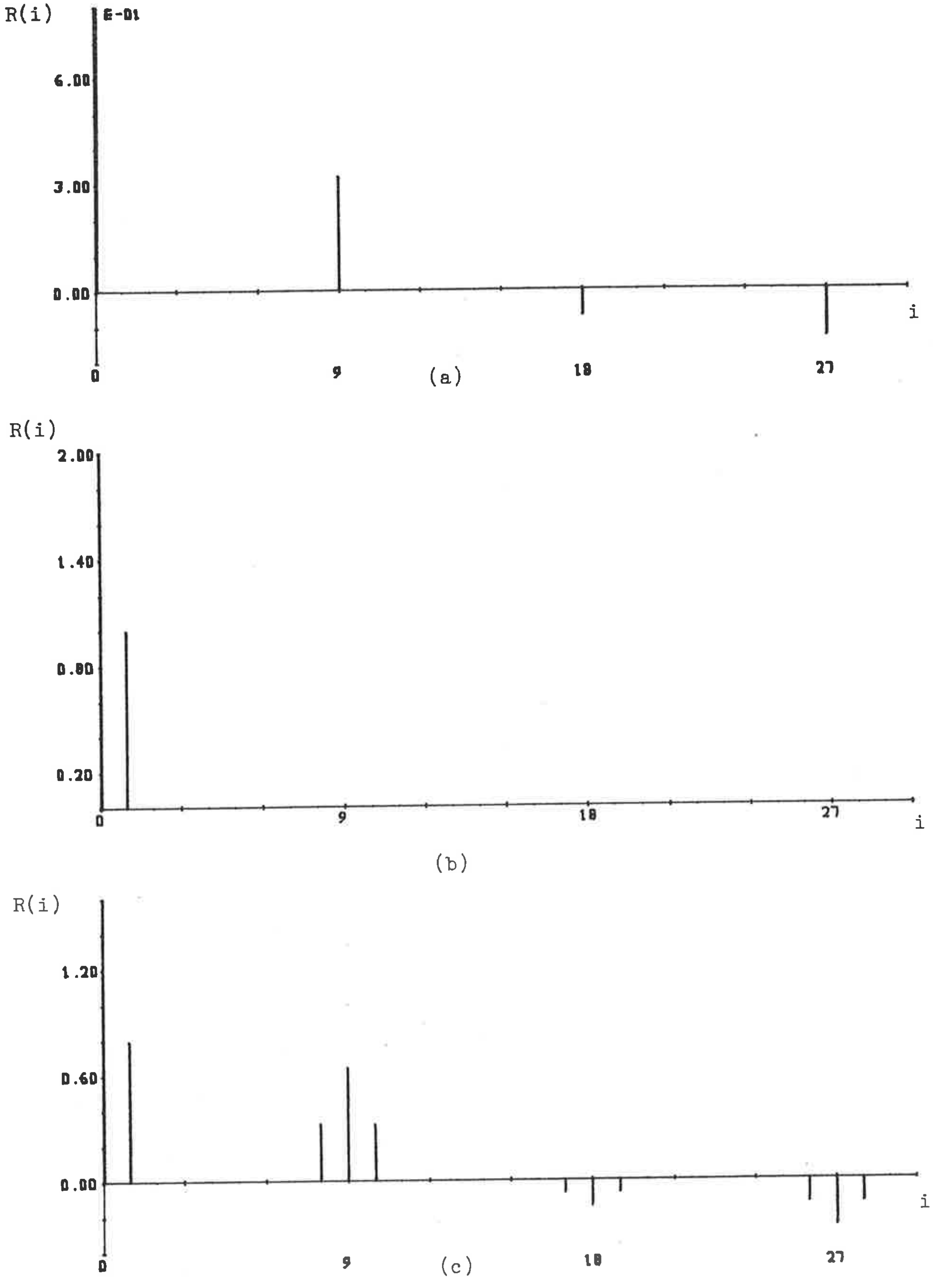
The previous section provided an understanding of the linear prediction process in terms of the acoustic tube model. It was shown that during the identification of the ith acoustic tube an autocorrelation from the (i-1)th order acoustic tube model, identified previously by linear prediction, is included in the identification process. This provides a memory path through which errors occurring at previous acoustic tube identifications may be corrected. Further investigations of this memory property of linear prediction are presented in this section for non-white excitation.

Identification of the ith acoustic tube is performed by the summation of $-\bar{R}(i)$ and $\bar{A}^{(i-1)}(i)$ and a division by the scale factor $\prod_{j=1}^{i-1}(1-k_j^2)$ (see Equation 3.25). In general, $\bar{R}(i)$ and $\bar{A}^{(i-1)}(i)$ are of similar magnitude, and so the error correction that occurs through $\bar{A}^{(i-1)}(i)$ is masked by $R(i)$. Therefore, to observe the error correction process it is necessary to choose an example where $\bar{R}(i)$ is zero for a range of i. Such a situation occurs in Figure 3.4 of Section 3.2. Figure 3.6 presents the autocorrelation functions for the impulse response of the original acoustic tubes of Figure 3.4, the non-white excitation function and the response of the original acoustic tubes when excited by the non-white excitation.

From Figure 3.6(c) and Equation 3.13 the first acoustic tube is identified by $k_1 = -0.5$, which is an incorrect identification of the original acoustic tube shape. The non-zero value of $R(1)$ is due to the non-white excitation and not the impulse response of the original acoustic tubes. The incorrect identification of the first acoustic tube causes non-zero autocorrelations $\bar{A}^{(1)}(j)$ for $j > 1$, as shown in Figure 3.7(a), and hence causes linear prediction to correct for the incorrect identification at subsequent acoustic tube identifications.

Identification of the second acoustic tube by k_2 is obtained by a scaling of $\bar{A}^{(1)}(2)$ $\left(\bar{R}(2)\right.$ is zero from Figure 3.6(c)$\left.\right)$ which from Figure 3.7(a) has $k_2 = +1/3$. The autocorrelation function of $\hat{s}(n)$ for the impulse response of the two acoustic tubes identified, i.e. $\bar{A}^{(2)}(j)$, is presented in Figure 3.7(b). A comparison of $\bar{A}^{(2)}(j)$ in Figure 3.7(b) with $\bar{R}(j)$ in Figure 3.6(c) shows autocorrelation matching has only occurred for $0 \leqslant j \leqslant 2$.

The identification of the third acoustic tube is obtained by scaling $\bar{A}^{(2)}(3)$ $\left(\bar{R}(3)\right.$ is zero from Figure 3.6(c)$\left.\right)$ which from Figure



(a)

(b)

(c)

FIGURE 3.6: Autocorrelation functions of (a) impulse response of acoustic tubes, (b) consecutive equal height impulses and (c) set of acoustic tubes excited by consecutive equal height impulses.

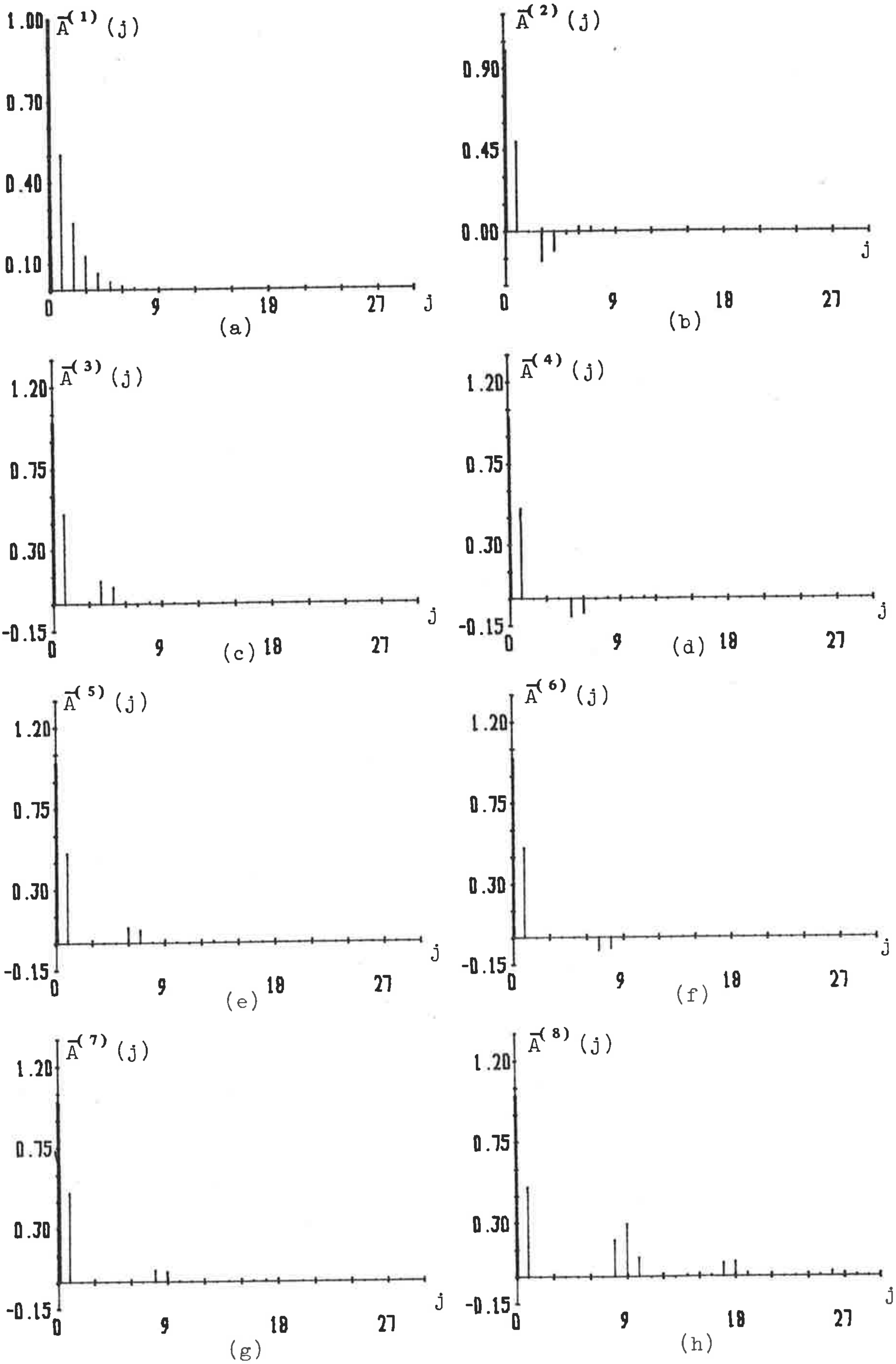


FIGURE 3.7: Autocorrelation functions for acoustic tubes identified by a linear predictive analysis.

3.7(b) defines $k_3 = -\frac{1}{4}$. Figure 3.7(c) presents $A^{(3)}(j)$, the autocorrelation function for $\hat{s}(n)$ derived from the three acoustic tubes identified. Comparison of Figures 3.7(c) with the normalized autocorrelation function $\bar{R}(j)$ in Figure 3.6(c) shows autocorrelation matching for only $0 \leqslant j \leqslant 3$. Hence, in general, the identification of the jth acoustic tube only corrects the value of $A^{(j)}(j)$ so that it matches $R(j)$. Therefore, it is necessary to identify an infinite number of acoustic tubes to completely correct for the original error in k_1 (due to the non-white excitation) which caused a mismatch of $A^{(1)}(j)$ to $R(j)$ for $j \geqslant 2$.

As more acoustic tubes are identified, it is found that the sign of k_j alternates and the modulus of k_j decreases for increasing j. From the properties of autocorrelation functions (e.g. $R(0) > R(j)$ for $j \geqslant 1$) it can be shown that, in general, the sign of k_j is opposite to that of k_{j+1} and $|k_j| > |k_{j+1}|$, for acoustic tubes such as those presented in Figure 3.4 and the positive excitation used. A translation of k_j's into a cross-sectional area of an acoustic tube, i.e. an A_i, is performed by

$$A_{j+1} = A_j \left(\frac{1 + k_j}{1 - k_j} \right) \tag{3.26}$$

which is derived from Equation 2.16. The properties of k_j given above translate via Equation 3.26 to a decaying oscillation of acoustic tube cross-sectional areas with a period of oscillation equal to the length of two acoustic tubes, i.e. as seen in Figure 3.4. The effect of the linear prediction correction process on the recovered acoustic tube shape is described above for a non-white excitation defined by

$$U_0(n) = \begin{cases} 1 & n=0 \\ a & n=1 \\ 0 & n\neq 0 \text{ or } 1 \end{cases} \tag{3.27}$$

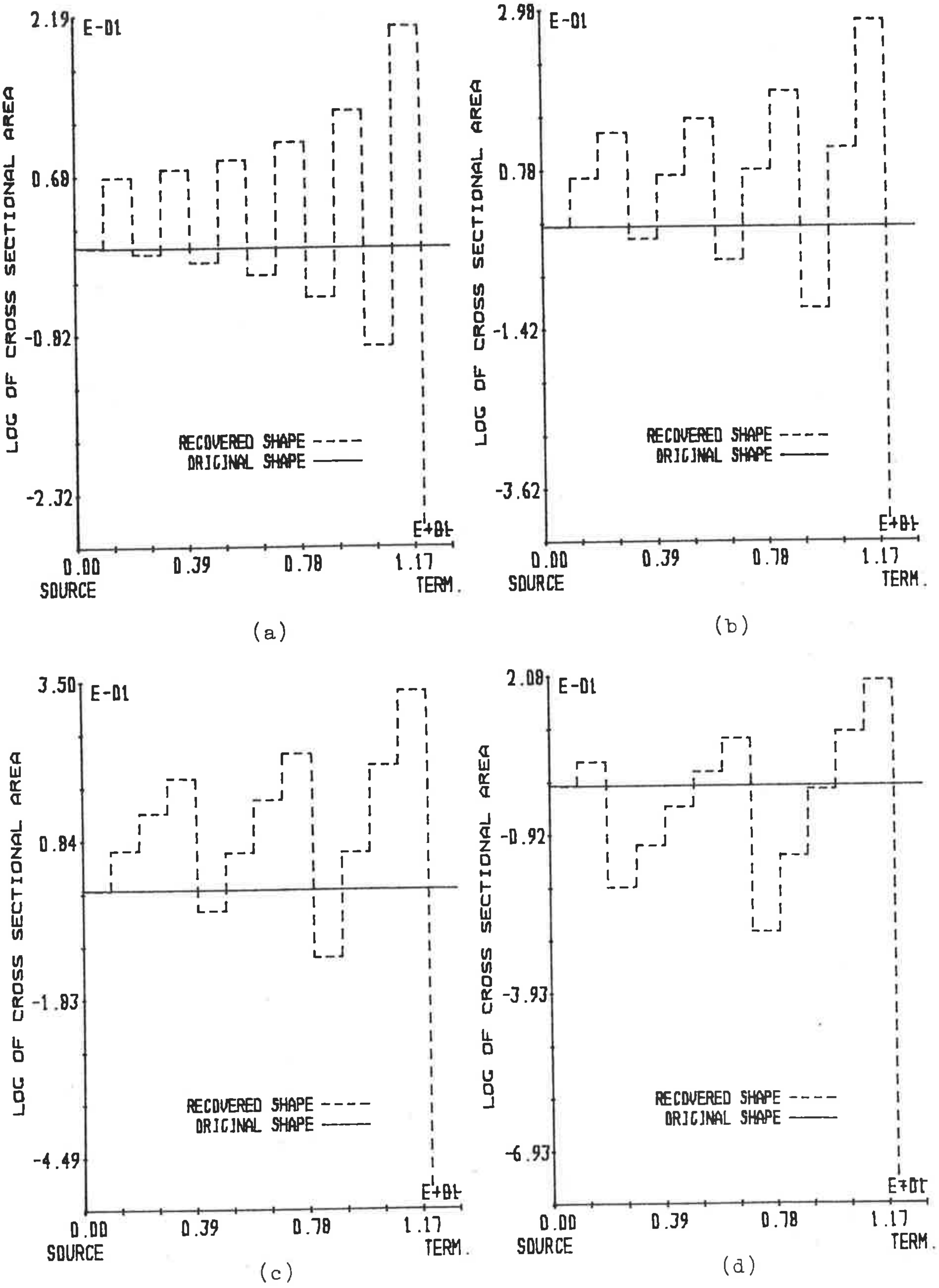
(note that this is a generalization of the non-white excitation considered above). The rate of decay of the oscillating acoustic tube shape is found to be only dependent on the value of a, and when a is small then the decay rate is quick with the slowest decay for an a of unity.

Figure 3.8 presents the recovered and original acoustic tube shapes for an excitation defined by

$$U_0(n) = \begin{cases} 1 & 0\leqslant n\leqslant N \\ 0 & n<0,\ n>N \end{cases} \tag{3.28}$$

with N having values from 1 to 4. The period of the decaying oscillations in the recovered acoustic tube shape is found from Figure 3.8 to be dependent on the duration of the excitation. The rate of decay of the recovered acoustic tube shape oscillations decreases as the duration of the excitation increases. In general, it is difficult to predict the period of oscillation and its decay rate, as both are very sensitive to changes in the shape of the non-white excitation.

The discussion so far has only considered the situation where there is no contribution to the autocorrelation function $\bar{R}(j)$ from a change in cross-sectional area in the acoustic tubes. In the example of Figure 3.4 the contribution of $\bar{R}(j)$ due to a change in acoustic tube shape occurs at $R(9)$, as seen in Figure 3.6. The non-white excitation causes non-zero autocorrelations around $R(9)$, namely at $R(8)$ and $R(10)$.



(a)

(b)

(c)

(d)

FIGURE 3.8: Comparison of recovered and original acoustic tube shapes for an excitation with (a) two, (b) three, (c) four and (d) five consecutive equal height impulses.

Identification of the eighth acoustic tube (via k_8) uses $\bar{R}(8)$ which is non-zero due to the non-white excitation and so causes an incorrect acoustic tube identification. This error causes linear prediction to perform the same type of correction procedure as detailed above at later acoustic tube identifications. The correction procedure for the error in identifying the first acoustic tube still occurs when the eighth acoustic tube is being identified; hence the exact error in k_8 due to the non-zero value of $\bar{R}(8)$ is masked by the correction procedure for the error in k_1.

From Equation 3.25 the identification of the ninth acoustic tube (via k_9) is performed by

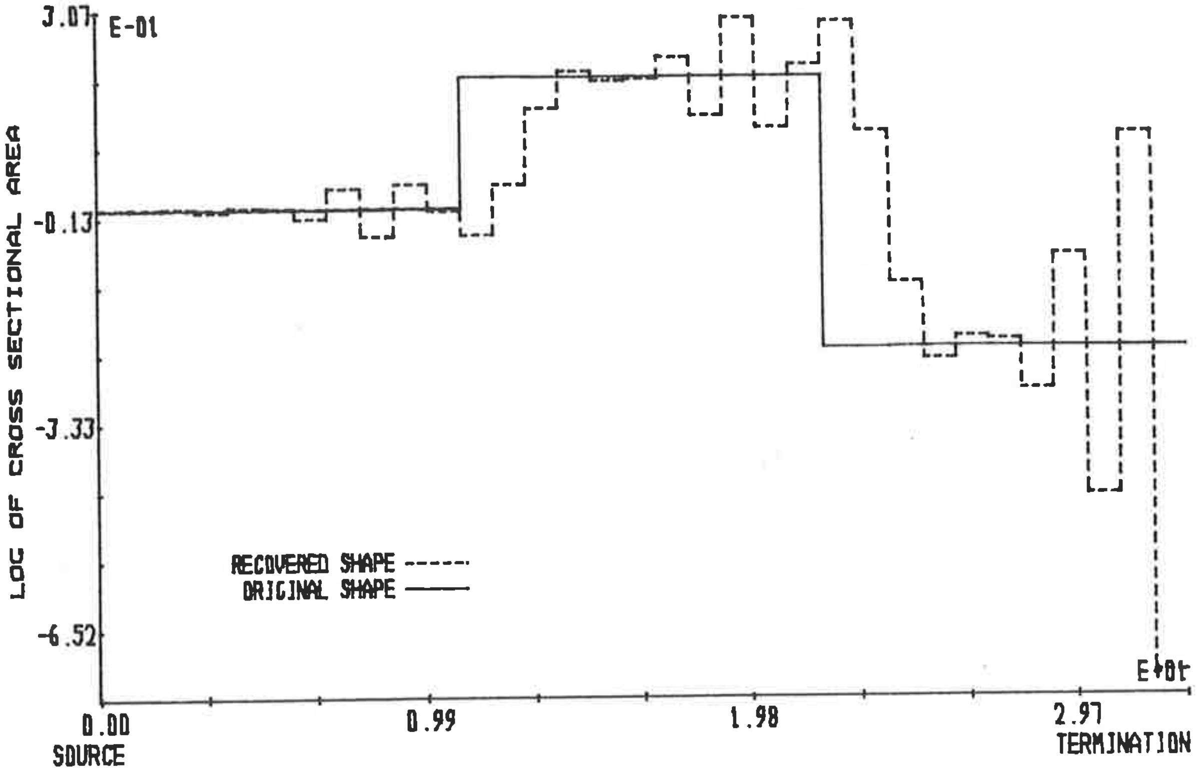
$$k_9 = \frac{-\bar{R}(9) + \bar{A}^{(8)}(9)}{\prod_{j=1}^{8}(1-k_j^2)} \tag{3.29}$$

The value of k_9 is correctly identified if $\bar{A}^{(8)}(9)$ is zero (which is the case for a white excitation), and so the correct identification of k_9 is masked by the correction of errors that occurred at previous acoustic tube identifications, via a non-zero $\bar{A}^{(8)}(9)$. The error in k_9 is corrected at subsequent acoustic tube identifications, and the correction process results in an oscillating recovered acoustic tube shape, as seen in Figure 3.4. If the decaying oscillation is permitted to continue on the recovered acoustic tube shape until it is zero, then all errors would have been corrected. Therefore, any change in acoustic tube shape can only be due to k_9 and, hence, k_9 can be identified if the error correction process is permitted to continue until all errors have been corrected.

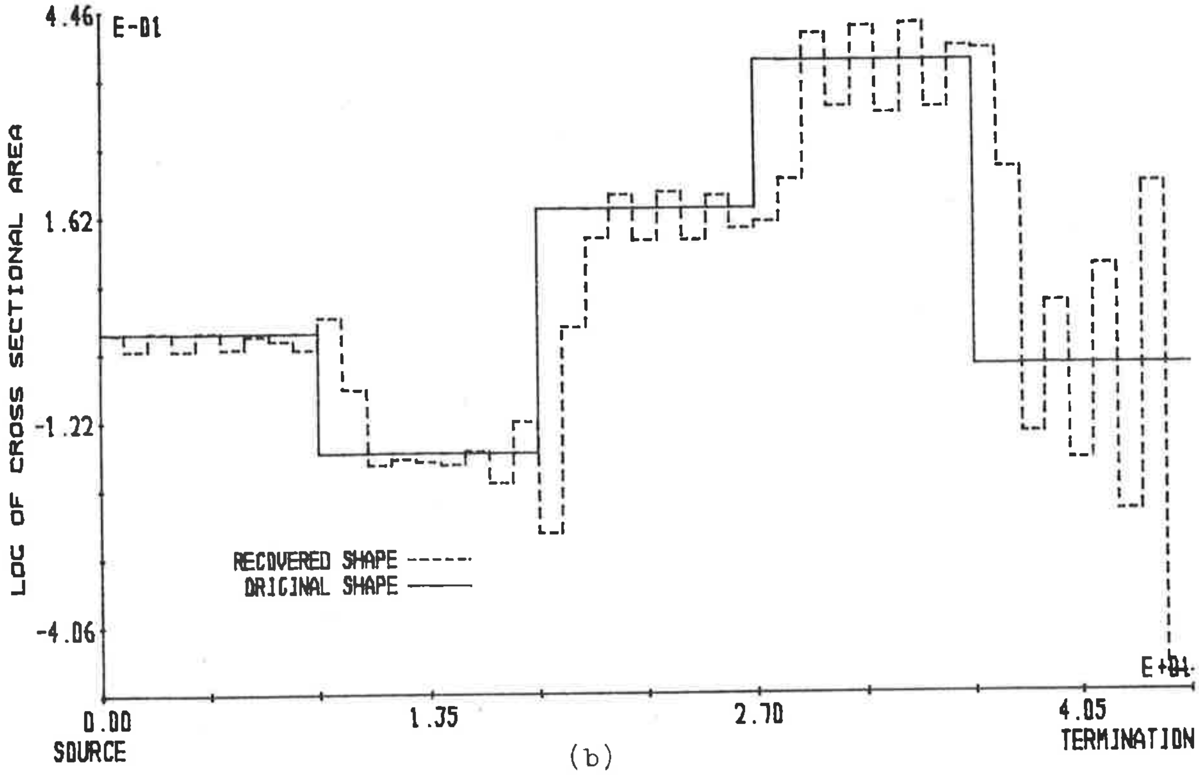
Each time an incorrect identification is made by linear prediction, a correction process continues at the following acoustic tube identifications. Therefore, if m incorrect identifications are made up until the identification of k_j the correction term used to calculate k_j, via $\bar{A}^{(j-1)}(j)$, contains m components of decaying acoustic tube shape oscillations. This, in general, has the effect of decreasing the rate of decay of the acoustic tube oscillations as more and more incorrect acoustic tube identifications are made, as observed in Figure 3.4.

From the above discussion it is concluded that the original acoustic tube shape may be recovered if the recovered acoustic tube oscillations are allowed to decay away completely. However, in general, the oscillating acoustic tube shape does not decay completely to a constant area value. Therefore, to recover the original acoustic tube shape, a procedure is necessary to determine the constant value that the oscillating acoustic tube shape decays to. A complete discussion of some methods for determining the constant value and the difficulties encountered is presented in Section 3.4.

As discussed earlier, the shape of the non-white excitation determines the rate of decay and period of the oscillating recovered acoustic tube shape. Figure 3.9 presents recovered acoustic tube shape for two different original acoustic tube shapes excited by more complex non-white waveforms than those used in Figure 3.4. Different decay rates of the oscillating recovered acoustic tube shape are illustrated in Figure 3.9.



(a)



(b)

FIGURE 3.9: Comparison of recovered and original acoustic tube shapes for excitation by (a) four consecutive impulses with heights 0.27, 0.76, 0.94 and 0.32 and (b) three consecutive impulses with heights 0.54, 0.93 and 0.37.

In the examples presented so far, i.e. Figure 3.4 and 3.9, the relatively few changes in original acoustic tube cross-sectional area are separated by a number of consecutive acoustic tubes with the same cross-sectional area. The effect on the recovered acoustic tube shape of reducing the number of consecutive acoustic tubes with the same cross-sectional area is presented in Figure 3.10. It is observed in Figure 3.10 that for decaying oscillations to be observed in the recovered acoustic tube shape a relatively large number of consecutive acoustic tubes with the same cross-sectional area must separate acoustic tube shape changes on the original acoustic tubes.

Therefore, the recovery of the original acoustic tube shape by removing decaying oscillations from the recovered acoustic tube shape caused by positive non-white excitations requires the original acoustic tube shape to satisfy special conditions. Thus, the relaxation of the constraint on the type of excitation is replaced by a constraint on the original acoustic tube shape. A full discussion and definition of the necessary constraints on the original acoustic tube shape is presented in the following section.

3.3 SPECIAL ACOUSTIC TUBE SHAPES

In the previous sections it has been shown that for positive non-white excitations and special conditions on the original acoustic tube shape decaying oscillations are observed on the recovered acoustic tube shape. It was suggested that the original acoustic tube shape can be recovered from such a situation. This section defines and discusses the necessary conditions that must occur on the original acoustic tube shape, hence allowing a recovery of the original acoustic tube shape.

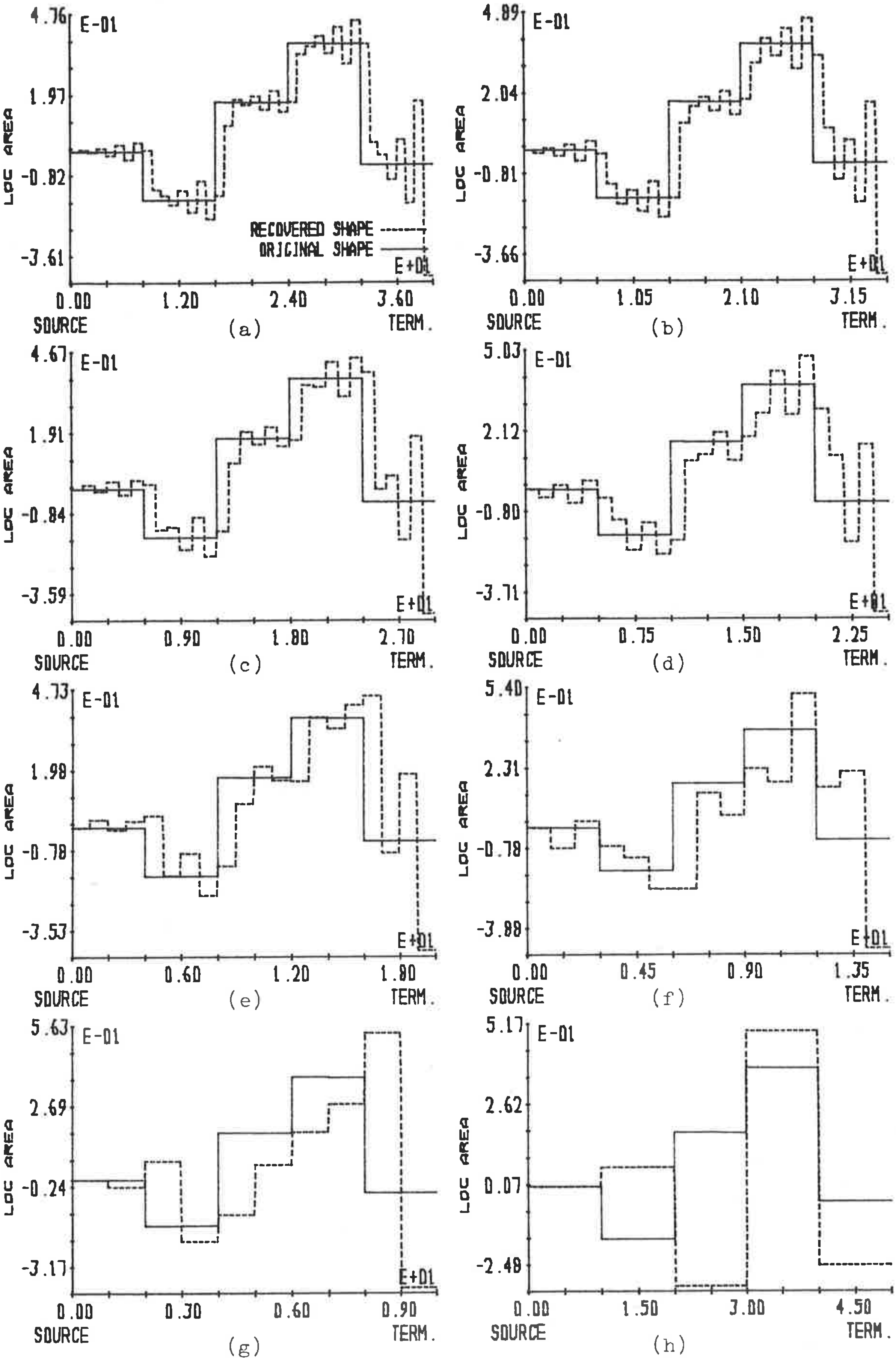


FIGURE 3.10: Comparison of recovered and original acoustic tube shape for different numbers of constant cross sectional area acoustic tubes between acoustic tube shape changes for an excitation of consecutive impulses with heights 0.46, 1.0 and 0.34.

Consider a set of acoustic tubes as illustrated in Figure 3.11. A typical normalized autocorrelation function for a signal generated from the acoustic tubes of Figure 3.11 with a positive non-white excitation of NT is presented in Figure 3.12. The period of oscillations on the recovered acoustic tube shape was shown, in Section 3.2.2, to depend on the shape of the non-white excitation, and for a positive non-white excitation of duration NT the maximum period of the oscillations on the recovered acoustic tube shape is N. If the number of acoustic tubes with the same cross-sectional area between the $(i\text{-}1)$th and ith change in acoustic tube shape from the termination is denoted by M_i then it is found empirically that M_i must be greater than or equal to $2N$ for at least one oscillation to be observed on the recovered acoustic tube shape.

Acoustic tube shapes which satisfy the above condition, i.e.

$$M_i \geqslant 2N \qquad (3.30a)$$

over all i are referred to as SPECIAL ACOUSTIC TUBE SHAPES. The condition described by Equation 3.30a restricts the value of M_i for a particular positive non-white excitation of duration NT. An inversion of Equation 3.30a, i.e.

$$N \leqslant \frac{M_s}{2} \qquad (3.30b)$$

where

$$M_s = \min_i M_i \qquad (3.30c)$$

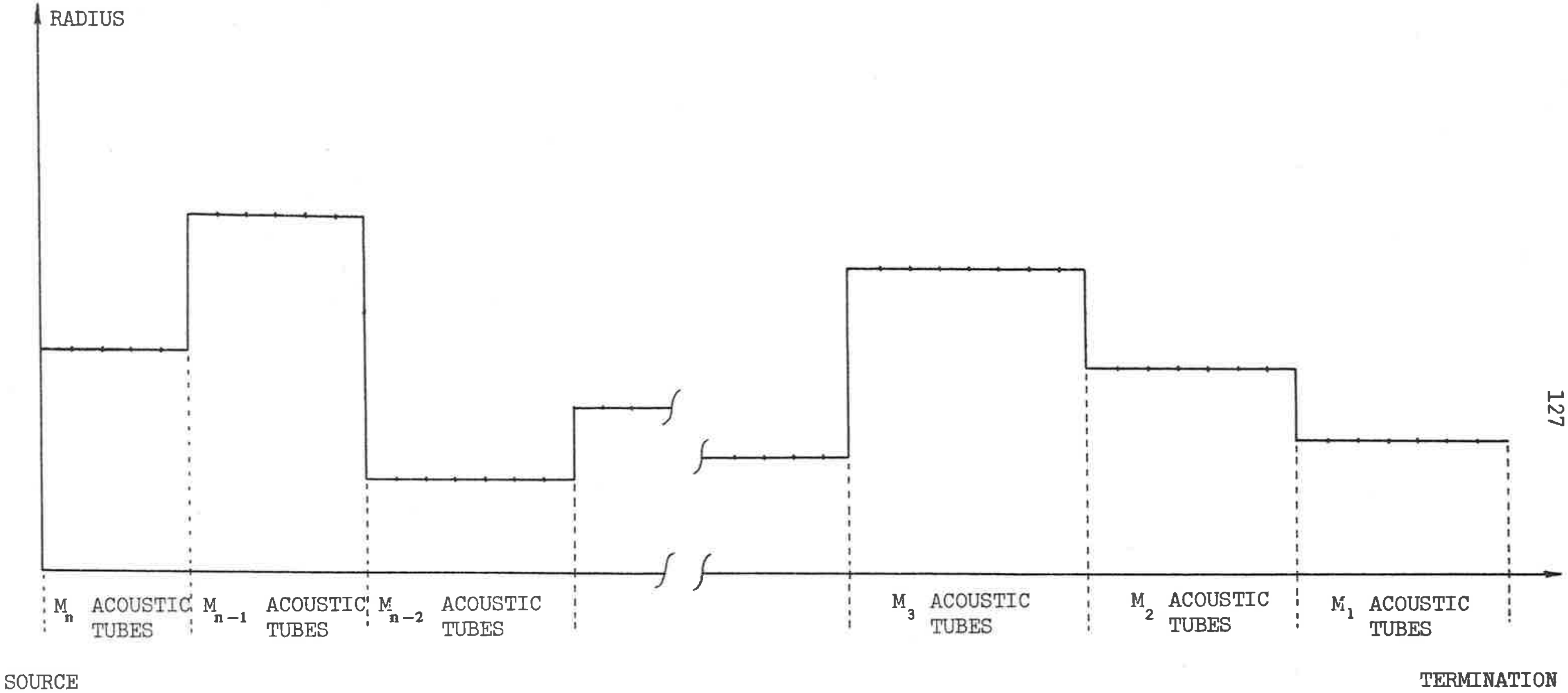


FIGURE 3.11: Set of acoustic tubes with sparse changes in cross sectional shape.

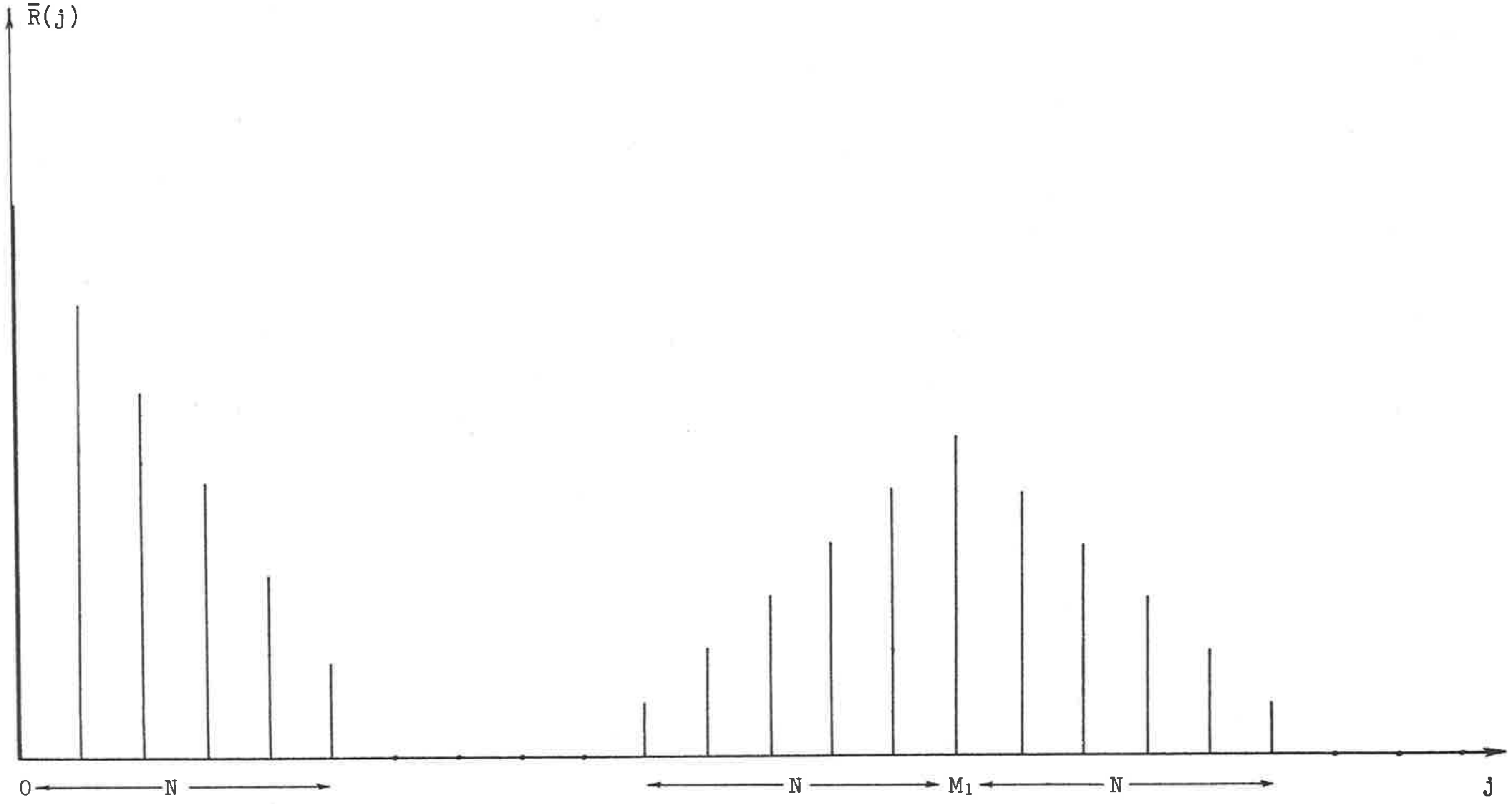
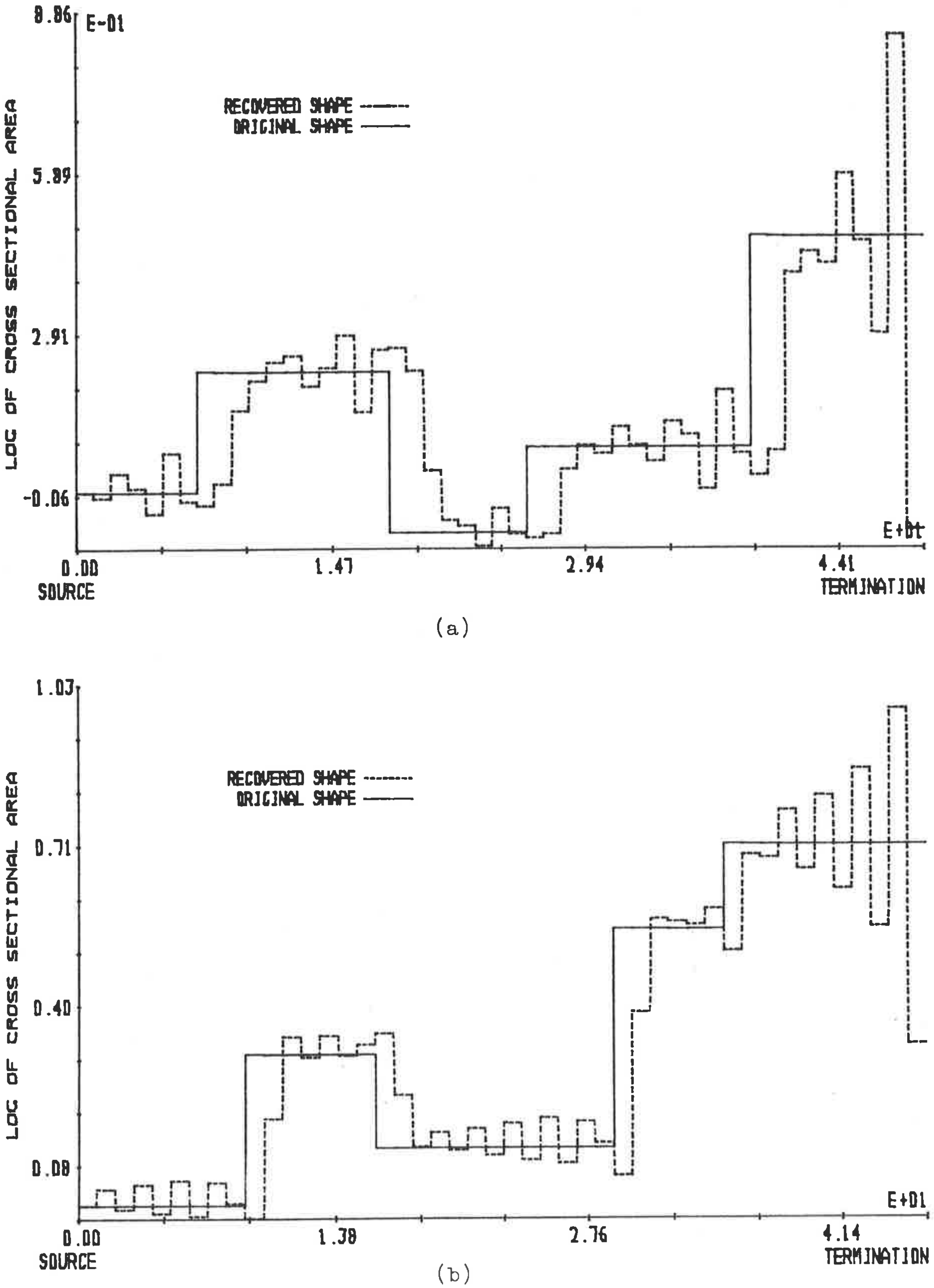


FIGURE 3.12: Typical autocorrelation function for non-white excitation of acoustic tubes with sparse changes in cross sectional shape.

now places a restriction on the duration of the positive non-white excitation for a particular acoustic tube shape. Equation 3.30 is referred to as the SPECIAL ACOUSTIC TUBE CRITERION. Figure 3.11 defines a typical special acoustic tube shape if the special acoustic tube criterion is satisfied. Figure 3.13 shows some examples of recovered acoustic tube shapes where the original acoustic tube shapes were special and excited by non-white waveforms. The constraint of Equation 3.30a only represents a lower limit on M_i and, in general, M_i is different for different i, which is important when the applications of special acoustic tubes are considered in Section 3.5.

The first decaying oscillation on the recovered acoustic tube shape occurs because of the non-zero autocorrelations $\bar{R}(j)$ for $0 \leq j \leq N-1$ (see Figure 3.12) and, while $\bar{R}(j)$ is zero for $j > N-1$, then the decaying oscillation continues on the recovered acoustic tube shape. However, the non-zero $\bar{R}(j)$ for $M_1-(N-1) \leq j \leq M_1-1$ disrupt the decaying oscillation so that the oscillation occurs for $M_1-(N-1)$ acoustic tubes. In general, the observed decaying oscillations on the recovered acoustic tube shape occur for $M_i-(N-1)$ acoustic tubes (M_i is the number of acoustic tubes with the same cross-sectional area between the $(i-1)$th and ith change in original acoustic tube shape) and start at the $(\sum_{j=1}^{i-1} M_j)$th acoustic tube from the termination.

From the discussions in Section 3.2, an acoustic tube which is M acoustic tubes from the termination cannot provide a contribution to $\bar{R}(j)$ until $j = M$. Therefore, if the first M acoustic tubes satisfy the special acoustic tube criterion, then the recovered acoustic tube shape for the first M acoustic tubes is not affected



(a)

(b)

FIGURE 3.13: Comparison of recovered and original special acoustic tube shapes excited by consecutive impulses with heights (a) 0.76, 1.03, 0.88 and 0.32, and (b) 0.34, 1.06, 0.79.

by any of the original acoustic tubes greater than M acoustic tubes from the termination. Hence, it is not necessary, in many cases, for the special acoustic tube shape criterion to be satisfied over the full original acoustic tube concatenation for oscillating recovered acoustic tube shapes to occur. In the example shown in Figure 3.14, recovered and original acoustic tube shapes are presented when only the original acoustic tubes near the termination satisfy the special acoustic tube criterion.

Investigations have shown that even if the special acoustic tube criterion is satisfied after M acoustic tubes from the termination decaying oscillations still occur in most situations where the special acoustic tube criterion is satisfied. Figure 3.15 presents the original and recovered acoustic tube shapes for the non-white excitation of a set of acoustic tubes which satisfy the special acoustic tube criterion except for five acoustic tubes at the termination. The portion of the original acoustic tube shape in Figure 3.15 which satisfies the special acoustic tube criterion is the same as the acoustic tube shape which satisfies the special acoustic tube criterion in Figure 3.14. The non-white excitations are also the same. A comparison of Figure 3.14 and 3.15 shows that the effect of acoustic tubes at the termination which do not satisfy the special acoustic tube criterion is to significantly reduce the decay rate of the oscillating recovered acoustic tube shape.

The above result is found to occur in most situations, and so, whenever the special acoustic tube criterion is satisfied, decaying oscillations are, in general, observed on the recovered acoustic tube shape. Figures 3.16 and 3.17 present recovered and original

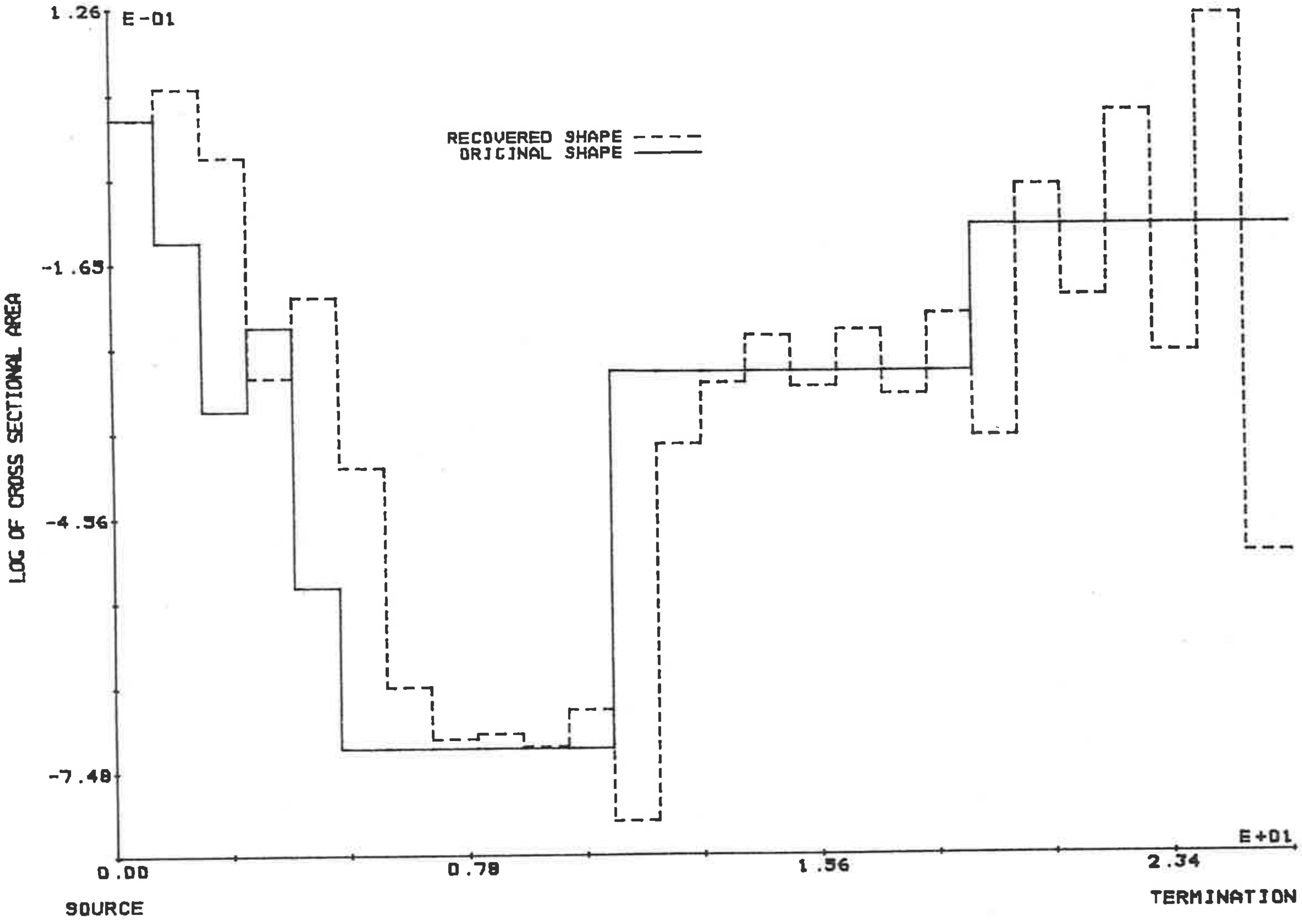


FIGURE 3.14: Comparison of recovered and original acoustic tube shapes when the original acoustic tube shape does not satisfy the special acoustic tube criterion over its full length. Excitation used has consecutive impulses of heights 0.24, 1.0 and 0.78.

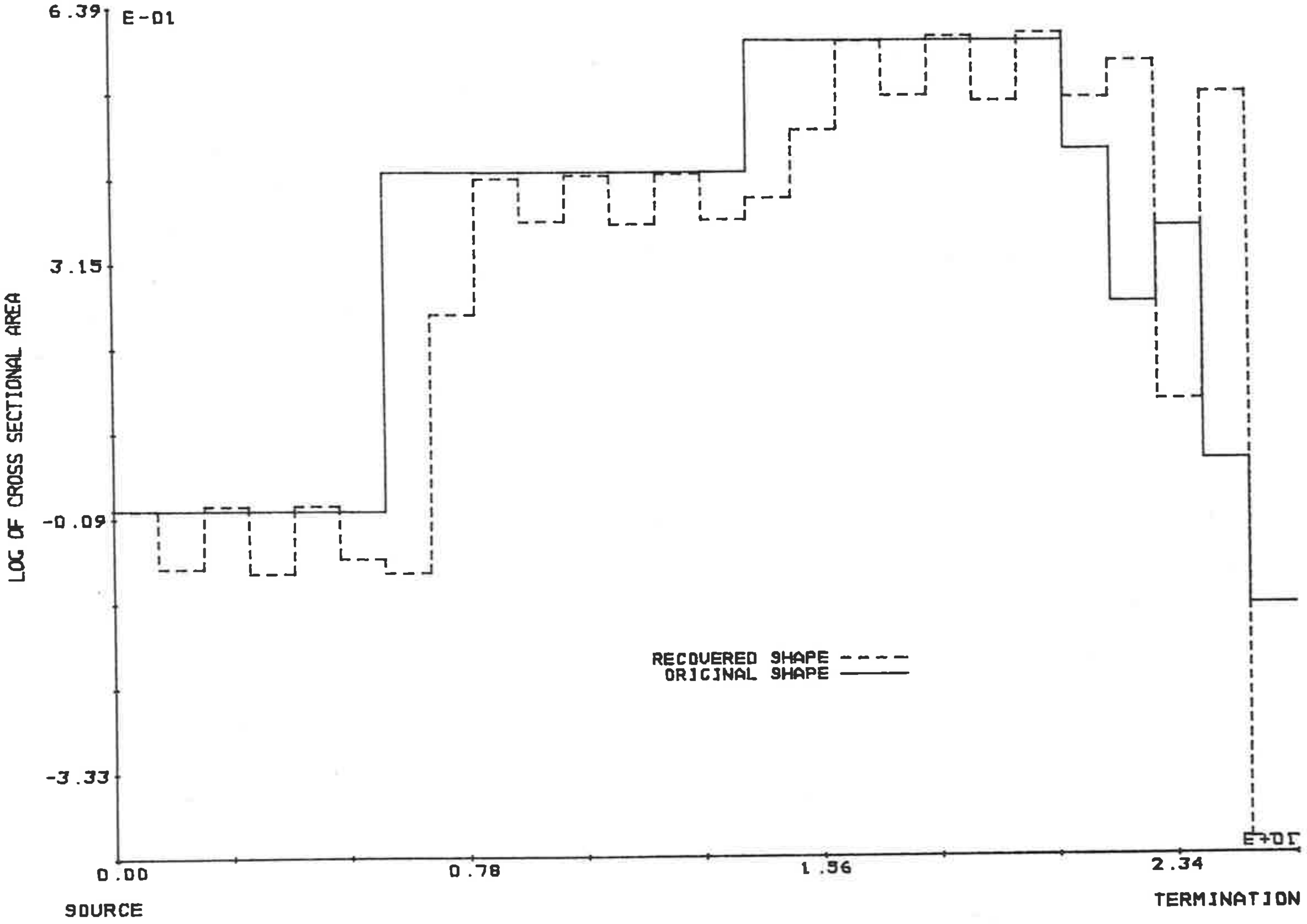


FIGURE 3.15: Comparison of recovered and original acoustic tube shapes when the original acoustic tube shape does not satisfy the special acoustic tube criterion over its full length. Excitation used has consecutive impulses of heights 0.24, 1.0 and 0.78.

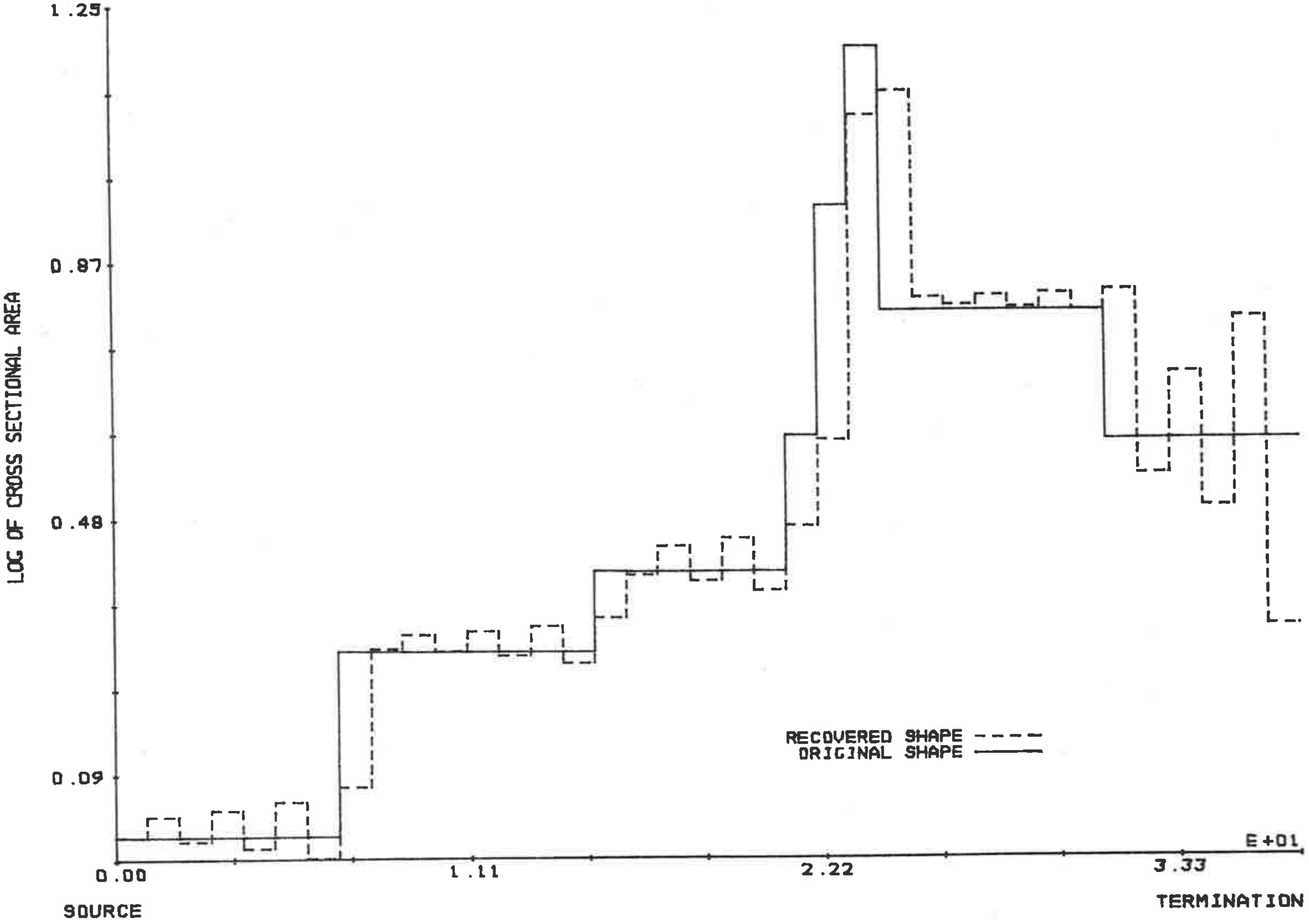


FIGURE 3.16: Comparison of recovered and original acoustic tube shapes when the original acoustic tube shape does not satisfy the special acoustic tube criterion over its full length. Excitation used has consecutive impulses of heights 1.0 and 0.83.

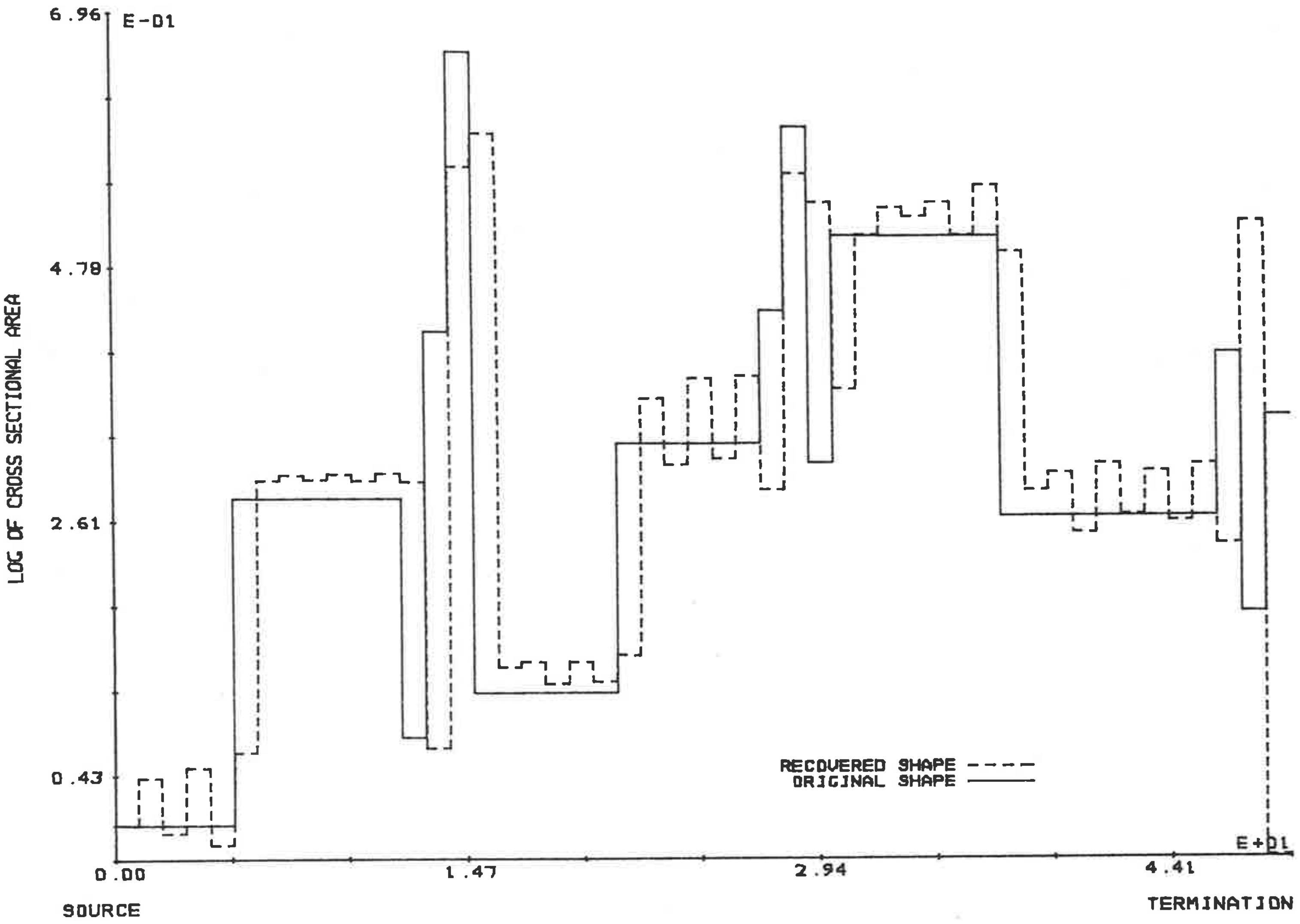


FIGURE 3.17: Comparison of recovered and original acoustic tube shapes when the original acoustic tube shape only satisfies the special acoustic tube criterion in small regions. Excitation used has consecutive impulses of heights 1.0 and 0.86.

acoustic tube shapes for situations where the special acoustic tube criterion is satisfied at various places along the original acoustic tubes.

3.4 RECOVERY OF SPECIAL ACOUSTIC TUBE SHAPES

In the previous section, special acoustic tubes were defined which produce a decaying oscillating recovered acoustic tube shape when excited by a positive non-white excitation and analysed by a conventional linear prediction. This section examines the recovery of the special acoustic tube shapes from the decaying oscillating recovered acoustic tube shape and develops the necessary additional processing and procedures. Throughout this section, the excitation of the original acoustic tubes is assumed to be by a positive non-white waveform.

If the original acoustic tube shape has all its changes in cross-sectional area separated by a number of acoustic tubes, then the recovered acoustic tube shape consists of a series of decaying oscillations, as shown by the example in Figure 3.4. In Figure 3.4, the recovered acoustic tube shape appears to have decaying oscillations about a shape which closely resembles the original acoustic tube shape. Therefore, removal of the decaying oscillations from the recovered acoustic tube shape may permit a good recovery of the original acoustic tube shape. Hence, a process is developed in this section which removes the decaying oscillations from the recovered acoustic tube shape when the original acoustic tube satisfies the special acoustic tube criterion.

A close examination of Figure 3.4 shows that what is implied by removing the decaying oscillations from the recovered acoustic tube shape is to determine the constant value to which each of the decaying oscillations decays. This constant acoustic tube shape is referred to as the decayed acoustic tube cross-sectional area, or simply the decayed area value. The decaying oscillation on the recovered acoustic tube shape has been shown to be caused by incorrect acoustic tube shape identifications by linear prediction due to the non-white excitation. It has also been shown that the decaying oscillation due to a single incorrect acoustic tube identification only decays to zero after an infinite number of acoustic tubes have been identified.

Only a finite number of oscillations can be observed in a recovered decaying oscillating acoustic tube shape and, therefore, an extrapolation of the available recovered oscillating acoustic tube shape to infinity may be necessary to determine the required decayed area value. Investigations of extrapolation techniques and their application to the observed decaying oscillating acoustic tube shapes discover two major difficulties. Firstly, the accuracy of any extrapolation process is highly dependent on the amount and accuracy of data available. If only a single oscillation is observed, as in the case where the special acoustic tube criterion (i.e. Equation 3.30) is just satisfied, the extrapolation results in poor estimation of the decayed area value.

Another difficulty with extrapolation techniques for determining decayed area values is the rate of decay of the recovered acoustic tube shape oscillations which is, in general, slow, e.g. see Figure 3.15. In these situations, the slow rate of decay causes large errors in the estimation of the decayed area values. Hence, in general, extrapolation methods are found to be unsuitable for determining decayed area values and, hence, permitting special acoustic tube shape recovery.

Slow decay of recovered acoustic tube shape oscillations causes difficulties with extrapolation techniques to determine the decayed area value, but is an advantage for other techniques. For instance, a simple averaging (or finding of the mean) of the envelope of the decaying oscillating recovered acoustic tube shape produces, in most cases, accurate decayed area values when the excitation is positive and non-white and the decay rate of the oscillation is slow. Investigations have shown that averaging of the envelope of the relatively fast decaying recovered acoustic tube oscillations at the termination also provides, in most cases, accurate decayed area values. Hence, an averaging of the envelope of the decaying oscillating recovered acoustic tube shape appears to be a simple and reliable process for most situations.

The first decaying oscillation from the termination of the recovered acoustic tube shape decays quickly, and so the decayed area value may be determined accurately if the oscillation is permitted to continue for a number of acoustic tube identifications. A typical autocorrelation function for a set of special acoustic tubes with a postivie non-white excitation is shown in Figure 3.12, and the discussions of Section 3.2.2 showed that the non-zero $\bar{R}(j)$ for $1 \leqslant j \leqslant N-1$ cause the recovered acoustic tube shape to exhibit a decaying oscillation. This decaying oscillation continues while $\bar{R}(j)$ is zero, i.e. until the (M_1-N+1)th acoustic tube identification when

$\bar{R}(M_1-N+1)$ is non-zero. Therefore, the decaying oscillation will continue if all the $\bar{R}(j)$ are set to zero for $j \geqslant N$.

The above discussion shows that the first decayed area value from the termination is obtained by setting the autocorrelation function of the signal $s(n)$ being analysed, i.e. $\bar{R}(j)$, to zero for $j \geqslant N$ and then performing a linear prediction until the recovered acoustic tube oscillations decay away sufficiently to determine the required decayed area value. If the linear predictive analysis is performed via the autocorrelation formulation, then $\bar{R}(j)$ is easily set to zero for $j \geqslant N$ and then the normal equations are solved via the Levinson [LEVINSON 1947, ROBINSON 1967], Durbin [DURBIN 1960], Leroux and Gueguen [LEROUX and GUEGUEN 1977] or the many other algorithms (see Chapter 2). Figure 3.18 shows recovered and original acoustic tube shapes where $\bar{R}(j)$ is set to zero for $j \geqslant 2$, where the excitation has a duration of $2T$. For the case presented in Figure 3.18, the recovered acoustic tube oscillations decay quickly so that the decayed area value is determined accurately after relatively few acoustic tube identifications, which is not necessarily the case in general.

Although the special acoustic tube criterion may be satisfied at least until the second cross-sectional area change from the termination, a procedure similar to the one described above, which continues the second decaying recovered acoustic tube shape from the termination, is very difficult to implement. The difficulty arises because a cross-sectional area change M_1 acoustic tube from the termination causes non-zero autocorrelations $\bar{R}(j)$ for j equal to integer multiples of M_1 instead of zero $\bar{R}(j)$ as seen above. These non-zero $\bar{R}(j)$ are difficult to predict and include in an analysis procedure.

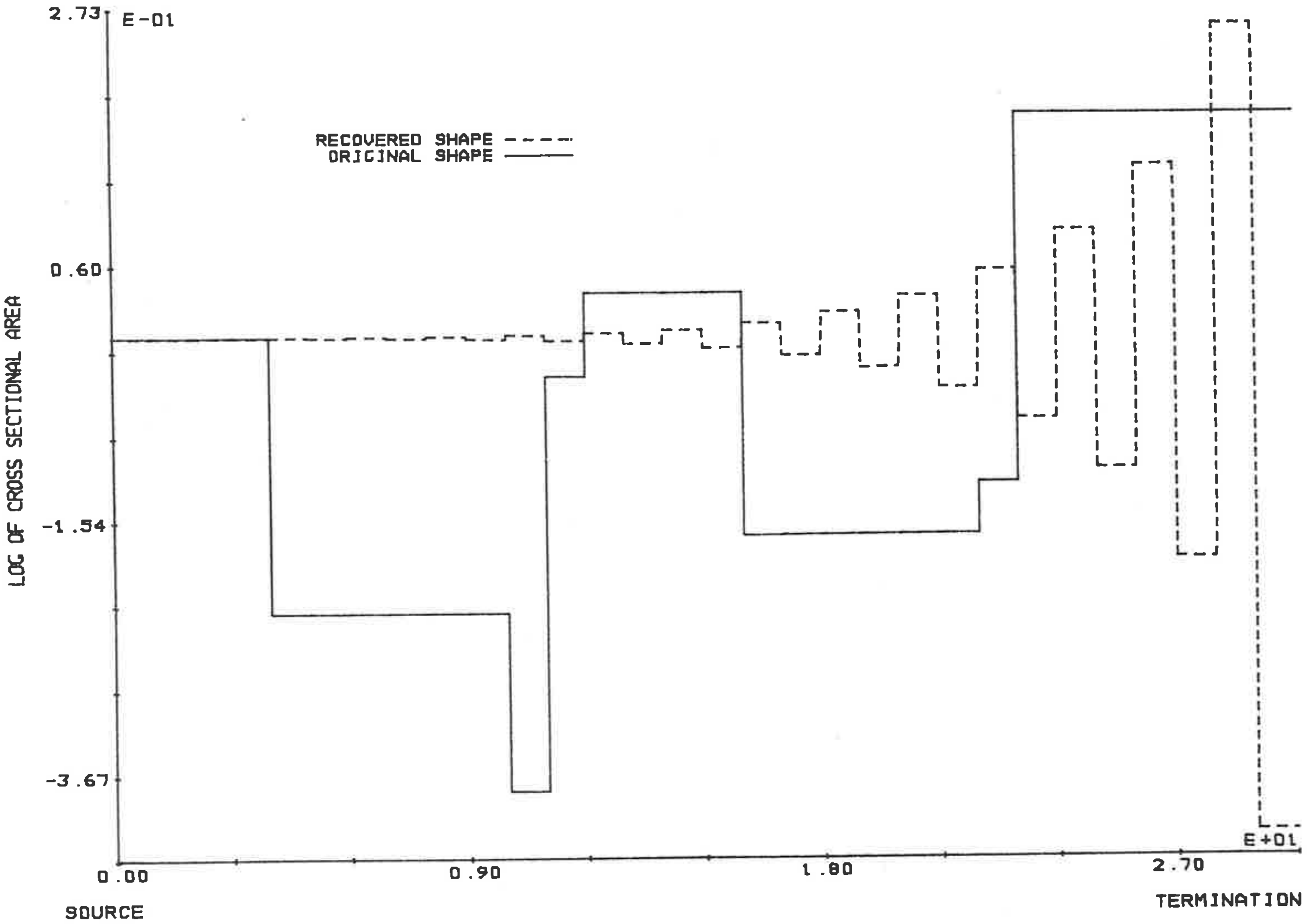


FIGURE 3.18: Comparison of recovered and original acoustic tube shapes when the autocorrelation function, R(i), set to zero for i > 2. Excitation used has consecutive impulses of heights 0.47, 1.0 and 0.47.

If the special acoustic tube criterion is satisfied at least until the second cross-sectional area change M_2 acoustic tubes from the termination, then the only non-zero values of $\bar{R}(j)$ for $j \leqslant M_2$ are at $j = 0$ and M_1. For the acoustic tubes up to M_2 acoustic tubes from the termination to be special, then the non-white excitation has a duration of less than $M_1/2$ and $M_2/2$. In this situation, the ratio of $\bar{R}(M_1)$ to $\bar{R}(0)$ is the same for an impulse excitation as for the above non-white excitation. This result is easily shown by consideration of the autocorrelation functions of the impulse response of the special acoustic tubes, of the non-white excitation and the output of the special acoustic tubes excited by the non-white excitation in a manner similar to that shown in Figure 3.6. Since the value of k_{M_1} is equal to $-\bar{R}(M_1)/\bar{R}(0)$, then k_{M_1} is recovered directly from the autocorrelation function, provided the acoustic tube shape up to M_2 acoustic tubes from the termination and the non-white excitation satisfy the special acoustic tube criterion.

Continuation of the above process to recover the special acoustic tube shape is possible provided the special acoustic tube criterion is satisfied over all the original acoustic tube shape. In general, the changes in cross-sectional area are separated by a different number of consecutive acoustic tubes and, hence, a special analysis process is required. The special analysis process must be able to identify acoustic tube cross-sectional area changes which have an unequal propagation delay between those area changes. The work on a d-step ahead predictor by GEVERS and WERTZ [1980] is a starting point for the development of such an analysis process.

Determining, from the recovered acoustic tube shape, the position at which a cross-sectional area change occurred on the original acoustic tube shape can be difficult. Investigations have shown that, in general, the position at which the original special acoustic tube shape changes corresponds to the position at which the decaying oscillation starts on the recovered acoustic tube shape. Unfortunately, in some situations, the exact position at which a decaying acoustic tube oscillation starts is difficult to determine.

Any analysis process which is developed to cope with unequal propagation delays between acoustic tube cross-sectional area changes is highly dependent on the knowledge of the position at which the area changes occur on the original acoustic tube shape. It has been found that identification of the position at which cross-sectional area changes occur on the original acoustic tube shape is very difficult to obtain from the recovered acoustic tube shape. Therefore, even if an analysis process that permits unequal propagation delays between changes in acoustic tube cross-sectional areas were developed, the uncertainty of the position at which the change in cross-sectional area occurs would not, in general, allow special acoustic tube shape recovery.

It was suggested earlier that a simple averaging of the envelope of the decaying oscillating recovered acoustic tube shape may provide accurate decayed area values. To determine the accuracy of the averaging process, an acoustic tube shape is considered which satisfies the special acoustic tube criterion over its full length. Figure 3.19 presents the special acoustic tube shape, the recovered acoustic tube shape by linear prediction, and

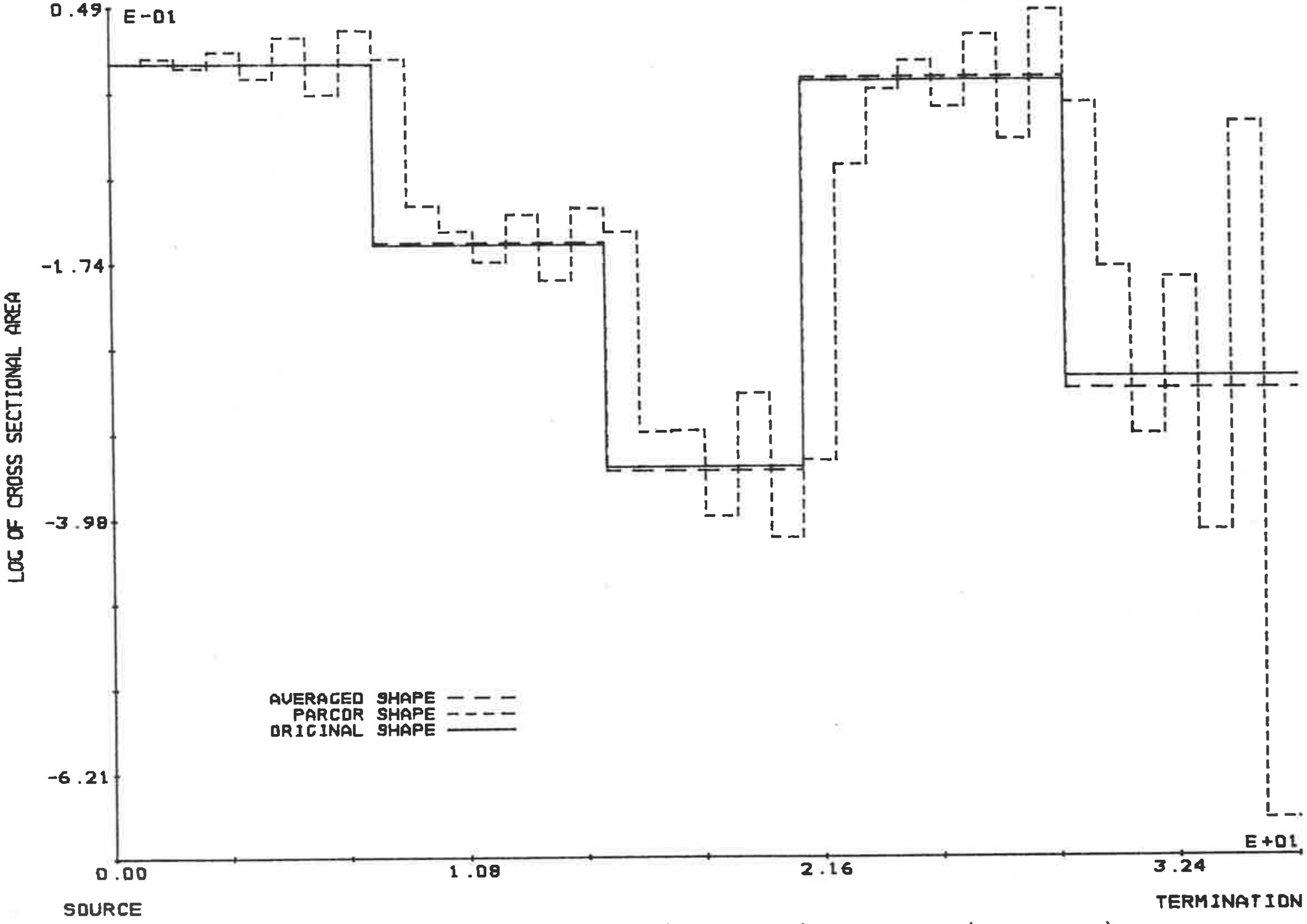


FIGURE 3.19: Comparison of recovered (by parcor), averaged (of parcor) and original acoustic tube shapes for an excitation having consecutive impulses of heights 0.4, 1.0 and 0.4.

the acoustic tube shaped recovered by using an averaging of the envelope of the decaying oscillations of the acoustic tube shape recovered by linear prediction. The excitation used has $U_0(0) = 0.4$, $U_0(1) = 1.0$ and $U_0(2) = 0.4$ with $U_0(n) = 0$ for all n other than $n = 0$, 1 or 2.

The special acoustic tube criterion is satisfied at the termination, and so the first cross-sectional area change that occurs on the original acoustic tube shape can be correctly identified by the process described above, i.e. the first original acoustic tube cross-sectional area change occurs seven sections from the termination, and so $k_7 = -\bar{R}(7)/\bar{R}(0)$. Using the averaging of the envelope of the decaying recovered acoustic tube shape, k_7 is identified as -.305, instead of the correct value, -.29, which only produces a small error in acoustic tube shape recovery. In general, Figure 3.16 shows that good special acoustic tube shape recovery is achieved by averaging the envelope of the decaying oscillating acoustic tube shape recovered by a linear predictive analysis.

A more realistic acoustic tube shape is considered where only a portion of the original acoustic tube shape satisfies the special acoustic tube criterion. Figure 3.20 presents the acoustic tube shape, and the excitation used in this case is the same as that used for Figure 3.19. The recovered acoustic tube shapes by using a linear prediction and an averaging of the envelope of the decaying oscillations of the acoustic tube shape recovered by linear prediction are presented in Figure 3.20.

A comparison of the acoustic tube shapes presented in Figure 3.20 shows a that the averaging process recovers the original

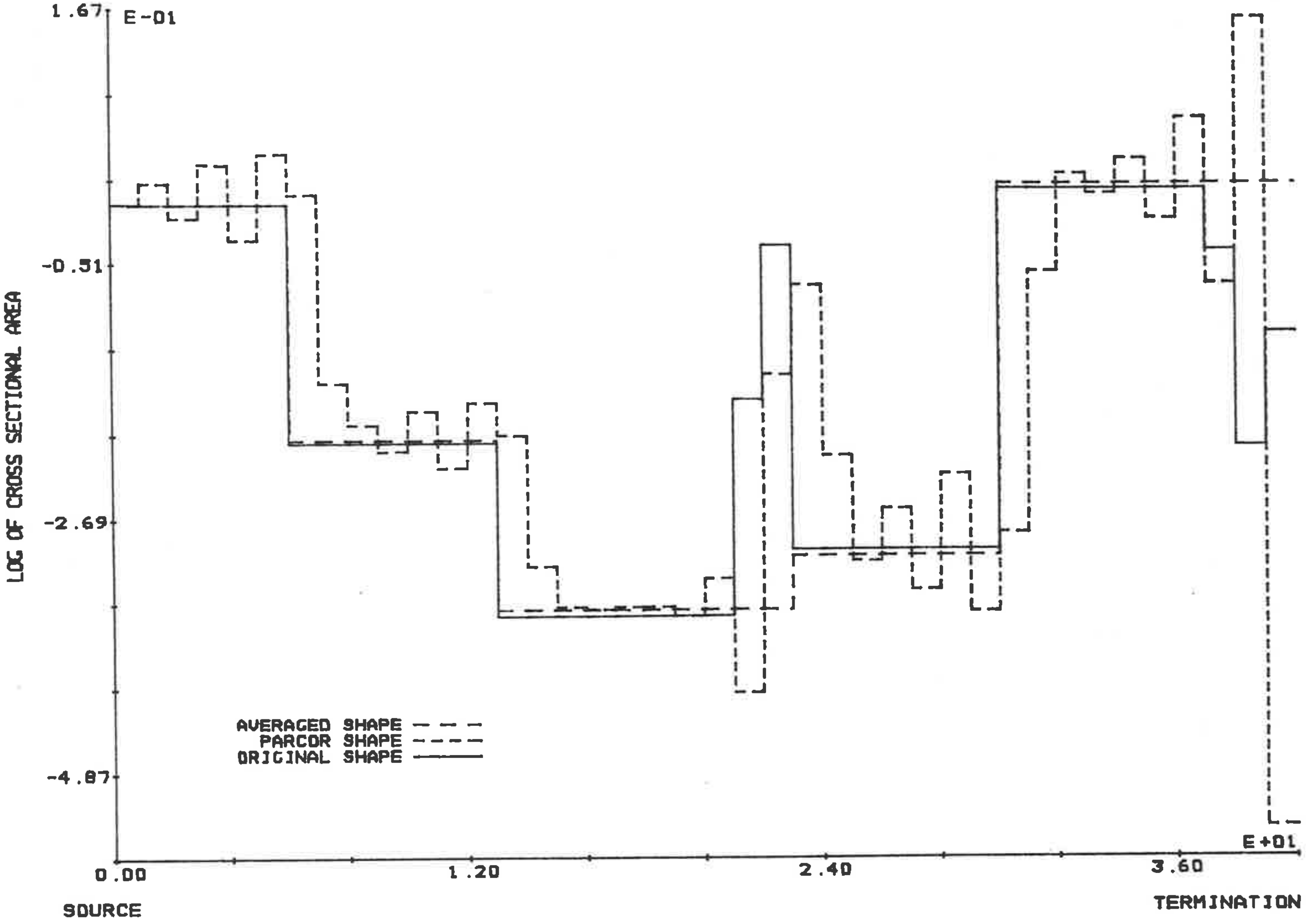


FIGURE 3.20: Comparison of recovered (by parcor), averaged (of parcor) and original acoustic tube shapes for an excitation having consecutive impulses of heights 0.4, 1.0 and 0.4.

acoustic tube shape with good accuracy whenever the special acoustic tube criterion is satisfied. The largest differences between original and recovered (by the averaging process) acoustic tube shapes occurs near the termination where the oscillating acoustic tube shape recovered by a linear predictive analysis decays the quickest. This observation supports the discussions presented earlier, where it was concluded that the recovered acoustic tube oscillations decay slower the further they are from the termination, and that the averaging process estimates the decayed area value more accurately for slow decays of the oscillating acoustic tube shape.

The examples considered show that an averaging of the envelope of the decaying recovered acoustic tube oscillations defines decayed area values which provide a good recovery of the original special acoustic tube shape. The accuracy of the recovered special acoustic tube shape may be increased by using a more complex process than the simple averaging considered above. However, in most situations, the extra computations required by a more complex process would not justify the small increase in the accuracy of the recovered special acoustic tube shapes.

The major difficulty with determining the decayed area values by the averaging process described above, or for that matter any other process, is the high dependence of the results on the available decayed recovered acoustic tube oscillations. In situations where the special acoustic tube criterion is just satisfied, then only a single oscillation of the recovered acoustic tube shape is observed and, hence, a poor recovery of the original special acoustic shape occurs.

Another analysis procedure for recovering special acoustic tube shapes is developed by examination of the autocorrelation function $\bar{R}(j)$ of an original acoustic tube shape which satisfies the special acoustic tube criterion over its full length, e.g. the autocorrelation function presented in Figure 3.12. The autocorrelation function of Figure 3.12 is the convolution of the autocorrelation function of the non-white excitation, which has a duration of NT, with the autocorrelation function of the impulse response of the special acoustic tube shape. The requirements of the special acoustic tube criterion ensure that the autocorrelation function $\bar{R}(j)$ for $0 \leqslant j \leqslant N-1$ is the autocorrelation function of the non-white excitation. Therefore, deconvolving $\bar{R}(j)$ for $0 \leqslant j \leqslant N-1$ from the autocorrelation function $\bar{R}(j)$, for all j, results in the autocorrelation function of the impulse response of the original special acoustic tubes, which can be analysed by a conventional linear prediction to recover the original special acoustic tube shape.

The above discussion only requires $\bar{R}(j)$ for $0 \leqslant j \leqslant N-1$ to be the autocorrelation function of the non-white excitation for the deconvolution process to apply, and this is true provided the special acoustic tube criterion is satisfied at the termination of the original acoustic tubes. Hence, provided the first cross-sectional area change occurs at least $2N$ acoustic tubes from the termination for a non-white excitation of duration NT, then the above deconvolution process followed by a linear predictive analysis recovers the original acoustic tube shape.

In summary, two procedures have been defined for recovering part or all (depending on where the special acoustic tube criterion is satisfied) of the original acoustic tube shape when excited by non-white waveforms of certain duration. If the first change in cross-sectional area from the termination of the original acoustic tube shape satisfies the special acoustic tube criterion, then a deconvolution of autocorrelation functions permits the recovery of the original acoustic tube shape. In situations where the special acoustic tube criterion is not satisfied at the termination, then a process was defined which provides good acoustic tube shape recovery when the special acoustic tube criterion is satisfied. The recovery process was a simple averaging of the observed decaying oscillations of the acoustic tube shape recovered by linear prediction, and the accuracy of this process was shown to be dependent on the available recovered acoustic tube oscillation, its rate of decay and relative position from the termination.

3.5 AREAS OF APPLICATION

Application of the analysis procedures developed in the previous section are limited because the special acoustic tube criterion is not satisfied in most practical situations. The practical situations where the analysis procedures can be applied are grouped into two distinct areas. The first application area is the non-white excitation of acoustic tube shapes where the acoustic tubes have lengths which are integer multiples of some common factor. Since, in practical situations, the excitation is most likely to be non-white and changes in cross-sectional area most likely to occur at random, then the above situation is more realistic than the assumptions of the linear prediction/acoustic tube model permit.

The second application area is to situations where both the forward and backward travelling waveforms at the termination of the acoustic tubes are available or can be made available. In such a situation, the original acoustic tube shape can be, in theory, recovered regardless of the type of finite duration excitation used. The occurrence of the necessary conditions for this application area are rare, and so only a brief description is given in this section.

For acoustic tubes of different lengths, the propagation delay of the waveforms travelling within each acoustic tube is different and, therefore, violates a fundamental assumption of linear prediction. If a linear prediction is performed in such a situation, then a cross-sectional area change is identified at positions where none occurred on the original acoustic tube shape. Hence, the first application area of the analysis process developed for special acoustic tube shape recovery is one where conventional linear prediction fails to recover the original acoustic tube shape.

A simple example of the situation described above is an acoustic tube shape which has only one cross-sectional area change a distance q from the termination. The impulse response of such an acoustic tube shape has an autocorrelation function defined by

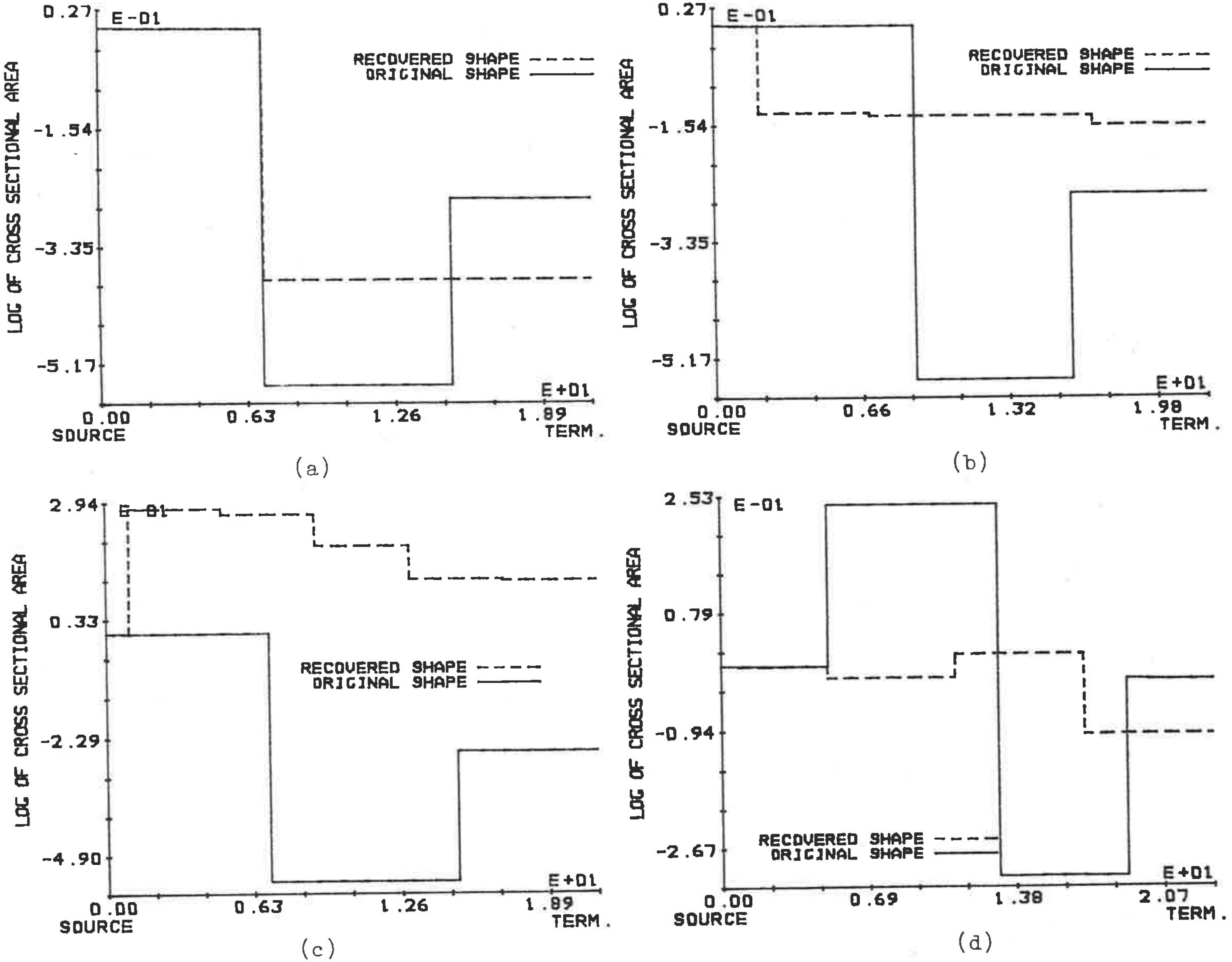
$$R\left(\frac{jq}{c}\right) = \begin{cases} R(0)(-\mu)^{j} & j \text{ integer} \\ 0 & j \text{ not integer} \end{cases} \qquad (3.31)$$

where μ is the reflection coefficient for the change in cross-sectional area and c is the velocity of waveform propagation in the acoustic tubes.

If a sampling period of r/c is used, then linear prediction only recovers a change in cross-sectional area if r and q have a common factor. When r is a multiple of q, i.e. $r = iq$ with $i > 0$ and integer, then linear prediction recovers a reflection coefficient equal to $(-\mu)^i$. Hence, for i not equal to one, i.e. a correct recovery, a significant error in acoustic tube shape recovery is made by a linear predictive analysis. Not only is the size of the cross-sectional area incorrect, but the position of the change is identified as a distance iq rather than q from the termination.

The above example shows that large errors may occur for acoustic tube shape recovery by linear prediction when the sampling period is not equal to the propagation delay of waveforms within the original acoustic tubes. Figure 3.21 shows some original and recovered acoustic tube shapes when linear prediction analyses the impulse response of the original acoustic tubes when the sampling period for the linear predictive analysis is, in general, not equal to the propagation delays within the original acoustic tubes. The recovered acoustic tube shapes shown in Figure 3.21 have little resemblance to the original acoustic tube shape and, hence, it is concluded that the acoustic tube shape recovered by linear prediction is highly dependent on the sampling period if the original acoustic tube shape contains non-commensurate discrete acoustic tubes.

For situations where the original acoustic tubes have different lengths but with a common factor, then a linear prediction with a sampling period equal to that common factor permits the recovered acoustic tube shape to have changes in cross-sectional areas



(a)

(b)

(c)

(d)

FIGURE 3.21: Comparison of recovered and original acoustic tube shapes when the sampling period is not equal to the common factor of the propagation delays in the original acoustic tube shape. An impulse is used for the excitation in all cases.

at the same position as the changes in cross-sectional area which occurred on the original acoustic tube shape. In general, the common factor implies a very small sampling period and, since real excitation waveforms have a finite duration, then the necessity for a small sampling period may cause the excitation to appear non-white. Therefore, in many practical situations, a contradiction of requirements can arise when the sampling period needs to be large so that finite duration excitations appear white, but a small sampling period is necessary to ensure a correct identification of cross-sectional area changes and their relative positions.

In situations where the above contradiction occurs, the procedures described in the previous section can be used to recover the original acoustic tube shape whenever that shape and the excitation satisfy the special acoustic tube criterion. The recovery precedure would choose the sampling period small enough for linear prediction to identify a change in cross-sectional area at every position where a change of acoustic tube cross-sectional area occurred on the original acoustic tubes. Once the sampling period has been chosen, then a recovered acoustic tube shape is determined by first performing a linear predictive analysis and then using the special acoustic tube shape recovery procedures detailed in the previous section.

An example of the above procedure, and the result of an incorrect choice of sampling frequency when only linear prediction is used, is shown in Figure 3.22. The excitation of the original acoustic tube shape is positive with a duration of $3(r/c)$ where r is the common factor for all the original acoustic tube lengths. For sampling periods of greater than $3(r/c)$, i.e. as used in

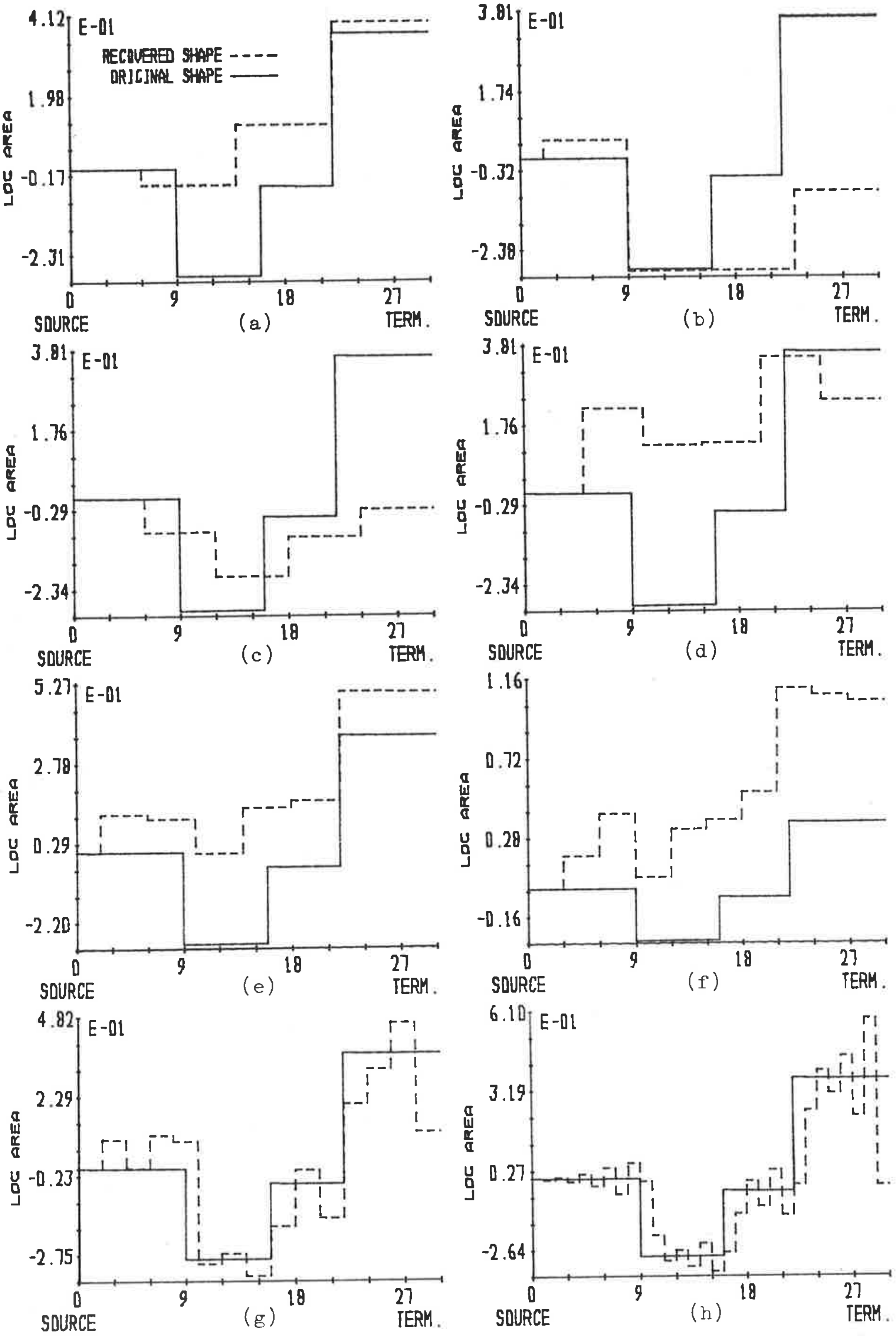
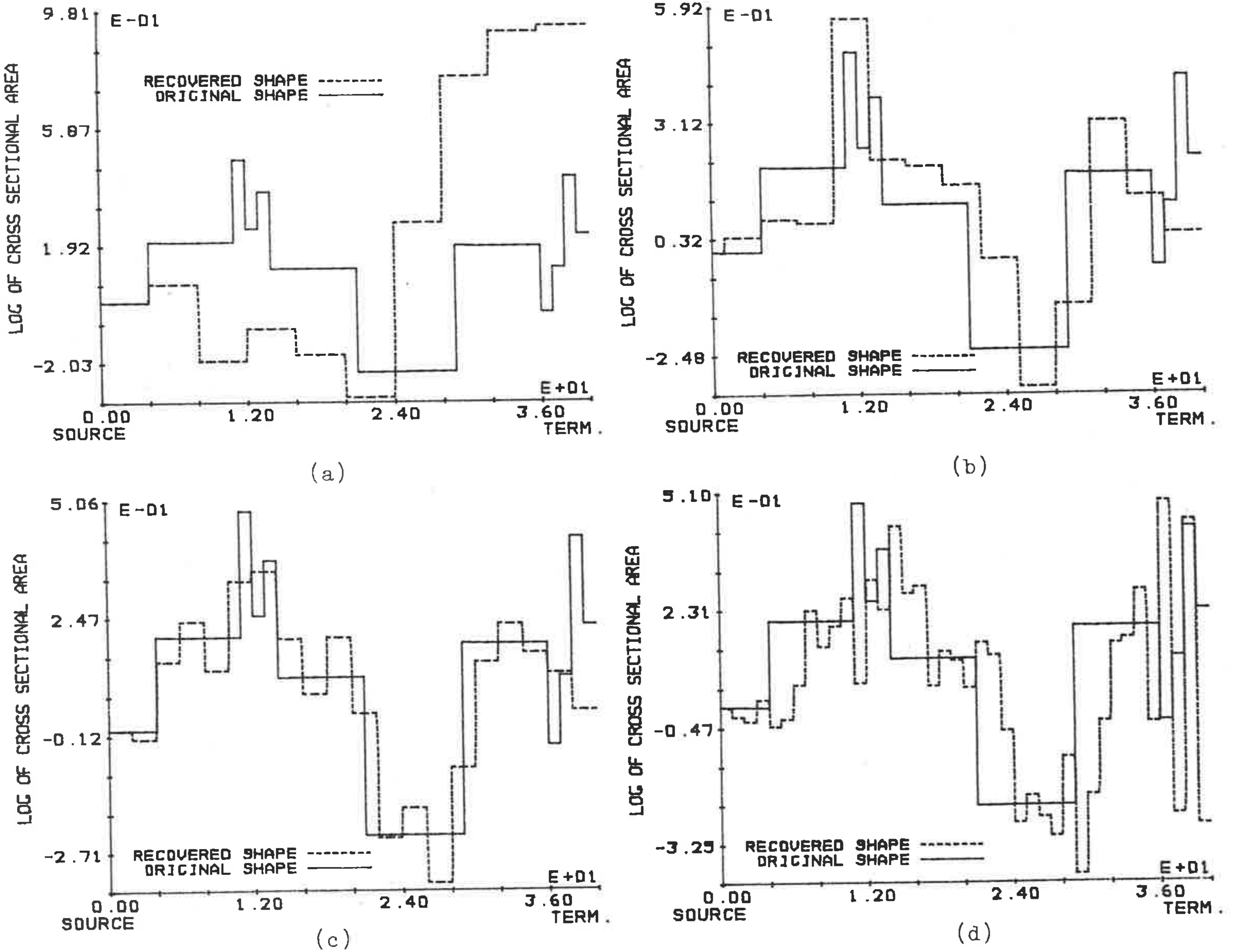


FIGURE 3.22: Comparison of recovered and original acoustic tube shapes for a sampling period of (a) 8T, (b) 7T, (c) 6T, (d) 5T, (e) 4T, (f) 3T, (g) 2T, (h) T. Impulse excitation and T is common factor of propagation delays in original acoustic tube shape.

Figure 3.22(a) to (e), the excitation appears white, but large errors in the recovered acoustic tube shape are observed. Using a sampling period of (r/c), the recovered acoustic tube shape is shown in Figure 3.22(h) and, by averaging the envelope of the decaying recovered acoustic tube oscillations, the original reflection coefficients, describing the original acoustic tube shape, are recovered with an error of less than 2%. For this example, the positions of the original acoustic tube cross-sectional area changes are identified correctly.

The example presented in Figure 3.22 had the special acoustic tube criterion satisfied over the full length of the original acoustic tube shape, but the example presented in Figure 3.23 only has the original acoustic tube shape satisfying the special acoustic tube criterion over portions of its length. The excitation of the original acoustic tube shape presented in Figure 3.23 is positive and has a duration of $4(r/c)$, where r is the common factor of the original acoustic tube lengths. Large errors are observed in the acoustic tube shapes recovered by linear prediction for all the sampling frequencies considered in Figure 3.23. Averaging the envelope of the decaying recovered acoustic tube shape to determine decayed area values for the sampling period of (r/c) (i.e. Figure 2.32(d)) recovers the reflection coefficients describing the original acoustic tube shape, where the special acoustic tube criterion is satisfied, with an error of less than 6%. The relative positions of the original acoustic tube cross-sectional area changes, where the special acoustic tube criterion is satisfied, are also recovered correctly.



(a)

(b)

(c)

(d)

FIGURE 3.23: Comparison of recovered and original acoustic tube shapes for a sampling period of (a) 4T, (b) 3T, (c) 2T, (d) **T**. Excitation has consecutive impulses of heights 0.5, 1.0, 0.8 and 0.3 and T is common factor of propagation delays in original acoustic tube shape.

An application has been shown for the special acoustic tube shape recovery procedures developed in the previous section, i.e. to situations where the requirements of sampling period to make the excitation appear white conflict with the requirement of sampling period to permit an acoustic tube shape change to be identified at positions where acoustic tube shape changes occur on the original acoustic tube shape. Two examples which have the above conflict were considered, and it was shown that whenever the special acoustic tube criterion is satisfied on the original acoustic tube shape then the recovery procedures developed in the previous section can recover, with reasonable accuracy, the original special acoustic tube shape.

A practical situation which may have the conflict of sampling period requirements described above is the identification of multi-layered structures where each layer has a different "impedance," i.e. layers of the earth's crust. In such situations, where short duration excitations are difficult to produce, large sampling periods are necessary for the excitation to appear white. However, in the general practical situation, the propagation times between layers are unequal, and so a small sampling period is necessary to identify the layer boundaries. Thus, a conflict of sampling period requirements and a practical application for the analysis processes defined in the previous section occurs.

The second application area for the special acoustic tube shape recovery procedures developed in the previous section is to situations where both the forward and backward travelling waveforms at the termination of the acoustic tubes are available. In such situations, the special acoustic tube criterion can be forced to to be satisfied by the addition of an extension tube (physically or via simulations) which has a constant cross-sectional area over its full length. For the extension acoustic tube to satisfy the special acoustic tube criterion, its length must be at least $2N/c$ where the duration of the excitation is NT (see Equation 3.30 and BOGNER and DAVIS [1983]).

In the situations where an extension tube is used, the autocorrelation function of the output waveform from the extension tube, $R(j)$, for j from zero to $(N-1)$ is the autocorrelation function of the excitation (see Section 3.4). Hence, a deconvolution of the autocorrelation function of the excitation, $R(j)$, for j from zero to $(N-1)$ from the autocorrelation function $R(j)$ for all j produces the autocorrelation function of the impulse response of the original acoustic tubes plus the extension tubes. A linear prediction following the deconvolution then recovers the original acoustic tube shape.

The deconvolution process described above requires the extension tube to be lossless if the autocorrelation function of the impulse response of the original acoustic tubes plus the extension tube is to be recovered. If the excitation of the original acoustic tubes has a long duration, then a long extension tube is required. In practice, a long acoustic tube has a significant loss over its length, and so significant errors may occur in the recovery of the original acoustic tube shape due to the losses in the extension tube. However, in most practical situations, the properties of an extension acoustic tube can be measured and, hence, losses and other non-ideal properties corrected for.

The small decay constant of a long extension acoustic tube causes problems with accurate calculation of the autocorrelation function, $R(j)$, of the output waveform, i.e. the addition of an extension tube increases the effective length of the output waveform. This implies that the original acoustic tubes must not be re-excited until sufficient output waveform from the extension acoustic tube has been observed to accurately calculate the required autocorrelation function. This is not the case for voiced speech and, therefore, the above procedure cannot be used to recover vocal tract shapes.

Accurate acoustic tube shape recovery has been shown, in Section 3.4, to be possible if the special acoustic tube criterion is satisfied at the termination of the original acoustic tubes. If both the forward and backward travelling waveforms at the termination of the original acoustic tubes are known, then an extension acoustic tube can be added to ensure that the special acoustic tube criterion is satisfied at the termination of the extension acoustic tube. The accuracy of the recovered acoustic tube shape with the addition of an extension tube and a deconvolution depends on the properties of the extension acoustic tube, which ideally is lossless, and the duration of the output acoustic waveform before the original acoustic tubes are re-excited.

3.6 CONCLUDING COMMENTS

The research presented in this chapter investigated the effects of non-white excitation on the acoustic tube shape recovered by linear prediction. As a result of these investigations a description, in terms of the autocorrelation function, was obtained for the linear prediction process. Using the knowledge of the manner in which a linear prediction is performed permits an understanding of some of the effects of non-white excitation on the acoustic tube shape recovered by a linear prediction.

The understanding of the effects of non-white excitation on a recovered acoustic tube shape permitted the relaxation of the white excitation requirement of linear prediction provided a restriction is placed on the original acoustic tube shape. This restriction is defined by the special acoustic tube criterion (Equation 3.30), and the acoustic tubes satisfying this criterion are called special acoustic tubes. Two analysis processes were presented which recover special acoustic tube shapes, with the choice of which procedure should be used in a particular situation depending on the extent and where the special acoustic tube criterion is satisfied.

In general, the special acoustic tube criterion is very restrictive and, therefore, special aocustic tube shapes are the exceptions rather than the rule in practical situations. Therefore, the practical applications of the special aocustic tube shape recovery procedures developed in this chapter are limited. Unfortunately, the special acoustic tubes do not occur in the acoustic tube model of the vocal tract and, hence, the research presented in this chapter is not directly applicable to speech analysis. However, the understanding of the linear prediction process, a major speech analysis tool, from an atuocorrelation function point of view is a significant contribution.

One application area for the special acoustic tube shape recovery procedures is to the problem of incorrect sampling of the output waveform of non-commensurate acoustic tubes or, in practice, general multilayered media where distinct boundaries occur between each layer. It was shown that large errors in acoustic tube shape (or multilayer "impedances") occur when the sampling period is not a common factor of the propagation times within the non-commensurate acoustic tubes (or multilayered media). The analysis procedures developed in this chapter permit the recovery of the non-commensurate acoustic tubes (or multilayered media) when the special acoustic tube criterion is satisfied.

The major contribution of this chapter to speech analysis is the understanding of the manner in which linear predictive analysis uses an autocorrelation function to identify commensurate acoustic tubes. Another contribution is the understanding of the manner in which an acoustic tube shape recovered by linear prediction is affected by certain non-white excitations.

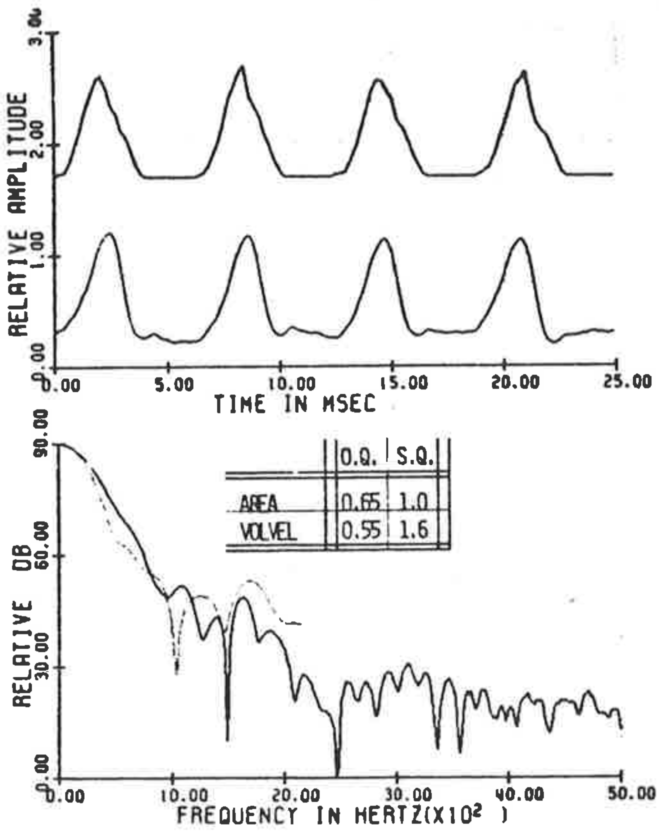
CHAPTER 4

GLOTTAL PULSE EXCITATION OF ACOUSTIC TUBES

4.1 INTRODUCTION

The investigations performed in Chapter 3 were for general positive non-white excitations of the acoustic tube model, i.e. the excitation has an arbitrary positive shape and no information is known about the excitation at the time of analysis. In contrast, there does exist some knowledge of the excitation waveform when analysing voiced speech sounds. This chapter investigates the possibility of using this a priori knowledge of the voiced sounds' excitation to reduce the effects of the non-white excitation on the recovered acoustic tube shape.

For voiced sounds, the glottal excitation waveform has a pulse shape and frequency spectrum similar to those shown in Figure 4.1. Since the frequency spectrum of the glottal pulse is non-white, conventional linear predictive analysis procedures cannot recover the vocal tract shape accurately. However, most glottal pulse shapes do not differ greatly from that shown in Figure 4.1, and so there exists some a priori knowledge of the excitation at the time of analysis. Some presently available speech analysis techniques use this a priori knowledge of the glottal pulse excitation to improve the accuracy of the recovered vocal tract shape.



	O.Q.	S.Q.
AREA	0.65	1.0
VOLVEL	0.55	1.6

FIGURE 4.1: Glottal area (top) and glottal volume velocity waveform (middle) and their associated spectra (bottom), (after Berouti, Childers and Paige 1977).

The first method used by speech researchers to remove the effects of glottal pulse excitation from the recovered acoustic tube shape was a constant dB per octave pre-emphasis of the speech waveform. This type of pre-emphasis has been discussed in Section 1.4, with a +12 dB per octave pre-emphasis being added to the speech spectrum in an attempt to remove glottal pulse excitation effects from the recovered acoustic tube shape. Recently, a number of researchers have disputed the effectiveness of this pre-emphasis and SONDHI [1979] has shown that the recovered area function is highly dependent on the amount of pre-emphasis applied. This is contradictory to the comments of MARKEL and GRAY ([1976], p. 81) but is supported by the experimental work of ENGEBRETSON and VEMULA [1974].

A study of glottal pulse excitation spectra by MONSEN and ENGEBRETSON [1977] showed that a conceptualization of the slope of the glottal pulse spectrum in terms of a single dB per octave is unrealistic. Their experimental work showed a variation between -12 dB and -18 dB per octave in the glottal pulse spectral slope. This variation is supported by the work of CARR and TRILL [1964] and MILLER [1959] who found a variation of between -8 dB and -17 dB per octave. MONSEN and ENGEBRETSON concluded that, in all cases observed, the spectral slope differs in steepness from octave to octave, and the spectrum changes over time in different ways, depending upon the linguistic content. A number of glottal pulses and their spectra are shown in Figure 4.2 to illustrate these points.

In an attempt to overcome the variability of the glottal pusle spectral slope, GRAY and MARKEL [1974] and MAKHOUL and VISWANATHAN [1974] developed an adaptive pre-emphasis filter.

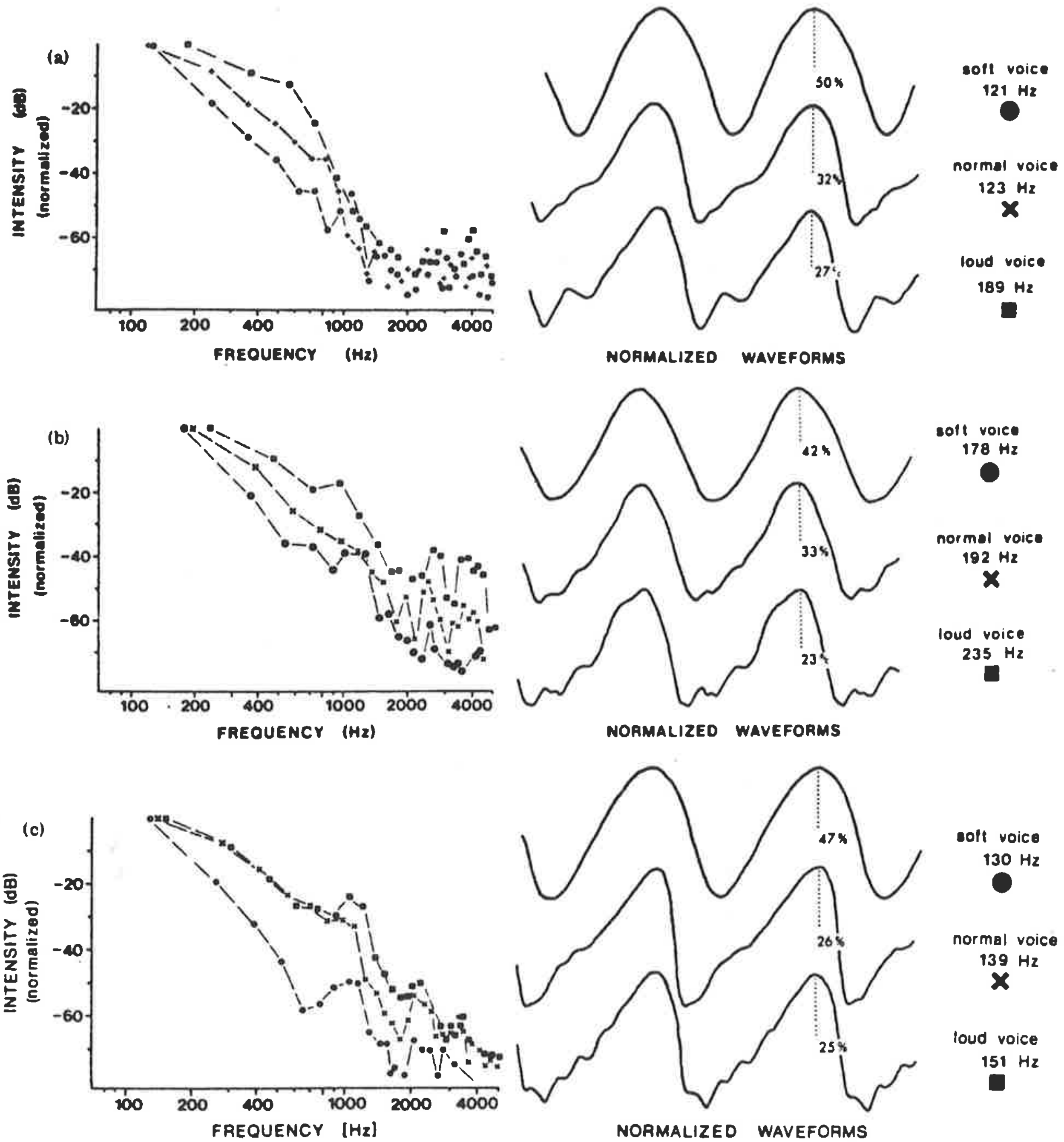


FIGURE 4.2: Waveforms and corresponding spectra for normal, loud and soft voice for (a) male subject, (b) female subject and (c) male subject (after Monsen and Engebretson 1977).

The adaptive pre-emphasis filter was not only intended to account for glottal pulse spectral slope, but other non-idea vocal tract properties such as lip radiation and losses, also. Hence, the effectiveness of their adaptive pre-emphasis filter to account for glottal pulse spectral slope variation cannot be determined from their work. A study of the usefulness of the adaptive pre-emphasis to account for glottal pulse spectral slopes is presented later in this chapter.

Another adaptive pre-emphasis filter has been developed by NAKAJIMA et al [1974] to account for glottal excitation and radiation effects. The pre-emphasis filter is based on a second and third order critical damping filter (a constrained auto-regressive model). The exact values of the filter coefficients are determined at the time of analysis by solving two polynomials of order three and five. This causes problems with real time applications which are discussed later in this chapter. A complete discussion of the form and procedures required to use this adaptive pre-emphasis filter is presented in Section 4.3.

Another technique which has been used to overcome the problems of glottal pulse excitation is to analyse the voiced sound when the glottis is closed. During the closed glottis interval, excitation of the vocal tract may cease, at which time the vocal tract spectrum is not being modified by the glottal excitation spectrum. Hence, accurate vocal tract shape recovery is possible, in principle, by analysis in the closed glottis interval. The concept of analysis in the closed glottis interval was first presented by MATHEWS, MILLER and DAVID [1961] and, since then, has been a popular analysis technique.

There are a number of disadvantages with closed glottis interval analysis, with the obvious one being the difficulty of making an accurate determination of that interval. Difficulties also arise for high pitch sounds, where the closed glottis interval is small, and so accurate covariances or autocorrelations cannot be found. For weakly voiced sounds, the vocal folds do not close completely, due to a lack of pressure reduction in the glottis to cause vocal fold closure. ROSENBERG [1971] concluded that vocal fold closure does not occur as a general rule, and so significant numbers of voiced sounds cannot be analysed by closed glottis interval analysis.

The above discussion of the techniques presently used to overcome the presence of glottal pulse excitation in the speech waveform suggests that significant errors are introduced by the glottal pulse excitation. Section 4.2 investigates the effects of glottal pulse excitation on acoustic tube shapes recovered by conventional linear predictive procedures, and establishes the importance of correcting for glottal pulse excitation effects on acoustic tube shape recovery.

Once the effects of glottal pulse excitation have been determined, the effectiveness of the available pre-emphasis techniques is determined in Section 4.3. The potential of the closed glottis interval analysis is also presented in Section 4.3, with an indication of the errors produced when the interval is incorrectly identified. Therefore, Section 4.3 provides a comparison of available techniques for removing glottal pulse effects from the speech waveform and determines if improvements are necessary for accurate acoustic tube shape recovery.

A new adaptive pre-emphasis filter is developed in Section 4.4, and its design includes the strengths of the available pre-emphasis filters, as discovered in Section 4.3. Thus, the new pre-emphasis filter retains the effective features of available pre-emphasis filters and attempts to overcome their deficiencies. In the design of the new adaptive pre-emphasis filter, only glottal pulse excitation effects are considered, unlike others which also attempt to remove lip radiation effects.

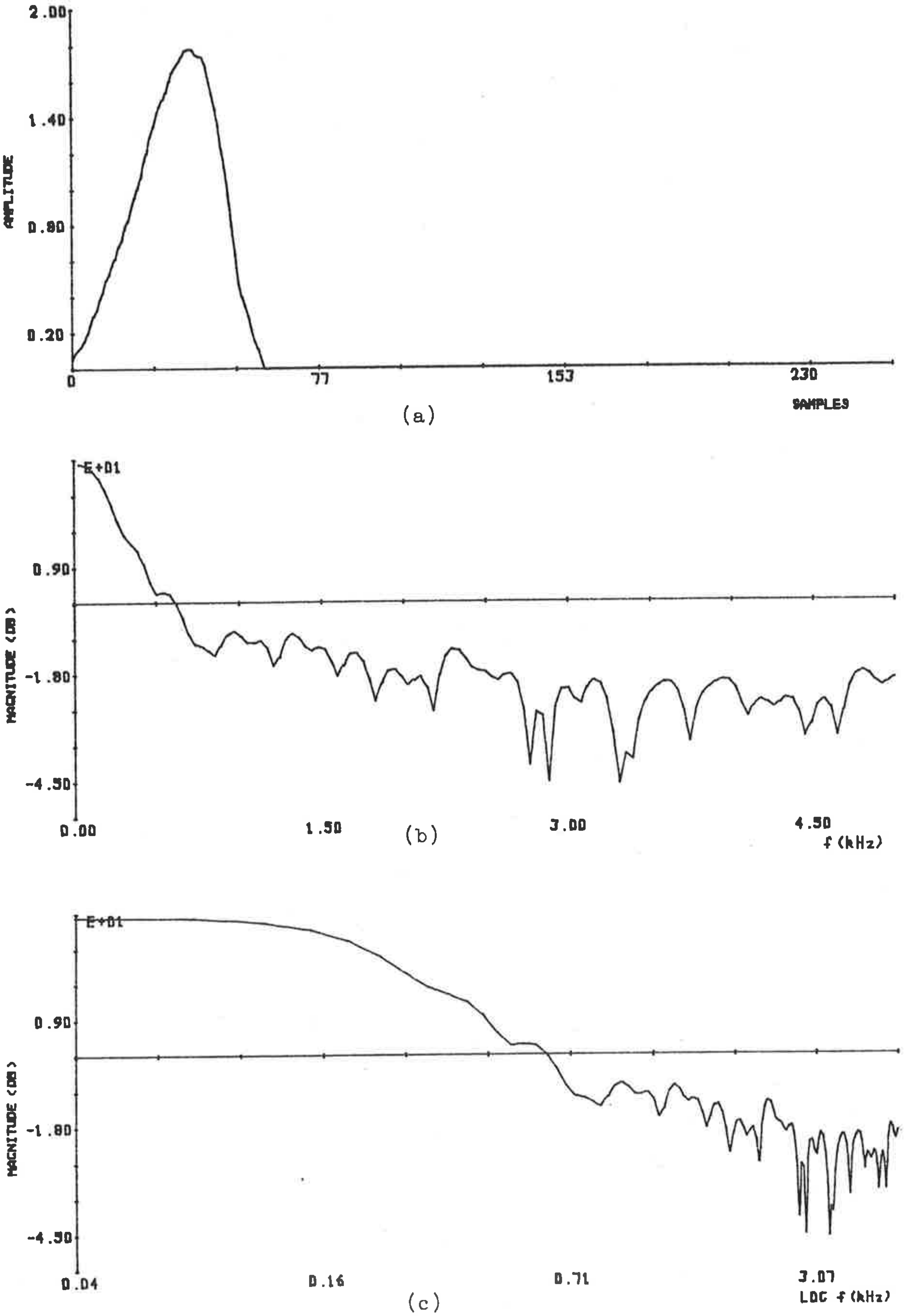
4.2 EFFECTS OF GLOTTAL PULSE EXCITATION

Accurate recovery of acoustic tube shapes from a time waveform by a conventional linear predictive analysis procedure requires an excitation which is white, i.e. has a flat spectrum. For voiced speech, the excitation of the vocal tract is by pulses of air at the glottis which have a non-white spectrum, as shown in the previous section. Therefore, errors are expected in the recovered vocal tract shape for a conventional linear predictive analysis of voiced speech. This section investigates the errors that occur in acoustic tube shape recovery when a conventional linear predictive analysis is used to recover ideal acoustic tubes excited by glottal pulse waveforms.

The analysis procedure used in this section is restricted to the autocorrleation formulation of linear prediction. If the errors in acoustic tube shape recovery due to glottal pulse excitation are to be determined, then no other non-ideal effects, such as losses and lip radiation, should be present to cloud the results. Hence, real speech waveforms cannot be used, and so all the data analysed in this section is synthetic. The main advantage of synthetic data is that the original acoustic tube shape and glottal pulse structure is known; therefore, accurate comparisons of analysis results can be made. The generation of synthetic data from ideal lossless acoustic tubes is presented in Appendix B, and no lip radiation effects are present, so the termination reflection coefficient, μ_M, is unity.

For the results of this section to be applicable to voiced speech, the glottal pulse and acoustic tube shapes used to generate synthetic data must be as close as possible to those occurring in reality. Therefore, the first glottal pulse considered is derived from a glottal pulse measured by SONDHI [1975] that occurred during normal phonation of a vowel. The real glottal pulse is plotted in Figure 4.3, along with its frequency spectrum. The procedure for determining this glottal pulse shape is by inverse filtering of the speech waveform [MILLER 1959, LINDQVIST 1965, ROSENBERG 1973, SUNDBERG and GAUFFIN 1978, FANT 1979, 1980]. Five acoustic tube shapes are used for generating the synthetic waveforms, and these shapes approximate the vocal tract shapes for five Russian vowels, |a|, |e|, |i|, |o| and |u|, as measured with X-ray photography by FANT [1960] (see Appendix C).

The results, in the form of area distances, for a Parcor analysis of the synthetic data produced by the excitation of five acoustic tube shapes approximating the Russian vowels |a|, |e|, |i|, |o| and |u| with the real glottal pulse shape are found in Table 4.1. The recovered and original acoustic tube shapes for the vowels with the largest and smallest area distances in Table 4.1 are plotted in Figure 4.4. Examination of Figure 4.4 reveals little similarity be-



(a)

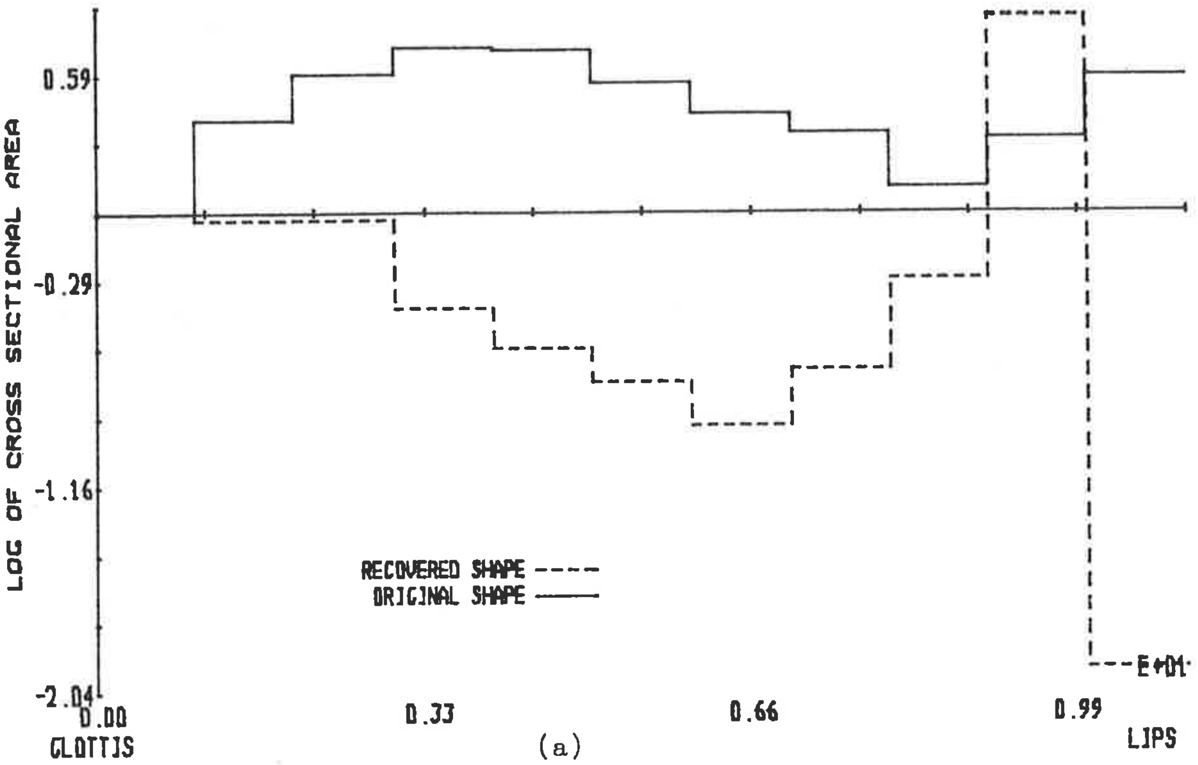
(b)

(c)

FIGURE 4.3: (a) Real glottal pulse waveform and its (b) linear and (c) logarithmic spectra.

VOWEL	AREA DISTANCES
/a/	27.87
/e/	3.48
/i/	13.33
/o/	25.38
/u/	162.95

TABLE 4.1: Area distances for a parcor linear predictive analysis of synthetic speech for five vowels generated with the real glottal pulse waveform.



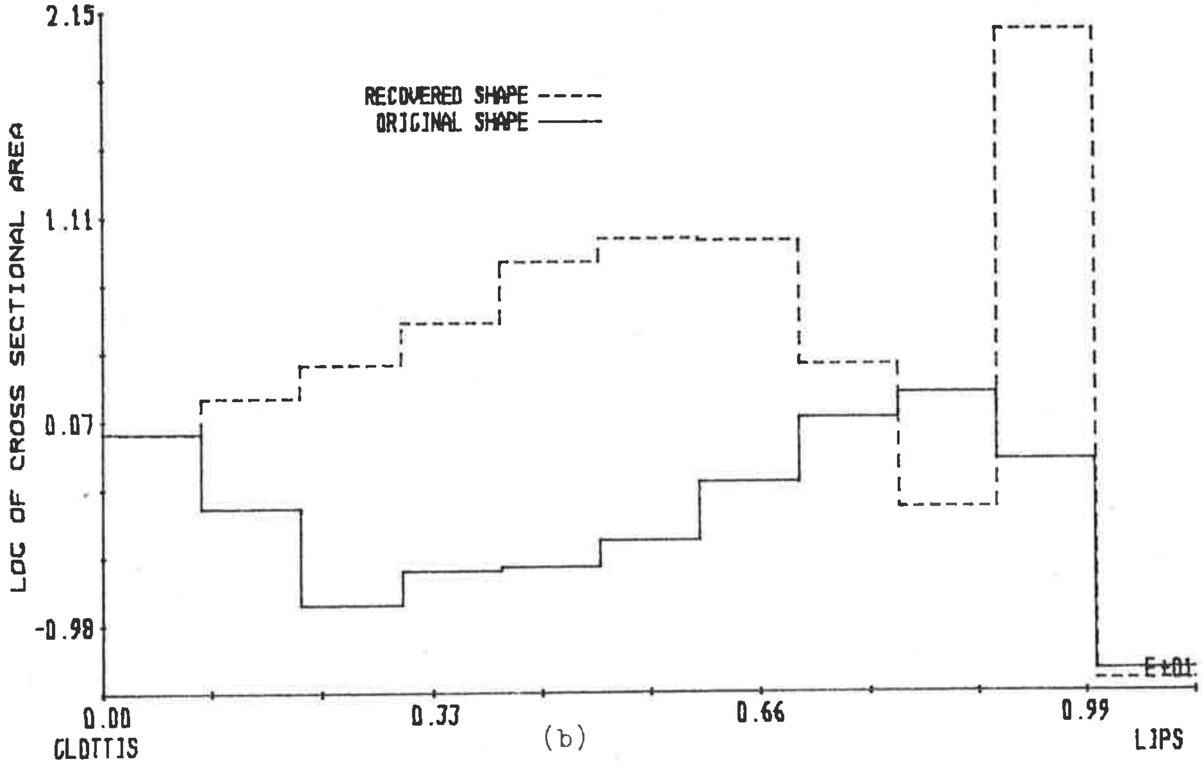


FIGURE 4.4: Comparison of recovered and original acoustic tube shape for the (a) vowel /e/ and (b) vowel /u/, for excitation by the real glottal pulse waveform.

tween the original and recovered acoustic tube shapes. Therefore, in this case, the glottal pulse excitation causes gross errors in acoustic tube shape recovery by an autocorrelation linear predictive analysis procedure.

A complete study of the effects of glottal excitation on recovered acoustic tube shapes requires an investigation for the large range of glottal shapes and spectral slopes that occur in reality. This study is achieved more effectively by using the glottal pulse models developed by a number of researchers [ROSENBERG 1971, FANT 1979], due to the ease with which the glottal pulse shape and spectral slope can be controlled.

The studies of glottal pulse models made by ROSENBERG [1971] were performed with six glottal pulse models based on either triangular, trapezoidal, polynomial or trigonometric functions. To determine which model was the best model of real glottal pulses, subjective testing was performed on synthetic speech excited by the six different glottal pulse models. Results of the subjective evaluations, carried out by means of A-B comparison tests, showed that the polynomial derived glottal pulse was a good perceptual replacement for real glottal pulses. When used in sentences, the preference score was greater than real speech, while in CVC syllables, its preference score was marginally less than real speech. Therefore, using the polynomial glottal pulse model of ROSENBERG provides glottal pulse waveforms which produce perceptually similar results to real glottal pulse waveforms.

The ROSENBERG polynomial glottal pulse model defines the glottal pulse by a rising branch

$$U_0(t) = \lambda\left[3\left(\frac{t}{T_p}\right)^2 - 2\left(\frac{t}{T_p}\right)^3\right] \qquad 0 \leqslant t \leqslant T_p \qquad (4.1)$$

and a falling branch

$$U_0(t) = \lambda\left[1 - \left(\frac{t-T_p}{T_N}\right)^2\right] \qquad T_p \leqslant t \leqslant T_p + T_N \qquad (4.2)$$

where T_p is the period during which the pulse has a positive slope, and T_N is the period during which the pulse has a negative slope. The pitch period is denoted as T by ROSENBERG, and the parameters of the polynomial pulse are then T_p/T, T_N/T and λ. The parameter λ is a scale factor and, since a scaling does not affect a linear predictive analysis, λ is set equal to unity.

The subjective testing performed by ROSENBERG on the glottal pulse models determined the best range of the parameters T_p/T and T_N/T to maximise the perceptual similarities between synthetic and real glottal pulses. It was found that there exists a fairly large tolerance for different combinations of the glottal pulse parameters. However, very small opening and closing times, or opening times less than or approximately equal to closing times were not preferred. The subjective testing found that $T_p/T = 0.40$ and $T_N/T = 0.16$ provided the best perceptual results with (T_p/T, T_N/T) equal to (.33, .09) and (.33, .19) also providing good results.

In order to present results which can be compared with those for the real glottal pulse already considered (i.e. Table 4.1 and Figure 4.3) the pitch period, T, of the ROSENBERG polynomial pulse model is chosen to be the same as that for the real glottal pulse, i.e. $T = 8$ msec. Using this value of T and the above three sets of values for the parameters T_p/T and T_N/T, three glottal pulse waveforms are generated and displayed in Figure 4.5. The frequency spectra for these three glottal pulses are presented in Figure 4.6.

Exciting the five acoustic tube shapes which approximate the five Russian vowels of FANT [1960] (see Appendix C) by the glottal pulse shapes presented in Figure 4.5 and then performing a Parcor analysis on the resultant synthetic speech data produces the area distances found in Table 4.2. Comparison of Tables 4.1 and 4.2 shows similar area distances for each vowel, and so the glottal pulses derived by ROSENBERG's polynomial pulse model are good substitutes for real glottal pulses. The similarity of area distances in Tables 4.1 and 4.2 also shows that poor acoustic tube shape recovery still occurs when synthetic glottal pulse excitation is present.

The recovered and original acoustic tube shapes for the largest and smallest area distance in Table 4.2 are plotted in Figure 4.7. The recovered acoustic tube shapes in Figure 4.7 have very little similarity to the original, as was the case for the real glottal pulse excitation in Figure 4.4. Hence, the results presented so far indicate that gross errors in acoustic tube shape recovery, by an autocorrelation linear predictive analysis, occur when glottal pulse excitation is present.

Another glottal pulse model which provides different variations of glottal pulse shape and spectral slope from the ROSENBERG polynomial glottal pulse model is one derived by FANT [1979]. The FANT glottal pulse model has a rising branch defined by

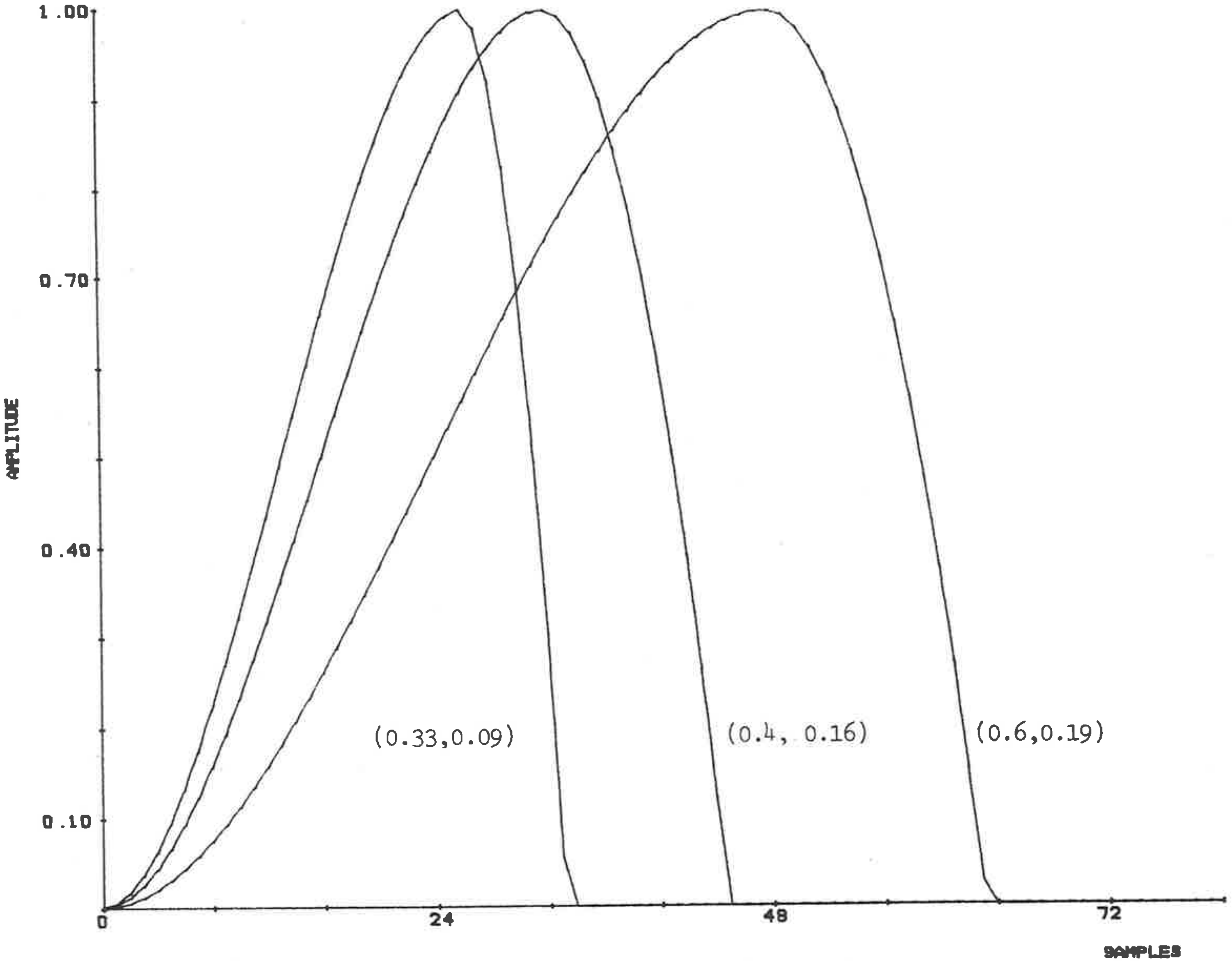
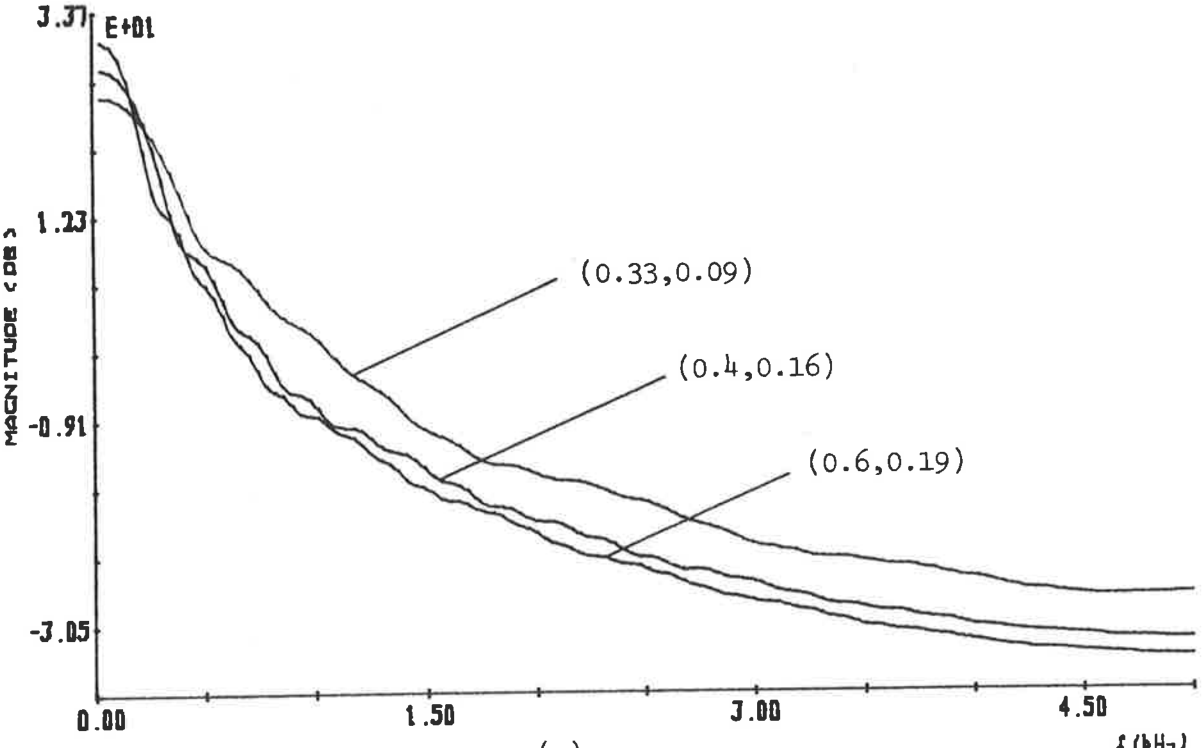
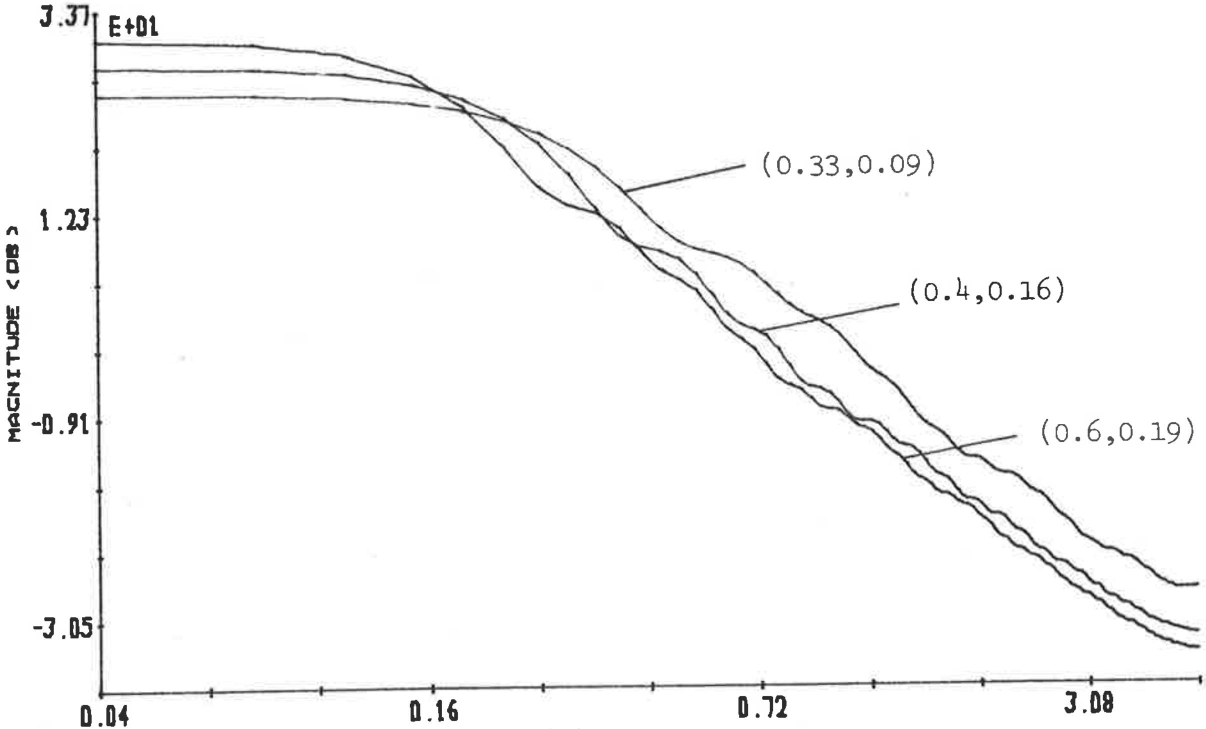


FIGURE 4.5: Glottal pulse waveforms generated from the Rosenberg polynomial glottal pulse model for various values of the parameters $(T_p/T, T_N/T)$.



(a)

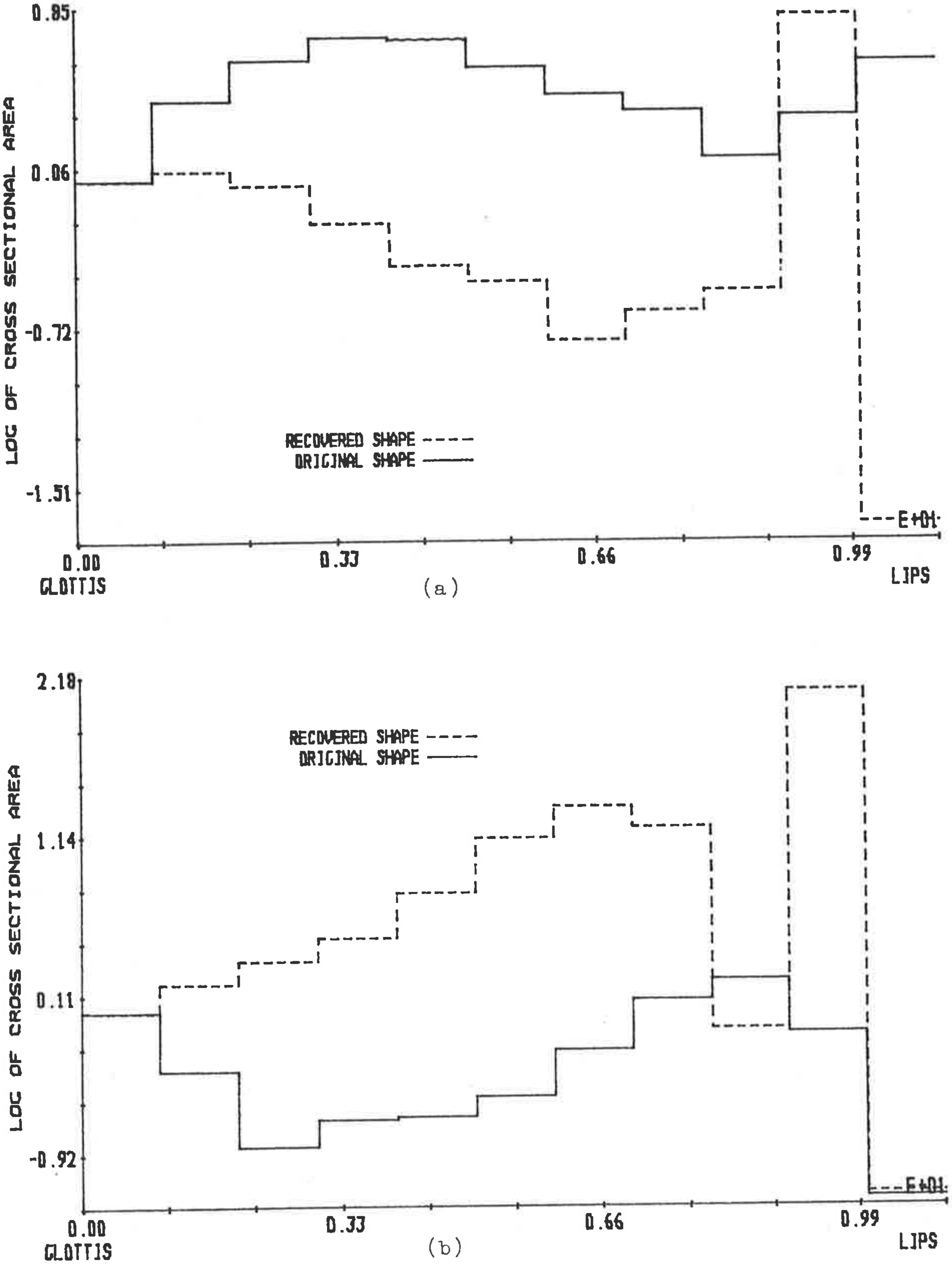


(b)

FIGURE 4.6: (a) Linear and (b) logarithmic spectra for glottal pulse waveforms generated from the Rosenberg polynomial glottal pulse model for various values of parameters $(T_p/T, T_N/T)$.

VOWEL	AREA DISTANCES		
	$T_p/T=0.33$, $T_N/T=0.09$	$T_p/T=0.4$, $T_N/T=0.16$	$T_p/T=0.6$, $T_N/T=0.19$
/a/	15.68	15.40	20.54
/e/	3.03	2.77	2.83
/i/	27.90	25.41	24.16
/o/	26.67	29.30	33.91
/u/	77.74	124.80	202.17

TABLE 4.2: Area distances for a parcor linear predictive analysis of synthetic speech for five vowels generated with glottal pulses from the Rosenberg polynomial glottal pulse model.



(a)

(b)

FIGURE 4.7: Comparison of recovered and original acoustic tube shapes for (a) the vowel /e/ and Rosenberg polynomial glottal pulse model excitation with parameters $(T_p/T,\ T_N/T)=(0.4,0.16)$, and (b) the vowel /u/ and Rosenberg polynomial glottal pulse model excitation with parameters $(T_p/T,\ T_N/T)=(0.6,0.19)$.

$$U_0(t) = \lambda\left[\tfrac{1}{2} - \cos\frac{\pi t}{T_p}\right] \qquad 0 \leqslant t \leqslant T_p \qquad (4.3)$$

and a falling branch defined by

$$U_0(t) = \lambda\left[K\cos\left(\frac{t-T_p}{T_p}\,\pi\right) - K + 1\right] \qquad T_p \leqslant t \leqslant T_p + T_N \quad (4.4)$$

where

$$T_N = \frac{T_p}{\pi}\,\mathrm{arcos}\left(\frac{K-1}{K}\right) \qquad (4.5)$$

or

$$K = \frac{1}{1 - \cos\left(\dfrac{\pi T_N}{T_p}\right)} \qquad (4.6)$$

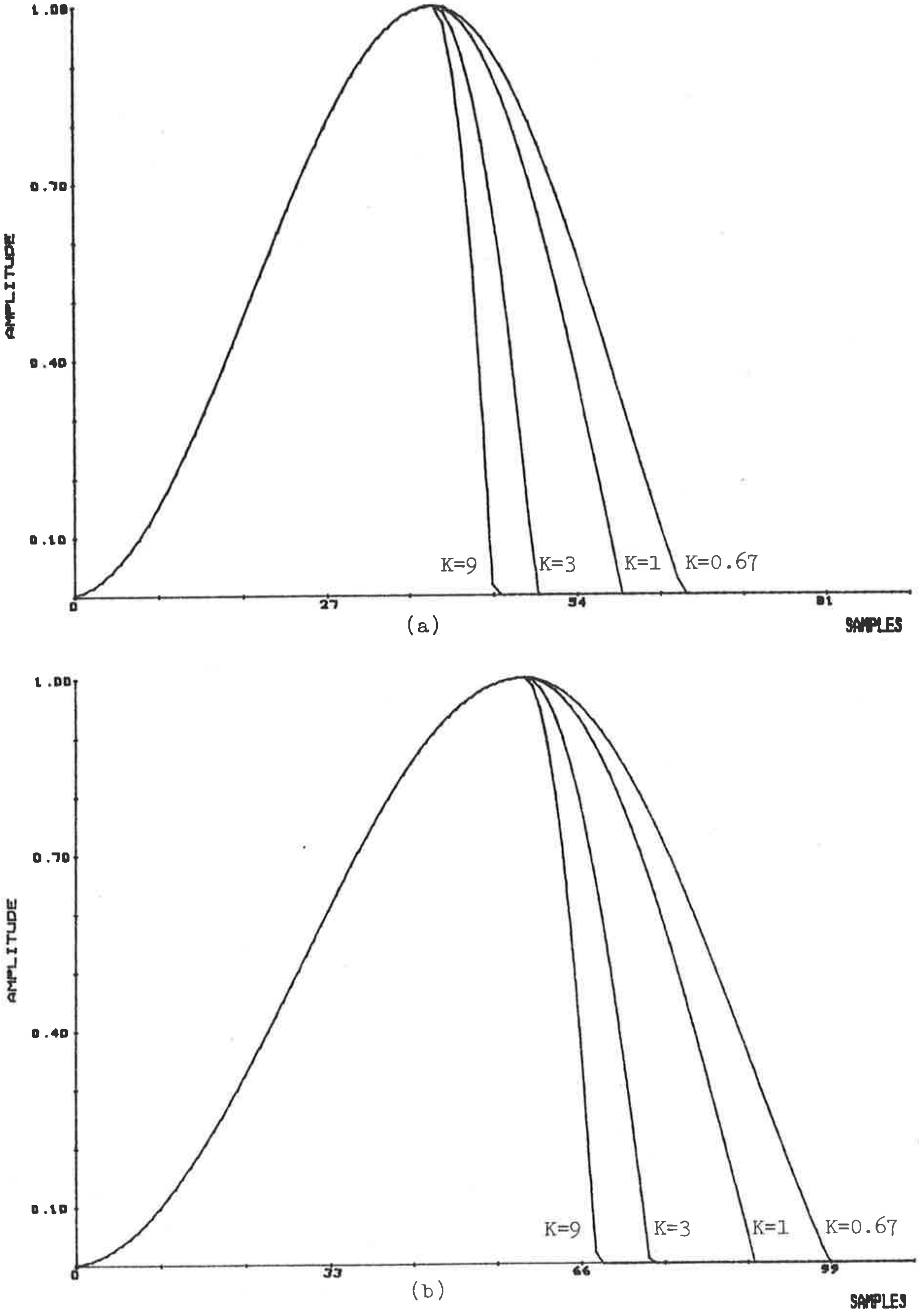
The parameter λ is a scale factor and is chosen as unity, since a scaling does not affect the results of an autocorrelation linear predictive analysis. The FANT glottal pulse model has three parameters, i.e. T_p, T_N and K, but any one can be determined from the other two; therefore, only the parameters T_p and K are used to completely define the glottal pulse shape throughout this thesis.

The range of parameters T_p and K of the FANT glottal pulse model is chosen to be consistent with the preferences found by the subjective testing of ROSENBERG [1971]. While this does not guarantee acceptable synthetic speech when using the FANT glottal pulse model, it keeps the shape of the pulses similar to real glottal pulses. ROSENBERG determined that acceptable pulse durations are between 30% and 100% of the pitch period, and opening to closing time ratios should be at least 1.5 and not greater than 6.0. Equation 4.6 translates these requirements into a restriction on the range of K to lie within 0.67 to 9.0. The restriction on the range of T_p was determined by ROSENBERG to be at least 18%, but not more than 60% of the pitch period, i.e. T_p/T restricted to range between .18 and .60.

To determine the effects of glottal pulse excitation with the glottal pulse defined by the FANT model, two values of T_p/T, namely 0.40 and 0.60, are used with the pitch period the same as before, i.e. $T = 8$ msec. The parameter K takes on the values 0.67, 1.0, 3.0 and 9.0 to provide a spread over the preferred range of K. Figure 4.8 presents plots of the glottal pulses derived from the FANT model for the above values of T_p/T and K, and the frequency spectra for each glottal pulse is presented in Figure 4.9.

Observation of the pulse shapes produced by ROSENBERG's polynomial glottal pulse model in Figure 4.5 and those by FANT's glottal pulse model in Figure 4.8 shows a significant difference in pulse shapes. For the ROSENBERG glottal pulses in Figure 4.5, the overall shape does not change dramatically, while the pulse duration does. This contrasts with the marked change in pulse shape and relatively small pulse duration changes for the FANT glottal pulses of Figure 4.8. Hence, the two glottal pulse models present different aspects of glottal pulse variations, and are not duplicating the results of each other.

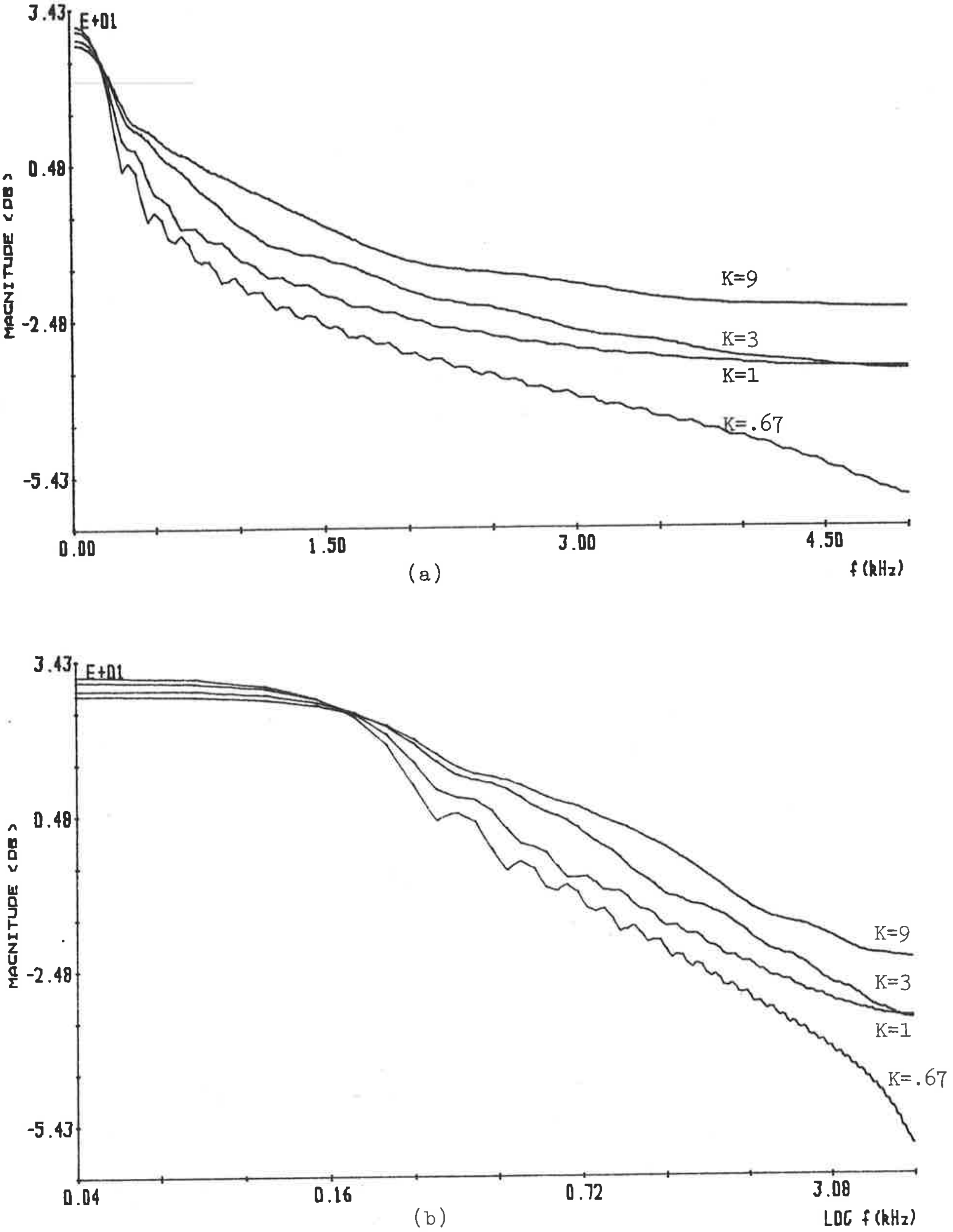
The FANT glottal pulses shown in Figure 4.8 are used to generate synthetic speech by exciting the same acoustic tube shapes approximating five Russian vocal tract shapes, as detailed in Ap-



(a)

(b)

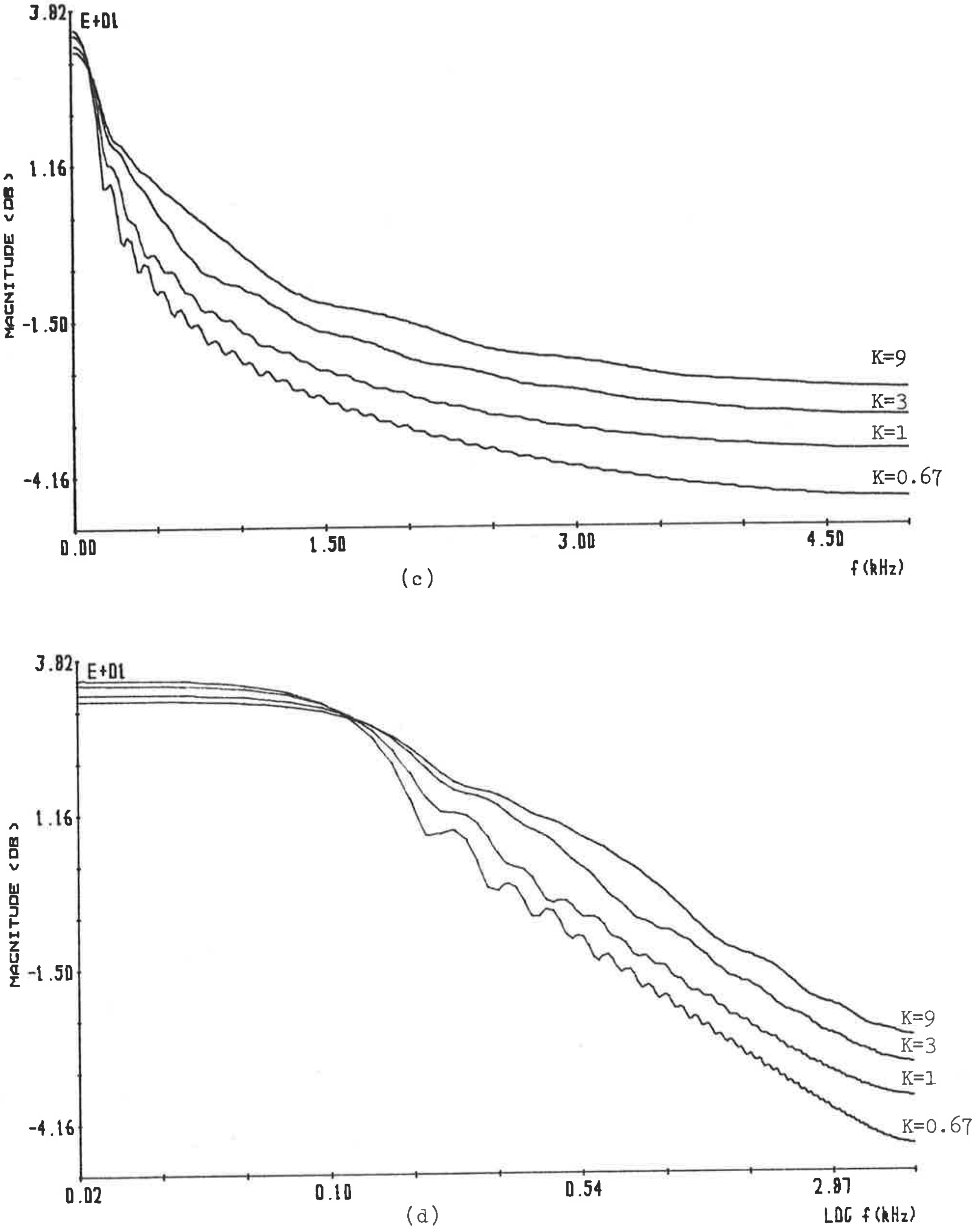
FIGURE 4.8: Glottal pulse waveforms generated from the fant glottal pulse model for various values of parameter K and (a) $T_p/T = 0.4$ and (b) $T_p/T = 0.6$.



(a)

(b)

FIGURE 4.9: (a) Linear and (b) logarithmic spectra for glottal pulse waveforms generated from the fant glottal pulse model for various values of parameter K and $T_p/T = 0.4$.



(c)

(d)

FIGURE 4.9: (c) Linear and (d) logarithmic spectra for glottal pulse waveforms generated from the fant glottal pulse model for various values of parameter K at $T_p/T = 0.6$.

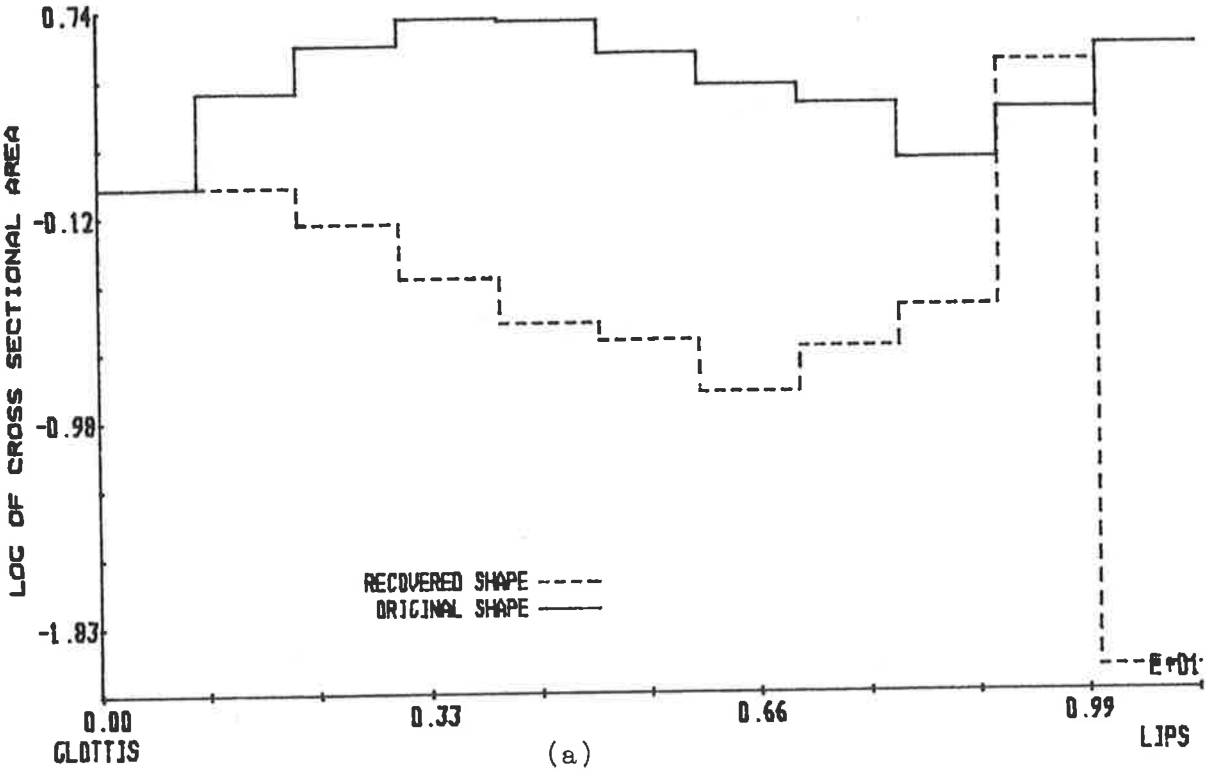
pendices B and C. Analysis of this synthetic speech by a Parcor analysis produces the area distances presented in Table 4.3. The recovered and original acoustic tube shapes for the largest and smallest area distances in Table 4.3 are plotted in Figure 4.10.

Comparison of Table 4.3 with Tables 4.1 and 4.2 shows that similar area distance values occur for each particular vowel. This indicates that poor acoustic tube shape recovery still occurs for excitation by glottal pulses generated from the FANT glottal pulse model. Examination of the acoustic tube shapes in Figure 4.10 verifies that there exists very little similarity between the recovered and original acoustic tube shapes. These results are consistent with the gross errors in acoustic tube shape recovery observed for excitations by the ROSENBERG glottal polynomial pulse model and real glottal pulse.

Table 4.3 shows a trend of decreasing area distances (i.e. improved acoustic tube shape recovery) for increasing values of K for a particular vowel, which is also observed in Table 4.2 for decreasing values of T_p/T and T_N/T. Study of the frequency spectra for the FANT and ROSENBERG glottal pulses in Figure 4.9 and 4.6, respectively, shows that the spectral slope of the glottal pulses decreases for increasing K and decreasing T_p/T and T_N/T. Therefore, the magnitude of the errors for acoustic tube shape recovery depend on the spectral slope of the glottal pulse used as an excitation, i.e. better acoustic tube shape recovery occurs for smaller glottal pulse spectral slope. This suggests that improved acoustic tube shape recovery is realised by removing from the speech waveform the spectral slope that is added by the glottal pulse excitation.

VOWEL	AREA DISTANCES							
	$T_p/T=0.4$				$T_p/T=0.6$			
	K=0.67	K=1	K=3	K=9	K=0.67	K=1	K=3	K=9
/a/	29.2	36.5	14.2	13.5	99.4	66.9	22.1	16.2
/e/	11.2	4.1	3.0	2.8	16.6	14.1	2.9	2.7
/i/	75.0	20.5	26.6	34.9	32.9	28.6	27.5	29.9
/o/	27.8	42.0	20.4	23.9	87.4	77.6	32.1	28.9
/u/	78.7	194.4	107.4	90.5	263.2	336.7	224.9	147.8

TABLE 4.3: Area distances for a parcor linear predictive analysis of synthetic speech for five vowels generated with glottal pulses from the fant glottal pulse model.



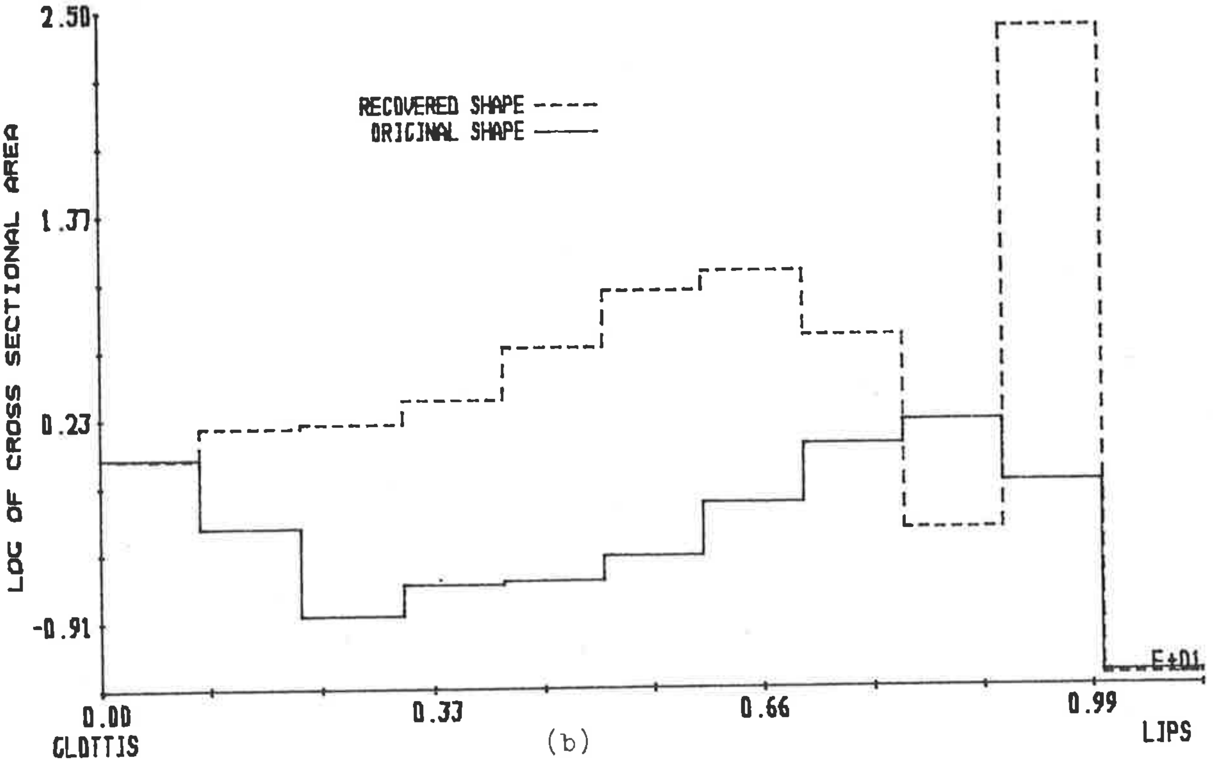


FIGURE 4.10: Comparison of recovered and original acoustic tube shapes for (a) the vowel /e/ and the fant glottal pulse model excitation with $(K,T_p/T) = (9,0.6)$ and (b) the vowel /u/ and the fant glottal pulse model excitation with $(K,T_p/T) = (1,0.6)$.

The results presented in this section lead to the conclusion that gross errors in acoustic tube shape recovery are made by autocorrelation linear predictive procedures when analysing acoustic waveforms from acoustic tubes excited by glottal pulse type waveforms. The magnitude of the errors in acoustic tube shape recovery are so large that little similarity occurs between the recovered and original acoustic tube shapes. The magnitude of the errors is dependent on the spectral slope of the glottal pulse, and the results presented in this section indicate that improved acoustic tube shape recovery is achieved by removal of the glottal pulse spectral slope from the acoustic waveform. This section has shown that there exists a need for a procedure or method which can account for glottal pulse excitation if accurate acoustic tube shape recovery is to be achieved.

4.3 EFFECTIVENESS OF CONVENTIONAL MEANS TO OVERCOME GLOTTAL PULSE EXCITATION

The glottal pulse excitation of voiced speech has been studied by many researchers, and resulted in the shape and spectral variations of the glottal pulse excitation being well documented. Therefore, at the time of analysis, there does exist some a priori knowledge of the glottal pulse excitation. It has been shown in the previous section that the glottal pulse excitation causes gross errors in acoustic tube shape recovery and, in the past, researchers have exploited the a priori knowledge of the glottal pulse excitation to improve acoustic tube shape recovery. One of the procedures that has been used is pre-emphasis, and this section evaluates the effectiveness of pre-emphasis techniques to produce accurate acoustic tube shape recovery when glottal pulse excitation is present.

Another method that has been used to overcome glottal pulse excitation effects is to perform the analysis during the closed glottis interval. During the process of vocal cord vibration, a period may exist when air flow into the vocal tract is stopped by closure of the vocal folds. This period is referred to as the closed glottis interval, and analysis of speech during the closed glottis interval has been shown to provide accurate acoustic tube shape recovery of the vocal tract [PINSON 1963, STEIGLITZ and DICKINSON 1977, ROGERS 1974, MARKEL and GRAY 1976]. Historically, the closed glottis interval anlaysis concept was introduced by MILLER et al [1961].

Despite the obvious advantages of closed glottis interval analysis, there does not exist a satisfactory method for determinining that interval from the speech waveform. Direct extraction of glottal waveforms and, hence, the determination of the closed glottis interval, has been considered by HOLMES [1976], but shown to be computationally tedious and to produce many ambiguities. The difficulty with determining the closed glottis interval led to the formulation of empirical rules [RABINER et al, 1976], but did not provide a reliable process for finding the interval.

Reasonable success has been had by using the largest value of the linear prediction residual (or error signal) to indicate the start of the closed glottis interval. Improvements on this basic concept have been made by STRUBE [1974], who modified a measure for linear predictability proposed by SOBAKIN [1972]. STRUBE's procedure has been found to determine the closed glottis interval reliably, but is computationally complex, and so is only suitable for special investigations. A less computationally demanding procedure, but with similar reliability, has been produced by ANANTHAPADMANABHA et al [1975, 1977, 1979]. The procedure applies epoch filter theory to the linear prediction residual, and has sufficient resolution to identify multiple pulse excitations within a single pitch period.

Apart from the problem of correctly identifying the closed glottis interval, others exist which restrict the general application of closed glottis interval analysis to all types of voiced sounds. The closed glottis interval is generally short, and so covariance rather than autocorrelation analyses must be used. As discussed in Chapter 2, there exist some distinct disadvantages with covariance analyses. Even when a covariance analysis is used, sufficient data must be available to provide accurate covariances, and so closed glottis interval analysis is restricted to relatively low pitch voiced sounds.

Another problem is that vocal fold closure may not occur if insufficient vocal effort is used and, hence, the vocal tract is always excited during the production of that sound. The research of ROSENBERG [1971] led him to conclude that vocal fold closure does not even occur as a general rule. MARKEL and GRAY [1976] state that the success of closed glottis interval analysis is strongly dependent upon the speaker, utterance and speaking conditions. Speakers or sounds with a low pitch tend to ensure glottis closure for a sufficiently long period of time and, therefore, produce the best results. Conditions of high pitch (e.g. women and child speakers) or low vocal effort lead to unreliable results.

An indication of the effectiveness of a closed glottis interval analysis on acoustic tube shape recovery is obtained by considering synthetic data for the vowel |e|. A single real glottal pulse shape is used to excite a set of eleven commensurate acoustic tubes approximating the vocal tract shape for the Russian vowel |e| (see Appendices B and C). Figure 4.11 presents the recovered and original acoustic tube shapes for a covariance analysis (using COVAR [GRAY and MARKEL, 1979]) of the synthetic data during the closed glottis interval and with various portions of the glottis pulse included in the covariance analysis.

When no portion of the glottal pulse is included in the analysis, then accurate acoustic tube shape recovery is observed (Figure 4.11(a)). However, Figure 4.11 shows that large errors occur when small portions of the glottis waveform are present during the analysis period. If a significant portion of the glottis pulse is included in the analysis, then the recovered acoustic tube shape may have no resemblance to the original. Hence, it can be concluded that the effectiveness of closed glottis interval analysis is highly dependent on the accuracy with which the interval is determined, and the existence of a sufficiently long closed glottis interval.

In contrast to closed glottis interval analysis, which finds a region where the glottal pulse can be ignored, pre-emphasis techniques attempt to filter out the glottal pulse from the speech waveform. The available pre-emphasis techniques generally attempt

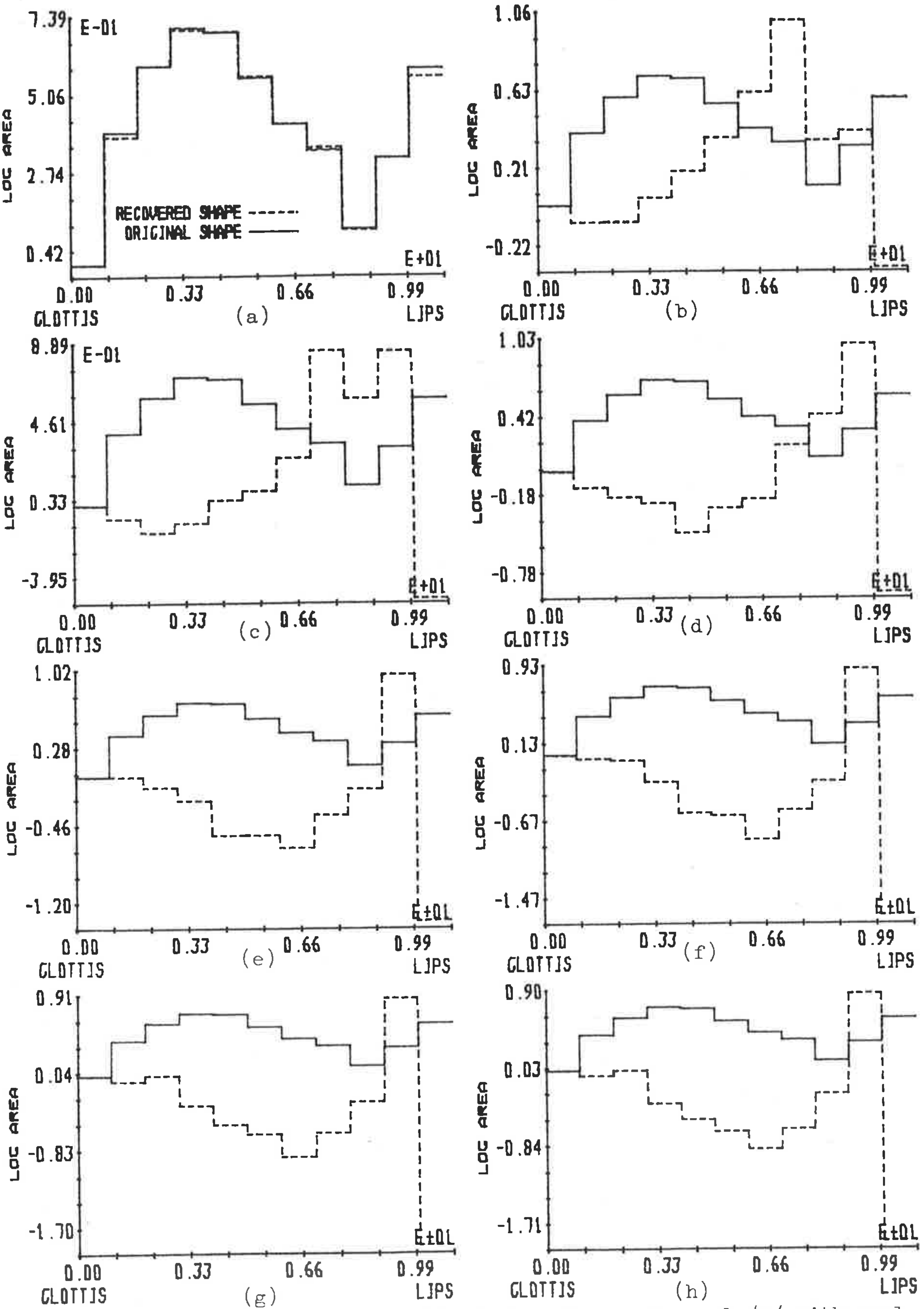


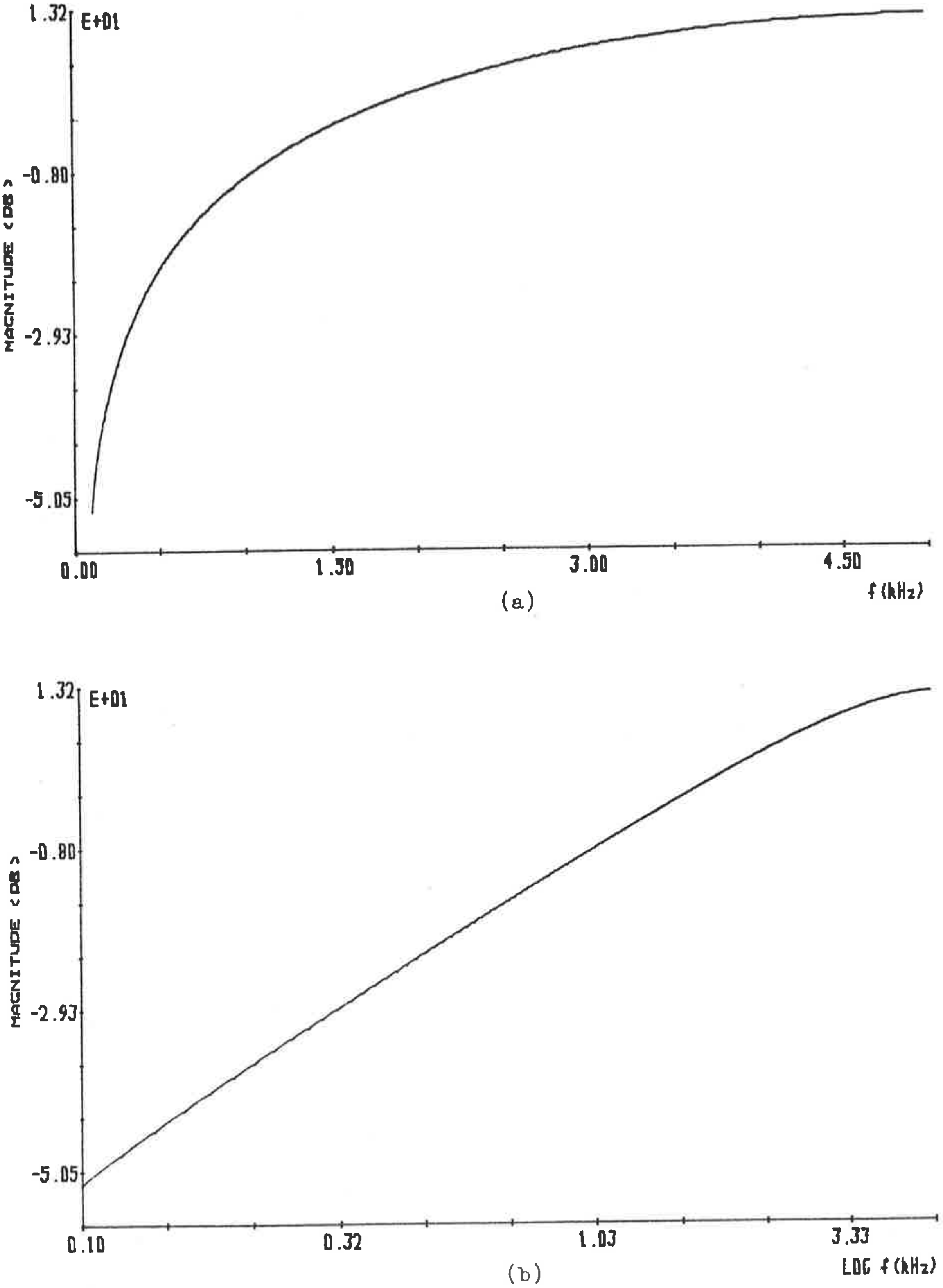
FIGURE 4.11: Closed glottis interval analysis of vowel /e/ with real glottal pulse excitation when the analysis window contains (a) 0%, (b) 5%, (c) 10%, (d) 30%, (e) 50%, (f) 70%, (g) 90% and (h) 100% of the glottal pulse waveform.

to remove glottal pulse and lip radiation effects simultaneously, but this section only considers the effectiveness of the pre-emphasis techniques to remove glottal pulse effects. Where possible, only that part of the pre-emphasis technique which deals with glottal pulse effects is considered.

When determining the effectiveness of pre-emphasis techniques to remove glottal pulse excitation, it is necessary that lip radiation, losses and other non-ideal linear prediction speech model properties do not cloud the results. Hence, all the speech data used here is synthetic, with all but the excitation assumption of the acoustic tube/linear prediction model of speech being satisfied. The analysis procedure used to obtain the recovered acoustic tube shapes is the autocorrelation formulation of linear prediction, in particular the Parcor formulation (see Chapter 2). Since synthetic data is used, the whole synthetic speech waveform is available, and so no window is used.

The first and most elementary form of pre-emphasis is to add in +12 dB per octave pre-emphasis to the speech waveform before analysis. In the z domain, the +12 dB per octave pre-emphasis is applied by a filter of the form $(1-z^{-1})^2$, i.e. two coincident real zeros at z equal to unity. Figure 4.12 presents the linear and logarithmic frequency spectrum for this pre-emphasis filter, which shows a +12 dB per octave spectral slope except near the half sampling frequency.

Evaluation of the +12 dB per octave pre-emphasis is performed with synthetic speech data generated with acoustic tube shapes approximating five Russian vowels (see Appendix C) and a glottal pulse



(a)

(b)

FIGURE 4.12: (a) Linear and (b) logarithmic spectra for the +12dB per octave pre-emphasis filter, $(1-z^{-1})^2$.

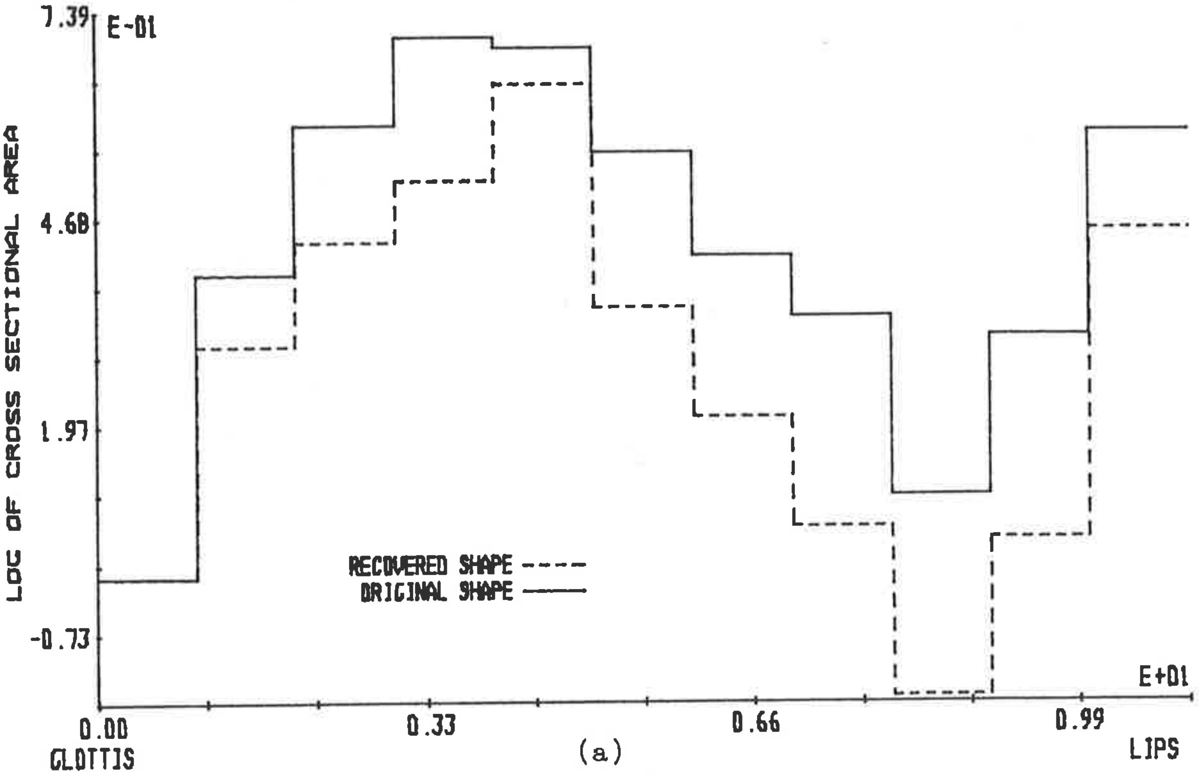
excitation, as defined in Figure 4.3. The results of pre-emphasizing the synthetic speech data with the filter $(1-z^{-1})^2$ and then performing a Parcor analysis are presented in Table 4.4 in the form of area distances. For comparison purposes, the results of Table 4.1, i.e. for no pre-emphasis, are reproduced in Table 4.4. Figure 4.13 presents a comparison of recovered and original acoustic tube shapes for two vowels from Table 4.4.

Comparison of the results in Table 4.4 for no pre-emphasis and a +12 dB per octave pre-emphasis shows a significant reduction in the area distances for a constant pre-emphasis of +12 dB per octave. From the recovered and original acoustic tube shapes presented in Figure 4.13, it is concluded that reasonable and also poor acoustic tube shape recovery can occur. In general, a +12 dB per octave pre-emphasis provides smaller area distance and, hence, better acoustic tube shape recovery than no pre-emphasis, but does not guarantee accurate acoustic tube shape recovery.

To permit a general conclusion to be made on the effectiveness of a constant +12 dB per octave pre-emphasis, to provide improved acoustic tube shape recovery, additional synthetic data using the glottal pulse models of FANT and ROSENBERG are considered. These glottal pulse models produce a wide range of excitation spectral slopes and pulse shapes, similar to those occurring for real speech waveforms. The same values of T_p/T, T_N/T and K of the glottal pulse models as used in Section 4.2 are used here, so that a comparison of results can be made. The area distances for a +12 dB per octave pre-emphasis of synthetic speech data followed by a Parcor analysis is presented in Tables 4.5 and 4.6, for ROSENBERG and FANT glottal pulse models, respectively.

VOWEL	AREA DISTANCES	
	+12dB Per Octave	No Pre-Emphasis
/a/	1.46	27.87
/e/	1.11	3.48
/i/	2.72	13.33
/o/	1.93	25.38
/u/	4.50	162.95

TABLE 4.4: Area distances for a parcor analysis of synthetic speech for five vowels generated with the real glottal pulse and pre-emphasized by a +12dB per octave pre-emphasis and no pre-emphasis.



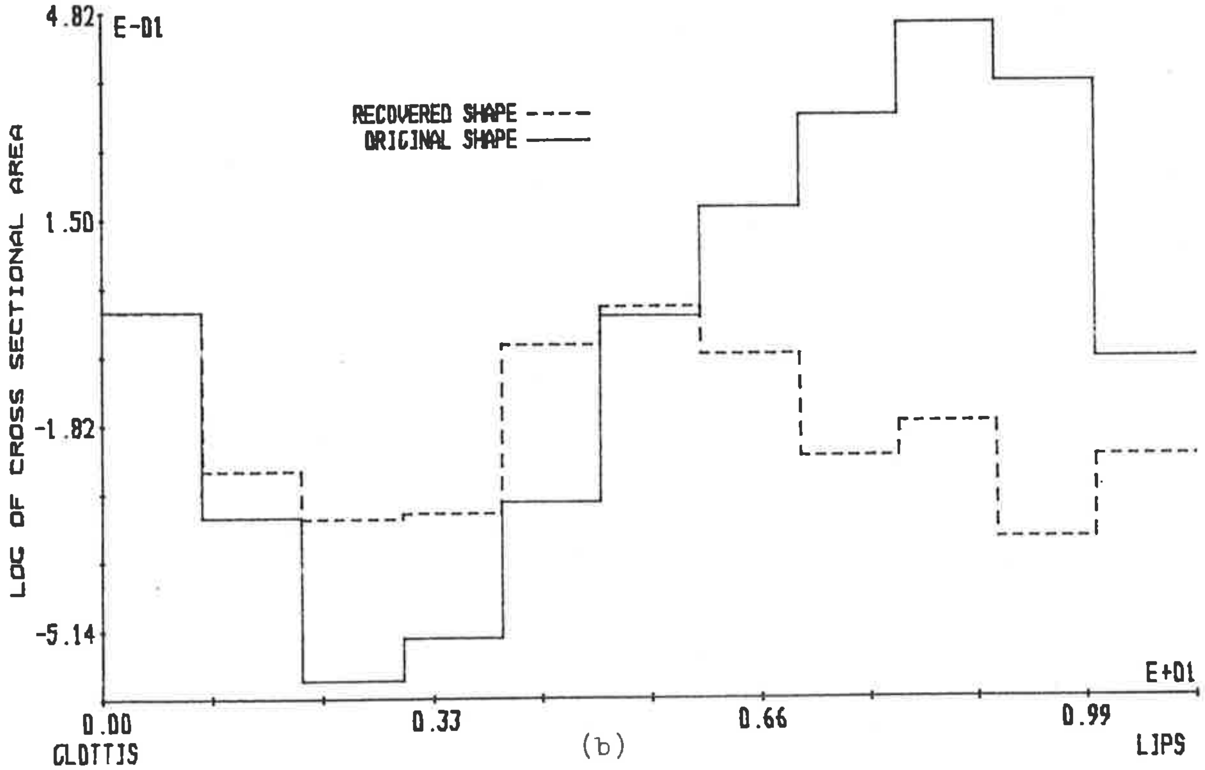


FIGURE 4.13: Comparison of recovered and original acoustic tube shapes after a +12dB per octave pre-emphasis of (a) the vowel /e/ with real glottal pulse excitation and (b) the vowel /o/ with real glottal pulse excitation.

VOWEL	AREA DISTANCES		
	T_p/T=0.33, T_N/T=0.09	T_p/T=0.4, T_N/T=0.16	T_p/T=0.6, T_N/T=0.19
/a/	6.28	6.48	1.55
/e/	1.13	1.19	0.64
/i/	10.95	6.85	2.75
/o/	7.21	8.71	2.50
/u/	21.11	23.07	9.01

TABLE 4.5: Area distances for a parcor analysis of synthetic speech for five vowels generated with glottal pulses from the Rosenberg polynomial glottal pulse model and pre-emphasized by a +12dB per octave pre-emphasis.

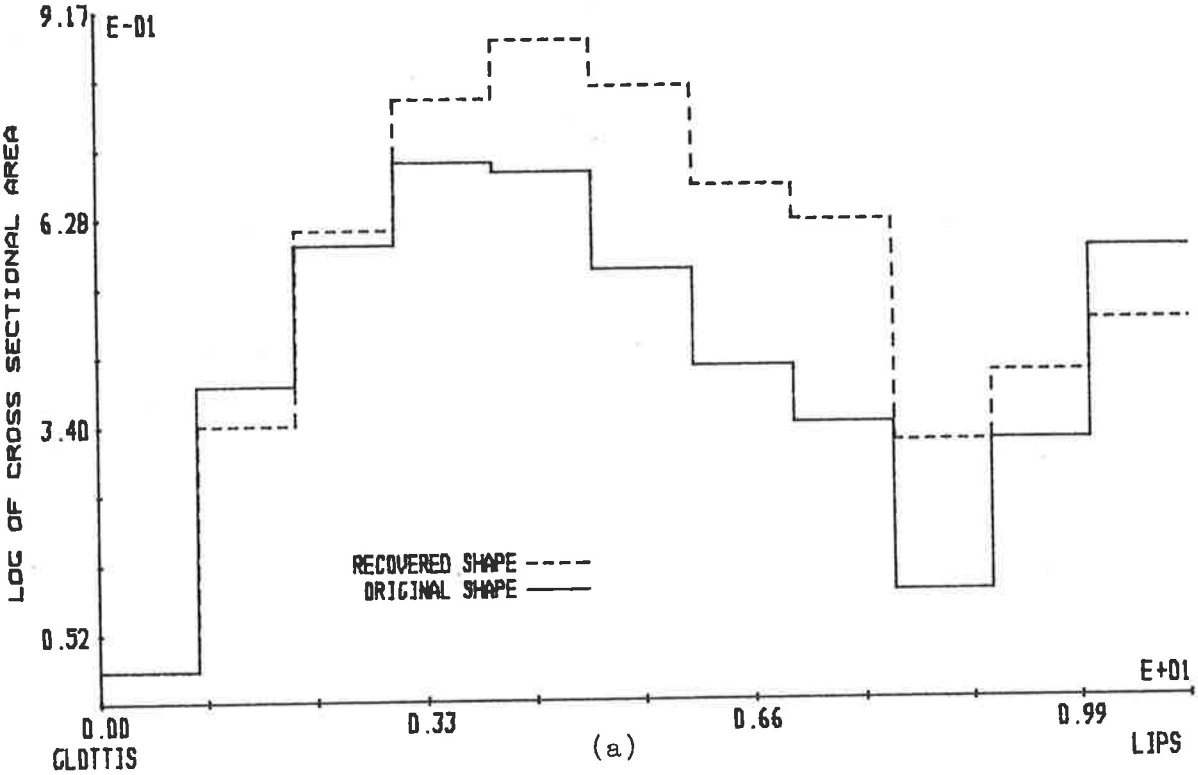
VOWEL	AREA DISTANCES							
	T_p/T=0.4				T_p/T=0.6			
	K=0.67	K=1	K=3	K=9	K=0.67	K=1	K=3	K=9
/a/	2.33	1.17	2.32	8.64	2.09	1.22	0.17	4.63
/e/	1.58	0.97	2.81	1.13	0.89	0.51	1.76	1.61
/i/	46.52	4.82	3.21	7.66	7.33	2.29	2.11	5.01
/o/	2.48	1.11	4.13	13.92	2.13	1.26	0.66	8.17
/u/	4.49	1.44	15.55	23.26	1.41	0.27	6.88	16.87

TABLE 4.6: Area distances for a parcor analysis of synthetic speech for five vowels generated with glottal pulses from the fant glottal pulse model and preemphasized by a +12dB per octave pre-emphasis.

The area distances presented in Tables 4.5 and 4.6 show a wide range of area distance values within each column. This can be explained by considering the large variations, in local regions, of both the excitation and acoustic tube shape impulse response spectra. When a constant dB per octave spectral slope is applied, various degrees of spectral correction occur over the frequency range considered. The positions at which good or poor matching occurs depend on the acoustic tube shape and the excitation waveform, and so variations in area distances can be expected when the excitation and/or acoustic tube shape changes. Hence, it is necessary to use a wide variety of excitation waveforms and acoustic tube shapes, such as used here, if a general conclusion is to be drawn on the effectiveness of a +12 dB per octave in providing accurate acoustic tube shape recovery.

Comparing area distances for no pre-emphasis and a +12 dB per octave pre-emphasis of synthetic speech where ROSENBERG's polynomial pulse is used for the excitation, i.e. comparing Tables 4.2 and 4.5 respectively, reveals smaller area distances occurring when the pre-emphasis is used. The synthetic speech waveforms used to generate Figure 4.7 were pre-emphasized by +12 dB per octave, and then analysed by a Parcor linear predictive analysis to produce the recovered and original acoustic tubes presented in Figure 4.14. On comparing Figures 4.7 and 4.14, an obvious improvement in acoustic tube shape recovery is found, but accurate acoustic tube shape recovery is not necessarily achieved.

The smallest area distances in Table 4.5 occur for each acoustic tube shape when the excitation pulse parameters (T_p/T, T_N/T) are



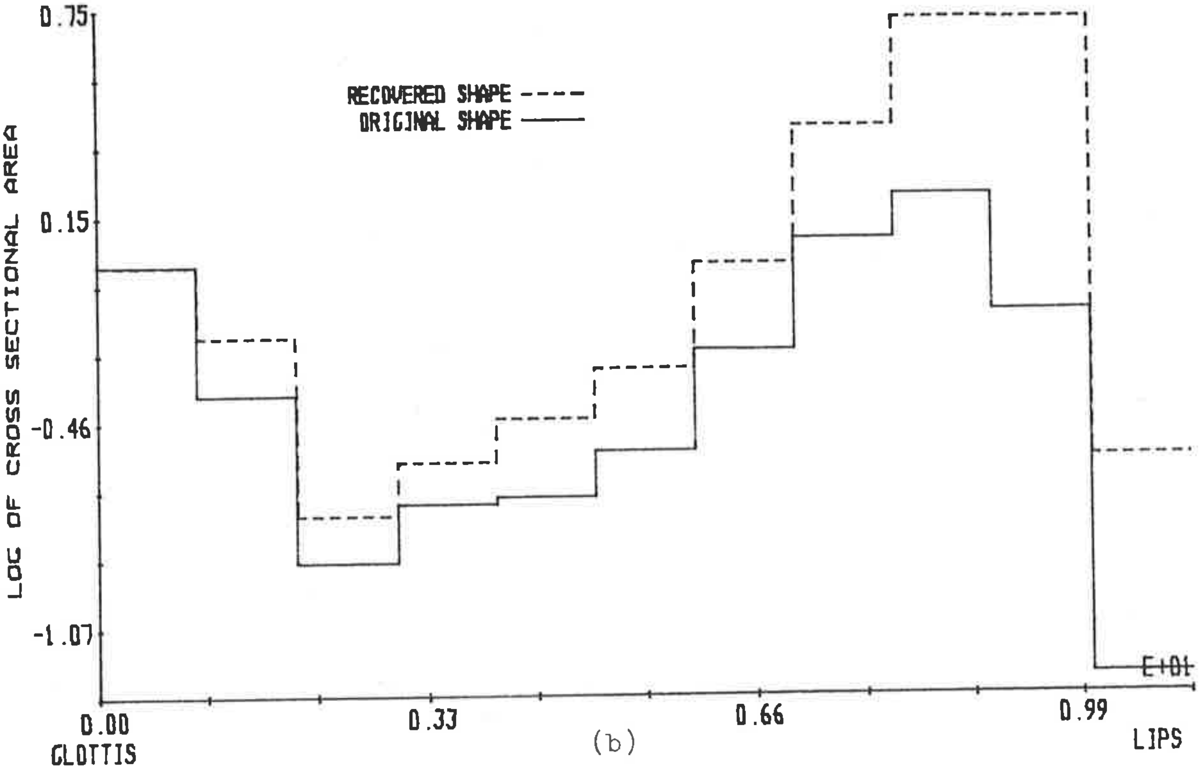


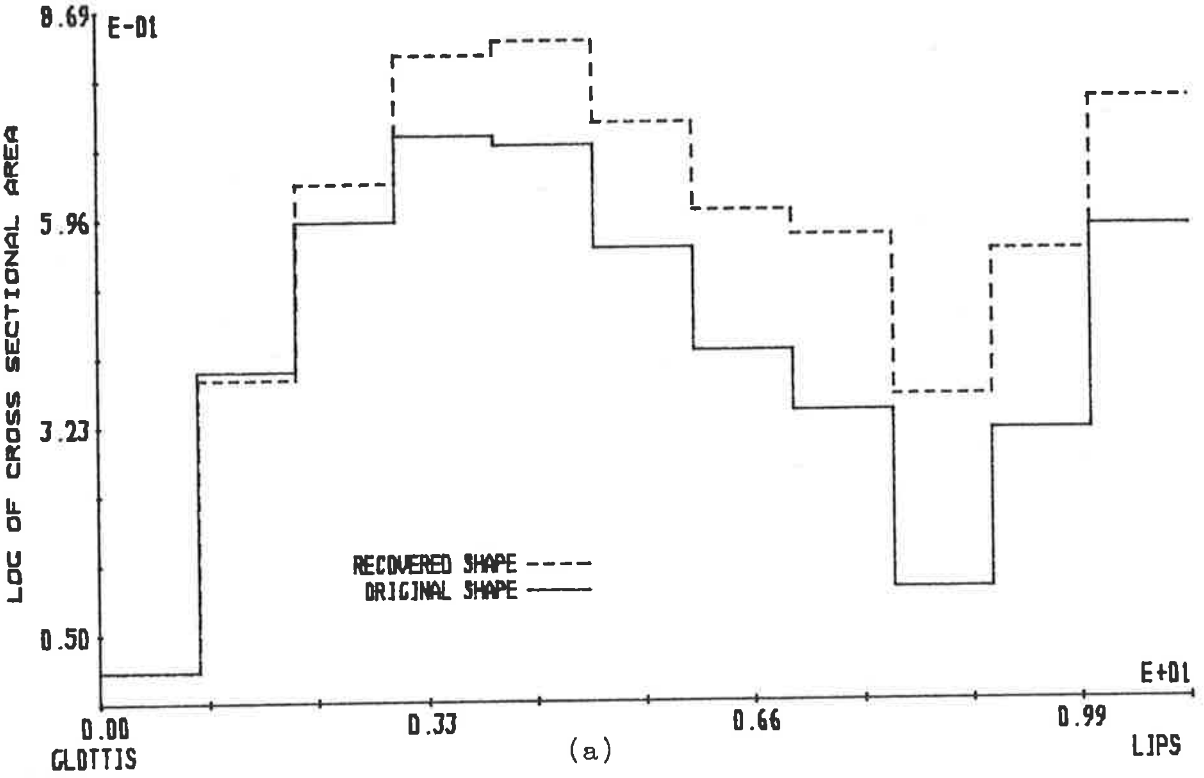
FIGURE 4.14: Comparison of recovered and original acoustic tube shapes after a +12dB per octave pre-emphasis of (a) the vowel /e/ excited by the Rosenberg polynomial glottal pulse model with parameters $(T_p/T, T_N/T) = (0.4, 0.16)$ and (b) the vowel /u/ excited by the Rosenberg polynomial glottal pulse model with parameters $(T_p/T, T_N/T) = (0.6, 0.19)$

equal to (0.6, 0.19). The spectral slopes* for the three excitation pulses used are -10.6 dB, -9.2 dB and -9.7 dB per octave for $(T_p/T, T_N/T)$ equal to (0.6, 0.19), (0.4, 0.16) and (0.33, 0.09), respectively. Hence, the application of a +12 dB per octave pre-emphasis provides more effective glottal pulse removal when the glottal pulse spectral slope is closest to -12 dB per octave.

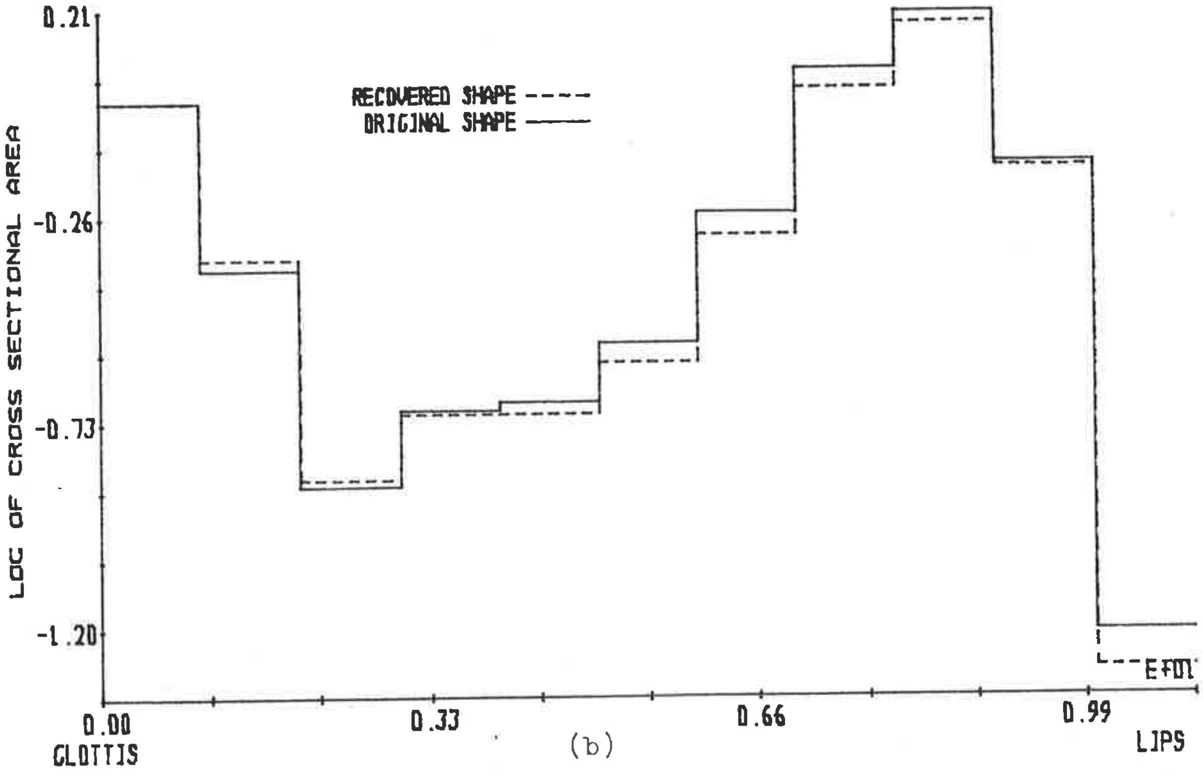
A comparison of Figures 4.3 and 4.6, where the FANT glottal pulse model was used, shows a +12 dB per octave pre-emphasis produces smaller area distances than no pre-emphasis. The synthetic speech waveforms used to generate Figure 4.10 were pre-emphasized by +12 dB per octave and then analysed by a Parcor linear predictive analysis to produce the recovered and original acoustic tube shapes presented in Figure 4.15. A comparison of Figures 4.10 and 4.15 shows an improvement in acoustic tube shape recovery when a +12 dB per octave pre-emphasis is used.

Of the glottal pulses used as excitation waveforms to generate the data of Table 4.6, the one with $(T_p/T, K)$ equal to (0.6, 3.0) has a spectral slope which is closest to -12 dB per octave, i.e. -11.9 dB per octave. From the discussions above, the effects of this glottal pulse on the recovered acoustic tube shape should be alsmot compeletely removed by a +12 dB per octave pre-emphasis, and observation of Table 4.6 shows this to be the case. Hence, it is concluded that a +12 dB per octave pre-emphasis is most effective when the excitation waveform has a spectral slope close to -12 dB per octave.

*Average over the range 0.1 to 4.5 kilohertz.



(a)



(b)

FIGURE 4.15: Comparison of recovered and original acoustic tube shapes after a +12dB per octave pre-emphasis of (a) the vowel /e/ excited by the fant glottal pulse model with parameters $(K, T_p/T) = (9, 0.6)$ and (b) the vowel /u/ excited by the fant glottal pulse model with parameters $(K, T_p/T) = (1, 0.6)$.

To observe the effects of changing the spectral slope of the excitation on the accuracy of acoustic tube shape recovery after a +12 dB per octave pre-emphasis and a Parcor analysis, Figure 4.16 plots the recovered and original acoustic tube shapes for the vowel |a| excited by the glottal pulses used in Table 4.6. The spectral slopes of the glottal pulse waveforms used to generate Figure 4.16(a) to (h) are -14.6 dB, -13.1 dB, -12.6 dB, -12.4 dB, -12.4 dB, -11.9 dB, -11.2 dB and -10.4 dB per octave, respectively. Figure 4.16 shows that, whenever the glottal pulse spectral slope is near -12 dB per octave, then good acoustic tube shape recovery occurs. However, only one or two dB variation from -12 dB per octave causes large errors in acoustic tube shape recovery.

The above results lead to the conclusion that a +12 dB per octave pre-emphasis (supplied by the filter $(1-z^{-1})^2$) produces smaller area distances and, hence, improved acoustic tube shape recovery in comparison with no pre-emphasis. In most cases, however, accurate acoustic tube shape recovery is not achieved by using a +12 dB per octave pre-emphasis. The best results for a +12 dB per octave pre-emphasis were found when the excitation glottal pulse spectral slope is close to -12 dB per octave. Hence, a +12 dB per octave pre-emphasis does not, in general, effectively remove glottal pulse excitation effects from speech waveforms and permit accurate acoustic tube shape recovery.

Despite the inadequacies of the fixed +12 dB per octave pre-emphasis, it was shown that, when the pre-emphasis has a spectral slope opposite to that of the glottal pulse, then good acoustic tube shape recovery resulted. To determine if this is a general result for all spectral slopes, a pre-emphasis is applied, the

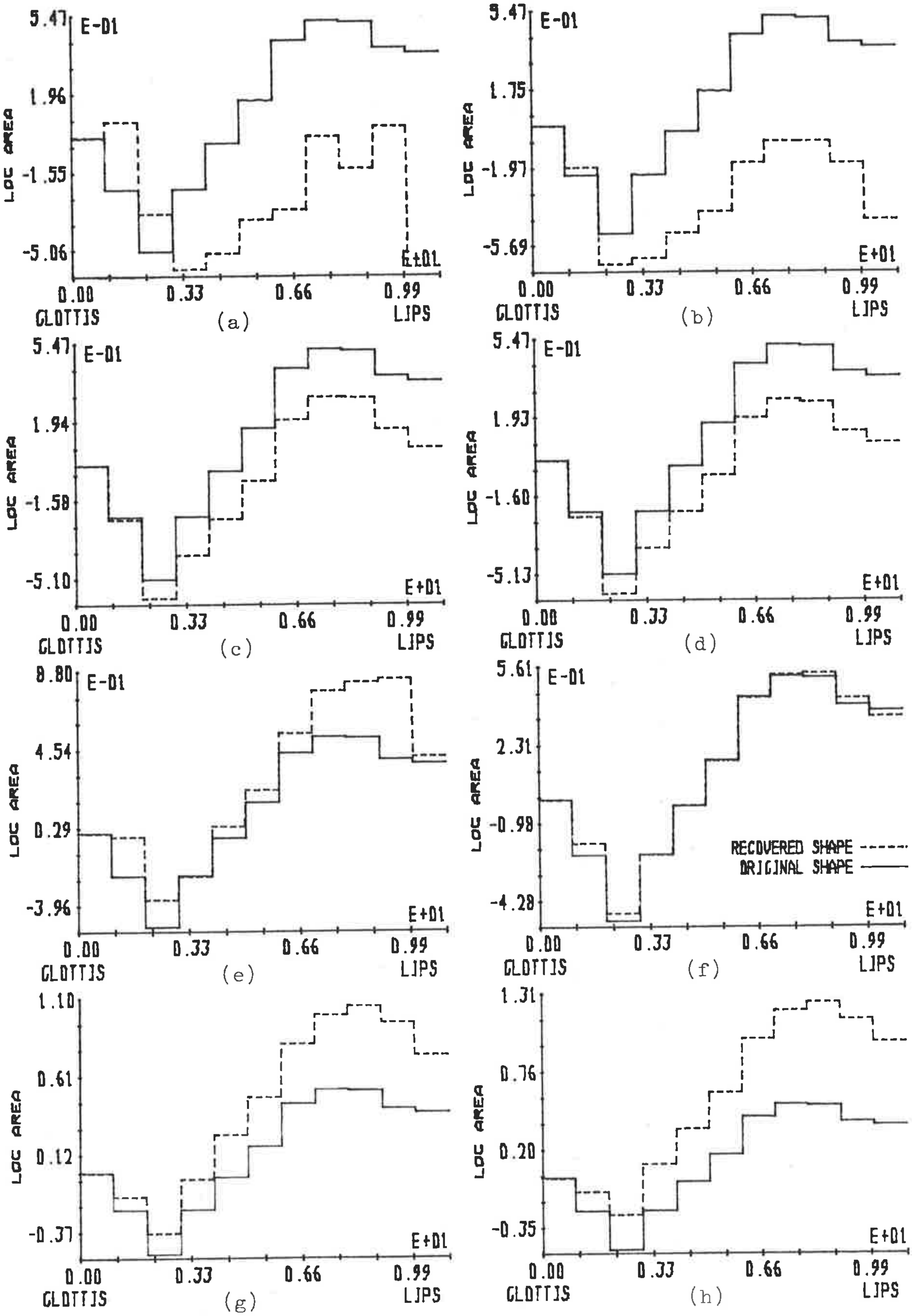


FIGURE 4.16: Comparison of recovered and original acoustic tube shapes after a +12dB per octave pre-emphasis of the vowel /a/ excited by the fant glottal pulse model with parameters (K, T_p/T) equal to (a) (0.67,0.4), (b) (0.67,0.6), (c) (1,0.4), (d) (1,0.6), (e) (3,0.4), (f) (3,0.6) (g) (9,0.6) and (h) (9,0.4).

spectral slope of which is the exact opposite to that of the glottal pulses used in Tables 4.4, 4.5 and 4.6, and the results are presented as area distances in Table 4.7.

Comparison of the area distances in Table 4.7 with those of Tables 4.2 and 4.3, i.e. for no pre-emphasis, shows a large reduction in area distances, and a comparison of Table 4.7 with Tables 4.5 and 4.6, i.e. for a +12 dB per octave pre-emphasis, shows a significant reduction in area distances for most cases. Since smaller area distances generally imply improved acoustic tube shape recovery, applying an opposite sign spectral slope pre-emphasis to that of the glottal pulse has an excellent potential for accurate acoustic tube shape recovery.

Figure 4.17 presents the recovered and original acoustic tube shapes for the pre-emphasis conditions used to generate Table 4.7 and the glottal pulse models and parameters used in Figure 4.14 and 4.15. A comparison of Figure 4.17 with Figures 4.14 and 4.15 shows that improved acoustic tube shape recovery is achieved by using the opposite sign spectral slope pre-emphasis to that of the glottal pulse excitation.

An adaptive filter which attempts to supply a variable spectral slope correction, to adapt to the changing spectral slope of the speech waveform, is that of GRAY and MARKEL [1974] and MAKHOUL and VISWANATHAN [1974]. This adaptive filter was designed to remove glottal pulse excitation and lip radiation effects, and has the form $(1-\gamma z^{-1})$. The parameter γ is chosen as

VOWEL	AREA DISTANCES						
	ROSENBERG POLYNOMIAL MODEL (T_p/T, T_N/T)			FANT PULSE MODEL (K, T_p/T)			
	(.33,.09)	(.4,.16)	(.6,.19)	(3,.4)	(9,.4)	(1,.6)	(9,.6)
/a/	0.18	1.31	0.79	1.08	0.74	0.87	0.74
/e/	0.24	1.15	0.48	2.54	0.53	0.68	0.89
/i/	0.95	1.16	1.48	4.56	0.67	0.91	0.95
/o/	0.17	1.27	1.24	1.50	0.67	0.95	0.91
/u/	0.58	1.98	4.20	4.11	0.60	0.50	1.23

TABLE 4.7: Area distances for a parcor analysis of synthetic speech for five vowels generated with glottal pulses from the Rosenberg polynomial pulse model and the fant glottal pulse model when an exact opposite spectral slope pre-emphasis to that of the glottal pulse is applied.

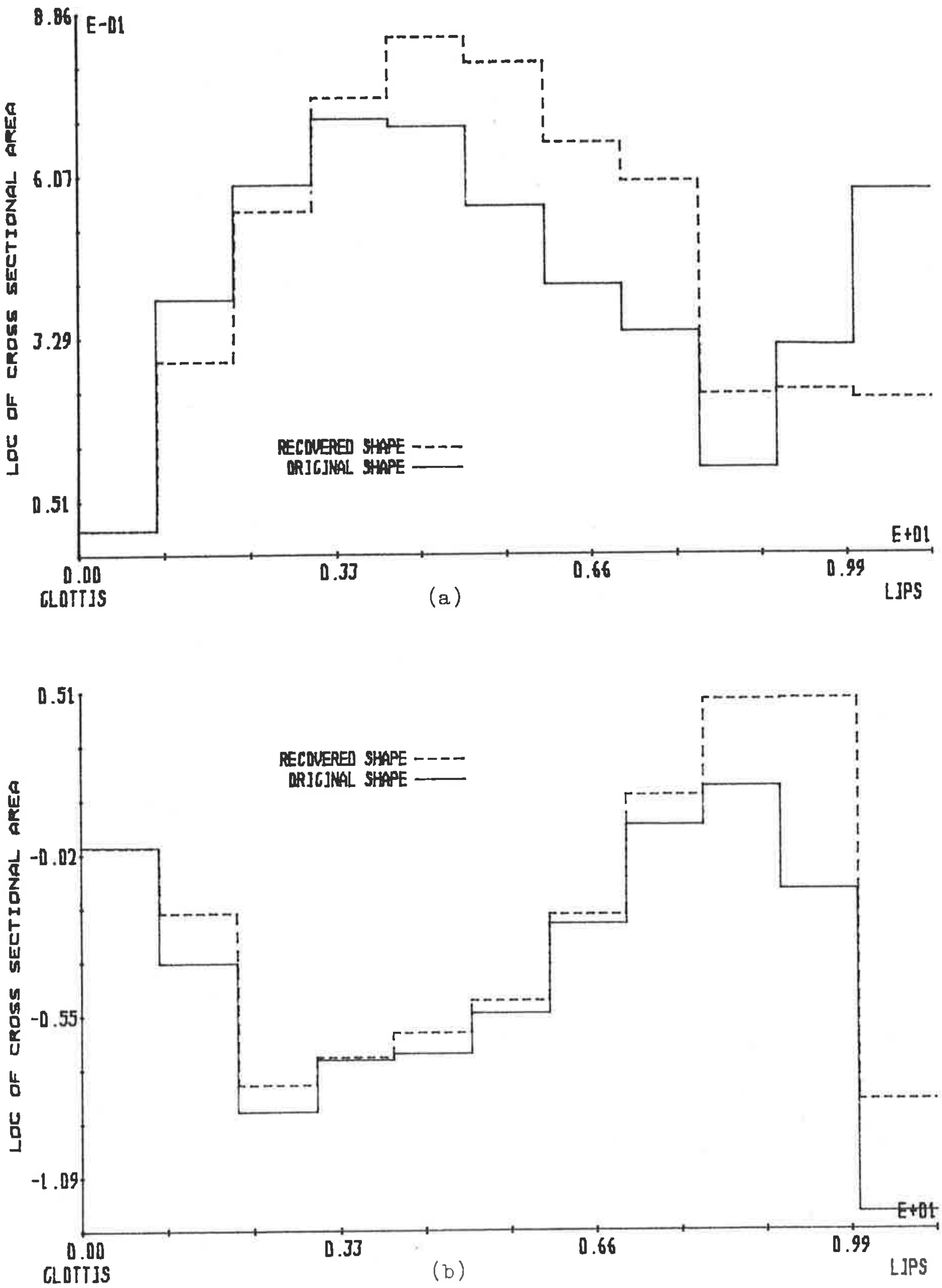
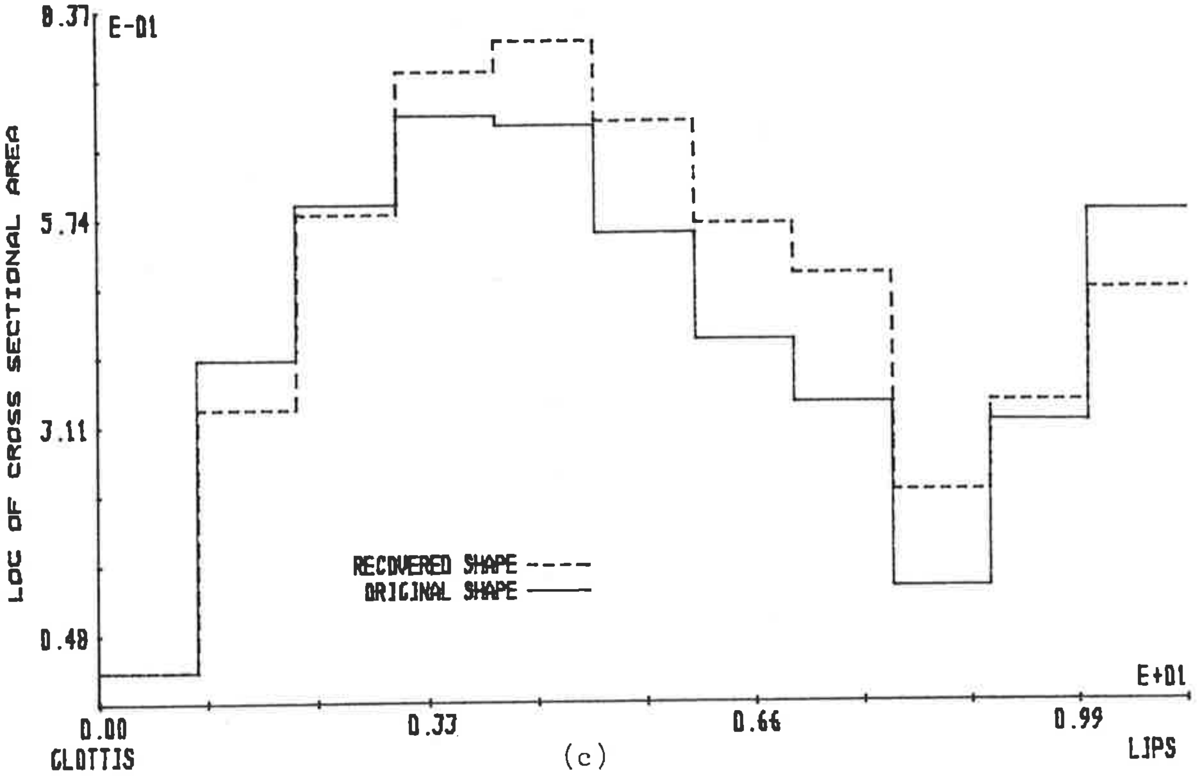
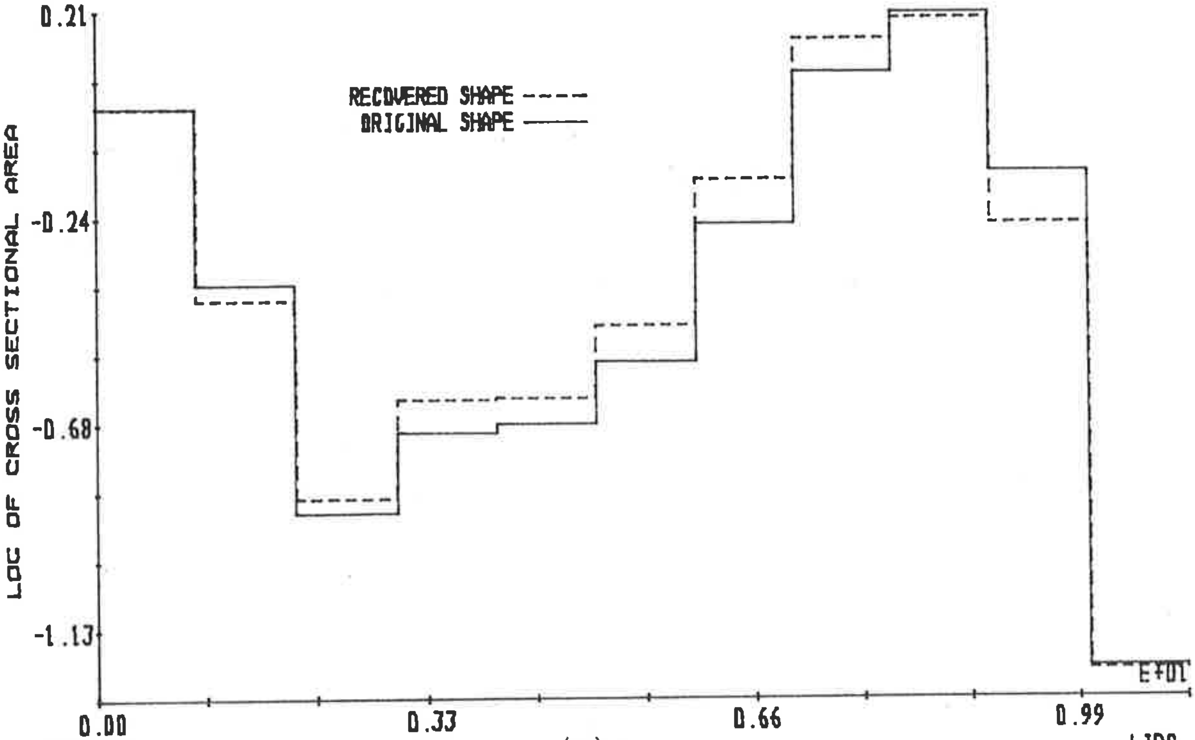


FIGURE 4.17: Comparison of recovered and original acoustic tube shapes for an equal but opposite sign spectral slope pre-emphasis to that of the glottal excitation for (a) the vowel /e/ and Rosenberg polynomial glottal pulse model with parameters (T_p/T, T_N/T)=(0.4,0.16) and (b) the vowel /u/ and Rosenberg polynomial glottal pulse model with parameters (T_p/T, T_N/T) = (0.6,0.19).



(c)



(d)

FIGURE 4.17: Comparison of recovered and original acoustic tube shapes for an equal but opposite sign spectral slope pre-emphasis to that of the glottal pulse excitation for (c) the vowel /e/ and fant glottal pulse model with parameters (K, T_p/T)=(9,0.6) and (d) the vowel /u/ and fant glottal pulse model with parameters (K, T_p/T)=(1,0.6).

$$\gamma = R(1)/R(0) \qquad (4.7)$$

to permit adaption to unvoiced and voiced sounds. For voiced sounds, γ is near unity, and so an approximate +6 dB per octave spectral pre-emphasis is applied to the speech waveform. The excitation for unvoiced sounds is noise-like and, hence, a nearly flat spectrum which does not require any spectral correction. For unvoiced sounds, $R(1)/R(0)$ is generally near zero, and so no pre-emphasis is applied by the adaptive filter.

This chapter only considers glottal pulse excitation effects on recovered acoustic tube shapes, and so the part of the adaptive filter of GRAY et al [1974] which corrects for lip radiation is removed. A -6 dB per octave spectral correction is generally applied to remove lip radiation effects and, in general, a +12 dB per octave pre-emphasis is applied to remove glottal pulse excitation effects. Therefore, an appropriate adaptive filter which has the adaptive properties of the GRAY et al [1974] adaptive filter and can be used to pre-emphasize for glottal pulse excitation only has the form $(1-\gamma z^{-1})^2$, where the parameter γ is defined by Equation 4.7. This filter is referred to as the unvoiced/voiced adaptive pre-emphasis filter, and applies approximately zero dB per octave pre-emphasis to unvoiced sounds and approximately +12 dB per octave pre-emphasis to voiced sounds.

Evaluation of the unvoiced/voiced adaptive pre-emphasis filter is performed with the same synthetic waveforms as used to generate Tables 4.4 to 4.7. The area distances for the unvoiced/voiced adaptive pre-emphasis followed by a Parcor analysis of the synthetic waveforms are presented in Table 4.8. The values of γ chosen by the unvoiced/voiced adaptive pre-emphasis filter for each

VOWEL	AREA DISTANCES							
	REAL GLOTTAL PULSE	ROSENBERG POLYNOMIAL MODEL (T_p/T, T_N/T)			FANT PULSE MODEL (K, T_p/T)			
		(.33,.09)	(.4,.16)	(.6,.19)	(.67,.4)	(1,.4)	(3,.4)	(9,.4)
/a/	1.46	6.06	6.43	1.55	2.33	1.17	2.31	8.46
/e/	1.11	1.11	1.89	0.64	1.57	0.97	2.80	1.13
/i/	2.72	10.88	6.82	2.74	4.65	4.82	3.21	7.63
/o/	1.92	10.15	8.68	2.49	2.48	1.11	4.11	13.73
/u/	4.50	21.06	23.05	9.00	4.49	1.44	15.53	23.22

TABLE 4.8: Area distances for a parcor analysis of synthetic speech generated with glottal pulses from the Rosenberg polynomial and fant glottal pulse models and the real glottal pulse waveform when the unvoiced/voiced adaptive pre-emphasis filter is used.

VOWEL	VALUE OF γ							
	REAL GLOTTAL PULSE	ROSENBERG POLYNOMIAL MODEL (T_p/T, T_N/T)			FANT PULSE MODEL (K, T_p/T)			
		(.33,.09)	(.4,.16)	(.6,.19)	(.67,.4)	(1,.4)	(3,.4)	(9,.4)
/a/	.99787	.98029	.99280	.99696	.99861	.99786	.99445	.98408
/e/	.99666	.98495	.99103	.99567	.99822	.99708	.99354	.98925
/i/	.99812	.99416	.99596	.99813	.99857	.99824	.99765	.99630
/o/	.99837	.98586	.99519	.99715	.99876	.99825	.99551	.98870
/u/	.99901	.99586	.99791	.99887	.99924	.99903	.99809	.99721

TABLE 4.9: Values of γ used by the unvoiced/voiced adaptive pre-emphasis filter to generate data presented in Table 4.8.

of the synthetic waveforms analysed are presented in Table 4.9. Only a small variation of γ, i.e. from 0.999 to 0.98, is observed in Table 4.9, which translates to a variation in pre-emphasis spectral slope of less than 0.1 dB per octave. Hence, only a small change in area distances should be observed for the unvoiced/ voiced adaptive pre-emphasis filter in comparison with the constant +12 dB per octave pre-emphasis. Comparison of Tables 4.8 with Tables 4.4, 4.5 and 4.6 verifies this conclusion.

To completely evaluate the unvoiced/voiced adaptive pre-emphasis filter requires the range of γ from zero to unity to be considered. Figure 4.18 presents the area distances for no pre-emphasis and an unvoiced/voiced adaptive pre-emphasis followed by a Parcor analysis of synthetic speech for the vowel |e| versus γ, which varies from zero to unity. The excitation used to generate the synthetic speech is derived from ROSENBERG's polynomial glottal pulse model (other glottal pulse models and real glottal pulses produce similar results to those presented in Figure 4.18).

A significant decrease in area distance is achieved by the unvoiced/voiced adaptive pre-emphasis in comparison with no pre-emphasis, from the results presented in Figure 4.18, when γ is close to unity and less than 0.7. When γ is between 0.7 and 0.92, i.e. values which occur in reality for voiced speech since glottal pulse spectral slopes as low as -8 dB per octave have been reported [MONSEN and ENGEBRETSON 1977, CARR and TRILL 1964, MILLER 1959], the area distances for the unvoiced/voiced adaptive pre-emphasis are greater than for no pre-emphasis. Therefore, the unvoiced/ voiced adaptive pre-emphasis filter only provides a reduction in

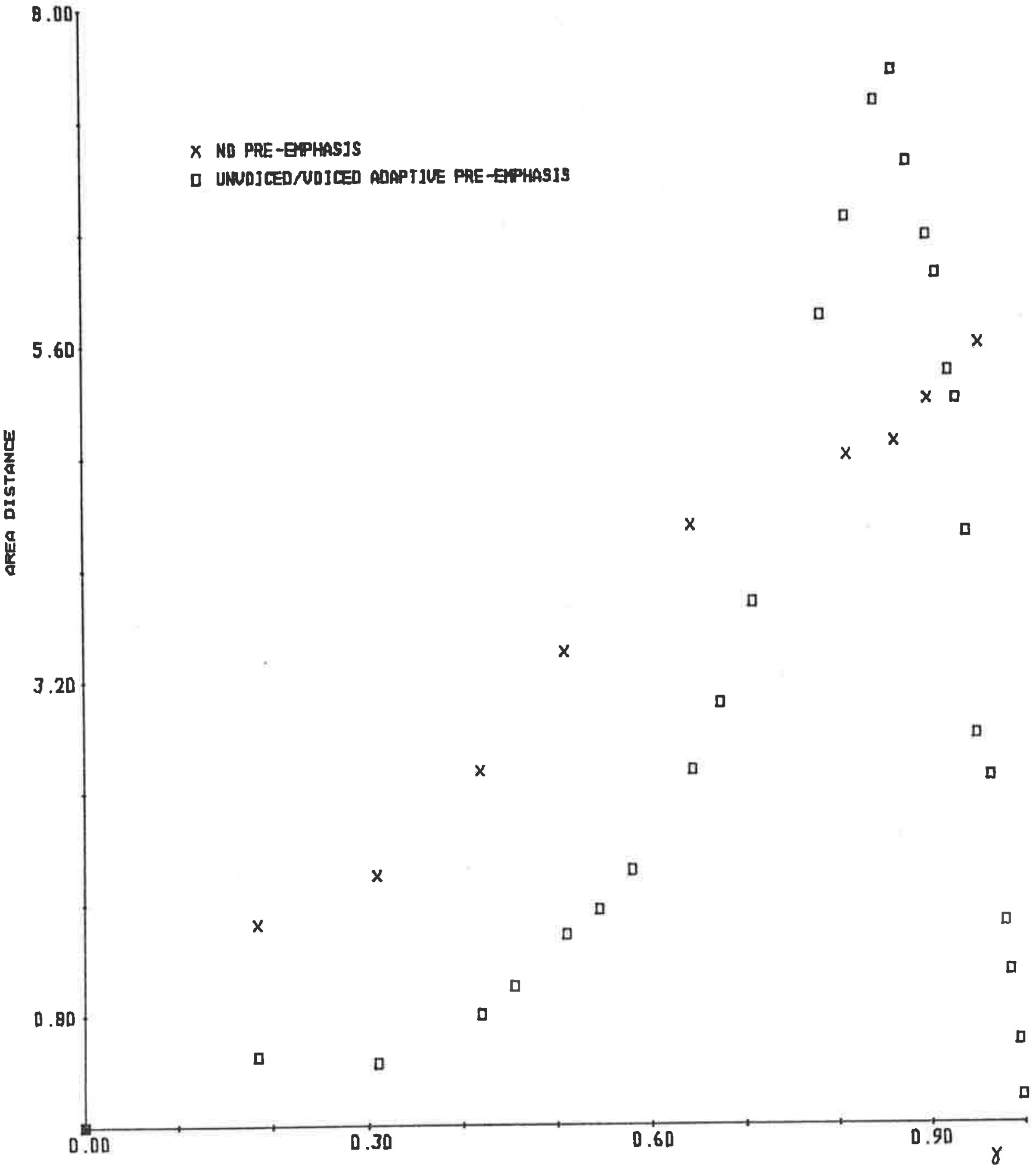


FIGURE 4.18: Area distances between recovered and original acoustic tube shapes for no pre-emphasis and an unvoiced/voiced adaptive pre-emphasis of synthetic speech for the vowel /e/.

area distances and, hence, improved acoustic tube shape recovery, when γ is close to and much less than unity (or the glottal spectral slope is close to or much less than -12 dB per octave).

An undesirable feature of the unvoiced/voiced adaptive pre-emphasis filter is its dependence on the waveform sampling frequency. This is due to the value of $R(1)$ changing and, therefore, the value of γ changing, with waveform sampling frequency. An evaluation of the unvoiced/voiced adaptive pre-emphasis filter for waveform sampling frequencies of 6 to 9 kilohertz is presented in Table 4.10. The area distances of Table 4.10 result from synthetic speech of the vowel $|a|$ generated with real and synthetic glottal pulses, which is pre-emphasized by the unvoiced/voiced adaptive pre-emphasis filter and then analysed by a Parcor analysis. The values of γ used by the unvoiced/voiced adaptive pre-emphasis filter to produce the results in Table 4.10(a) are presented in Table 4.10(b). If Tables 4.8 and 4.9 are considered with Table 4.10, then the range of waveform sampling frequencies presented is from 6 to 10 kilohertz.

Comparison of Tables 4.8, 4.9 and 4.10 shows that the range of γ for any one sampling frequency increases as the sampling frequency decreases, with the smallest range of γ occurring for a sampling frequency of 10 kilohertz, i.e. from 0.98 to 0.999, and the largest range of γ occurring for a sampling frequency of 6 kilohertz, i.e. from 0.93 to 0.996. For any particular glottal pulse, the value of γ decreases for decreasing sampling frequency. These results show that the unvoiced/voiced adaptive pre-emphasis filter applies a different pre-emphasis to the same waveform if the sampling frequency changes.

SAMPLING FREQUENCY (kHz)	AREA DISTANCES							
	REAL GLOTTAL PULSE	ROSENBERG POLYNOMIAL MODEL (T_p/T, T_N/T)			FANT PULSE MODEL (K, T_p/T)			
		(.33,.09)	(.4,.16)	(.6,.19)	(.67,.4)	(1,.4)	(3,.4)	(9,.4)
6	1.39	3.93	3.38	0.64	1.59	0.92	2.24	4.51
7	1.33	2.97	5.33	3.58	1.86	0.98	2.10	4.73
8	1.37	9.22	1.33	1.04	1.92	1.02	2.54	6.16
9	1.63	7.55	3.22	4.81	1.98	1.10	2.35	10.32

(a)

SAMPLING FREQUENCY (kHz)	VALUE OF γ							
	REAL GLOTTAL PULSE	ROSENBERG POLYNOMIAL MODEL (T_p/T, T_N/T)			FANT PULSE MODEL (K, T_p/T)			
		(.33,.09)	(.4,.16)	(.6,.19)	(.67,.4)	(1,.4)	(3,.4)	(9,.4)
6	.99403	.92974	.97514	.98989	.99582	.99351	.98287	.95834
7	.99485	.95914	.97997	.99089	.99702	.99532	.98774	.96882
8	.99665	.95781	.99028	.99489	.99777	.99653	.99125	.97501
9	.99748	.97079	.99153	.99516	.99824	.99730	.99306	.98187

(b)

TABLE 4.10: (a) Area distances and (b) value of γ for an unvoiced/voiced adaptive pre-emphasis and a parcor analysis of synthetic speech for the vowel /a/ generated with a real glottal pulse and glottal pulses from the Rosenberg polynomial and fant glottal pulse models and sampled at frequencies of 9,8,7 and 6 kilohertz.

A variation in the area distances presented in Table 4.10 is expected, since the discrete representation of the vocal tract shape by a set of acoustic tubes changes as the sampling frequency changes. Allowing for this variation, a general trend of decreasing area distances with decreasing sampling frequency is observed in Table 4.10(a). This result, coupled with the variation in γ for changes in sampling frequency discussed above, leads to the conclusion that the effectiveness of the unvoiced/voiced adaptive pre-emphasis filter to provide improved acoustic tube shape recovery is dependent on the waveform sampling frequency.

Figure 4.19 presents the recovered and original acoustic tube shapes for two columns of Table 4.10(a), one for a ROSENBERG polynomial glottal pulse excitation, and one for a FANT glottal pulse excitation. The recovered acoustic tube shapes in Figure 4.19 show the errors due to incorrectly accounting for the glottal pulse excitation outweighing the errors due to changes in sampling frequency. Hence, sampling frequency effects are only important when good acoustic tube shape recovery is achieved by the unvoiced/voiced adaptive pre-emphasis.

The evaluations presented in this section have shown that the unvoiced/voiced adaptive pre-emphasis filter provides an improvement in acoustic tube shape recovery wherever the glottal pulse excitation spectral slope is close or much less than -12 dB per octave. Choosing the parameter of the unvoiced/voiced adaptive filter, i.e. γ, equal to $R(1)/R(0)$ permitted an adaption to unvoiced and voiced sounds, but does not provide the correct amount of spectral correction to remove glottal pulse excitation effects from the recovered acoustic tube shape. The effectiveness of the

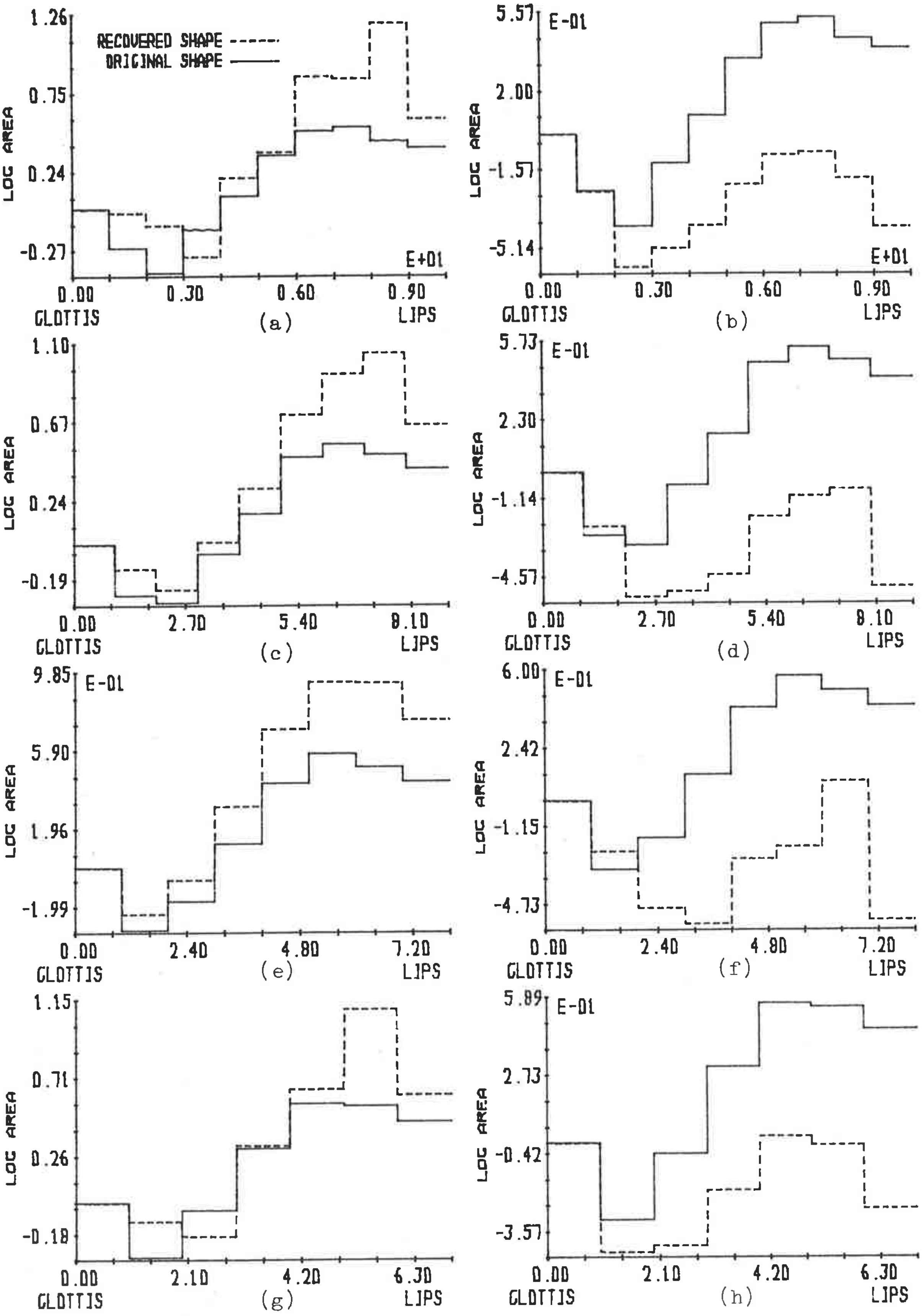


FIGURE 4.19: Comparison of recovered and original acoustic tube shapes for an unvoiced/voiced adaptive pre-emphasis of the vowel /a/ excited by the Rosenberg polynomial glottal pulse model with parameters $(T_p/T,\ T_N/T)=(.33,0.9)$ for (a) to (d), and the fant glottal pulse model with parameters $(K,\ T_p/T)=(0.67,0.4)$ for (e) to (g). Sampling frequencies of 9,8,7 and 6 kilohertz are used in (a) and (e), (b) and (f), (c) and (g) and, (d) and (h) respectively.

unvoiced/voiced adaptive filter was shown to have a small dependence on the sampling frequency. In general, the unvoiced/voiced adaptive pre-emphasis filter does not provide accurate acoustic tube shape recovery.

Another pre-emphasis filter which attempts to adapt to changing spectral slopes of speech waveforms is that of NAKAJIMA et al [1974]. The NAKAJIMA adaptive pre-emphasis filter is designed to remove both the glottal pulse excitation and radiation characteristics from the speech waveform to produce the impulse response of the vocal tract. This is achieved by second and third order critical damping digital filters in cascade. Determination of the parameters of the second and third order filters requires the calculation of covariances and the solution of polynomials of order three and five. Hence, this filter is computationally more complex to implement than the adaptive filter of GRAY and MARKEL [1974] and MAKHOUL and VISWANATHAN [1974].

The second order critical damping digital filter which corrects for the overall frequency slope of the speech spectrum is defined by

$$\hat{s}_i = \varepsilon_1 s_{i-1} - \frac{\varepsilon_1^2}{4} s_{i-2} \qquad (4.8)$$

where s_i is the discrete speech signal at time i, $\hat{s}_i$ is the estimate of s_i and ε_1 is the coefficient of the filter. Assuming steady state conditions and a window length of p speech samples, then NAKAJIMA shows that the method of least squares gives ε_1 as the real root of the third order polynomial

$$C_{22}\varepsilon_1^3 - 6C_{21}\varepsilon_1^2 + (4C_{02} + 8C_{11})\varepsilon_1 - 8C_{01} = 0 \qquad (4.9)$$

with the requirement that $|\varepsilon_1| < 2$. The C_{jk} in Equation 4.9 are covariances and are defined as

$$C_{jk} = \sum_{\ell=0}^{p-1} s_{i-j-\ell} \cdot s_{i-k-\ell} \qquad (4.10)$$

with i such that only speech samples within a window are considered.

The third order critical damping digital filter which corrects for mid-frequency effects of the speech spectrum is defined by

$$\hat{s}_i = \varepsilon_2 s_{i-1} - \frac{\varepsilon_2^2}{3} s_{i-2} + \frac{\varepsilon_2^3}{27} s_{i-3} \qquad (4.11)$$

with ε_2 being the coefficient of the filter. If steady state conditions occur and the window length is p speech samples, then NAKAJIMA showed the method of least squares gives ε_2 as the real root of the fifth order polynomial

$$C_{33}\varepsilon_2^5 - 15C_{23}\varepsilon_2^4 + (54C_{22} + 36C_{13})^2\varepsilon_2^3 - (243C_{12} + 27C_{03})\varepsilon_2^2 + (243C_{11} + 162C_{02})\varepsilon_2 - 243C_{01} = 0 \qquad (4.12)$$

with the requirement that $|\varepsilon_2| < 3$. The covariances of Equation 4.12 are defined by Equation 4.10.

Pre-emphasis of the speech waveform by the NAKAJIMA adaptive filter is performed by first applying the second order filter and then the third order filter. Thus, the covariances required to define the second order filter through Equation 4.8 are evaluated from the speech waveform, while those required to define the third order filter through Equation 4.11 are evaluated from the output waveform of the second order filter. Both Equations 4.8 and 4.11 require solution by a root finding algorithm, and a simple NEWTON-RAPHSON algorithm has been found by NAKAJIMA et al [1974] to be sufficient. There does not exist any guarantee that the roots of Equations 4.8 and 4.11 satisfy the requirements $|\varepsilon_1| < 2$ and $|\varepsilon_2| < 3$, but experimental investigations by NAKAJIMA et al [1974] have shown acceptable solutions exist in the majority of cases.

The evaluation of the NAKAJIMA adaptive pre-emphasis filter is performed with the same procedure and synthetic waveforms used for the evaluation of the unvoiced/voiced adaptive pre-emphasis filter, i.e. as used to generate Table 4.8, and Table 4.11 contains the results of this evaluation. Comparison of Tables 4.8 and 4.11 shows that the NAKAJIMA adaptive pre-emphasis filter is vastly inferior to the unvoiced/voiced adaptive filter. Observation of the results for a constant +12 dB per octave pre-emphasis, i.e. the results of Tables 4.4, 4.5 and 4.6, also reveals that the NAKAJIMA adaptive filter is inferior for all the synthetic waveforms considered.

The area distances presented in Table 4.11 for the NAKAJIMA adaptive pre-emphasis filter are found to be comparable to the area distances obtained for no pre-emphasis of the synthetic speech. In fact, some cases provide a worse (i.e. larger) area distance than the no pre-emphasis results. Hence, it can be concluded that the NAKAJIMA adaptive filter does not provide a general improvement in acoustic tube shape recovery, and is inferior to other conventional pre-emphasis procedures.

VOWEL	AREA DISTANCES							
	REAL GLOTTAL PULSE	ROSENBERG POLYNOMIAL MODEL (T_p/T, T_N/T)			FANT PULSE MODEL (K, T_p/T)			
		(.33,.09)	(.4,.16)	(.6,.19)	(.67,.4)	(1,.4)	(3,.4)	(9,.4)
/a/	19.22	16.20	13.35	16.91	36.70	31.06	18.80	14.65
/e/	3.07	2.77	2.61	2.63	5.36	3.21	2.73	2.56
/i/	14.36	32.69	37.00	30.27	36.89	18.53	22.01	43.79
/o/	19.53	23.56	19.79	27.99	55.44	41.70	34.19	23.73
/u/	106.10	47.42	54.62	124.90	54.02	195.80	134.50	63.96

TABLE 4.11: Area distances for a parcor analysis of synthetic speech for five vowels generated with glottal pulses from the Rosenberg polynomial and fant flottal pulse models with a pre-emphasis by the NAKAJIMA adaptive pre-emphasis filter.

Evaluation results presented in this section have shown that a fixed +12 dB per octave pre-emphasis of the speech waveform produced a reduction in area distances whenever the glottal spectral slope is near -12 dB per octave. However, in general, the accuracy of acoustic tube shape recovery was poor, and in many cases the recovered acoustic tube shape has little resemblance to the original acoustic tube shape. It was shown that the poor performance of a constant dB per octave pre-emphasis is due to its inability to adapt to the variations in glottal pulse spectral shape that occur in reality. If a spectral slope pre-emphasis opposite in sign to that of the glottal pulse is applied, then it was shown that a potential exists for accurate acoustic tube shape recovery.

Two pre-emphasis filters which adapt to a change in glottal pulse spectral slope were evaluated. The unvoiced/voiced adaptive pre-emphasis filter, derived from the adaptive filter of GRAY and MARKEL [1974] and MAKHOUL and VISWANATHAN [1974], showed an improvement in acoustic tube shape recovery in comparison with a constant +12 dB per octave pre-emphasis. However, accurate acoustic tube shape recovery is not, in general, achieved with this filter. The other pre-emphasis filter evaluated was that of NAKAJIMA et al [1974] and its performance was found to be poor and inferior to the other available pre-emphasis methods.

Techniques which perform analysis during the closed glottis interval were investigated and found to provide accurate acoustic tube shape recovery. However, if any error is made in estimating the closed glottis interval, then large errors in acoustic tube shape were shown to occur. These closed glottis interval analyses are not applicable to all voiced sounds, as in many cases a closed glottis interval does not exist or it is too small to perform an analysis. Therefore, closed glottis interval analysis is not applicable to the wide range of speech sounds that may occur.

In conclusion, this section has shown that accurate acoustic tube shape recovery cannot be achieved with available pre-emphasis and other techniques from waveforms generated with ideal acoustic tubes and excited by glottal pulse type waveforms. Therefore, there exists a need for further investigations if accurate vocal tract shapes are to be recovered directly from the speech waveform.

4.4 IMPROVED ADAPTIVE PRE-EMPHASIS

The investigations presented so far in this chapter have shown that poor acoustic tube shape recovery occurs if glottal pulse excitation is not taken into account. The previous section showed that the available methods which have been designed to remove glottal pulse excitation effects from the recovered acoustic tube shape perform poorly, and do not provide accurate acoustic tube shape recovery. This section designs a new adaptive pre-emphasis filter to overcome the inadequacies of presently available pre-emphasis filters and provide improved acoustic tube shape recovery.

The form of the new adaptive pre-emphasis filter is designed to use the advantages, but overcome the disadvantages, of presently available pre-emphasis techniques. Once the form of the new adaptive pre-emphasis is defined, then the parameters of that filter are determined empirically from a large number of glottal pulse waveforms, both syntehtic and real. Defining the parameters empirically permits the new adaptive pre-emphasis filter to be tailored to remove glottal pulse excitation effects from the recovered acoustic tube shape. Therefore, in a loose sense, the new adaptive pre-emphasis attempts to provide an optimal reduction in area distances for the new adaptive pre-emphasis followed by an autocorrelation linear predictive analysis of acoustic waveforms from acoustic tubes excited by glottal pulse waveforms.

4.4.1 A NEW ADAPTIVE PRE-EMPHASIS FILTER

The requirements of the new adaptive pre-emphasis filter are determined from the investigations presented in the previous sectons, i.e. the effects of glottal pulse excitation on the recovered acoustic tube shape and the effectiveness of existing methods for removing these glottal pulse excitation effects. Thus, the new adaptive pre-emphasis filter includes the successful features of existing pre-emphasis methods, while attempting to overcome their inadequacies, as identified in the previous section. In some cases, conflicting requirements arise, and a compromise solution is necessary.

Improved acoustic tube shape recovery was shown to occur when the spectral slope of the pre-emphasis is opposite in sign to that of the glottal pulse excitation. The unvoiced/voiced adaptive pre-emphasis filter was shown to satisfy this situation whenever the glottal pulse spectral slope is near zero or -12 dB per octave. Adequate spectral slope correction is not provided at other glottal pulse spectral slopes, because the unvoiced/voiced adaptive pre-emphasis filter does not attempt to determine the spectral slope of the glottal pulse excitation. Since the unvoiced/voiced adaptive pre-emphasis filter only contains two zeros, it cannot provide more than +12 dB per octave spectral correction, which is insufficient to cover the range of observed glottal pulse excitation spectral slopes.

The new adaptive pre-emphasis filter should therefore be able to approximate the glottal pulse excitation spectral slope from the acoustic synthetic, or real, speech waveform and apply an equal but opposite sign spectral slope correction. The new adaptive pre-emphasis filter should be able to apply a spectral slope correction of at least +17 dB per octave to cover the range of observed glottal pulse excitation spectral slopes [MONSEN and ENGEBRETSON 1977, CARR and TRILL 1964, MILLER 1959]. To provide a spectral slope of of +17 dB per octave, the new adaptive pre-emphasis filter must contain at least three zeros but, as the number of zeros increases, the range of frequencies over which a constant dB per octave spectral slope is available decreases. Therefore, the presence of a third zero when less than +12 dB per octave is required jeopardises the new filter's effectiveness. The compromise solution is to only introduce the third zero when greater than +12 dB per octave spectral slope correction is required.

The adaptive pre-emphasis filter of GRAY and MARKEL [1974] and MAKHOUL and VISWANATHAN [1974] was originally designed to adapt to the quasi-periodic glottal pulse excitation of voiced sounds and the noise-like excitation of unvoiced sounds. This is a desirable feature of a pre-emphasis filter, since it eliminates the need for separate unvoiced/voiced detection and decision systems to apply or not apply pre-emphasis. The basis for the unvoiced/voiced decision is made by the ratio $R(1)/R(0)$, which is considered by GRAY and MARKEL [1974] to provide a reliable decision. Therefore, use of the ratio $R(1)/R(0)$ to provide an unvoiced/voiced decision in the new adaptive pre-emphasis filter should be considered.

To completely remove the glottal pulse frequency spectrum from the speech waveform frequency spectrum requires consideration of global trends, i.e. overall spectral slope, as well as the local variations in the spectrum. A complex pre-emphasis filter with relatively large numbers of parameters is required to correct for local variations in the glottal pulse spectrum. Experiments have shown that the major errors in acoustic tube shape recovery occur because of the global, i.e. overall spectral slope, rather than the local variations in the glottal pulse spectrum. Hence, a simple pre-emphasis filter with a small number of parameters is desirable and cost effective for simplicity of implementation and improvements gained in acoustic tube shape recovery.

The requirements presented above lead to the new adaptive pre-emphasis filter having the form

$$(1 - \alpha z^{-1})^2 \tag{4.13}$$

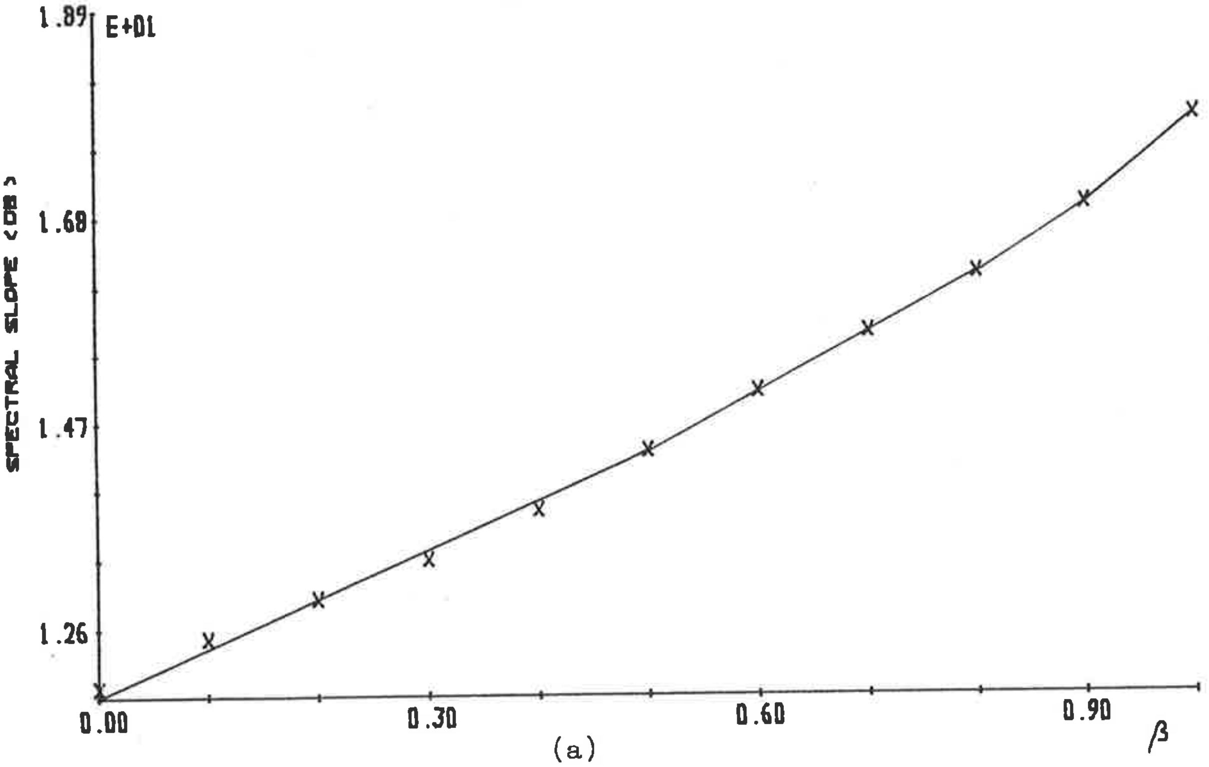
when less than +12 dB per octave spectral correction is required, and

$$(1 - z^{-1})^2 (1 - \beta z^{-1}) \tag{4.14}$$

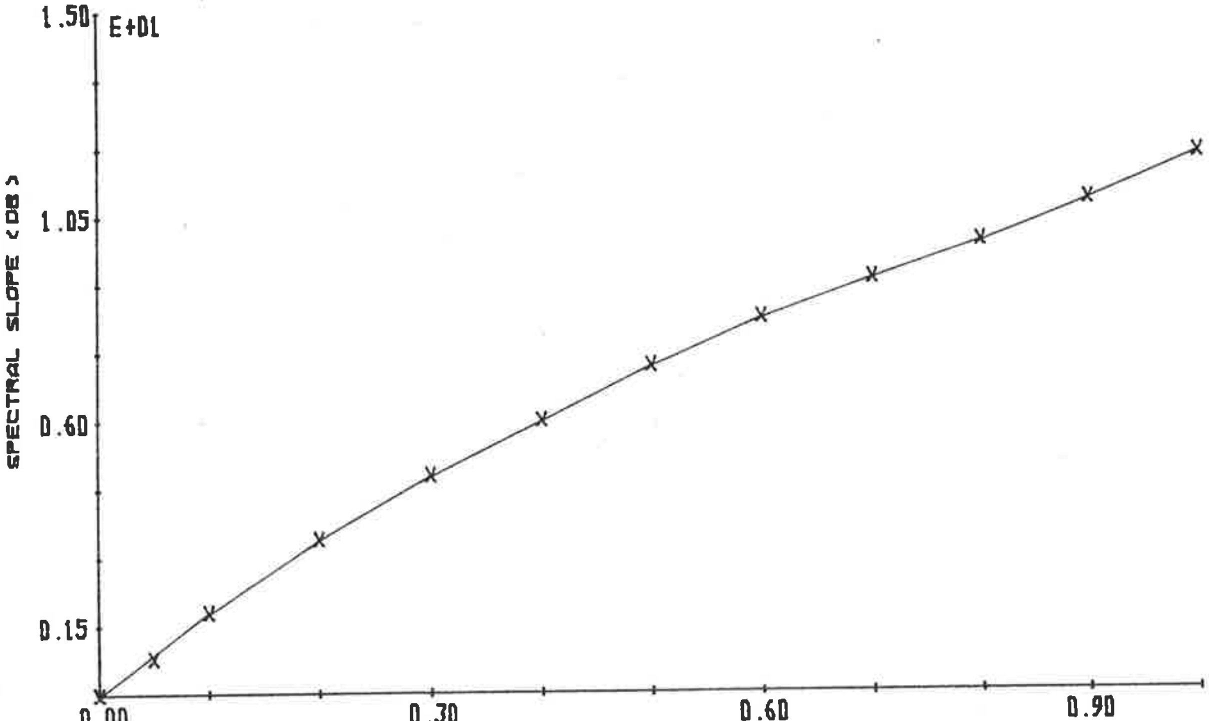
when greater than +12 dB per octave spectral correction is required. The new adaptive pre-emphasis filter can provide a spectral slope correction between zero and +18 dB per octave by the appropriate selection of the parameters α or β. The new adaptive pre-emphasis filter is referred to as the two/three adaptive pre-emphasis filter, because of its two or three zero structure. Figure 4.20 presents a plot of the spectral slope provided by the two/three adaptive pre-emphasis filter as a function of the parameters α and β.

The spectral slope of a number of glottal pulse waveforms and the value of $R(1)/R(0)$ of synthetic speech for a number of vowels generated using these glottal pulses as excitations is presented in Figure 4.21. Allowing for perturbations due to acoustic tube shape changes, the value of $R(1)/R(0)$ is seen to provide some indication of the glottal pulse excitation spectral slope, from Figure 4.21. Therefore, using $R(1)/R(0)$ to determine the values of α and β parameters may not only provide an unvoiced/voiced detection but also an indication of the glottal pulse spectral slope. The ease with which $R(1)/R(0)$ can be calculated and the above discussions lead to $R(1)/R(0)$ being the quantity which is extracted from the waveform being analysed and used to define the two/three adaptive pre-emphasis filter for that situation.

The area distance measure has been designed to indicate the similarity of two acoustic tube shapes and, if the area distance between a recovered and original acoustic tube shape is small, then good acoustic tube shape recovery has occurred. Since the goal of the two/three adaptive pre-emphasis filter is to provide accurate acoustic tube shape recovery, then the parameters α and/or β should be chosen such that the smallest area distance is obtained. A procedure for determining this smallest or minimum area distance is presented in Figure 4.22.



(a)



(b)

FIGURE 4.20: Spectral slope provided by the two/three adaptive pre-emphasis filter for values of its parameters α and β .

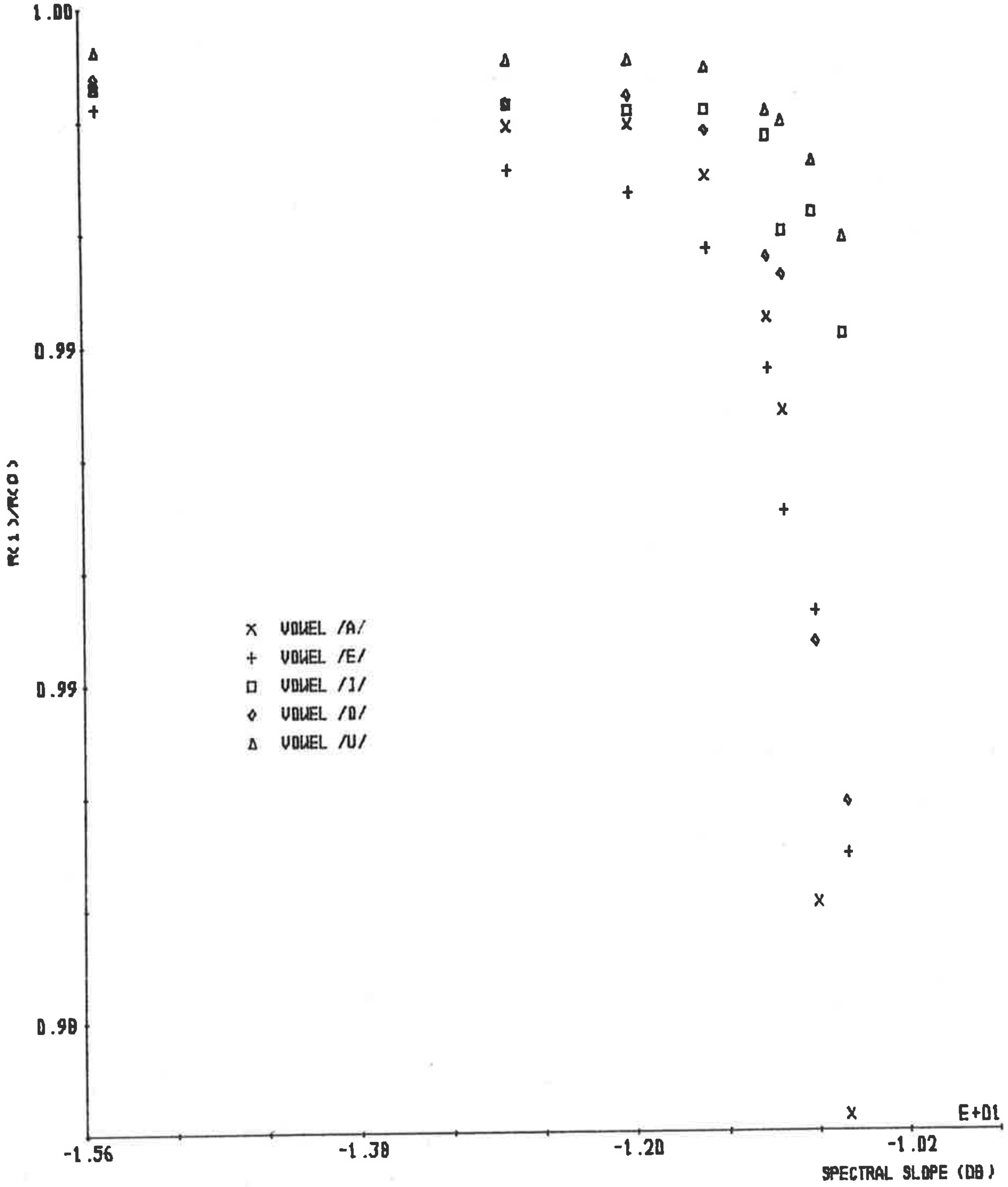


FIGURE 4.21: Spectral slope and value of R(1)/R(0) for synthetic speech of the vowels /a/, /e/, /i/, /o/ and /u/ with excitation by glottal pulse waveforms.

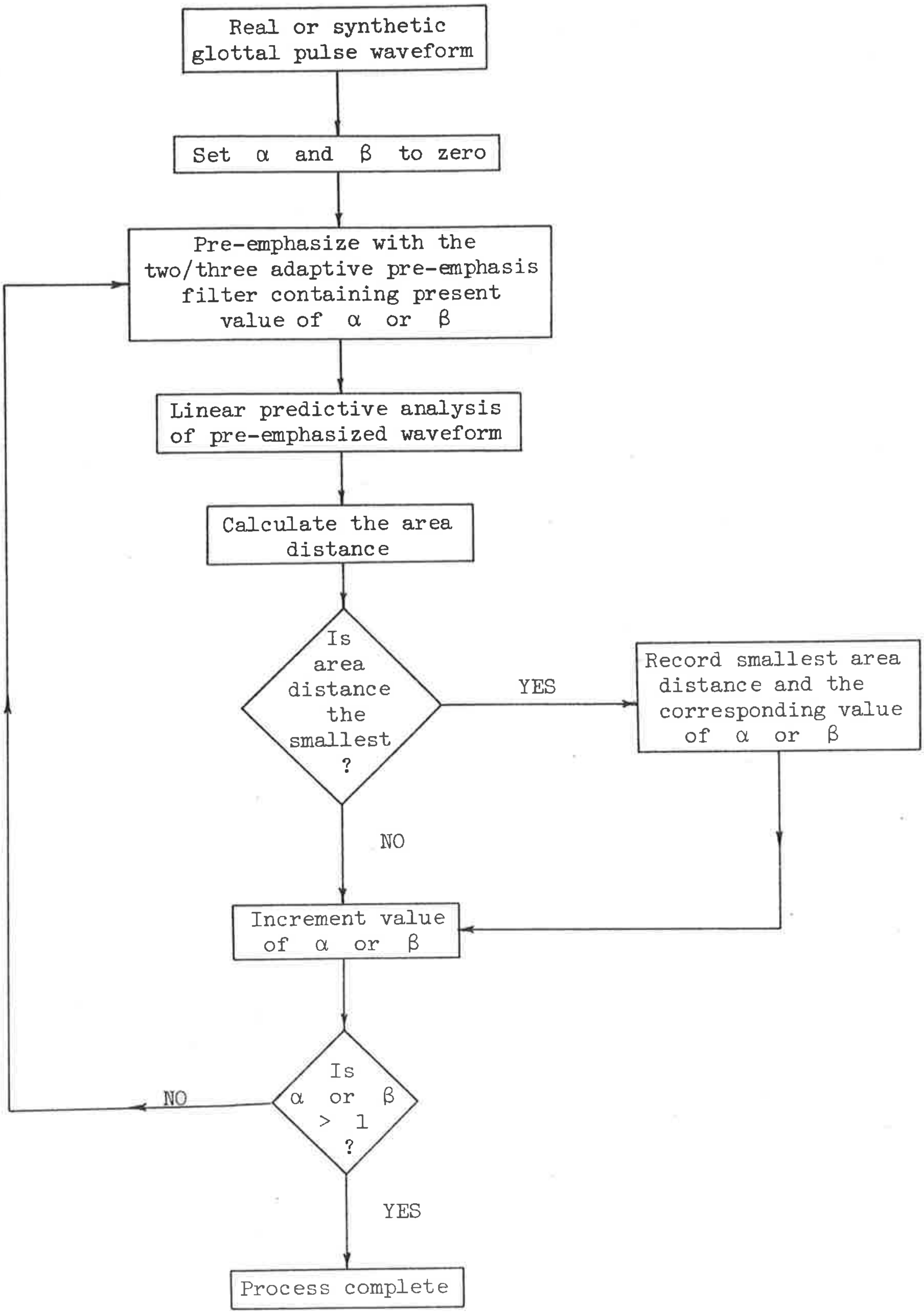

Real or synthetic glottal pulse waveform
Set α and β to zero
Pre-emphasize with the two/three adaptive pre-emphasis filter containing present value of α or β
Linear predictive analysis of pre-emphasized waveform
Calculate the area distance
Is area distance the smallest ?
YES
Record smallest area distance and the corresponding value of α or β
NO
Increment value of α or β
Is α or β > 1 ?
NO
YES
Process complete

FIGURE 4.22: Procedure for determining the value of α or β so that a pre-emphasis by the two/three adaptive pre-emphasis filter produces a minimum area distance.

By determining the values of $R(1)/R(0)$ for a large number of different glottal pulse waveforms and the corresponding α or β parameters of the two/three adaptive pre-emphasis filter that provide a minimum area distance, a relationship between $R(1)/R(0)$ and α and β can be defined. This relationship tailors the two/three adaptive filter so that the best possible reduction in glottal pulse excitation effects on the recovered acoustic tube shape is achieved by the two/three adaptive pre-emphasis filter. The necessary experiments to determine the relationships between $R(1)/R(0)$ and the parameters α and β are performed in Section 4.4.2.

Figure 4.23 presents the glottal pulse spectral slopes of synthetic speech for the vowel |a| and the values of $R(1)/R(0)$ of this synthetic speech for sampling frequencies of 6, 7, 8, 9 and 10 kilohertz. The results presented in Figure 4.23 show the necessity to have different relationships between $R(1)/R(0)$ and the parameters α and β for different sampling frequencies, even though it appears that the difference between the relationships will be small. The definition of these relationships for different sampling frequencies is presented in Section 4.4.3.

The complete form of the two/three adaptive pre-emphasis filter and the relationships between $R(1)/R(0)$ and the parameters α and β are presented in Section 4.4.4.

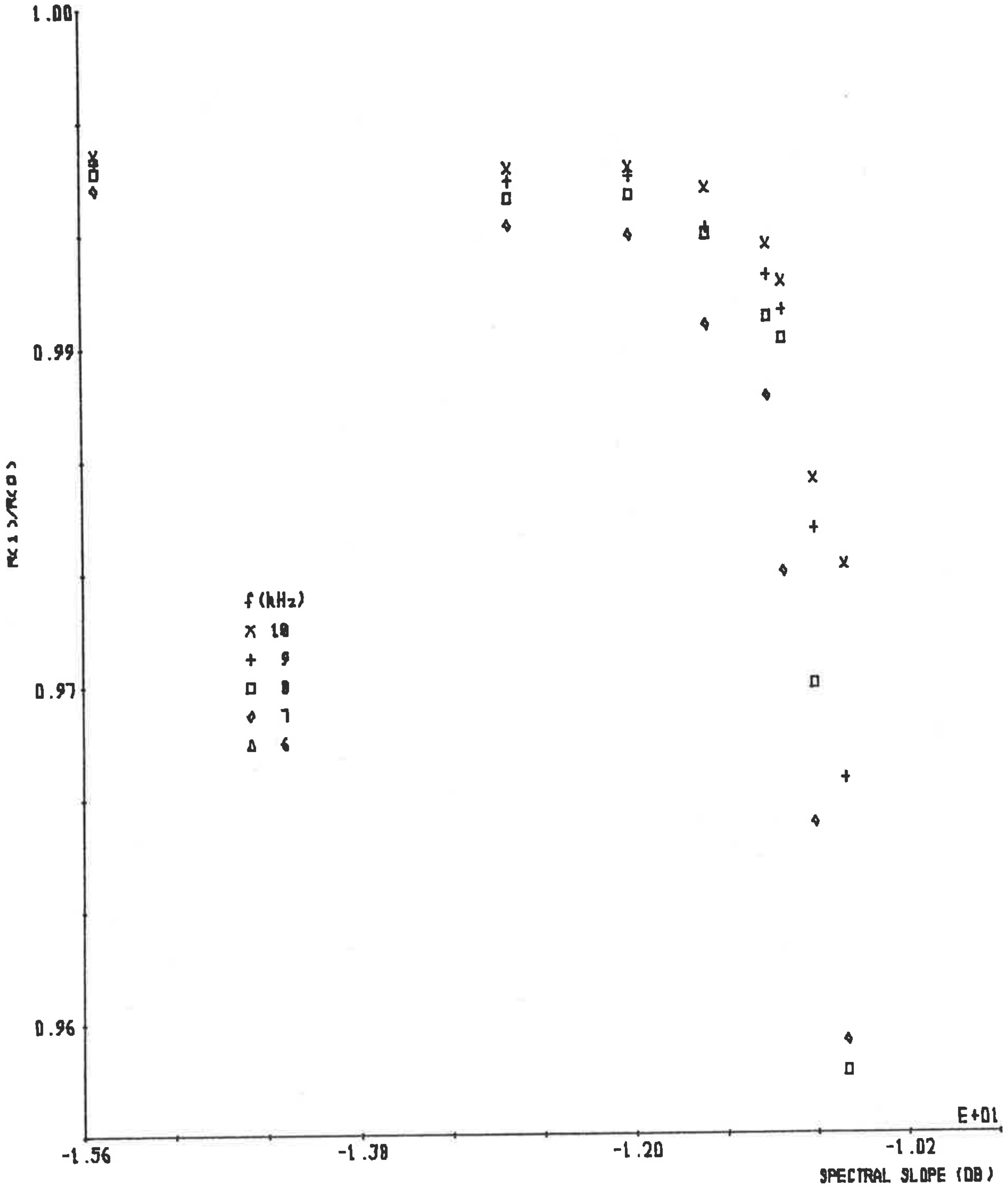


FIGURE 4.23: Spectral slope and value of R(1)/R(0) for synthetic speech of the vowel /a/, for sampling frequencies of 10,9,8,7 and 6 kilohertz.

4.4.2 DETERMINATION OF PARAMETERS

In the previous section, the form of the two/three adaptive pre-emphasis filter was defined and the manner in which the filter parameters are to be determined. It was concluded that the value of $R(1)/R(0)$ of the waveform being analysed should be used to define the parameters of the two/three adaptive pre-emphasis filter for a particular situation. The necessary experiments to define a relationship between $R(1)/R(0)$ and the two/three adaptive pre-emphasis filter parameters α and β are performed in this section for a waveform sampling frequency of 10 kilohertz.

The criterion used to determine the value of α or β for a particular waveform is to produce a minimum area distance after a two/three adaptive pre-emphasis followed by a Parcor linear predictive analysis. The values of α and β which produce this minimum area distance are denoted as α' and β', respectively. Figure 4.22 details the method used to determine either α' or β' for a particular glottal pulse waveform. Repeating this method for a large number of different glottal pulses permits the required relationship between $R(1)/R(0)$ and the parameters α' and β' to be determined. Determination of the relationships in this manner permits the linear predictive process to tailor the two/three adaptive pre-emphasis filter so that improved acoustic tube shape recovery may be achieved.

For applications to speech waveforms, the value of $R(1)/R(0)$ of the glottal pulse excitation is not available, and so $R(1)/R(0)$ of the speech waveform must be used. A brief study of synthetic speech waveforms (see Figure 4.21) showed that the value of $R(1)/R(0)$ for a glottal pulse is slightly different from the value of $R(1)/R(0)$ evaluated from a synthetic speech waveform when that glottal pulse was used as the excitation. Therefore, the values of α' and β' may not necessarily provide the minimum area distances for a two/three adaptive pre-emphasis and a linear predictive analysis of synthetic speech when the glottal pulse used to determine α' or β' is used as the excitation. However, investigations show that this does not significantly affect the performance of the two/three adaptive pre-emphasis filter (see Chapter 5).

The accuracy of the relationships between $R(1)/R(0)$ and the two/three adaptive pre-emphasis filter parameters α and β depends on the number and range of different glottal pulse excitation waveforms considered. It is important for the large range of glottal pulse shapes and spectral slopes that occur for real speech to be considered. To achieve this, a number of glottal pulse models are used, with an emphasis on those satisfying subjective testing.

To ensure that the relationships between $R(1)/R(0)$ and the parameters α and β are suitable for application to real speech analysis, a number of glottal pulses derived from published glottal pulse waveforms measured during phonation are considered. These glottal pulses presented in published material are digitized by using a data tablet and a small minicomputer to oversee and store the digitized data. The separation of the waveform points which are digitized is determined by the sampling period or frequency required.

Although the glottal pulse waveforms used to define the relationships between α' and $R(1)/R(0)$ or β' and $R(1)/R(0)$ are derived from the same glottal pulse models, each glottal pulse waveform can only be used to determine one of these relationships. Therefore, the large number of glottal pulse waveforms used can be divided into two mutually exclusive groups, one for determining each relationship. This division of glottal pulse waveforms is performed here, and Section 4.4.2.1 determines the relationship between $R(1)/R(0)$ and α', while Section 4.4.2.2 determines the relationship between $R(1)/R(0)$ and β'.

4.4.2.1 Parameter α

Repeated use of the procedure defined in Figure 4.22 for a large number of different glottal pulses generates a set of data points $(R(1)/R(0), \alpha')$ from which a relationship between $R(1)/R(0)$ and α' can be determined. This relationship is defined by first choosing an appropriate function and then using a best fit algorithm to fit that function to the data points. The choice of an appropriate function must consider the requirements of simplicity, accuracy within important regions, and continuity of the function over the full ranges of $R(1)/R(0)$ and α'. The choice of the best fit algorithm depends on the function chosen to fit the data points and the criterion used to define the best fit, e.g. least squares.

A typical set of data points $(R(1)/R(0), \alpha')$ is presented in in Figure 4.24, and an observation of these data points suggests that a simple function may provide an accurate fit to all the data points. The concentration of data points in Figure 4.24 when $R(1)/R(0)$ is close to unity is typical of that found for voiced sounds in real speech. Therefore, it is important that the function

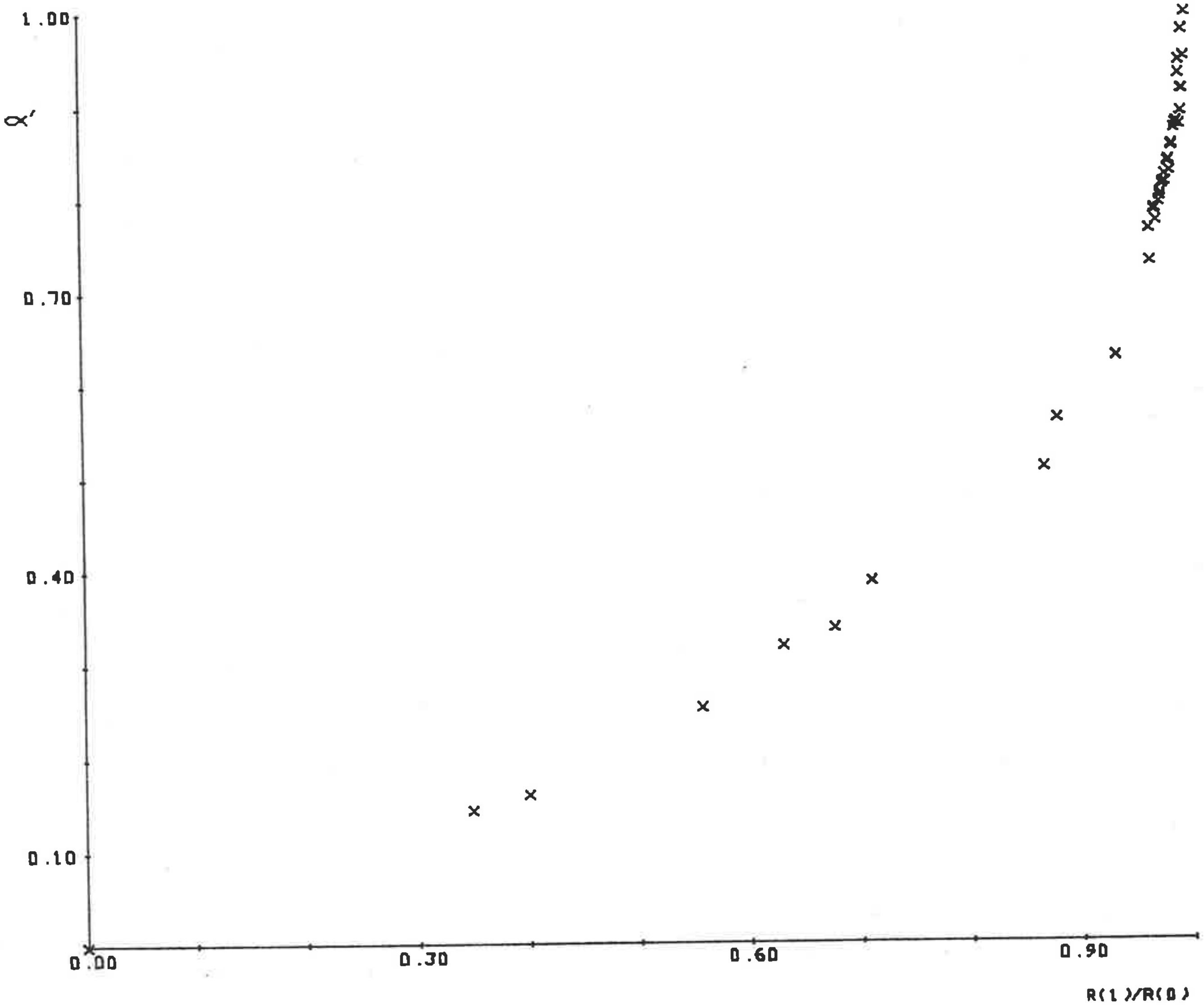


FIGURE 4.24: Typical set of data points (R(1)/R(0), α′).

chosen to fit the data points provides an accurate fit when $R(1)/R(0)$ is near unity. Unfortunately, the steep slope that occurs when $R(1)/R(0)$ is near unity is difficult to fit with a simple function without causing very poor fitting to the data points when $R(1)/R(0)$ is less than unity. In practical applications of the two/three adaptive pre-emphasis filter to real time situations, it is desirable to use only simple functions so that large numbers of complex computations are avoided.

One solution to fitting a function accurately to a steep slope is to determine the inverse relationship, i.e. $R(1)/R(0)$ in terms of α', from the data points $\left(\alpha', R(1)/R(0)\right)$. Hence, in the region where $R(1)/R(0)$ is near unity, the data points $\left(\alpha', R(1)/R(0)\right)$ have approximately zero slope, which permits simple functions to provide an accurate fit to the data points. In general, it is not possible to invert functions, and so the inverse relationship is used to define a lookup table. The major advantage of a lookup table is the very small number of elementary computations that are necessary to determine the value of α' for any particular situation. A full discussion of the implementation of the inverse function, i.e. $R(1)/R(0)$ in terms of α', as a lookup table is presented in Appendix D.

Investigations have shown that using a polynomial function fit to the data points $\left(\alpha', R(1)/R(0)\right)$ satisfies the necessary requirements of simplicity, accuracy and continuity over the full range of $R(1)/R(0)$. Hence, the relationship found in this section between α' and $R(1)/R(0)$ is a best fit polynomial with the form

$$\frac{R(1)}{R(0)} = \sum_{i=0}^{n} c_i (\alpha')^i \qquad (4.15)$$

where n is the order of the best fit polynomial. Investigations with various curve fitting algorithms have shown that a least squares curve fitting algorithm [DANIEL and WOOD 1980, LEONARD 1965] provides an accurate fit of the best fit polynomial to the sets of data points considered in this chapter. The order, n, of the best fit polynomial is determined as that which minimizes the deviation between the data points and the best fit polynomial. A weighting function is used when determining the best fit polynomial to ensure the polynomial is as close as possible to the end points of the data, i.e. (0,0) and (1,1). This has been found to be necessary especially near the data point (0,0) where only a few data points exist and, hence, the polynomial fit near (0,0) can be poor without a weighting function.

Generation of the sets of data points $(\alpha', R(1)/R(0))$ is firstly performed with synthetic glottal pulses generated from ROSENBERG's glottal pulse models [1971]. ROSENBERG performed subjective testing with six glottal pulse models, and found two to have almost zero preference scores and another to have consistently low scores. Hence, only the three remaining glottal pulse models, designated models B, C and E by ROSENBERG, which provide realistic glottal pulse shapes, are used. Model B is the polynomial glottal pulse model used in previous sections, and defined by Equations 4.1 and 4.2. Both models C and E are trigonometric pulses, with model C being defined by

$$u_0(t) = \frac{\lambda}{2}\left[1 - \cos\frac{t\pi}{T_p}\right] \qquad 0 \leqslant t \leqslant T_p \qquad (4.16)$$

and

$$u_0(t) = \lambda\cos\left(\frac{t-T_p}{T_N}\right)\frac{\pi}{2} \qquad T_p \leqslant t \leqslant T_p+T_N \qquad (4.17)$$

with model E being defined by

$$u_0(t) = \lambda\sin\frac{t}{T_p}\cdot\frac{\pi}{2} \qquad 0 \leqslant t \leqslant T_p \qquad (4.18)$$

and

$$u_0(t) = \lambda\cos\left(\frac{t-T_p}{T_N}\right)\frac{\pi}{2} \qquad T_p \leqslant t \leqslant T_p+T_N \qquad (4.19)$$

The parameters T_p and T_N are defined in Section 4.2 as the period during which the glottal pulse has a positive and negative slope, respectively. The scaling performed by the parameter λ does not affect the results of an analysis process, and so for simplicity λ is assumed to be unity.

The accuracy of the relationship between α' and $R(1)/R(0)$ is dependent on the number and range of different glottal pulse shapes used. A large number of different pulse model parameters are used to satisfy this requirement. In general, the range of parameters is restricted to the preferred range determined by the subjective testing of ROSENBERG [1971], to ensure results which are close to reality. However, to cover the full range of $R(1)/R(0)$ values some glottal pulse shapes have their parameters marginally outside this preferred range.

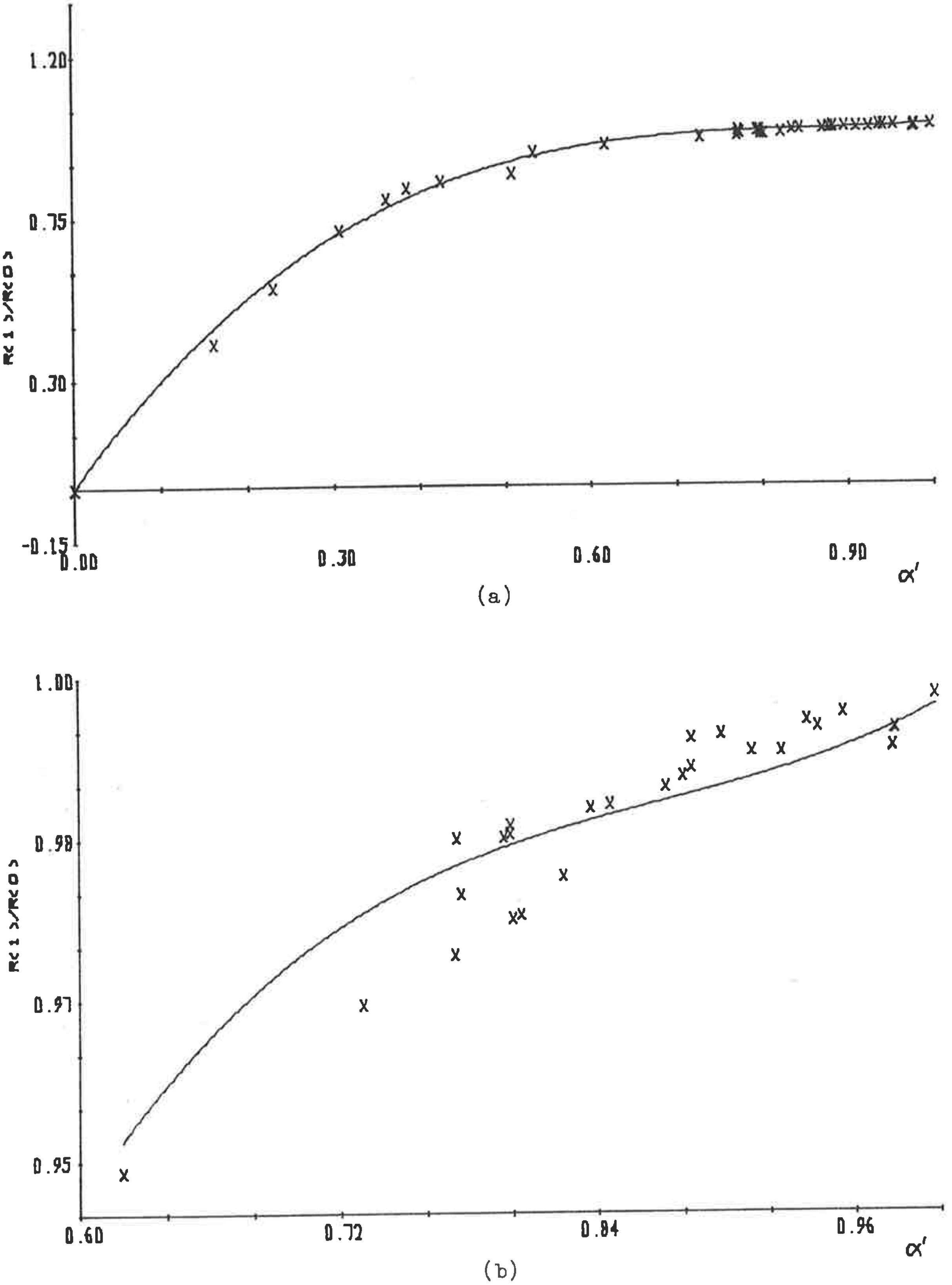
The glottal pulse model B of ROSENBERG [1971] is used to generate the first set of data points $(\alpha', R(1)/R(0))$. Each data point is obtained by choosing the parameters T_p/T and T_N/T of the glottal pulse model B, generating the glottal pulse waveform, and then using the procedure defined in Figure 4.22 to determine α'. Calculation of $R(0)$ and $R(1)$ directly from the glottal pulse waveform then defines the data point $(\alpha', R(1)/R(0))$ for that glottal pulse waveform. Figure 4.25 plots the set of data points obtained in this manner, for the ROSENBERG glottal pulse model B. The best fit polynomial found by the least squares curve fitting algorithm [LEONARD 1965] is also plotted in Figure 4.25, and defines a relationship between α' and $R(1)/R(0)$ as

$$\frac{R(1)}{R(0)} = 3.286(\alpha') - 3.709(\alpha')^2 + 1.425(\alpha')^3 \qquad (4.20)$$

Using the same procedure and glottal pulse model parameters, T_p/T and T_N/T, as for the glottal pulse model B above, ROSENBERG's glottal pulse model C generates the data points $(\alpha', R(1)/R(0))$ plotted in Figure 4.26. The best fit polynomial, which is also plotted in Figure 4.26, is determined by the least squares curve fitting algorithm [LEONARD 1965] and defines a relationship between α' and $R(1)/R(0)$ as

$$\frac{R(1)}{R(0)} = 3.305(\alpha') - 3.732(\alpha')^2 + 1.429(\alpha')^3 \qquad (4.21)$$

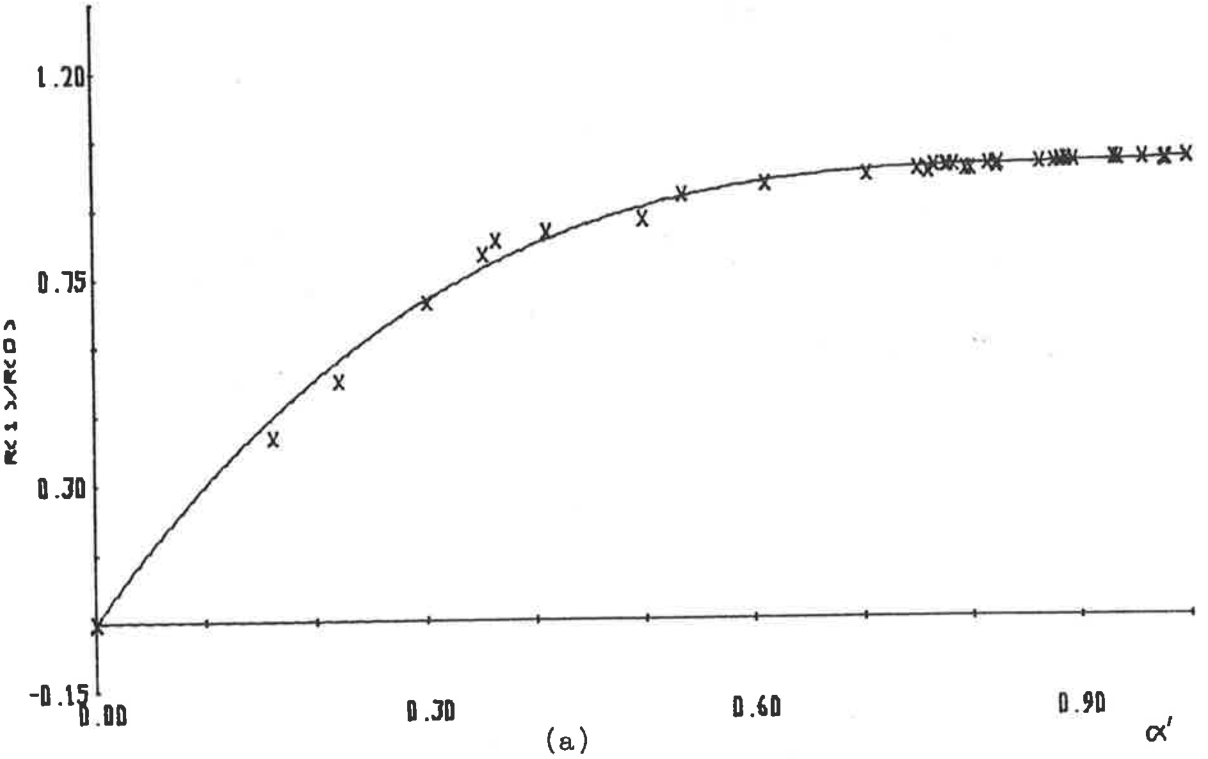
On comparing the best fit polynomials for glottal pulses generated from ROSENBERG's glottal pulse models B and C, i.e. Equations 4.20 and 4.21, a similarity in polynomial coefficients suggests that a single best fit polynomial may fit all the data points with little error.

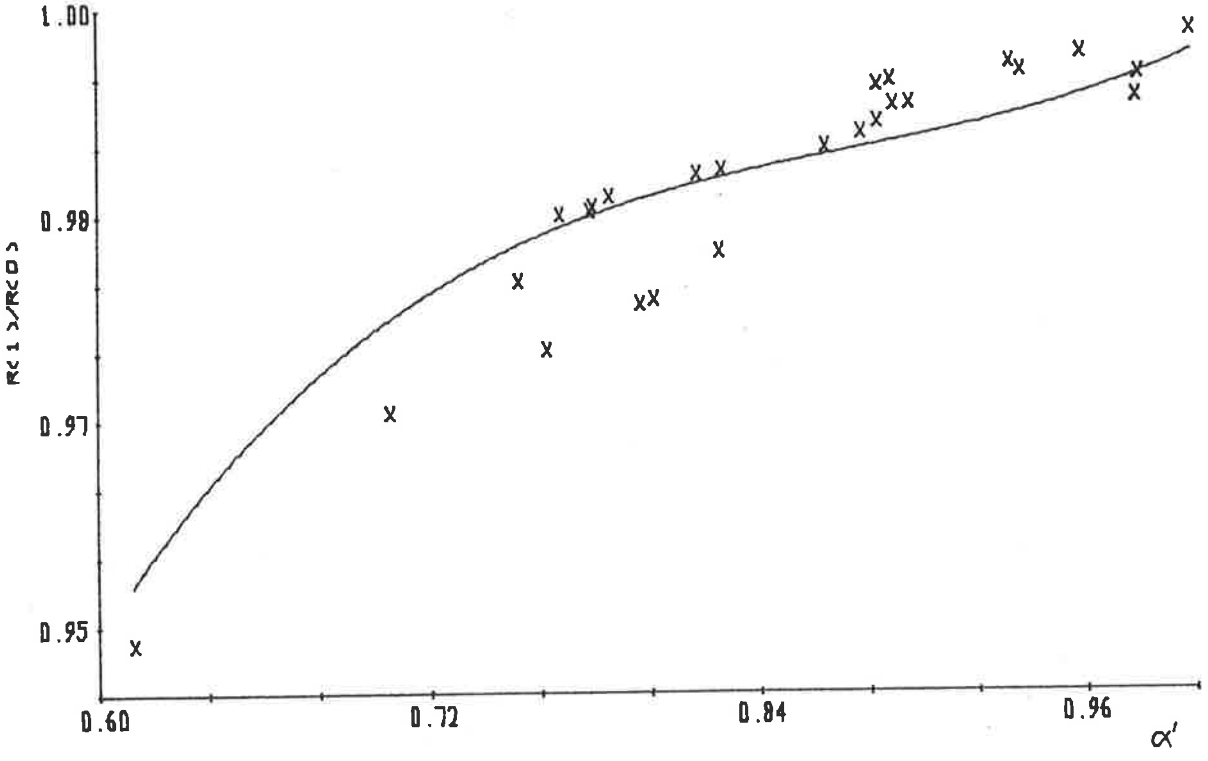


(a)

(b)

FIGURE 4.25: Data points (α´, R(1)/R(0)) and the best fit polynomial, for glottal pulses generated from the Rosenberg glottal pulse model B for α´ between (a) zero and unity and (b) 0.6 and unity.





(b)

FIGURE 4.26: Data points (α´, R(1)/R(0)) and the best fit polynomial, for glottal pulses generated from the Rosenberg glottal pulse model C for α´ between (a) zero and unity and (b) 0.6 and unity.

The data points $(\alpha', R(1)/R(0))$ plotted in Figure 4.27 are those generated by glottal pulses from the ROSENBERG glottal pulse model E using the same procedure and glottal pulse model parameters, T_p/T and T_N/T, as for models B and C above. Figure 4.27 also plots the best fit polynomial to this data which defines a relationship between α' and $R(1)/R(0)$ as

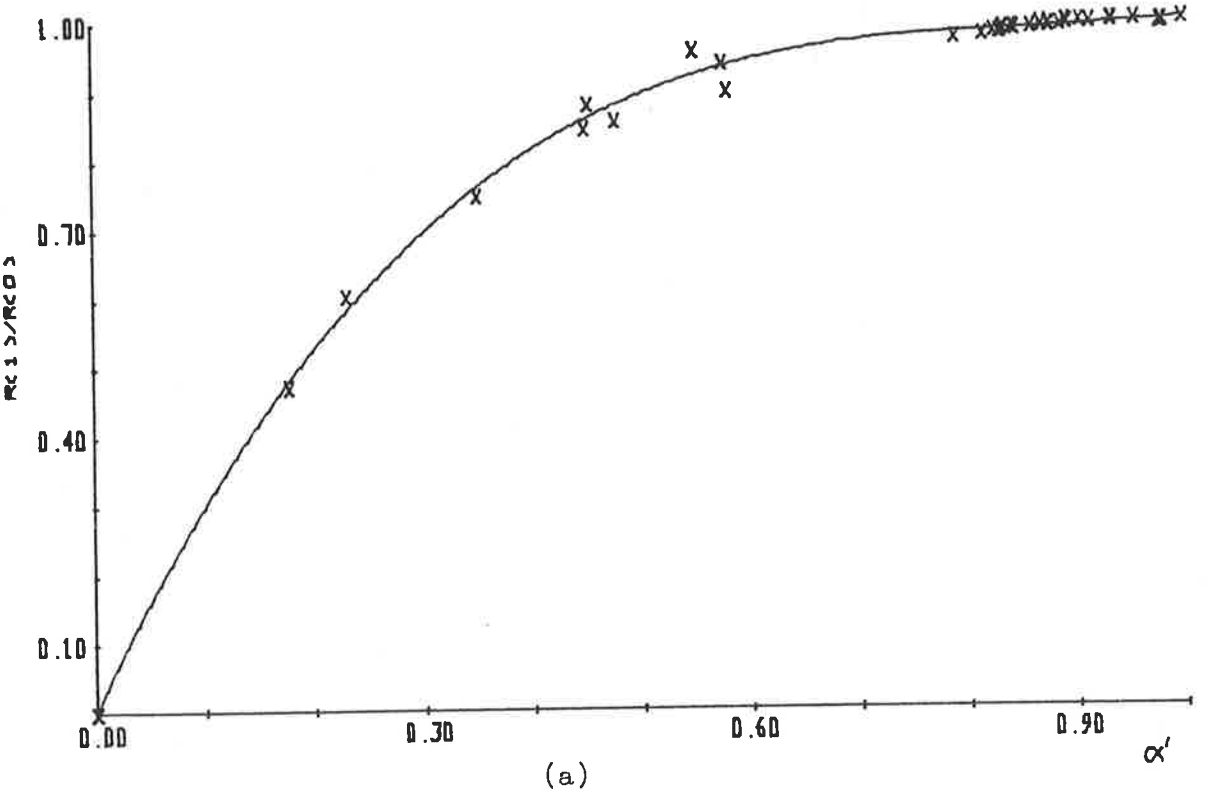
$$\frac{R(1)}{R(0)} = 3.292(\alpha') - 3.738(\alpha')^2 + 1.446(\alpha')^3 \qquad (4.22)$$

Comparison of the best fit polynomial for the glottal pulse model E, i.e. Equation 4.22, with those for ROSENBERG glottal pulse models B and C, i.e. Equations 4.20 and 4.21, respectively, shows a similarity in polynomial coefficients.

A combination of all the data points $(\alpha', R(1)/R(0))$ generated from the ROSENBERG glottal pulse models B, C and E and an application of the least squares curve fitting algorithm [LEONARD 1965] produces the best fit polynomials for the combined data as

$$\frac{R(1)}{R(0)} = 3.303(\alpha') - 3.741(\alpha')^2 + 1.441(\alpha')^3 \qquad (4.23)$$

A plot of this best fit polynomial (i.e. Equation 4.23) and the data points generated from the ROSENBERG glottal pulse models is presented in Figure 4.28. Different plotting symbols are used in Figure 4.28 to indicate which glottal pulse model was used to generate the data points $(\alpha', R(1)/R(0))$. The even mixture of plotting symbols and the even spread about the best fit polynomial indicates a consistency of data points $(\alpha', R(1)/R(0))$ generated from different glottal pulse models. This consistency of data points for a wide range of glottal pulse waveforms is necessary for a single



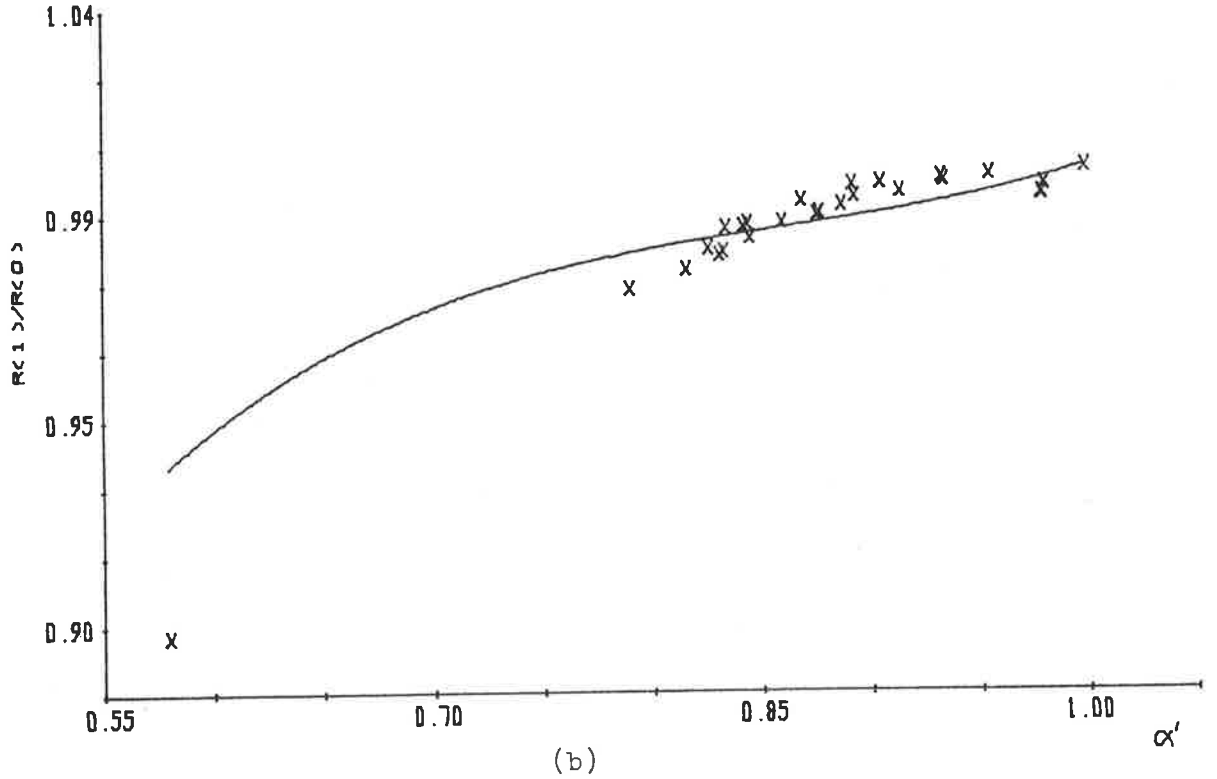
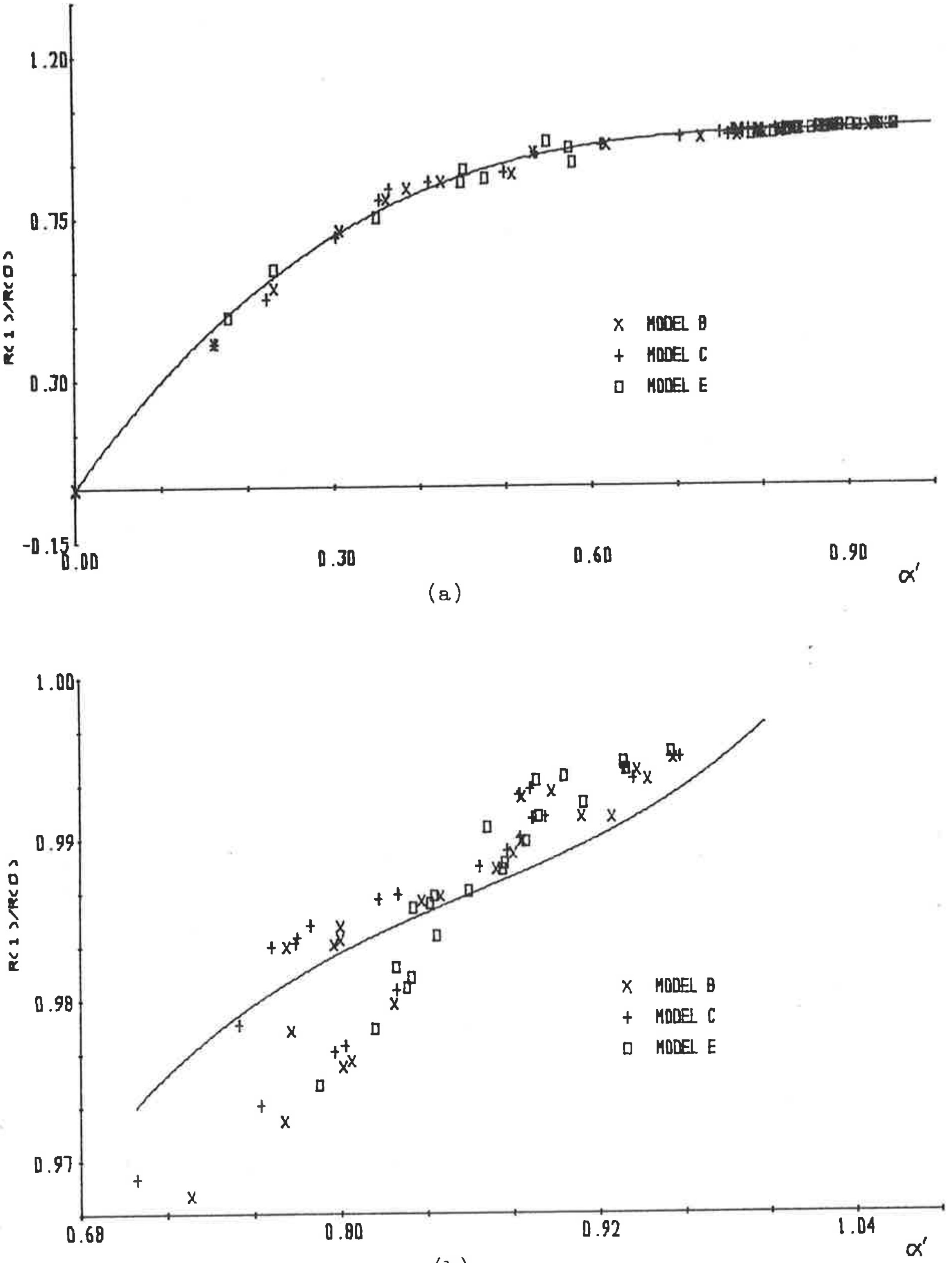


FIGURE 4.27: Data points (α′, R(1)/R(0)) and the best fit polynomial, for glottal pulses generated from the Rosenberg glottal pulse model E for α′ between (a) zero and unity and (b) 0.55 and unity.



(a)

(b)

FIGURE 4.28: Data points (α′, R(1)/R(0)) and best fit polynomial, for glottal pulses generated from the Rosenberg glottal pulse models B, C and E for α′ between (a) zero and unity and (b) 0.68 and unity.

best fit polynomial to define an adaptive pre-emphasis filter which effectively removes, or accounts for, glottal pulse waveform effects in speech analysis problems.

The spread of data points that is observed around the best fit polynomial in Figure 4.28 indicates that the value of α' obtained from the best fit polynomial does not necessarily provide a true minimum area distance. However, investigations show that the change from the true minimum is relatively small when compared with the area distances obtained when no pre-emphasis or other available pre-emphases are used. These investigations also show that reduced area distances occur when the best fit polynomial is used to define the two/three adaptive pre-emphasis filter in comparison with available pre-emphasis techniques.

Additional glottal pulse shape variation is provided by using the FANT glottal pulse model, defined by Equations 4.3 to 4.6. A set of data points $(\alpha', R(1)/R(0))$ is generated using glottal pulses from the FANT glottal pulse model, in the same manner as for the ROSENBERG glottal pulse models. The range of parameters T_p/T and K are mainly kept within the preferred ranges found by ROSENBERG's [1971] subjective testing. However, as with the ROSENBERG glottal pulse models, some parameter values outside this preferred range are used to provide data points which cover the range of $R(1)/R(0)$ from zero to unity.

The set of data points $(\alpha', R(1)/R(0))$ generated by using the FANT glottal pulse model is plotted in Figure 4.29. The best fit polynomial found by the least squares curve fitting algorithm

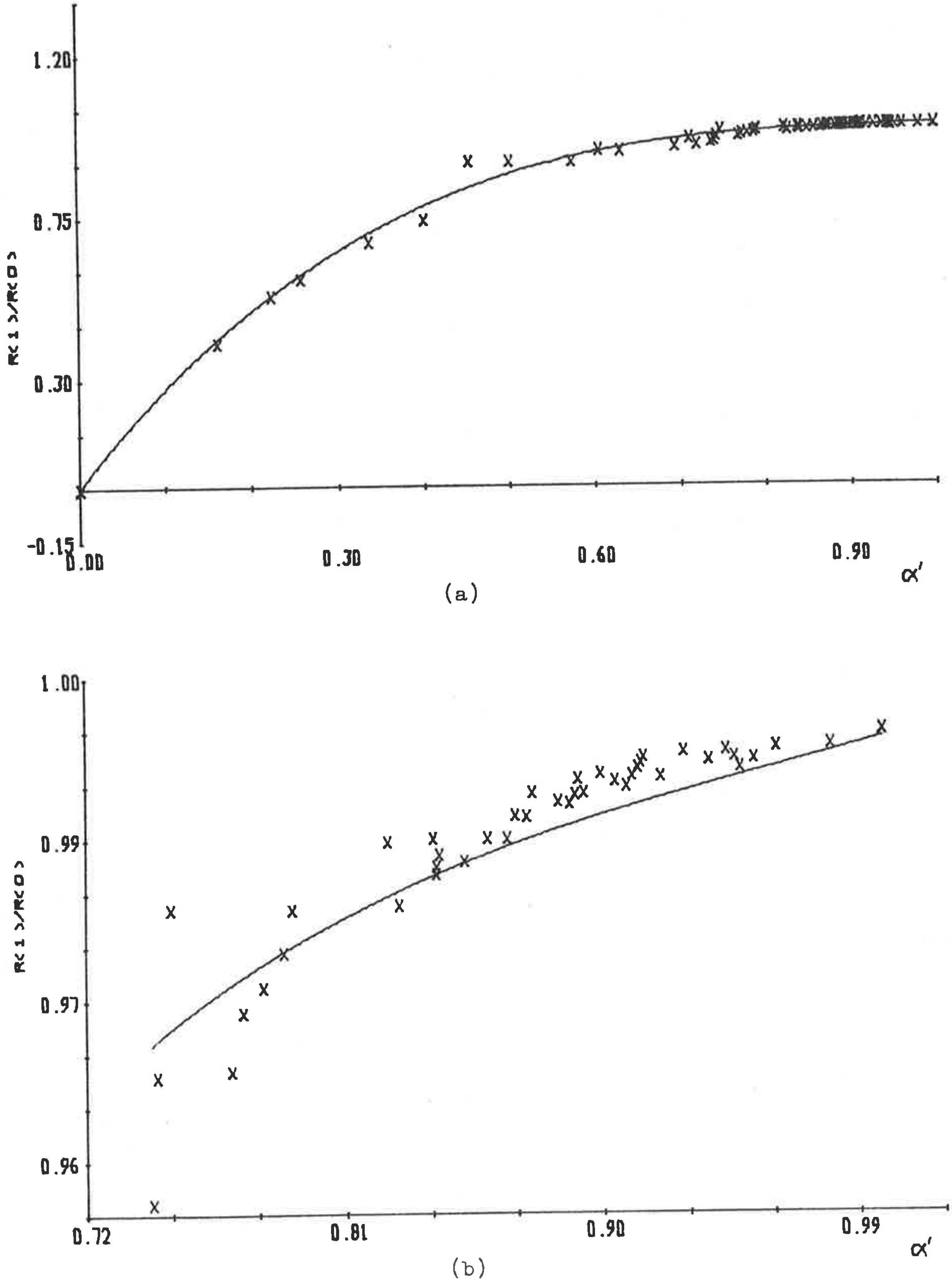


FIGURE 4.29: Data points (α´, R(1)/R(0)) and the best fit polynomial, for glottal pulses generated from the fant glottal pulse model for α´ between (a) zero and unity and (b) 0.72 and unity.

[LEONARD 1965] is also plotted in Figure 4.29 and defines a relationship between α' and $R(1)/R(0)$ as

$$\frac{R(1)}{R(0)} = 2.982(\alpha') - 3.044(\alpha')^2 + 1.063(\alpha')^3 \qquad (4.24)$$

This best fit polynomial is similar to that for ROSENBERG's glottal pulse models B, C and E, i.e. Equation 4.23. Hence, completely different glottal pulse models provide similar relationships between α' and $R(1)/R(0)$, and so it appears plausable that a single best fit polynomial can be found from which α' is accurately determined for any glottal pulse slope.

Although the glottal pulses considered so far satisfy the requirements for realistic glottal pulse shapes as determined by ROSENBERG's [1971] subjective testing, real glottal pulses need to be considered also. Due to the difficulty in measuring real glottal pulse waveforms, those published by other researchers [ROTHENBERG 1973, ENGEBRETSON and VEMULA 1974, MILLER 1959, SONDHI 1975, SUNDBERG and GAUFFIN 1978] are used. These published glottal pulse waveforms were obtained by a number of different methods, and under a variety of conditions for both male and female subjects, to provide a wide range of different glottal pulse shapes.

The glottal pulse waveforms presented by MILLER [1959] are determined from the radiated speech pressure waveform by filtering with a network whose characteristic is the inverse of the first vocal resonance. A number of voiced sounds with various fundamental frequencies and relative vocal effort are used to produce the published glottal pulse waveforms. MILLER also presents measurements of vocal cord opening as a function of time from motion picture studies of the glottis, and these are used in this section.

A major effort to accurately record glottal pulse waveforms was made by ROTHENBERG [1973] who designed a special mask to allow the volume velocity at the lips to be measured accurately down to zero frequency. Inverse filtering of the volume velocity waveform at the lips then produces glottal pulse waveforms. ROTHENBERG presents glottal pulse waveforms obtained by this procedure for two male speakers with various fundamental frequencies and vocal effort. Studies by SUNDBERG and GAUFFIN [1978] of the glottal pulse waveform and its spectra used the ROTHENBERG mask, but with an improved inverse filter. The glottal pulse waveforms published by SUNDBERG and GAUFFIN for different types of phonation are also used in this section.

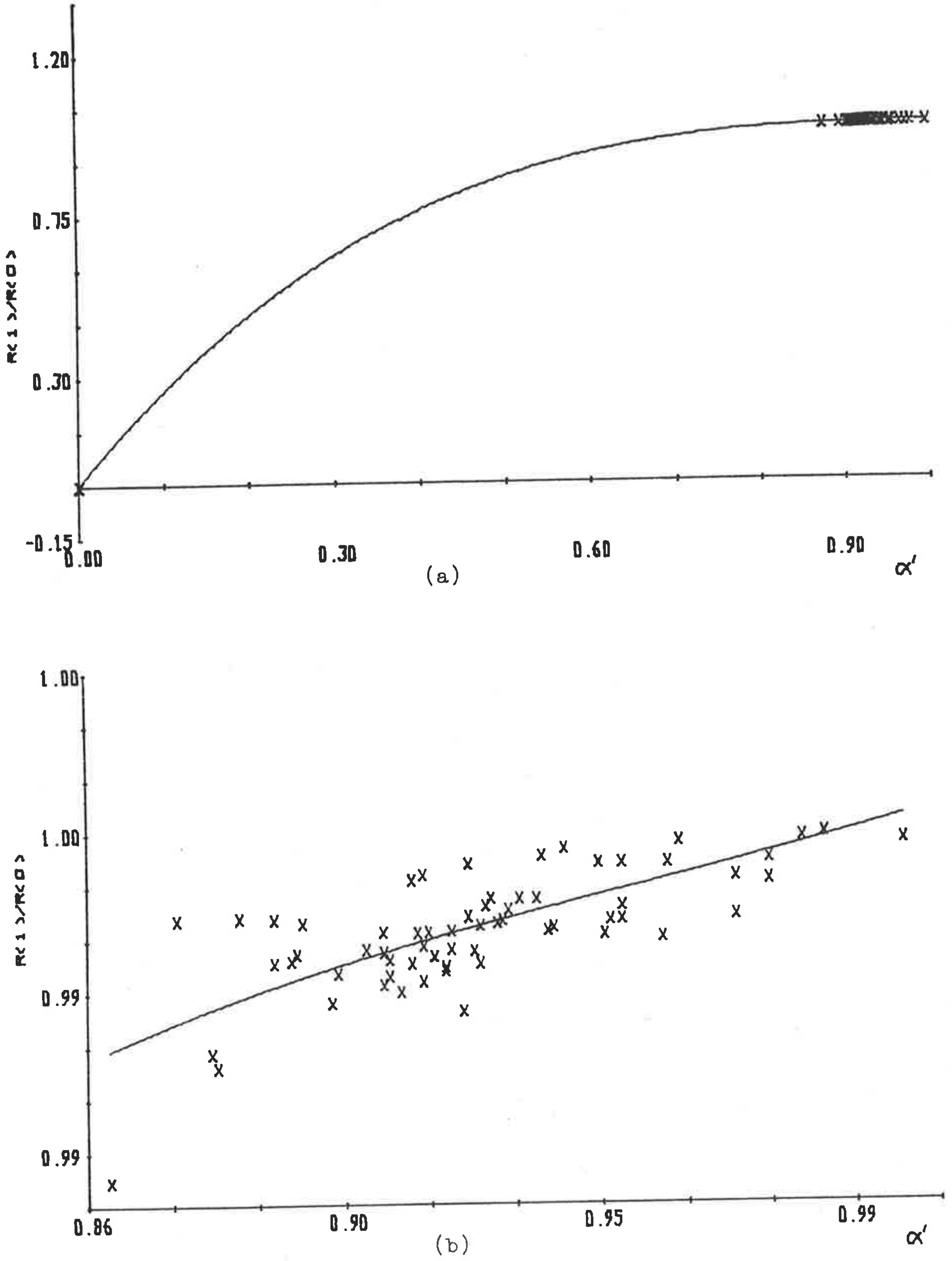
Glottal pulse waveform measurement by SONDHI [1975] and MONSEN and ENGEBRETSON [1977] used a reflectionless metal tube instead of a mask, to measure the volume velocity at the lips. Terminating the vocal tract at the lips by a reflectionless tube significantly reduces the effect of the vocal tract resonances on the glottal pulse waveform. Therefore, the volume velocity measured in the metal tube is a reasonable approximation of the glottal pulse volume velocity. The glottal pulse waveforms presented by MONSEN and ENGEBRETSON [1977] are for 10 male and female speakers with phonation over a wide range of vocal effort and fundamental frequencies, for monosyllable and three-syllable utterances.

The various descriptions of the procedures used to measure glottal pulse volume velocities indicate that none of the published glottal pulse waveforms are true excitation waveforms for the vocal tract. Each of the procedures used does not completely remove the effects of the vocal tract through which the glottal volume velocity passes. However, the errors in the glottal pulse waveforms measured by these procedures is generally small when comparied with measurements taken of vocal cord vibration from motion picture studies.

Using the published glottal pulse waveforms, a set of data points $(\alpha', R(1)/R(0))$ is obtained by repeated use of the procedure defined in Figure 4.22, and these data points are plotted in Figure 4.30. The best fit polynomial is determined by the least squares curve fitting algorithm [LEONARD 1965] as

$$\frac{R(1)}{R(0)} = 3.122(\alpha') - 3.305(\alpha')^2 + 1.183(\alpha')^3 \qquad (4.25)$$

and is plotted in Figure 4.30. Comparing this best fit polynomial with those obtained for the synthetic glottal pulses, i.e. Equations 4.20 and 4.24, shows reasonable similarity. In fact, the best fit polynomial for real glottal pulses, i.e. Equation 4.25, lies between the best fit polynomials for the ROSENBERG and FANT glottal pulse models. Hence, it appears that the data points $(\alpha', R(1)/R(0))$ generated from synthetic and real glottal pulses are consistent with one another and, therefore, a single best fit polynomial can be found which provides α' accurately for real or synthetic glottal pulses.



(a)

(b)

FIGURE 4.30: Data points (α´, R(1)/R(0)) and the best fit polynomial, for glottal pulses derived from measured glottal pulse waveforms for α´ between (a) zero and unity and (b) 0.86 and unity.

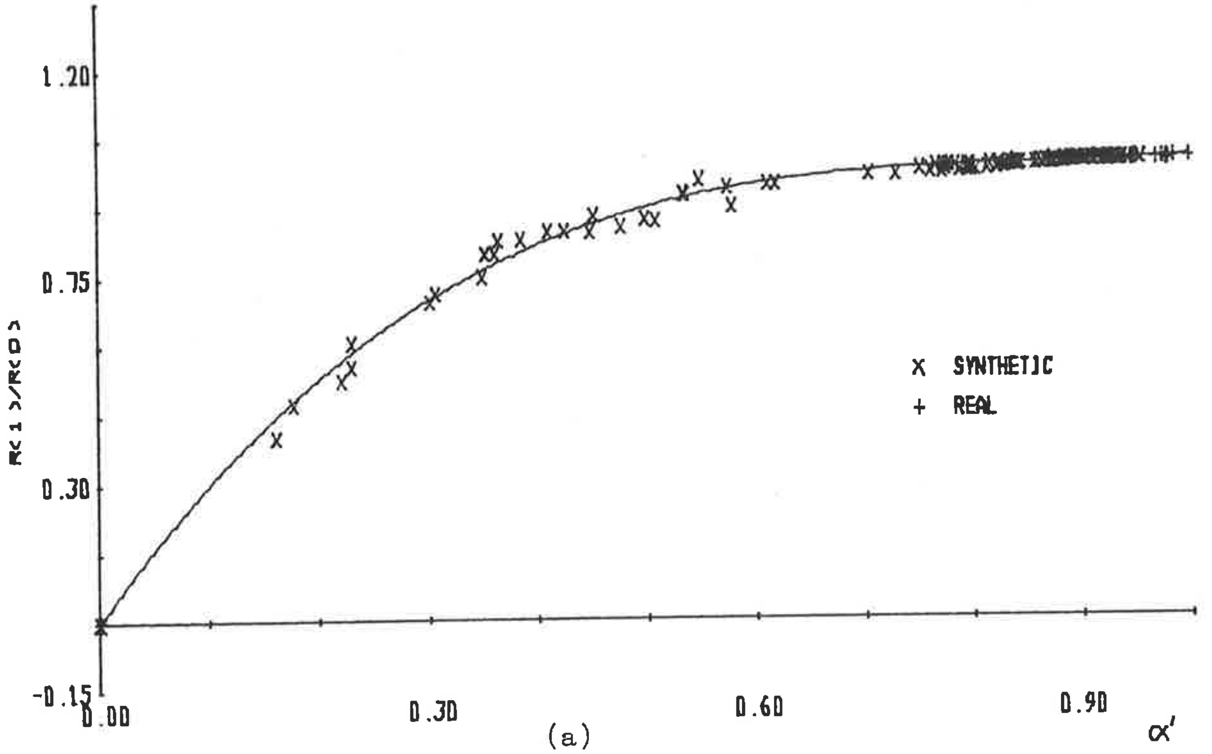
Figure 4.31 presents a plot of all the data points $(\alpha', R(1)/R(0))$ generated to this point, from synthetic and real glottal pulses, with different symbols being used to indicate if the data point was generated from a real or synthetic glottal pulse. The even spread of different symbols in Figure 4.31 verifies the above conclusion that the data points $(\alpha', R(1)/R(0))$ generated from real or synthetic glottal pulses are consistent with one another.

The best fit polynomial to the data points in Figure 4.31 is found by the least squares curve fitting algorithm [LEONARD 1965] as

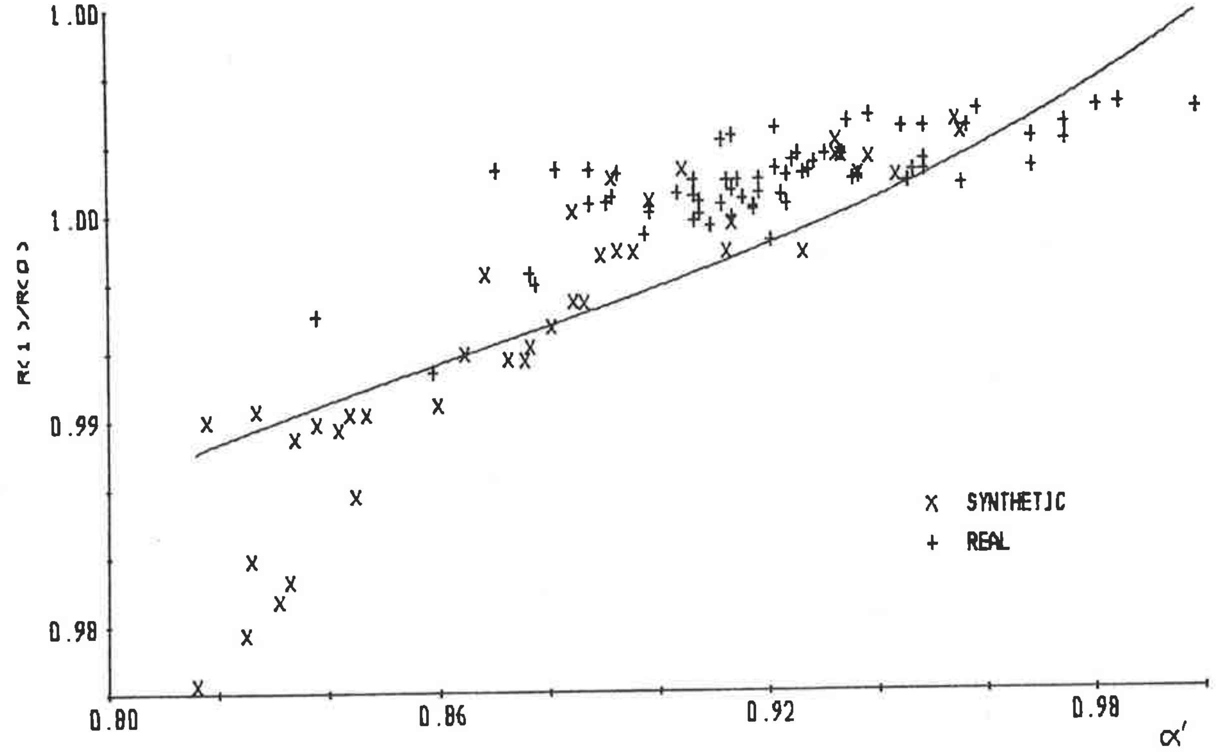
$$\frac{R(1)}{R(0)} = 3.217(\alpha') - 3.552(\alpha')^2 + 1.339(\alpha')^3 \qquad (4.26)$$

This best fit polynomial is determined from a total of 244 data points generated from 69 real and 175 synthetic glottal pulses, covering the broad range of glottal pulse shapes found in voiced speech. Equation 4.26 is the required relationship between α' and $R(1)/R(0)$ which defines the two/three adaptive pre-emphasis filter when a sampling frequency of 10 kilohertz is used and not more than +12 dB per octave pre-emphasis is required.

Equation 4.26 successfully satisfies the necessary requirements set out for the relationship between α' and $R(1)/R(0)$ at the start of this section, i.e. simplicity, accuracy of fit and continuity over the range of $R(1)/R(0)$ values. A direct implementation of Equation 4.26 is not possible, since a value of α' is required from a known $R(1)/R(0)$, and so, as it was determined earlier, Equation 4.26 must be implemented as a lookup table. Appendix D presents the details and the reduction in computations (hence



(a)



(b)

FIGURE 4.31: Data points (α′, R(1)/R(0)) and the best fit polynomial, for synthetic glottal pulses generated from glottal pulse models and real glottal pulses from measured glottal pulse waveforms for α′ from (a) zero to unity and (b) 0.80 and unity.

faster evaluation of α') that occurs through implementation of Equation 4.26 as a lookup table.

4.4.2.2 Parameter β

When greater than +12 dB per octave pre-emphasis is required to remove glottal pulse excitation effects from the recovered acoustic tube shape, the two/three adaptive pre-emphasis filter uses three zeros. Two of these three zeros are fixed at $z = 1$, while the third is determined by a parameter β, via a relationship with $R(1)/R(0)$ of the waveform being analysed. The purpose of this section is to determine the relationship between $R(1)/R(0)$ and β which produces a minimum area distance between the recovered and original acoustic tube shapes.

Repeated use of the procedure defined in Figure 4.22 for different glottal pulse shapes generates a set of data points $(R(1)/R(0), \beta')$ from which the required relationship between $R(1)/R(0)$ and β' is determined. The requirements that this relationship must satisfy are the same as those for the relationship between α' and $R(1)/R(0)$ developed in the previous section, i.e. simplicity, accuracy of fit to the data points and continuity over the range of $R(1)/R(0)$ considered. Investigations have shown that fitting a polynomial function to the data points $(R(1)/R(0), \beta')$ satisfies the necessary requirements of the relationship between $R(1)/R(0)$ and β'.

The best fit polynomial to the data points $(R(1)/R(0), \beta')$ has the form

$$\beta' = \sum_{i=0}^{n} d_i \left(\frac{R(1)}{R(0)} \right)^i \tag{4.27}$$

where n is the order of the best fit polynomial. Investigations with various curve fitting algorithms have shown that the same least squares curve fitting algorithm [LEONARD 1965] used to determine the relationship between α' and $R(1)/R(0)$ also provides an accurate polynomial fit to the data points $(R(1)/R(0),\ \beta')$. The order, n, of the best fit polynomial is that which minimizes the deviation between the data points and the best fit polynomial.

The data points $(R(1)/R(0),\ \beta')$ plotted in Figure 4.23 are generated by using glottal pulses derived from ROSENBERG's glottal pulse model B [ROSENBERG 1971] (defined by Equations 4.1 and and 4.2) with the procedure detailed in Figure 4.22. A large spread of data points is found in Figure 4.32 which causes large errors for any type of function fitting those data points. Hence, the data points presented in Figure 4.32 are not acceptable for determining a relationship between $R(1)/R(0)$ and β'.

Investigations show that the data points in Figure 4.32 are widespread because of a slope discontinuity that occurs in the synthetic glottal pulse at closure. Whenever a signal with a discontinuity is passed through a filter containing a zero (i.e. a differentiation is performed) then the relative magnitude of the discontinuity increases with respect to the rest of the signal. In the calculation of β' the two/three adaptive pre-emphasis filter uses three zeros, and it is found that the magnitude of the discontinuity becomes large enough for the procedure detailed in Figure 4.22 to find a β' which attempts to remove the discontinuity, instead of the glottal pulse shape.

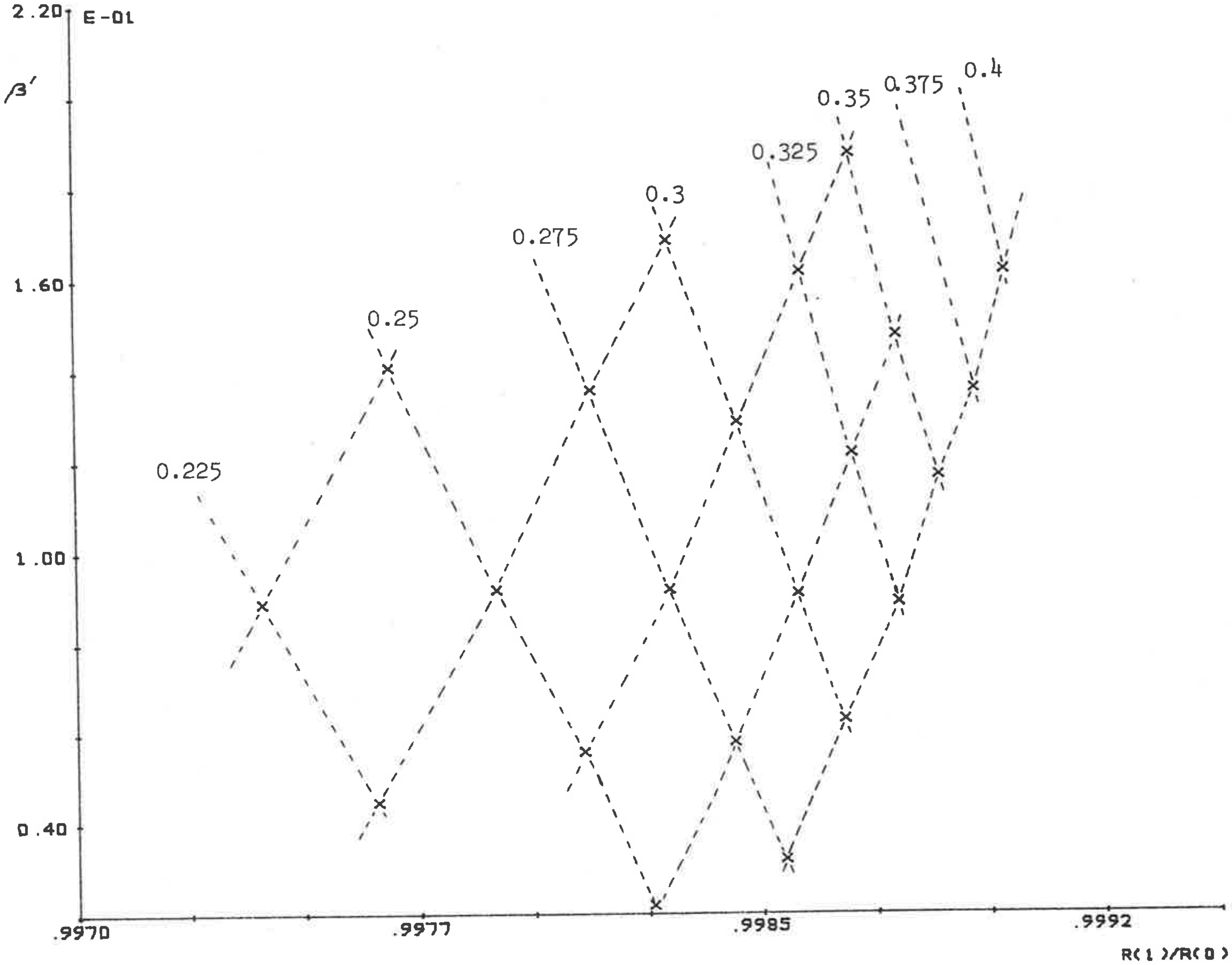


FIGURE 4.32: Data points $(R(1)/R(0), \beta')$ for glottal pulses generated from the Rosenberg glottal pulse model B for various values of parameter T_N/T.

The value of the slope discontinuity in the glottal pulse derived by ROSENBERG's glottal pulse model B is totally dependent on the pulse parameter T_N/T, and Figure 4.32 shows that whenever T_N/T changes, the value of β' changes significantly. While T_N/T is constant, an approximately linear relationship between $R(1)/R(0)$ and β' occurs. Therefore, it can be concluded that whenever a synthetic glottal pulse has a slope discontinuity, a spreading of data points $(R(1)/R(0),\ \beta')$ occurs, which is due to the value of β' being highly dependent on the magnitude of the slope discontiuity.

The other glottal pulse models considered in the previous section, i.e. ROSENBERG's glottal pulse models C and E and FANT's glottal pulse model, all have slope discontinuities, and so cannot be used to find a relationship between $R(1)/R(0)$ and β'. ROSENBERG [1971] considered a glottal pulse model which has no slope discontinuities (i.e. model D), but subjective testing found that synthetic speech generated by using this glottal pulse model resulted in almost zero preference scores. Hence, all the glottal pulse models do not appear to provide acceptable data points $(R(1)/R(0),\ \beta')$ and, therefore, only real glottal pulses derived from published glottal pulse waveforms are considered in this section.

Figure 4.33 presents a plot of the data points $(R(1)/R(0),\ \beta')$ generated by the procedure detailed in Figure 4.22 with glottal pulse waveforms derived from the published glottal pulse waveforms of MONSEN and ENGEBRETSON [1977], ROTHENBERG [1973], SONDHI [1975], SUNBERG and GAUFFIN [1978], MILLER [1959], and FLANAGAN and LANDGRAF [1968]. The best fit polynomial to the data points

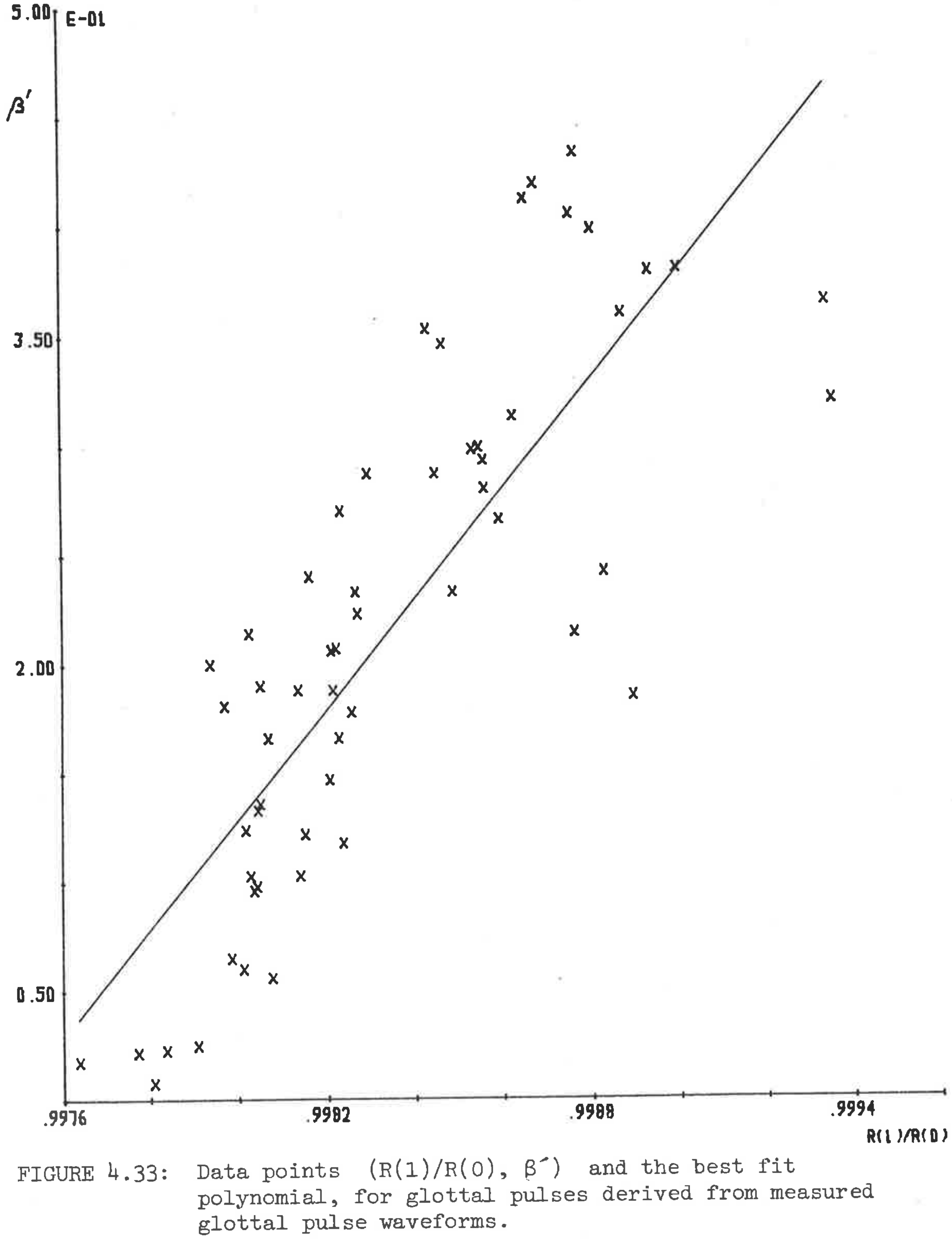


FIGURE 4.33: Data points (R(1)/R(0), β´) and the best fit polynomial, for glottal pulses derived from measured glottal pulse waveforms.

of Figure 4.33 is plotted in that figure, and determined by the least squares curve fitting algorithm [LEONARD 1965] as

$$\beta' = -246.791 + 247.416\frac{R(1)}{R(0)} \qquad (4.28)$$

Equation 4.28 defines the relationship between the parameter β of the two/three adaptive pre-emphasis filter and $R(1)/R(0)$ of the glottal pulse waveform, when its spectral slope has a magnitude of approximately +12 dB per octave or greater.

The coefficients of Equation 4.28 are similar and, since the value of $R(1)/R(0)$ is close to unity, some problems may be encountered when implementing this equation on a limited word length computer or specialized machine, i.e. large errors can result when subtracting two numbers which are approximately the same. An alternative form of Equation 4.28 which increases the accuracy of the resultant β' for limited/small word length calculations is

$$\beta' = 2.4741\ (R-0.7474) \qquad (4.29a)$$

where

$$R = \frac{\frac{R(1)}{R(0)} - 0.99}{100} \qquad (4.29b)$$

The relationship between β' and $R(1)/R(0)$ defined by either Equation 4.28 or Equation 4.29 is of simple form, requiring only one multiplication and some subtractions; hence, in most situations, there is no necessity for a lookup table, as was required for the determination of α' from $R(1)/R(0)$ (see Section 4.4.2.1). If Equation 4.28 is used to calculate β' from $R(1)/R(0)$, then there is a necessity to perform a subtraction which requires long word lengths for an accurate result, and so in this situation a lookup table would be an advantage.

The number of glottal pulse waveforms used to define the relationship between β and $R(1)/R(0)$, i.e. Equations 4.28 or 4.29, is much smaller than the number of glottal pulses used to define the relationship between α and $R(1)/R(0)$ (see Section 4.4.2.1). This is due to the smaller occurrence of glottal pulse waveforms with a spectral slope magnitude greater than +12 dB per octave. Hence, the accuracy of the relationship between β' and $R(1)/R(0)$ defined by Equations 4.28 or 4.29 is less than the accuracy of the relationship between α' and $R(1)/R(0)$ determined in Section 4.4.2.1 over all possible glottal pulse waveforms.

Figure 4.33 shows an apparently large spread of data points around the best fit polynomial curve, but a similar spread of data points is observed in Section 4.4.2.1 for a similar range of $R(1)/R(0)$. The small range of $R(1)/R(0)$ presented in Figure 4.33 indicates that the value of $R(1)/R(0)$ must be calculated to a greater accuracy than in the case where α is determined from $R(1)/R(0)$. Hence, the implementation of Equations 4.28 or 4.29 requires $R(1)/R(0)$ to be calculated on relatively long word length machines with relatively large computational accuracy, i.e. at least four significant figures.

The full range of β' from zero to unity is not presented in Figure 4.33 because glottal pulse waveforms derived from measured glottal pulse waveforms could not be found with spectral slope magnitudes approaching -18 dB per octave. Hence, an extrapolation of the best fit polynomial curve, i.e. Equations 4.28 or 4.29, is used when β' is between 0.5 and unity. The possibility that the best fit polynomial curve is not representive of the relationship between β' and $R(1)/R(0)$ for β' between 0.5 and unity is investigated when the two/three adaptive pre-emphasis filter is evaluated in Chapter 5.

The transition point where the two/three adaptive pre-emphasis filter changes from being defined by α to β or vice versa is determined from Equations 4.28 or 4.29 as the value of $R(1)/R(0)$ when β' is zero. Equation 4.28 or Equation 4.29 determines the transition point as $R(1)/R(0) = 0.99747$, so that the two/three adaptive pre-emphasis filter is defined by the parameter α when $R(1)/R(0)$ is less than 0.99747, and by the parameter β when $R(1)/R(0)$ is greater than 0.99747.

4.4.3 INVESTIGATION OF SAMPLING FREQUENCY EFFECTS

The relationships developed in the previous section between $R(1)/R(0)$ and the parameters α and β of the two/three adaptive pre-emphasis filter are only for one sampling frequency, i.e. 10 kilohertz. A change in sampling frequency causes the value of $R(1)$ to change; hence, the relationships between parameters α and β and the value of $R(1)/R(0)$ also changes. This section repeats the procedures used in Section 4.4.2, and investigates the relationships between $R(1)/R(0)$ and the parameters α and β for a range of sampling frequencies which are commonly used in speech analysis.

Sampling frequencies used in speech analysis vary over a wide range and depend on the application. However, the upper and lower bounds for the sampling frequency are determined by a number of basic criteria. For voiced speech, a significant loss of information occurs if bandlimiting to less than 3 kilohertz is performed, while very little information exists above 5 kilohertz. NYQUIST's sampling theorem requires that the sampling frequency be at least twice the highest frequency component of the signal being sampled. Hence, for voiced sounds, the lowest sampling frequency considered is around 6 kilohertz, and the largest around 10 kilohertz. Due to the finite roll-off of real low pass filters used to bandlimit signals, a practical minimum sampling frequency for voiced speech is around 6.5 kilohertz.

For the application of speech analysis to vocal tract shape recovery, the sampling frequency is determined by the amount of detail required in the recovered vocal tract shape. For a sampling frequency of f_s (f_s an integer) then f_s+1* discrete cross-sectional areas are recovered from the speech waveform. Using the upper and lower bounds of sampling frequency determined above, then between 8 and 11 discrete vocal tract cross-sectional areas can be recovered from the speech waveform. In most cases, this is sufficient to enable all but the fine detail around the teeth to be observed. Hence, only the range of sampling frequencies between 7 and 10 kilohertz is considered for vocal tract shape recovery in this thesis.

*Assuming the velocity of sound is 34 cms/msec and a vocal tract tract length of 17 cms.

The preliminary investigations of sampling frequency effects performed in Section 4.3 showed that changes in $R(1)/R(0)$ are significant, but not large, for sampling frequency changes from 6 to 10 kilohertz. Therefore, the relationships between $R(1)/R(0)$ and the parameters α and β are expected to change only slightly for the range of sampling frequencies defined above, i.e. 7 to 10 kilohertz. Hence, this section only investigates the relationships between $R(1)/R(0)$ and the parameters α and β of the two/three adaptive pre-emphasis filter for integer sampling frequencies of 7, 8 and 9 kilohertz.

The relationships between $R(1)/R(0)$ and the parameters α and β for the various sampling frequencies are determined in the same manner as those in Section 4.4.2, i.e. for a sampling frequency of 10 kilohertz. This procedure determines data points $\left(R(1)/R(0),\ \alpha'\right)$ or $\left(R(1)/R(0),\ \beta'\right)$ for a particular glottal pulse waveform, where α' and β' are the parameter values of the two/three adaptive pre-emphasis filter which provide a minimum area distance after a Parcor analysis of the pre-emphasized glottal pulse waveform (see Figure 4.22). The required relationships are then found by curve fitting to a large set of data points $\left(R(1)/R(0),\ \alpha'\right)$ or $\left(R(1)/R(0),\ \beta'\right)$. This procedure enables linear prediction to tailor the two/three adaptive pre-emphasis filter to remove the most significant portions of the glottal pulse excitation from the speech waveform, and in a loose sense enables an optimal reduction in area distances.

The glottal pulse waveforms used in the following sections are derived from the same glottal pulse models and published glottal pulse waveforms as used in Section 4.4.2. Each glottal pulse waveform can only be used to determine one of the relationships between α' and $R(1)/R(0)$ or β' and $R(1)/R(0)$. Hence, the glottal pulses are divided into two groups, with one group used in the following section (i.e. Section 4.4.3.1) to determine the relationship between α' and $R(1)/R(0)$, and the other group being used in Section 4.4.3.2 to determine the relationship between β' and $R(1)/R(0)$.

4.4.3.1 Parameter α

This section determines the relationships between α' of the two/three adaptive pre-emphasis filter and $R(1)/R(0)$ of real and synthetic glottal pulse waveforms for sampling frequencies of 7, 8 and 9 kilohertz. Repeated use of the process defined in Figure 4.22 for a large number of different glottal pulse waveforms generates a set of data points $(R(1)/R(0), \alpha')$ for each sampling frequency. Determination of a best fit function to each of these sets of data points then produces the required relationships between α' and $R(1)/R(0)$.

Investigations show that the sets of data points $(R(1)/R(0), \alpha')$ for the different sampling frequencies considered have the same form as those for a sampling frequency of 10 kilohertz. Therefore, any best fit function has difficulty in fitting the steep slope for $R(1)/R(0)$ near unity while still providing an accurate fit for the rest of the range of $R(1)/R(0)$. The solution to this problem is the same one taken in Section 4.4.2.1, that is to find the inverse relationship, i.e. $R(1)/R(0)$ in terms of α', from the data points $(\alpha', R(1)/R(0))$. This provides an approximately zero slope for $R(1)/R(0)$ near unity, which permits a simple best fit function to be used. The inverse relationship is then used to define a lookup table (see Appendix D) so that α' is found quickly and easily from a known $R(1)/R(0)$ value.

The best fit function that satisfies the requirements of simplicity, accuracy of fit and continuity over the range of $R(1)/R(0)$ values is a polynomial function. This best fit polynomial is the same as that used in Section 4.4.2.1, which is defined by

$$\frac{R(1)}{R(0)} = \sum_{i=0}^{n} c_i(\alpha')^{i} \qquad (4.30)$$

The order, n, of the best fit polynomial is determined as that which minimizes the deviation between the data points and the best fit polynomial. As found in Section 4.4.2, the least squares curve fitting algorithm [LEONARD 1965] provides an accurate fit of the data points to the best fit polynomial. As in Section 4.4.2.1, a weighting function is used to ensure the polynomial is as close to the end points (0,0) and (1,1) as possible. The sparsity of data points near (0,0) requires such a weighting function for the best fit polynomial to be close to (0,0).

The glottal pulse waveforms used to define the required relationship between α' and $R(1)/R(0)$ are generated from the same sources as used in Section 4.4.2.1. All the synthetic glottal pulses are derived from the glottal pulse models B, C and E of ROSENBERG [1971] and FANT's glottal pulse model [1979], with the same parameter values as used in Section 4.4.2.1. The real glottal pulse waveforms are generated from the same published glottal pulse waveforms [ROTHENBERG 1973, MONSEN and ENGEBRETSON 1977, MILLER 1959, SONDHI 1975, SUNDBERG and GAUFFIN 1978] as used in Section 4.4.2.1.

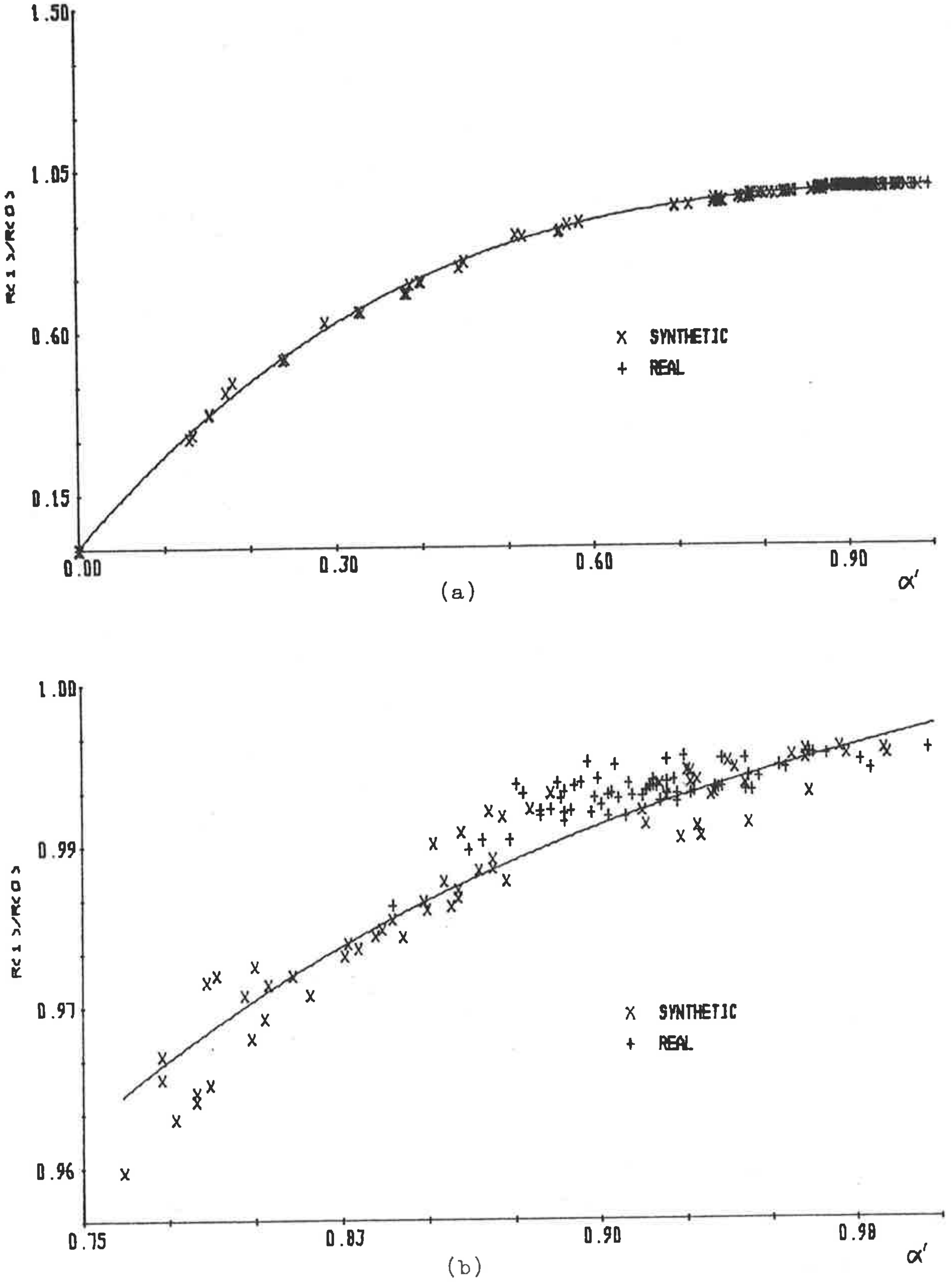
For a sampling frequency of 9 kilohertz, the set of data points $(\alpha', R(1)/R(0))$ generated from the wide range of real and synthetic glottal pulses is plotted in Figure 4.34. Different plotting symbols are used in Figure 4.34 to indicate which data points are generated from real and which from synthetic glottal pulse waveforms. The even spread of different plotting symbols in Figure 4.34(b) shows that the synthetically generated glottal pulse waveforms are consistent with the results obtained from real glottal pulse waveforms, as found for the 10 kilohertz sampling frequency case considered in Section 4.4.2.1.

The best fit polynomial to the data points in Figure 4.34 is found by the least squares curve fitting algorithm [LEONARD 1965] as

$$\frac{R(1)}{R(0)} = 3.244(\alpha') - 3.610(\alpha')^2 + 1.371(\alpha')^3 \qquad (4.31)$$

A comparison of the above best fit polynomial with that for a 10 kilohertz sampling frequency (i.e. Equation 4.26) reveals only a small variation in polynomial coefficients, c_i. Therefore, it is concluded that a waveform sampling frequency variation from 10 to 9 kilohertz has only a small effect on the relationship between α' and $R(1)/R(0)$.

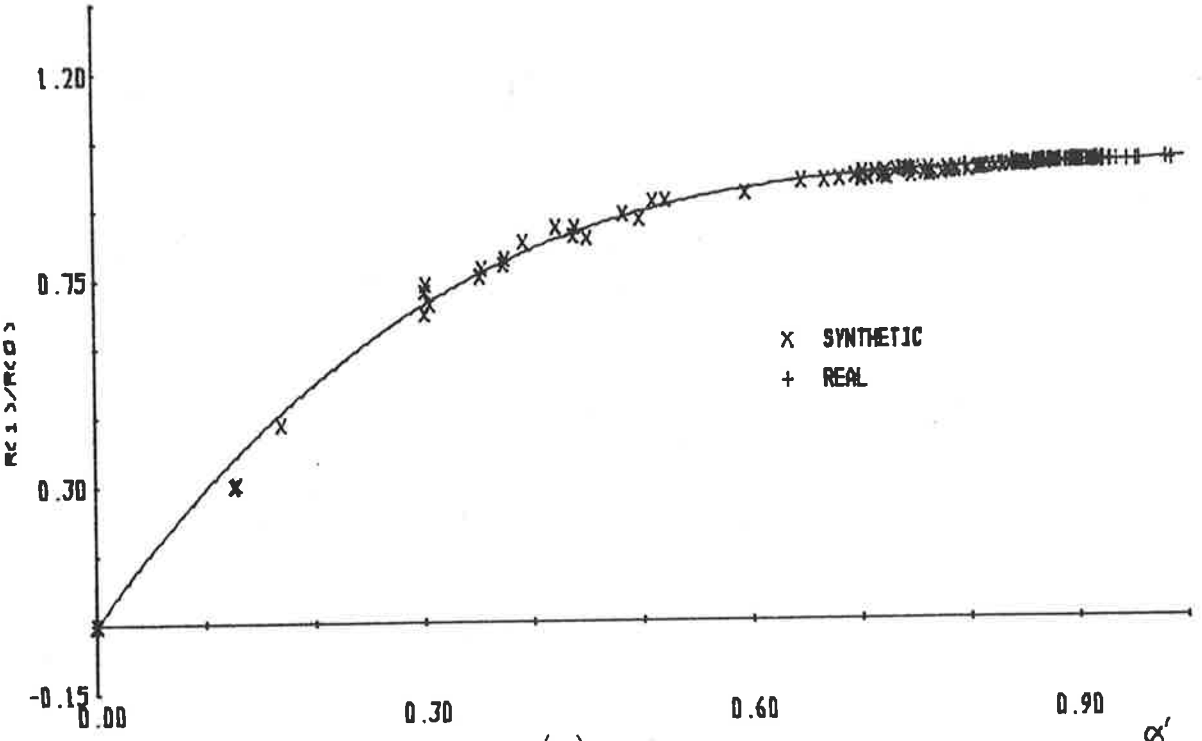
The data points plotted in Figure 4.35 are generated with the same glottal pulse waveforms used to generate Figure 4.34, but sampled at a frequency of 8 kilohertz. Different plotting symbols are used to indicate those data points generated from real and those from synthetic glottal pulse waveforms, an an even spread of those symbols is observed, as in the case of 10 and



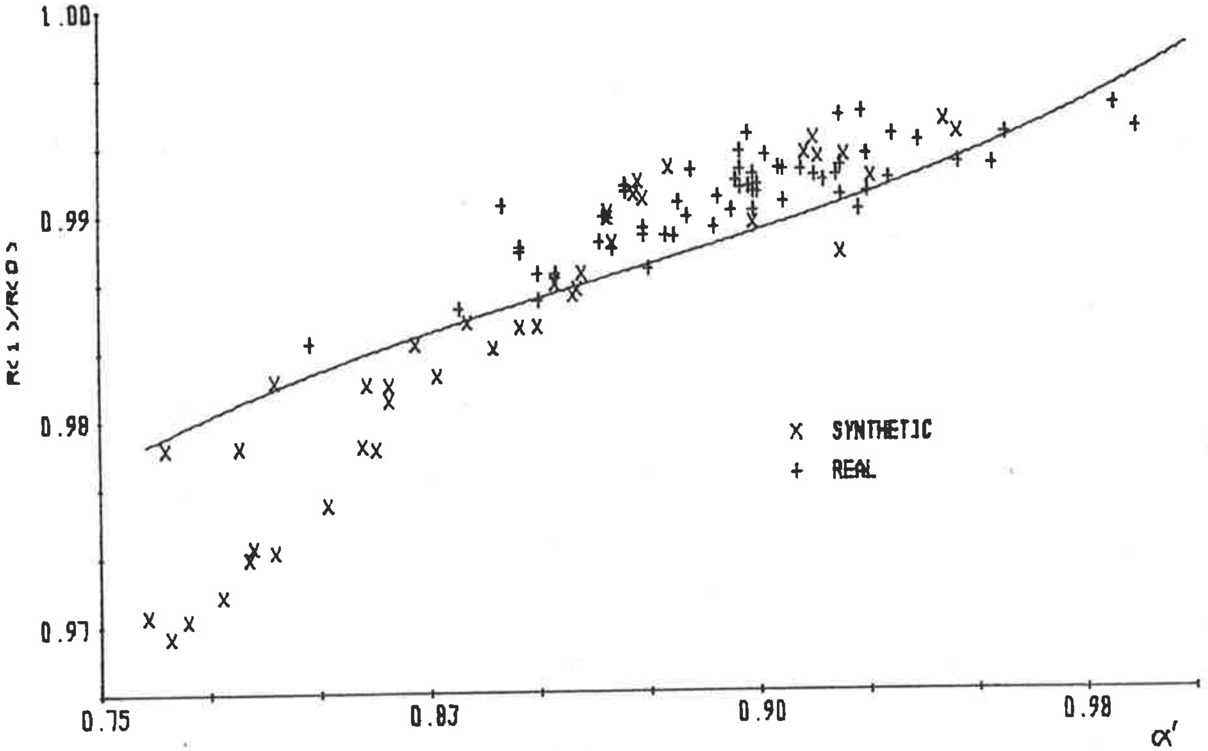
(a)

(b)

FIGURE 4.34: Data points (α′, R(1)/R(0)) and the best fit polynomial, for glottal pulses sampled at a frequency of 9 kilohertz and derived synthetically from glottal pulse models or from real glottal pulses for α′ between (a) zero and unity and (b) 0.75 and unity.



(a)



(b)

FIGURE 4.35: Data points (α′, R(1)/R(0)) and the best fit polynomial, for glottal pulses sampled at a frequency of 8 kilohertz and derived synthetically from glottal pulse models or from real glottal pulses for α′ between (a) zero and unity and (b) 0.75 and unity.

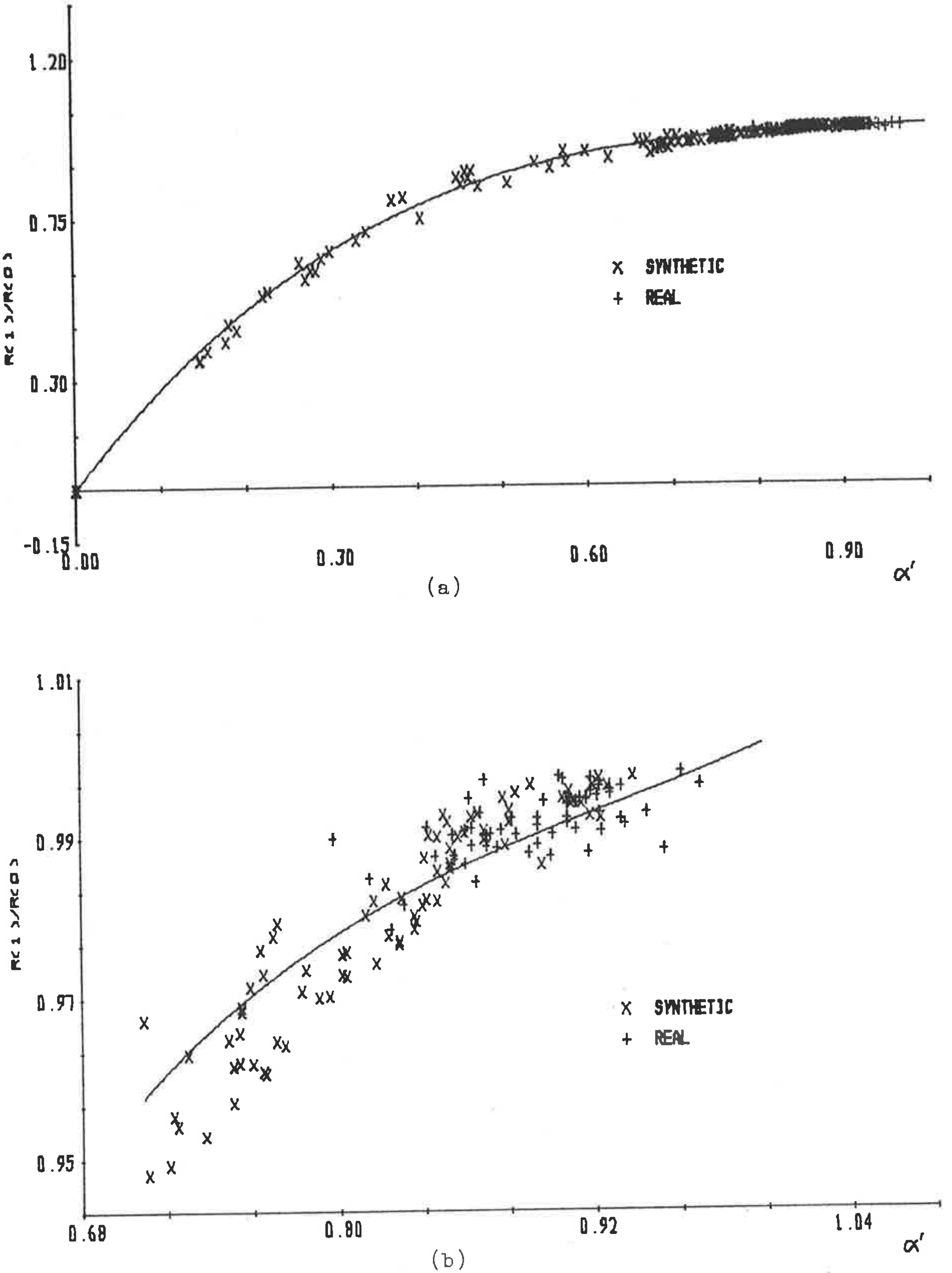
9 kilohertz sampling frequencies (see Figure 4.31 and 4.34, respectively). The best fit polynomial to the data points in Figure 4.35 is found by the least squares curve fitting algorithm [LEONARD 1965] as

$$\frac{R(1)}{R(0)} = 3.211(\alpha') - 3.545(\alpha')^2 + 1.339(\alpha')^3 \qquad (4.32)$$

Comparison of this best fit polynomial with those for a sampling frequency of 10 or 9 kilohertz (i.e. Equations 4.26 and 4.31, respectively) reveals only small variations in the best fit polynomial coefficients, c_i. Therefore, a waveform sampling frequency variation between 10 and 8 kilohertz (inclusive) has only a small effect on the relationship between α' and $R(1)/R(0)$.

The last waveform sampling frequency considered in this section is one of 7 kilohertz. Figure 4.36 presents the data points $\left(\alpha', R(1)/R(0)\right)$ generated from the same glottal pulse waveforms and by the same analysis procedure as used to generate Figures 4.31, 4.34 and 4.35 (i.e. those for 10, 9 and 8 kilohertz sampling frequencies, respectively), but sampled at 7 kilohertz. Data points in Figure 4.36 generated from real or synthetic glottal pulse waveforms are distinguished by the use of different plotting symbols. An even spread of plotting symbols in Figure 4.36 shows a consistency between the real and synthetic glottal pulse waveforms.

Using the least squares curve fitting algorithm [LEONARD 1965], the best fit polynomial to the data points of Figure 4.36 is



(a)

(b)

FIGURE 4.36: Data points (α′, R(1)/R(0)) and the best fit polynomial, for glottal pulses sampled at a frequency of 7 kilohertz and derived synthetically from glottal pulse models or from real glottal pulses for α′ between (a) zero and unity and (b) 0.75 and unity.

$$\frac{R(1)}{R(2)} = 3.205(\alpha') - 3.148(\alpha')^2 + 1.127(\alpha')^3 \qquad (4.33)$$

A significant difference in best fit polynomial coefficients, c_i, is observed when those for a sampling frequency of 7 kilohertz (i.e. Equation 4.33) are compared with those for sampling frequencies of 10, 9 and 8 kilohertz (i.e. Equations 4.26, 4.31 and 4.32, respectively). Therefore, the relationship between α' and $R(1)/R(0)$ is only significantly affected for a waveform sampling frequency of less than 8 kilohertz.

The similarity of best fit polynomial coefficeints for sampling frequencies of 10, 9 and 8 kilohertz (see Equations 4.26, 4.31 and 4.32, respectively) suggests that a single best fit polynomial can be used to describe the relationship between α' and $R(1)/R(0)$. The significant spread of data points about the best fit polynomials is found to be larger than the variation in best fit polynomials for sampling frequencies of 10, 9 and 8 kilohertz; hence, it is not critical to know the best fit polynomial coefficients to a high degree of accuracy, i.e. two significant figures would be sufficient. In general, the relatively large spread of data points about the best fit polynomial indicates that the error in using a best fit polynomial to determine a value of α, for which the two/three adaptive pre-emphasis filter produces a minimum area distance, is much larger than the errors produced by changing the sampling frequency.

In conclusion, the investigations of this section have shown that sampling frequencies between 10 and 8 kilohertz, inclusive, do not have a significant effect on the relationship between α' and $R(1)/R(0)$. For sampling frequencies of less than 8 kilohertz, the relationship between α' and $R(1)/R(0)$ is affected slightly. However, the change in the relationship between α' and $R(1)/R(0)$ for all sampling frequencies considered is much smaller than the spread of data points $(\alpha', R(1)/R(0))$ used to generate this relationship. Hence, for sampling frequencies between 10 and 8 kilohertz, inclusive, a suitable relationship between α' and $R(1)/R(0)$ is

$$\frac{R(1)}{R(0)} = 3.217(\alpha') - 3.552(\alpha')^2 + 1.339(\alpha')^3 \qquad (4.34)$$

For a sampling frequency of 7 kilohertz, the relationship between α' and $R(1)/R(0)$ is

$$\frac{R(1)}{R(0)} = 3.025(\alpha') - 3.148(\alpha')^2 + 1.127(\alpha')^3 \qquad (4.35)$$

although the investigations in this section suggest that using Equation 4.34 in the 7 kilohertz sampling frequency case would produce similar values of α'.

4.4.3.2 Parameter β

This section investigates the effects on the relationship between β' and $R(1)/R(0)$, for the two/three adaptive pre-emphasis filter, when sampling frequencies of 9, 8 and 7 kilohertz are used. The glottal pulse waveforms used to generate the relationship between β' and $R(1)/R(0)$ are derived only from the published glottal waveforms of MONSEN and ENGEBRETSON [1977], ROTHENBERG [1973], SONDHI [1975], SUNBERG and GAUFFIN [1978], MILLER [1959], and FLANAGAN and LANDGRAF [1968]. The reason for not using glottal pulse waveforms derived from the glottal pulse models of ROSENBERG

[1971] and FANT [1979] is presented in Section 4.4.2.2. The same procedure is used to generate the relationships between β' and $R(1)/R(0)$ for the different sampling frequencies as was used in Section 4.4.2.2 for a sampling frequency of 10 kilohertz (see Figure 4.22).

A best fit function defines a relationship between β' and $R(1)/R(0)$ from the data points $\left(\beta', R(1)/R(0)\right)$, and the form of that function is a polynomial defined by

$$\beta' = \sum_{i=0}^{n} d_i \left(\frac{R(1)}{R(0)}\right)^i \qquad (4.36)$$

i.e. the same polynomial form as used in Section 4.4.2.2 for a sampling frequency of 10 kilohertz. The order, n, of the polynomial function is determined as that which minimizes the deviation between the data points and the polynomial function values. A least squares curve fitting algorithm [LEONARD 1965] is used to determine the polynomial function which best fits the data and, hence, is referred to as the best fit polynomial.

All the glottal pulse waveforms used to define the relationship between β' and $R(1)/R(0)$ in Section 4.4.2.2 are resampled at 9 kilohertz and, by using the procedure defined in Figure 4.22, a set of data points $\left(\beta', R(1)/R(0)\right)$ is generated and plotted in Figure 4.37. The best fit polynomial to the data points of Figure 4.37 is plotted in that figure, and determined by the least sqaures curve fitting algorithm [LEONARD 1965] as

$$\beta' = -294.826 + 295.625\frac{R(1)}{R(0)} \qquad (4.37)$$

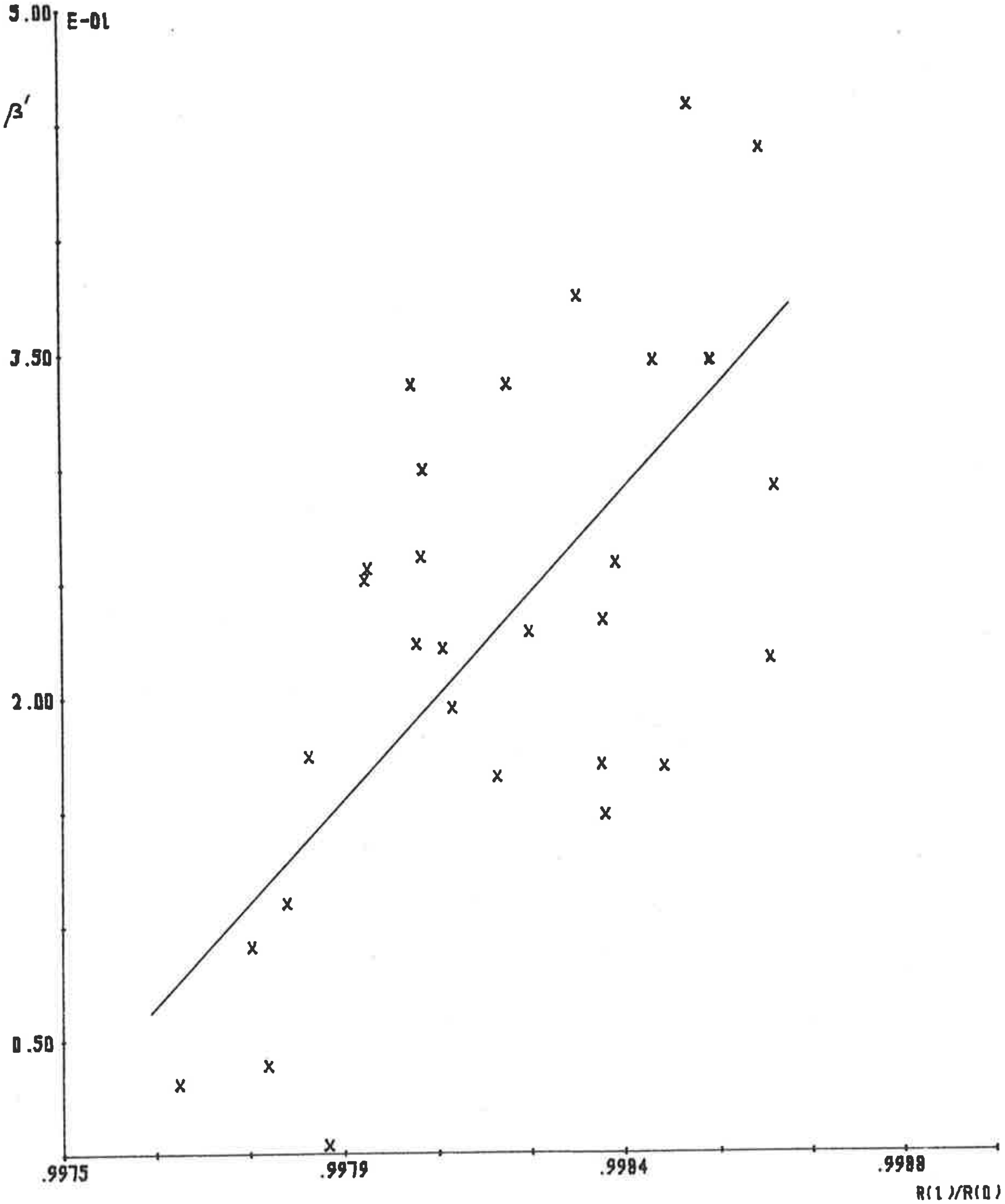


FIGURE 4.37: Data points (R(1)/R(0), β´) and the best fit polynomial, for glottal pulses sampled at a frequency of 9 kilohertz.

The transition point, which is the value of $R(1)/R(0)$ above which Equation 4.37 is used to define the two/three adaptive pre-emphasis filter is 0.99730. An alternative form of Equation 4.37, which increases the accuracy of the calculated β' for limited/small word length computations is

$$\beta' = 2.9563\ (R-0.7297) \qquad (4.38a)$$

where

$$R = \frac{\dfrac{R(1)}{R(0)} - 0.99}{100} \qquad (4.38b)$$

Resampling the glottal pulse waveforms used in Figure 4.37 (i.e. where a sampling frequency of 9 kilohertz is used) at a sampling frequency of 8 kilohertz, and using the procedure defined in Figure 4.22, a set of data points $(\beta', R(1)/R(0))$ is generated and plotted in Figure 4.38. Using the least squares fitting algorithm of LEONARD [1965], the best fit polynomial to the data points of Figure 4.38, which is plotted in that figure, is

$$\beta' = -341.561 + 342.549\frac{R(1)}{R(0)} \qquad (4.39)$$

The value of $R(1)/R(0)$ above which Equation 4.39 is used to define the two/three adaptive pre-emphasis filter, i.e. the transition point, is 0.99712. An alternative form of Equation 4.39 which increases the accuracy of β' calculated in situations where limited/small word length computations are performed is

$$\beta' = 3.4255\ (R -0.7116) \qquad (4.40a)$$

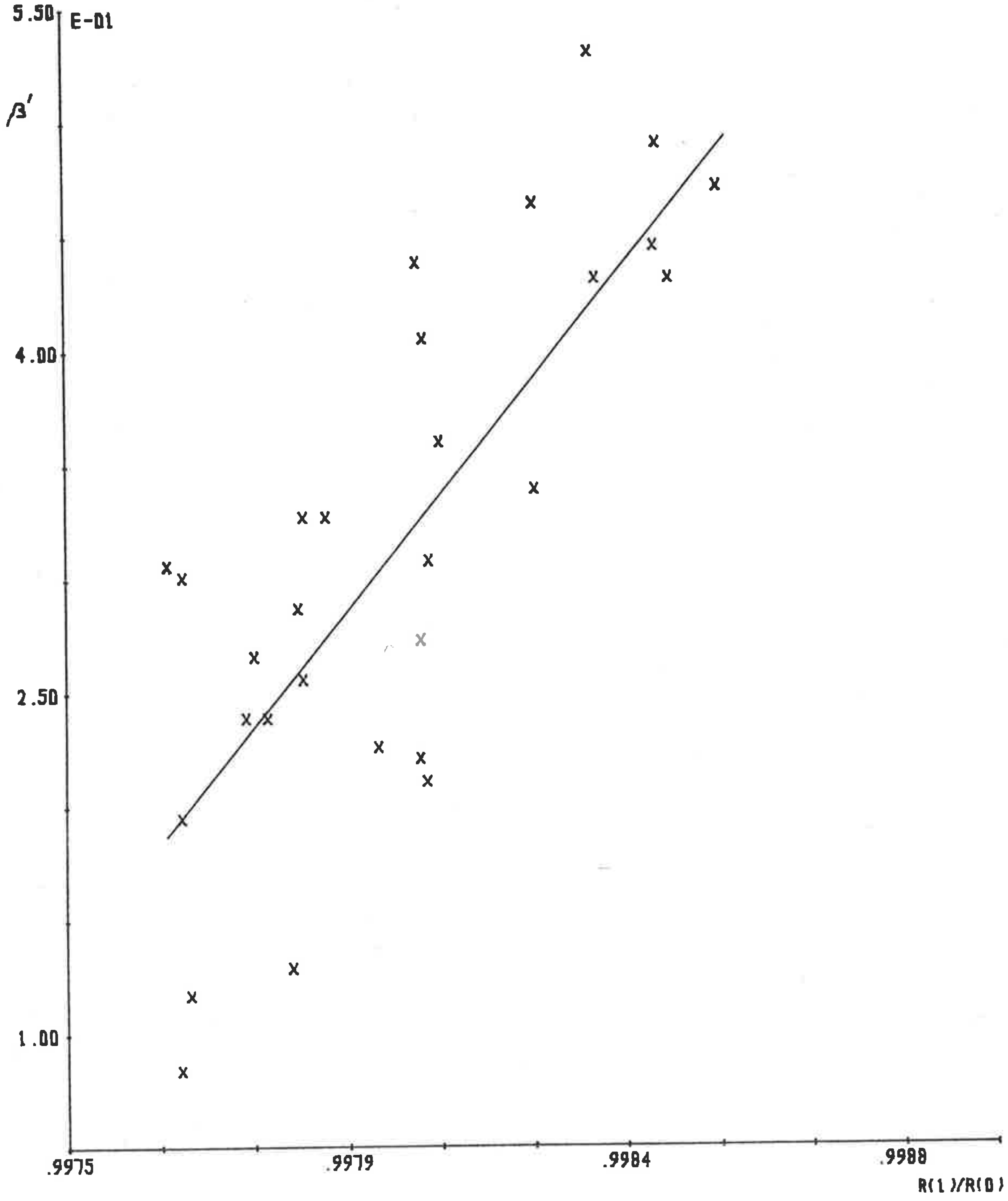


FIGURE 4.38: Data points (R(1)/R(0), β´) and the best fit polynomial, for glottal pulses sampled at a frequency of 8 kilohertz.

where

$$R = \frac{\frac{R(1)}{R(0)} - 0.99}{100} \qquad (4.40b)$$

Figure 4.39 presents the data points $\left(\beta', R(1)/R(0)\right)$ generated by sampling the glottal pulse waveforms used to generate Figures 4.33, 4.37 and 4.38, at a sampling frequency of 7 kilohertz, followed by the procedure defined in Figure 4.22. A best fit polynomial (which is plotted in Figure 4.39) to the data points of Figure 4.39 is obtained by using the least squares curve fitting algorithm of LEONARD [1965], and has the form

$$\beta' = -369.261 + 370.372\frac{R(1)}{R(0)} \qquad (4.41)$$

The value of $R(1)/R(0)$ above which the two/three adaptive preemphasis filter is defined by Equation 4.41, i.e. the transition point, is 0.99700. For situations where limited/small word lengths are used to perform computations, an alternative form of Equation 4.41 is

$$\beta' = 3.7037\ (R - 0.7000) \qquad (4.42a)$$

where

$$R = \frac{\frac{R(1)}{R(0)} - 0.99}{100} \qquad (4.42b)$$

The spread of data points about the best fit polynomial in Figures 4.37, 4.38 and 4.39 appears large, but is similar to the

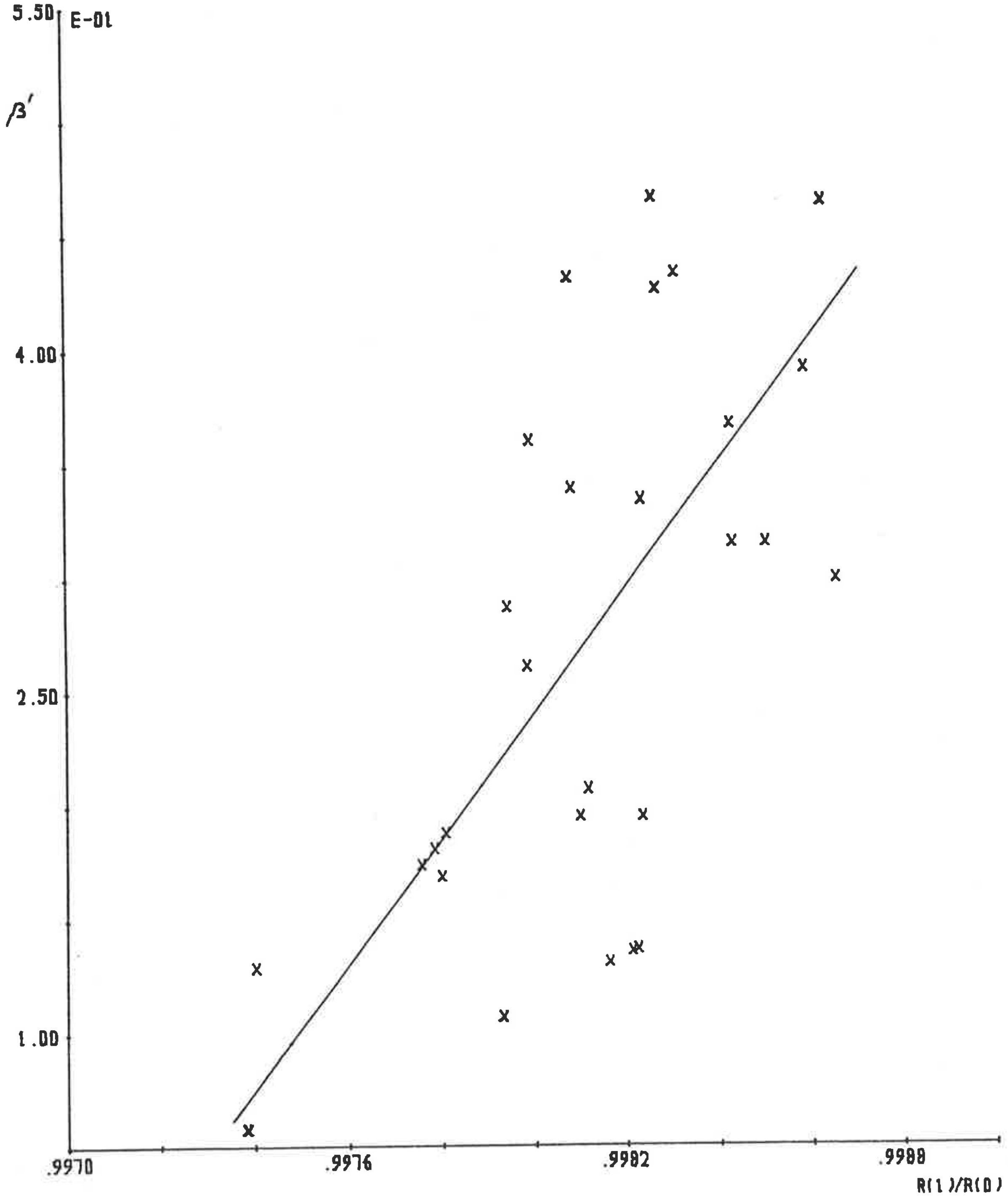


FIGURE 4.39: Data points (R(1)/R(0), β´) and the best fit polynomial, for glottal pulses sampled at a frequency of 7 kilohertz.

spread of data points observed in Figure 4.33 for the 10 kilohertz sampling frequency case. The full range of β' from zero to unity is not presented in Figures 4.37, 4.38 and 4.39 because of the lack of glottal pulse waveforms the spectral slope of which is approximately -18 dB per octave. Hence, an extrapolation of the best fit polynomials is used for values of β' that are greater than the range of β' presented in Figure 4.37, 4.38 and 4.39. For a particular glottal pulse waveform, the values of $R(1)/R(0)$ decrease with decreasing sampling frequency, and so the range of $R(1)/R(0)$ presented in Figure 4.37, 4.38 and 4.39 is smaller for smaller sampling frequencies.

A comparison of the best fit polynomials for sampling frequencies of 10, 9, 8 and 7 kilohertz is presented in Figure 4.40, with the data points used to define those best fit polynomials. The need for different relationships between β' and $R(1)/R(0)$ for different sampling frequencies is shown in Figure 4.40. In general, Figure 4.40 shows that, as the sampling frequency decreases, the slope of the best fit polynomial increases and the transition point (i.e. the value of $R(1)/R(0)$ when β' is zero) decreases in value.

4.4.4 THE TWO/THREE ADAPTIVE PRE-EMPHASIS FILTER

A new adaptive pre-emphasis filter, called the two/three adaptive pre-emphasis filter, has been defined to account for glottal pulse type excitations of acoustic tubes, and therefore provide improved acoustic tube shape recovery. The form of the two/three adaptive pre-emphasis filter is defined to overcome the inadequacies of conventional pre-emphasis filters but retain their advantages or successful features.

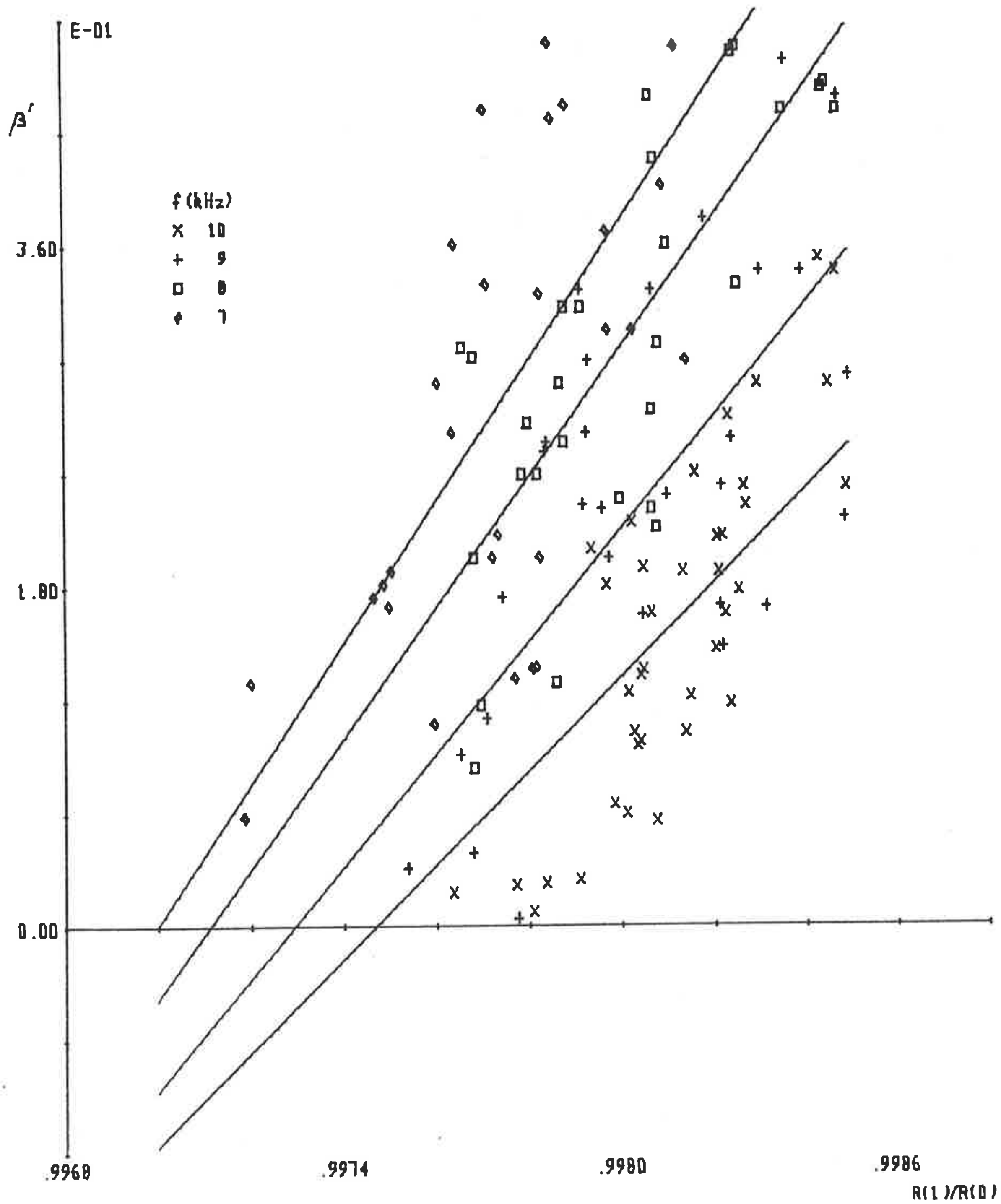


FIGURE 4.40: Data points (R(1)/R(0), β´) and the best fit polynomials, for glottal pulses sampled at frequencies of 10,9,8 and 7 kilohertz.

Experiments performed in the early part of this chapter showed that the recovered acoustic tube shape was close to the original acoustic tube shape when the spectral slope of the pre-emphasis filter and glottal pulse waveform have equal magnitudes but opposite signs. Therefore, the spectral slope variation of the two/three adaptive pre-emphasis filter must be at least the same as the observed spectral slope variation in glottal pulse waveforms, which is from -8 dB to -17 dB per octave. Hence, the form of the two/three adaptive pre-emphasis filter has two zeros when less than +12 dB per octave spectral correction is necessary, i.e.

$$(1 - \alpha z^{-1})^2 \qquad (4.43)$$

and three zeros, with two fixed, when between +12 dB and +18 dB per octave spectral correction is necessary, i.e.

$$(1 - z^{-1})^2 (1 - \beta z^{-1}) \qquad (4.44)$$

During the investigations of the conventional pre-emphasis techniques, it was found that a reasonable indication of the spectral slope of the glottal pulse waveform is given by the ratio $R(1)/R(0)$, which is evaluated from the autocorrelation function of the speech waveform. Using the ratio $R(1)/R(0)$ permits an automatic adaption to unvoiced and voiced sounds (an advantage of the adaptive pre-emphasis filters of MARKEL and GRAY [1974] and MAKHOUL and VISWANATHAN [1974]) to be incorporated into the two/three adaptive pre-emphasis filter.

The value of α or β used in the two/three adaptive pre-emphasis filter for a particular value of $R(1)/R(0)$ is obtained from empirical relationships between α and $R(1)/R(0)$ or β and $R(1)/R(0)$. If the two/three adaptive pre-emphasis filter completely accounts for a glottal pulse waveform, then an impulse results which, when analysed by a linear predictive analysis, recovers an acoustic tube with no cross-sectional area changes, called the ideal acoustic tube shape. Hence, the empirical relationships between α and $R(1)/R(0)$ and β and $R(1)/R(0)$ are determined from best fit curves to data points $(\alpha', R(1)/R(0))$ and $(\beta', R(1)/R(0))$, where α' and β' define the two/three adaptive pre-emphasis filter such that a glottal pulse waveform when pre-emphasized by the two/three adaptive pre-emphasis filter followed by a linear predictive analysis results in the smallest area distance between recovered and ideal acoustic tube shapes.

The glottal pulse waveforms used to define the relationship between α' and $R(1)/R(0)$ and β' and $R(1)/R(0)$ are generated from glottal pulse models or derived from measured glottal pulse waveforms published by various researchers. Due to difficulties in fitting curves to step slopes, the inverse relationship, i.e. $R(1)/R(0)$ in terms of α', was determined and used to generate a lookup table so that values of α' can be found for particular values of $R(1)/R(0)$.

For a waveform sampling frequency of 10 kilohertz, the relationship for $R(1)/R(0)$ in terms of α' was determined as

$$\frac{R(1)}{R(0)} = 3.217(\alpha') - 3.552(\alpha')^2 + 1.339(\alpha')^3 \qquad (4.45)$$

The relationship between β' and $R(1)/R(0)$ for a waveform sampling frequency of 10 kilohertz was found as

$$\beta' = -246.791 + 247.416\left(\frac{R(1)}{R(0)}\right) \qquad (4.46)$$

Equation 4.45 is used when $R(1)/R(0)$ is less than 0.99747, and Equation 4.46 is used when $R(1)/R(0)$ is greater than 0.99747. The value of $R(1)/R(0)$ equal to 0.99747 is called the transition value, or transition point.

Investigations were performed to determine if a change in waveform sampling frequency causes a change in the empirical relationships between α' and $R(1)/R(0)$ or β' and $R(1)/R(0)$. For sampling frequencies between 10 and 8 kilohertz, inclusive, it was shown that changes in the relationships between α' and $R(1)/R(0)$ were much smaller than the spread of data points used to define the relationships between α' and $R(1)/R(0)$. Hence, Equation 4.45 is used to define α' for a particular value of $R(1)/R(0)$ when the sampling frequency is less than or equal to 10 kilohertz, but greater than or equal to 8 kilohertz. For a sampling frequency of 7 kilohertz, the relationship between α' and $R(1)/R(0)$ is significantly different from the relationship for sampling frequencies between 10 and 8 kilohertz, inclusive. Therefore, if the sampling frequency is 7 kilohertz, then the relationship used to determine α' from a particular value of $R(1)/R(0)$ is

$$\frac{R(1)}{R(0)} = 3.025(\alpha') - 3.148(\alpha')^2 + 1.127(\alpha')^3 \qquad (4.47)$$

The relationship between β' and $R(1)/R(0)$ was found to change significantly as the sampling frequency changed. At sampling frequencies of 9, 8 and 7 kilohertz, the empirical relationships between β' and $R(1)/R(0)$ were determined to be

$$\beta' = -294.826 + 295.625\frac{R(1)}{R(0)} \qquad (4.48)$$

for a sampling frequency of 9 kilohertz,

$$\beta' = -341.561 + 342.549\frac{R(1)}{R(0)} \qquad (4.49)$$

for a sampling frequency of 8 kilohertz, and

$$\beta' = -369.261 + 370.372\frac{R(1)}{R(0)} \qquad (4.50)$$

for a sampling frequency of 7 kilohertz.

Since the relationship between β' and $R(1)/R(0)$ changes with sampling frequency, the transition point, at which the two/three adaptive pre-emphasis filter changes from being defined by the parameter α to being defined by the parameter β or vice versa, also changes. For sampling frequencies of 9, 8 and 7 kilohertz, the transition points were found to be 0.99730, 0.99712 and 0.99700, respectively.

4.5 SUMMARY

This chapter has investigated the excitation and recovery of a set of acoustic tubes by glottal pulse waveforms similar to those used to excite the vocal tract for the production of voiced speech. The glottal pulse waveforms were generated from glottal pulse models, or derived from glottal pulse waveforms measured by a number of speech researchers. A wide range of glottal pulse shapes were used so that the wide range of glottal pulse shapes that occur for the numerous voiced sounds and different speakers are considered.

The effects of glottal pulse waveform excitation on the acoustic tube shape recovered by a linear predictive analysis were presented, and it was shown that gross errors in acoustic tube shape recovery occur. These results support similar research presented by other speech researchers; therefore, there exists a need for a procedure or method which can account for glottal pulse excitation if accurate acoustic tube shape recovery is to be achieved.

A number of conventional means to overcome glottal pulse excitation effects on the recovered acoustic tube shape were investigated. Closed glottis interval analysis was discussed, and the high dependence of an accurate acoustic tube shape recovery on accurate determination of the closed glottis interval was shown. Other disadvantages of the closed glottis interval analysis were discussed, which included the inapplicability of the analysis method for weakly voiced and also high pitched sounds.

Conventional pre-emphasis techniques were discussed, and evaluated with synthetic speech waveforms to determine their effectiveness at removing glottal pulse excitation effects from the recovered acoustic tube shape. A constant +12 dB per octave pre-emphasis was found to provide improved acoustic tube shape recovery only when the glottal pulse waveform, used to excite the acoustic tube shape, has a spectral slope of approximately -12 dB per octave. Since glottal pulse waveforms have spectral slopes that vary widely from -12 dB per octave, a constant +12 dB per octave pre-emphasis does not, in general, provide good acoustic tube shape recovery.

An adaptive pre-emphasis filter, based on that presented by MARKEL and GRAY [1974] and MAKHOUL and VISWANATHAN [1974] which uses the ratio $R(1)/R(0)$ to define the amount of pre-emphasis applied to the acoustic waveform before analysis, was evaluated. An improvement in acoustic tube shape recovery was observed when compared with the acoustic tube shapes recovered by using no pre-emphasis or a +12 dB per octave pre-emphasis but, in general, good acoustic tube shape recovery does not occur. The adaptive pre-emphasis filter was shown to cause the lack of improvement in acoustic tube shape recovery, rather than the concept of using $R(1)/R(0)$ as an indicator for the amount of spectral slope correction that should be applied to the acoustic waveform before analysis.

Another adaptive pre-emphasis filter, that of NAKAJIMA et al [1974], was evaluated with synthetic speech waveforms. This adaptive pre-emphasis filter consists of second and third order critical damping digital filters in cascade, with the necessity of solving polynomials of order three and five to define the two digital filters. Evaluation of the adaptive pre-emphasis filter of NAKAJIMA et al [1974] showed that poor acoustic tube shape recovery resulted, which was similar to that obtained by using no pre-emphasis.

It was concluded that, in general, none of the available pre-emphasis techniques could guarantee good acoustic tube shape recovery when acoustic tubes are excited by glottal pulse waveforms. However, during the evaluation of the available pre-emphasis techniques, it was found that good acoustic tube shape recovery may be achieved if a spectral correction is used which is equal, but opposite in sign, to the spectral slope of the glottal pulse wavefrom. The value of $R(1)/R(0)$ was found to provide a useful indication of the glottal pulse waveform's spectral slope.

A new adaptive pre-emphasis filter, called the two/three adaptive pre-emphasis filter, has been defined to provide improved acoustic tube shape recovery when a set of acoustic tubes is excited by glottal pulse waveforms. The two/three adaptive pre-emphasis filter consists of two parts, one part defined by a parameter α corrects for glottal pulse spectral slopes between zero and -12 dB per octave, and the other part defined by the parameter β corrects for glottal pulse spectral slopes between -12 dB and -18 dB per octave. The values of α and β used in a particular situation are obtained from empirical expressions between parameter α and $R(1)/R(0)$, and parameter β and $R(1)/R(0)$. These empirical expressions are determined by finding a best fit curve to the data points $(\alpha', R(1)/R(0))$ and $(\beta', R(1)/R(0))$, where α' and β' define the two/three adaptive pre-emphasis filter such that a two/three adaptive pre-emphasis of the glottal pulse waveform, followed by a linear predictive analysis, results in the smallest area distance between recovered and original acoustic tube shapes.

A number of different waveform sampling frequencies between 10 and 7 kilohertz, inclusive, were used, and shown to affect the relationships between α' and $R(1)/R(0)$ and β' and $R(1)/R(0)$ to various degrees. The relationship between α' and $R(1)/R(0)$ only changed significantly when a waveform sampling frequency of 7 kilohertz was used, whereas the relationship between β' and $R(1)/R(0)$ changed significantly as the waveform sampling frequency changed.

CHAPTER 5

EVALUATION OF THE TWO/THREE ADAPTIVE PRE-EMPHASIS FILTER

5.1 INTRODUCTION

Investigations performed in Chapter 4 lead to the definition of a new adaptive pre-emphasis filter, called the two/three adaptive pre-emphasis filter, to remove glottal pulse excitation effects from an acoustic tube shape recovered by a linear predictive analysis. This chapter describes the evaluation of the two/three adaptive pre-emphasis filter to determine if improved acoustic tube/vocal tract shape recovery is achieved.

The two/three adaptive pre-emphasis filter is defined by two parameters, α and β, which are, in turn, evaluated via a relationship with $R(1)/R(0)$ of the waveform being analysed. Since the two parameters α and β and their relationships with $R(1)/R(0)$ are completely independent of one another, the evaluation of each is performed spearately. The waveform sampling frequency was shown in Chapter 4 to have a small effect on the relationships between the parameter α and $R(1)/R(0)$, and the parameter β and $R(1)/R(0)$.

The performance of the two/three adaptive pre-emphasis filter is compared with that of existing pre-emphasis techniques. For this comparison, no pre-emphasis, a constant +12 dB per octave pre-emphasis (as used in most speech analysis techniques) and the unvoiced/voiced adaptive pre-emphasis filter derived from the work of GRAY and MARKEL [1974] and MAKHOUL and VISWANATHAN

[1974] (see Chapter 4) are used. The desired goal of pre-emphasis techniques is to provide accurate acoustic tube shape recovery, and so the effectiveness of a pre-emphasis technique is determined by comparing recovered and original acoustic tube shapes. Comparisons of acoustic tube shapes are primarily made by the area distance measure defined in Chapter 2, and secondly from plots of the acoustic tube shapes when the acoustic tube shapes are similar or the area distance measure produces ambiguous results.

Evaluation of the two/three adaptive pre-emphasis filter is firstly performed with glottal pulse waveforms generated from the glottal pulse models of ROSENBERG [1971] and FANT [1979], and then with measured glottal pulse waveforms from published papers [ROSEN-1973, MONSEN and ENGEBRETSON 1977, MILLER 1959, SONDHI 1975, and SUNDBERG and GAUFFIN 1978]. The glottal pulse waveforms derived from published papers are digitized by using a data tablet under the control of a mini-computer to oversee and store the digitized data. The separation of waveform points which are digitized is determined by the sampling period or frequency required. Where possible, the glottal pulse waveforms used in this chapter are different from those used to determine the relationships between α and $R(1)/R(0)$ and β and $R(1)/R(0)$.

Following the evaluation of the two/three adaptive pre-emphasis filter with glottal pulse waveforms, i.e. in Section 5.2, an evaluation using synthetic speech data is presented in Section 5.3. The generation of the synthetic speech is detailed in Appendix B, with the glottal pulse waveforms of Section 5.2 being used as the excitation. Synthetic speech provides an accurate evaluation, because the original acoustic tube shape is known, therefore a comparison of recovered and original acoustic tube shapes is possible. To ensure that the results of the evaluations using synthetic speech are consistent with evaluations using real speech, the acoustic tube shapes used to generate synthetic speech are derived from real vocal tract shapes, as presented in Appendix C.

It is difficult to evaluate the two/three adaptive pre-emphasis filter with real speech, as the vocal tract shape used to produce a particular speech sound is difficult to obtain. However, an indication of an improvement in vocal tract shape recovery is obtained by comparing the recovered acoustic tube shape with measured vocal tract shapes for similar speech sounds. In this chapter, where an evaluation with real speech is performed, i.e. Section 5.4, recovered acoustic tube shapes are compared with vocal tract shapes measured by FANT [1960] (see Appendix C).

The evaluation results presented in this chapter are very detailed, and attempt to cover a wide range of situations and conditions that occur for five vowel sounds. If the reader does not wish to read the detailed evaluations presented in the following sections, it is suggested that he skip to the summary, where all the evaluation results are discussed and a table summarizing the evaluation results is presented.

5.2 EVALUATION WITH GLOTTAL PULSES

The design process used in Chapter 4 tailored the two/three adaptive pre-emphasis filter to the properties of glottal pulse type excitation waveforms. This approach was used so that the two/three adaptive pre-emphasis filter could remove the glottal pulse waveform properties from the speech waveform and so provide improved vocal tract shape recovery. Therefore, it is appropriate to perform the initial evaluation of the two/three adaptive pre-emphasis filter with glottal pulse waveforms.

All the glottal pulse waveforms used in this section are either synthetic or derived from measured glottal pulse waveforms [ROSENBERG 1973, MONSEN and ENGEBRETSON 1977, MILLER 1959, SONDHI 1975, SUNDBERG and GAUFFIN 1978]. The synthetic glottal pulses are generated from the glottal pulse models of ROSENBERG [1971] and FANT [1979]. Unless otherwise indicated, all the glottal pulse waveforms used in this section are different from those used in Chapter 4, where the two/three adaptive pre-emphasis filter was defined.

The process used to evaluate the two/three adaptive pre-emphasis filter in this section firstly passes the glottal pulse waveform through the adaptive pre-emphasis filter. The resulting output from the filter is then analysed by a Parcor analysis procedure, i.e. an autocorrelation formulation of linear prediction, to produce a recovered acoustic tube shape. If the two/three adaptive pre-emphasis filter completely corrects for the glottal pulse waveform, then the output waveform from the filter is an impulse, from which a Parcor analysis recovers an acoustic tube shape with no change in cross-sectional area over its length.

Hence, the effectiveness of the two/three adaptive pre-emphasis filter is measured by the similarity of the recovered acoustic tube shape (generated by the above process) to one with no cross-sectional area changes over its length. The area distance measure described in Section 2.5 is used to measure the similarity of acoustic tube shapes.

Two conventional pre-emphasis techniques are used to provide a comparison for the results of the two/three adaptive pre-emphasis evaluation and, therefore, to determine if any improvement in acoustic tube shape recovery is achieved by using the two/three adaptive pre-emphasis filter. The conventional pre-emphasis techniques are a +12 dB per octave pre-emphasis and/or the unvoiced/voiced adaptive pre-emphasis filter derived from the work of GRAY and MARKEL [1974] and MAKHOUL and VISWANATHAN [1974]. A discussion of both these conventional pre-emphasis techniques is presented in Section 4.3.

The two/three adaptive pre-emphasis filter consists of two independent parts, as defined in Section 4.4.4. Whenever the glottal pulse spectral slope is greater than approximately -12 dB per octave, then a parameter α defines the two/three adaptive pre-emphasis filter. For a glottal pulse spectral slope less than -12 dB per octave, the two/three adaptive pre-emphasis filter is defined by a parameter β. Section 5.2.1 contains the evaluation with glottal pulse waveforms of the two/three adaptive pre-emphasis filter when it is defined by the parameter α. The evaluation of the two/three adaptive pre-emphasis filter when defined by the parameter β is presented in Section 5.2.2.

For the evaluations performed in Sections 5.2.1 and 5.2.2, a sampling frequency of 10 kilohertz is used. Section 4.4.3 showed that different sampling frequencies had an effect on the form of the two/three adaptive pre-emphasis filter. Section 5.2.3 presents an evaluation of the two/three adaptive pre-emphasis filter with glottal pulse waveforms for the sampling frequencies of 9, 8 and 7 kilohertz.

5.2.1 PARAMETER α

This section evaluates the two/three adaptive pre-emphasis filter with glottal pulse waveforms when the filter is defined by the parameter α. The value of α used in each case is obtained from the ratio $R(1)/R(0)$, which is evaluated from the glottal pulse waveform, as described in Chapter 4 and Appendix D.

Figure 5.1 presents a plot of the area distances between the recovered and ideal (i.e. one with no cross-sectional area changes) acoustic tube shapes for no pre-emphasis, a constant +12 dB per octave pre-emphasis, an unvoiced/voiced adaptive pre-emphasis, and a two/three adaptive pre-emphasis of glottal pulse waveforms, versus the value of $R(1)/R(0)$ within the range of 0.95 to near unity. The glottal pulse waveforms used to generate the data points in Figure 5.1 are derived from the glottal pulse models B, C and E of ROSENBERG [1971].

Figure 5.1 shows the results of a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis to be similar. This is expected, since the unvoiced/voiced adaptive pre-emphasis filter was designed to provide a +12 dB per octave pre-

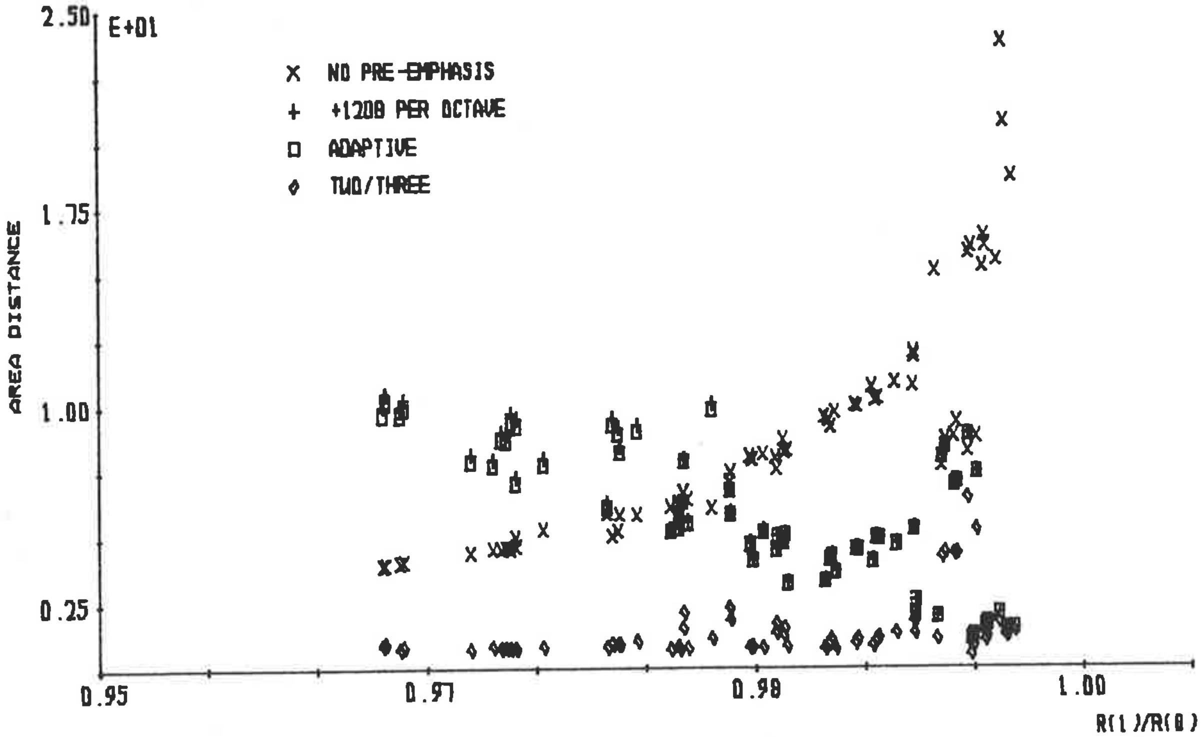


FIGURE 5.1: Area distances between recovered and ideal acoustic tube shapes for analysis of glottal pulse waveforms from the ROSENBERG glottal pulse model.

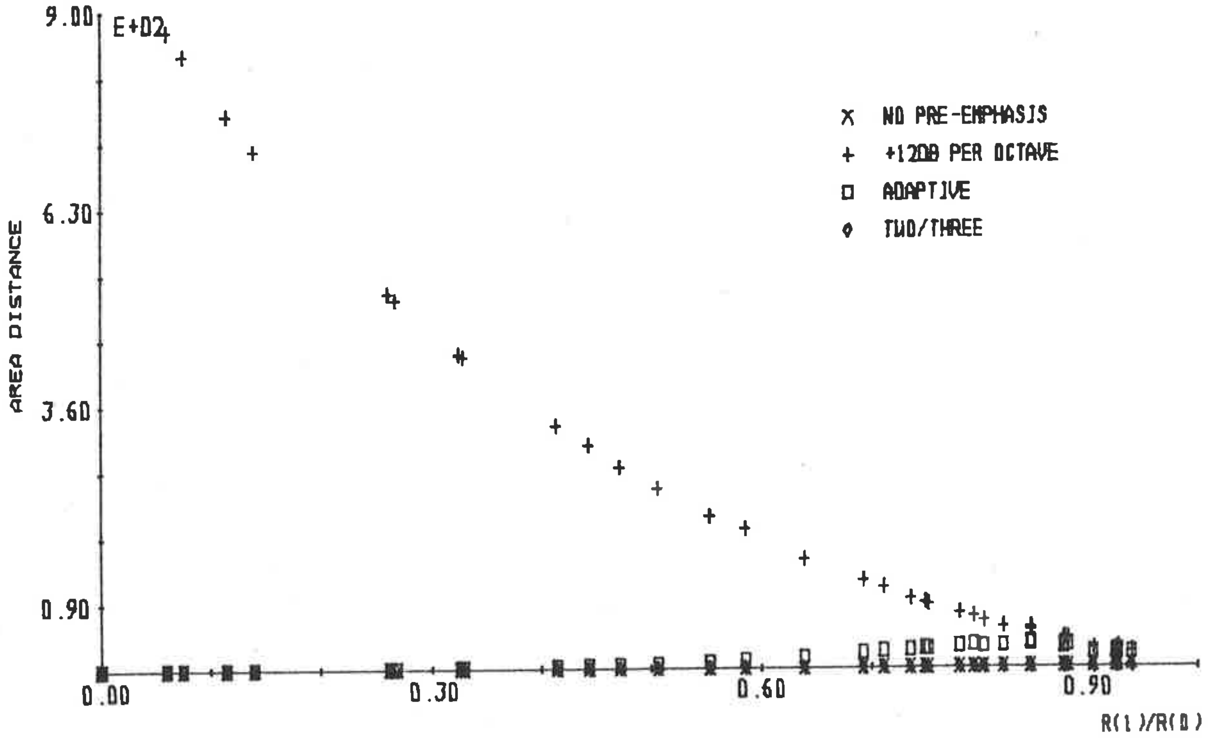


FIGURE 5.2: Area distances between recovered and ideal acoustic tube shapes for analysis of glottal pulse waveforms from the ROSENBERG glottal pulse model.

emphasis when $R(1)/R(0)$ is near unity. The constant +12 dB per octave and unvoiced/voiced adaptive pre-emphases only provide a reduction in area distances, when compared with no pre-emphasis, for $R(1)/R(0)$ close to unity. For all the results presented in Figure 5.1, the two/three adaptive pre-emphasis produces the smallest area distances.

Figure 5.2 plots the area distances between recovered and ideal acoustic tube shapes for no pre-emphasis, a +12 dB per octave pre-emphasis, and unvoiced/voiced and two/three adaptive pre-emphases of glottal pulse waveforms derived from ROSENBERG's [1971] glottal pulse models B, C and E, versus $R(1)/R(0)$ in the range zero to 0.95. The area distances for a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis are seen in Figure 5.2 to be much larger than for no pre-emphasis or a two/three adaptive pre-emphasis. Therefore, Figure 5.3 presents the data points of Figure 5.2 for only no pre-emphasis and a two/three adaptive pre-emphasis.

Observation of Figure 5.2 shows that a considerable reduction in area distances and, hence, effective removal of glottal pulses is achieved by the two/three adaptive pre-emphasis in comparison with the other pre-emphases considered, when $R(1)/R(0)$ is between zero and 0.95. In comparison with no pre-emphasis, the two/three adaptive pre-emphasis is shown in Figure 5.3 to produce a large reduction in area distances when $R(1)/R(0)$ is near 0.95, and similar area distances when $R(1)/R(0)$ is near zero.

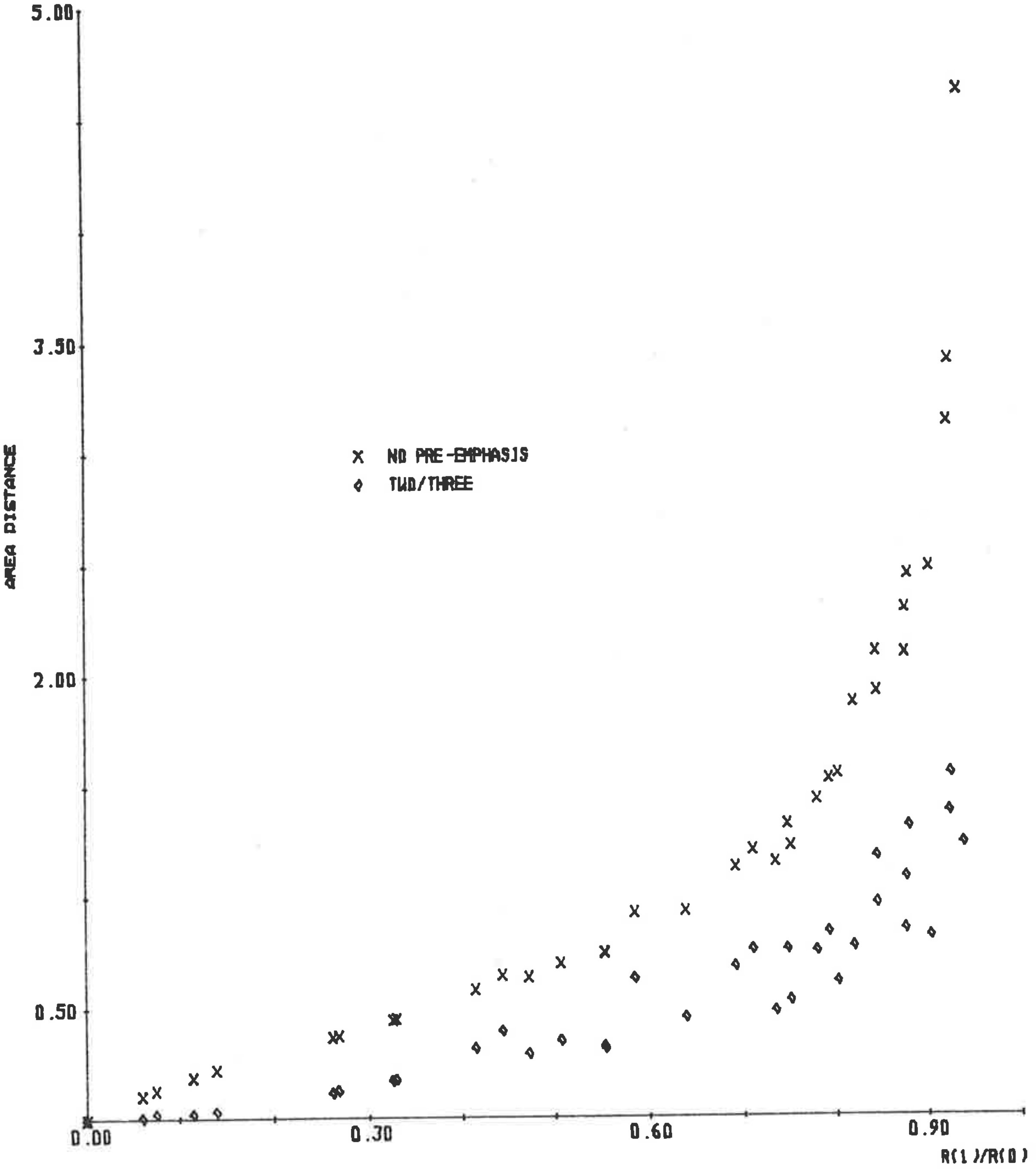


FIGURE 5.3: Area distances between recovered and ideal acoustic tube shapes for no pre-emphasis and a two/three adaptive pre-emphasis of glottal pulses from the ROSENBERG glottal pulse model.

Hence, for the range of $R(1)/R(0)$ from zero to unity, the two/three adaptive pre-emphasis of the glottal pulse waveforms derived from ROSENBERG's glottal pulse models B, C and E, provides a significant reduction in area distance when compared with no pre-emphasis. Over the same range of $R(1)/R(0)$, a very large reduction of area distances is achieved by a two/three adaptive pre-emphasis in comparison with a constant +12 dB per octave and unvoiced/voiced adaptive pre-emphases.

Figure 5.4 presents a plot of area distances between recovered and ideal acoustic tube shapes for no pre-emphasis, a constant +12 dB per octave pre-emphasis, and unvoiced/voiced and two/three adaptive pre-emphases of glottal pulse waveforms derived from FANT's [1979] glottal pulse model, versus $R(1)/R(0)$ in the range 0.95 to unity. The constant +12 dB per octave pre-emphasis and the unvoiced/voiced adaptive pre-emphasis produce similar area distances in Figure 5.4, as observed for glottal pulse waveforms derived from ROSENBERG's glottal pulse models. The constant +12 dB per octave pre-emphasis and unvoiced/voiced adaptive pre-emphasis only provide smaller area distances than no pre-emphasis when $R(1)/R(0)$ is close to unity, as observed in Figure 5.1.

Figure 5.5 plots the area distances between recovered and ideal acoustic tube shapes for no pre-emphasis, an unvoiced/voiced adaptive pre-emphasis and a two/three adaptive pre-emphasis of glottal pulse waveforms derived from FANT's glottal pulse model, versus $R(1)/R(0)$ in the range zero to 0.95. Area distances for a +12 dB per octave pre-emphasis are not presented, as they are much larger than the area distances of Figure 5.5, similar to those for the ROSENBERG glottal pulses in Figure 5.2. Figure

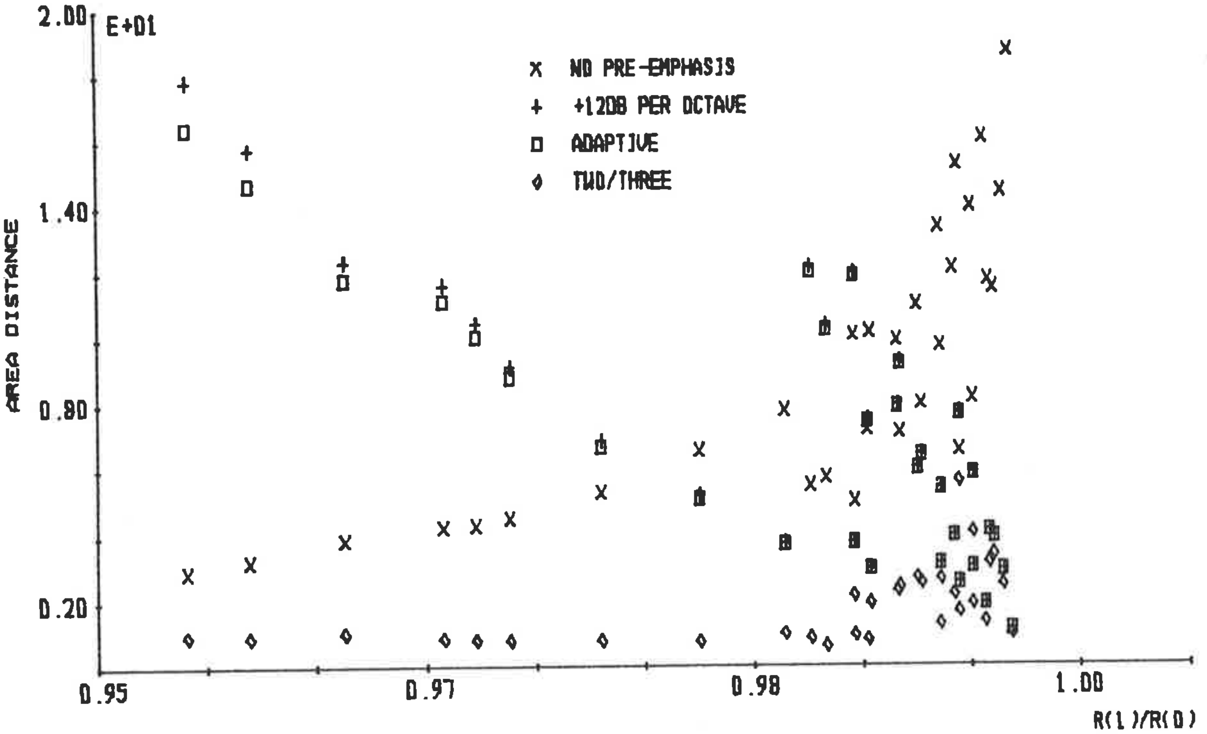


FIGURE 5.4: Area distances between recovered and ideal acoustic tube shapes for an analysis of glottal pulses from the FANT glottal pulse model.

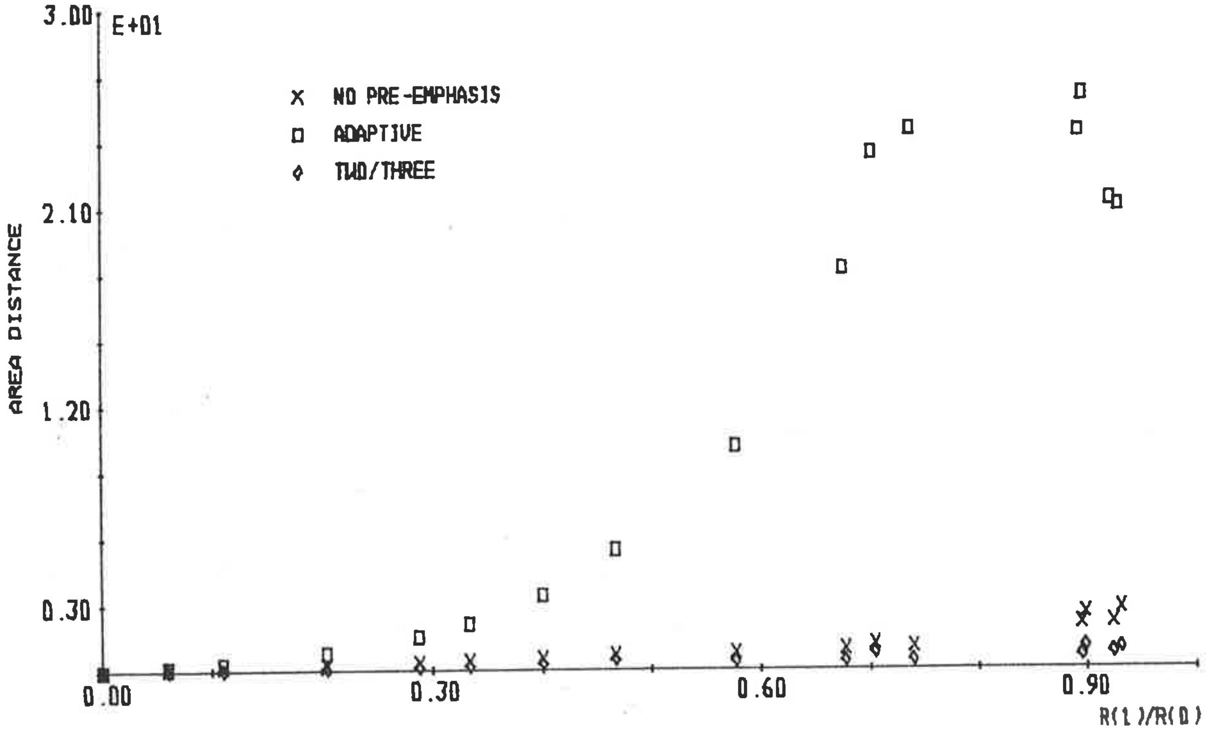


FIGURE 5.5: Area distances between recovered and ideal acoustic tube shapes for an analysis of glottal pulses from the FANT glottal pulse model.

5.5 shows that the area distances for the unvoiced/voiced adaptive pre-emphasis are much larger than for no pre-emphasis and a two/three adaptive pre-emphasis, except when they are similar for $R(1)/R(0)$ close to zero. The area distances for the two/three adaptive pre-emphasis are the smallest of all the pre-emphases considered.

Evaluations with glottal pulse waveforms derived from the glottal pulse models of ROSENBERG [1971] and FANT [1979] have shown that a two/three adaptive pre-emphasis filter provides consistently smaller area distances than the other pre-emphasis techniques considered, including no pre-emphasis, for $R(1)/R(0)$ ranging from zero to near unity. Over the same range of $R(1)/R(0)$, the existing pre-emphases, i.e. +12 dB per octave and unvoiced/voiced adaptive, produce much larger area distances than no pre-emphasis and the two/three adaptive pre-emphasis, with the only exceptions being when $R(1)/R(0)$ is close to zero or unity.

An evaluation of the two/three adaptive pre-emphasis filter is performed with a number of glottal pulse waveforms derived from real glottal pulse waveforms measured by ROSENBERG [1973], MONSEN and ENGEBRETSON [1977], MILLER [1959], SONDHI [1975], and SUNDBERG and GAUFFIN [1978]. The area distances versus $R(1)/R(0)$ for no pre-emphasis, a +12 dB per octave and unvoiced/voiced and two/three adaptive pre-emphases of glottal pulses derived from the above sources are presented in Figure 5.6.

When $R(1)/R(0)$ is near 0.998, Figure 5.6 shows that the +12 dB per octave and unvoiced/voiced adaptive pre-emphases provide a significant reduction in area distances when compared with no

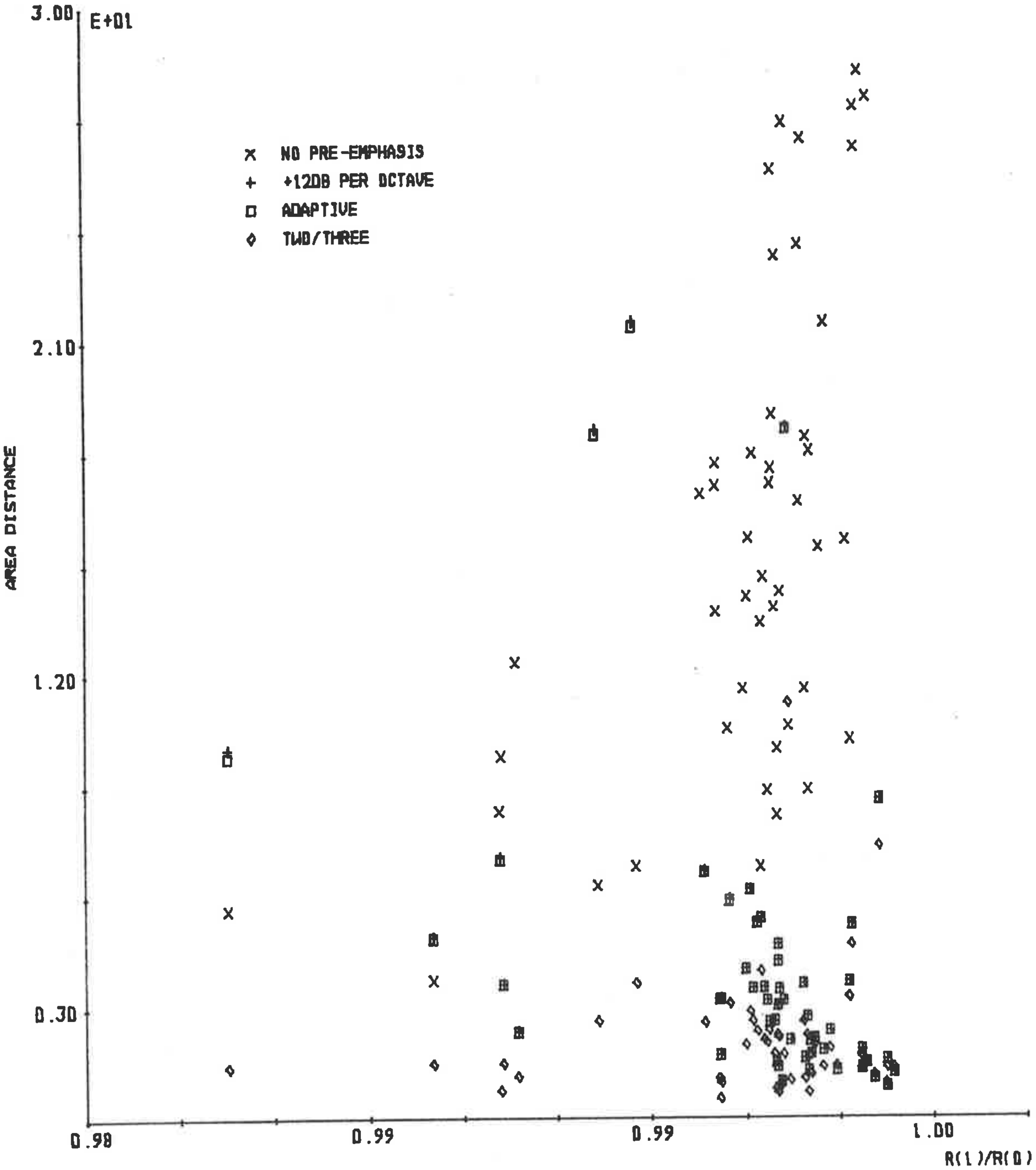


FIGURE 5.6: Area distances between recovered and ideal acoustic tube shapes for analysis of real glottal pulses.

pre-emphasis. However, as previously observed, the reduction in area distance decreases as $R(1)/R(0)$ decreases, until $R(1)/R(0)$ is near 0.95, when the area distances for no pre-emphasis are smaller. For the two/three adaptive pre-emphasis, a significant reduction in area distances is observed in Figure 5.6, in comparison with the other pre-emphasis techniques, including no pre-emphasis.

A comparison of the results obtained by using different glottal pulse models and glottal pulse waveforms derived from measured glottal pulses, i.e. a comparison of Figures 5.1 to 5.6, shows a consistency of general trends. This suggests that the evaluations presented in this section can be considered as a reliable representation of general trends that occur for all glottal pulse excitation waveforms when pre-emphasized by the techniques considered in this section.

The evaluations presented in this section have shown that the two/three adaptive pre-emphasis filter consistently provides a significant reduction in area distances in comparison with no pre-emphasis, a constant +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis of glottal pulse waveforms derived from glottal pulse models and measured glottal pulse waveforms. Both the constant +12 dB per octave pre-emphasis and the unvoiced/voiced adaptive pre-emphasis were shown to produce much larger area distances than no pre-emphasis and the two/three adaptive pre-emphasis, except when $R(1)/R(0)$ is near zero or unity.

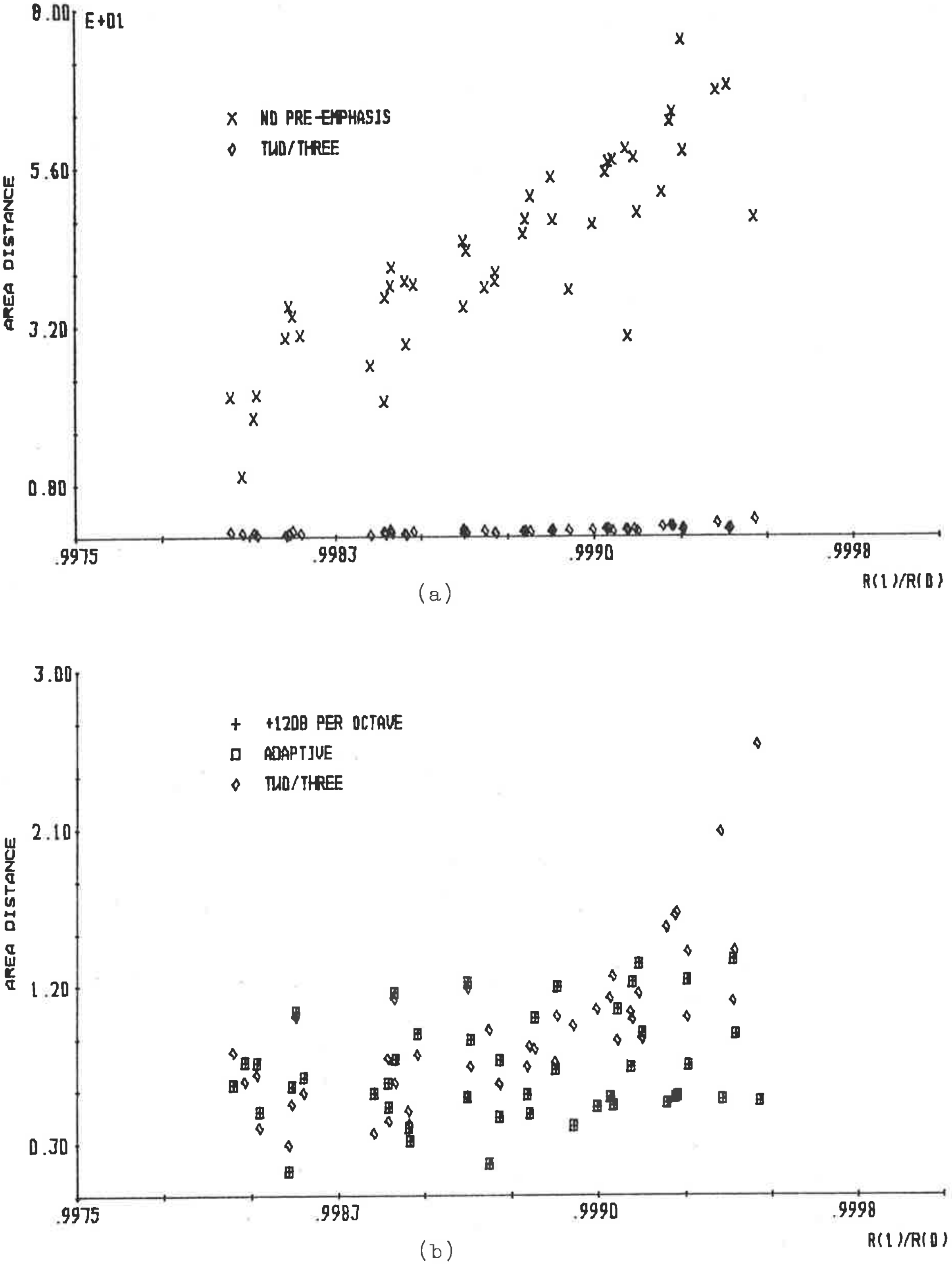
5.2.2 PARAMETER β

This section evaluates the two/three adaptive pre-emphasis filter with glottal pulse waveforms when the filter is defined by the parameter β. The value of β used in each case is obtained from the ratio $R(1)/R(0)$, which is evaluated from the glottal pulse waveform, as described in Chapter 4 and Appendix D.

Figure 5.7 presents the area distances between recovered and ideal (i.e. an acoustic tube shape with no cross-sectional area change) acoustic tube shapes for no pre-emphasis, a constant +12 dB per octave pre-emphasis, an unvoiced/voiced adaptive pre-emphasis and a two/three adaptive pre-emphasis of glottal pulse waveforms versus $R(1)/R(0)$. All the glottal pulse waveforms used to generate the data points of Figure 5.7 are derived from the ROSENBERG [1971] glottal pulse models B, C and E.

The area distances for no pre-emphasis of the glottal pulse waveforms are found in Figure 5.7 to be much larger than the area distances for the other pre-emphases considered in Figure 5.7. The area distances for the two/three adaptive pre-emphasis filter are shown in Figure 5.7 to be smaller than a constant +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis, when $R(1)/R(0)$ is near 0.9978, i.e. the smallest value of $R(1)/R(0)$ considered in Figure 5.7.

The range of $R(1)/R(0)$ considered in Figure 5.7 is close to unity, and so the unvoiced/voiced adaptive pre-emphasis filter provides approximately +12 dB per ocatve pre-emphasis; therefore, the area distances for a constant +12 dB per octave and unvoiced/voiced adaptive pre-emphases are similar in Figure 5.7. The evaluation of



(a)

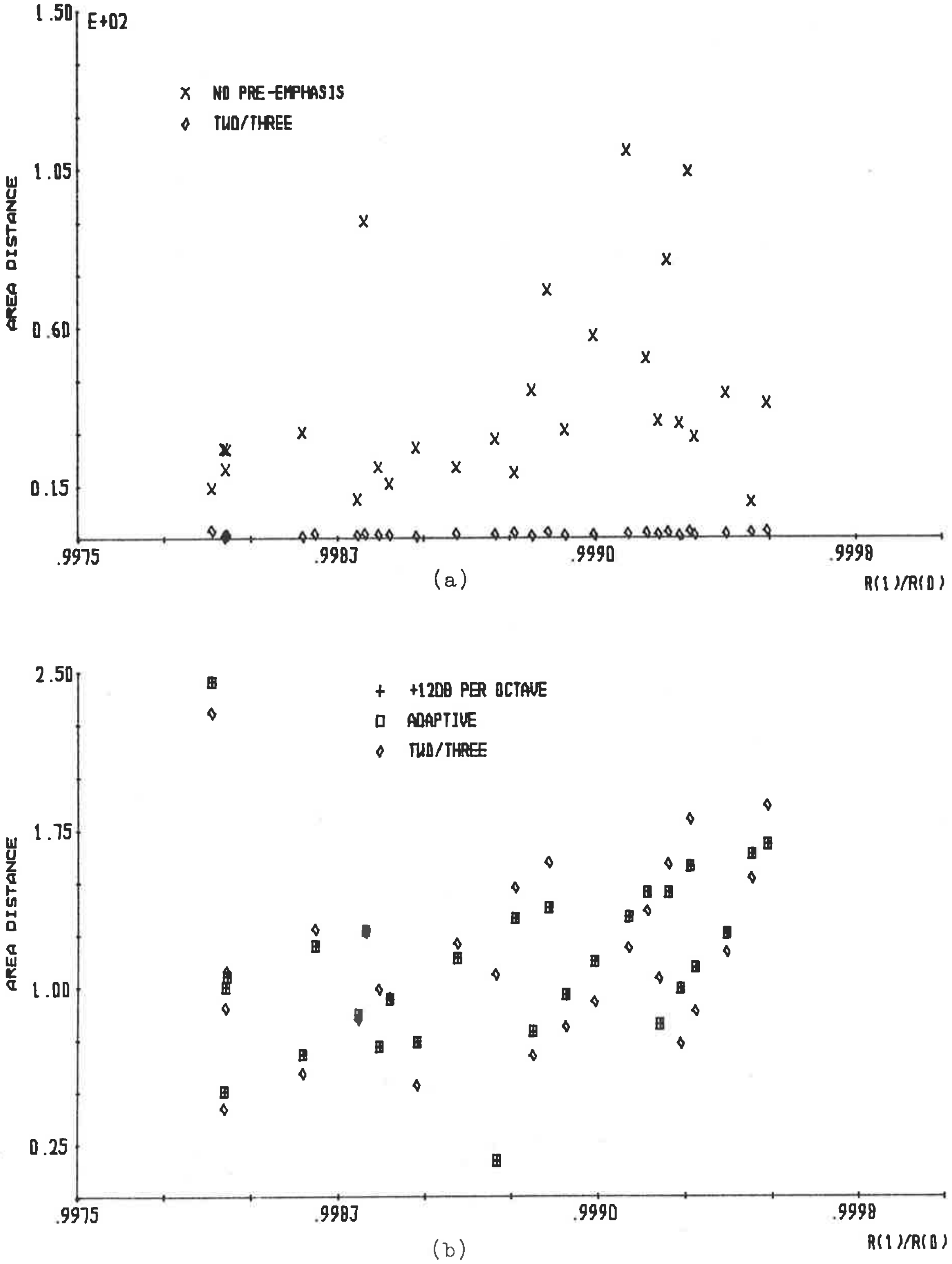
(b)

FIGURE 5.7: (a) Area distances for no pre-emphasis and a two/three adaptive pre-emphasis, and (b) area distances for a +12dB per octave, an unvoiced/voiced adaptive and two/three adaptive pre-emphases, of glottal pulses from the ROSENBERG glottal pulse model.

the two/three adaptive pre-emphasis filter with glottal pulse waveforms from ROSENBERG's glottal pulse models has shown that a reduction in area distances occurs in comparison with existing pre-emphasis techniques only when $R(1)/R(0)$ is near the transition value, i.e. the value of $R(1)/R(0)$ at which the two/three adaptive pre-emphasis filter changes from being defined by α to β or vice versa.

A plot of area distances versus $R(1)/R(0)$ for no pre-emphasis, a +12 dB per octave and unvoiced/voiced and two/three adaptive pre-emphases of glottal pulse waveforms derived from the glottal pulse model of FANT [1979] are presented in Figure 5.8. The area distances for no pre-emphasis of the glottal pulse waveforms are seen to be much larger than any of the other pre-emphases in Figure 5.8. This is consistent with the evaluation results presented in Figure 5.7, where glottal pulse waveforms derived from ROSENBERG's glottal models were used.

In general, Figure 5.8(b) shows that a two/three adaptive pre-emphasis of a glottal pulse waveform defined by the FANT glottal pulse model produces smaller area distances than the other pre-emphases used. This general reduction in area distances occurs over the full range of $R(1)/R(0)$ presented in Figure 5.8, which implies that the two/three adaptive pre-emphasis is more effective for glottal pulse waveforms derived from the FANT glottal pulse model than for those derived from the ROSENBERG glottal pulse model, i.e. by comparing Figure 5.7 and 5.8.



(a)

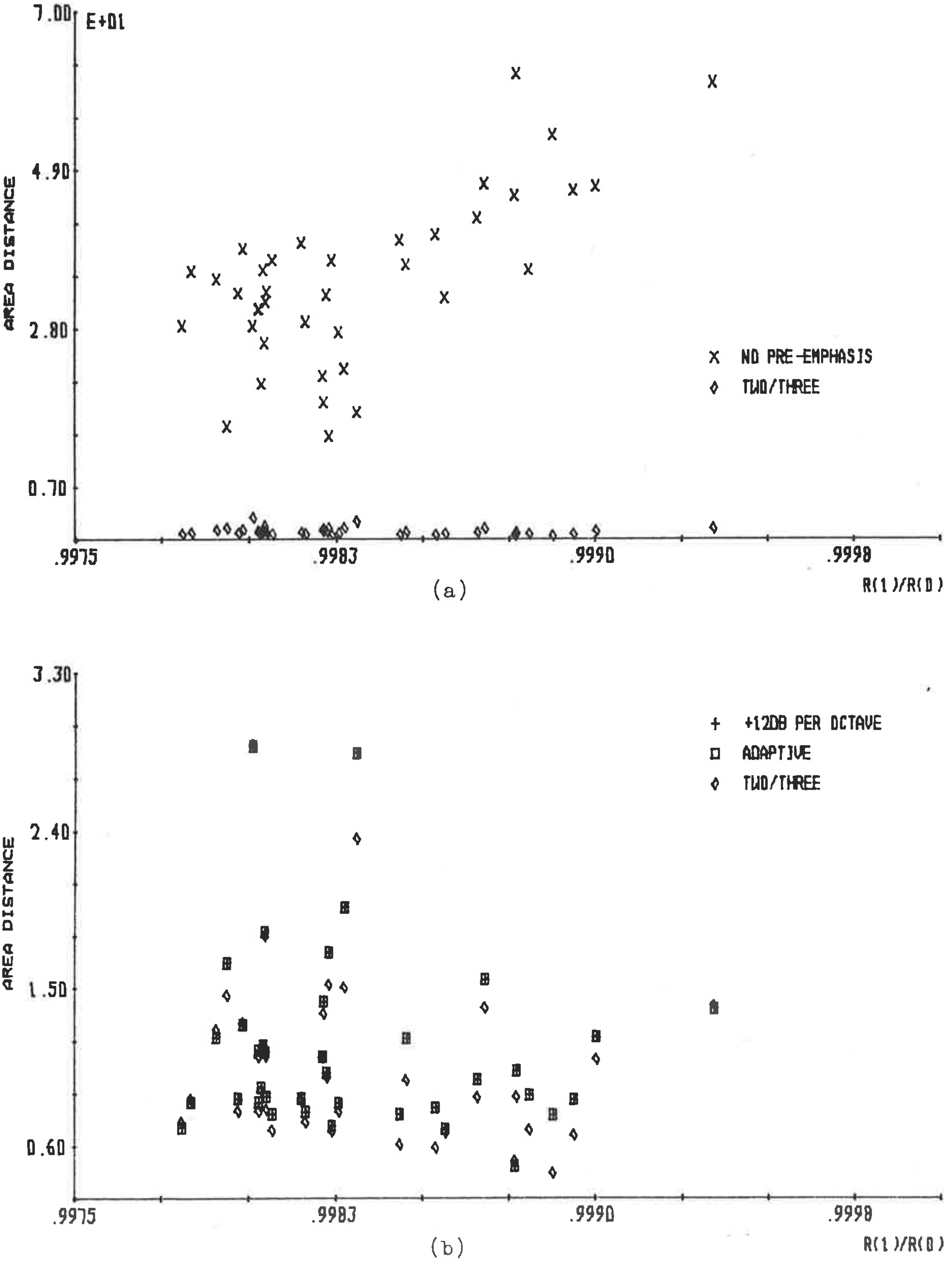
(b)

FIGURE 5.8: (a) Area distances for no pre-emphasis and a two/three adaptive pre-emphasis, and (b) area distances for a +12dB per octave, an unvoiced/voiced adaptive and two/three adaptive pre-emphases, of glottal pulses from the FANT glottal pulse model.

Using glottal pulse waveforms derived from published glottal pulse waveforms of MONSEN and ENGEBRETSON [1977], ROTHENBERG [1973], SONDHI [1975], SUNDBERG and GAUFFIN [1978], MILLER [1959], and FLANAGAN and LANDGRAF [1968], a plot of area distances versus $R(1)/R(0)$ for no pre-emphasis, a +12 dB per octave pre-emphasis, and unvoiced/voiced and two/three adaptive pre-emphases is presented in Figure 5.9. The area distances in Figure 5.9 resulting from no pre-emphasis of the glottal pulse waveforms are much larger than those for the other pre-emphases. This observation is consistent with the evaluation results for glottal pulse waveforms derived from glottal pulse models presented in Figure 5.7 and 5.8.

Figure 5.9(b) shows that, in general, a two/three adaptive pre-emphasis of glottal pulse waveforms derived from real glottal pulse waveforms results in smaller area distances than for the other pre-emphases considered. In the few cases where the above statement is incorrect, the area distances of all three pre-emphases have similar values. The reduction in area distances achieved by the two/three adaptive pre-emphasis filter, in Figure 5.9, occurs over the full range of $R(1)/R(0)$ considered in Figure 5.9, which is consistent with the evaluations presented in Figure 5.8 for glottal pulse waveforms from the FANT glottal pulse model.

In general, the evaluations presented in this section have shown that the two/three adaptive pre-emphasis filter, when defined by the parameter β, provides a small reduction in area distances when compared with other pre-emphasis techniques (including no pre-emphasis). The reduction in area distances is small when compared with that achieved by the two/three adaptive pre-emphasis filter when defined by the parameter α, i.e. as presented in Section



(a)

(b)

FIGURE 5.9: (a) Area distances for no pre-emphasis and a two/three adaptive pre-emphasis, and (b) area distances for a +12dB per octave, an unvoiced/voiced adaptive and two/three adaptive pre-emphases of real glottal pulses.

5.2.1 However, since a general reduction in area distances is achieved, even though it is small, it is an advantage to use the two/three adaptive pre-emphasis filter in preference to the other pre-emphasis techniques considered here.

5.2.3 SAMPLING FREQUENCIES

The relationships between the parameter α and $R(1)/R(0)$ and the parameter β and $R(1)/R(0)$, which define the two/three adaptive pre-emphasis filter, were shown in Chapter 4 to be dependent on the waveform sampling frequency used. These relationships have been defined in Chapter 4 for waveform sampling frequencies of 10, 9, 8 and 7 kilohertz. Evaluation of the two/three adaptive pre-emphasis filter has been performed with glottal pulse waveforms sampled at a frequency of 10 kilohertz in Sections 5.2.1 and 5.2.2. This section evaluates the two/three adaptive pre-emphasis filter with glottal pulse waveforms that have been sampled at frequencies of 9, 8 and 7 kilohertz.

The effects of waveform sampling frequency are different for the two independent parts of the two/three adaptive pre-emphasis filter, i.e. that defined by the parameter α and that defined by the parameter β; therefore, the performance of each part at different waveform sampling frequencies is presented separately. Sections 5.2.3.1 and 5.2.3.2 present the performance of the two/three adaptive pre-emphasis filter when defined by the parameters α and β, respectively, for pre-emphasis of glottal pulse waveforms sampled at a frequency of 9, 8 and 7 kilohertz.

5.2.3.1 Parameter α

It has been shown in Chapter 4 that the rate at which the analog waveform is sampled affects the relationship between the parameter α of the two/three adaptive pre-emphasis and the value of $R(1)/R(0)$. An evaluation of the two/three adaptive pre-emphasis filter for a sampling frequency of 10 kilohertz was presented in Section 5.2.1, and this section performs an evaluation for sampling frequencies of 9, 8 and 7 kilohertz. The relationship between α and $R(1)/R(0)$ is the same for sampling frequencies between 10 and 8 kilohertz, inclusive, but different for a sampling frequency of 7 kilohertz.

All the evaluations presented in this section are in the form of area distances between recovered and ideal (i.e. no cross-sectional area change over its length) acoustic tube shapes versus $R(1)/R(0)$, which is evaluated from the glottal pulse waveform. The glottal pulse waveforms are generated from the glottal pulse models of ROSENBERG [1971] and FANT [1979], or derived from the published glottal pulse waveforms measured by ROSENBERG [1973], MILLER [1959], SONDHI [1975], and SUNDBERG and GAUFFIN [1978].

Figures 5.10, 5.11 and 5.12 present area distances versus $R(1)/R(0)$ for no pre-emphasis, a +12 dB per octave pre-emphasis (Figures 5.10(a), 5.11(a) and 5.12(a) only), and unvoiced/voiced and two/three adaptive pre-emphases of glottal pulse waveforms, followed by a conventional Parcor analysis, for sampling frequencies of 9, 8 and 7 kilohertz, respectively. The area distances for a +12 dB per octave pre-emphasis are not presented when $R(1)/R(0)$ is less than 0.95, because they are very large.

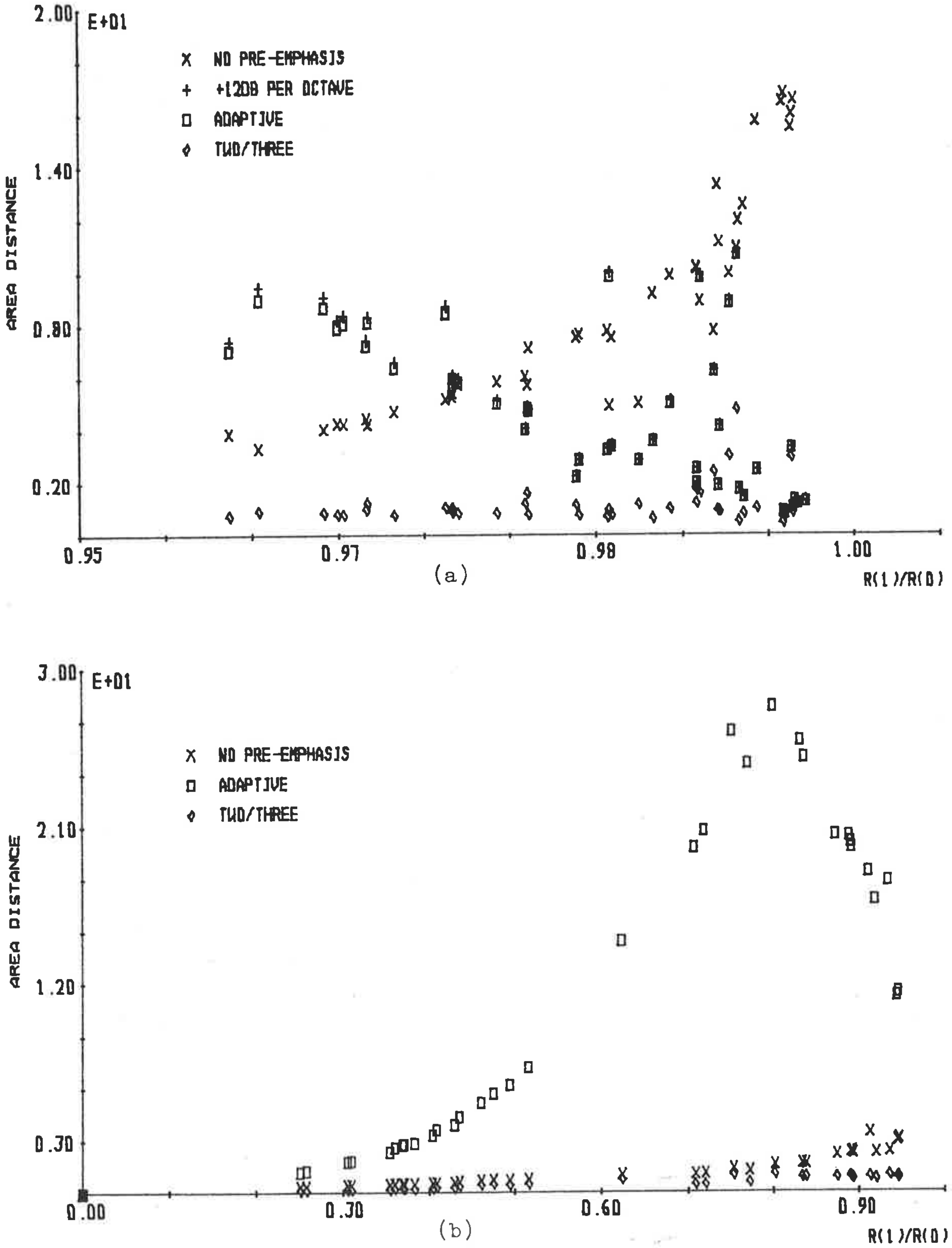
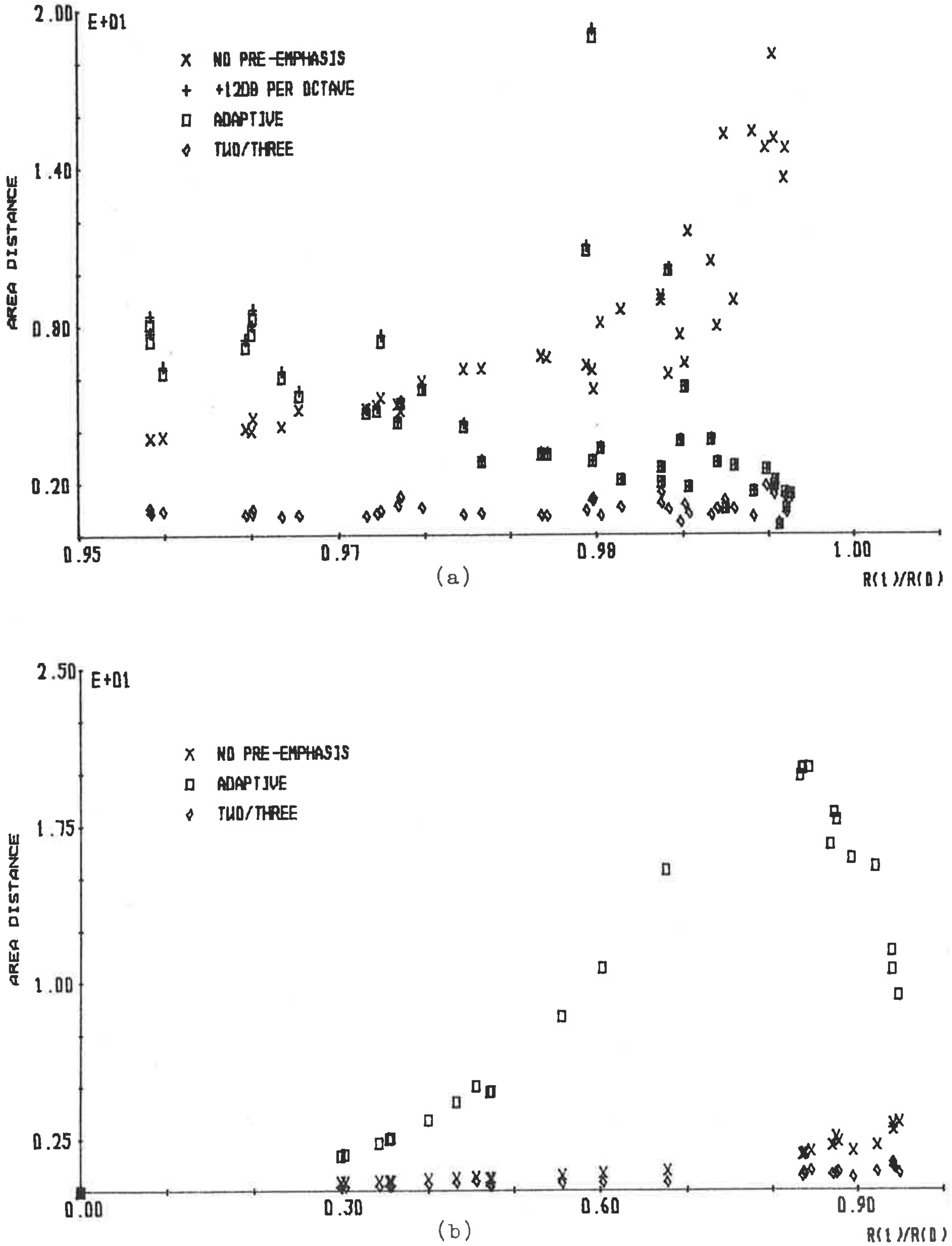


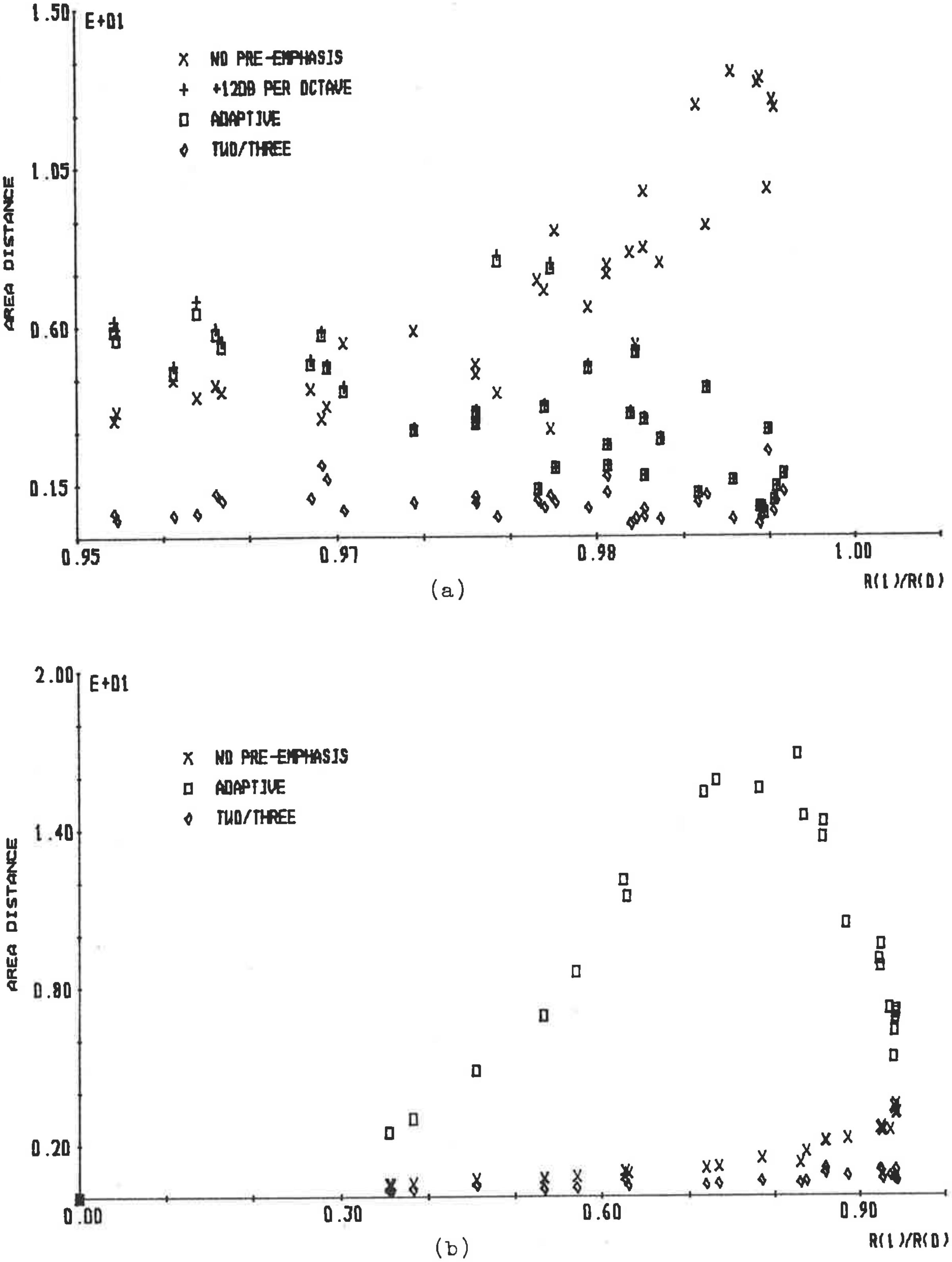
FIGURE 5.10: Area distances for analysis of glottal pulse waveforms sampled at 9 kilohertz with R(1)/R(0) (a) greater than 0.95 and (b) less than 0.95.



(a)

(b)

FIGURE 5.11: Area distances for analysis of glottal pulse waveforms sampled at 8 kilohertz with R(1)/R(0) (a) greater than 0.95 and (b) less than 0.95.



(a)

(b)

FIGURE 5.12: Area distances for analysis of glottal pulse waveforms sampled at 7 kilohertz with R(1)/R(0) (a) greater than 0.95 and (b) less than 0.95.

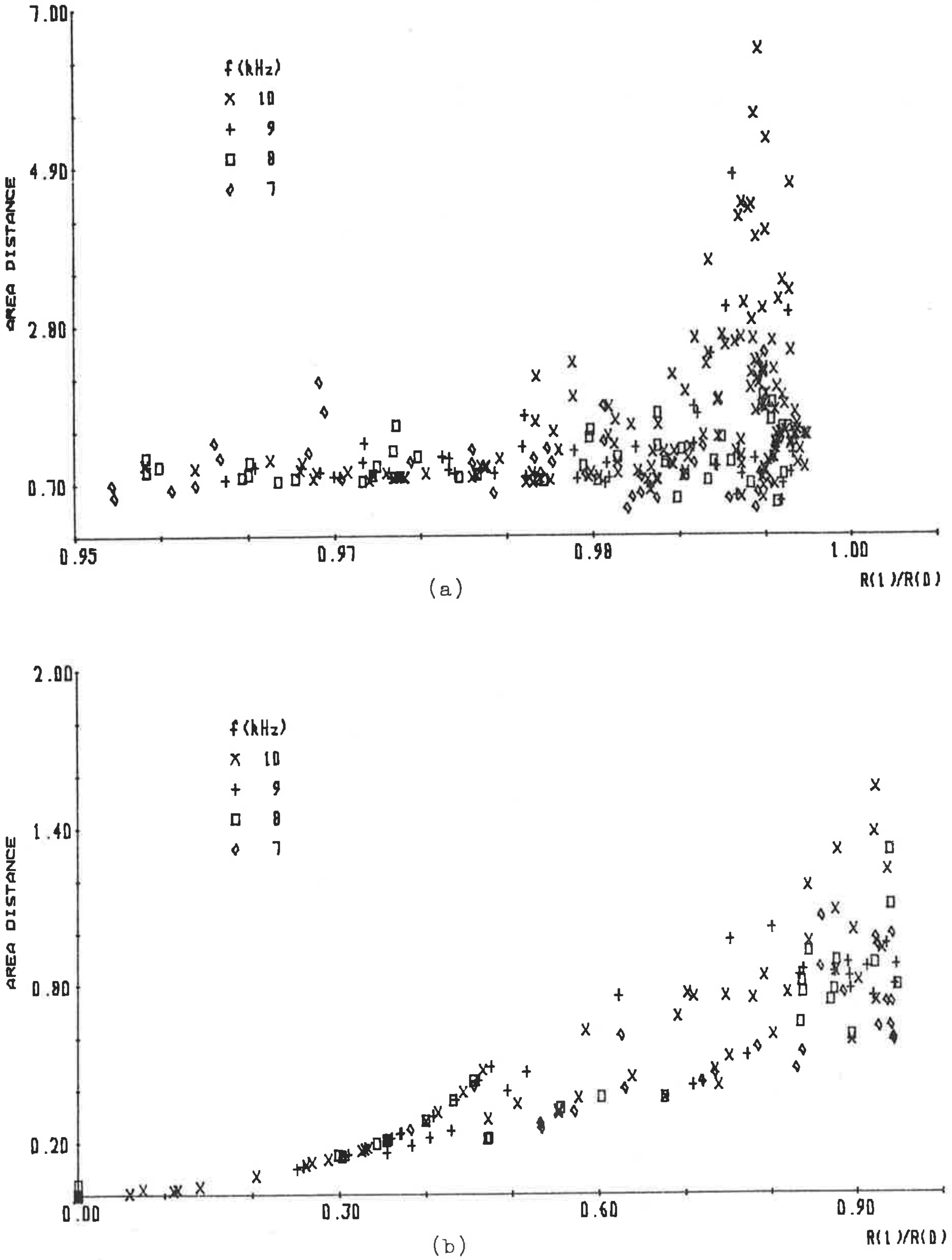
The trends of area distance versus $R(1)/R(0)$ are observed in Figure 5.10 to be similar to those occurring for a sampling frequency of 10 kilohertz (see Section 5.2.1). The two/three adaptive pre-emphasis filter is shown in Figure 5.10 to produce much smaller area distances than those for no pre-emphasis and an unvoiced/voiced adaptive pre-emphasis. When $R(1)/R(0)$ is close to unity, then similar area distances occur for a +12 dB per octave pre-emphasis and the unvoiced/voiced and two/three adaptive pre-emphases. These observations are consistent with those for a sampling frequency of 10 kilohertz (see Section 5.2.1); therefore, a sampling frequency between 10 and 9 kilohertz, inclusive, does not significantly affect the performance of the two/three adaptive pre-emphasis filter.

Trends of area distances versus $R(1)/R(0)$ are observed in Figure 5.11 to be similar to those presented in Section 5.2.1, where a sampling frequency of 10 kilohertz is used. The two/three adaptive pre-emphasis filter is shown in Figure 5.11 to provide a large reduction in area distances in comparison with those for no pre-emphasis and an unvoiced/voiced adaptive pre-emphasis. Similar area distances occur for the +12 dB per octave pre-emphasis and the unvoiced/voiced and two/three adaptive pre-emphases, when $R(1)/R(0)$ is near unity. The reduction in area distances by the two/three adaptive pre-emphasis in comparison with the other pre-emphases considered is smaller for a sampling frequency of 8 kilohertz than for sampling frequencies of 10 and 9 kilohertz. However, the reduction in area distances is still significant; therefore, there exists a definite advantage in using the two/three adaptive pre-emphasis filter in preference to other pre-emphasis techniques.

Observation of Figure 5.12 and a comparison with Figures 5.10 and 5.11 and the evaluations presented in Section 5.2.1 for a sampling frequency of 10 kilohertz shows similar trends in area distance versus $R(1)/R(0)$ occurs for sampling frequencies of 7, 8, 9 and 10 kilohertz. The reduction in area distances for a two/three adaptive pre-emphasis in comparison with the other pre-emphases considered is smaller for a sampling frequency of 7 kilohertz than occurs for sampling frequencies of 10, 9 and 8 kilohertz. However, the reduction in area distances is still large, and shows that the two/three adaptive pre-emphasis filter provides a significant improvement, in comparison with other pre-emphasis techniques, in accounting for glottal pulse excitation effects for sampling frequencies between 10 and 7 kilohertz, inclusive.

A comparison of area distances for a two/three adaptive pre-emphasis (when defined by the parameter α) of glottal pulse waveforms for sampling frequencies of 10, 9, 8 and 7 kilohertz is presented in Figure 5.13. Observation of Figure 5.13 shows that the area distances for different sampling frequencies are similar. Therefore, it can be concluded that the relationships between α and $R(1)/R(0)$ defined in Chapter 4 have been successful in removing the effects of a change in sampling frequency between 10 and 7 kilohertz, inclusive.

The evaluations presented in this section have shown that, when the two/three adaptive pre-emphasis filter is defined by the parameter α, the performance of the two/three adaptive pre-emphasis filter in removing glottal pulse excitation effects on the recovered acoustic tube shape is not significantly affected by a change in sampling frequency between 10 and 7 kilohertz, inclusive.



(a)

(b)

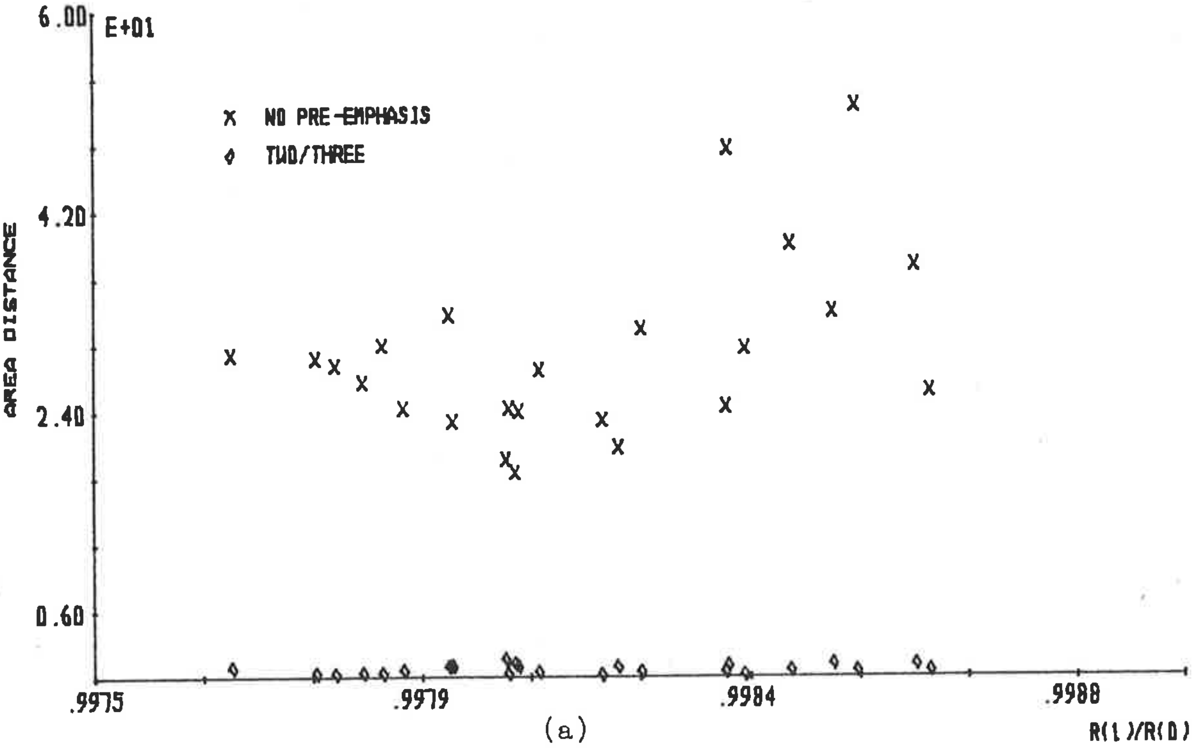
FIGURE 5.13: Comparison of area distances for a two/three adaptive pre-emphasis of glottal pulse waveforms sampled at 10, 9, 8 and 7 kilohertz with R(1)/R(0) (a) greater than 0.95 and (b) less than 0.95.

5.2.3.2 Parameter β

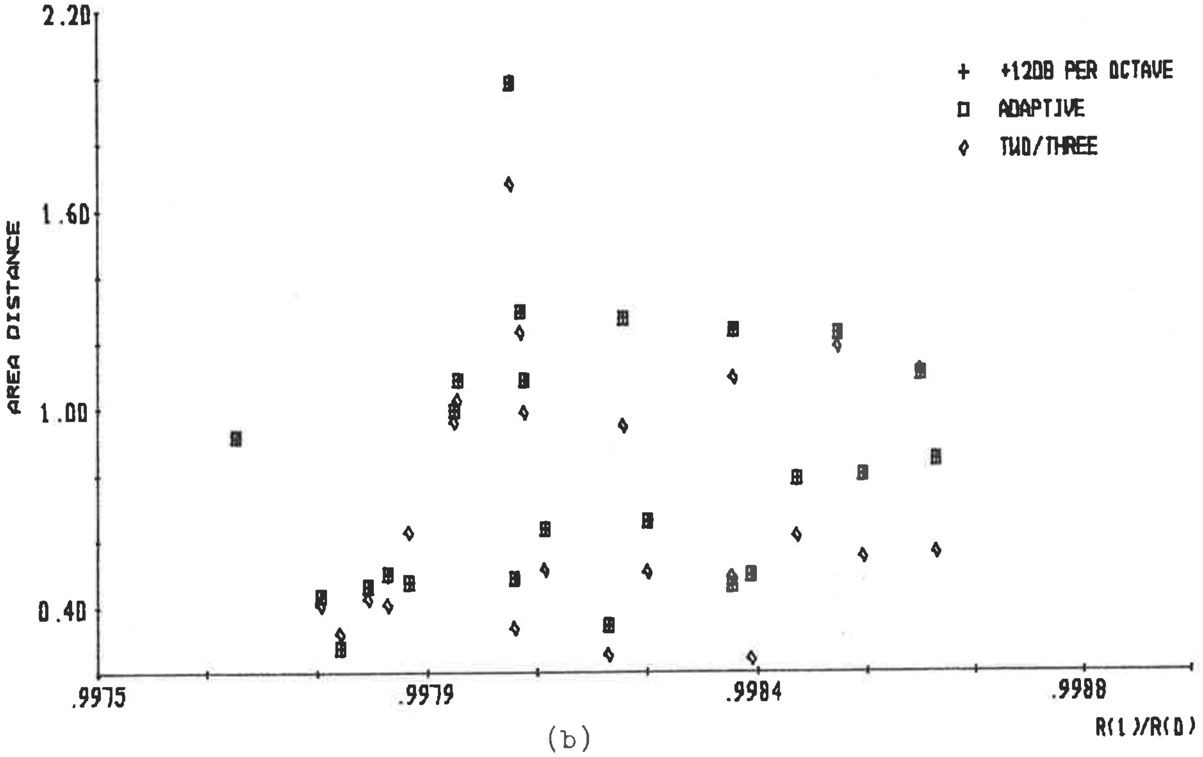
An evaluation of the two/three adaptive pre-emphasis filter when defined by the parameter β was presented in Section 5.2.2 for a waveform sampling frequency of 10 kilohertz. This section evaluates the two/three adaptive pre-emphasis when it is defined by the parameter β, and waveform sampling frequencies of 9, 8 and 7 kilohertz are used. The relationship between β and $R(1)/R(0)$ which is used to define the two/three adaptive pre-emphasis filter changes with different sampling frequencies, as discussed in Chapter 4.

All the evaluations of the two/three adaptive pre-emphasis are presented as area distnaces between recovered and ideal (i.e. no cross-sectional area change over its length) acoustic tube shapes versus the value of $R(1)/R(0)$, which is evaluated from the glottal pulse waveforms. The glottal waveforms used to perform the evaluations presented in this section are derived from published glottal pulse waveforms of MONSEN and ENGEBRETSON [1977], ROTHENBERG [1973], SONDHI [1975], SUNDBERG and GAUFFIN [1978], MILLER [1959], and FLANAGAN and LADNGRAF [1968].

For sampling frequencies of 9, 8 and 7 kilohertz, Figures 5.14, 5.15 and 5.16, respectively, present the area distances versus $R(1)/R(0)$ for no pre-emphasis, a +12 dB per octave pre-emphasis and the unvoiced/voiced and two/three adaptive pre-emphases of glottal pulse waveforms followed by a conventional Parcor linear predictive analysis. These results, coupled with those presented in Section 5.2.2, provide evaluations of the two/three adaptive pre-emphasis filter, when defined by the parameter β, for the range of sampling frequencies from 10 to 7 kilohertz, inclusive.

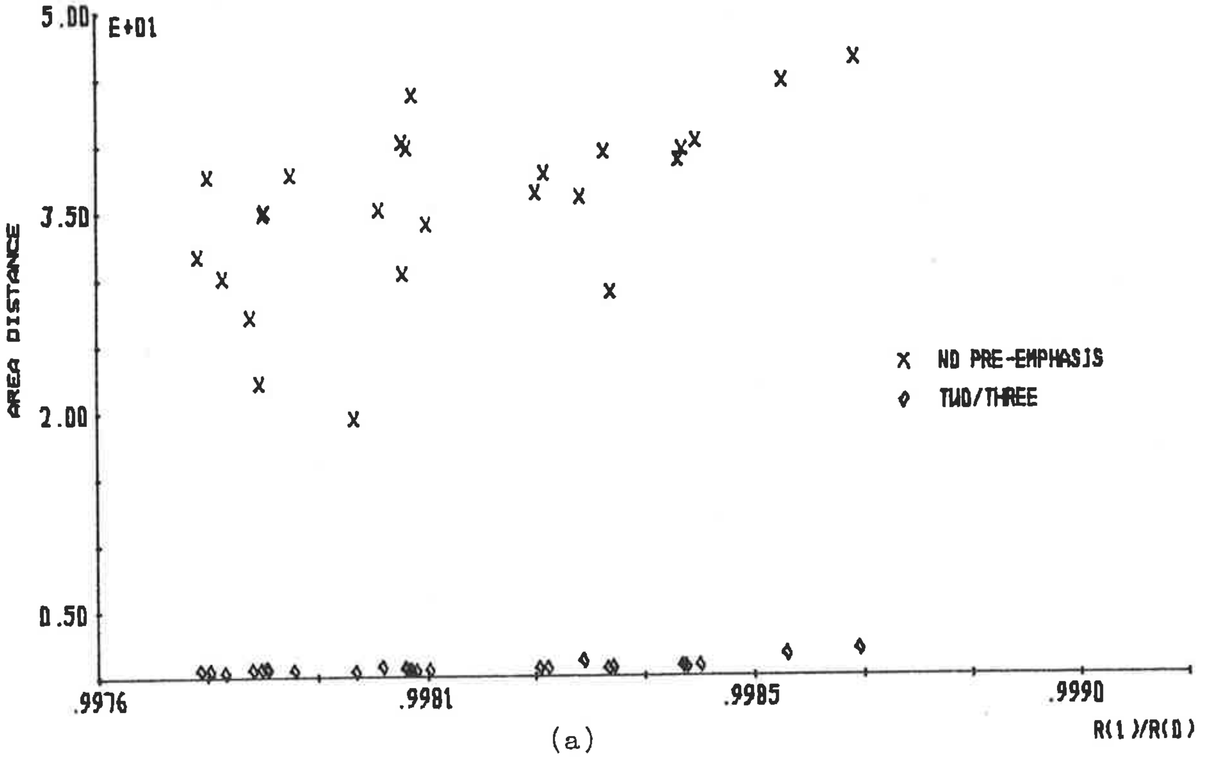


(a)



(b)

FIGURE 5.14: Area distances for an analysis of glottal pulse waveforms sampled at 9 kilohertz by (a) no pre-emphasis and a two/three adaptive pre-emphasis and (b) a +12dB per octave, an unvoiced/voiced adaptive and two/three adaptive pre-emphases.



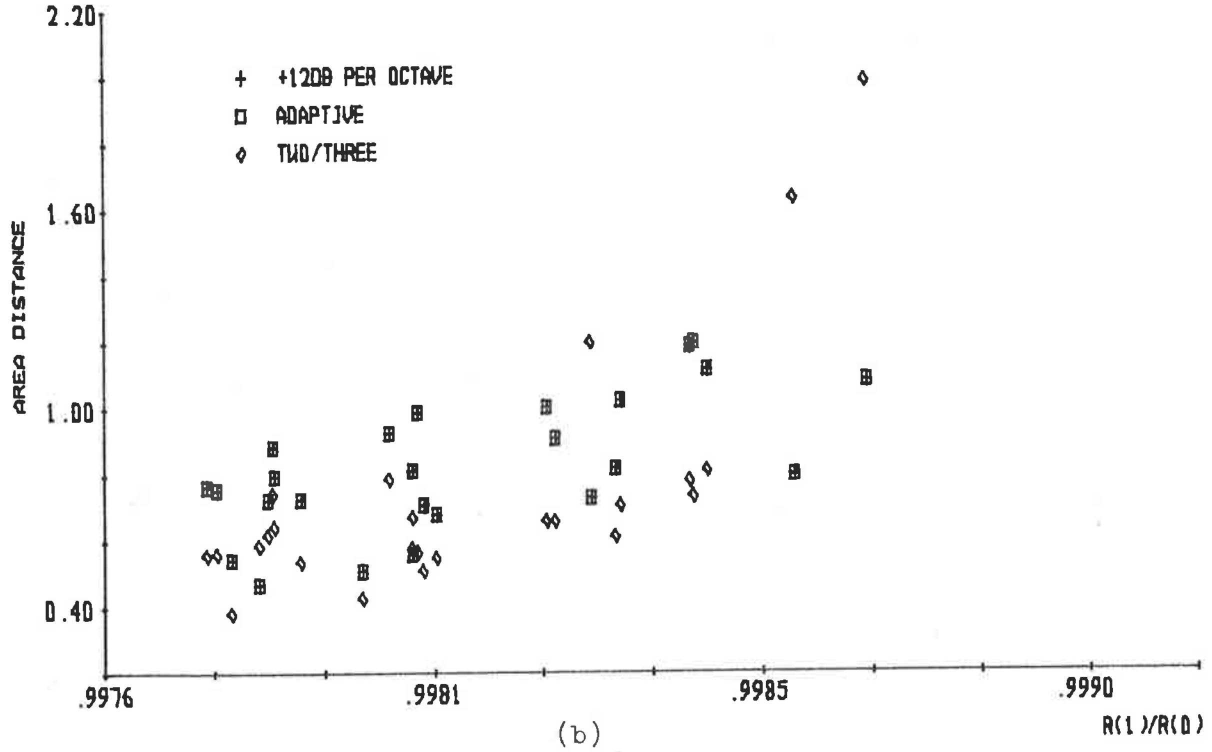
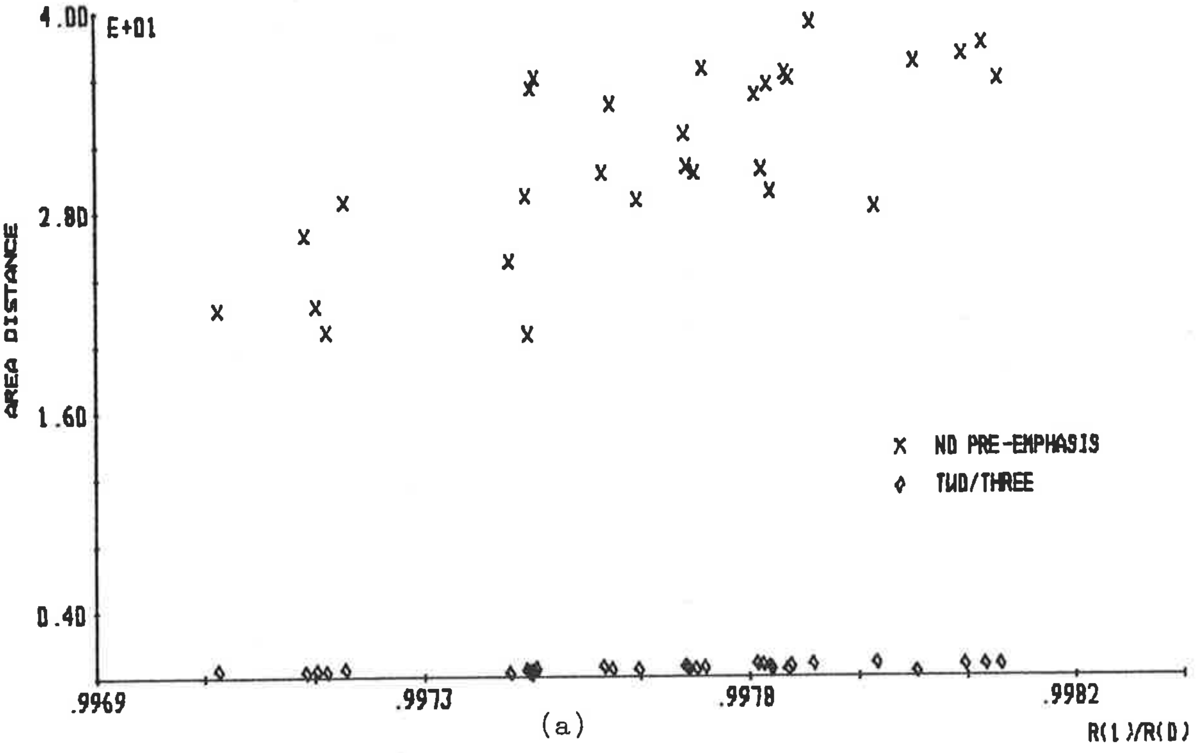
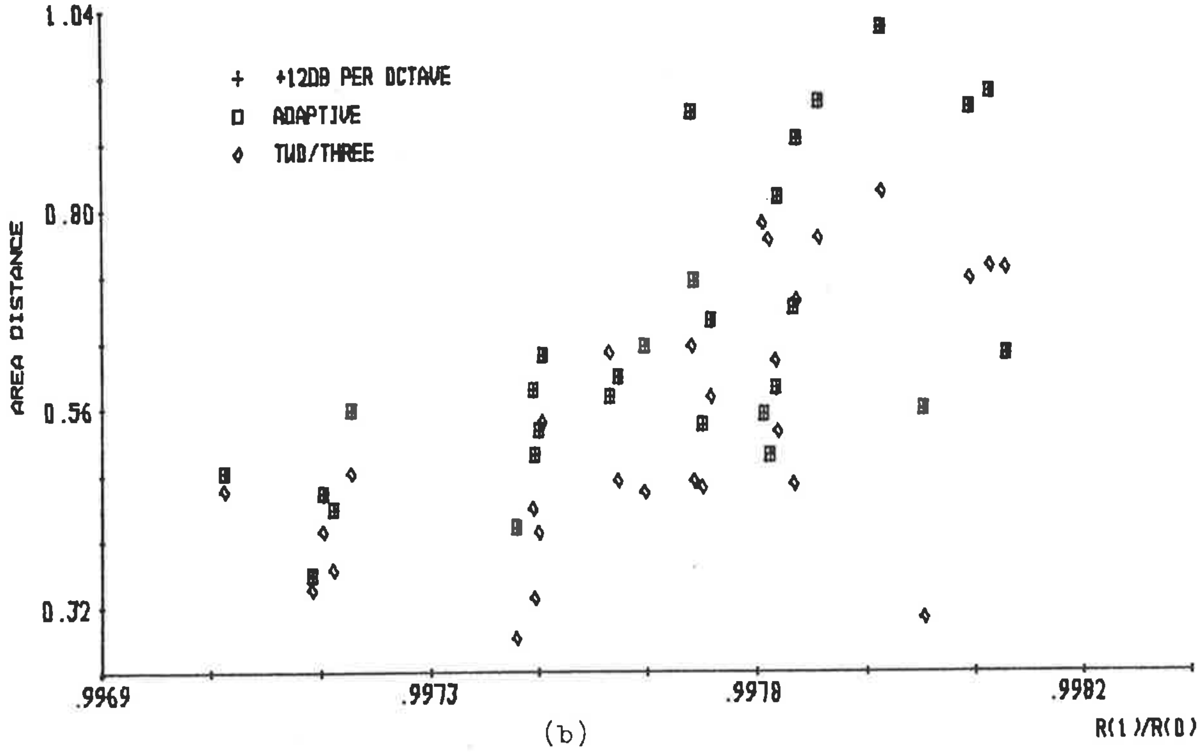


FIGURE 5.15: Area distances for an analysis of glottal pulse waveforms sampled at 8 kilohertz by (a) no pre-emphasis and a two/three adaptive pre-emphasis and (b) a +12dB per octave, an unvoiced/voiced adaptive and two/three adaptive pre-emphases



(a)



(b)

FIGURE 5.16: Area distances for an analysis of glottal pulse waveforms sampled at 7 kilohertz by (a) no pre-emphasis and a two/three adaptive pre-emphasis and (b) a +12dB per octave, an unvoiced/voiced adaptive and two/three adaptive pre-emphases.

Figures 5.14 to 5.16 show that the two/three adaptive pre-emphasis filter provides considerably smaller area distances than no pre-emphasis for sampling frequencies of 9, 8 and 7 kilohertz. Considerably smaller area distances for the two/three adaptive pre-emphasis filter are observed in Figure 5.9, where the waveform sampling frequency of 10 kilohertz is used. Therefore, the two/three adaptive pre-emphasis filter, when defined by the parameter β, produces a large reduction in area distances in comparison with no pre-emphasis when the waveform sampling frequencies are between 10 and 7 kilohertz, inclusive. The actual amount by which the area distances are reduced if a two/three pre-emphasis filter is used instead of no pre-emphasis increases as the sampling frequency decreases.

The area distances for a +12 dB per octave and unvoiced/voiced adaptive pre-emphases are similar in Figures 5.14, 5.15 and 5.16, which is consistent with the evaluation results of Figure 5.9, where the waveform sampling frequency is 10 kilohertz. In all but a few cases, the two/three adaptive pre-emphasis filter is shown in Figures 5.14, 5.15 and 5.16 to produce a reduction in area distances when compared with the area distances of a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis. For each sampling frequency, the amount by which an area distance reduction is achieved by using the two/three adaptive pre-emphasis filter varies a great deal but, in general, is small compared with the area distance reduction achieved by the two/three adaptive pre-emphasis filter when it is defined by the parameter α.

A comparison of the area distnaces for the two/three adaptive pre-emphasis filter (when defined by the parameter β) for the waveform sampling frequencies of 10, 9, 8 and 7 kilohertz is presented in Figure 5.17. As the sampling frequency decreases, a general reduction in area distance values is observed in Figure 5.17. As expected from the discussions of Section 4.4.3.2, the range of $R(1)/R(0)$ for a reduction in sampling frequency shifts towards smaller values of $R(1)/R(0)$. In general, these effects are small, and it can be concluded that only minor changes in area distances occur due to a change in sampling frequency.

Therefore, the results presented in this section and those presented in Figure 5.9 show that a reduction in area distances is achieved by the two/three adaptive pre-emphasis filter when compared with the area distances for no pre-emphasis, a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis for sampling frequencies between 10 and 7 kilohertz. A large reduction in area distances occurs for the two/three adaptive pre-emphasis in comparison with no pre-emphasis, but only a small reduction in area distances in comparison with +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis. In general, the results presented in this section show that the relationships between β and $R(1)/R(0)$ developed in Chapter 4 have been successful in removing the effects of a change in sampling frequency between 10 and 7 kilohertz, inclusive.

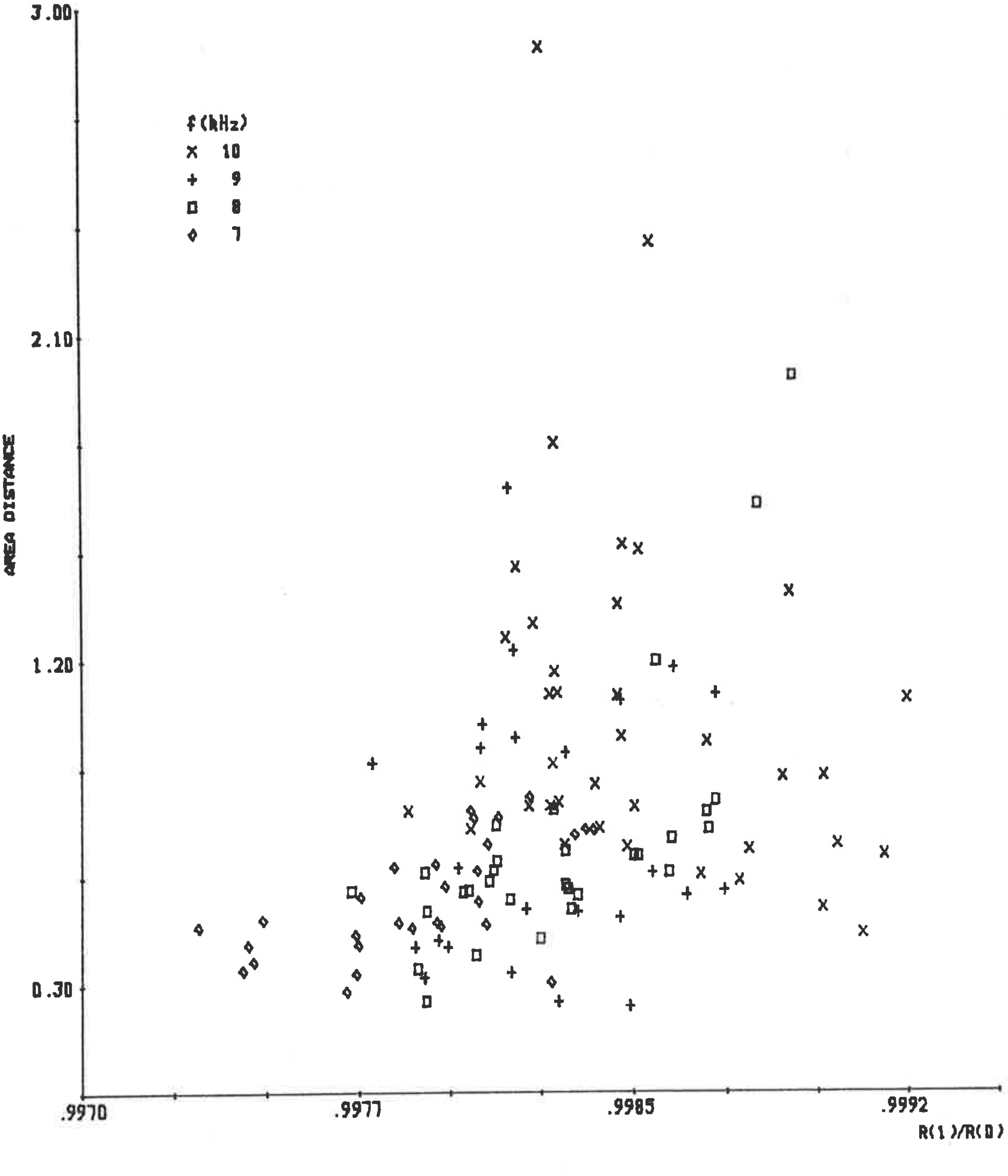


FIGURE 5.17: Comparison of area distances for a two/three adaptive pre-emphasis of glottal pulse waveforms sampled at 10, 9, 8 and 7 kilohertz.

5.3 EVALUATION WITH SYNTHETIC SPEECH DATA

The two/three adaptive pre-emphasis filter is designed to remove glottal pulse excitation effects from the speech waveform so that improved acoustic tube shape recovery may be achieved. An evaluation of the two/three adaptive pre-emphasis filter with real speech is difficult due to the problems of obtaining vocal tract shapes for the particular speech sounds being analysed. This section performs an evaluation of the two/three adaptive pre-emphasis filter with synthetic speech data, which overcomes the problems with using real speech data, since the original acoustic tube shape is known.

The procedure for generating the synthetic speech data is detailed in Appendix B. All the glottal pulse waveforms used in generating the synthetic speech are the same as those used in Section 5.2. The source of the glottal pulse waveforms is either the glottal pulse models of ROSENBERG [1971] and FANT [1979] or published glottal pulse waveforms [ROSENBERG 1973, MONSEN and ENGEBRETSON 1977, MILLER 1959, SONDHI 1975, SUNDBERG and GAUFFIN 1978]. The original acoustic tube shapes used are derived from FANT's [1970] measured vocal tract shapes for Russian vowels, as described in Appendix C. These original acoustic tube shapes are used to provide synthetic speech which is as close as possible to real speech. The two/three adaptive pre-emphasis filter does not correct for radiation from the termination of the acoustic tubes and, therefore, a lossless termination of the original acoustic tube, i.e. $\mu_M = +1$, is used to avoid radiation effects clouding the evaluation results presented in this section.

The process used to evaluate the two/three adaptive pre-emphasis filter with synthetic speech is to firstly pass the synthetic speech through the two/three adaptive pre-emphasis filter. The output waveform from the two/three pre-emphasis filter is then analysed by a Parcor linear predictive analysis procedure (see Chapter 2) to produce a recovered acoustic tube shape. A comparison of recovered and original acoustic tube shapes is then performed by an area distance measure (see Section 2.5), to determine the effectiveness of the two/three adaptive pre-emphasis filter in providing accurate acoustic tube shape recovery. In situations where the area distance measure provides ambiguous results, both the recovered and original acoustic tube shapes are plotted on the same graph.

To determine if the two/three adaptive pre-emphasis filter provides an improvement in acoustic tube shape recovery, the results of the evaluations for the two/three adaptive pre-emphasis filter are compared with the results for no pre-emphasis and two conventional pre-emphasis techniques applied to the same synthetic speech waveforms. The conventional pre-emphasis techniques are the same as those used in Section 5.2, i.e. a constant +12 dB per octave pre-emphasis and the unvoiced/voiced adaptive pre-emphasis filter derived from the work of GRAY and MARKEL [1974] and MAKHOUL and VISWANATHAN [1974].

The two/three adaptive filter has one form defined by a parameter α when the glottal pulse spectral slope is greater than approximately -12 dB per octave, and another form defined by a parameter β when the glottal pulse spectral slope is less than approximately -12 dB per octave. Since these two forms are independent of each other, the evaluation of the two/three adaptive pre-emphasis filter with synthetic speech waveforms is performed in two parts. Section 5.3.1 evaluates the two/three adaptive pre-emphasis filter with synthetic speech waveforms when the filter is defined by the parameter α. The evaluation of the two/three adaptive pre-emphasis filter with synthetic speech waveforms when the filter is defined by the parameter β is presented in Section 5.3.2.

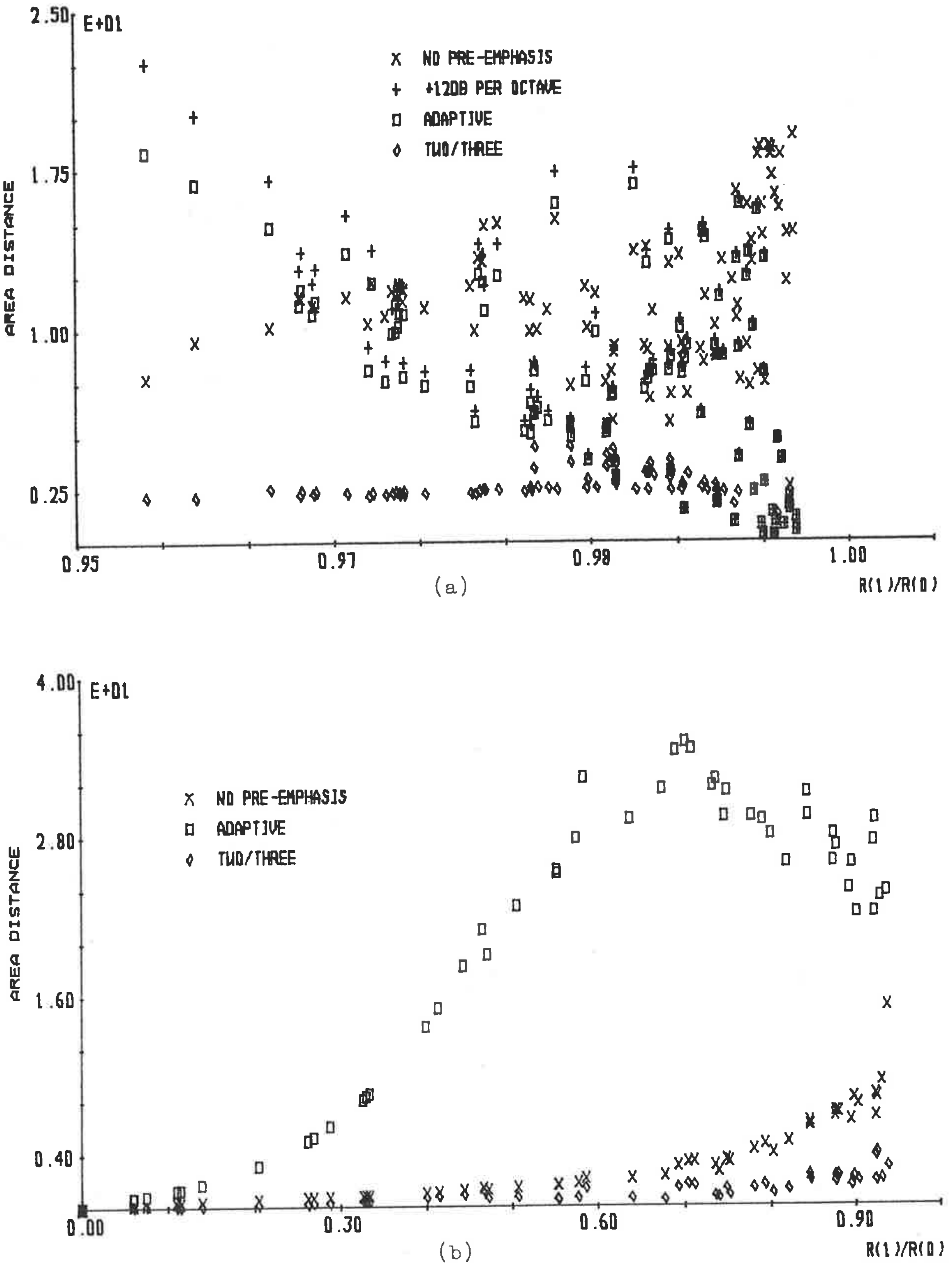
All the evaluations performed in Sections 5.3.1 and 5.3.2 use a sampling frequency of 10 kilohertz. In Section 4.4.3 it was shown that the form of the two/three adaptive pre-emphasis filter changes slightly for different sampling frequencies. Hence, Section 5.3.3 presents an evaluation of the two/three adaptive pre-emphasis filter with synthetic speech waveforms for sampling frequencies of 9, 8 and 7 kilohertz.

5.3.1 PARAMETER α

This section presents an evaluation of the two/three adaptive pre-emphasis filter, when defined by the parameter α, with synthetic speech data. The value of α used by the two/three adaptive pre-emphasis filter is determined via a relationship with $R(1)/R(0)$, which is evaluated from the synthetic speech data, as discussed and defined in Chapter 4 and Appendix D. The evaluation results are presented as area distances between recovered and original acoustic tube shapes versus $R(1)/R(0)$ of the glottal pulse waveform used to generate the synthetic speech waveform. The value of $R(1)/R(0)$ of the glottal pulse waveform is used so that all the results are related to the glottal pulse waveforms which the two/three adaptive pre-emphasis has been designed to remove.

The procedure used to generate the synthetic speech waveforms is described in Appendix B. To ensure that the synthetic speech waveforms are similar to real speech waveforms, the acoustic tube shapes used in the generation process approximate the vocal tract shapes for the five vowels |a|, |e|, |i|, |o| and |u|, as defined in Appendix C. The sampling frequency for all the synthetic waveforms generated and used to produce the evaluation results presented in this section is 10 kilohertz. The glottal pulse waveforms used to excite the acoustic tube model and generate the synthetic speech are the same as those used in Section 5.2.1, being generated from the glottal pulse models of ROSENBERG [1971] and FANT [1979] and the digitized glottal pulse waveforms measured by ROSENBERG [1973], MONSEN and ENGEBRETSON [1977], MILLER [1959], SONDHI [1975], and SUNDBERG and GAUFFIN [1978]. To avoid radiation effects clouding the evaluations presented in this section, the termination reflection coefficient of the acoustic tube model is unity, i.e. $\mu_M = 1.0$.

Evaluation results for synthetic speech of the vowel |a| are presented in Figure 5.18 as area distances versus $R(1)/R(0)$ for no pre-emphasis, a +12 dB per octave pre-emphasis (Figure 5.18(a) only), an unvoiced/voiced adaptive pre-emphasis and a two/three adaptive pre-emphasis. The area distances for a +12 dB per octave pre-emphasis are not presented when $R(1)/R(0)$ is less than 0.95, as they are considerably larger than the area distances for the other pre-emphases. Similar area distances are observed in Figure 5.18 for no pre-emphasis and the unvoiced/voiced and two/three adaptive pre-emphases when $R(1)/R(0)$ is near zero, and for a +12 dB per octave pre-emphasis and the unvoiced/voiced and two/three adaptive



(a)

(b)

FIGURE 5.18: Area distances for analysis of synthetic speech for the vowel /a/ with R(1)/R(0) (a) greater than 0.95 and (b) less than 0.95.

pre-emphases when $R(1)/R(0)$ is near unity. For $R(1)/R(0)$ greater than 0.95, the +12 dB per octave pre-emphasis and the unvoiced/voiced adaptive pre-emphasis produce similar area distances.

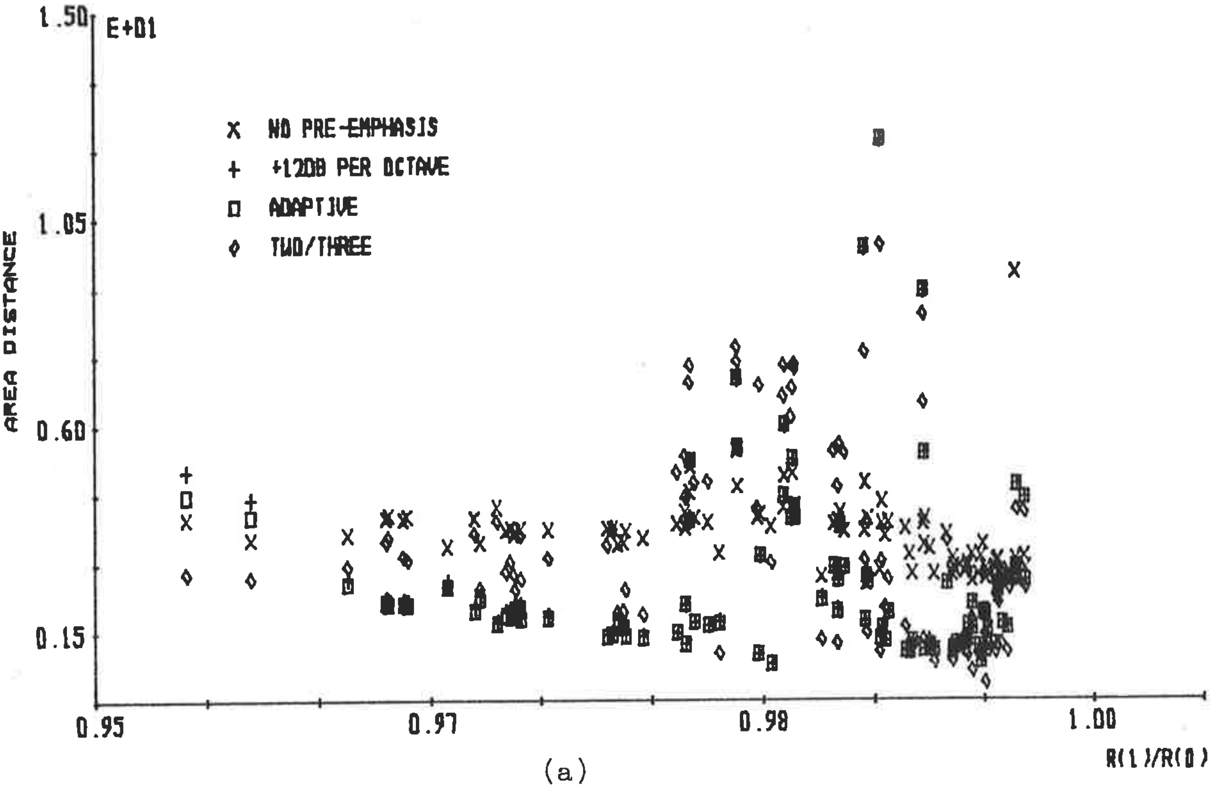
Figure 5.18 shows that, in general, the two/three adaptive pre-emphasis filter produces smaller area distances than no pre-emphasis or a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis. In comparison with the unvoiced/voiced adaptive pre-emphasis, the two/three adaptive pre-emphasis filter produces similar area distances when $R(1)/R(0)$ is near unity or zero, but significantly smaller area distances otherwise. Except for a few cases, when $R(1)/R(0)$ is near unity, the area distances for a two/three adaptive pre-emphasis are much less than the area distances for a +12 dB per octave pre-emphasis.

A reduction in area distances occurs for the two/three adaptive pre-emphasis filter in comparison with no pre-emphasis over the full range of $R(1)/R(0)$ presented in Figure 5.18, with the largest reduction in area distances occurring when $R(1)/R(0)$ is near unity and the smallest reduction when $R(1)/R(0)$ is near zero. In general, the results presented in Figure 5.18 show that significantly smaller area distances occur for the two/three adaptive pre-emphasis when compared with the area distances for no pre-emphasis and the other pre-emphases. The relative variation in area distances for a two/three adaptive pre-emphasis is observed to be smaller than for no pre-emphasis and the other pre-emphases.

A comparison of the evaluation results presented in Figure 5.18 with those presented in Section 5.2.1 for glottal pulse waveforms reveals a consistency of general trends. The relatively small variation in area distances observed in Figure 5.18, which is consistent with the observations in Section 5.2.1, for the two/three adaptive pre-emphasis filter indicates that the filter produces similar acoustic tube shapes irrespective of the glottal pulse waveform used to generate synthetic speech for the vowel |a|. Hence, a significant improvement in acoustic tube shape recovery occurs when the two/three adaptive pre-emphasis filter is used with synthetic speech of the vowel |a|.

Synthetic speech for the vowel |e| is generated in the same manner as for the vowel |a|, with the procedure being detailed in Appendix B. Figure 5.19 presents the area distances versus $R(1)/R(0)$ for no pre-emphasis, a +12 dB per octave pre-emphasis (Figure 5.19(a) only), an unvoiced/voiced adaptive pre-emphasis, and a two/three adaptive pre-emphasis of synthetic speech for the vowel |e|. The area distances for a +12 dB per octave pre-emphasis are very large when $R(1)/R(0)$ is less than 0.95, and so are not presented in Figure 5.19. The glottal pulse waveforms used to generate the synthetic speech for the vowel |e| are the same as those used to generate the synthetic speech for the vowel |a|.

The general trends of area distances versus $R(1)/R(0)$ observed in Figure 5.19(a), where $R(1)/R(0)$ is between 0.95 and approximately unity, are different from the general trends observed in Figure 5.18(a) for the vowel |a|. In most cases, Figure 5.19(a) shows that the +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis produce similar area distances, which are smaller than no pre-emphasis. The area distances for a two/three adaptive pre-emphasis are similar when $R(1)/R(0)$ is near unity, but slightly larger when $R(1)/R(0)$ is less than unity, when com-



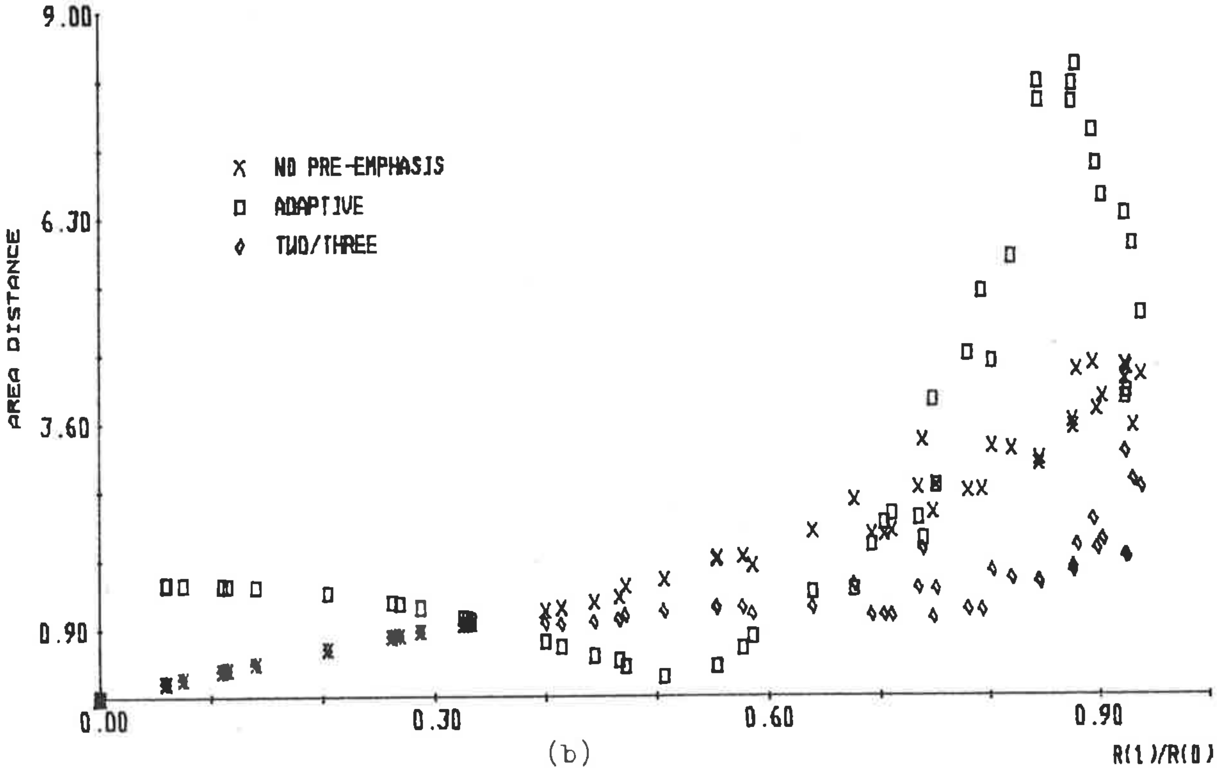


FIGURE 5.19: Area distances for the analysis of synthetic speech for the vowel /e/ with R(1)/R(0) (a) greater than 0.95 and (b) less than 0.95.

pared with the area distances of a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis.

For $R(1)/R(0)$ between zero and 0.95, the general trends of area distance versus $R(1)/R(0)$ presented in Figure 5.19(b) are different from those presented in Figure 5.18(b) for the vowel |a|. The unvoiced/voiced pre-emphasis filter is shown to provide larger area distances than no pre-emphasis for $R(1)/R(0)$ between 0.95 and 0.7 and between 0.35 and zero, but smaller area distances otherwise. A two/three adaptive pre-emphasis produces consistently smaller area distances than no pre-emphasis for $R(1)/R(0)$ between zero and 0.95, but smaller area distances in comparison with an unvoiced/voiced adaptive pre-emphasis only when $R(1)/R(0)$ is between zero and 0.35 and between 0.6 and 0.95.

In general, the two/three adaptive pre-emphasis filter produces area distances which have a relatively smaller variation than the area distances for no pre-emphasis, a +12 dB per octave pre-emphasis, and an unvoiced/voiced adaptive pre-emphasis. Marked contrasts are observed in the general trends of area distance versus $R(1)/R(0)$ for the vowel |a| and the vowel |e| but, in general, the two/three adaptive pre-emphasis filter produces a reduction in area distances compared with no pre-emphasis and the other pre-emphases when applied to synthetic speech of the vowel |e|.

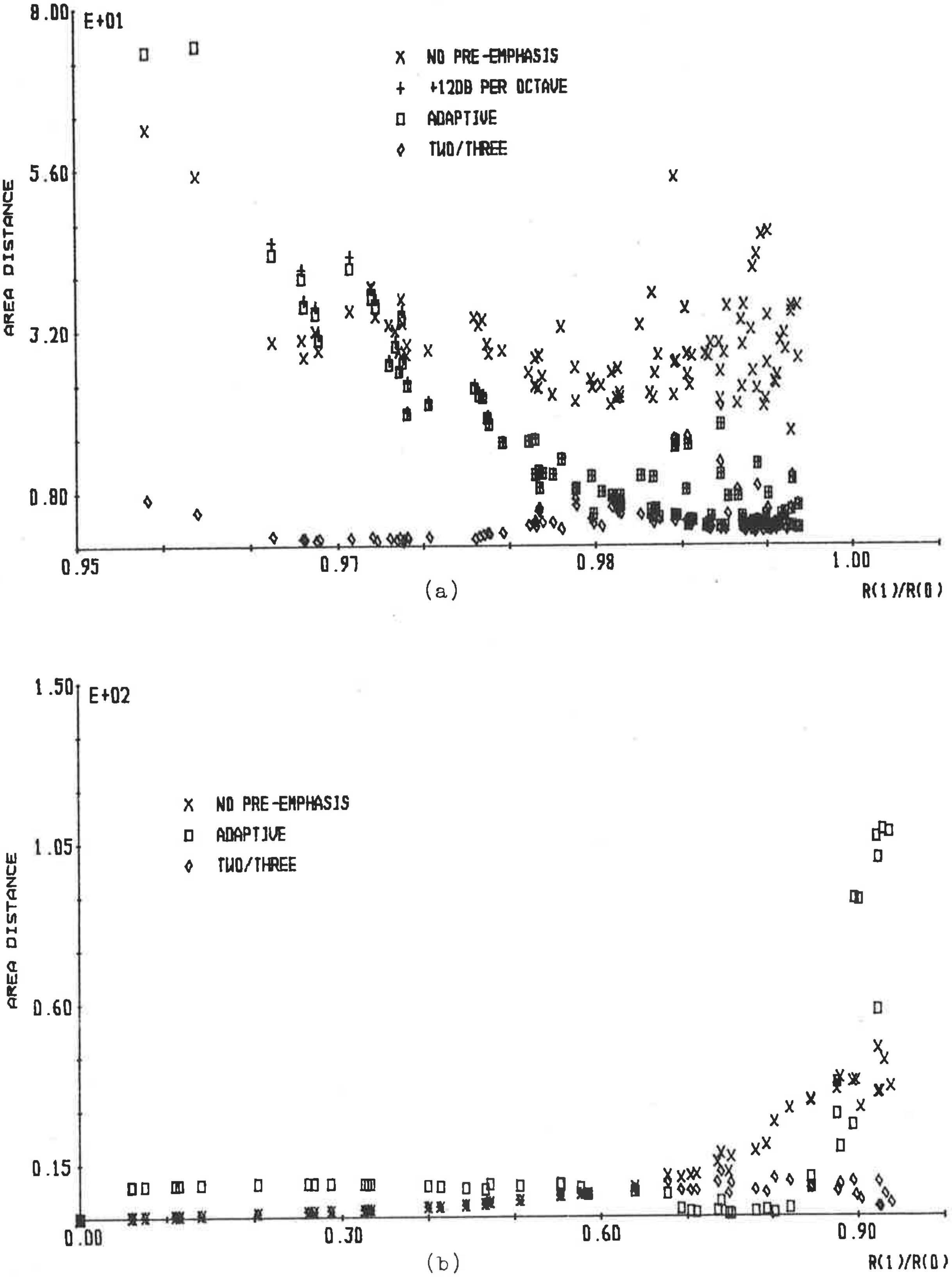
Synthetic speech for the vowel |i| is generated by the procedure defined in Appendix B, and used to evaluate the two/three adaptive pre-emphasis filter. The same glottal pulse waveforms are used to generate the synthetic speech for the vowel |i| as were used to generate the synthetic speech for the vowels |a|

and |e|. Figure 5.20 presents the area distances versus $R(1)/R(0)$ for no pre-emphasis, a +12 dB per octave pre-emphasis (Figure 5.20(a) only), an unvoiced/voiced adaptive pre-emphasis, and a two/three adaptive pre-emphasis of synthetic speech for the vowel |i|.

Figure 5.20(a) shows that a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis produce smaller area distances than no pre-emphasis when $R(1)/R(0)$ is between 0.95 and unity. The two/three adaptive pre-emphasis produces a reduction in area distances in comparison with no pre-emphasis and the other pre-emphases, which increases as the value of $R(1)/R(0)$ decreases towards 0.95. These observations for synthetic speech of the vowel |i| are consistent with those for synthetic speech of the vowel |a|.

The unvoiced/voiced adaptive pre-emphasis filter is shown in Figure 5.20(b) to produce larger area distances than no pre-emphasis when $R(1)/R(0)$ is between zero and 0.6 and between 0.9 and 0.95, but much smaller area distances than no pre-emphasis otherwise. A two/three adaptive pre-emphasis is shown in Figure 5.20(b) to produce smaller area distances than an unvoiced/voiced adaptive pre-emphasis when $R(1)/R(0)$ is between zero and 0.6 and between 0.85 and 0.95, but larger area distances otherwise. Smaller area distances are observed for the two/three adaptive pre-emphasis than no pre-emphasis when $R(1)/R(0)$ is between zero and 0.6.

The variation in area distances is shown to be much smaller in Figure 5.20 for a two/three adaptive pre-emphasis filter than for no pre-emphasis and the other pre-emphases. This observation



(a)

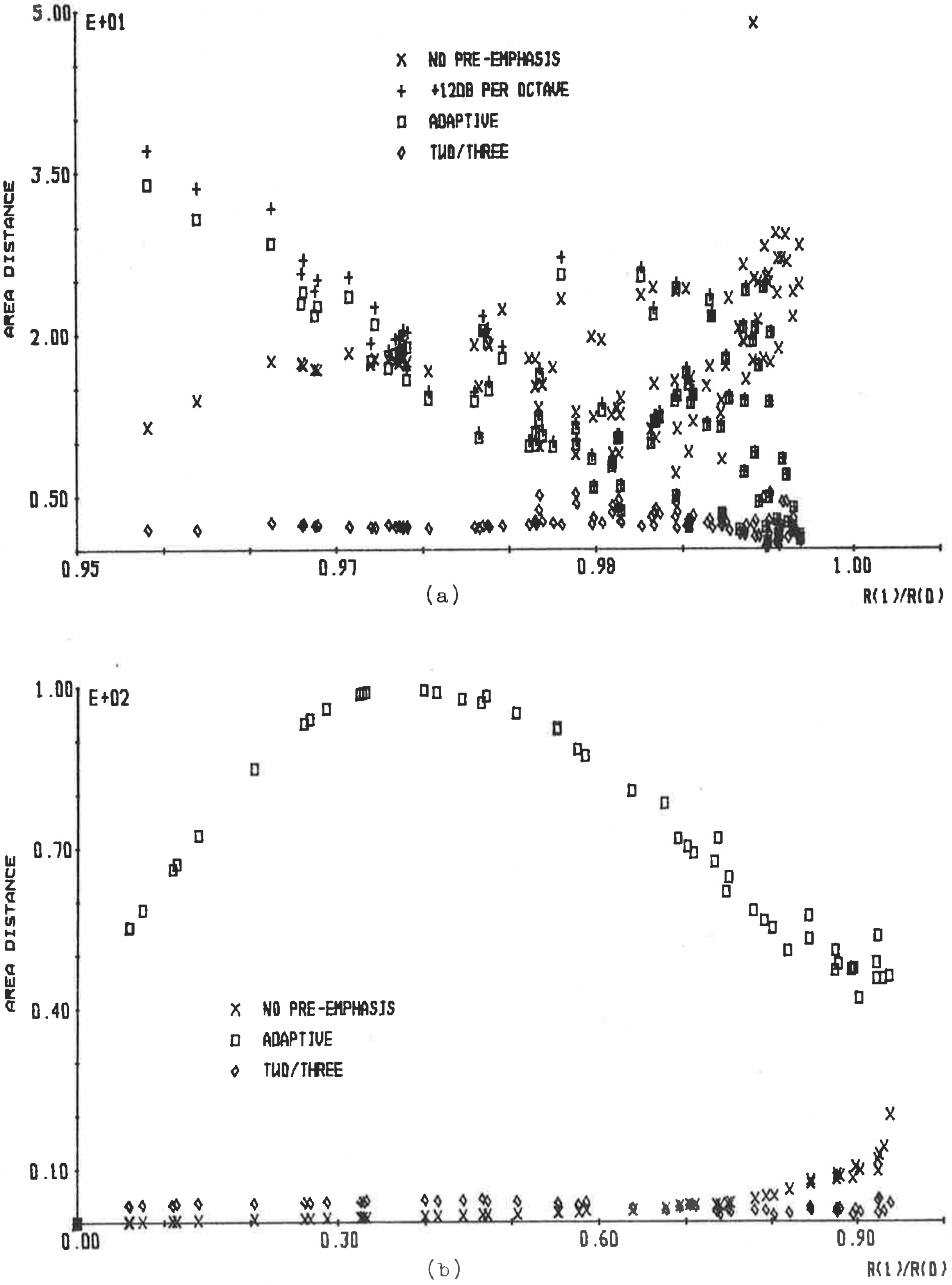
(b)

FIGURE 5.20: Area distances for the analysis of synthetic speech for the vowel /i/ with R(1)/R(0) (a) greater than 0.95 and (b) less than 0.95.

is consistent with the observations for synthetic speech of the vowels |a| and |e| presented previously in this section. In general, the two/three adaptive pre-emphasis filter has been shown to provide a significant reduction in area distances when compared with no pre-emphasis and a +12 dB per octave pre-emphasis of synthetic speech for the vowel |i|. Except for a small range of $R(1)/R(0)$, the two/three adaptive pre-emphasis produces smaller area distances than an unvoiced/voiced adaptive pre-emphasis of synthetic speech for the vowel |i|.

Synthetic speech for the vowel |o| is generated in the same manner as for the vowels |a|, |e| and |i|. The glottal pulse waveforms used to generate the synthetic speech for the vowel |o| are the same as those used to generate the synthetic speech for the vowels |a|, |e| and |i|. Figure 5.21 presents the area distances versus $R(1)/R(0)$ for no pre-emphasis, a +12 dB per octave pre-emphasis, an unvoiced/voiced adaptive pre-emphasis (Figure 5.21(a) only), and a two/three adaptive pre-emphasis of synthetic speech for the vowel |o|.

The area distances for a two/three adaptive pre-emphasis are shown in Figure 5.21(a) to be significantly less than no pre-emphasis and the other pre-emphases when $R(1)/R(0)$ is between 0.95 and approximately unity. The unvoiced/voiced adaptive pre-emphasis and a +12 dB per octave pre-emphasis are shown in Figure 5.21(a) to provide smaller area distnaces than no pre-emphasis only when $R(1)/R(0)$ is close to unity. The variation in area distances of the two/three adaptive pre-emphasis filter is much smaller than the other two pre-emphasis techniques and no pre-emphasis. The above observations are consistent with the observations presented



(a)

(b)

FIGURE 5.21: Area distances for the analysis of synthetic speech for the vowel /o/ with R(1)/R(0) (a) greater than 0.95 and (b) less than 0.95

previously in this section for synthetic speech of the vowels |a|, |e| and |i|, when $R(1)/R(0)$ is between 0.95 and unity.

Figure 5.21(b) shows that the unvoiced/voiced adaptive pre-emphasis of synthetic speech for the vowel |o| produces much larger area distances than a two/three adaptive pre-emphasis and no pre-emphasis. The two/three adaptive pre-emphasis filter is shown in Figure 5.21(b) to produce smaller area distances than no pre-emphasis when $R(1)/R(0)$ is between 0.6 and 0.95, but larger area distances when $R(1)/R(0)$ is between zero and 0.6. Figure 5.22 presents a comparison of recovered and original acoustic tube shapes for no pre-emphasis and a two/three adaptive pre-emphasis of synthetic speech for the vowel |o|, when a glottal pulse waveform with $R(1)/R(0)$ equal to 0.33 is used as the excitation. The recovered acoustic tube shape presented in Figure 5.22 is representative of those for $R(1)/R(0)$ in the range zero to 0.6. Figure 5.22 shows that, while the two/three adaptive pre-emphasis produces slightly larger area distances than no pre-emphasis, good acoustic tube shape recovery is achieved.

In general, the area distances for a two/three adaptive pre-emphasis have a much smaller variation than the area distances for no pre-emphasis or the other two pre-emphasis techniques. This observation is consistent with the observations for synthetic speech of the vowels |a|, |e| and |i| presented previously in this section. The two/three adaptive pre-emphasis of synthetic speech for the vowel |o| has been shown to be much smaller than the area distances for a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis. A reduction of area distances by the two/three adaptive pre-emphasis in comparison with no pre-emphasis only oc-

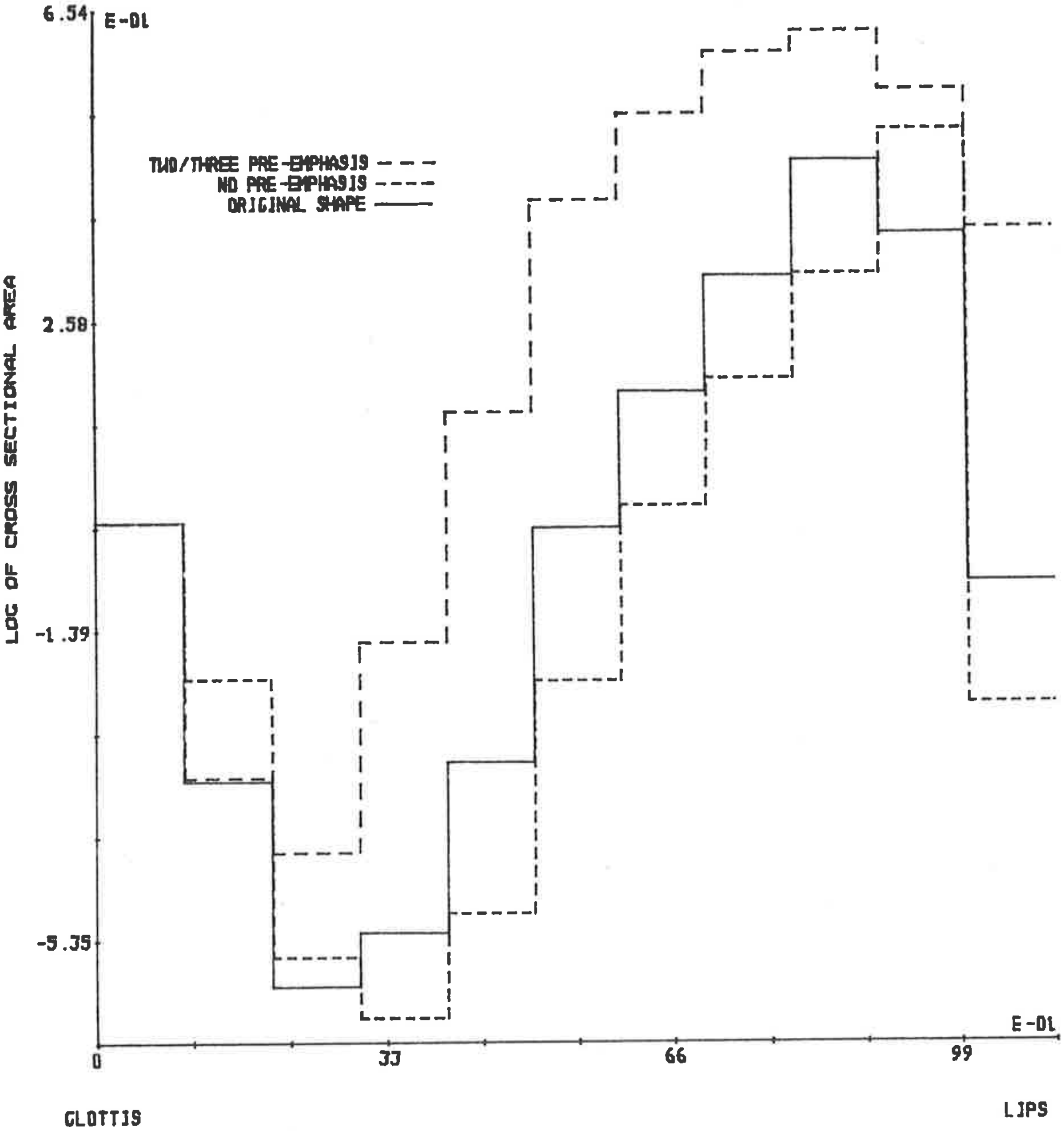


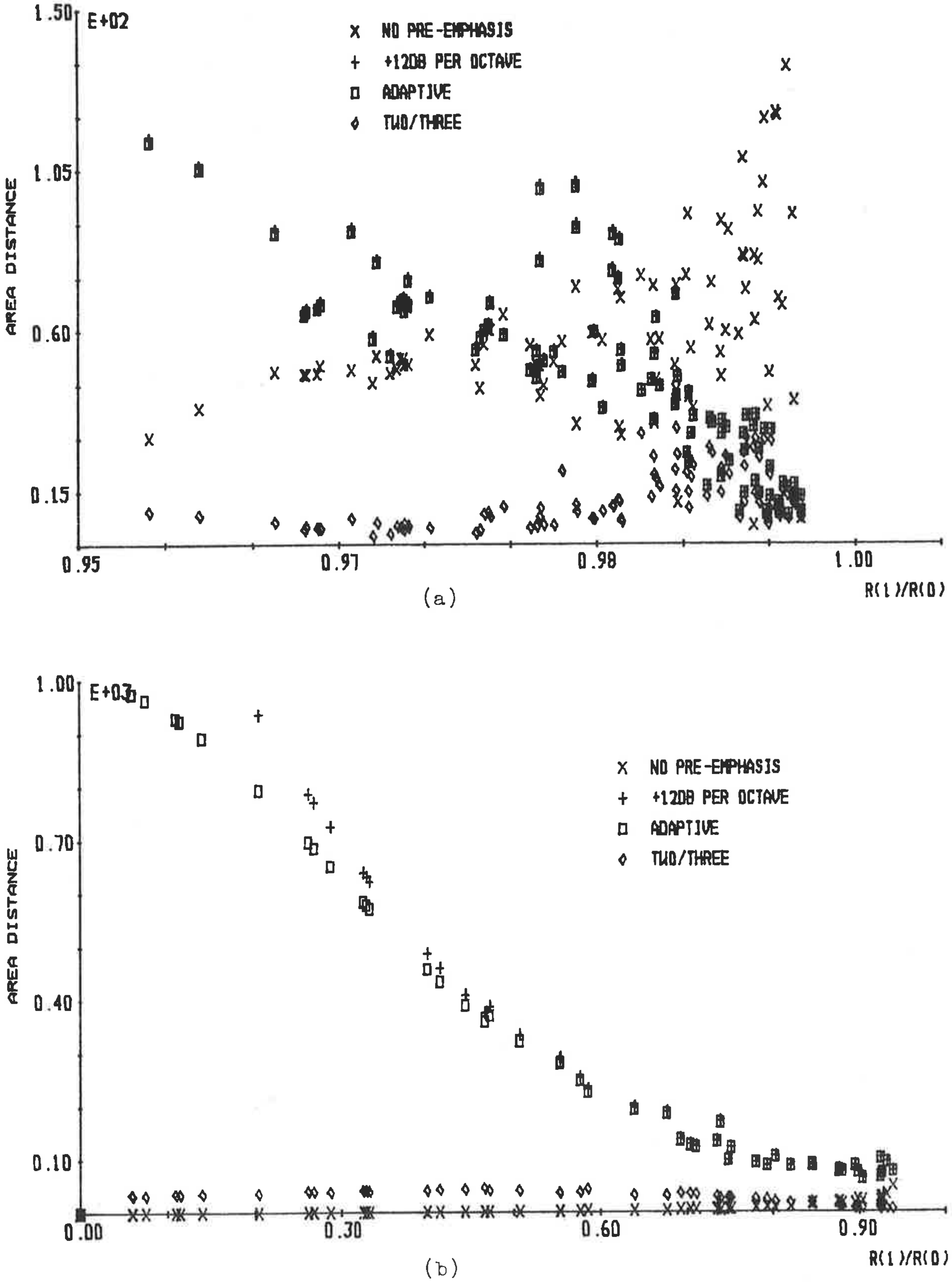
FIGURE 5.22: Comparison of original acoustic tube shape with those recovered by no pre-emphasis and a two/three adaptive pre-emphasis of synthetic speech for the vowel /o/ with R(1)/R(0) = 0.33.

curs when $R(1)/R(0)$ is between 0.6 and unity, but good acoustic tube shape recovery is still achieved by the two/three adaptive pre-emphasis when $R(1)/R(0)$ is between zero and 0.6.

Figure 5.23 presents the area distances versus $R(1)/R(0)$ for no pre-emphasis, a +12 dB per octave pre-emphasis, an unvoiced/voiced adaptive pre-emphasis, and a two/three adaptive pre-emphasis of synthetic speech for the vowel |u|. The glottal pulse waveforms used to generate the synthetic speech for the vowel |u| are the same as those used to generate synthetic speech for the vowels |a|, |e|, |i| and |o|.

Figure 5.23(a) shows that the area distances for a two/three adaptive pre-emphasis of synthetic speech for the vowel |u| are significantly smaller than the area distances for no pre-emphasis and the other two pre-emphasis techniques when $R(1)/R(0)$ is between 0.95 and 1.0. The unvoiced/voiced adaptive pre-emphasis and a +12 dB per octave pre-emphasis only provide smaller area distances than no pre-emphasis when $R(1)/R(0)$ is close to unity. The variation in area distances of the two/three adaptive pre-emphasis is much smaller than for no pre-emphasis and the other two pre-emphasis techniques. These observations are consistent with the observations presented previously for the vowels |a|, |e|, |i| and |o|.

The unvoiced/voiced adaptive pre-emphasis and the +12 dB per octave pre-emphasis are shown in Figure 5.23(b) to produce very much larger area distances than a two/three adaptive pre-emphasis and no pre-emphasis when $R(1)/R(0)$ is between zero and 0.95. Figure 5.23(b) shows that a reduction in area distances by the two/three adaptive pre-emphasis filter does not occur in compari-



(a)

(b)

FIGURE 5.23: Area distances for the analysis of synthetic speech for the vowel /u/ with R(1)/R(0) (a) greater than 0.95 and (b) less than 0.95.

son with no pre-emphasis when $R(1)/R(0)$ is between zero and 0.85. Original and recovered acoustic tube shapes are presented in Figure 5.24 for no pre-emphasis and a two/three adaptive pre-emphasis of synthetic speech for the vowel |u|, with the glottal pulse excitation having $R(1)/R(0) = .58$. The recovered acoustic tube shape of Figure 5.24 is representive of the acoustic tube shapes recovered by a two/three adaptive pre-emphasis of synthetic speech for the vowel |u| when $R(1)/R(0)$ is between zero and 0.85. Figure 5.24 shows that poor acoustic tube recovery occurs for the two/three adaptive pre-emphasis filter when $R(1)/R(0)$ is between zero and 0.85.

In general, the area distances for a two/three adaptive pre-emphasis of synthetic speech for the vowel |u| have a much smaller variation than the area distances for no pre-emphasis and the other two pre-emphasis techniques. This is consistent with the observations for synthetic speech of the vowels |a|, |e|, |i| and |o| presented previously. The area distances for a two/three adaptive pre-emphasis of synthetic speech for the vowel |u| have been shown to be consistently smaller than for a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis. A reduction in area distances by the two/three adaptive pre-emphasis in comparison with no pre-emphasis does not occur when $R(1)/R(0)$ is between zero and 0.85, and for $R(1)/R(0)$ in this range poor acoustic tube shape recovery occurs for the two/three adaptive pre-emphasis filter.

The evaluations performed in this section with synthetic speech for the vowels |a|, |e|, |i|, |o| and |u| have shown that, in all but a few cases, a significant reduction in area distances

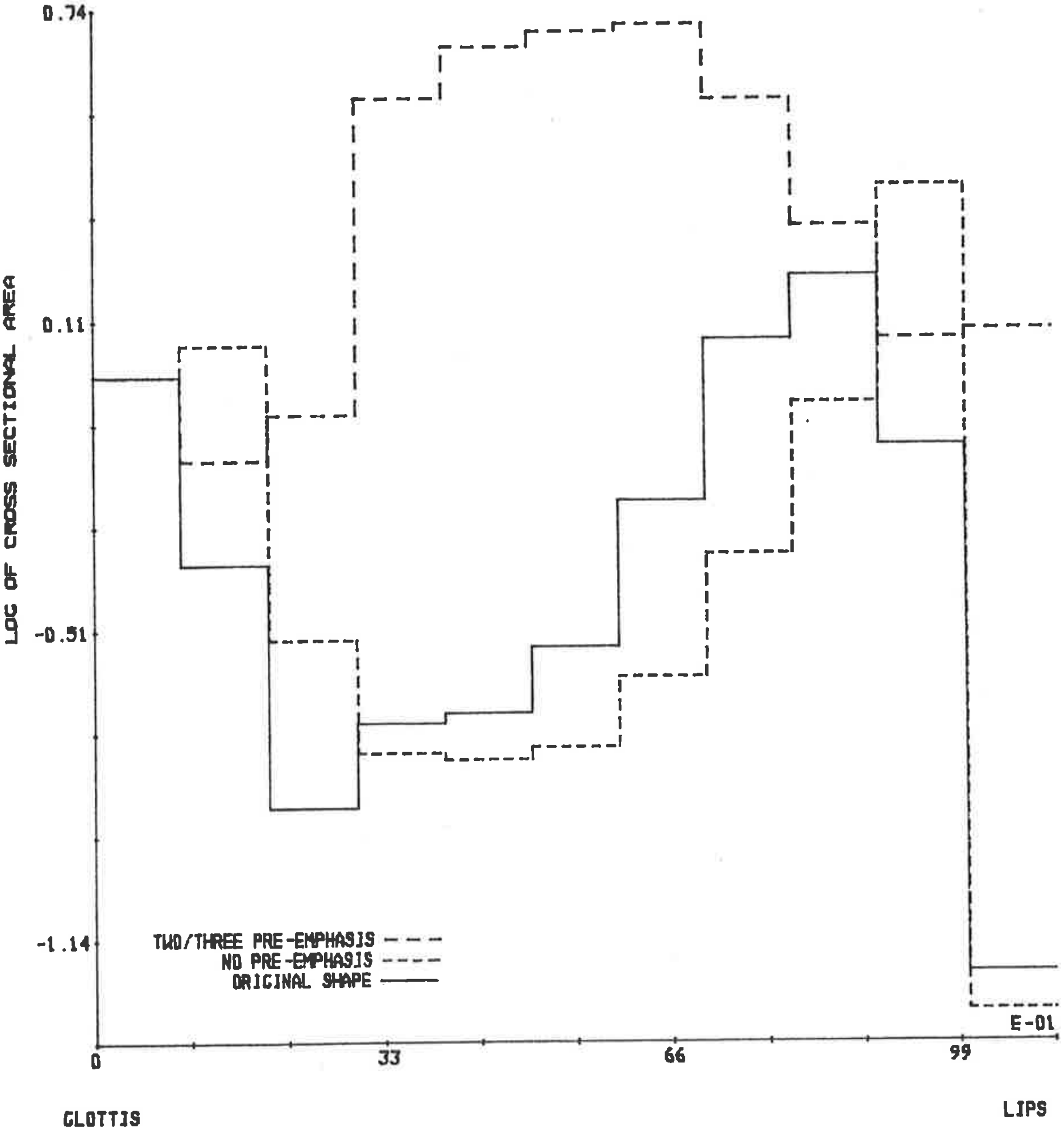


FIGURE 5.24: Comparison of original acoustic tube shape and those recovered by no pre-emphasis and a two/three adaptive pre-emphasis of synthetic speech for the vowel /u/, with R(1)/R(0) = 0.577.

is achieved by the two/three adaptive pre-emphasis filter in comparison with a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis. When $R(1)/R(0)$ is close to unity, then the two/three adaptive pre-emphasis, a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis all produce similar area distances. This is expected as, for $R(1)/R(0)$ close to unity, all apply approximately +12 dB per octave pre-emphasis.

For all the evaluation results presented in this section, the two/three adaptive pre-emphasis filter was shown to provide area distances which have a small variation in comparison with the area distances for no pre-emphasis, a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis when $R(1)/R(0)$ is between zero and approximately unity. Therefore, the acoustic tube shape recovered by a two/three adaptive pre-emphasis is affected less by changes in glottal pulse excitation, and so is better at removing glottal pulse excitation effects from the recovered acoustic tube shape than the other pre-emphases considered.

When $R(1)/R(0)$ is between 0.95 and approximately unity, then consistent general trends of area distances versus $R(1)/R(0)$ are observed for the all the pre-emphases of synthetic speech for the vowels |a|, |e|, |i|, |o| and |u|. One general trend is for the the two/three adaptive pre-emphasis to provide smaller area distances than no pre-emphasis, a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis. When $R(1)/R(0)$ is close to unity, then similar area distances occur for both a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis, which are less than the area distances for no pre-emphasis.

Different trends of area distance versus $R(1)/R(0)$ are observed for the two/three and the unvoiced/voiced adaptive pre-emphases when $R(1)/R(0)$ is between zero and 0.95. For synthetic speech of the vowels |a|, |e| and |i|, the two/three adaptive pre-emphasis filter provides a significant reduction in area distances compared with no pre-emphasis when $R(1)/R(0)$ is between zero and 0.95. However, larger area distances occur for a two/three adaptive pre-emphasis of synthetic speech for the vowels |o| and |u| when $R(1)/R(0)$ is less than approximately 0.8. In the case of synthetic speech for the vowel |o|, the two/three adaptive pre-emphasis still produces good acoustic tube shape recovery, but this is not the case for the vowel |u|. This may be remedied by redefining the relationship between the parameter α of the two/three adaptive pre-emphasis filter and $R(1)/R(0)$.

The two/three adaptive pre-emphasis filter was defined in Chapter 4 using $R(1)/R(0)$ evaluated from the glottal pulse waveform, but in this section it has been necessary to evaluate $R(1)/R(0)$ from the synthetic speech waveform. The results presented in this section have shown that the correctness of using $R(1)/R(0)$ from the autocorrelation function of the synthetic speech waveform is dependent on the acoustic tube shape used to generate the synthetic speech. However, except for a few cases, the two/three adaptive pre-emphasis filter has been shown to provide significant reductions in area distances compared with no pre-emphasis, a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis when $R(1)/R(0)$ is calculated from the synthetic speech waveform.

The +12 dB per octave pre-emphasis and the unvoiced/voiced adaptive pre-emphasis were also evaluated in this section by being compared with each other, no pre-emphasis and a two/three adaptive pre-emphasis. In general, the +12 dB per octave pre-emphasis only produces small area distances, which indicates good acoustic tube shape recovery when $R(1)/R(0)$ is near unity. The unvoiced/voiced adaptive pre-emphasis filter only produces small area distances when $R(1)/R(0)$ is near unity and close to zero, and very poor acoustic tube shape recovery occurs otherwise.

The evaluations presented in this section have shown that, when the two/three adaptive pre-emphasis is defined by the parameter α, then significant reductions in area distances occur when compared with no pre-emphasis, a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis of synthetic speech for the vowels |a|, |e|, |i|, |o| and |u|.

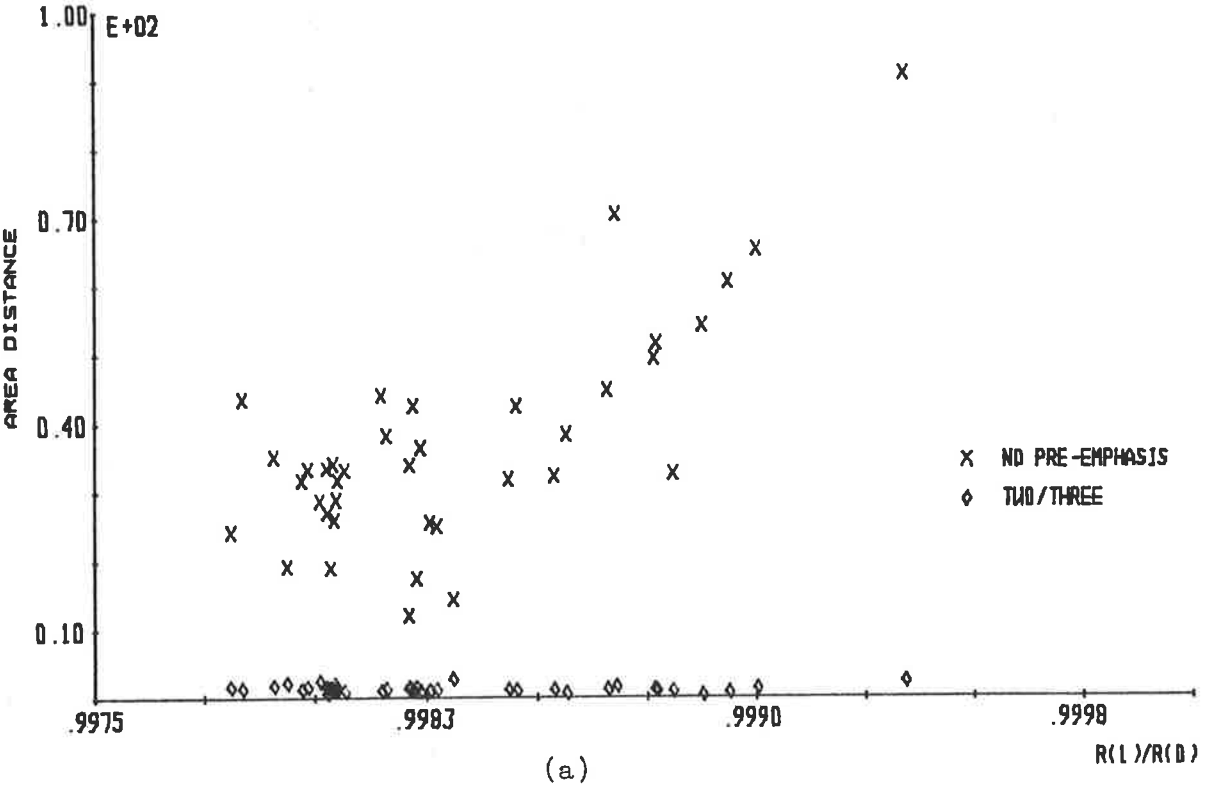
5.3.2 PARAMETER β

This section details an evaluation of the two/three adaptive pre-emphasis filter with synthetic speech data when the filter is defined by the parameter β. The values of β used by the two/three adaptive pre-emphasis filter are determined via a relationship with $R(1)/R(0)$, which is evaluated from the autocorrelation function of the synthetic speech waveform. The form and details of the relationship between β and $R(1)/R(0)$ are defined and discussed in Chapter 4.

All the evaluations recorded in this section are presented as area distances between recovered and original acoustic tube shapes and plotted against $R(1)/R(0)$, which is evaluated from the glottal pulse waveform used in the generation of the synthetic speech waveforms. The value of $R(1)/R(0)$ of the glottal pulse waveform is used so that all the results are related to the glottal pulse waveforms which the two/three adaptive pre-emphasis filter is designed to remove.

The synthetic speech waveforms used in the evaluations presented in this section are generated by the procedure detailed in Appendix B. To ensure the synthetic speech waveforms have properties which are as close as possible to real speech waveforms, the acoustic tube shape used in the synthetic speech generation process approximate the vocal tract shapes for five vowels, |a|, |e|, |i|, |o| and |u|, as measured by FANT [1970] (see Appendix C also). To avoid radiation effects clouding the evaluations presented in this section, the termination reflection coefficient of the acoustic tube model, used to generate the synthetic speech, is assumed as unity, i.e. $\mu_M = 1$. All the glottal pulse waveforms used to generate the synthetic speech for the evaluations presented in this section are digitized glottal pulse waveforms measured by MONSEN and ENGEBRETSON [1977], ROTHENBERG [1973], SONDHI [1975], SUNDBERG and GAUFFIN [1978], MILLER [1959], and FLANAGAN and LANDGRAF [1968].

Figure 5.25 presents area distances versus $R(1)/R(0)$ for synthetic speech of the vowel |a| for no pre-emphasis, a +12 dB per octave pre-emphasis, an unvoiced/voiced adaptive pre-emphasis and a two/three adaptive pre-emphasis followed by a



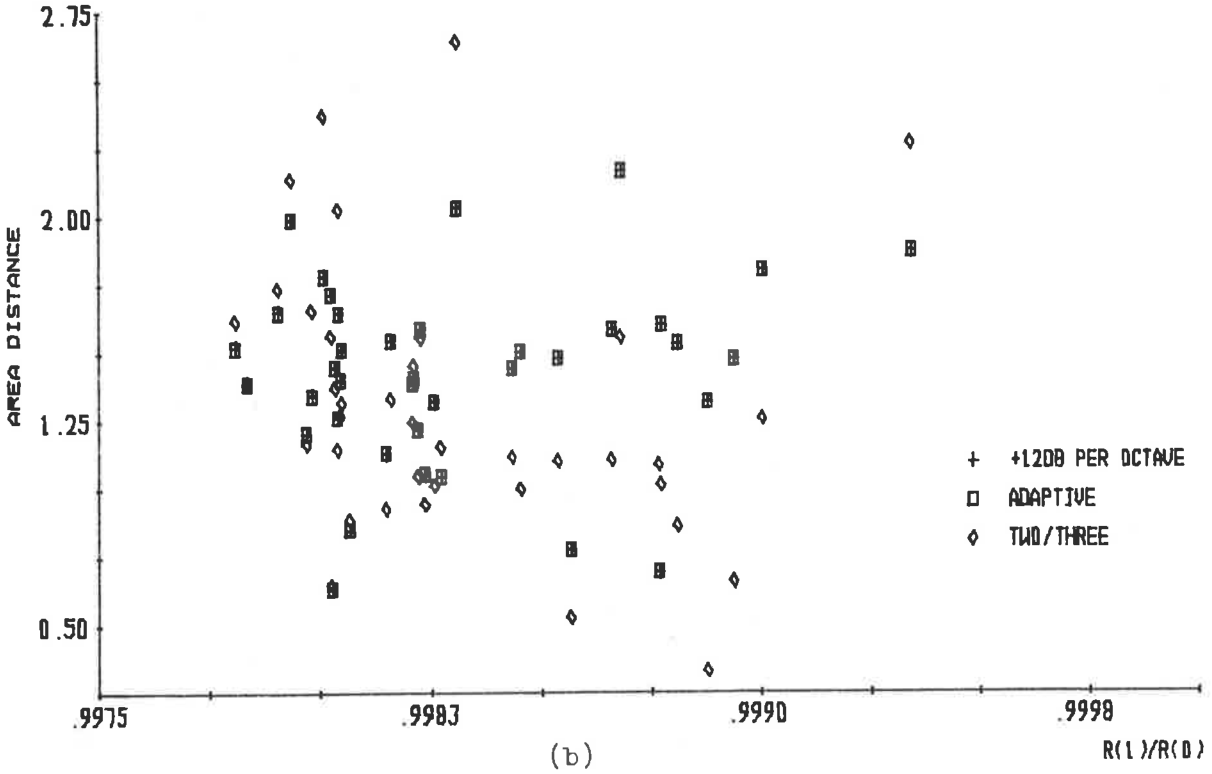


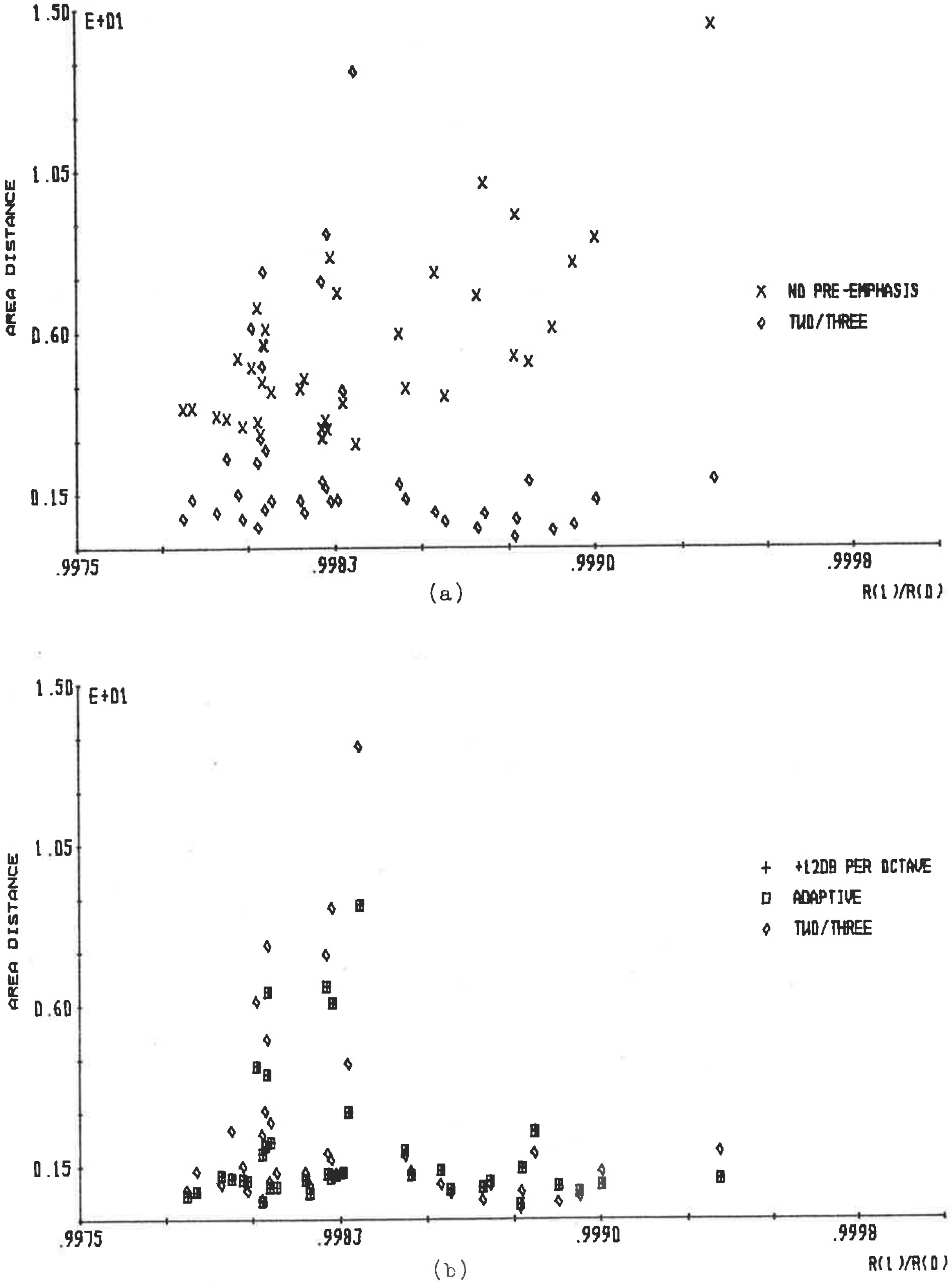
FIGURE 5.25: Area distances for an analysis of synthetic speech for the vowel /a/ by (a) no pre-emphasis and a two/three adaptive pre-emphasis and (b) a +12dB per octave, an unvoiced/voiced adaptive and two/three adaptive pre-emphases.

conventional Parcor analysis. Figure 5.25(a) shows that the two/three adaptive pre-emphasis provides a large reduction in area distances when compared with the area distances for no pre-emphasis of the synthetic speech for the vowel |a|.

A comparison of area distances for the two/three adaptive pre-emphasis filter and those for a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis presented in Figure 5.25(b) shows that, in most cases, a small reduction in area distances is achieved by the two/three adaptive pre-emphasis filter. This reduction in area distances by the two/three adaptive pre-emphasis filter increases as the value of $R(1)/R(0)$ increases.

Area distances versus $R(1)/R(0)$ for no pre-emphasis, a +12 dB per octave pre-emphasis, an unvoiced/voiced adaptive pre-emphasis and a two/three adaptive pre-emphasis of synthetic speech for the vowel |e| followed by a Parcor analysis are presented in Figure 5.26. Figure 5.26(a) shows that larger area distances may occur for a two/three adaptive pre-emphasis in comparison with no pre-emphasis but, in general, a significant reduction in area distances occurs for the two/three adaptive pre-emphasis of synthetic speech for the vowel |e|.

Figure 5.26(b) compares the area distances for a two/three adaptive pre-emphasis and the other pre-emphases, and shows that similar area distances occur for each pre-emphasis technique. In this situation, no overall increase or reduction in area distances is achieved by the two/three adaptive pre-emphasis filter. Hence, for synthetic speech of the vowel |e|, the only benefit gained by using the two/three adaptive pre-emphasis filter, when defined by



(a)

(b)

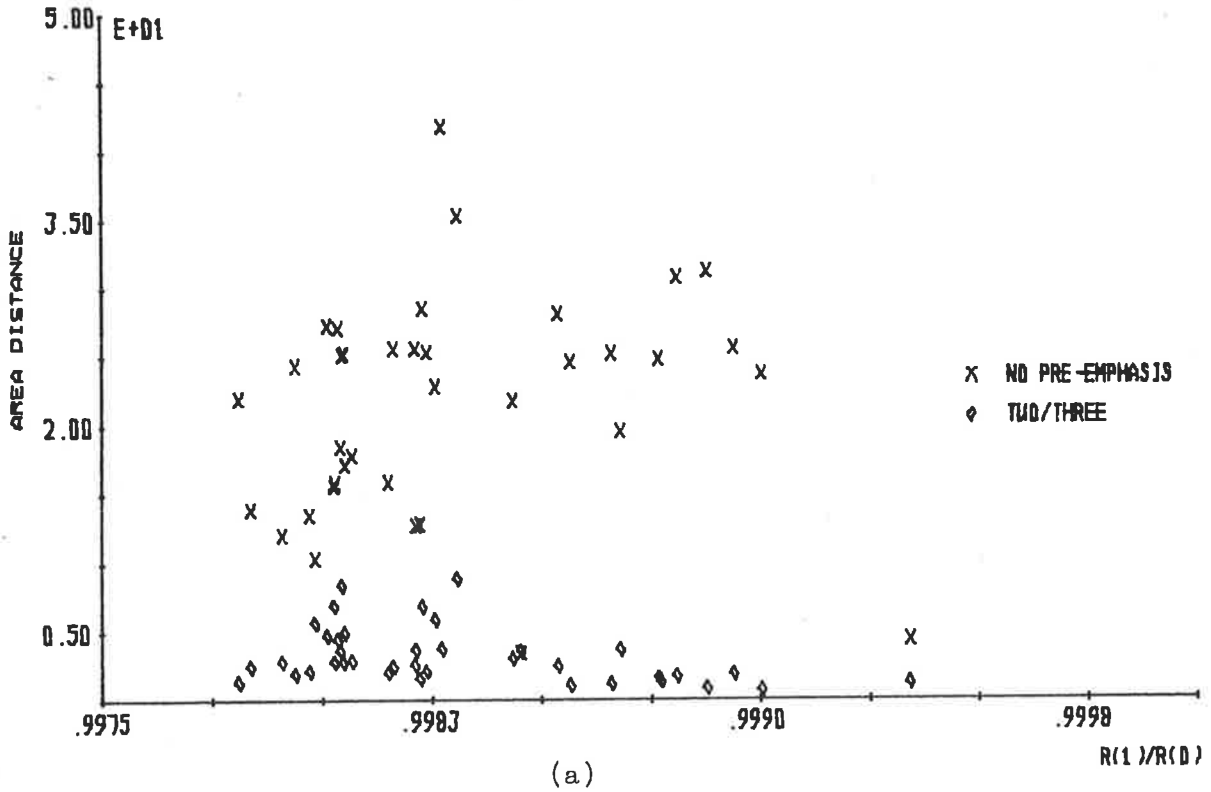
FIGURE 5.26: Area distances for analysis of synthetic speech for the vowel /e/ by (a) no pre-emphasis and a two/three adaptive pre-emphasis and (b) a +12dB per octave, an unvoiced/voiced adaptive and two/three adaptive pre-emphases.

parameter β, is a reduction in area distances when compared with no pre-emphasis.

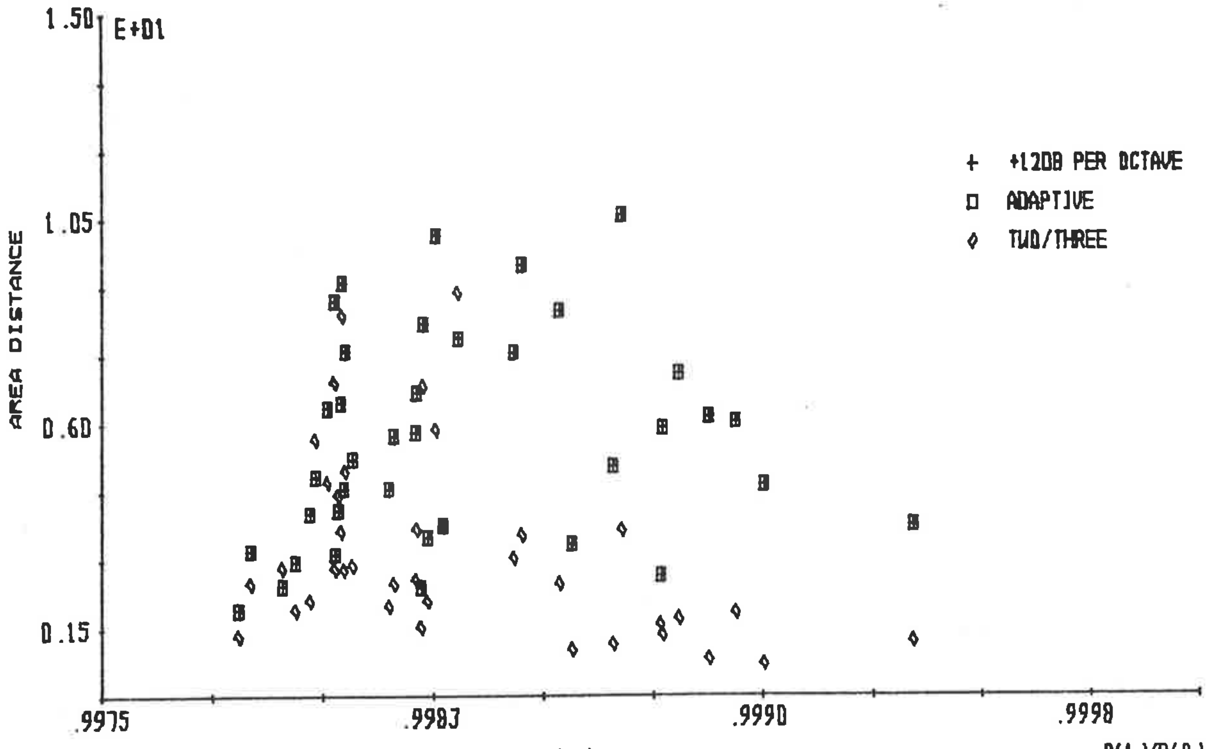
Figure 5.27 presents area distances versus $R(1)/R(0)$ for synthetic speech of the vowel |i| with no pre-emphasis, a +12 dB per octave pre-emphasis and unvoiced/voiced and two/three adaptive pre-emphases followed by a conventional Parcor analysis. The two/three adaptive pre-emphasis filter is shown in Figure 5.27(a) to produce a large reduction in area distances when compared with the area distances for no pre-emphasis.

A small reduction of area distances is observed in Figure 5.27(b) for the two/three adaptive pre-emphasis filter when compared with the area distances for a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis. The amount of area distance reduction is shown in Figure 5.27(b) to increase with increasing value of $R(1)/R(0)$, which is consistent with the evaluation results presented in Figure 5.25(b) for synthetic speech of the vowel |a|.

For synthetic speech of the vowel |o|, the area distances versus $R(1)/R(0)$ are presented in Figure 5.28 for no pre-emphasis, a +12 dB per octave pre-emphasis and unvoiced/voiced and two/three adaptive pre-emphases followed by a conventional Parcor analysis. A large reduction of area distances is observed in Figure 5.28(a) for the two/three adaptive pre-emphasis in comparison with no pre-emphasis. This observation is consistent with the results presented for synthetic speech of the vowels |a|, |e| and |i|.

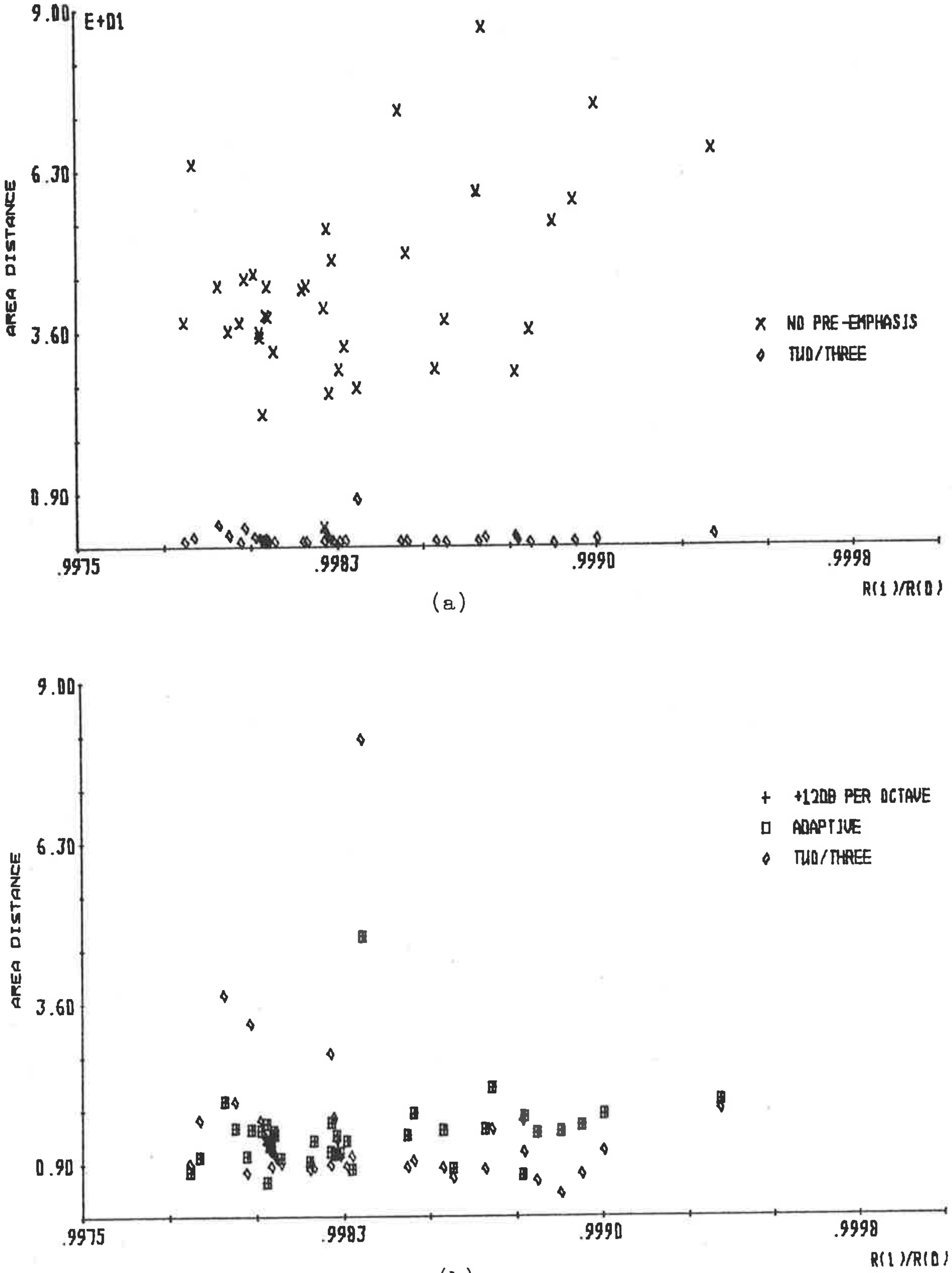


(a)



(b)

FIGURE 5.27: Area distances for analysis of synthetic speech for the vowel /i/ by (a) no pre-emphasis and a two/three adaptive pre-emphasis and (b) a +12dB per octave, an unvoiced/voiced adaptive and two/three adaptive pre-emphases.



(a)

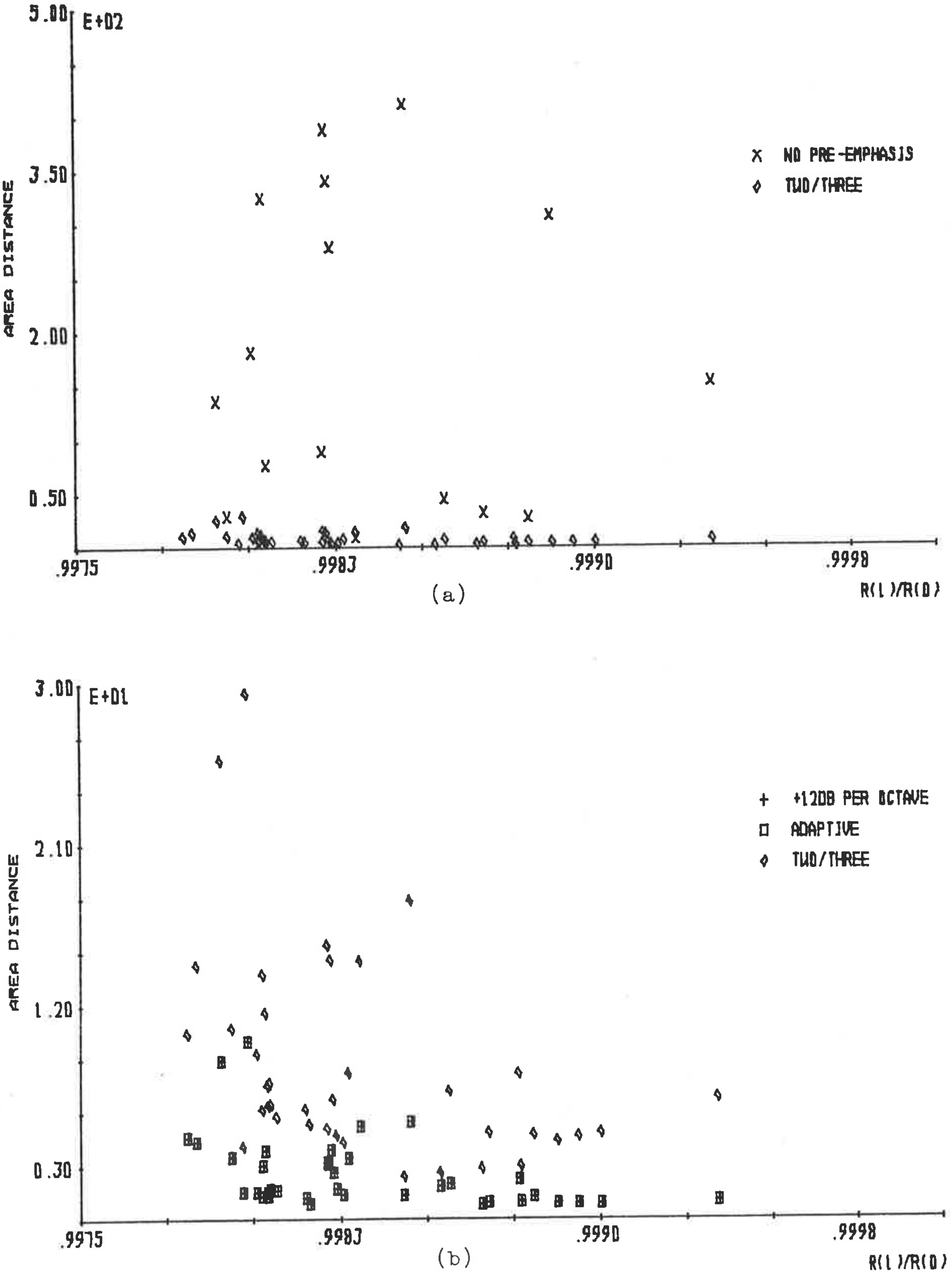
(b)

FIGURE 5.28: Area distances for analysis of synthetic speech for the vowel /o/ by (a) no pre-emphasis and a two/three adaptive pre-emphasis and (b) a +12dB per octave, an unvoiced/voiced adaptive and two/three adaptive pre-emphases.

In general, a small reduction in area distances occurs for the two/three adaptive pre-emphasis in comparison with a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis, as shown in Figure 5.28(b). In some cases, a slightly larger area distance is returned by the two/three adaptive pre-emphasis; therefore, a reduction in area distance by the two/three adaptive pre-emphasis filter in comparison with existing pre-emphasis techniques cannot always be guaranteed for the vowel |o|.

Area distances versus $R(1)/R(0)$ are presented in Figure 5.29 for synthetic speech of the vowel |u| pre-emphasized by no pre-emphasis, a +12 dB per octave pre-emphasis, an unvoiced/voiced and two/three adaptive pre-emphases, followed by a conventional Parcor analysis. A comparison of area distances in Figure 5.29(a) shows significantly smaller area distances for a two/three adaptive pre-emphasis than for no pre-emphasis.

The area distances presented in Figure 5.29(b) show that the two/three adaptive pre-emphasis filter provides significantly larger area distances than a +12 dB per octave pre-emphasis or an unvoiced/voiced adaptive pre-emphasis. This result contrasts with those found for synthetic speech of the vowels |a|, |e|, |i| and |o|, where the area distances for a two/three adaptive pre-emphasis are mostly smaller than those for the conventional pre-emphasis techniques. Hence, the two/three adaptive pre-emphasis filter, when defined by the parameter β, has an inferior performance when used on synthetic speech for the vowel |u|.



(a)

(b)

FIGURE 5.29: Area distances for analysis of synthetic speech for the vowel /u/ by (a) no pre-emphasis and a two/three adaptive pre-emphasis and (b) a +12dB per octave, an unvoiced/voiced adaptive and two/three adaptive pre-emphases.

The results presented in this section have shown that the two/three adaptive pre-emphasis filter, when defined by the parameter β, provides a large reduction in area distances when compared with the area distances for no pre-emphasis of synthetic speech for all the vowels |a|, |e|, |i|, |o| and |u|. For synthetic speech of the vowels |a|, |i| and |o|, a general reduction in area distances is found when compared with the area distances for a +12 dB per octave and unvoiced/voiced adaptive pre-emphases. However, the two/three adaptive pre-emphasis filter does not provide any significant reduction of area distances for the synthetic speech of the vowel |e|, and provides a significant increase in area distances for the synthetic speech of the vowel |u|.

Comparing the evaluation results for the two/three adaptive pre-emphasis of synthetic speech, when the filter is defined by the parameter α and then the parameter β, reveals that the improvements gained in comparison with a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis are more significant when the filter is defined by the parameter α than when defined by the parameter β. In general, the two/three adaptive pre-emphasis filter performs worse for synthetic speech of the vowel |u|.

5.3.3 SAMPLING FREQUENCIES

The relationships between the parameter α and $R(1)/R(0)$ and the parameter β and $R(1)/R(0)$ which define the two/three adaptive pre-emphasis filter have been shown, in Chapter 4, to be dependent on the waveform sampling frequency. The relationships between α and $R(1)/R(0)$ and between β and $R(1)/R(0)$ were defined in Chapter 4 for waveform sampling frequencies of 10, 9, 8 and 7 kilohertz.

The two/three adaptive pre-emphasis filter has been evaluated for a sampling frequency of 10 kilohertz in Sections 5.3.1 and 5.3.2, and this section performs an evaluation for waveform sampling frequencies of 9, 8 and 7 kilohertz.

Since the effects of waveform sampling frequency changes are different for the two independent parts of the two/three adaptive pre-emphasis filter, i.e. that defined by the parameter α and that defined by the parameter β, each part is evaluated separately. Sections 5.3.3.1 and 5.3.3.2 present the evaluations of the two/three adaptive pre-emphasis filter when defined by the parameters α and β, respectively, for synthetic speech of five vowels sampled at 9, 8 and 7 kilohertz.

5.3.3.1 Parameter α

The relationship between the parameter α of the two/three adaptive pre-emphasis filter and $R(1)/R(0)$ was shown in Chapter 4 to be dependent on the waveform sampling frequency. This section presents an evaluation of the two/three adaptive pre-emphasis filter when defined by the parameter α and used to pre-emphasize synthetic speech of five vowel sounds which are sampled at frequencies of 9, 8 and 7 kilohertz.

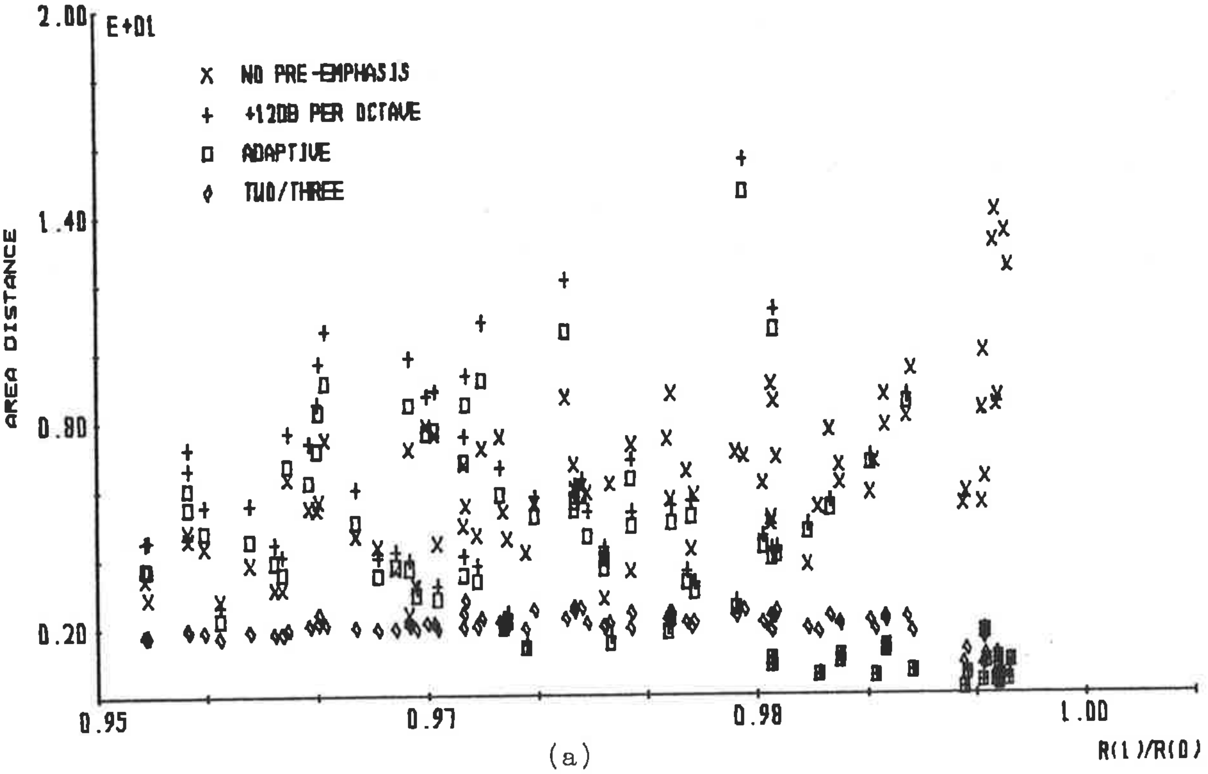
The purpose of this section is to determine whether the general trends observed for the two/three adaptive pre-emphasis filter, in comparison with existing pre-emphasis techniques for a waveform sampling frequency of 10 kilohertz, as presented in Section 5.3.1, occur for waveform sampling frequencies of 9, 8 and 7 kilohertz. The variations in the area distances for a particular waveform pre-emphasized by the two/three adaptive pre-emphasis filter and then analysed by a Parcor analysis, for a waveform sampling frequency change, is presented. The relationship between parameter α and $R(1)/R(0)$ used to define the two/three adaptive pre-emphasis filter is the same for waveform sampling frequencies of 10, 9 and 8 kilohertz, but different for a waveform sampling frequency of 7 kilohertz.

All the evaluations of the two/three adaptive pre-emphasis filter are presented as area distances between recovered and original acoustic tube shapes versus the value of $R(1)/R(0)$, evaluated from the glottal pulse waveform used to generate the synthetic speech. The value of $R(1)/R(0)$ of the glottal pulse waveform is used so that all the results are related to the glottal pulse waveforms which the two/three adaptive pre-emphasis filter is designed to remove. The autocorrelation function of the synthetic speech waveform is used to determine the value of $R(1)/R(0)$ which defines the two/three adaptive pre-emphasis filter.

The synthetic speech waveforms are generated by the procedure described in Appendix B. To ensure the synthetic speech waveforms are as close as possible to real speech waveforms, the acoustic tube shape used to generate the synthetic speech approximates the vocal tract shapes for the five vowels |a|, |e|, |i|, |o| and |u|, as defined in Appendix C. The glottal pulse waveforms used to excite the acoustic tube shape are the same as those used in Sections 5.2.1 and 5.3.1, being generated from the glottal pulse models of ROSENBERG [1973] and FANT [1979], or derived from the measured glottal pulse waveforms of ROSENBERG [1973], MONSEN and ENGEBRETSON [1977], MILLER [1959], SONDHI [1975], and SUNDBERG and GAUFFIN [1978]. To avoid radiation effects clouding the evaluations presented in this section, the termination reflection coefficient of the acoustic tube model is unity, i.e. $\mu_M = 1$.

Figure 5.30 presents the area distances versus $R(1)/R(0)$ for no pre-emphasis, a +12 dB per octave pre-emphasis, an unvoiced/voiced adaptive pre-emphasis and a two/three adaptive pre-emphasis of synthetic speech for the vowel |a| for sampling frequencies of 9, 8 and 7 kilohertz. The results presented in Figure 5.30 show that the two/three adaptive pre-emphasis filter produces a reduction in area distances compared with those of no pre-emphasis for $R(1)/R(0)$ between zero and approximately unity, and the +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis for $R(1)/R(0)$ between zero and approximately 0.97. When $R(1)/R(0)$ is greater than 0.97, the +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis produce smaller area distances than the two/three adaptive pre-emphasis. Although the amount by which the two/three adaptive pre-emphasis filter provides a reduction in area distance when compared with other pre-emphases may change slightly, the above observations are consistent with the evaluations presented in Figure 5.18 for a waveform sampling frequency of 10 kilohertz.

Figure 5.31 presents the area distances for the two/three adaptive pre-emphasis of synthetic speech for the vowel |a| versus $R(1)/R(0)$ for the four waveform sampling frequencies 10, 9, 8 and 7 kilohertz. Different symbols are used to indicate the different sampling frequencies for the data points in Figure 5.31, and the even spread of symbols indicates that similar ranges of area distances are returned by the two/three adaptive pre-emphasis filter if the sampling frequency is between 10 and 7 kilohertz, inclusive.



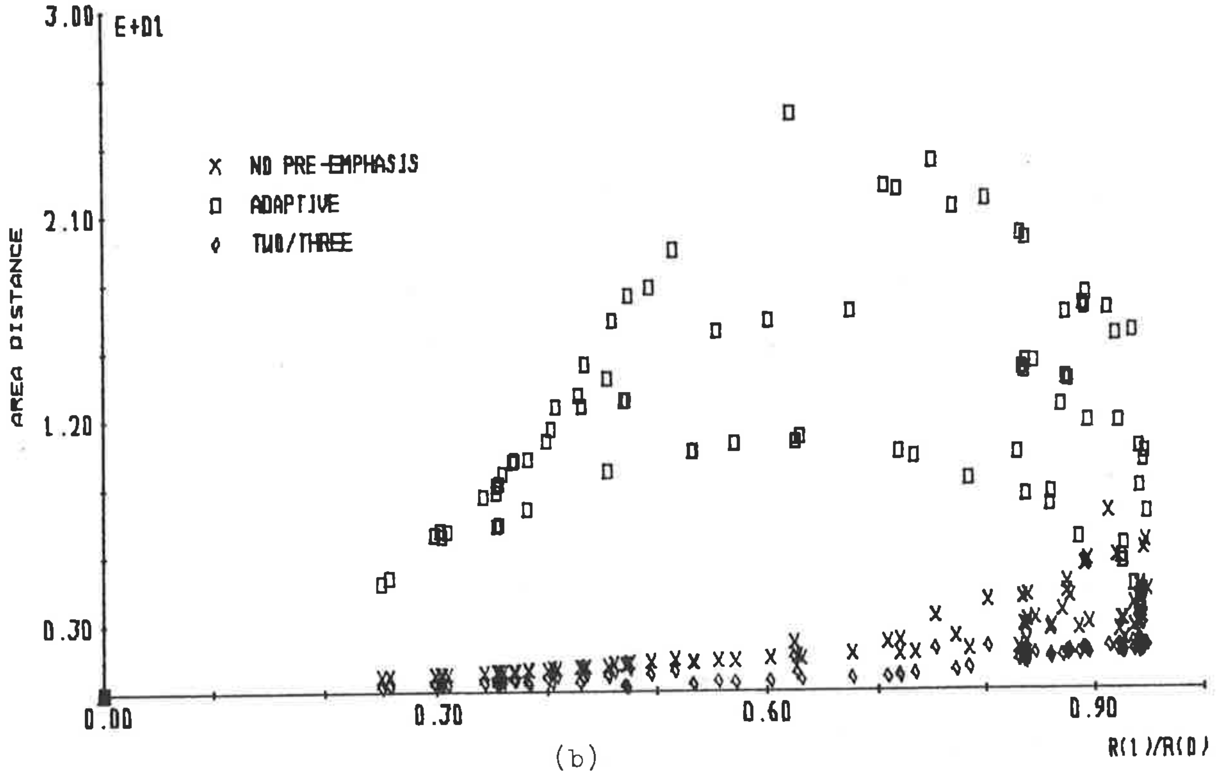
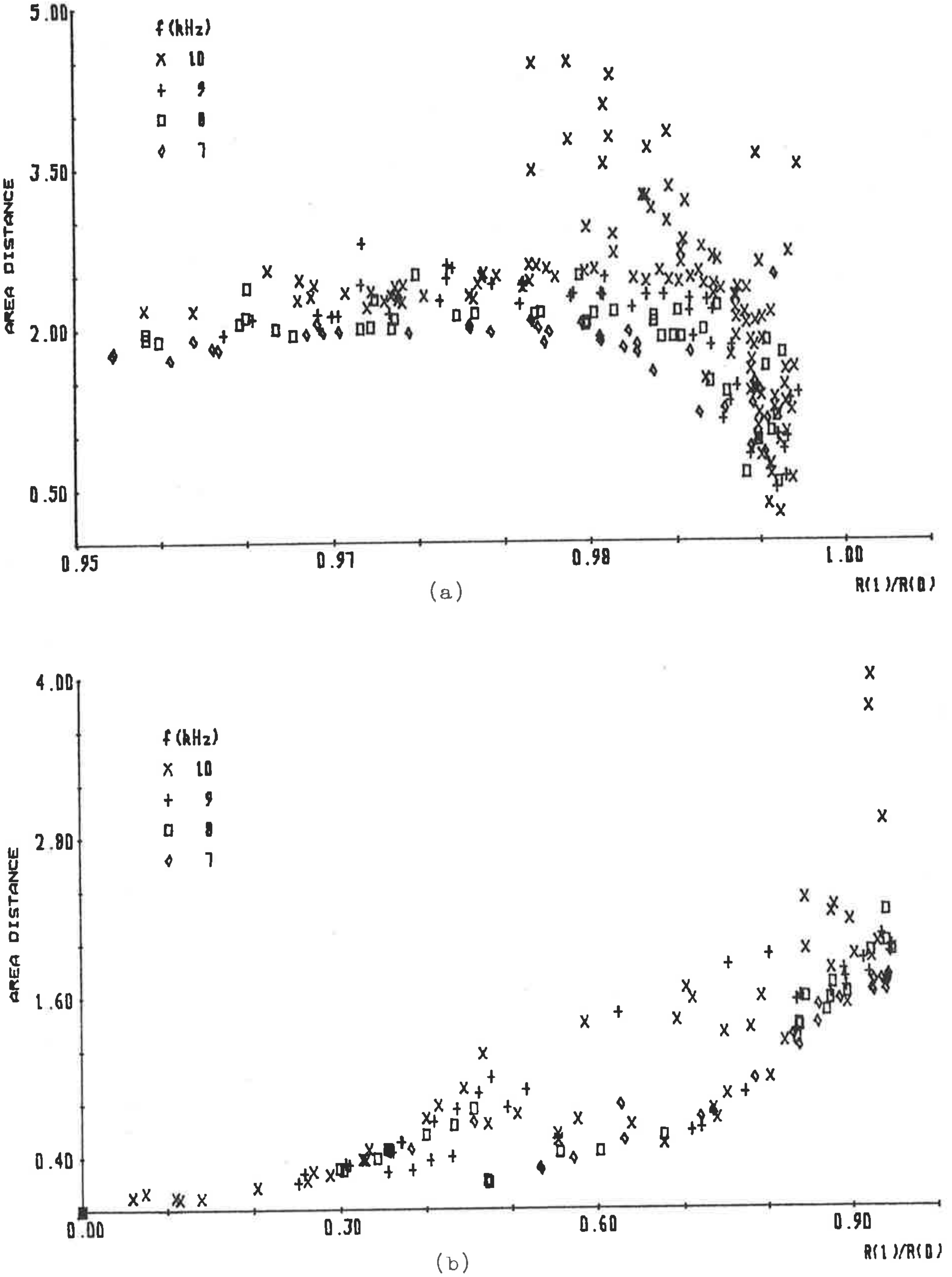


FIGURE 5.30: Area distances for analysis of synthetic speech for the vowel /a/ sampled at frequencies of 9, 8 and 7 kilohertz with R(1)/R(0) (a) greater than 0.95 and (b) less than 0.95.



(a)

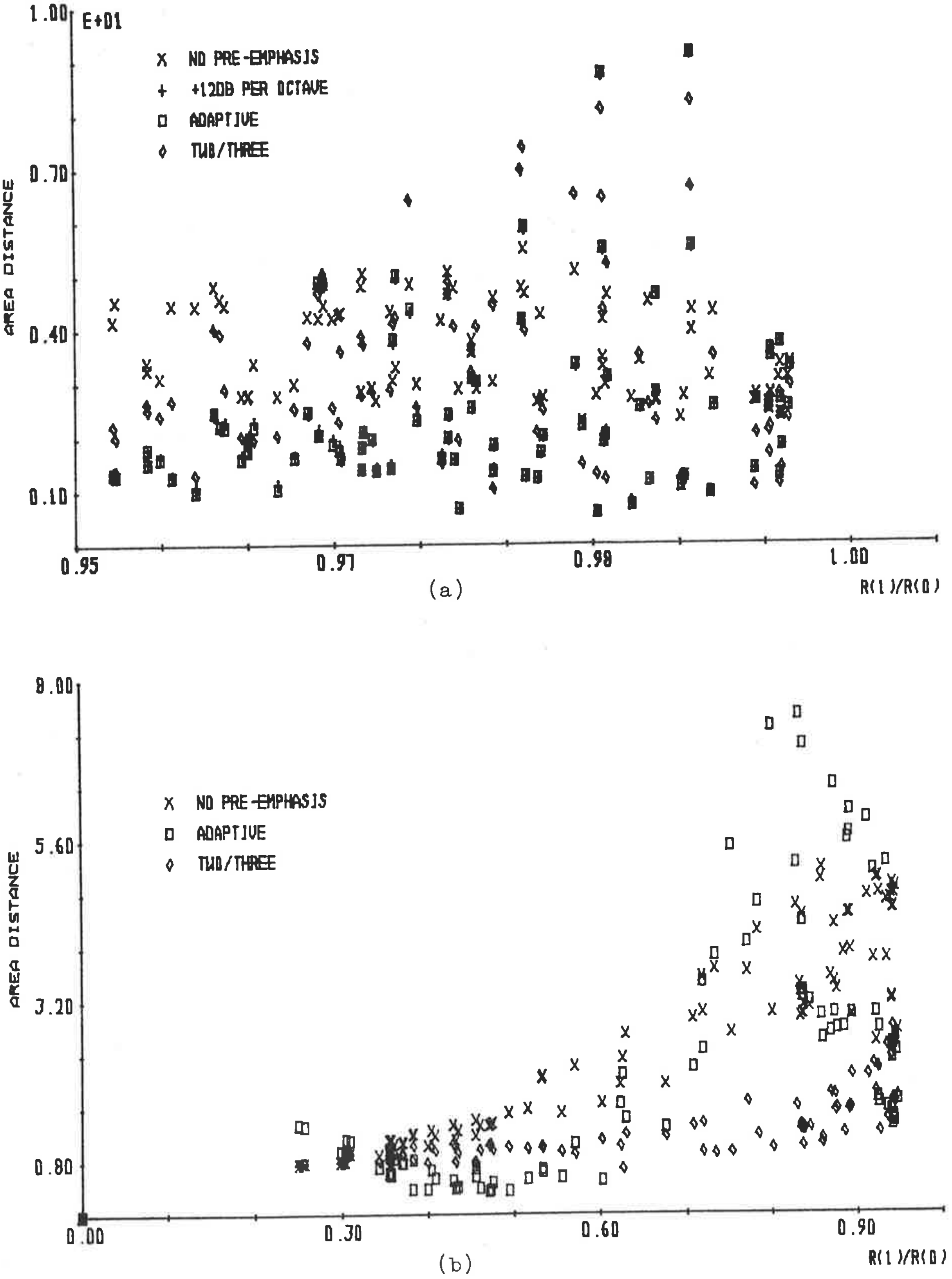
(b)

FIGURE 5.31: Area distances for the two/three adaptive pre-emphasis of synthetic speech for the vowel /a/ sampled at frequencies of 10, 9, 8 and 7 kilohertz with R(1)/R(0) (a) greater than 0.95 and (b) less than 0.95.

Hence, for synthetic speech of the vowel |a|, the performance of the two/three adaptive pre-emphasis to provide a significant reduction in area distances compared with those for no pre-emphasis, a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis is not significantly affected by sampling frequencies between 10 and 7 kilohertz, inclusive. The amount by which the two/three adaptive pre-emphasis filter provides a reduction in area distances compared with existing pre-emphases may be affected by changes in sampling frequency between 10 and 7 kilohertz, inclusive.

The area distances versus $R(1)/R(0)$ presented in Figure 5.32 are for no pre-emphasis, +12 dB per octave pre-emphasis, an unvoiced/voiced adaptive pre-emphasis, and a two/three adaptive pre-emphasis of synthetic speech for the vowel |e|, when sampled at frequencies of 9, 8 and 7 kilohertz. The data points of Figure 5.32(a) show that the two/three adaptive pre-emphasis filter does not always produce a reduction in area distances when compared with those of no pre-emphasis, a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis for $R(1)/R(0)$ between 0.95 and approximately unity. In general, when $R(1)/R(0)$ is between 0.95 and approximately unity, similar area distances occur for a two/three adaptive pre-emphasis, a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis.

For $R(1)/R(0)$ less than 0.95, the two/three adaptive pre-emphasis produces area distances which are significantly less than those for no pre-emphasis until $R(1)/R(0)$ is less than 0.3, when similar area distances occur. A reduction in area distances by the two/three adaptive pre-emphasis filter occurs when compared



(a)

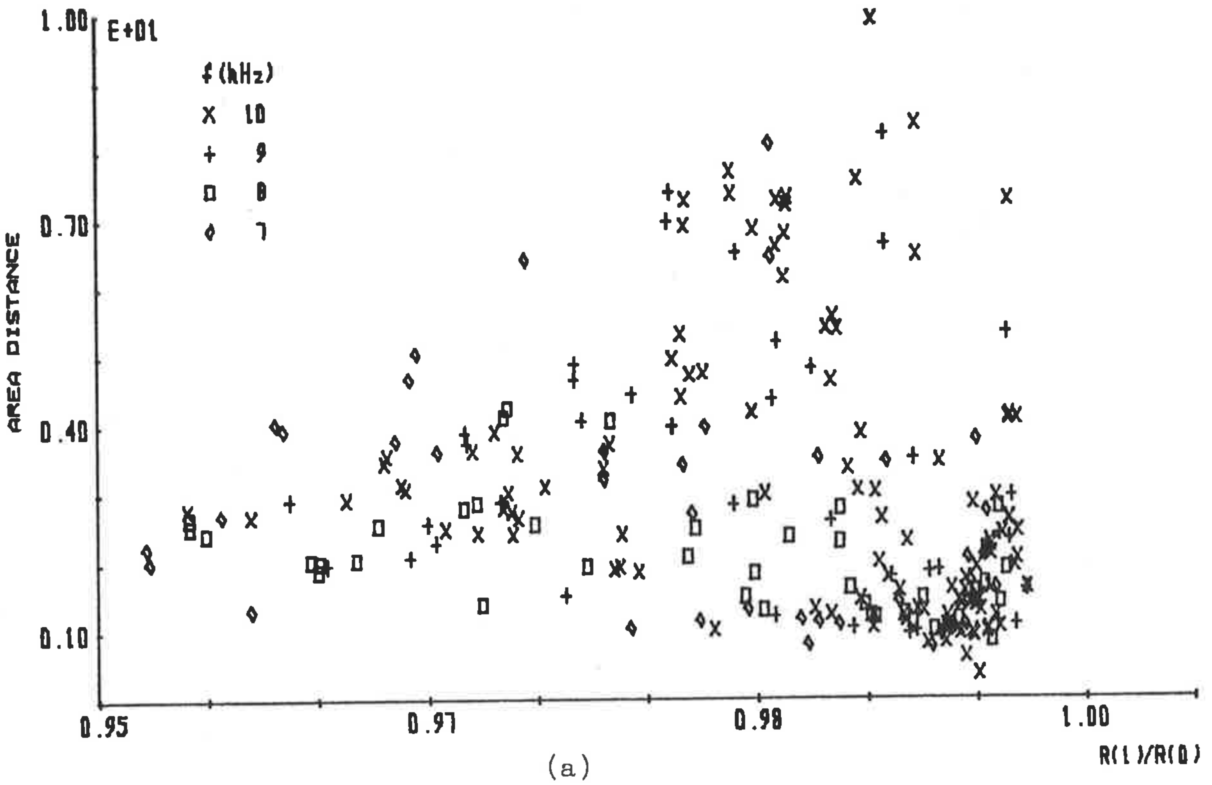
(b)

FIGURE 5.32: Area distances for analysis of synthetic speech for the vowel /e/ sampled at frequencies of 9,8 and 7 kilohertz with R(1)/R(0) (a) greater than 0.95 and (b) less than 0.95.

with the area distances for a +12 dB per octave pre-emphasis when $R(1)/R(0)$ is between zero and 0.95, and when compared with the area distances for an unvoiced/voiced adaptive pre-emphasis when $R(1)/R(0)$ is between 0.9 and 0.6, and less than 0.3. The amount by which a reduction in area distance occurs for the two/three adaptive pre-emphasis compared with the area distances of existing pre-emphasis techniques changes for different sampling frequencies, but the above observations and general trends are consistent with those observed for a 10 kilohertz waveform sampling frequency (see Figure 5.19).

The area distances versus $R(1)/R(0)$ for the two/three adaptive pre-emphasis of synthetic speech for the vowel |e| at sampling frequencies of 10, 9, 8 and 7 kilohertz are presented in Figure 5.33. Different symbols indicate the different sampling frequencies in Figure 5.33, and the even spread of data points shows that similar ranges of area distances occur when the two/three adaptive pre-emphasis is used with waveform sampling frequencies between 10 and 7 kilohertz, inclusive.

Hence, for synthetic speech of the vowel |e|, using a sampling frequency between 10 and 7 kilohertz, inclusive, does not significantly affect the performance of the two/three adaptive pre-emphasis filter to provide a reduction in area distances when compared with those for existing pre-emphasis techniques. In the case of the vowel |e|, similar reductions in area distances occur for any waveform sampling frequency between 10 and 7 kilohertz, inclusive.



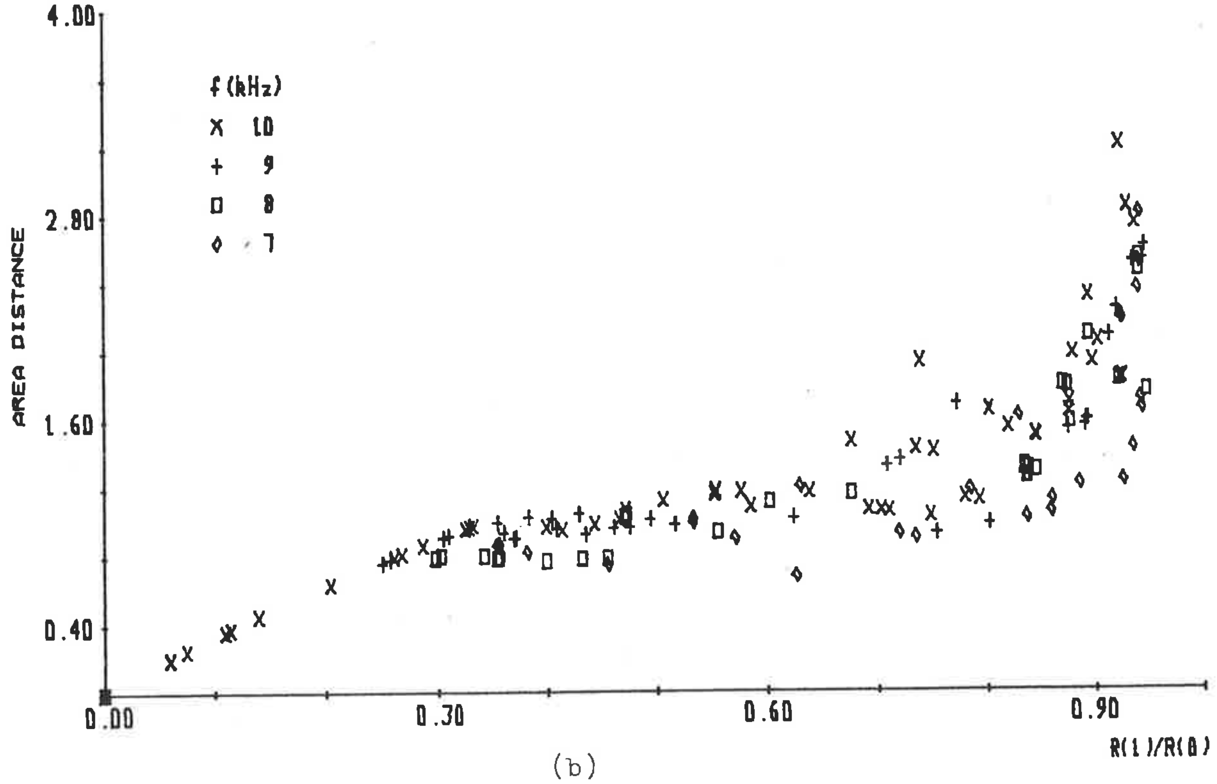
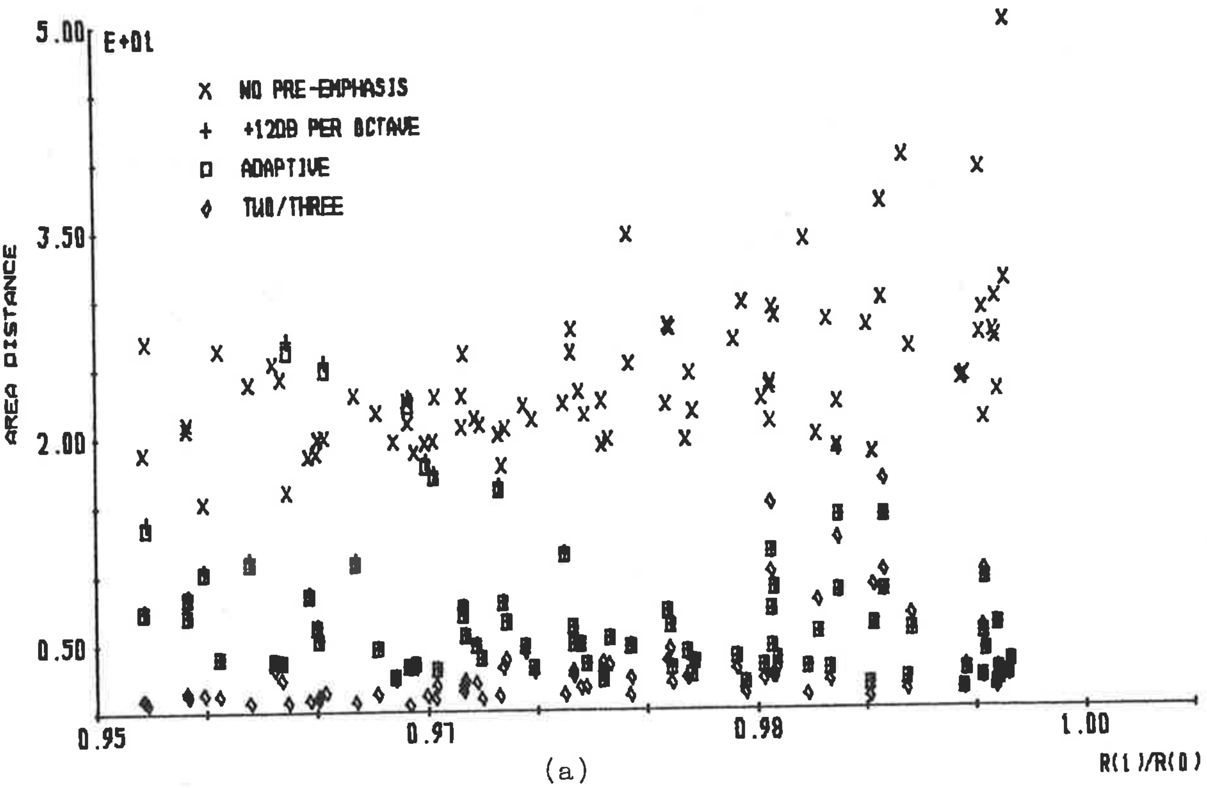


FIGURE 5.33: Area distances for the two/three adaptive pre-emphasis of synthetic speech for the vowel /e/ sampled at frequencies of 10, 9, 8 and 7 kilohertz with R(1)/R(0) (a) greater than 0.95 and (b) less than 0.95

For synthetic speech of the vowel |i| pre-emphasized by no pre-emphasis, +12 dB per octave pre-emphasis, an unvoiced/voiced adaptive pre-emphasis and a two/three adaptive pre-emphasis, the area distances versus $R(1)/R(0)$ are presented in Figure 5.34 for sampling frequencies of 9, 8 and 7 kilohertz. Figure 5.34 shows that a large reduction in area distances occurs for the two/three adaptive pre-emphasis when compared with the area distances for no pre-emphasis when $R(1)/R(0)$ is greater than 0.5, but similar area distances when $R(1)/R(0)$ is less than 0.5. Similar, but smaller, area distances occur for the two/three adaptive pre-emphasis filter in comparison with a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis when $R(1)/R(0)$ is greater than 0.95. For $R(1)/R(0)$ between 0.75 and 0.95 and less than 0.5, the two/three adaptive pre-emphasis produces smaller area distances than an unvoiced/voiced adaptive pre-emphasis.

The above observations differ from those for a waveform sampling frequency of 10 kilohertz (see Figure 5.20), where a much larger reduction in area distance occurs for the two/three adaptive pre-emphasis in comparison with a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis when $R(1)/R(0)$ is greater than 0.95. A comparison of Figures 5.20 and 5.34 shows that the area distances for a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis have decreased from the 10 kilohertz sampling frequency case to the 9, 8 and 7 kilohertz sampling frequencycase, rather than a significant change in area distances being produced by a two/three adaptive pre-emphasis.



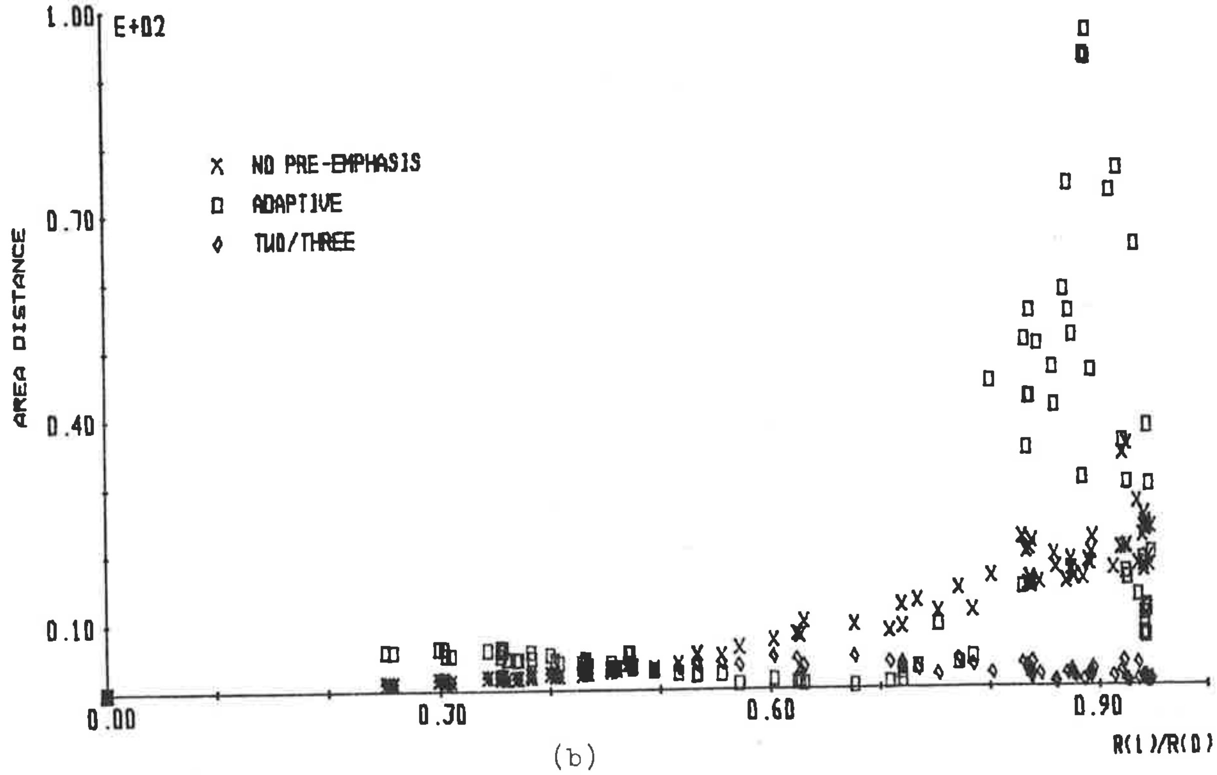
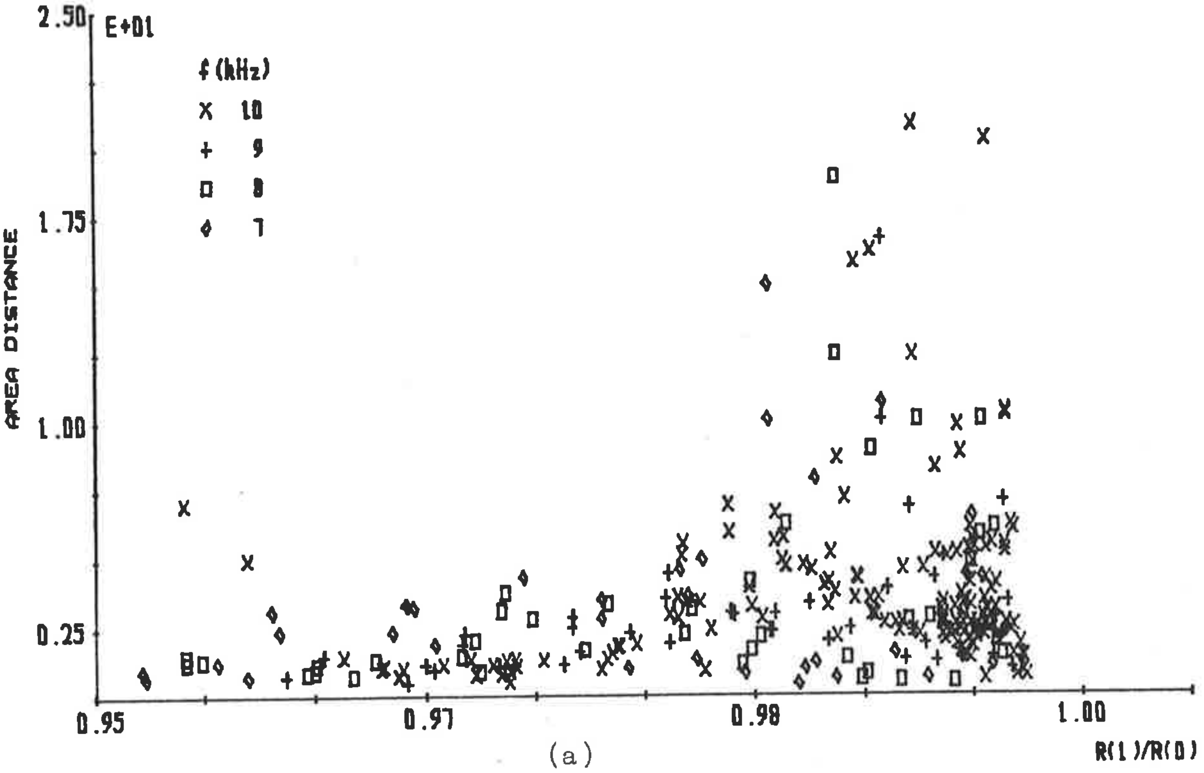


FIGURE 5.34: Area distances for analysis of synthetic speech for the vowel /i/ sampled at frequencies of 9, 8 and 7 kilohertz with R(1)/R(0) (a) greater than 0.95 and (b) less than 0.95.

For the four waveform sampling frequencies considered, i.e. 10, 9, 8 and 7 kilohertz, the area distances for a two/three adaptive pre-emphasis of synthetic speech for the vowel |i| are presented in Figure 5.35. For each different sampling frequency, a different symbol is used, and an even spacing of symbols for $R(1)/R(0)$ greater than 0.95 indicates that the range of area distances produced by the two/three adaptive pre-emphasis filter is unaffected by sampling frequencies between 10 and 7 kilohertz, inclusive, when $R(1)/R(0)$ is greater than 0.95. When $R(1)/R(0)$ is between zero and 0.95, a reduction in area distances occurs for sampling frequencies of 9, 8 and 7 kilohertz when compared with the area distances for a 10 kilohertz sampling frequency.

Therefore, the only effect of changing the sampling frequencies between 10 and 7 kilohertz, inclusive, for synthetic speech of the vowel |i| is to produce a small reduction in area distances as the sampling frequency decreases. In general, the reduction in area distance compared with the area distances for no pre-emphasis and when a reduction in area distances occurs in comparison with those for a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis is not significantly affected by a change in sampling frequency between 10 and 7 kilohertz, inclusive.

The area distances for no pre-emphasis, a +12 dB per octave pre-emphasis, an unvoiced/voiced adaptive pre-emphasis and a two/three adaptive pre-emphasis of synthetic speech for the vowel |o| is presented in Figure 5.36 versus $R(1)/R(0)$, for waveform sampling frequencies of 9, 8 and 7 kilohertz. A reduction in area distances is observed in Figure 5.36 for the two/three adaptive pre-emphasis filter when compared with the area distances for a +12 dB per oc-



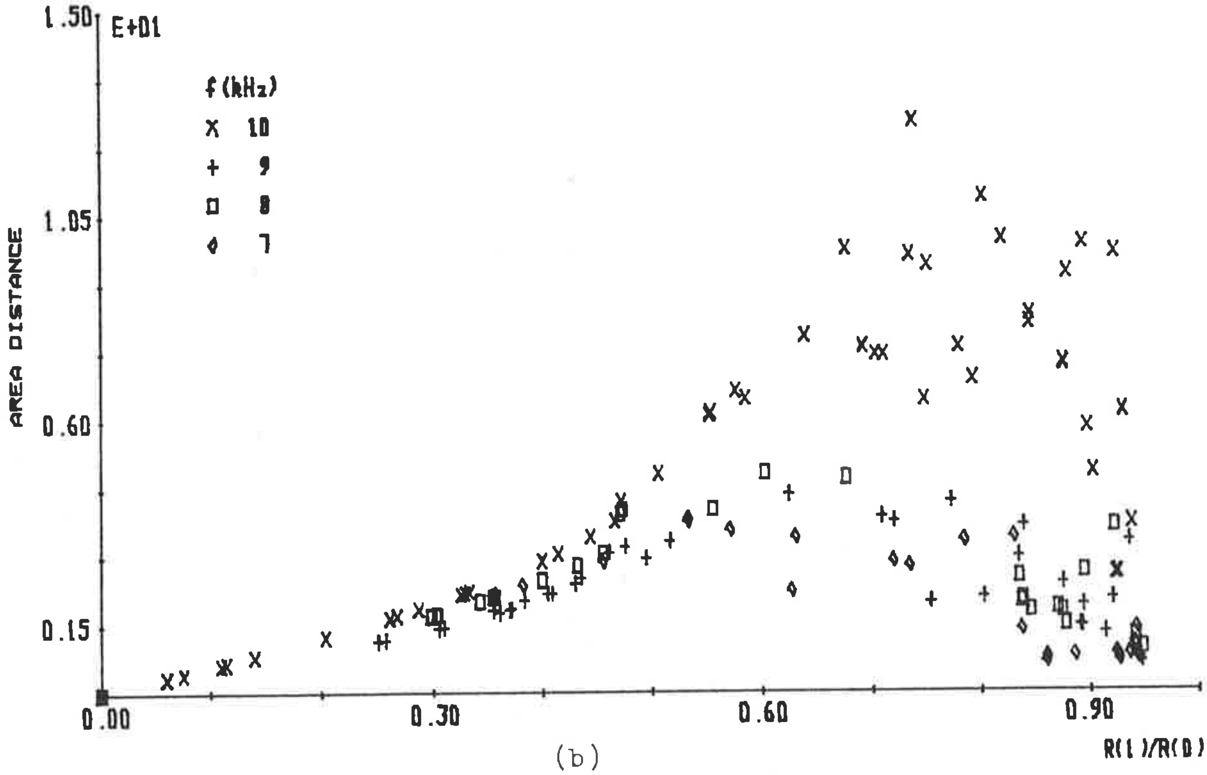
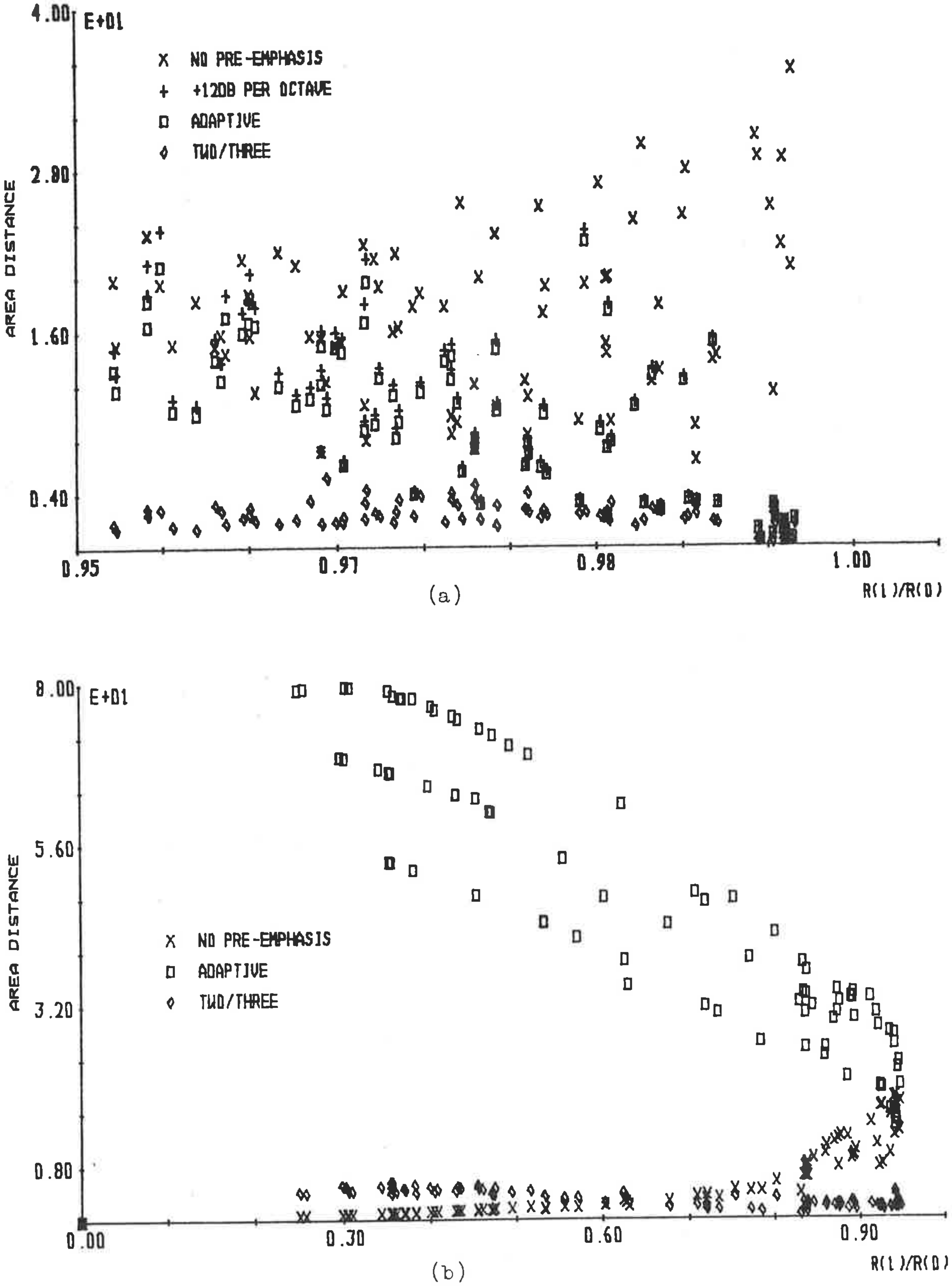


FIGURE 5.35: Area distances for the two/three adaptive pre-emphasis of synthetic speech for the vowel /i/ sampled at frequencies of 10, 9, 8 and 7 kilohertz with R(1)/R(0) (a) greater than 0.95 and (b) less than 0.95



(a)

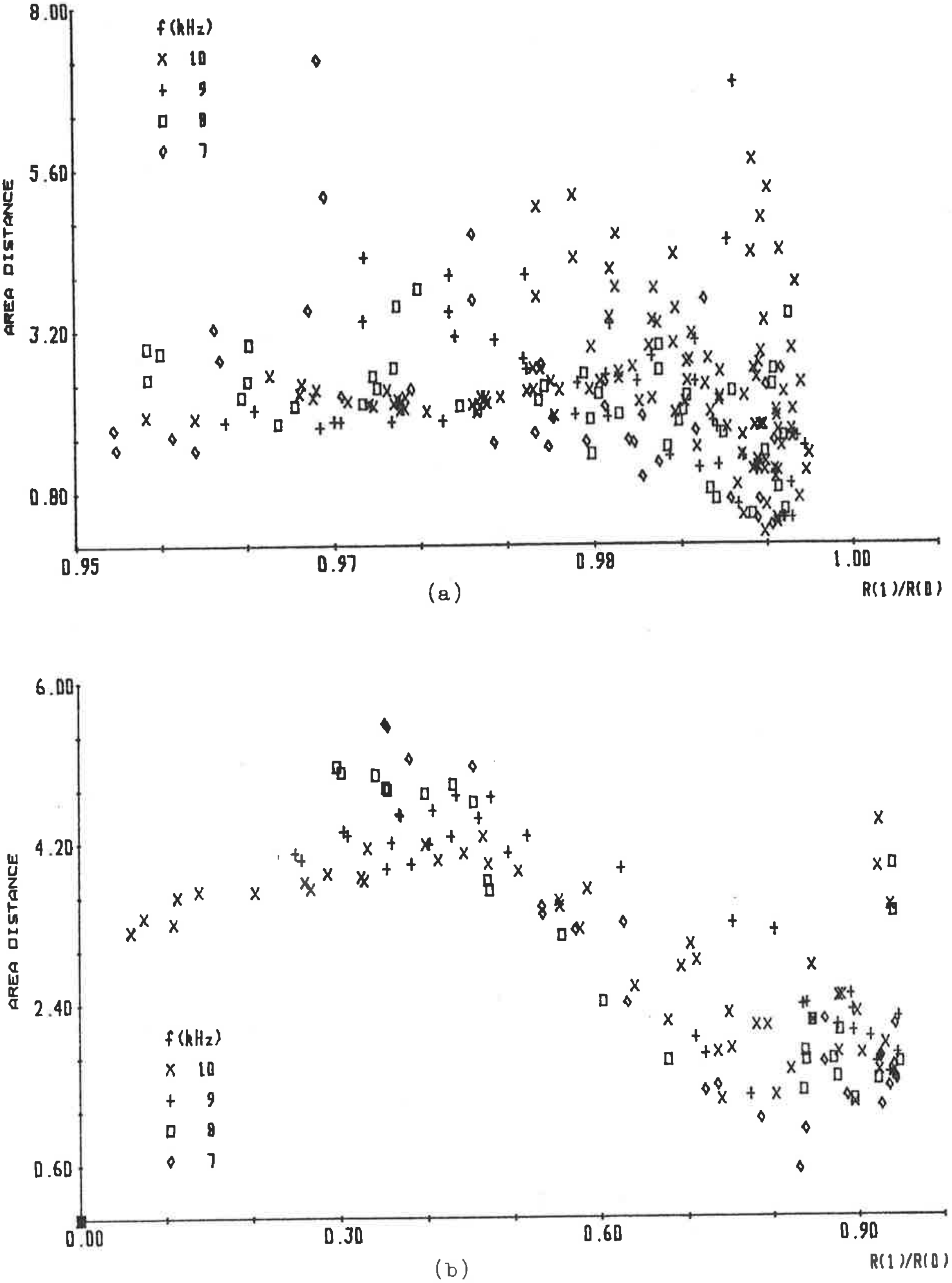
(b)

FIGURE 5.36: Area distances for analysis of synthetic speech for the vowel /o/ sampled at frequencies of 9, 8 and 7 kilohertz with R(1)/R(0) (a) greater than 0.95 and (b) less than 0.95

tave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis, with the reduction in area distances increasing markedly as $R(1)/R(0)$ decreases. A large reduction in area distances occurs for the two/three adaptive pre-emphasis when compared with the area distances for no pre-emphasis, when $R(1)/R(0)$ is near unity, but the area distances for no pre-emphasis are smaller when $R(1)/R(0)$ is less than 0.6.

The above observations follow the general trends for a waveform sampling frequency of 10 kilohertz for synthetic speech of the vowel |o|, as presented in Figure 5.21. However, the reduction in area distances by the two/three adaptive pre-emphasis filter when compared with the area distances for a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis is smaller for the sampling frequencies of 9, 8 and 7 kilohertz than for 10 kilohertz. As observed for synthetic speech of the vowel |i|, this smaller reduction in area distances is due to a reduction in area distances by the +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis as the sampling frequency decreases, rather than a change in the area distances for the two/three adaptive pre-emphasis filter.

Figure 5.37 presents the area distances for a two/three adaptive pre-emphasis of synthetic speech of the vowel |o| versus $R(1)/R(0)$ for the sampling frequencies 10, 9, 8 and 7 kilohertz. To indicate which area distances correspond to which sampling frequencies, different symbols are used in Figure 5.37, where the even spread of different symbols indicates that the range of area distances returned by a two/three adaptive pre-emphasis is not signi-



(a)

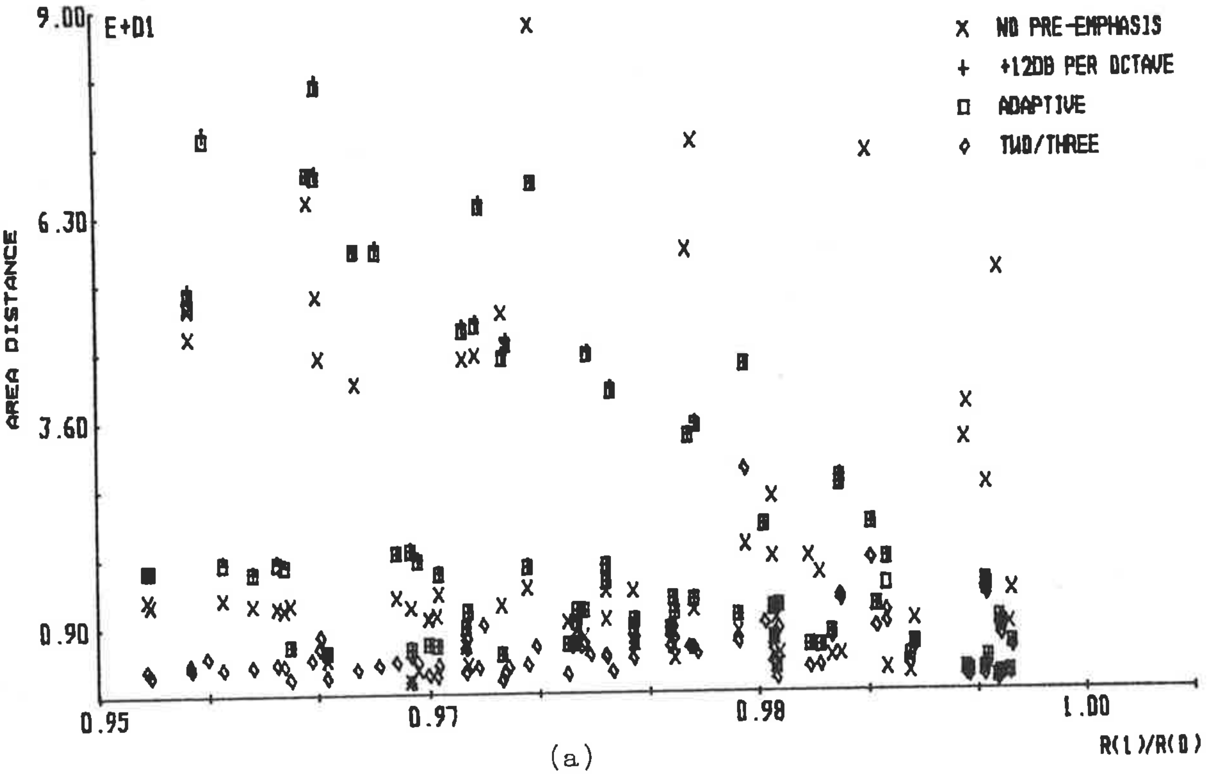
(b)

FIGURE 5.37: Area distances for the two/three adaptive pre-emphasis of synthetic speech for the vowel /o/ sampled at frequencies of 10, 9, 8 and 7 kilohertz with R(1)/R(0) (a) greater than 0.95 and (b) less than 0.95

ficantly affected by using sampling frequencies between 10 and 7 kilohertz, inclusive.

Hence, a two/three adaptive pre-emphasis of synthetic speech for the vowel |o| produces a range of area distances which is not significantly affected by a change in waveform sampling frequency between 10 and 7 kilohertz, inclusive. The reductions in area distances for the two/three adaptive pre-emphasis in comparison with the area distances for no pre-emphasis are approximately the same for different sampling frequencies. However, as the sampling frequency decreases, the reduction in area distances by the two/three adaptive pre-emphasis in comparison with the area distances for an unvoiced/voiced adaptive pre-emphasis decreases due to the area distances of the unvoiced/voiced adaptive pre-emphasis decreasing.

Figure 5.38 presents the area distances versus $R(1)/R(0)$ for no pre-emphasis, a +12 dB per octave pre-emphasis, an unvoiced/voiced adaptive pre-emphasis and a two/three adaptive pre-emphasis of synthetic speech for the vowel |u| for waveform sampling frequencies of 9, 8 and 7 kilohertz. In general, the two/three adaptive pre-emphasis filter is shown in Figure 5.38 to produce area distances which are smaller than no pre-emphasis, a +12 dB per octave pre-emphasis and an unvoiced/voiced pre-emphasis for $R(1)/R(0)$ greater than 0.95. The variation in area distances for the two/three adaptive pre-emphasis is much smaller than for the other pre-emphasis techniques.



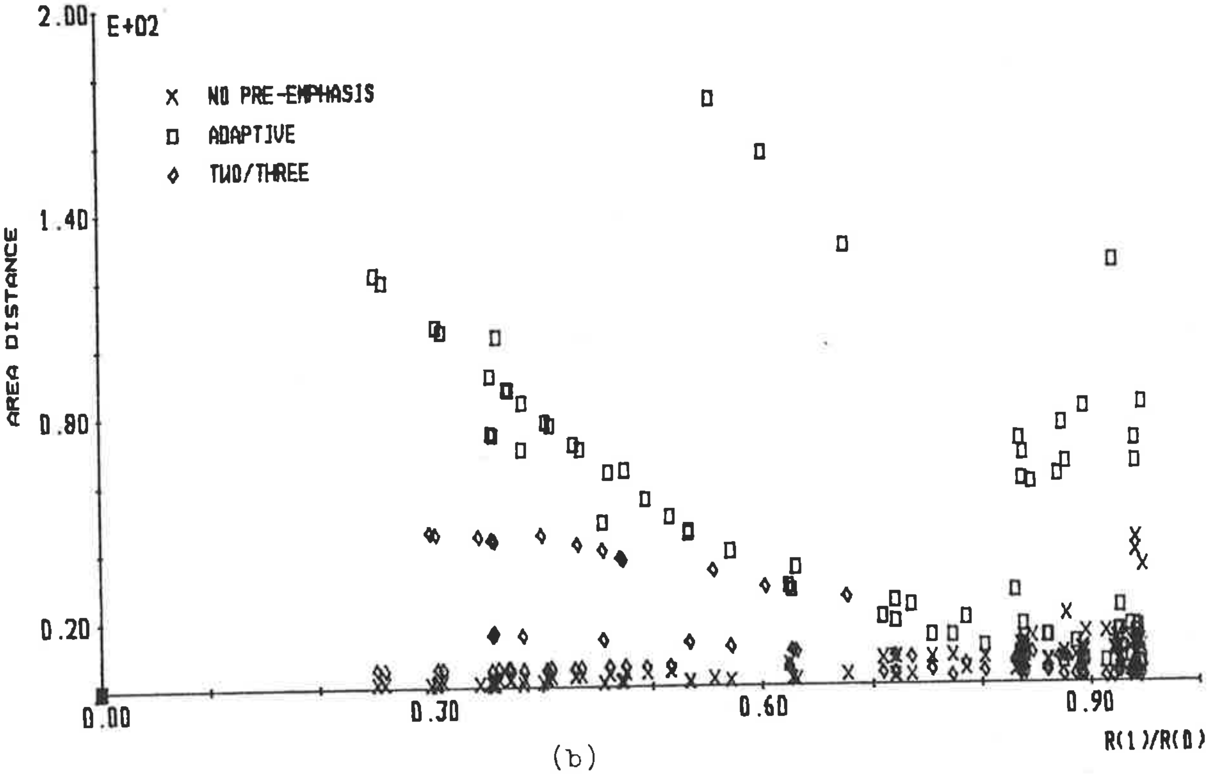
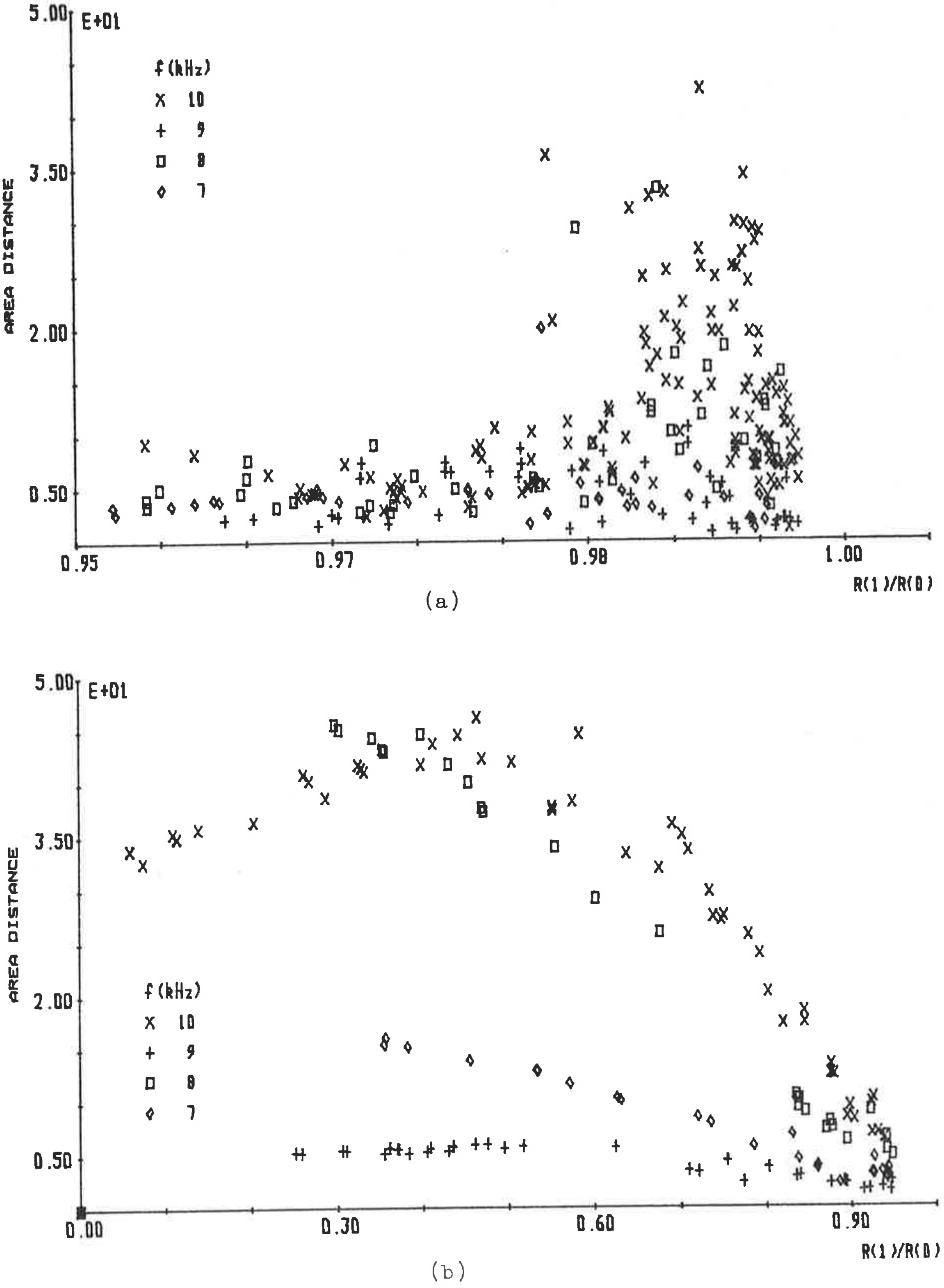


FIGURE 5.38: Area distances for analysis of synthetic speech for the vowel /u/ sampled at frequencies of 9, 8 and 7 kilohertz with R(1)/R(0) (a) greater than 0.95 and (b) less than 0.95

When $R(1)/R(0)$ is less than 0.95, the two/three adaptive pre-emphasis filter produces a reduction in area distances in comparison with the area distances for an unvoiced/voiced adaptive pre-emphasis, but in comparison with no pre-emphasis a reduction of area distance only occurs when $R(1)/R(0)$ is greater than 0.8. In contrast to the other vowels considered in this section, a significant change in area distances, produced by a two/three adaptive pre-emphasis, occurs when the sampling frequency changes, and $R(1)/R(0)$ is less than 0.9. This change in area distances does not affect the performance of the two/three adaptive pre-emphasis filter to provide a reduction in area distances, since for some sampling frequencies larger area distances occur for the two/three adaptive pre-emphasis filter than for no pre-emphasis when $R(1)/R(0)$ is less than 0.9.

Figure 5.39 presents the area distances for a two/three adaptive pre-emphasis of synthetic speech for the vowel $|u|$ sampled at frequencies of 10, 9, 8 and 7 kilohertz versus $R(1)/R(0)$. Different symbols are used to differentiate between the different sampling frequencies, and the even spread of symbols in Figure 5.39 for $R(1)/R(0)$ greater than 0.9 shows that the performance of the two/three adaptive pre-emphasis is not significantly affected by changing sampling frequencies between 10 and 7 kilohertz, inclusive, when $R(1)/R(0)$ is greater than 0.9. When $R(1)/R(0)$ is less than 0.9, then different ranges of area distances occur for different waveform sampling frequencies, with much larger area distances for sampling frequencies of 10 and 8 kilohertz than for sampling frequencies of 9 and 7 kilohertz. This change in area distances with sampling frequency is not consistent with the other vowels considered in this section.



(a)

(b)

FIGURE 5.39: Area distances for the two/three adaptive pre-emphasis of synthetic speech for the vowel /u/ sampled at frequencies of 10, 9, 8 and 7 kilohertz with R(1)/R(0) (a) greater than 0.95 and (b) less than 0.95.

Therefore, for synthetic speech of the vowel |u|, a change in sampling frequency between 10 and 7 kilohertz, inclusive, only affects the performance of the two/three adaptive pre-emphasis filter to provide a reduction in area distances when $R(1)/R(0)$ is less than 0.9.

The evaluations presented in this section have shown that, in general, when the two/three adaptive pre-emphasis filter is defined by the parameter α, and the relationships between α' and $R(1)/R(0)$, as presented in Chapter 4, the area distances for a two/three adaptive pre-emphasis of synthetic speech waveforms are not significantly affected by a change in sampling frequency between 10 and 7 kilohertz, inclusive. The only exception to this statement is for synthetic speech of the vowel |u| when $R(1)/R(0)$ is less than 0.9.

5.3.3.2 Parameter β

The relationship between the parameter β of the two/three adaptive pre-emphasis filter and $R(1)/R(0)$ has been shown, in Chapter 4, to be dependent on the waveform sampling frequency. Section 5.3.2 presented an evaluation of the two/three adaptive pre-emphasis filter, when defined by the parameter β, with synthetic speech of five vowel sounds sampled at a frequency of 10 kilohertz. This section presents an evaluation of the two/three adaptive pre-emphasis filter, when defined by the parameter β, with synthetic speech of five vowel sounds sampled at frequencies of 9, 8 and 7 kilohertz.

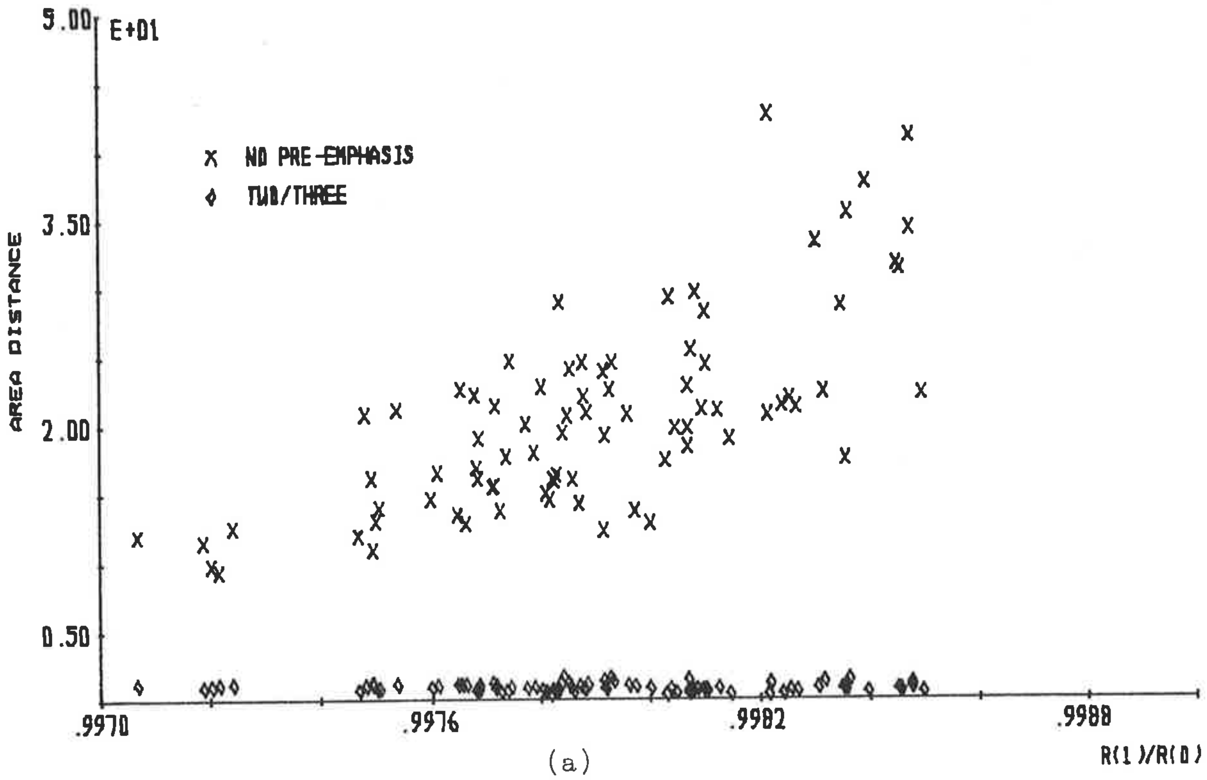
This section determines whether the general trends observed for the two/three adaptive pre-emphasis filter in comparison with the existing pre-emphasis techniques for a waveform sampling frequency of 10 kilohertz, i.e. as presented in Section 5.3.2, are consistent with the general trends for waveform sampling frequencies of 9, 8 and 7 kilohertz. The variations that occur in area distances for the synthetic speech waveforms when pre-emphasized by the two/three adaptive pre-emphasis and then analysed by a Parcor analysis are presented for changes in sampling frequency. For all the waveform sampling frequencies considered in this section, a different relationship between β and $R(1)/R(0)$ is required, and these relationships are defined in Chapter 4.

All the evaluations of the two/three adaptive pre-emphasis filter are presented as plots of area distances between recovered and original acoustic tube shapes versus the value of $R(1)/R(0)$, which is evaluated from the glottal pulse waveform used to generate the synthetic speech. The reason for using $R(1)/R(0)$ of the glottal pulse waveform is to relate the results obtained to the glottal pulse waveforms which the two/three adaptive pre-emphasis filter is designed to remove. However, the two/three adaptive filter is defined via $R(1)/R(0)$ evaluated from the autocorrelation function of the synthetic speech waveform.

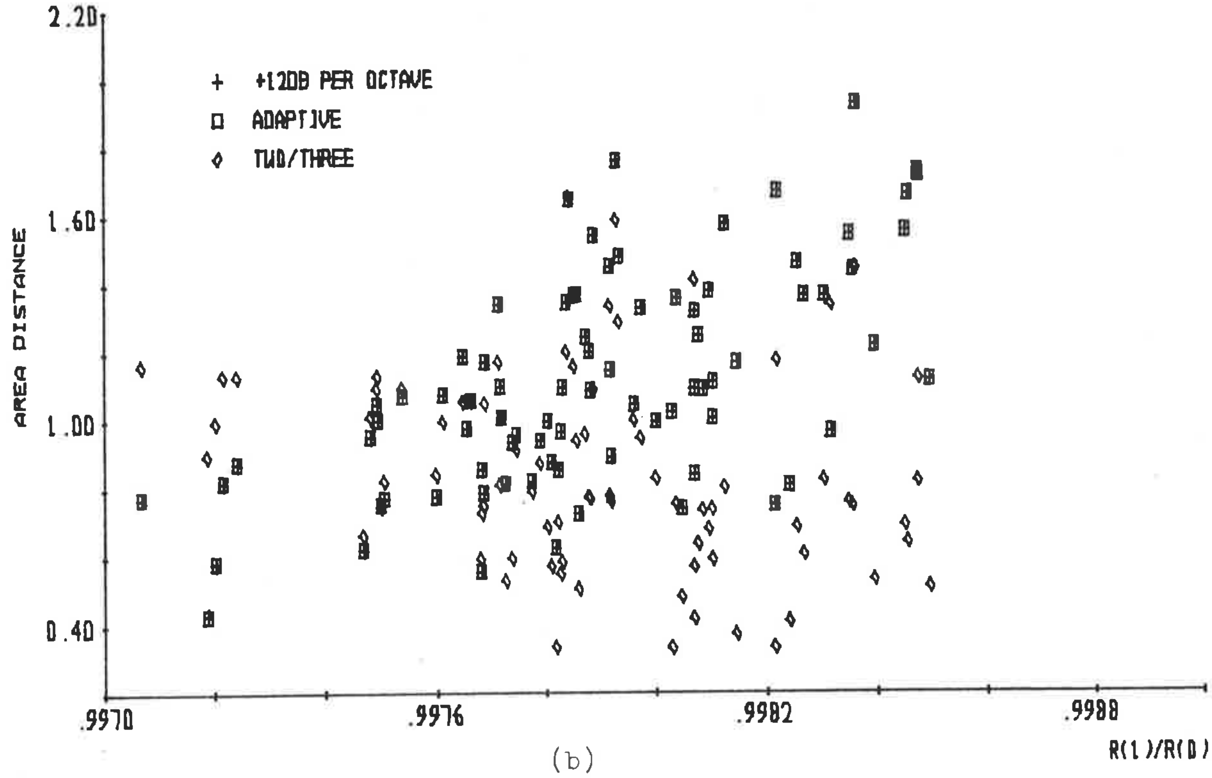
The procedure used to generate the synthetic speech used for the evaluations presented in this section is described in Appendix B. To ensure the synthetic speech is as close as possible to real speech waveforms, the acoustic tube shape used to generate the synthetic speech approximates the vocal tract shapes for five vowels, |a|, |e|, |i|, |o| and |u|, as defined in Appendix C. To avoid radiation effects clouding the evaluations presented in this section, the termination reflection coefficient of the acoustic tube model used to generate the synthetic speech is assumed as unity, i.e. $\mu_M = 1$. All the glottal pulse waveforms used to generate synthetic speech are derived from the measured glottal pulse waveforms of MONSEN and ENGEBRETSON [1977], ROTHENBERG [1973], SONDHI [1975], SUNDBERG and GAUFFIN [1978], MILLER [1959], and FLANAGAN and LANDGRAF [1968].

Figure 5.40 presents area distances versus $R(1)/R(0)$ for no pre-emphasis, a +12 dB per octave, an unvoiced/voiced adaptive pre-emphasis and a two/three adaptive pre-emphasis of synthetic speech for the vowel |a|, for sampling frequencies of 9, 8 and 7 kilohertz. The two/three adaptive pre-emphasis is shown in Figure 5.40(a) to provide a large reduction in area distances in comparison with no pre-emphasis. Except when $R(1)/R(0)$ is near 0.9970, i.e. the transition point, the two/three adaptive pre-emphasis filter provides a reduction in area distances when compared with a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis. The amount by which a reduction in area distance occurs for the two/three adaptive pre-emphasis filter increases as $R(1)/R(0)$ increases.

The area distances for the two/three adaptive pre-emphasis of synthetic speech for the vowel |a| versus $R(1)/R(0)$ is presented in Figure 5.41 for four waveform sampling frequencies, namely 10, 9, 8 and 7 kilohertz. Different symbols are used to differentiate between different sampling frequencies and, except for the sampling frequency of 10 kilohertz, for which larger area distances occur, an even spread of symbols is observed in Figure 5.41. However, in



(a)



(b)

FIGURE 5.40: Area distances for analysis of synthetic speech for the vowel /a/, sampled at frequencies of 9, 8 and 7 kilohertz, by (a) no pre-emphasis and a two/three adaptive pre-emphasis and (b) a +12dB per octave, an unvoiced/voiced adaptive and two/three adaptive pre-emphases

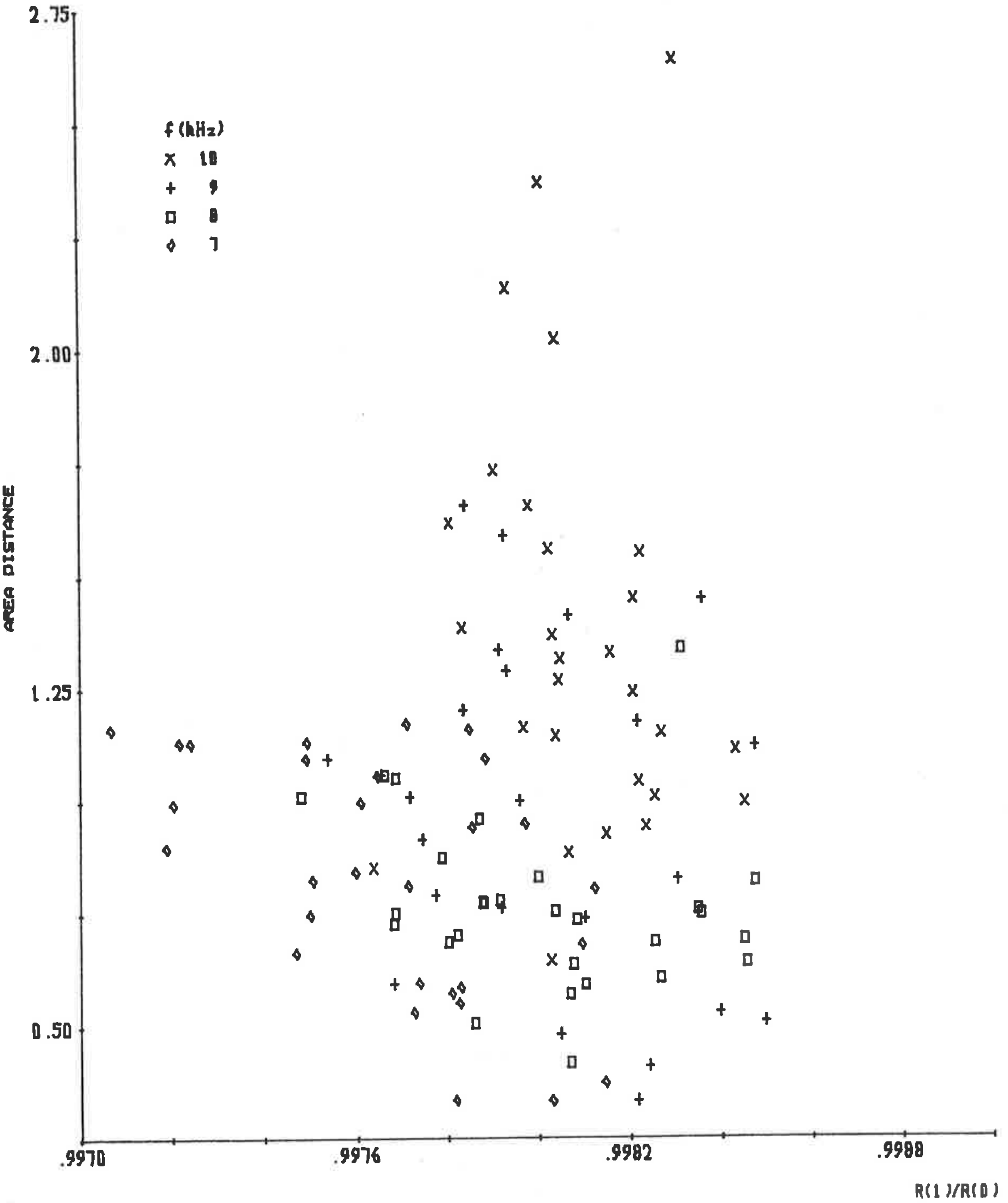
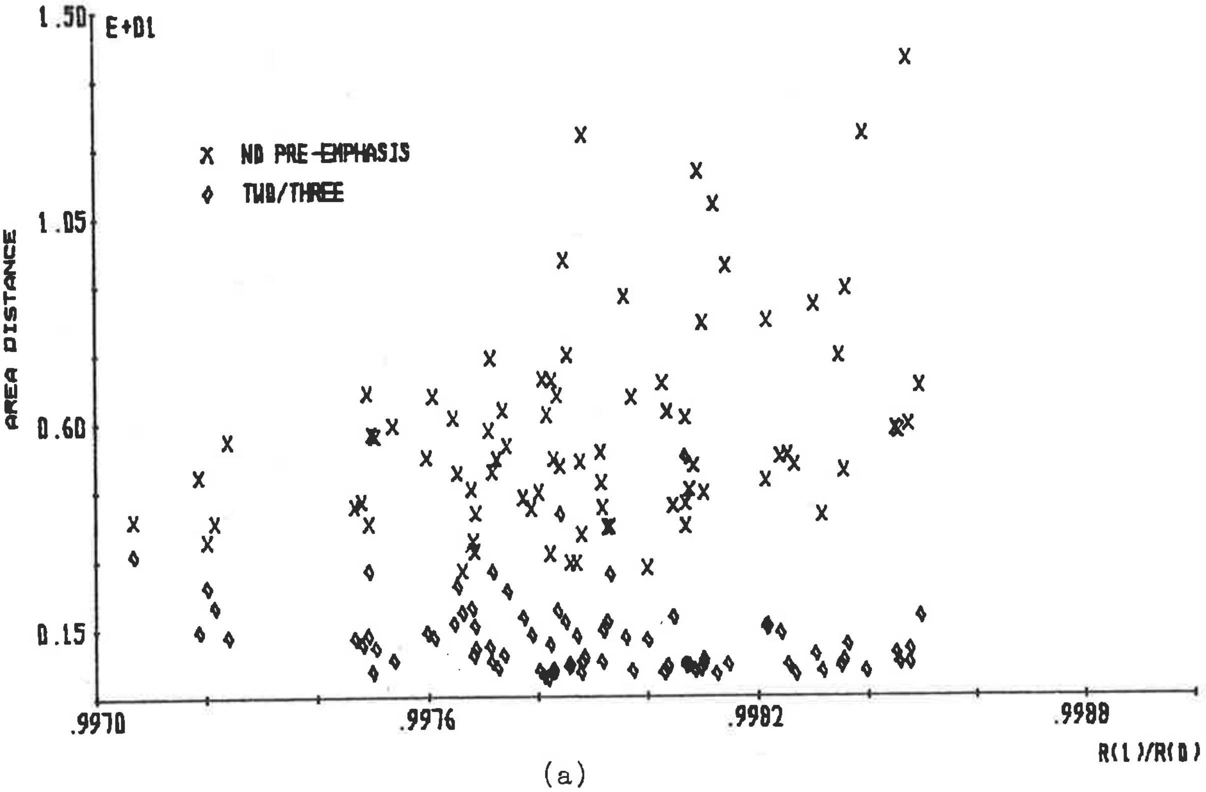


FIGURE 5.41: Area distances for the two/three adaptive pre-emphasis of synthetic speech for the vowel /a/ sampled at frequencies of 10, 9, 8 and 7 kilohertz

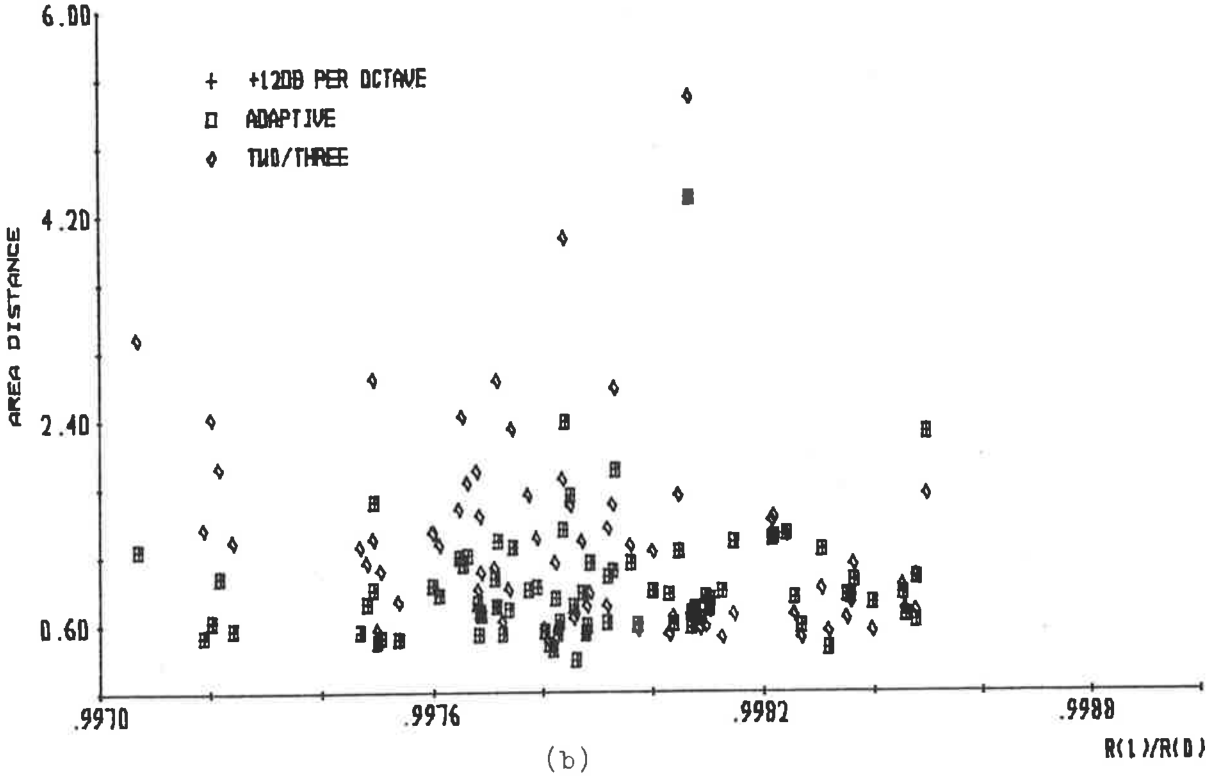
general, the performance of the two/three adaptive pre-emphasis filter to provide a reduction in area distances for synthetic speech of the vowel |a| is only slightly affected by waveform sampling frequencies between 10 and 7 kilohertz, inclusive.

The results of no pre-emphasis, a +12 dB per octave pre-emphasis, an unvoiced/voiced adaptive pre-emphasis and a two/three adaptive pre-emphasis of synthetic speech for the vowel |e| sampled at frequencies of 9, 8 and 7 kilohertz are presented as area distances versus $R(1)/R(0)$ in Figure 5.42. In general, the results of Figure 5.42(a) show that a reduction in area distance occurs when the two/three adaptive pre-emphasis is used instead of no pre-emphasis. However, Figure 5.42(b) shows that the two/three adaptive pre-emphasis produces slightly larger area distances than a +12 dB per octave and an unvoiced/voiced adaptive pre-emphasis when $R(1)/R(0)$ is near 0.9970, but similar area distances when $R(1)/R(0)$ approaches 0.9985. The improvement in the performance of the two/three adaptive pre-emphasis as $R(1)/R(0)$ increases is consistent with the evaluations presented for synthetic speech of the vowel |a|.

Figure 5.43 presents, for waveform sampling frequencies of 10, 9, 8 and 7 kilohertz, the area distances versus $R(1)/R(0)$ for a two/three adaptive pre-emphasis of synthetic speech for the vowel |e|. The use of different symbols for each sampling frequency shows that, except for a sampling frequency of 10 kilohertz, similar area distances occur for different sampling frequencies. Hence, the performance of the two/three adaptive pre-emphasis filter is not significantly affected by sampling frequencies between 9 and 7 kilohertz, inclusive, and only slightly affected by a sampling



(a)



(b)

FIGURE 5.42: Area distances for analysis of synthetic speech for the vowel /e/, sampled at frequencies of 9, 8 and 7 kilohertz, by (a) no pre-emphasis and a two/three adaptive pre-emphasis and (b) a +12dB per octave, an unvoiced/voiced adaptive and two/three adaptive pre-emphasis.

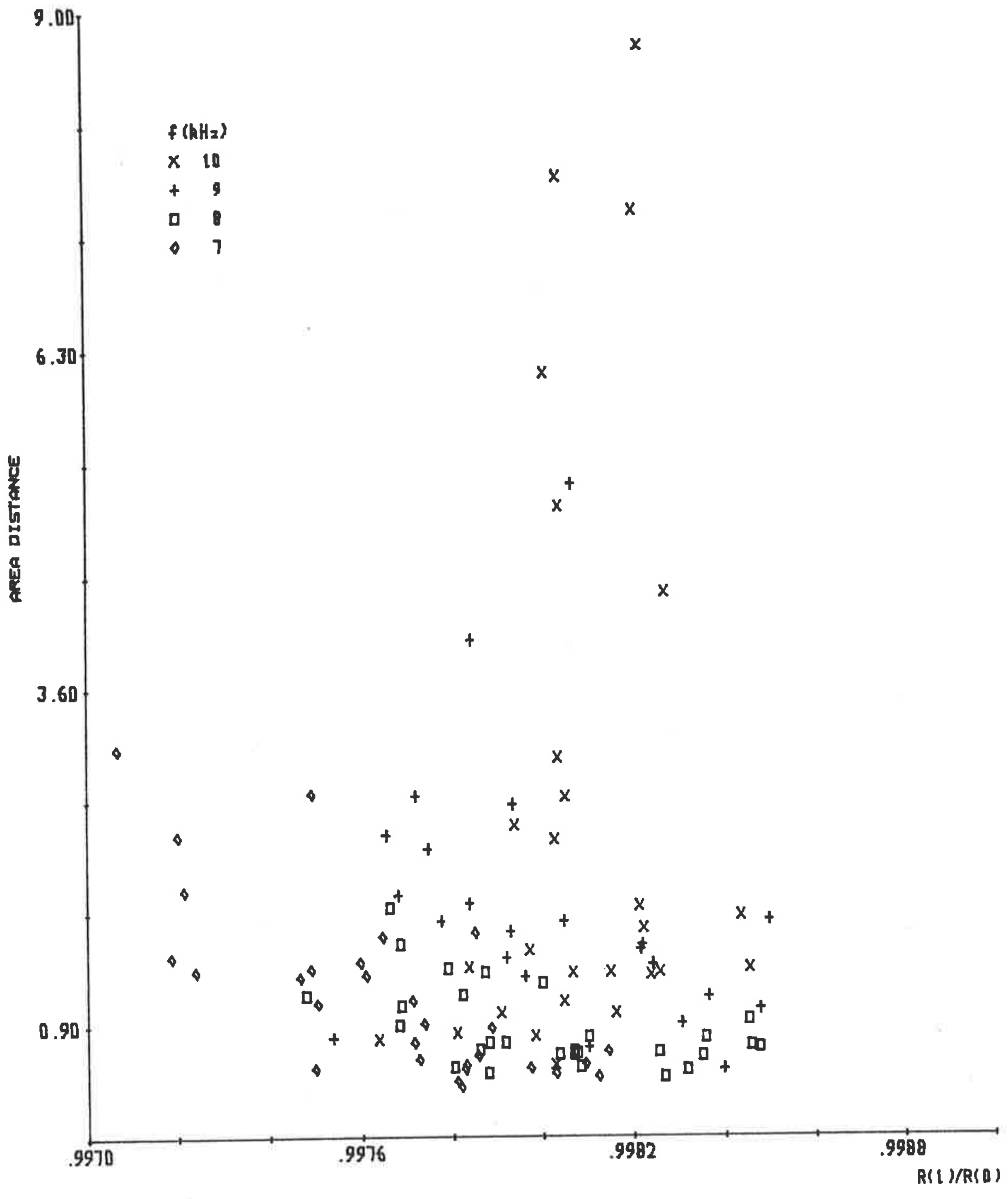
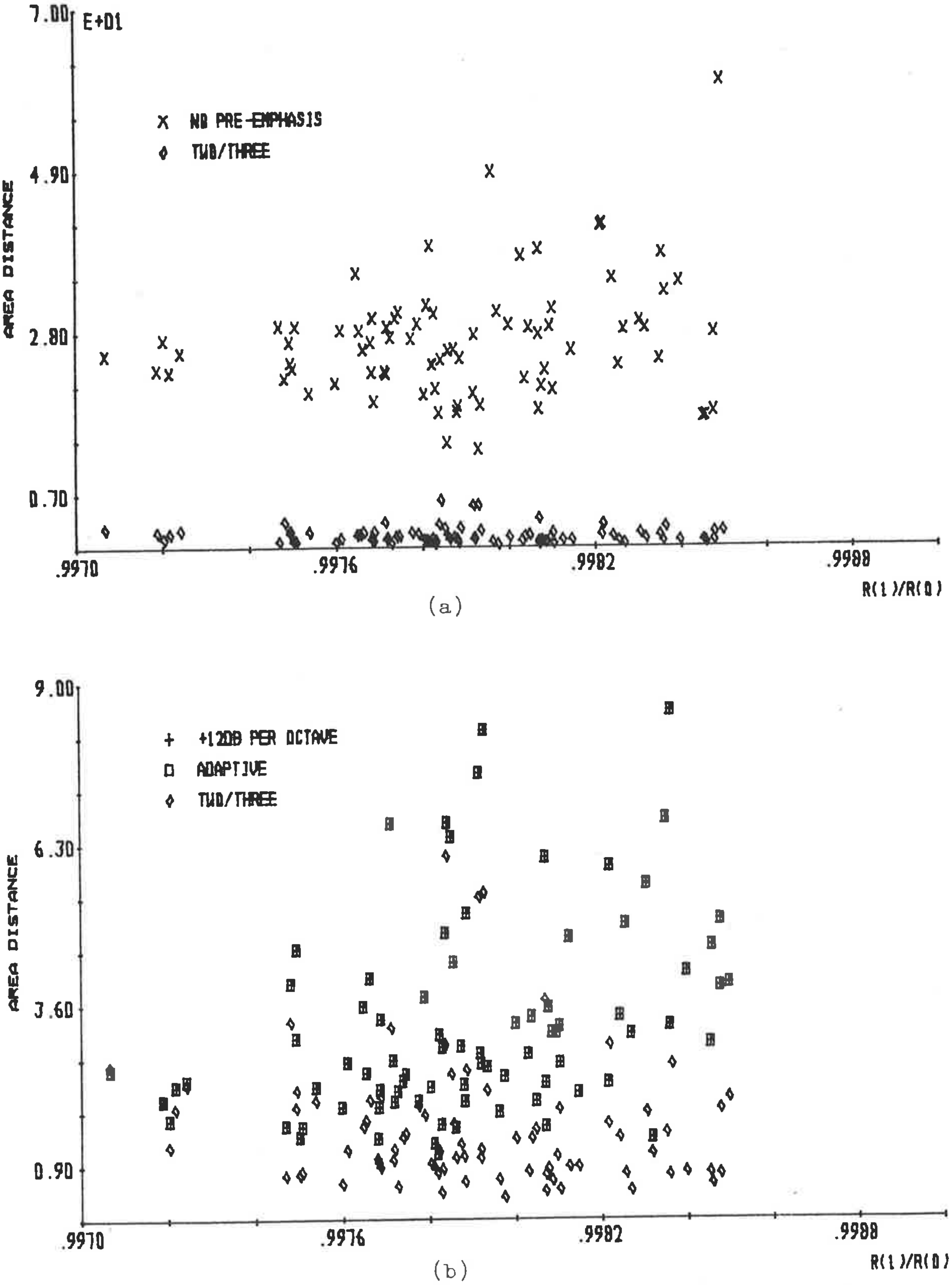


FIGURE 5.43: Area distances for the two/three adaptive pre-emphasis of synthetic speech for the vowel /e/ sampled at frequencies of 10, 9, 8 and 7 kilohertz

frequency of 10 kilohertz, when used to pre-emphasize synthetic speech of the vowel |e|.

The results presented in Figure 5.44 are for synthetic speech of the vowel |i| which have been sampled at frequencies of 9, 8 and 7 kilohertz with no pre-emphasis, a +12 dB per octave pre-emphasis, an unvoiced/voiced adaptive pre-emphasis and a two/three adaptive pre-emphasis. A large reduction in area distances is shown in Figure 5.44(a) for the two/three adaptive pre-emphasis in comparison with no pre-emphasis. A significant reduction in area distances is shown in Figure 5.44(b) for the two/three adaptive pre-emphasis filter in comparison with a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis. As observed for the synthetic speech for the vowels |a| and |e|, the performance of the two/three adaptive pre-emphasis to provide a reduction in area distance improves as $R(1)/R(0)$ increases.

The area distances versus $R(1)/R(0)$ for the two/three adaptive pre-emphasis of synthetic speech for the vowel |i| sampled at frequencies of 10, 9, 8 and 7 kilohertz are presented in Figure 5.45. Using different symbols to show the results for different sampling frequencies, Figure 5.45 shows that, except for a sampling frequency of 10 kilohertz, no significant change in area distances occurs for different sampling frequencies. Hence, for sampling frequencies between 9 and 7 kilohertz, inclusive, a similar performance of the two/three adaptive pre-emphasis filter to provide a reduction in area distances occurs, with a slightly poorer performance for a sampling frequency of 10 kilohertz.



(a)

(b)

FIGURE 5.44: Area distances for analysis of synthetic speech for the vowel /i/, sampled at frequencies of 9, 8 and 7 kilohertz, by (a) no pre-emphasis and a two/three adaptive pre-emphasis and (b) a +12 dB per octave, an unvoiced/voiced adaptive and two/three adaptive pre-emphasis

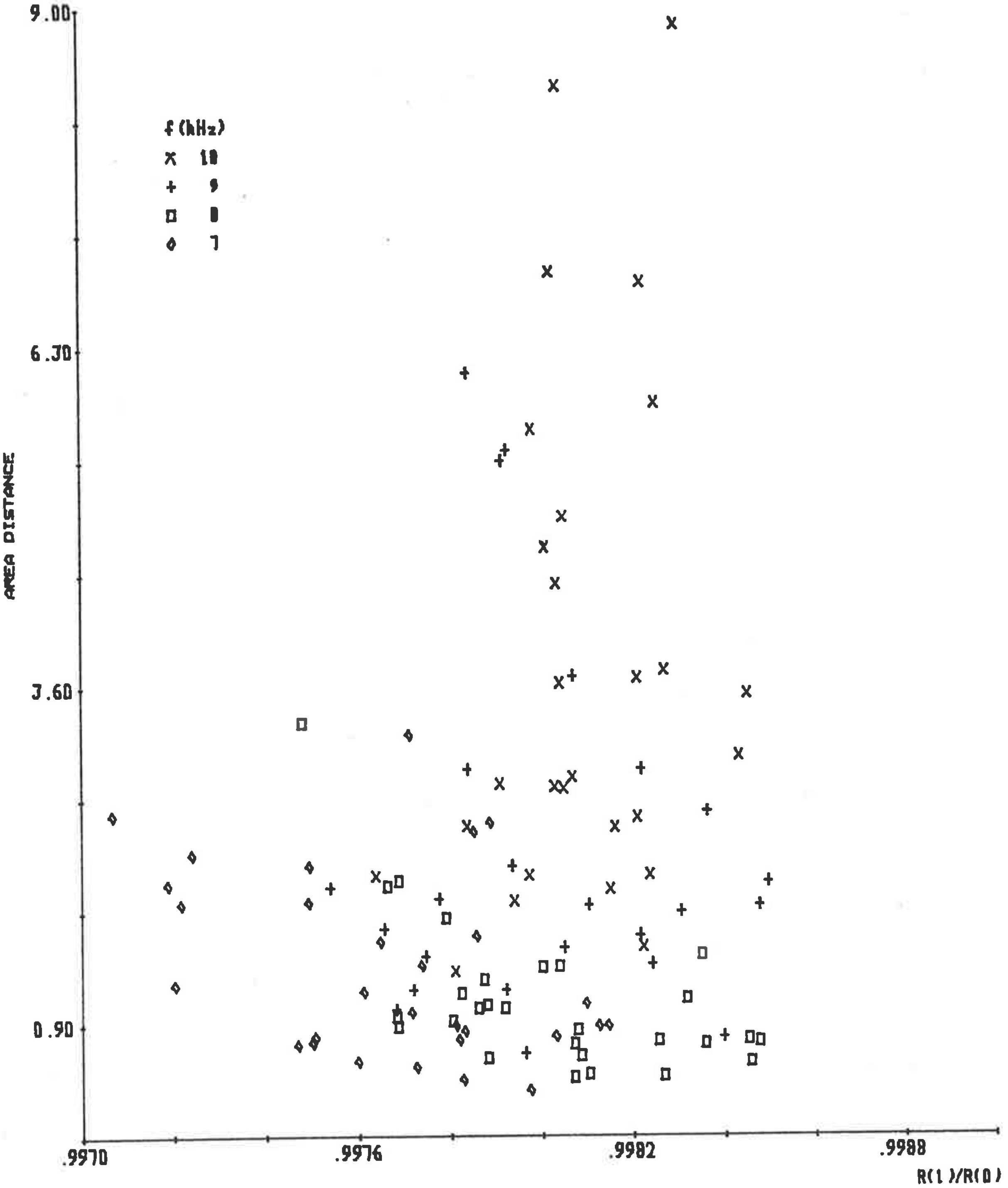


FIGURE 5.45: Area distances for the two/three adaptive pre-emphasis of synthetic speech for the vowel /i/ sampled at frequencies of 10, 9, 8 and 7 kilohertz

Figure 5.46 presents, for synthetic speech of the vowel |o|, the area distances versus $R(1)/R(0)$ for no pre-emphasis, a +12 dB per octave pre-emphasis, an unvoiced/voiced adaptive pre-emphasis and a two/three adaptive pre-emphasis, when sampling frequencies of 9, 8 and 7 kilohertz are used. The two/three adaptive pre-emphasis is shown in Figure 5.46(a) to produce a large reduction in area distances when compared with no pre-emphasis. In general, Figure 5.46(b) shows a small reduction in area distance by the two/three adaptive pre-emphasis in comparison with a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis. A general improvement in the performance of the two/three adaptive pre-emphasis filter when $R(1)/R(0)$ increases is observed in Figure 5.46, which is consistent with the observations for the two/three adaptive pre-emphasis of synthetic speech for the vowels |a|, |e| and |i|.

Figure 5.47 presents the area distances versus $R(1)/R(0)$ for a two/three adaptive pre-emphasis of synthetic speech for the vowel |o| sampled at frequencies of 10, 9, 8 and 7 kilohertz. Different symbols are used in Figure 5.47 to indicate the different sampling frequencies, and an even spread of data points (except for a few at a sampling frequency of 10 kilohertz) is observed in Figure 5.47. Hence, for sampling frequencies between 10 and 7 kilohertz, inclusive, the performance of the two/three adaptive pre-emphasis filter to provide a reduction in area distances is not significantly affected.

For synthetic speech of the vowel |u|, Figure 5.48 presents area distances versus $R(1)/R(0)$ for no pre-emphasis, a +12 dB per octave pre-emphasis, an unvoiced/voiced adaptive pre-emphasis and

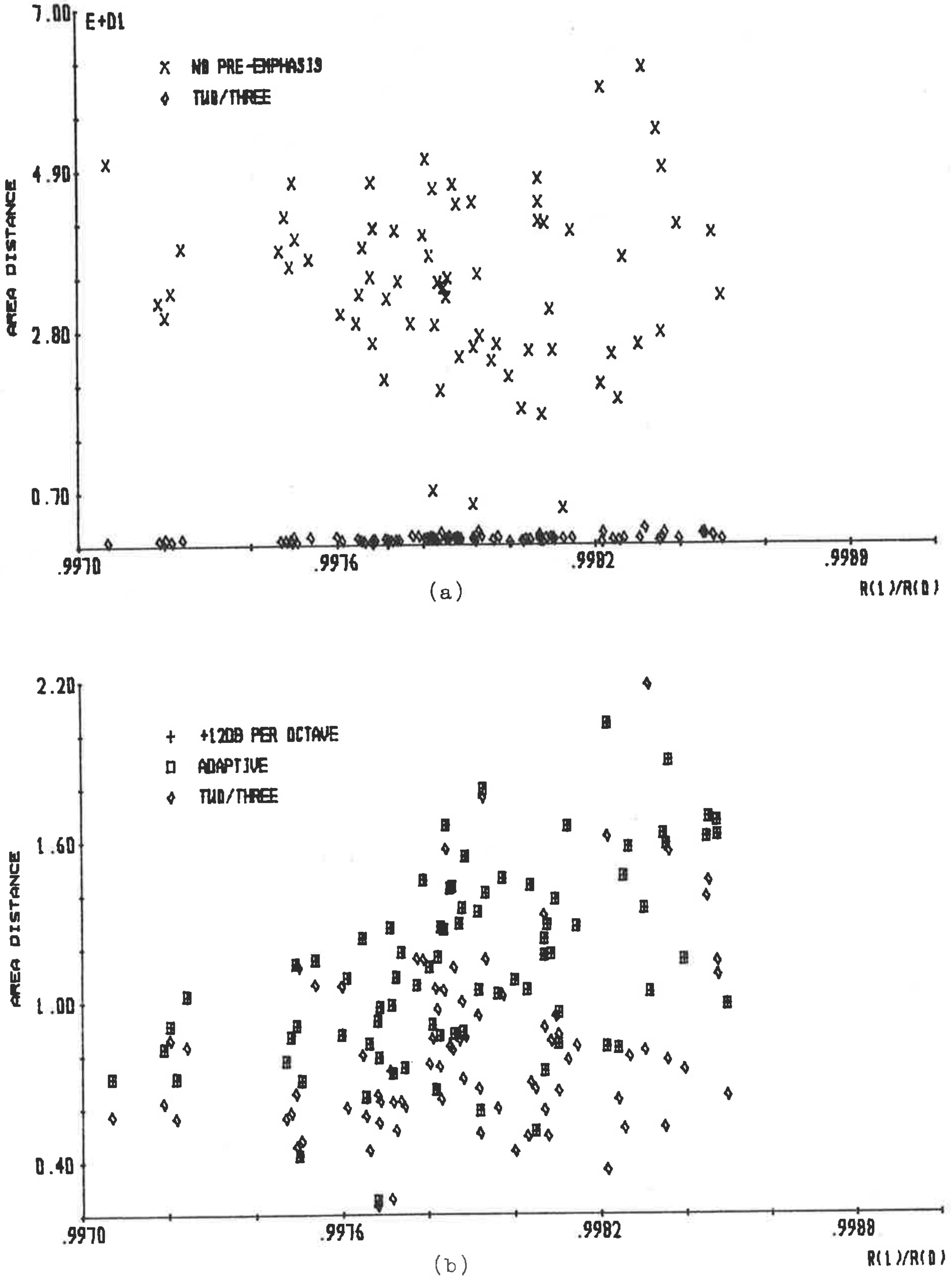


FIGURE 5.46: Area distances for analysis of synthetic speech for the vowel /o/, sampled at frequencies of 9, 8 and 7 kilohertz, by (a) no pre-emphasis and a two/three adaptive pre-emphasis and (b) a +12 dB per octave, an unvoiced/voiced adaptive and two/three adaptive pre-emphasis

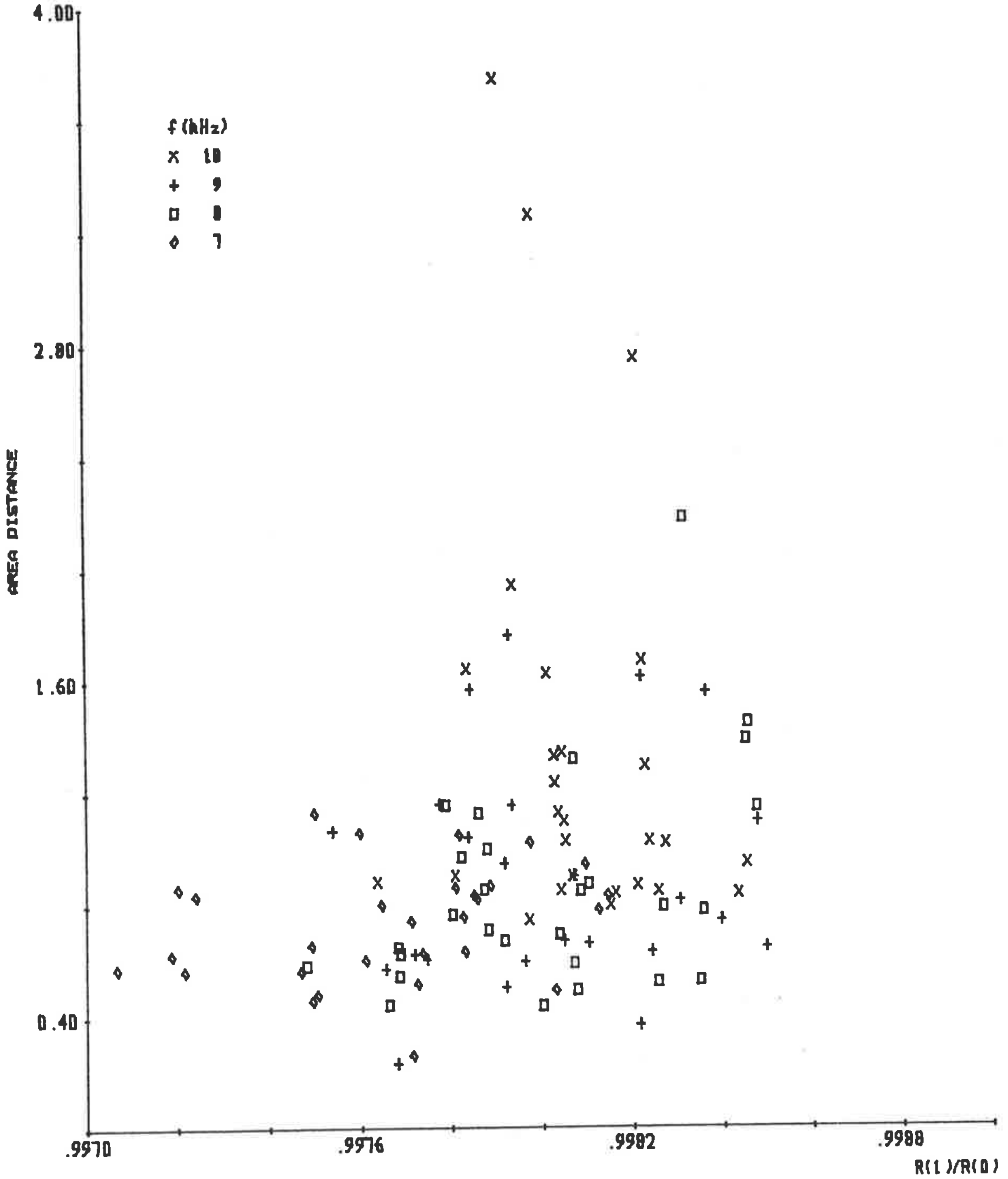
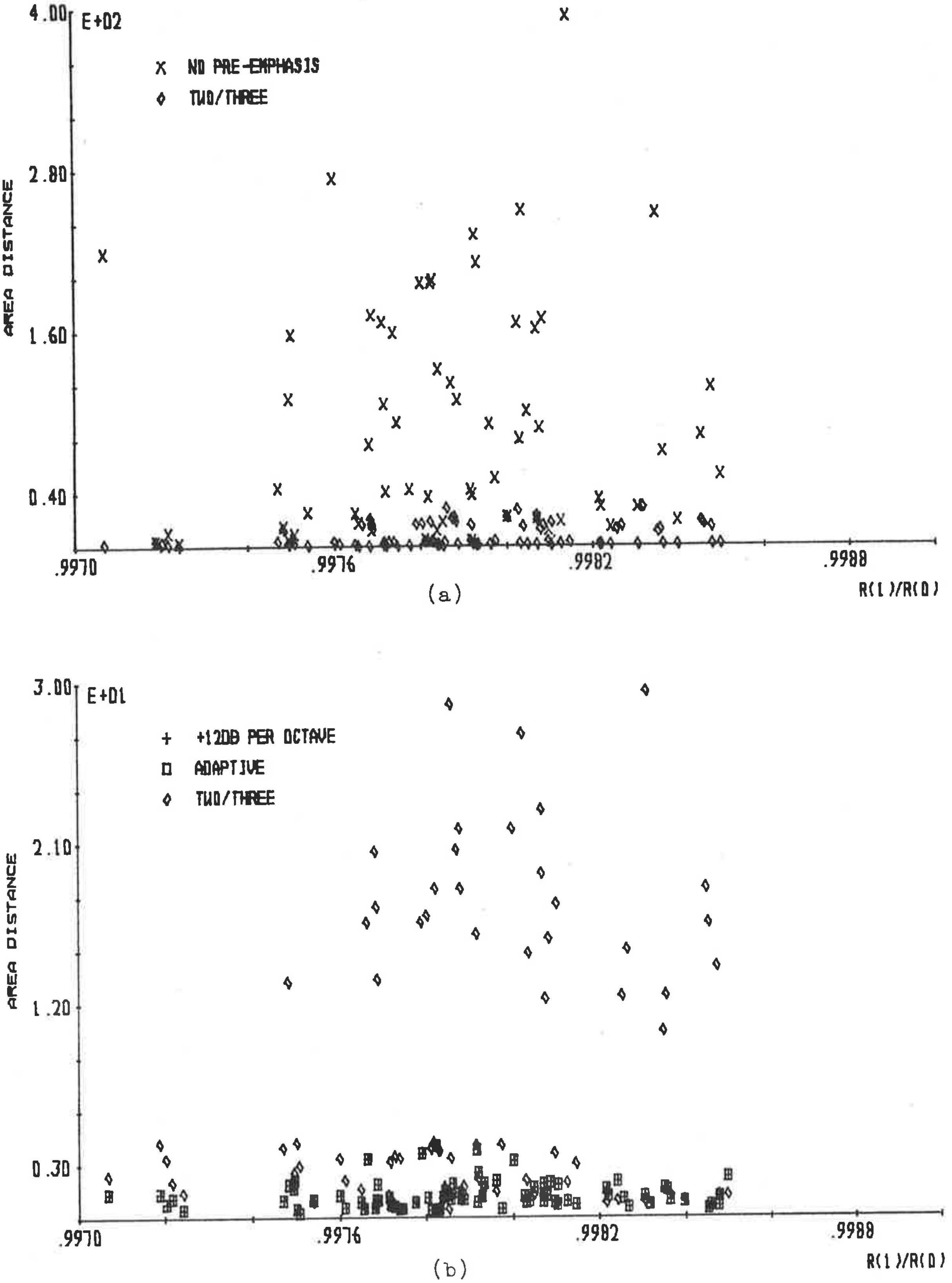


FIGURE 5.47: Area distances for the two/three adaptive pre-emphasis of synthetic speech for the vowel /o/ sampled at frequencies of 10, 9, 8 and 7 kilohertz



(a)

(b)

FIGURE 5.48: Area distances for analysis of synthetic speech for the vowel /u/, sampled at frequencies of 9, 8 and 7 kilohertz, by (a) no pre-emphasis and a two/three adaptive pre-emphasis and (b) a + 12 dB per octave, an unvoiced/voiced adaptive and two/three adaptive pre-emphasis

a two/three adaptive pre-emphasis when sampling frequencies of 9, 8 and 7 kilohertz are used. Figure 5.48(a) shows that the two/three adaptive pre-emphasis produces a reduction in area distances when compared with a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis.

For the four sampling frequencies of 10, 9, 8 and 7 kilohertz, Figure 5.49 presents the area distances versus $R(1)/R(0)$ for a two/three adaptive pre-emphasis of synthetic speech for the vowel $|u|$. The area distances for sampling frequencies of 10 and 8 kilohertz are shown in Figure 5.49 to be much larger than the area distances for sampling frequencies of 9 and 7 kilohertz. Hence, for synthetic speech of the vowel $|u|$, a change in sampling frequency between 10 and 7 kilohertz, inclusive, may affect the performance of the two/three adaptive pre-emphasis filter to provide a reduction in area distances.

The evaluations performed in this section have shown that, in general, when the two/three adaptive pre-emphasis filter is defined by the parameter β and the relationships between β' and $R(1)/R(0)$ presented in Chapter 4, the area distances for a two/three adaptive pre-emphasis of synthetic speech waveforms are not significantly affected by a change in sampling frequency between 10 and 7 kilohertz, inclusive. An exception to this statement occurs for synthetic speech of the vowel $|u|$.

Similar conclusions are drawn for the evaluations presented in Section 5.3.3.1, where the two/three adaptive pre-emphasis filter was defined by the parameter α, and a close comparison of the evaluations of this section and Section 5.3.3.1 shows that they

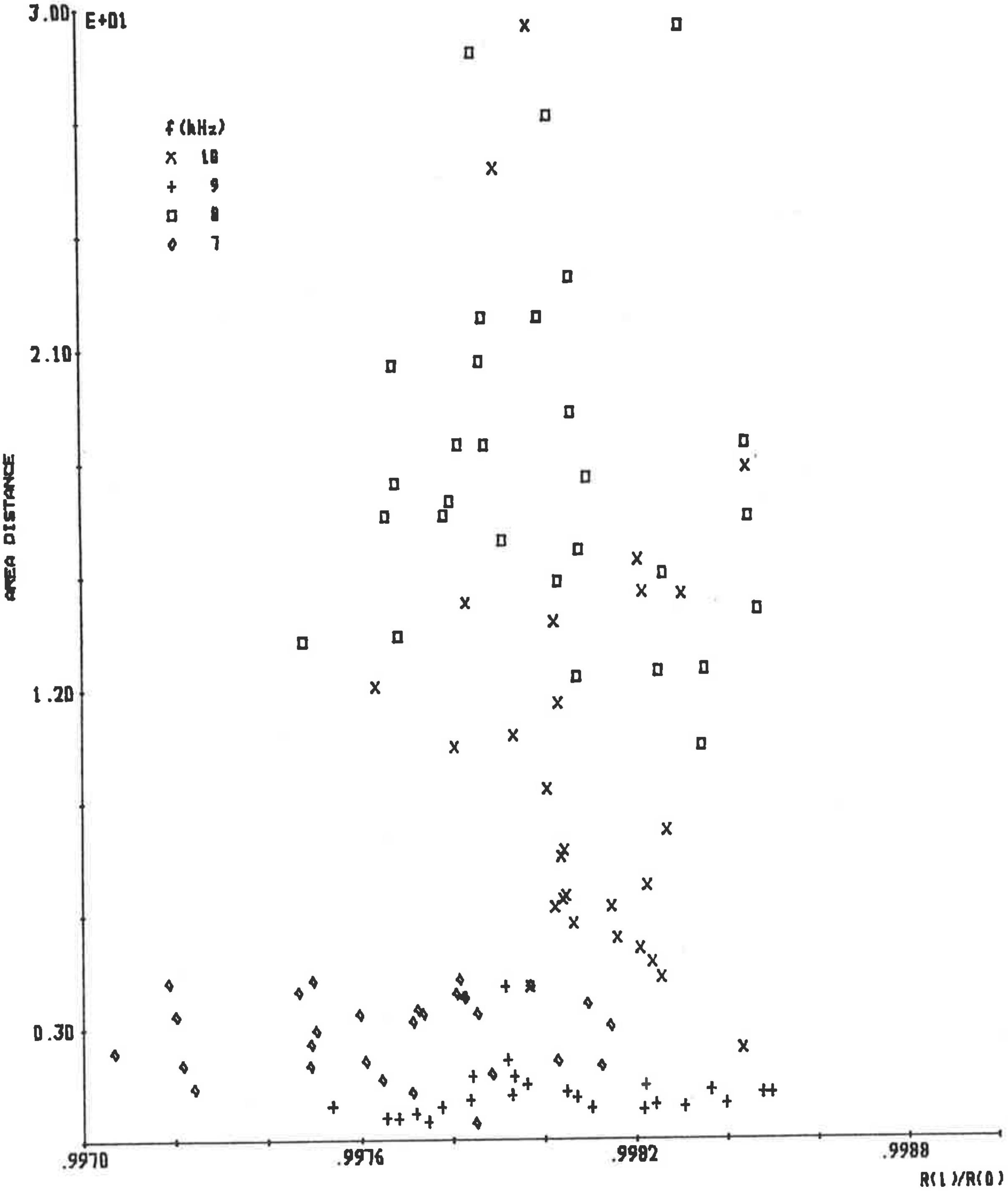


FIGURE 5.49: Area distances for the two/three adaptive pre-emphasis of synthetic speech for the vowel /u/ sampled at frequencies of 10, 9, 8 and 7 kilohertz

are consistent with each another. Hence, a general conclusion can be made that, except for synthetic speech of the vowel $|u|$, the area distances resulting from a two/three adaptive pre-emphasis of synthetic speech are not significantly affected by using a sampling frequency between 10 and 7 kilohertz, inclusive.

5.4 EVALUATION WITH REAL SPEECH DATA

This chapter has so far described the evaluation of the two/three adaptive pre-emphasis filter with glottal pulse waveforms (both synthetic and derived from measured glottal pulses) and synthetic speech data. In the case of synthetic speech, many properties of the vocal tract and speech production system have been ignored, e.g. lip radiation and vocal tract losses. This section evaluates the performance of the two/three adaptive pre-emphasis filter with real speech data for several vowel sounds.

A number of problems exist for evaluation of the two/three adaptive pre-emphasis filter with real speech waveforms. One problem is the difficulty of determining the vocal tract shape for a particular speech sound so that a comparison can be made with the recovered acoustic tube shape. Another problem is the clouding of the results by other non-ideal vocal tract properties such as lip radiation and vocal tract losses. Therefore, even if the two/three adaptive pre-emphasis completely removes glottal pulse excitation effects from the recovered acoustic tube shape, lip radiation and vocal tract losses prevent the vocal tract shape from being recovered. Hence, a careful interpretation of the recovered acoustic tube shape is necessary.

Since the non-ideal properties of the vocal tract cloud the effectiveness of the two/three adaptive pre-emphasis filter in removing glottal pulse excitation effects from the acoustic tube shape recovered by linear prediction, accurate vocal tract shapes for the speech sounds being analysed are not necessary. Hence, the vocal tract shapes determined from X-ray photographs by FANT [1970] and presented in Appendix C are compared with the recovered acoustic tube shapes. The real speech waveforms used in this section closely approximate the Russian vowels |a|, |e|, |i|, |o| and |u|.

A full description of the procedure and conditions under which the real speech waveforms were digitally recorded is presented in Appendix H. Briefly, the speech waveforms were recorded from seven Australian male speakers, each phonating the vowels |a|, |e|, |i|, |o| and |u| in |h-d| frames. A number of recording sessions were used, separated by at least one day, but not more than thirty days. At each recording session, the five vowels were spoken in the |h-d| frame, in a random order which changed for each recording session.

The digital recording of the real speech was performed by first filtering the speech by a low pass filter with a cut-off frequency of 4.5 kilohertz, and then sampling at a frequency of 10 kilohertz. The sampled speech waveform was stored on magnetic tape under the control of a minicomputer. Assuming the male vocal tract length is 17 cms, then using a sampling frequency of 10 kilohertz permits a 10 acoustic tube representation of the vocal tract shape.

The analysis procedure for the real speech data was to first choose 300 samples, representing a 30 msec time interval, of the vowel sound and window this data with a Hamming window [BLACKMAN and TUKEY 1958, MARKEL 1971, MAKHOUL and WOLF 1972]. The windowed data was then pre-emphasized with either a +12 dB per octave pre-emphasis, an unvoiced/voiced adaptive pre-emphasis, a two/three adaptive pre-emphasis or no pre-emphasis at all. A Parcor linear predictive analysis is then used to recover an acoustic tube shape which is compared by the area distance measure with the vocal tract shape measured by FANT [1970] for the corresponding vowel sound.

All the evaluation results presented in this section are in the form of area distances between the recovered acoustic tube shape and the corresponding vocal tract shape measured by FANT [1970] for the corresponding vowel sound plotted against $R(1)/R(0)$, which is evaluated from the autocorrelation function of the speech waveform. The reason for plotting area distances versus $R(1)/R(0)$ is that the value of $R(1)/R(0)$ determines the amount of pre-emphasis applied by the two/three adaptive pre-emphasis filter and the unvoiced/voiced adaptive pre-emphasis filter.

Figure 5.50 presents the area distances versus $R(1)/R(0)$ for no pre-emphasis, a +12 dB per octave pre-emphasis, an unvoiced/voiced adaptive pre-emphasis and a two/three adaptive pre-emphasis of real speech for the vowel |a|. A large reduction in area distances is observed in Figure 5.50 for the two/three adaptive pre-emphasis filter when compared with the +12 dB per octave pre-emphasis and unvoiced/voiced adaptive pre-emphasis filter. Similar, but slightly larger, area distances occur for the two/three adaptive pre-emphasis filter when compared with no pre-emphasis.

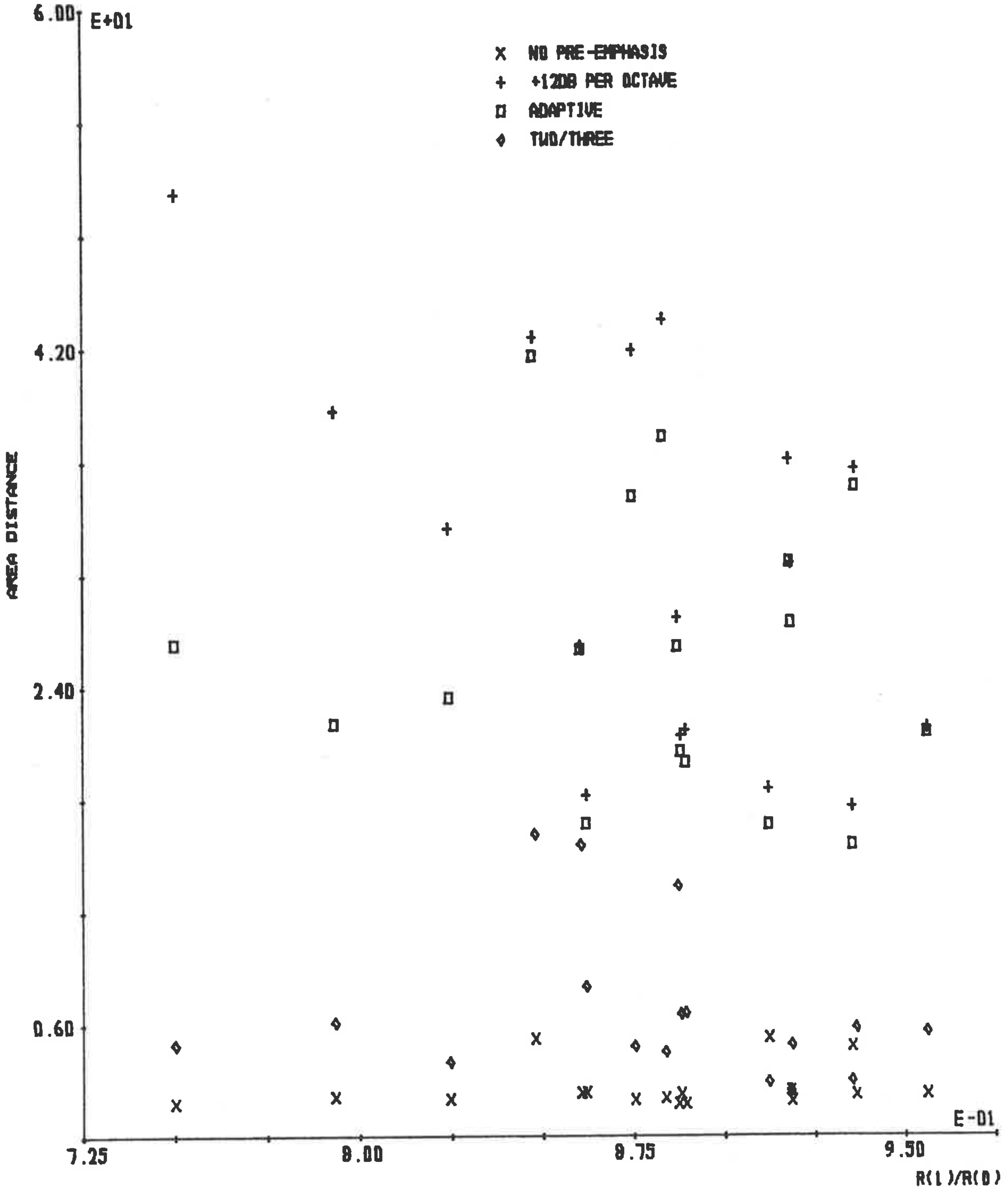


FIGURE 5.50: Area distances for analysis of real speech for the vowel /a/ by no pre-emphasis, +12 dB per octave, unvoiced/voiced adaptive and two/three adaptive pre-emphases

Since the true vocal tract shape is not known for the real speech being analysed, the above observation does not necessarily imply that a two/three adaptive pre-emphasis of the real speech provides poorer vocal tract shape recovery than when no pre-emphasis is used.

The general trends of area distance versus $R(1)/R(0)$ for the various pre-emphasis techniques presented in Figure 5.50 are similar to the general trends observed when synthetic speech for the vowel |a| was used in Section 5.3. That is, a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis provided much larger area distances than no pre-emphasis and a two/three adaptive pre-emphasis. Hence, for the range of $R(1)/R(0)$ presented in Figure 5.50 (i.e. the real speech situation) the evaluation results for synthetic and real speech of the vowel |a| are consistent with one another.

For real speech of the vowel |e|, the area distances versus $R(1)/R(0)$ for no pre-emphasis, a +12 dB per octave pre-emphasis, an unvoiced/voiced adaptive pre-emphasis and a two/three adaptive pre-emphasis are presented in Figure 5.51. For $R(1)/R(0)$ approaching unity, similar area distances occur for no pre-emphasis and the three pre-emphasis techniques. For $R(1)/R(0)$ less than unity, the area distances for a +12 dB per octave are much larger than the other two pre-emphasis techniques and no pre-emphasis. The area distances for a two/three adaptive pre-emphasis are much smaller than for an unvoiced/voiced adaptive pre-emphasis, and similar to no pre-emphasis.

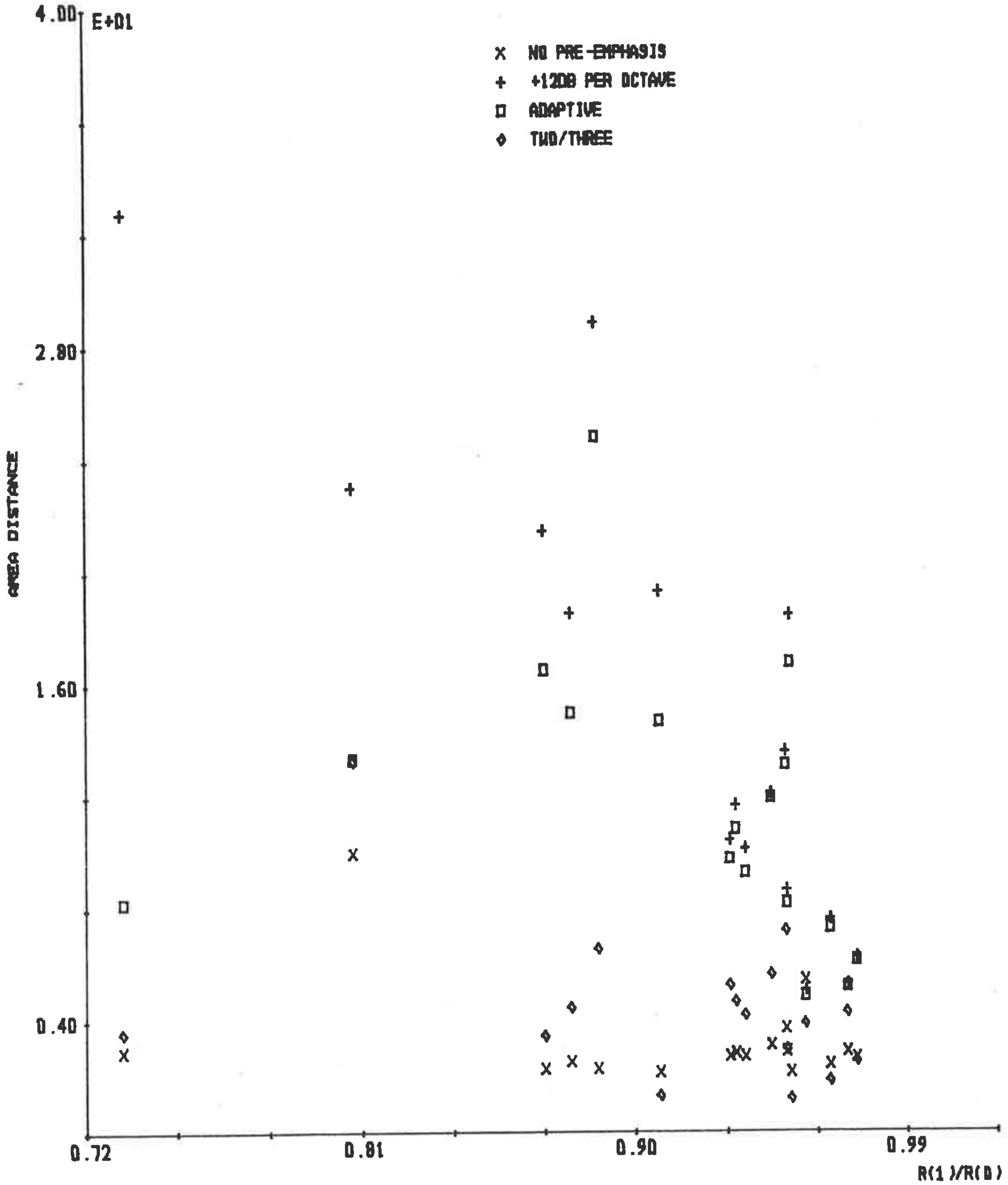


FIGURE 5.51: Area distances for analysis of real speech for the vowel /e/ by no pre-emphasis, +12 dB per octave, unvoiced/voiced adaptive and two/three adaptive pre-emphases

Therefore, for real speech of the vowel |e|, a reduction in area distances is achieved by the two/three adaptive pre-emphasis when compared with a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis. Since the true vocal tract shape is not know for the real speech being analysed, it is not possible to conclude, from the results presented in Figure 5.51, that improved vocal tract shape recovery is achieved by the two/three adaptive pre-emphasis filter in comparison with the vocal tract shape recovered when no pre-emphasis is used. The general trends observed in Figure 5.51 for real speech of the vowel |e| are found to be consistent with the general trends observed for synthetic speech of the vowel |e| presented in Section 5.3.

The area distances versus $R(1)/R(0)$ for no pre-emphasis, a +12 dB per octave pre-emphasis, an unvoiced/voiced adaptive pre-emphasis and a two/three adaptive pre-emphasis of real speech for the vowel |i| are presented in Figure 5.52. For $R(1)/R(0)$ close to unity, Figure 5.52 shows that a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis produce similar area distances, which are smaller than a two/three adaptive pre-emphasis and much smaller than no pre-emphasis. When $R(1)/R(0)$ is much less than unity, the +12 dB per octave pre-emphasis produces much larger area distances than no pre-emphasis and the other two adaptive pre-emphasis techniques. The two/three adaptive pre-emphasis produces the smallest area distances when $R(1)/R(0)$ is between 0.77 and 0.90.

The general trends observed in Figure 5.52 are consistent with the general trends presented in Section 5.3 for synthetic speech of the vowel |i|, except for the two/three adaptive pre-

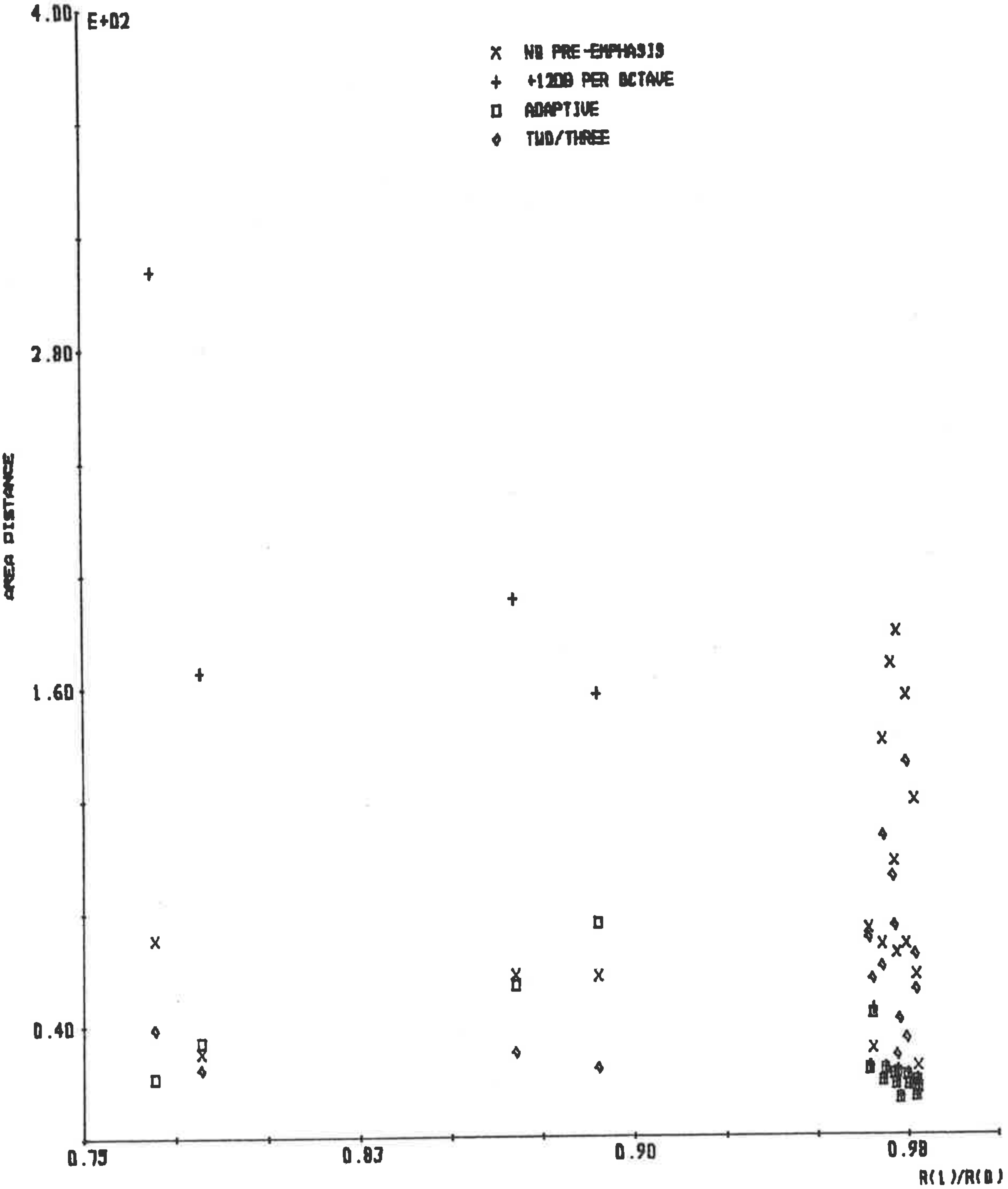


FIGURE 5.52: Area distances for analysis of real speech for the vowel /i/ by no pre-emphasis, +12 dB per octave, unvoiced/voiced adaptive and two/three adaptive pre-emphases

emphasis when $R(1)/R(0)$ is near unity. In comparison with the +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis, the two/three adaptive pre-emphasis filter produces larger area distances for real speech than for synthetic speech when $R(1)/R(0)$ is near unity.

Figure 5.53 presents the area distances versus $R(1)/R(0)$ for no pre-emphasis, a +12 dB per octave pre-emphasis, an unvoiced/voiced adaptive pre-emphasis and a two/three adaptive pre-emphasis of real speech for the vowel |o|. The area distances for the two/three adaptive pre-emphasis are consistently smaller than the area distances for a +12 dB per octave and an unvoiced/voiced adaptive pre-emphasis, but larger than the area distances for no pre-emphasis. Except for the smaller area distances returned by no pre-emphasis in comparison with the other three pre-emphasis techniques, the general trends observed in Figure 5.53 for real speech of the vowel |o| are consistent with the evaluation results presented in Section 5.3 for synthetic speech of the vowel |o|.

The results of no pre-emphasis, a +12 dB per octave pre-emphasis, an unvoiced/voiced adaptive pre-emphasis and a two/three adaptive pre-emphasis of real speech for the vowel |u| are present-as area distances versus $R(1)/R(0)$ in Figure 5.54. The range of $R(1)/R(0)$ presented in Figure 5.54 is much smaller and closer to unity than for the other figures presented in this section, and shows the tendancy for $R(1)/R(0)$ to be much higher for real speech of the vowel |u| than for the vowels |a|, |e|, |i| and |o|.

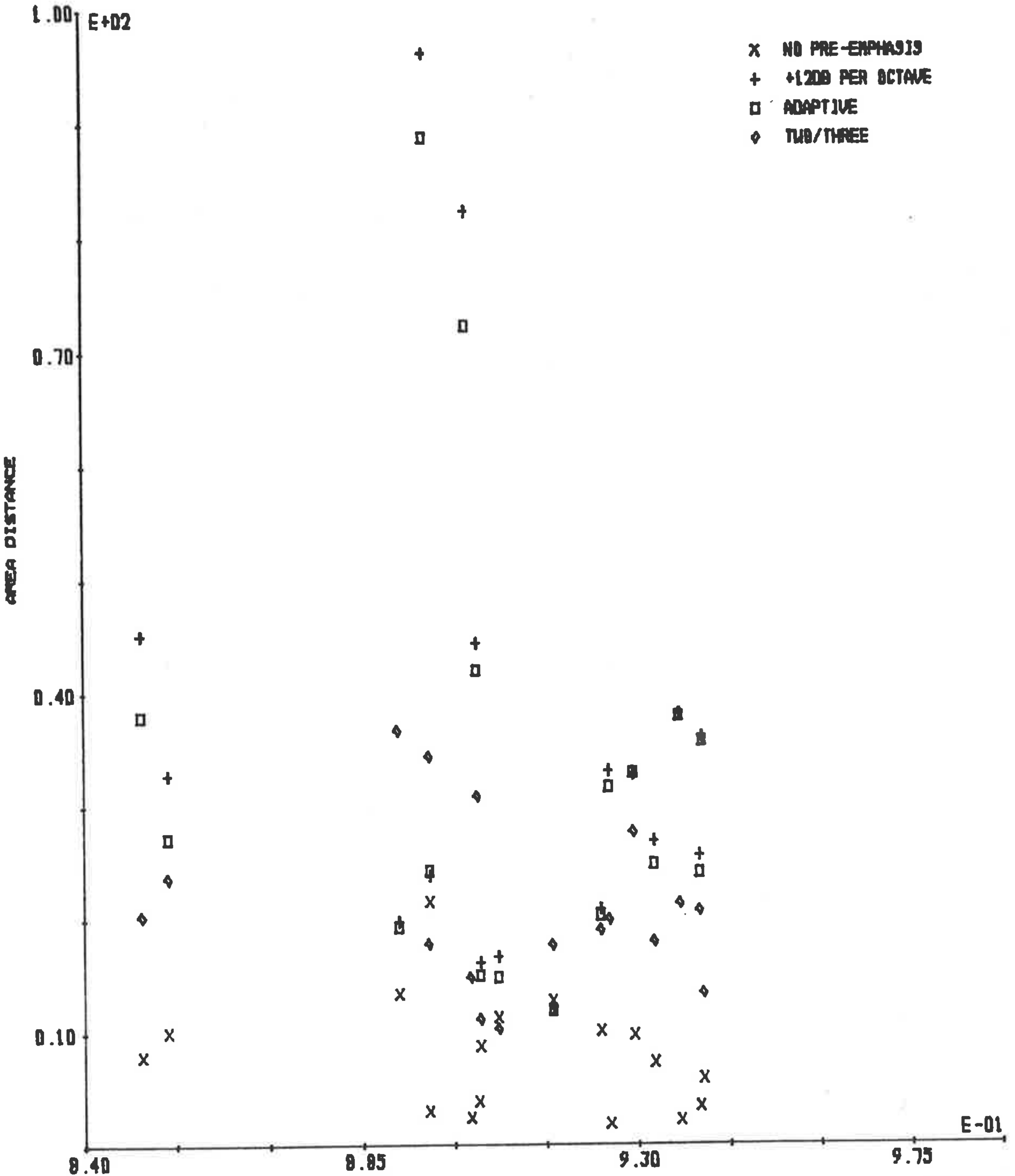


FIGURE 5.53: Area distances for analysis of real speech for the vowel /o/ by no pre-emphasis, +12 dB per octave, unvoiced/voiced adaptive and two/three adaptive pre-emphases

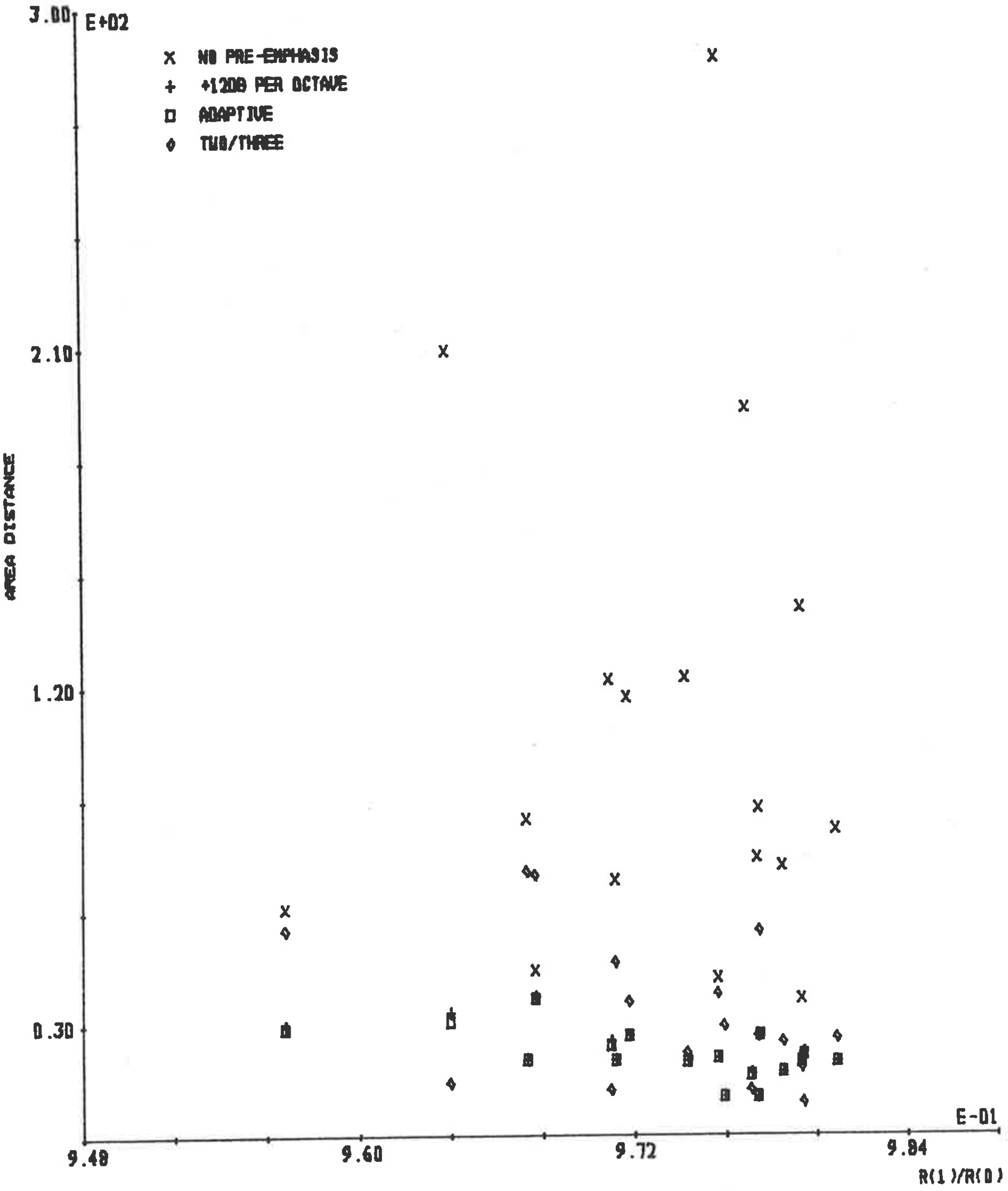


FIGURE 5.54: Area distances for analysis of real speech for the vowel /u/ by no pre-emphasis, +12 dB per octave, unvoiced/voiced adaptive and two/three adaptive pre-emphases

Figure 5.54 shows that the area distances produced by the three pre-emphasis techniques are, in general, smaller than the area distances for no pre-emphasis, a result which is consistent with the evaluation results presented in Section 5.3 for synthetic speech of the vowel |u|. When $R(1)/R(0)$ approaches unity, then the area distances for the three pre-emphasis techniques are similar. Slightly larger area distances occur for a two/three adaptive pre-emphasis than for a +12 dB per octave pre-emphasis and an unvoiced/voiced adaptive pre-emphasis when $R(1)/R(0)$ is less than approximately 0.97. Since the true vocal tract shape is not known, the latter observation does not necessarily imply that poorer vocal tract shape recovery occurs for a two/three adaptive pre-emphasis than for a +12 dB per octave pre-emphasis or an unvoiced/voiced adaptive pre-emphasis.

The evaluation results presented in this section appear to show that, in many cases, smaller area distances occur for no pre-emphasis than when any of the three pre-emphasis techniques are used. However, a careful interpretation of the results must be made, since the true vocal tract shape for each of the speech waveforms is not known; therefore, the area distances calculated only offer an indication of the accuracy of the recovered acoustic tube shape. Only the glottal pulse excitation is corrected for, and not all of the non-ideal properties of the vocal tract, which include lip radiation and vocal tract losses; hence, accurate vocal tract shape recovery cannot be expected.

A comparison of evaluations performed in this section for real speech of five vowel sounds with the evaluations presented in Section 5.3 for synthetic speech of the same five vowel sounds shows a similarity in the general trends of area distances versus $R(1)/R(0)$ for each pre-emphasis technique. Hence, in general, for the ranges of $R(1)/R(0)$ presented in the figures of this section, the evaluations for real and synthetic speech are consistent with one another.

In general, the area distances produced by using a two/three adaptive pre-emphasis of real speech for the vowels |a|, |e| and |o| are significantly smaller, and for the vowels |i| and |u| similar to the area distances produced by using a +12 dB per octave pre-emphasis or an unvoiced/voiced adaptive pre-emphasis. In conclusion, the evaluations presented in this section, using real speech, have shown that there is a significant advantage in using a two/three adaptive pre-emphasis instead of a +12 dB per octave pre-emphasis or an unvoiced/voiced adaptive pre-emphasis, if vocal tract shape recovery is desired.

5.5 SUMMARY

A new pre-emphasis filter, referred to as the two/three adaptive pre-emphasis filter, has been defined in Chapter 4 to permit improved acoustic tube/vocal tract shape recovery in comparison with existing pre-emphasis techniques. This chapter has presented a detailed evaluation of the two/three adaptive pre-emphasis filter in comparison with existing pre-emphasis techniques for both synthetic and real waveforms.

The evaluations presented in this chapter attempt to cover the many situations that occur in reality for the speech sounds of five vowels and, as a consequence, a very large number of evaluation results have been presented. As a guide to the reader who has not read this chapter in detail, Table 5.1 presents the overall performance of the two/three adaptive pre-emphasis filter for a sampling frequency of 10 kilohertz. The performance of the two/three adaptive pre-emphasis filter in comparison with other pre-emphases is indicated in Table 5.1 by an improvement factor. The improvement factor is defined as the ratio of area distances for an existing pre-emphasis technique to the area distance for a two/three adaptive pre-emphasis. Hence, an improvement factor of greater than unity implies that the two/three adaptive pre-emphasis filter produces smaller area distances and, in most cases, this implies an improvement in acoustic tube/vocal tract shape recovery.

All the improvement factors presented in Table 5.1 only represent maximum, typical and minimum values for the two/three adaptive pre-emphasis filter's performance. As seen in this chapter, it is difficult to judge the performance of any pre-emphasis techniques over the wide range of conditions presented in this chapter, from just the maximum, typical and minimum values of area distances it produces. Therefore, Table 5.1 should only be treated as a guide to the performance of the two/three adaptive pre-emphasis filter in comparison with the other pre-emphasis techniques.

		REAL GLOTTAL PULSES	SYNTHETIC SPEECH					REAL SPEECH				
			/a/	/e/	/i/	/o/	/u/	/a/	/e/	/i/	/o/	/u/
NO PRE-EMPHASIS	MAX.	15.5	40.0	7.5	20.6	43.0	24.3	1.0	1.4	4.5	1.2	10.1
	TYP.	2.0	2.0	2.0	4.0	2.0	1.0	1.0	1.0	1.5	0.5	4.0
	MIN.	1.0	1.0	1.0	1.0	0.1	0.1	0.3	0.3	1.0	0.1	0.7
+12dB PER OCTAVE PRE-EMPHASIS	MAX.	900.	384.	281.	84.5	133.	37.0	10.4	9.6	8.1	2.8	2.3
	TYP.	-	-	-	-	-	10.0	5.0	4.0	1.0	3.0	0.5
	MIN.	1.1	0.7	0.7	0.3	0.3	0.4	1.6	1.2	0.2	0.5	0.3
UNVOICED/ VOICED ADAPTIVE PRE-EMPHASIS	MAX.	32.8	30.8	4.9	42.7	25.1	28.7	7.1	13.8	3.2	2.6	2.3
	TYP.	15.0	10.0	3.0	4.0	10.0	10.0	5.0	3.0	4.0	3.0	0.5
	MIN.	1.0	0.7	0.2	0.3	0.3	0.4	1.5	1.0	0.2	0.5	0.3

TABLE 5.1: Improvement factor for the two/three adaptive pre-emphasis in comparison with other pre-emphases as applied to glottal pulse waveforms and both synthetic and real speech of five vowels.

The adaptive characteristic of the two/three adaptive pre-emphasis filter introduced in Chapter 4 was empirically determined by finding the necessary pre-emphasis to whiten a large variety of glottal pulse waveforms. Hence, the initial evaluation of the two/three adaptive pre-emphasis filter was performed with glottal pulse waveforms. Evaluations with simulated and measured glottal pulse waveforms sampled at a frequency of 10 kilohertz showed the two/three adaptive pre-emphasis filter to be significantly better at compensating for glottal pulse waveforms than existing pre-emphasis techniques.

During the empirical formulation of the two/three adaptive pre-emphasis filter, it was found that a change in sampling frequency caused changes in the area distances resulting from a two/three adaptive pre-emphasis. Therefore, a two/three adaptive pre-emphasis filter was defined for each of the sampling frequencies, 10, 9, 8 and 7 kilohertz, so that similar area distances occur for different waveform sampling frequencies. The evaluations presented in this section showed that defining a two/three adaptive pre-emphasis filter for different sampling frequencies permitted similar results to occur regardless of a change in sampling frequency between 10 and 7 kilohertz, inclusive.

Synthetic speech waveforms were used to study the performance of the two/three adaptive pre-emphasis filter in recovering acoustic tube shapes when those acoustic tube shapes are excited by glottal pulse waveforms. The synthetic speech waveforms are generated from the acoustic tube model the shape of which approximates five different vocal tract shapes for the vowel sounds |a|, |e|, |i|, |o| and |u|. The only violation of the ideal acoustic tube model is the non-white glottal pulse excitation. The area distance measure defined in Chapter 2 was used to provide a quantitative measure of the similarity between recovered and original acoustic tube shapes.

Evaluations performed with synthetic speech, generated for a sampling frequency of 10 kilohertz, showed that the two/three adaptive pre-emphasis filter, in general, provides a significant reduction in area distances when compared with conventional pre-emphasis techniques for the vowels |a|, |e|, |i|, |o| and |u|. The two/three adaptive pre-emphasis, when compared with no pre-emphasis, provides reduced area distances for the vowels |a|, |e| and |i|, but slightly larger area distances for the vowels |o| and |u|. Good acoustic tube shape recovery is achieved by the two/three adaptive pre-emphasis for the vowel |o|, even though the area distances are larger than for no pre-emphasis, but poor acoustic tube shape recovery occurs for the vowel |u|.

In general, the variation in area distances for a two/three adaptive pre-emphasis is much smaller than the variation in area distances for existing pre-emphasis techniques and no pre-emphasis. Therefore, the acoustic tube shapes resulting from the use of a two/three adaptive pre-emphasis are less sensitive to changes in glottal pulse excitation than other pre-emphasis techniques and no pre-emphasis. Hence, the two/three adaptive pre-emphasis filter is more effective at removing glottal pulse excitation effects from the recovered acoustic tube shape than existing techniques. The small variation in acoustic tube shapes has important applications in many areas, such as vocoders, where stable slowly varying parameters are required to represent the speech waveform.

The performance of the two/three adaptive pre-emphasis filter to recover improved acoustic tube shapes for sampling frequencies of 9, 8 and 7 kilohertz was also investigated. For synthetic speech of the vowels |a|, |e|, |i| and |o|, it was found that the two/three adaptive pre-emphasis filter has been defined such that no significant change in area distances occurs when the sampling frequency is between 10 and 7 kilohertz, inclusive. The area distances for a two/three adaptive pre-emphasis of synthetic speech for the vowel |u| are an exception, since they change significantly for changes in sampling frequency between 10 and 7 kilohertz, inclusive.

An evaluation of the two/three adaptive pre-emphasis filter was performed with real speech waveforms. However, a careful interpretation of the results is necessary, since the original vocal tract shape is not known and non-ideal properties of the vocal tract, such as lip radiation and vocal tract losses, cloud the results. Speech waveforms for the five vowels |a|, |e|, |i|, |o| and |u| were obtained from seven male speakers. The acoustic tube shapes used to determine the accuracy of the recovered acoustic tube shapes were derived from the X-ray measurements of the vocal tract shape for the equivalent Russian vowels.

In general, the two/three adaptive pre-emphasis is found to provide a large reduction in area distances in comparison with the existing pre-emphasis techniques for the vowels |a|, |e| and |o|, and similar area distances for the vowels |i| and |u|. When the evaluations for real speech are compared with those of synthetic speech, a similarity in the general trends of area distances versus $R(1)/R(0)$ for each pre-emphasis technique is found. Hence, the evaluations with real and synthetic speech are found to be consistent with one another.

The evaluations performed in this chapter with glottal pulse waveforms and then synthetic and real speech waveforms of five vowels have, in general, shown that the two/three adaptive pre-emphasis filter provides a general improvement in removing glottal pulse excitation effects from the recovered acoustic tube shape than existing pre-emphasis techniques. The use of the two/three adaptive pre-emphasis filter produces improved acoustic tube shape recovery with a small sensitivity to changes in the glottal pulse excitation and, except for the vowel |u|, a small sensitivity to changes in sampling frequencies between 10 and 7 kilohertz, inclusive.

CHAPTER 6

ACOUSTIC TUBES WITH A LOSSY TERMINATION

6.1 INTRODUCTION

Previous chapters have considered the violation of the excitation assumption of the linear prediction/acoustic tube model. However, this is not the only assumption of the linear prediction/ acoustic tube model that is violated in real situations. Investigations such as those performed by FLANAGAN [1972] have shown that radiation from the termination of the acoustic tubes or the vocal tract is significant, and should be taken into account. If radiation from the lips did not occur, there would be no point in speaking. This chapter considers the radiation from the termination of the acoustic tubes/vocal tract, and investigates procedures that may provide improved acoustic tube/vocal tract shape recovery when radiation occurs.

The violation of the ideal terminating condition of the linear prediction/acoustic tube model considered in this chapter is in the form of a lossy termination. The ideal termination condition is a completely lossless termination which does not allow any waveform to be radiated which, as discussed above, is not the case in most practical situations, especially the speech case. The amount of loss at the termination of acoustic tubes is defined by a radiation impedance which is defined later in this and also the following sections.

An exact model of radiation and the loss from an open-ended acoustic tube is complex, and difficult to represent. However, an approximate relationship between the pressure and volume velocity at the open-ended acoustic tube exists, if the relative size of the opening is small in comparison with the dimensions of the surroundings. This approximate relationship, from the work of MORSE and INGARD [1968], is

$$P_M(\ell,\omega) = Z_{M+1}(\omega)W_M(\ell,\omega) \tag{6.1}$$

where $P_M(\ell,\omega)$ and $W_M(\ell,\omega)$ are the frequency domain pressure and volume velocities at the termination, respectively, and $Z_{M+1}(\omega)$ is the radiation impedance. When the cross-sectional area of the radiating acoustic tube is small in relation to the dimensions of the surroundings, FLANAGAN [1972] has shown that the radiation impedance may be represented by a parallel combination of a normalized resistance, R_r, and a normalized inductance, L_r. Therefore, the normalized radiation impedance, $Z_r(\omega)$, may be expressed as

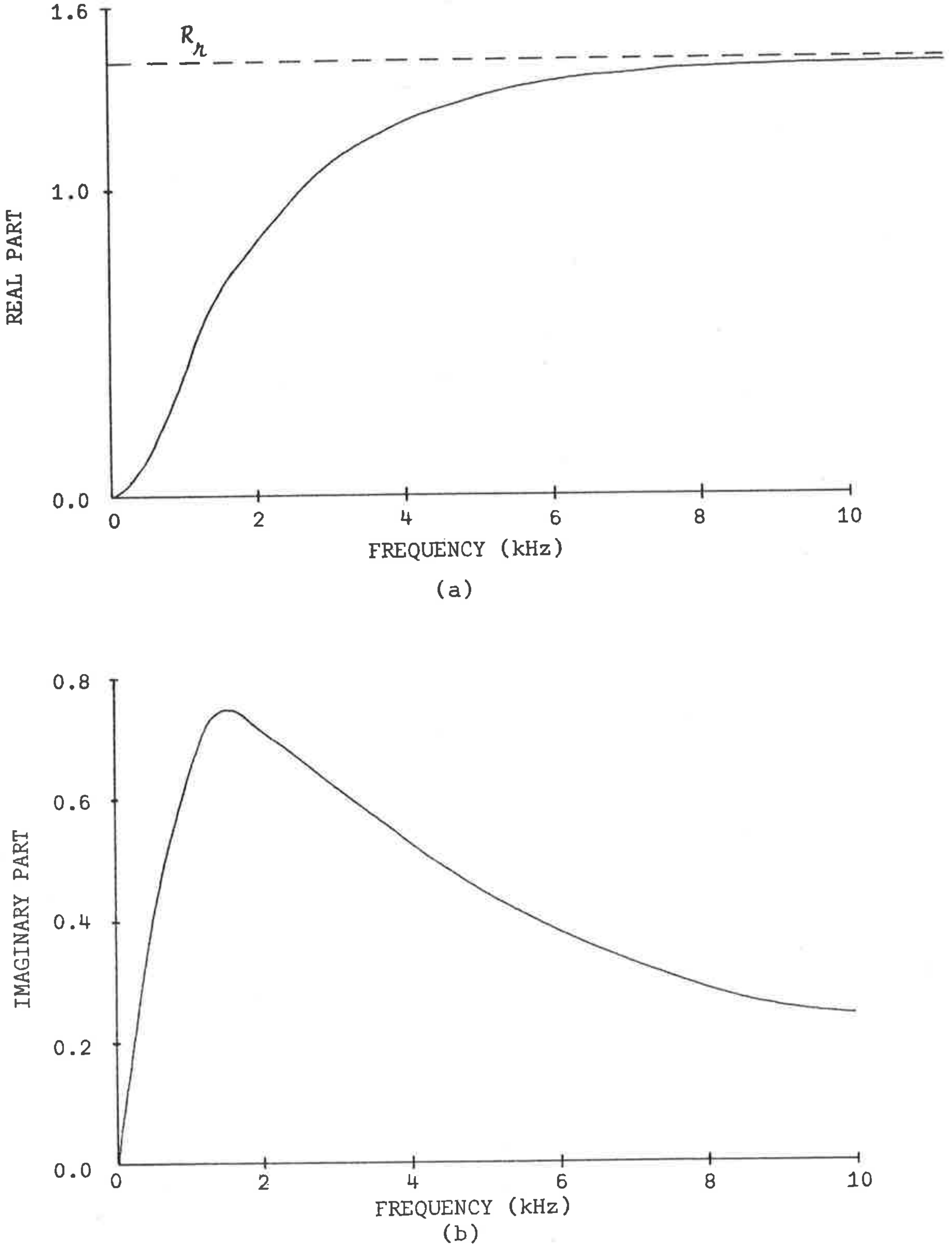
$$Z_r(\omega) = \frac{j\omega L_r R_r}{R_r + j\omega L_r} \tag{6.2}$$

The normalized resistance, R_r, and normalized inductance, L_r, are frequency independent quantities which are determined by the dimensions of the radiating acoustic tube and its surroundings. Since R_r and L_r are frequency independent, the radiation impedance, $Z_r(\omega)$, is a simple function of frequency. The radiation impedance is small, and tends to zero for low frequencies, which is the ideal condition required by the linear prediction/acoustic tube model, i.e. zero radiation impedance implies no radiated waveform. When $\omega L_r \ll R_r$ the radiation impedance, $Z_r(\omega)$, approximates the pure inductance L_r, and at high frequencies where $\omega L_r \gg R_r$ the radiation impedance, $Z_r(\omega)$, approximates the resistance R_r.

Figure 6.1 shows the real and imaginary parts of $Z_r(\omega)$ as a function of frequency for values of R_r and L_r that are typical of speech radiation. The energy radiated from the acoustic tube is proportional to the real part of the radiation impedance. Therefore, Figure 6.1(a) shows that the energy loss is frequency dependent, and most significant at high frequencies. In speech analysis, a -6 dB per octave pre-emphasis is applied to the speech waveform in an attempt to account for this frequency dependent radiation loss from the vocal tract at the lips.

In the speech application, the radiation loss cannot be separated from other losses in the vocal tract or from other non-ideal properties of the vocal tract; therefore, the effectiveness of a -6 dB per octave pre-emphasis to account for a radiation loss is difficult to obtain from speech waveforms. A study, using synthetic speech waveforms, to determine the effectiveness of a conventional -6 dB per octave pre-emphasis to account for radiation loss is presented in Section 6.3.

The development of analysis procedures that provide improved acoustic tube shape recovery in comparison with conventional procedures for a lossy termination requires an accurate model of real acoustic tube termination conditions. Section 6.2 presents the development of the approximations to realistic radiation presented above, into a realistic model of radiation loss. From this model and tractable mathematical requirements, a generalization of the termination assumption of linear prediction is obtained.



(a)

(b)

FIGURE 6.1: (a) Real and (b) Imaginary parts of the radiation impedance.

This generalized termination model is used to develop an analysis procedure which allows for a loss of energy at the termination of acoustic tubes.

The requirements for an improved analysis of a lossy termination of acoustic tubes are presented in Section 6.4. These requirements are used to define the basis for an analysis procedure which is presented in Section 6.5 which recovers an acoustic tube shape from the autocorrelation function of a waveform radiated from the lossy termination model described in Section 6.2. Another analysis procedure is presented in Section 6.6 which uses a transfer function of the acoustic tube model and places constraints on the recovered acoustic tube shape to overcome an ambiguity problem.

6.2 TERMINATION MODEL DEFINITION

An exact model of the radiation from an open-ended acoustic tube is complex, and cannot be represented in closed form. Therefore, it is necessary to consider simplifying assumptions which produce a tractable mathematical description of the termination model. The assumption of the linear prediction/acoustic tube model that requires the transverse dimensions of the acoustic tubes to be small when compared with the wavelength of the acoustic waveforms permits a significant reduction in the complexity of the radiation, or acoustic tube termination, model. For the speech application, this assumption permits the radiation from the lips to be modelled by a vibrating piston in a spherical baffle. MORSE and INGARD [1968] have shown that, when the radius of the piston is small in comparison with the radius of the spherical baffle, then the radiation impedance approximates that presented by a piston in an infinite baffle. FLANAGAN [1972] has shown that, for the vocal tract, the lip opening is small compared with the radius of the head; therefore, the piston in an infinite baffle is a realistic model for the radiation of speech at the lips.

The normalized radiation impedance for a piston in an infinite baffle is well known [MORSE and INGARD 1968, FLANAGAN 1972, BERANEK 1954], and has the closed form

$$Z_r(\omega) = \left[1 - \frac{J_1(2ka)}{ka}\right] + j\left[\frac{K_1(2ka)}{2(ka)^2}\right] \tag{6.3}$$

where $Z_r(\omega)$ is the normalized radiation impedance, and is related to $Z_{M+1}(\omega)$, the termination impedance of the acoustic tube model, by

$$Z_{M+1}(\omega) = \frac{\pi a^2}{\rho c} Z_r(\omega) \tag{6.4}$$

$J_1(2ka)$ and $K_1(2ka)$ are the first order Bessel functions [MORSE and INGARD 1968], ρ is the density of the acoustic medium, a is the radius of the piston, c is the velocity of sound, and k is the wavenumber defined by

$$k = \frac{\omega}{c} \tag{6.5}$$

If ka is small when compared with unity, then terms having exponents of $(2ka)$ greater than three are negligible, so that

$$J_1(2ka) \approx ka - \frac{(ka)^2}{2} \tag{6.6}$$

and

$$B_1(2ka) \approx \frac{16}{3\pi}(ka)^3 \qquad (6.7)$$

Therefore, by using Equation 6.6 and 6.7, the normalized impedance is approximated by

$$Z_r(\omega) \approx \frac{(ka)^2}{2} + j\,\frac{8ka}{3\pi} \qquad (6.8a)$$

$$= \frac{(a\omega)^2}{2c^2} + j\,\frac{8a\omega}{3\pi c} \qquad (6.8b)$$

when ka $(= \frac{\omega a}{c})$ is small compared with unity.

In circuit component terms, Equation 6.8 shows the normalized radiation impedance to be a resistance, proportional to ω^2, in series with an inductance which is independent of ω. Transforming the series combination of Equation 6.8 into a parallel one permits the normalized radiation impedance to be written as

$$Z_r(\omega) = \frac{j\omega R_r L_r}{R_r + j\omega L_r} \qquad (6.9)$$

(i.e. Equation 6.2) where

$$R_r = \frac{128}{9\pi^2} \qquad (6.10)$$

and

$$L_r = \frac{8a}{3\pi c} \qquad (6.11)$$

for ka $(= \frac{\omega a}{c})$ small compared with unity.

Equations 6.10 and 6.11 permit Equation 6.8(b) to be rewritten as

$$Z_r(\omega) = \frac{\omega^2 L_r^2}{R_r} + j\omega L_r \tag{6.12}$$

for ka small compared with unity. The normalized resistance, R_r, is a constant, and the normalized inductance, L_r, is dependent on only the radius of the piston in the radiation model; hence, both R_r and L_r are constants for a particular situation. From Equations 6.9 and 6.12, the normalized radiation impedance, $Z_r(\omega)$, is seen to approach zero or an ideal short circuit for small ω and $Z_r(\omega)$ approximates the pure normalized resistance, R_r, for large ω.

From the definition of cross-sectional area, i.e. Equation 2.19, Equation 6.4 simplifies to

$$Z_{M+1}(\omega) = \frac{\rho c}{A_M} Z_r(\omega) \tag{6.13}$$

and, hence, Equation 6.1 is rewritten as

$$P_M(\ell,\omega) = \frac{\rho c}{A_M} Z_r(\omega)\ W_M(\ell,\omega) \tag{6.14}$$

and, on substitution of Equation 6.8, is

$$P_M(\ell,\omega) = \frac{\rho c}{A_M} \cdot \frac{j\omega R_r L_r}{R_r + j\omega L_r} W_M(\ell,\omega) \tag{6.15}$$

Equations 2.6 and 2.7 permit the time domain equivalent of Equation 6.15 to be written as

$$L_r(R_r-1)\,\frac{dW_M^+(t-\ell/c)}{dt} - R_r W_M^-(t-\ell/c)$$

$$= L_r(R_r+1)\,\frac{dW_M^-(t-\ell/c)}{dt} + R_r W_M^-(t-\ell/c) \qquad (6.16)$$

Equation 6.16 is only valid when ka is small compared with unity, and when R_r and L_r are constants.

The simplifications used to derive Equation 6.16 are based on the requirement that ka is small with respect to unity, i.e.

$$ka = \frac{\omega a}{c} = \frac{2\pi a}{c}\, f \ll 1 \qquad (6.17a)$$

which defines an upper limit to the allowable frequencies as

$$f \ll \frac{c}{2\pi a} \qquad (6.17b)$$

The maximum frequency for which Equation 6.16 is valid is, therefore, inversely proportional to the radius of the piston in a spherical baffle model, i.e. the larger the cross-sectional area of the piston, the smaller the maximum allowable frequency. FLANAGAN [1972] reports that a typical extreme lip opening area is approximately 5 cms^2, which represents a piston radius, a, of 1.3 cms, and so the maximum allowable frequency is 4 kilohertz, via Equation 6.17(b). This is not an unrealistic limit, since the linear prediction/acoustic tube model of the vocal tract requirement that the transverse dimensions of the acoustic tubes modelling the vocal tract be small compared with the wavelength of the acoustic waveforms breaks down for frequencies above 4 kilohertz. Therefore, Equation 6.16 may be used to represent the radiation impedance for the termination of the vocal tract at the lips, if the speech signal is low pass filtered to 4 kilohertz.

The analysis procedures used in this thesis assume discrete sampled waveforms, and so Equation 6.16 must be converted into a discrete sampled waveform equation. This is achieved by firstly converting Equation 6.1 into the time domain, and using Equations 2.6 and 2.7 in conjunction with the z domain form of the normalized radiation impedance, i.e. $Z_r(z)$. Of these steps, the most difficult is the determination of the z domain form of $Z_r(\omega)$.

There exists a large number of methods for converting frequency domain quantities into the z domain [OPPENHEIM and SCHAFER 1975, GOLD and RADER 1969, RADER and GOLD 1967, KAISER 1963 and 1966] with the bilinear transform [GOLDEN and KAISER 1964, STEIGLITZ 1965, GIBBS 1970, RADER and GOLD 1967, OPPENHEIM and SCHAFER 1975] being a well-established and widely used procedure. The bilinear transform performs the transformation from the s plane to the z plane via

$$s = \frac{2}{T}\left(\frac{1-z^{-1}}{1+z^{-1}}\right) \qquad (6.18)$$

where T is the sampling period. Equation 6.18 produces a one to one mapping of the left half s plane into the unit circle of the z domain, with the $j\omega$ or s axis mapped onto the unit circle. The right hand s plane is mapped to the exterior of the unit circle. Application of the bilinear transform to convert the quantity $H(s)$ to $H'(z)$ is performed by a simple substitution of Equation 6.18, i.e.

$$H'(z) = H(s)\Bigg|_{s = \frac{2}{T}\left(\frac{1 - z^{-1}}{1 + z^{-1}}\right)} \tag{6.19}$$

A disadvantage of the simple bilinear transform described by Equation 6.19 is a non-linear frequency warping. An exact form of the frequency warping is obtained by a direct substitution of the definition of the variables s and z, i.e. $s = j\omega$ and $z = e^{j\omega_1 T}$, into Equation 6.18, to produce

$$\frac{\omega T}{2} = \tan\left(\frac{\omega_1 T}{2}\right) \tag{6.20}$$

When $\omega_1 T$ is small, then $\tan\left(\frac{\omega_1 T}{2}\right)$ is approximately $\frac{\omega_1 T}{2}$, and so an approximately linear relationship exists between ω and ω_1, i.e. frequency warping is negligible. For $\omega_1 T$ near the half sampling frequency (i.e. $\omega_1 = \pi/T$) considerable frequency warping occurs, which requires compensation via Equation 6.20. In practice, the bilinear transform is only used in situations where the range of frequencies considered is less than the half sampling frequency.

If the normalized radiation impedance is defined in the z domain as $Z_r(z)$, then Equation 6.19 allows $Z_r(z)$ to be evaluated as

$$Z_r(z) = Z_r(s)\Bigg|_{s = \frac{2}{T}\left(\frac{1 - z^{-1}}{1 + z^{-1}}\right)} \tag{6.21}$$

Equations 6.9 and 6.21 enable the normalized radiation impedance to be written, in the z domain, as

$$Z_r(z) = b\left[\frac{1 - z^{-1}}{1 - pz^{-1}}\right] \tag{6.22a}$$

where

$$b = \frac{2R_r L_r}{2L_r + TR_r} \tag{6.22b}$$

and

$$p = \frac{2L_r - TR_r}{2L_r + TR_r} \tag{6.22c}$$

Equations 6.22b and 6.22c are rewritten in terms of the sampling period, T, the velocity of sound, c, and the piston radius a, by the use of Equations 6.10 and 6.11, i.e.

$$b = \frac{128a}{3\pi(3\pi a + 8Tc)} \tag{6.23a}$$

and

$$p = \frac{3\pi a - 8Tc}{3\pi a + 8Tc} \tag{6.23b}$$

Equation 6.19 shows that the normalized radiation impedance in the z domain, $Z_r(z)$, is defined by a fixed zero at

$$z = 1 \tag{6.24a}$$

and a real variable pole at

$$z = p \tag{6.24b}$$

and a magnitude term, b. Equation 6.23 shows the position of the real pole and the value of the magnitude term to be variable, and dependent on the sampling frequency, T, the velocity of sound, c (which is generally constant), and the piston radius, a.

The position of the real pole of $Z_r(z)$ is important for a further simplification of the radiation model, and from Equation 6.23b the relative magnitudes of $3\pi a$ and $8Tc$ determine on which side of the origin the real pole lies, i.e.

$$p < 0 \quad \text{for} \quad a < \frac{8Tc}{3\pi} \qquad (6.25a)$$

and

$$p > 0 \quad \text{for} \quad a > \frac{8Tc}{3\pi} \qquad (6.25b)$$

In the situation where the sampling rate is 10 kilohertz, Equation 6.25 becomes

$$p < 0 \quad \text{for} \quad a < 2.89 \text{ cms} \qquad (6.26a)$$

and

$$p > 0 \quad \text{for} \quad a > 2.89 \text{ cms} \qquad (6.26b)$$

Investigations by FLANAGAN [1972] show that for the speech application, the maximum lip opening has a radius of around 1.3 cms. Hence, from Equation 6.26 it can be expected that the real pole of the radiation model is always to the left of the origin in the z plane.

Conventionally, the pole of $Z_r(\omega)$ is neglected, so that the radiation model is described by a single zero at $z = 1$. Using the radiation model developed in this section and described by Equations 6.22 and 6.23, an evaluation of the validity of neglecting the pole of $Z_r(z)$ is performed. One approach to performing this evaluation is presented below, and that is to compare the frequency response of $Z_r(\omega)$ with and without the pole present. Another approach, which is presented later, is to determine the relative positions of the pole and zero of $Z_r(\omega)$ and their relative contributions to $Z_r(\omega)$ as a function of frequency.

The frequency response of $Z_r(\omega)$ is plotted in Figure 6.2 for the piston radius, a, having values 0.3, 0.7, 1.1 and 1.5 centimeters. Figure 6.2 shows a constant +6 dB per octave spectral slope and 90° phase angle for low frequencies, which is the frequency response of a single zero. For frequencies near 4 kilohertz, a departure from the expected frequency response of a single zero is observed, varying from 0.1 dB and 7° to 1.5 dB and 30°, at 4 kilohertz, when a is 0.3 and 1.5 centimeters, respectively. Therefore, the pole of $Z_r(\omega)$ only provides significant change to the response of a single zero at frequencies near 4 kilohertz, and then only when a approaches 1.5 centimeters. Figure 6.2 also shows that the pole of $Z_r(\omega)$ produces a scaling of the magnitude of $Z_r(\omega)$.

The assumption used to reduce Equation 6.3 to Equation 6.9 requires a limitation of frequencies to less than 4 kilohertz, and for this limitation the above discussion indicated that the pole of $Z_r(\omega)$ has only a small effect on the frequency response of $Z_r(\omega)$. Therefore, a restriction of allowable frequencies to

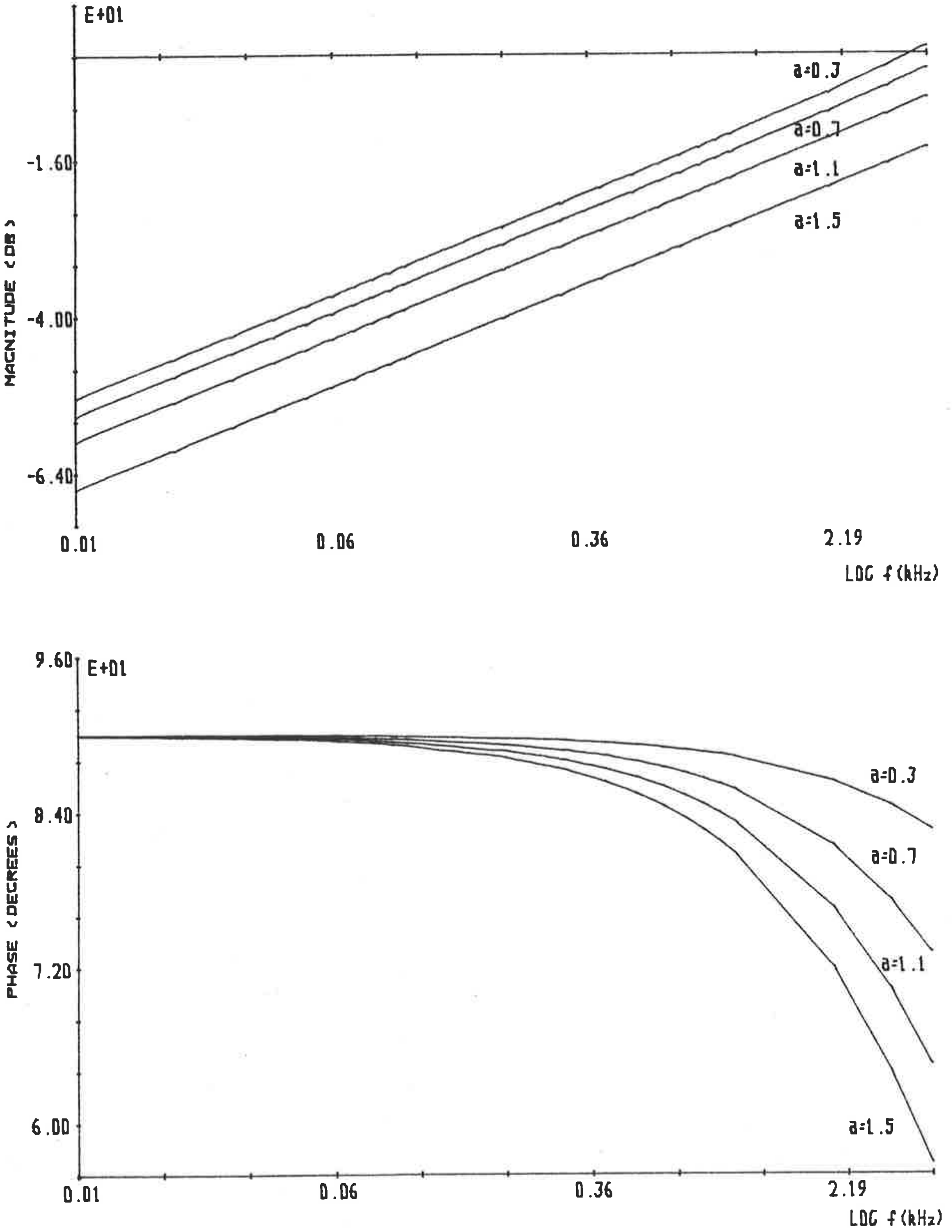


FIGURE 6.2: Frequency response of $Z_4(\omega)$ for various values of piston radius, a.

less than 4 kilohertz permits $Z_r(\omega)$ to be approximated by a single zero and magnitude term.

The pole of the normalized radiation impedance, $Z_r(\omega)$, from Equation 6.9 occurs at

$$j\omega = -\frac{R_r}{L_r} \tag{6.27a}$$

which, by using Equations 6.10 and 6.11, is rewritten as

$$j\omega = -\frac{16c}{3\pi a} \tag{6.27b}$$

Equation 6.27 shows that the pole of $Z_r(\omega)$ lies on the negative $j\omega$ axis, and its position is inversely proportional to the piston radius, a (see also Equation 6.25). For the extremes of piston radius, a, as suggested by FLANAGAN [1972] for normal vowel production, i.e. 0.3 and 1.3 centimeters, the pole of $Z_r(\omega)$ occurs at $j\omega$ equal to $-(2\pi)\times 3\times 10^4$ and $-(2\pi)\times 0.7\times 10^4$ radians/sec, respectively. Clearly, with $Z_r(\omega)$ being evaluated for ω in the range zero to $(2\pi)\times 4\times 10^4$ radians/sec, the pole of $Z_r(\omega)$ when the piston radius is 0.3 centimeters has little effect. When the piston radius, a, is near 1.3 centimeters, the pole is significant, but its largest contribution to $Z_r(\omega)$ is less than half that provided by the zero of $Z_r(\omega)$.

The above discussion, in association with Figure 6.2, leads to the conclusion that the pole of $Z_r(\omega)$ has only a small effect on the shape of the magnitude of $Z_r(\omega)$ versus frequency. However, the pole of $Z_r(\omega)$ does produce a scaling of the magnitude of $Z_r(\omega)$, and to determine the significance of this scaling, the z domain equivalent of $Z_r(\omega)$, i.e. $Z_r(z)$, is examined. Neglecting any contribution from the pole of $Z_r(z)$ implies neglection of the denominator of $Z_r(z)$, and so Equation 6.22 becomes

$$Z_r(z) = b(1-z^{-1}) \tag{6.28}$$

For a piston radius, a, of 0.7 centimeters, the frequency response of $Z_r(z)$, as defined by Equation 6.28, is plotted with $Z_r(\omega)$ in Figure 6.3. Observation of Figure 6.3 shows a large difference in the magnitudes of $Z_r(z)$ and $Z_r(\omega)$, and so the magnitude term associated with the pole of $Z_r(\omega)$ must be taken into account.

For frequencies less than 4 kilohertz, it can be shown that the contribution to the magnitude of $Z_r(\omega)$ of the pole is approximately (1-p) (see Equation 6.22). Therefore, $Z_r(z)$ may be approximated by

$$Z_r(z) = d(1-z^{-1}) \tag{6.29a}$$

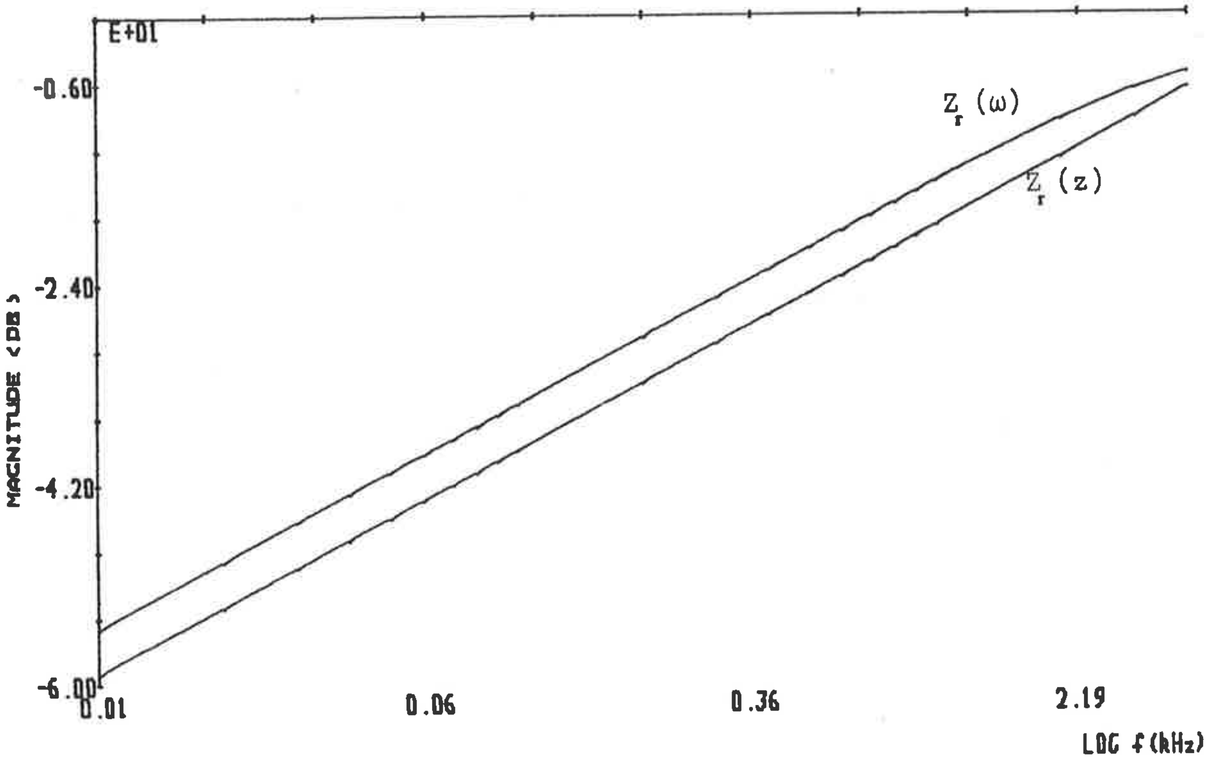
where

$$d = \frac{b}{1-p} \tag{6.29b}$$

which, by the use of Equation 5.23, may be rewritten as

$$d = \frac{8a}{3\pi Tc} \tag{6.29c}$$

The frequency responses of $Z_r(z)$ as defined by Equation 6.29, and $Z_r(\omega)$ are plotted in Figure 6.4 for values of piston radius, a, of 0.3, 0.7, 1.1 and 1.5 centimeters. Figure 6.4 shows that, in terms of magnitude, $Z_r(z)$ as defined by Equation 6.29 is an excel-



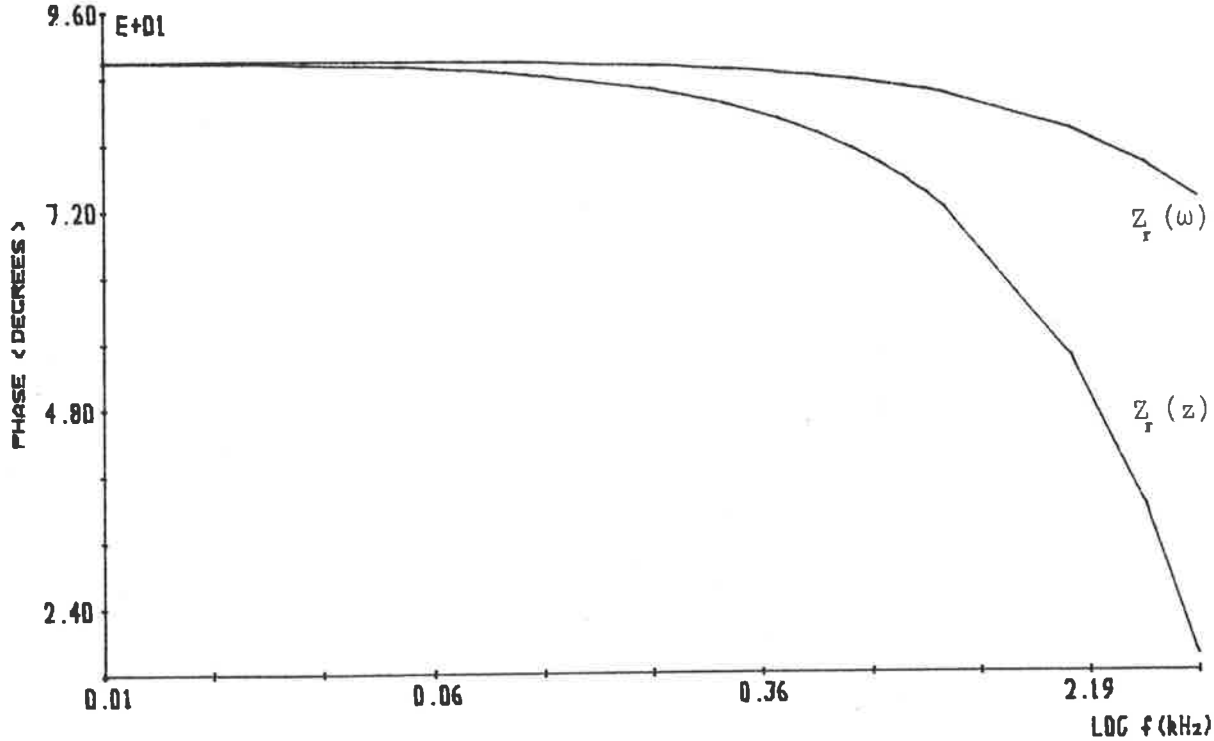


FIGURE 6.3: Comparison of frequency responses of $Z_r(\omega)$ and $Z_r(z)$ when the piston radius, a, is 0.7

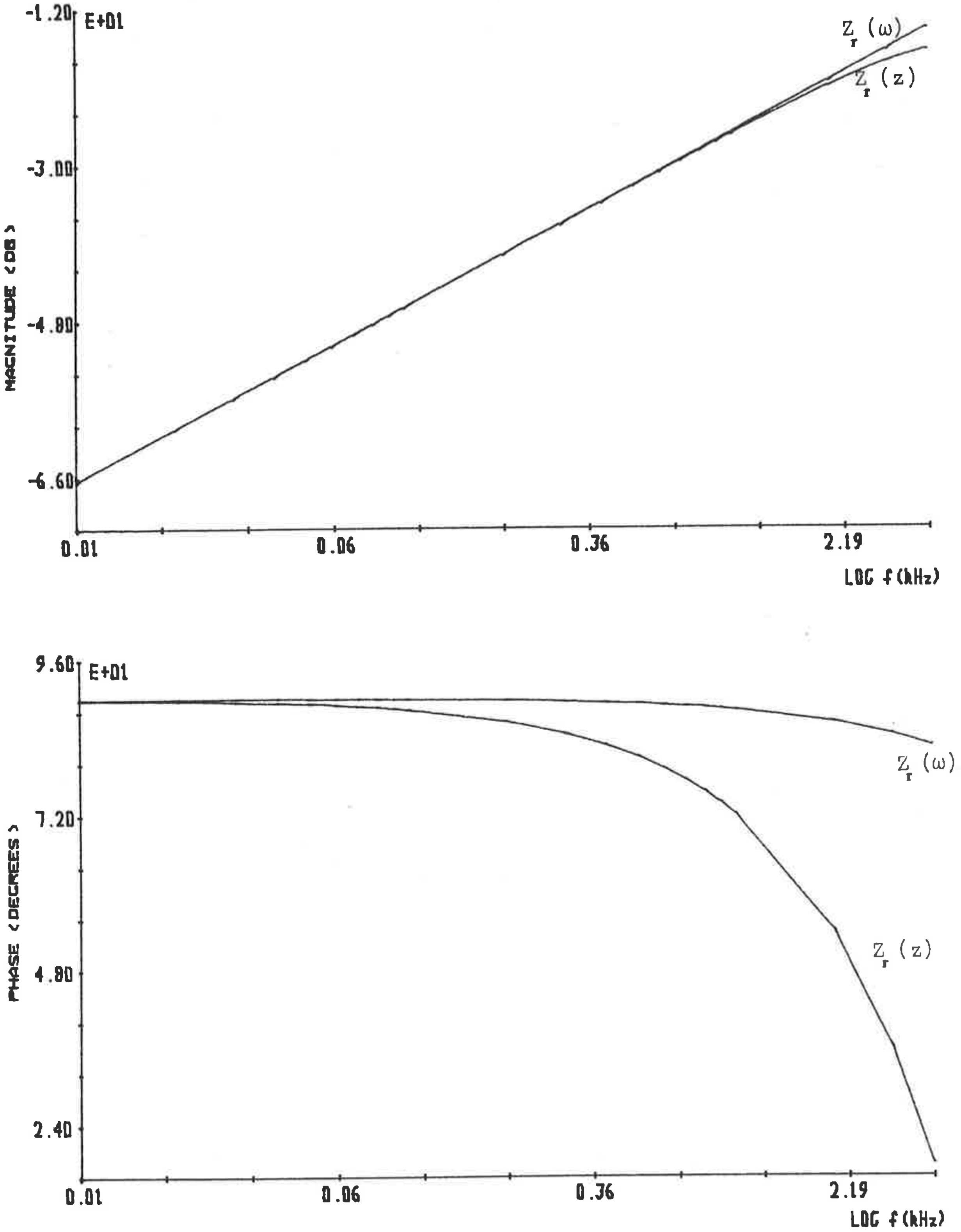
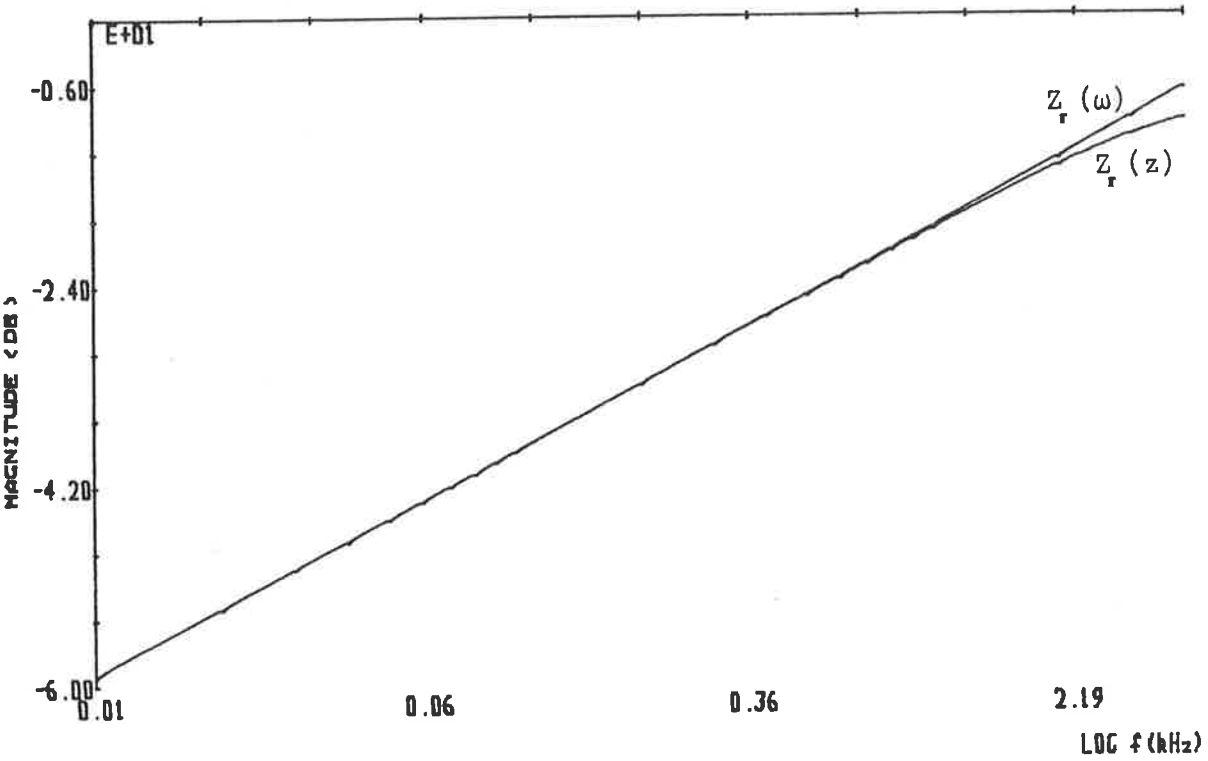


FIGURE 6.4(a): Comparison of frequency responses of $Z_r(\omega)$ and $Z_r(z)$ when the piston radius, a, is 0.3.



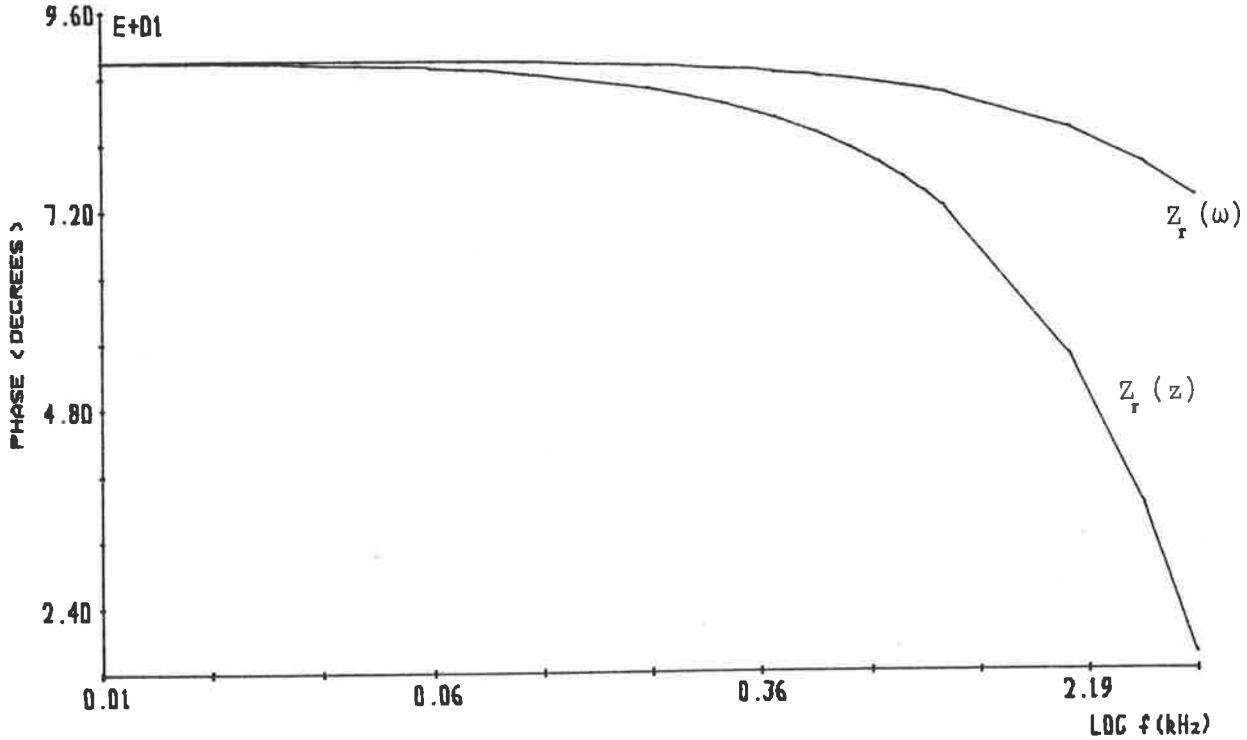
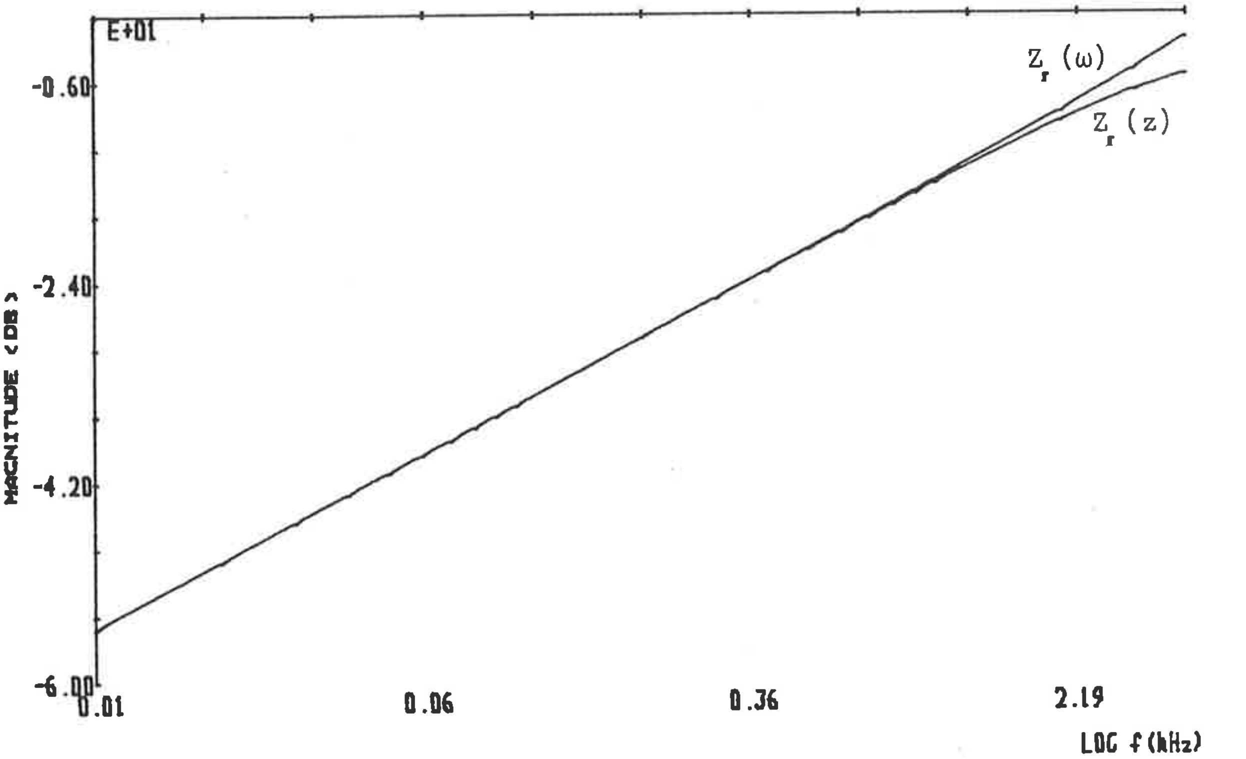


FIGURE 6.4(b): Comparison of frequency responses of $Z_r(\omega)$ and $Z_r(z)$ when the piston radius, a, is 0.7.



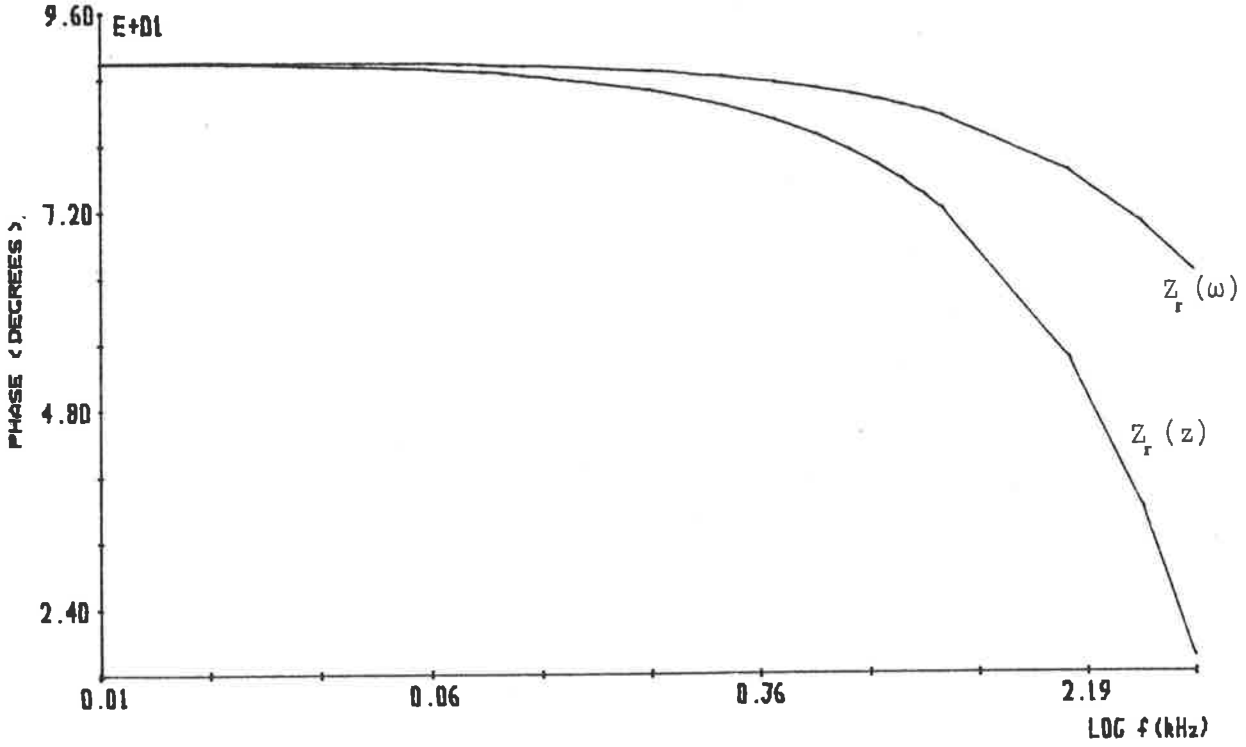
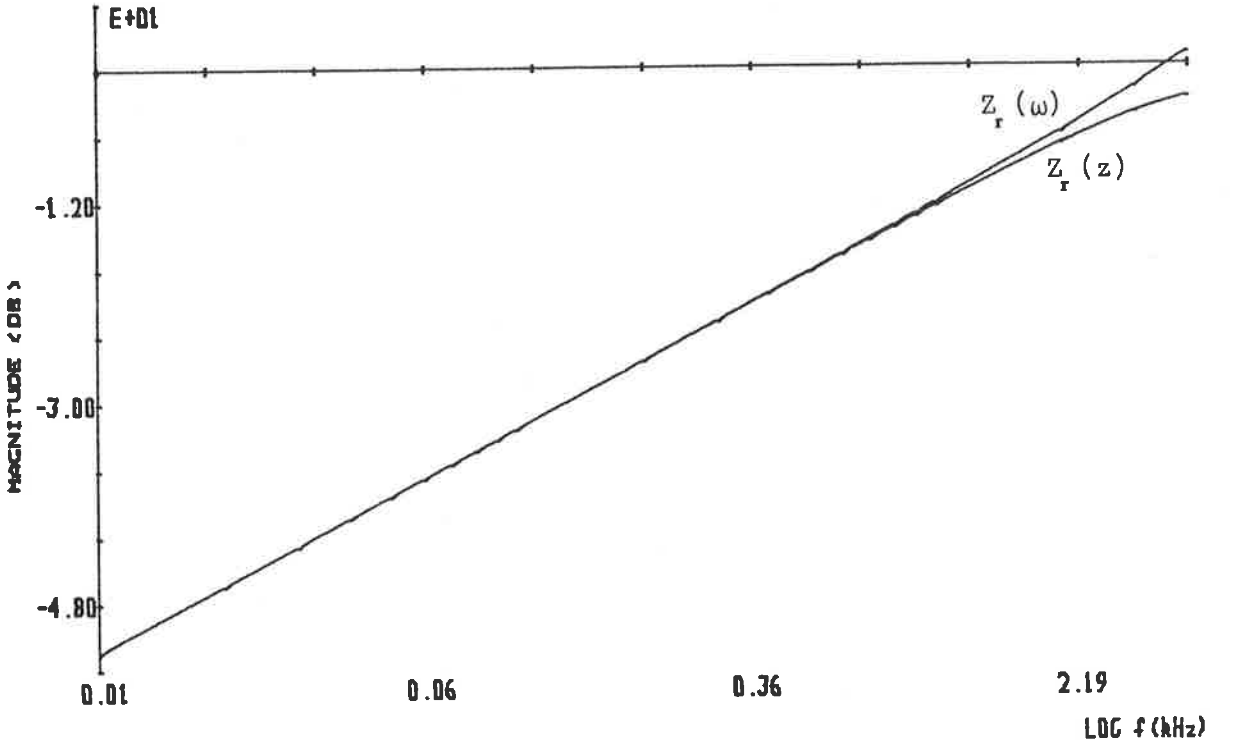


FIGURE 6.4(c): Comparison of frequency responses of $Z_r(\omega)$ and $Z_r(z)$ when the piston radius, a, is 1.1.



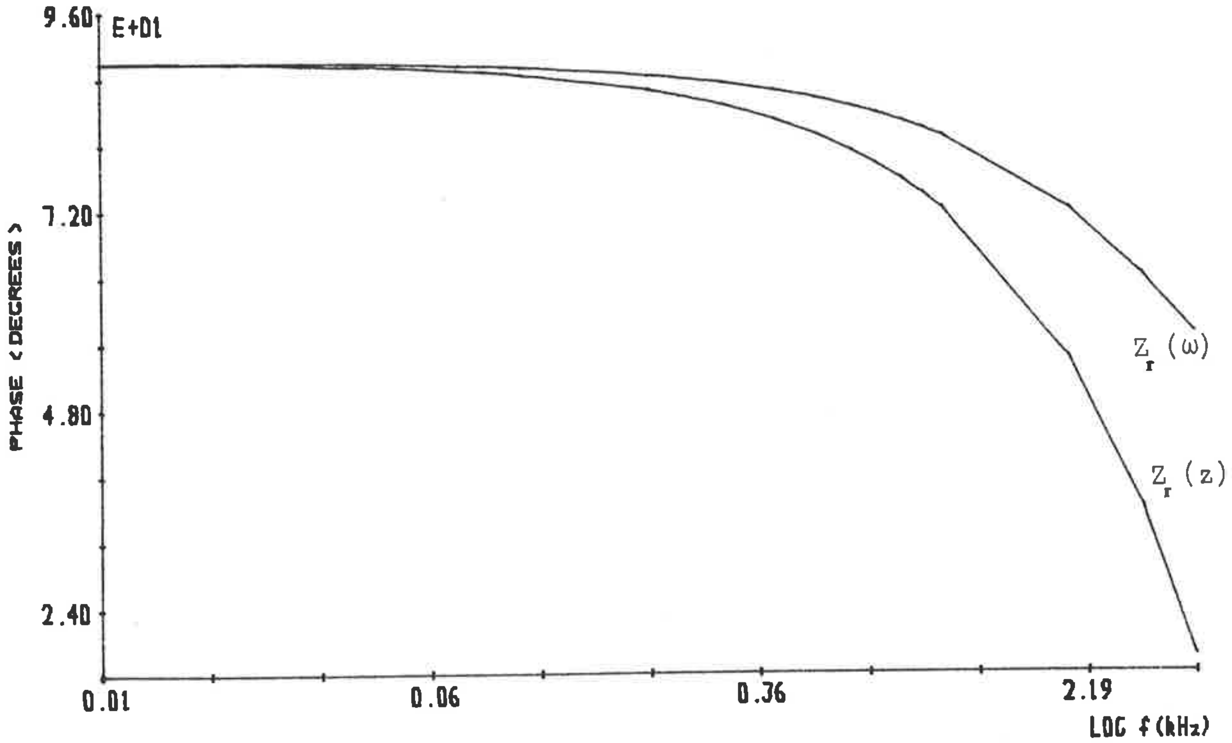


FIGURE 6.4(d): Comparison of frequency responses of $Z_r(\omega)$ and $Z_r(z)$ when the piston radius, a, is 1.5.

lent approximation to $Z_r(\omega)$ at low frequencies, and only a small error occurs at high frequencies. Therefore, except for the phase response, the $Z_r(z)$ as defined by Equation 6.29 is a good approximation to $Z_r(\omega)$. This approximation is performed by modelling with a single zero at $z = 1$ and a magnitude term which is dependent on the mouth or lip opening area.

The discrete time sampled equivalent of Equation 6.16 is determined using the known form of the radiation model in the z domain. Transforming Equation 6.1 into the time domain, using Equation 2.6 and 2.7, followed by a transformation into the z domain, produces

$$U_M(z) + V_M(z) = Z_r(z)[U_M(z) - V_M(z)] \tag{6.30}$$

and an application of Equation 6.29 produces

$$U_M(z)[(1-d) + dz^{-1}] = V_M(z)[-(1+d) + dz^{-1}] \tag{6.31}$$

Equation 6.31 is the termination equation of the acoustic tube model relating the forward and backward volume velocities at the termination.

In the z domain, the radiated volume velocity waveform is denoted by $V_r(z)$, and defined as

$$U_r(z) = U_M(z) - V_M(z) \tag{6.32}$$

and the radiated pressure waveform is denoted by $P_r(z)$, and defined as

$$P_r(z) = \frac{\rho c}{A_M}\left(U_M(z) + V_M(z)\right) \tag{6.33}$$

Therefore, in the z domain, Equation 6.1 is written as

$$P_r(z) = \frac{\rho c}{A_M} Z_r(z) U_r(z) \tag{6.34}$$

and using the approximation to the radiation impedance as described by Equation 6.29

$$P_r(z) = \frac{\rho c}{A_M} d(1 - z^{-1}) U_r(z) \tag{6.35}$$

The time domain equivalents of Equations 6.31, 6.32 and 6.35 are

$$u_M(n) = -\frac{d}{1-d} u_M(n-1) - \left(\frac{1+d}{1-d}\right) v_M(n) + \frac{d}{1-d} v_M(n-1) \tag{6.36}$$

$$u_r(n) = u_M(n) - v_M(n) \tag{6.37}$$

and

$$p_r(n) = \frac{\rho c d}{A_M}\left(u_r(n) - u_r(n-1)\right) \tag{6.38}$$

respectively. Figure 6.5 shows the signal flow diagram for the termination of the acoustic tube model when radiation is described by Equations 6.36, 6.37 and 6.38.

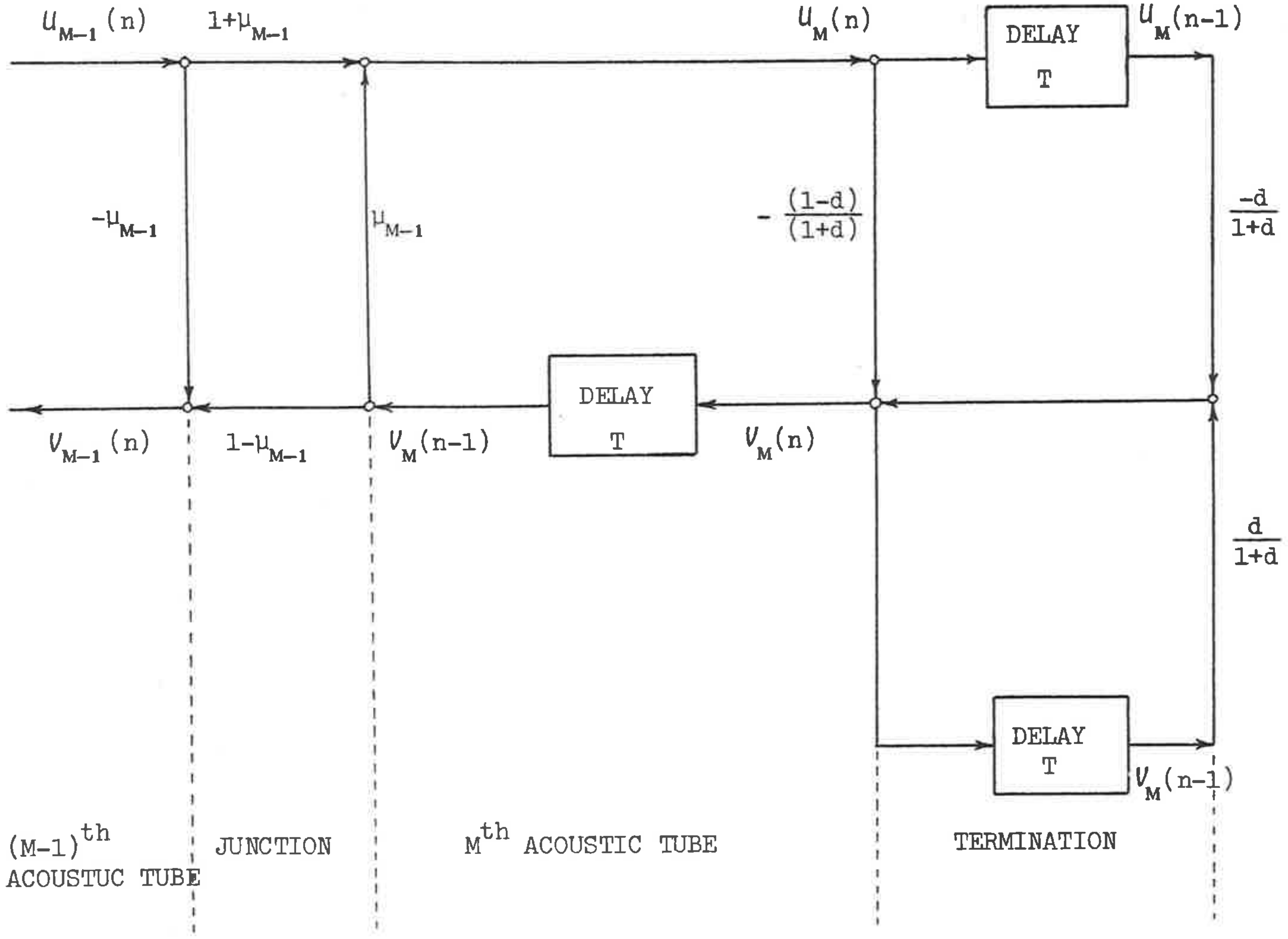


FIGURE 6.5: Signal flow diagram for the termination of the acoustic tube model by a radiation model.

In this section, a model of the radiation impedance presented to the termination of the human vocal tract, i.e. at the lips, has been derived. This model approximates the radiation from the lips by a piston in an infinite plane baffle. It was shown that, when the assumptions

a) the mouth or lip opening has an effective radius which is small compared with the radius of the head

b) the acoustic waveforms only contain frequencies which are less than 4 kilohertz

are satisfied, then an approximation to the radiation impedances is described by Equation 6.29. This approximation has the simple form of a single zero at $z = 1$ (or $j\omega = 0$) and a magnitude term which is dependent on the effective radius of the mouth opening.

From the realistic radiation model developed in this section, it is possible to evaluate the effectiveness of conventional procedures to overcome radiation effects. This evaluation is performed in the following section. A realistic radiation model also enables an analysis process to be developed which takes radiation into account, to provide improved acoustic tube/vocal tract shape recovery. The development of such an analysis process is presented later in this chapter.

6.3 THE LOSSY TERMINATION AND CONVENTIONAL ANALYSIS PROCEDURES

Accurate acoustic tube shape recovery from the output or radiated acoustic waveform of a set of acoustic tubes requires radiation to be taken into account. This section investigates the effectiveness of conventional pre-emphasis procedures in providing acoustic tube shape recovery when realistic radiation occurs.

Using real speech data to perform the evaluations in this section is not possible, as the glottal excitation, vocal tract losses and other real vocal tract properties cloud the results. Hence, all the data used for the evaluations presented in this section are synthetic and generated using the ideal acoustic tube model of Chapter 2 and the radiation model presented in Figure 6.5. The procedure for generating the synthetic data waveforms is presented in Appendix E, where the only violation of the ideal linear prediction/acoustic tube model assumptions is the radiation of an acoustic waveform, described by the radiation model presented in Figure 6.5.

The conventional approach to removing radiation effects from recovered acoustic tube shapes is to apply a -6 dB per octave pre-emphasis to the radiated pressure waveform before analysis. This conventional pre-emphasis is generally applied by using a digital filter, $H_r(z)$, which has a single pole at $z = 1$, i.e.

$$H_r(z) = \frac{1}{1 - z^{-1}} \qquad (6.39)$$

For the analog case, the pre-emphasis filter is an integrator. Conventionally, the glottal and radiation pre-emphasis filters are combined when pre-emphasising speech, and so the pre-emphasis by $H_r(z)$ is implicit in a conventional pre-emphasis of speech waveforms.

An indication of the effectiveness of pre-emphasizing by $H_r(z)$, to remove radiation effects, is obtained by applying the filter $H_r(z)$ directly to the radiation model of Section 6.2, i.e. multiplying Equation 6.35 by Equation 6.39 to produce

$$H_r(z)\ P_r(z) = \frac{\rho c}{A_M}\ dU_r(z) \tag{6.40}$$

Equation 6.40 shows that conventional pre-emphasis results in the radiated volume velocity, $U_r(z)$, being determined to within a normalization constant, $\frac{\rho c}{A_M}$, and the magnitude term, d, of the normalized radiation impedance, $Z_r(z)$. Hence, the zero of the radiation impedance is accounted for and, therefore, if the zero is the most significant part of $Z_r(z)$, then accurate acoustic tube shapes are recovered.

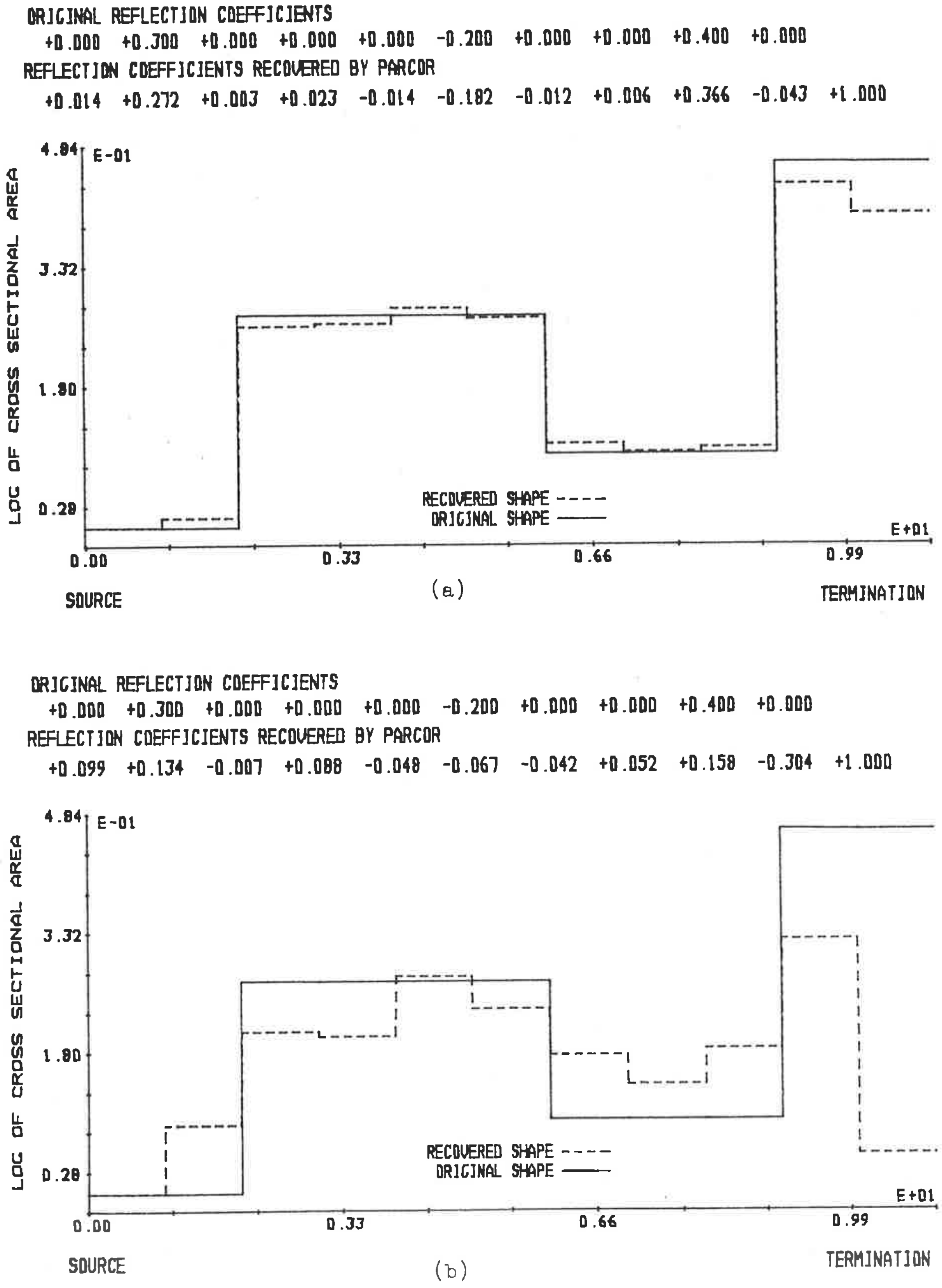
The effectiveness of conventional pre-emphasis in accounting for radiation effects on the recovered acoustic tube shape is firstly performed with synthetic radiated pressure waveforms generated as described in Appendix E, for two simple acoustic tube shapes. These two acoustic tube shapes differ only at the terminating acoustic tube where one has a relatively large cross-sectional area while the other has a relatively small cross-sectional area, and are referred to as open and closed terminating acoustic tubes, respectively. The analysis procedure used to generate evaluation results presented in this section firstly passes the synthetic radiated pressure waveform through $H_r(z)$, and then performs a Parcor analysis on the resultant volume velocity waveform, to produce a recovered acoustic tube shape.

The radiation impedance of each acoustic tube shape is varied by changing the piston radius, a, of the radiation model, defined in Section 6.2, from 0.2 to 2.0 centimeters. This range of piston radius exceeds that suggested by FLANAGAN [1972] (i.e. 0.5 to 1.3 centimeters) as the extremes for lip opening of typically phonated vowels. Table 6.1 contains the area distances of the recovered acoustic tube shapes for both the open and closed terminating acoustic tubes and the value of the magnitude term, d, of the radiation model, as a function of piston radius, a. The best and worst acoustic tube shapes recovered for the open and closed terminating acoustic tubes are presented in Figure 6.6 and 6.7, respectively.

The non-zero area distances in Table 6.1 imply that inaccurate acoustic tube shape recovery occurs when using conventional pre-emphasis to remove radiation effects from the recovered acoustic tube shape. Figures 6.6 and 6.7 show that good acoustic tube shape recovery occurs when the piston radius, a, is small. This is expected, since linear prediction ideally requires a piston radius, a, of zero. When the piston radius, a, is large, i.e. near 2.0 centimeters, then a significant difference is observed between the recovered and original acoustic tube shapes of Figures 6.6 and 6.7. Hence, the conventional -6 dB per octave pre-emphasis does not completely remove all the radiation effects from recovered acoustic tube shape.

VALUE OF PISTON RADIUS, a	VALUE OF RADIATION GAIN, d	OPEN TERMINATING ACOUSTIC TUBES	CLOSED TERMINATING ACOUSTIC TUBES
0.2	0.05	0.164	0.157
0.4	0.10	0.292	0.276
0.6	0.15	0.395	0.370
0.8	0.20	0.481	0.448
1.0	0.25	0.555	0.514
1.2	0.30	0.618	0.570
1.4	0.35	0.674	0.619
1.6	0.40	0.724	0.662
1.8	0.45	0.769	0.699
2.0	0.50	0.809	0.733

TABLE 6.1: Area distances for conventional linear predictive analysis of acoustic tubes with radiation at the termination.



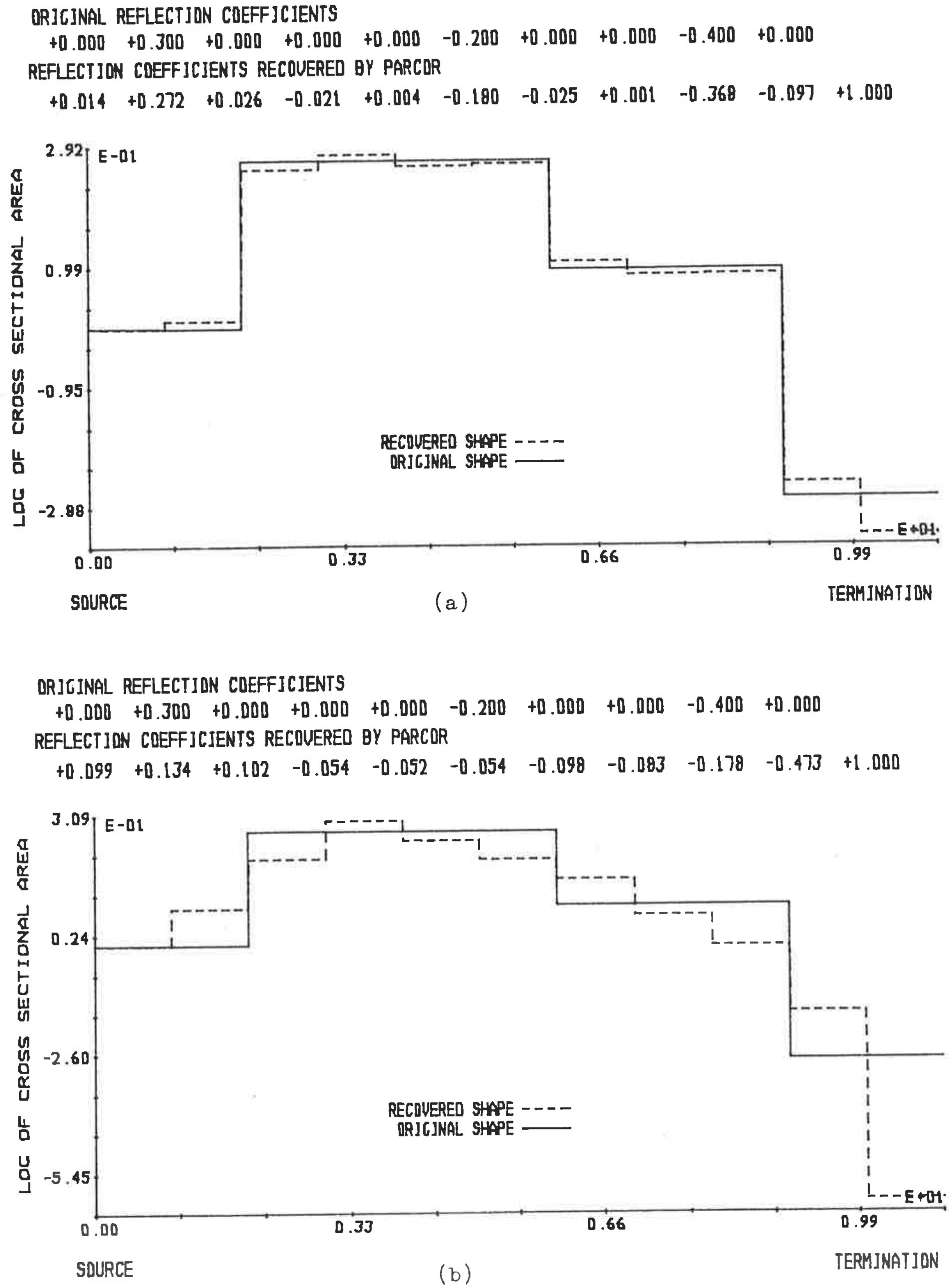
```
ORIGINAL REFLECTION COEFFICIENTS
  +0.000  +0.300  +0.000  +0.000  +0.000  -0.200  +0.000  +0.000  +0.400  +0.000
REFLECTION COEFFICIENTS RECOVERED BY PARCOR
  +0.014  +0.272  +0.003  +0.023  -0.014  -0.182  -0.012  +0.006  +0.366  -0.043  +1.000
```

(a)

```
ORIGINAL REFLECTION COEFFICIENTS
  +0.000  +0.300  +0.000  +0.000  +0.000  -0.200  +0.000  +0.000  +0.400  +0.000
REFLECTION COEFFICIENTS RECOVERED BY PARCOR
  +0.099  +0.134  -0.007  +0.088  -0.048  -0.067  -0.042  +0.052  +0.158  -0.304  +1.000
```

(b)

FIGURE 6.6: Comparison of recovered and original open termination acoustic tubes with radiation when the piston radius, a, is (a) 0.2 and (b) 2.0.



ORIGINAL REFLECTION COEFFICIENTS

+0.000 +0.300 +0.000 +0.000 +0.000 -0.200 +0.000 +0.000 -0.400 +0.000

REFLECTION COEFFICIENTS RECOVERED BY PARCOR

+0.014 +0.272 +0.026 -0.021 +0.004 -0.180 -0.025 +0.001 -0.368 -0.097 +1.000

(a)

ORIGINAL REFLECTION COEFFICIENTS

+0.000 +0.300 +0.000 +0.000 +0.000 -0.200 +0.000 +0.000 -0.400 +0.000

REFLECTION COEFFICIENTS RECOVERED BY PARCOR

+0.099 +0.134 +0.102 -0.054 -0.052 -0.054 -0.098 -0.083 -0.178 -0.473 +1.000

(b)

FIGURE 6.7: Comparison of recovered and original closed termination acoustic tubes with radiation when the piston radius, a, is (a) 0.2 and (b) 2.0

Two acoustic tube shapes are used in the above evaluations, in order to determine the effects of radiation on the recovered acoustic tube shape for open and closed terminating acoustic tubes. Table 6.1 indicates that approximately the same area distances occur for each shape, and the comparison of original and recovered acoustic tube shapes of Figures 6.6 and 6.7 suggests that similar errors are made in the acoustic tube shape recovered for both cases. Hence, the errors in acoustic tube shape recovery, even when conventional pre-emphasis is used, appear to be independent of the absolute value of the cross-sectional area of the terminating acoustic tube.

Further evaluation of the effectiveness of conventional pre-emphasis to account for radiation effects on acoustic tube shape recovery is performed with more complex acoustic tube shapes which approximate real vocal tract shapes. These acoustic tube shapes approximate five Russian vowel shapes as measured by FANT [1960] (see Appendix C), and are used to generate radiated pressure waveforms as described in Appendix E. Analysis of the radiated pressure waveforms is performed by firstly passing the data through the pre-emphasis filter, $H_r(z)$, and then performing a Parcor analysis on the resultant volume velocity waveform, to produce a recovered acoustic tube shape.

As with the two simple acoustic tubes used earlier, the radiation impedance for the five vowel shapes is varied by changing the piston radius, a, of the radiation model from 0.2 to 2.0 centimeters. Table 6.2 presents the area distances between the original and recovered acoustic tube shapes as a function of piston radius, a. The original and recovered acoustic tubes shapes for

PISTON RADIUS, a	VOWEL				
	/a/	/e/	/i/	/o/	/u/
0.2	0.166	0.280	0.584	0.274	0.247
0.4	0.303	0.468	0.894	0.469	0.465
0.6	0.419	0.610	1.089	0.628	0.664
0.8	0.522	0.724	1.222	0.764	0.845
1.0	0.615	0.818	1.319	0.889	1.019
1.2	0.701	0.898	1.392	1.005	1.180
1.4	0.780	0.967	1.451	1.114	1.331
1.6	0.855	1.027	1.503	1.218	1.474
1.8	0.925	1.080	1.549	1.318	1.608
2.0	0.992	1.128	1.594	1.415	1.737

TABLE 6.2: Area distances for conventional linear predictive analysis of synthetic speech for five vowels and a range of piston radius, a, of the radiation model.

piston radii of 0.2 and 2.0 centimeters (i.e. the best and worst recovered acoustic tube shapes) are presented in Figures 6.8 to 6.12 for each of the five vowel shapes.

Observation of Table 6.2 reveals a small change in the area distances for the same piston radius, a, from one vowel shape to another. Thus, similar area distances are found for recovering acoustic tube shapes for open terminating vowel shapes such as |a| as for closed terminating vowel shapes such as |u|. Hence, the conclusion made earlier that the radiation effects are similar regardless of the absolute cross-sectional area of the terminating acoustic tube is reinforced by the results of Table 6.2.

The non-zero area distances of Table 6.1 imply that inaccurate tube shape recovery occurs when using the conventional -6 dB per octave pre-emphasis to account for radiation from the termination of acoustic tubes. The recovered acoustic tube shapes displayed in Figures 6.8 to 6.12 for a piston radius, a, of 0.2 centimeters show that good acoustic tube shape recovery is achieved when the piston radius, a, is small. This agrees with the previous results for the simple acoustic tube shapes.

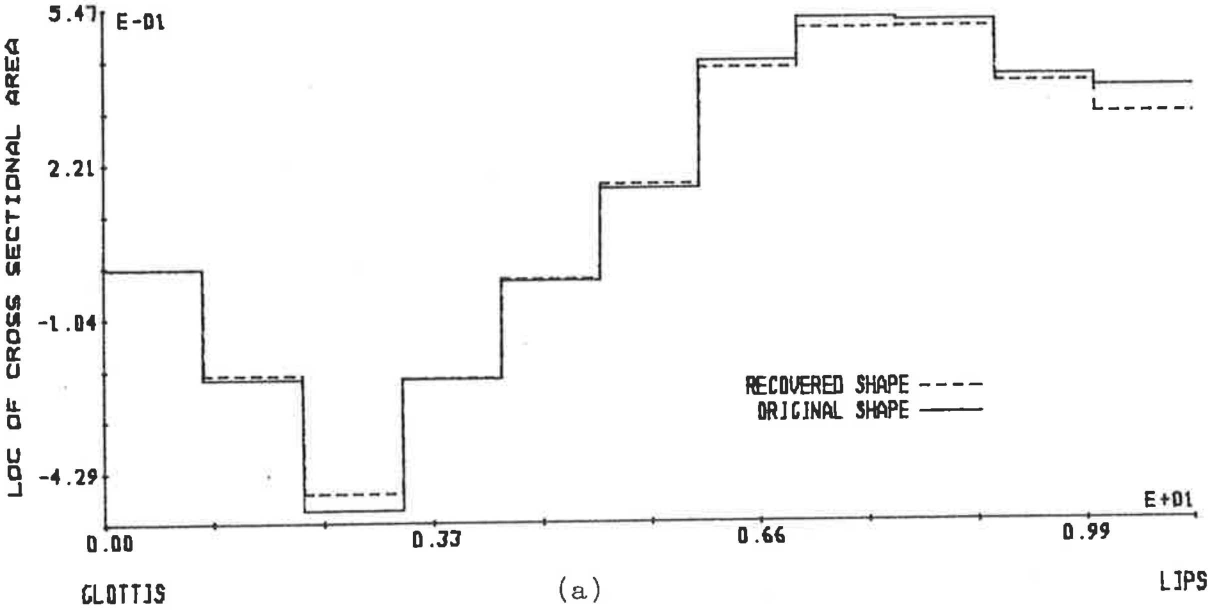
For the case where the piston radius, a, of the radiation model is 2.0 centimeters, a variety of reasonable and poor acoustic tube shape recoveries is observed from Figures 6.8 to 6.12. The recovered acoustic tube shapes for the vowels |a|, |o| and |u| have a close similarity to the original acoustic tube shape. In general, it is observed that the major departures between original and recovered acoustic tube shapes occur at the termination end of all the recovered acoustic tube shapes.

ORIGINAL REFLECTION COEFFICIENTS

-0.262 -0.310 +0.310 +0.232 +0.217 +0.298 +0.100 -0.006 -0.131 -0.026

REFLECTION COEFFICIENTS RECOVERED BY PARCOR

-0.252 -0.281 +0.273 +0.235 +0.222 +0.274 +0.093 +0.002 -0.132 -0.075 +1.000



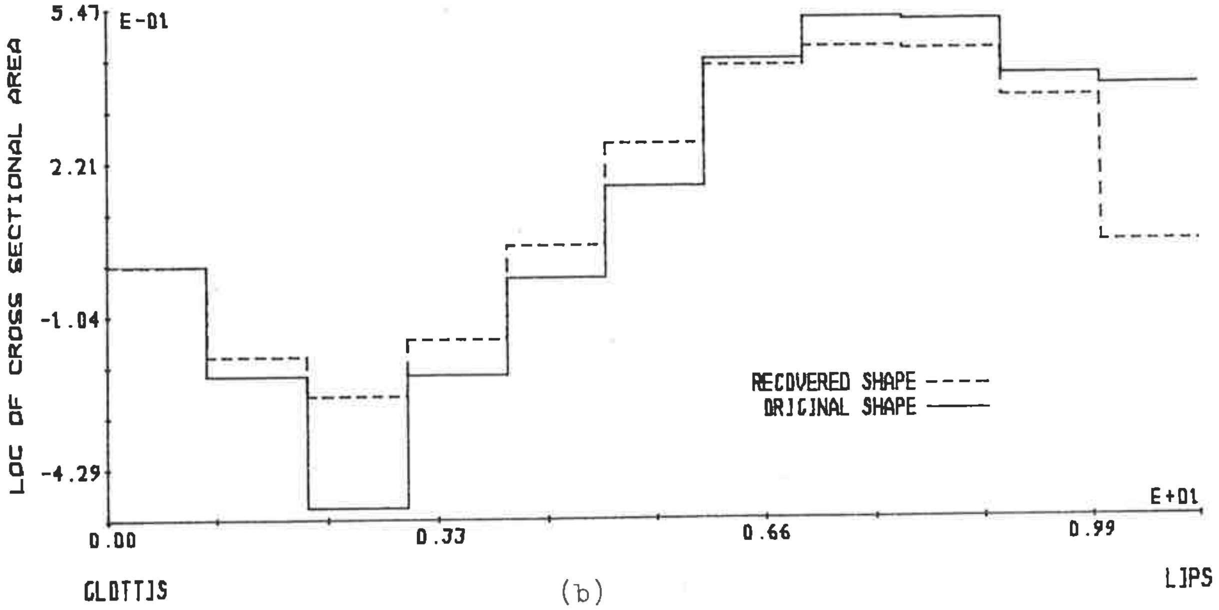
(a)

ORIGINAL REFLECTION COEFFICIENTS

-0.262 -0.310 +0.310 +0.232 +0.217 +0.298 +0.100 -0.006 -0.131 -0.026

REFLECTION COEFFICIENTS RECOVERED BY PARCOR

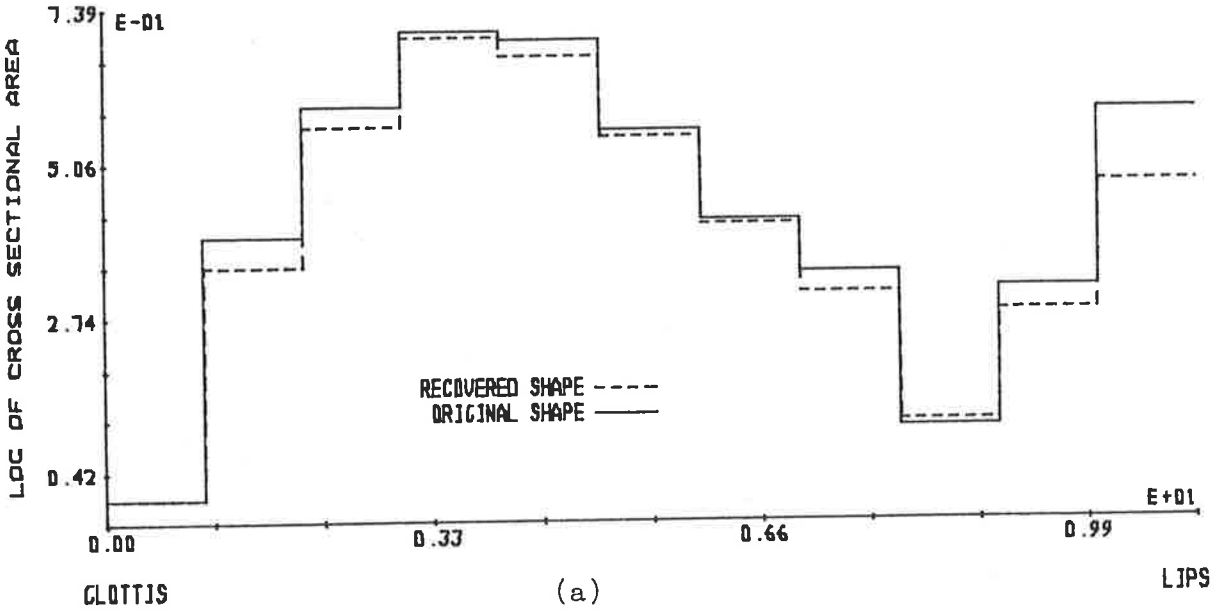
-0.217 -0.096 +0.137 +0.224 +0.243 +0.185 +0.044 -0.007 -0.115 -0.340 +1.000



(b)

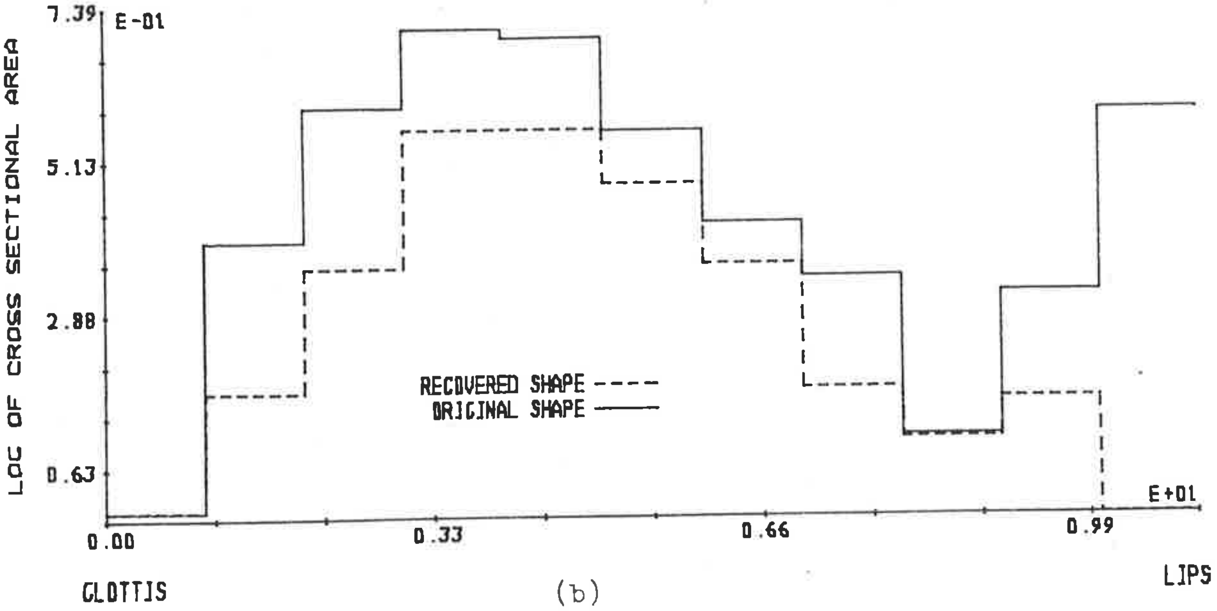
FIGURE 6.8: Comparison of recovered and original acoustic tube shapes for a linear predictive analysis of synthetic speech for the vowel /a/ when the radiation model has a piston radius, a, of (a) 0.2 and (b) 2.0.

```
ORIGINAL REFLECTION COEFFICIENTS
  +0.425  +0.221  +0.130  -0.014  -0.156  -0.154  -0.091  -0.262  +0.235  +0.297
REFLECTION COEFFICIENTS RECOVERED BY PARCOR
  +0.383  +0.236  +0.153  -0.032  -0.139  -0.150  -0.119  -0.219  +0.187  +0.216  +1.000
```



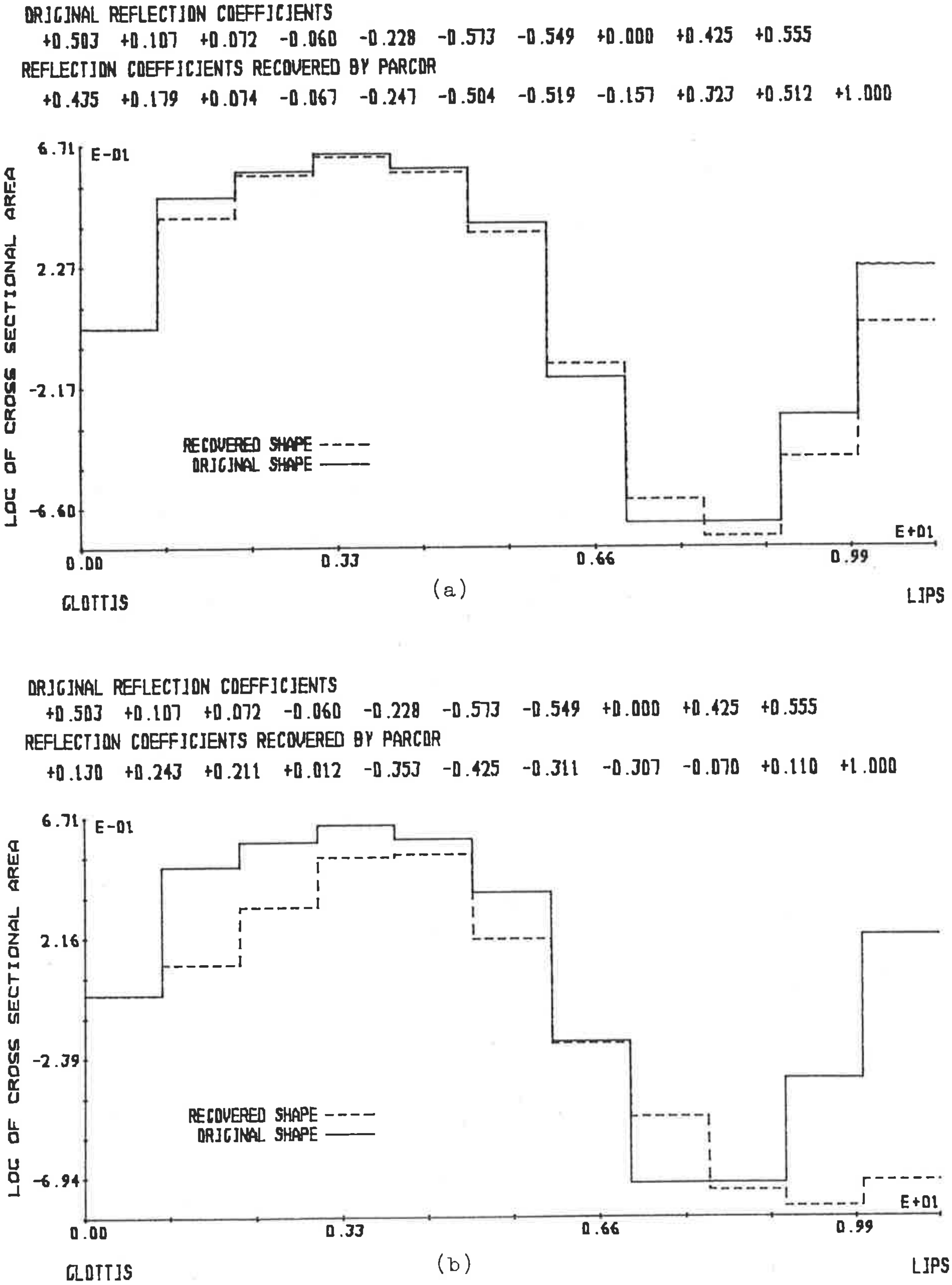
(a)

```
ORIGINAL REFLECTION COEFFICIENTS
  +0.425  +0.221  +0.130  -0.014  -0.156  -0.154  -0.091  -0.262  +0.235  +0.297
REFLECTION COEFFICIENTS RECOVERED BY PARCOR
  +0.197  +0.205  +0.228  +0.000  -0.091  -0.135  -0.206  -0.085  +0.067  -0.194  +1.000
```



(b)

FIGURE 6.9: Comparison of recovered and original acoustic tube shapes for a linear predictive analysis of synthetic speech for the vowel /e/ when the radiation model has a piston radius, a, of (a) 0.2 and (b) 2.0



ORIGINAL REFLECTION COEFFICIENTS
+0.503 +0.107 +0.072 -0.060 -0.228 -0.573 -0.549 +0.000 +0.425 +0.555
REFLECTION COEFFICIENTS RECOVERED BY PARCOR
+0.435 +0.179 +0.074 -0.067 -0.247 -0.504 -0.519 -0.157 +0.323 +0.512 +1.000

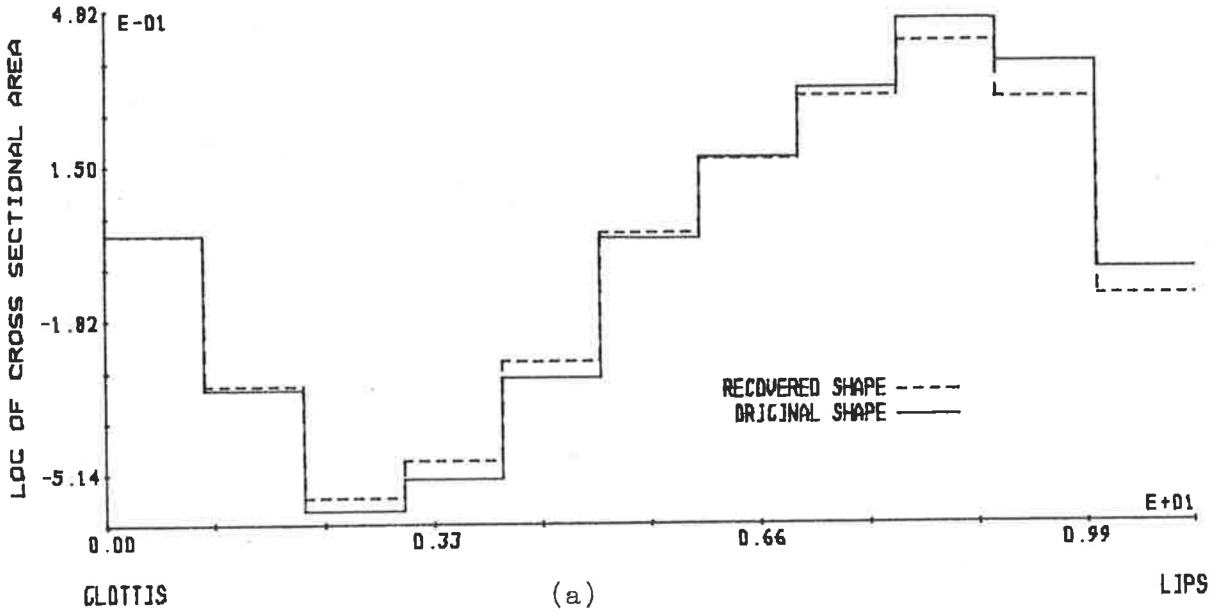
(a)

ORIGINAL REFLECTION COEFFICIENTS
+0.503 +0.107 +0.072 -0.060 -0.228 -0.573 -0.549 +0.000 +0.425 +0.555
REFLECTION COEFFICIENTS RECOVERED BY PARCOR
+0.130 +0.243 +0.211 +0.012 -0.353 -0.425 -0.311 -0.307 -0.070 +0.110 +1.000

(b)

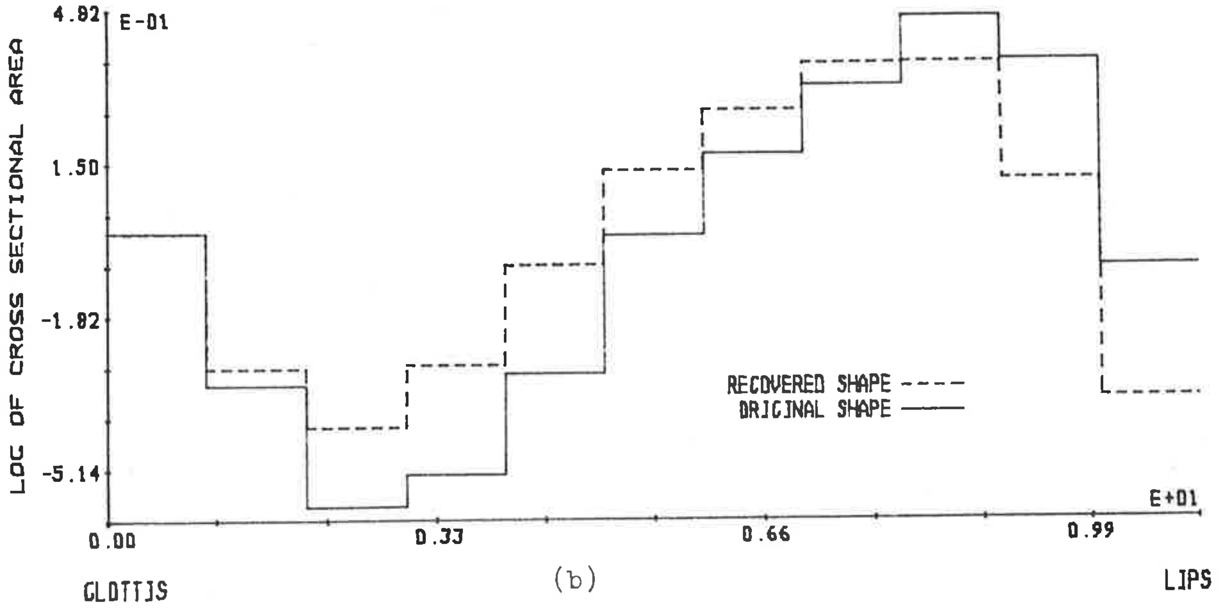
FIGURE 6.10: Comparison of recovered and original acoustic tube shapes for a linear predictive analysis of synthetic speech for the vowel /i/ when the piston radius, a, is (a) 0.2 and (b) 2.0

```
ORIGINAL REFLECTION COEFFICIENTS
  -0.364  -0.293  +0.078  +0.245  +0.331  +0.197  +0.168  +0.167  -0.105  -0.471
REFLECTION COEFFICIENTS RECOVERED BY PARCOR
  -0.358  -0.269  +0.091  +0.241  +0.307  +0.181  +0.152  +0.133  -0.140  -0.454  +1.000
```



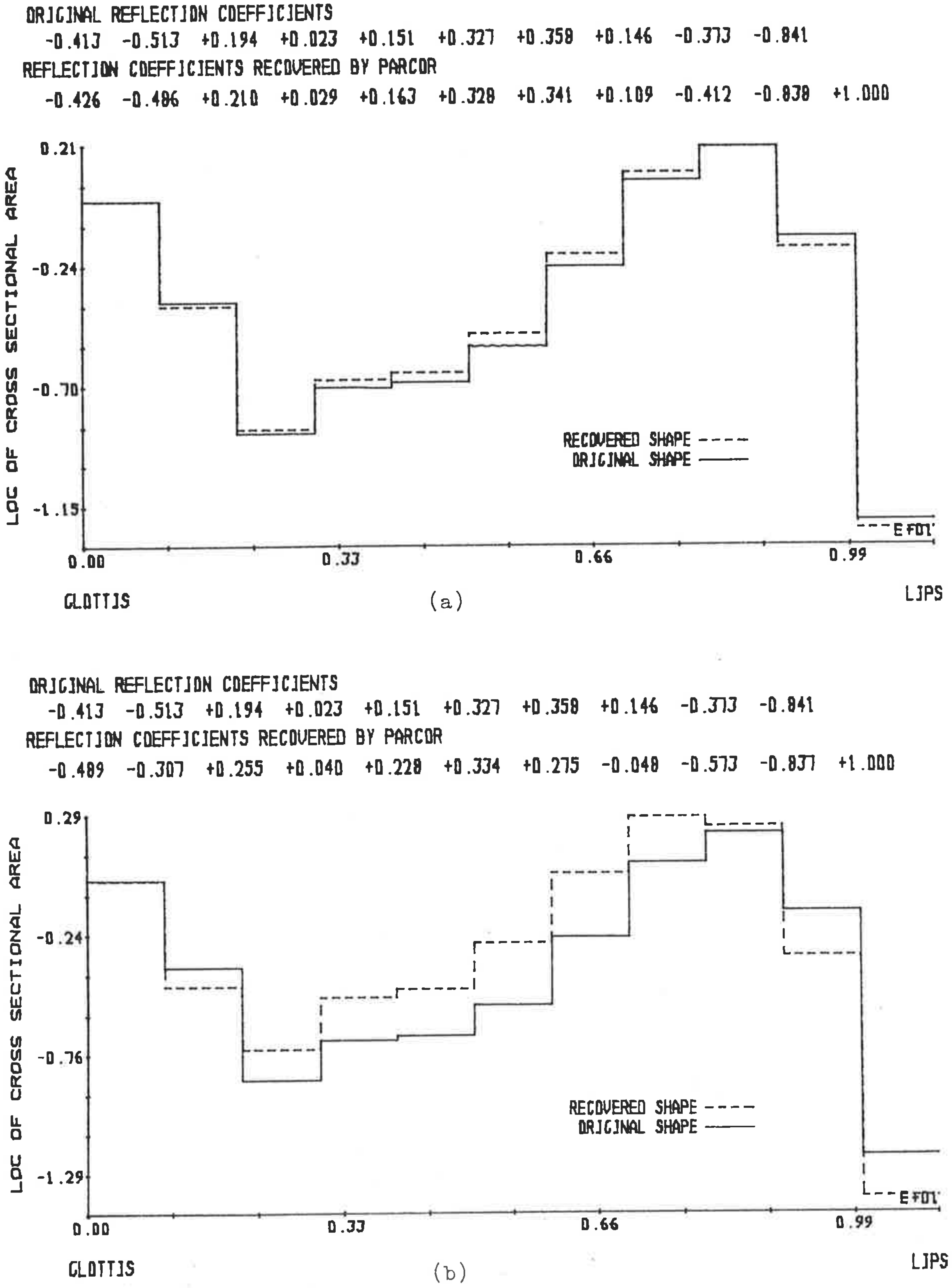
(a)

```
ORIGINAL REFLECTION COEFFICIENTS
  -0.364  -0.293  +0.078  +0.245  +0.331  +0.197  +0.168  +0.167  -0.105  -0.471
REFLECTION COEFFICIENTS RECOVERED BY PARCOR
  -0.328  -0.144  +0.154  +0.242  +0.229  +0.148  +0.115  +0.001  -0.282  -0.495  +1.000
```



(b)

FIGURE 6.11: Comparison of original and recovered acoustic tube shapes for a linear predictive analysis of synthetic speech for the vowel /o/ when the radiation model has a piston radius, a, of (a) 0.2 and (b) 2.0



ORIGINAL REFLECTION COEFFICIENTS

-0.413 -0.513 +0.194 +0.023 +0.151 +0.327 +0.358 +0.146 -0.373 -0.841

REFLECTION COEFFICIENTS RECOVERED BY PARCOR

-0.426 -0.486 +0.210 +0.029 +0.163 +0.328 +0.341 +0.109 -0.412 -0.838 +1.000

(a)

ORIGINAL REFLECTION COEFFICIENTS

-0.413 -0.513 +0.194 +0.023 +0.151 +0.327 +0.358 +0.146 -0.373 -0.841

REFLECTION COEFFICIENTS RECOVERED BY PARCOR

-0.489 -0.307 +0.255 +0.040 +0.228 +0.334 +0.275 -0.048 -0.573 -0.837 +1.000

(b)

FIGURE 6.12: Comparison of original and recovered acoustic tube shapes for a linear predictive analysis of synthetic speech for the vowel /u/ when the radiation model has a piston radius, a, of (a) 0.2 and (b) 2.0

A comparison of the relative effects of glottal pulse excitation and radiation on the recovered acoustic tube shape is possible by comparing the results of this section with those of Section 4.3. The area distances resulting from a Parcor analysis of the synthetic speech with radiation present vary from 5 to 170 when the piston radius, a, is 2 centimeters. For synthetic speech generated with glottal pulse excitation and no radiation, and analysed by a Parcor analysis, Section 4.3 showed that area distances between 3 and 250 resulted. Therefore, glottal pulse excitation and lip radiation can cause similar effects on the recovered acoustic tube shape.

This section has shown that, when a conventional -6 dB per octave pre-emphasis is used on synthetic speech with radiation included in the generation process, the resultant area distances vary between approximately zero and unity. For a conventional +12 dB per octave pre-emphasis of synthetic speech with glottal pulse excitations, Section 4.3 shows that area distances between approximately zero and nine occur. Hence, it appears that the conventional pre-emphasis of -6 dB per octave is more effective at removing radiation effects than a conventional +12 dB per octave pre-emphasis is at removing glottal pulse excitation effects from the recovered acoustic tube shape.

The investigations performed in this section show that a conventional -6 dB per octave pre-emphasis provides, in some cases, acceptable area function recovery when radiation from the termination of the acoustic tubes is present. In situations where the piston radius, a, of the radiation model defined in Section 6.2 is small, good acoustic tube shape recovery is achieved by using a conventional -6 dB per octave pre-emphasis. However, for large piston radii, significant departures from the original acoustic tube shape are observed such that the recovered acoustic tube shape may not resemble the original acoustic tube shape. It was found that the most significant departures of the recovered acoustic tube shape from the original acoustic tube shape occur near the terminating rather than the source end of the recovered acoustic tube shape.

Even though the conventional -6 dB per octave pre-emphasis was shown to be more effective at removing radiation effects than the conventional +12 dB per octave pre-emphasis is at removing glottal pulse excitation effects, accurate acoustic tube shape recovery is not achieved. It was shown that the conventional pre-emphasis of -6 dB per octave completely accounts for the zero of the radiation model as defined in Section 6.2. Therefore, the remaining part of the radiation model, in particular the magnitude term, d, produces significant errors in acoustic tube shape recovery.

6.4 REQUIREMENTS FOR IMPROVED ANALYSIS OF A LOSSY TERMINATION

A model for the radiation from an open-ended acoustic tube was defined in Section 6.2. The radiation model is described by a radiation impedance with a single zero at $z = 1$ (or $j\omega = 0$), and a magnitude term, d. The effectiveness of a conventional -6 dB per octave pre-emphasis to account for radiation effects on the recovered acoustic tube shape was investigated in Section 6.3, and it was shown that the conventional -6 dB per octave completely removes the effects of the zero of the radiation impedance. However, inaccurate acoustic tube shape recovery still occurs, because the conventional -6 dB per octave pre-emphasis completely ignores the magnitude term, d, of the radiation impedance. This section considers the manner in which the magnitude term, d, of the radiation model may be included in the acoustic tube model via an acoustic tube termination equation.

A conventional -6 dB per octave pre-emphasis of the radiated pressure waveform produces the radiated volume velocity scaled by the constant $\frac{\rho c}{A_M} \cdot d$ (see Equation 6.40). It is well-known (see Chapter 2) that a scaling of the waveform before a linear predictive analysis does not affect the results of the analysis and, therefore, the constant $\frac{\rho c}{A_M} \cdot d$ may be ignored. However, a linear predictive analysis requires knowledge of the forward and backward travelling volume velocity waveforms, $U_M(n)$ and $V_M(n)$, respectively, which cannot be determined accurately from the radiated volume velocity if the magnitude term of the radiation impedance is ignored.

The relationship between the radiated pressure and volume velocity is defined by the radiation model of Section 6.2 as

$$P_r(z) = \frac{\rho c}{A_M} d(1 - z^{-1})U_r(z) \tag{6.41}$$

Application of a -6 dB per octave pre-emphasis, via the filter $H_r(z)$ (as defined by Equation 6.39) removes the zero in the right-hand side of Equation 6.41, i.e.

$$H_r(z)P_r(z) = \frac{\rho c}{A_M} dU_r(z) \tag{6.42}$$

Equations 6.32 and 6.33 enable Equation 6.42 to be rewritten as

$$[U_M(z) + V_M(z)] = d[U_M(z) - V_M(z)] \tag{6.43}$$

which, upon rearrangement, becomes

$$V_M(z) = -\left(\frac{1-d}{1+d}\right)U_M(z) \tag{6.44}$$

The relationship described by Equation 6.44 enables the radiated volume velocity to be described in terms of $U_M(z)$ as

$$U_r(z) = \frac{2d}{1+d} U_M(z) \tag{6.45}$$

To perform a linear predictive analysis, discrete time sampled waveform equations are required; therefore, a transformation of the above z domain equations into the discrete domain is necessary. In the discrete time domain, Equation 6.42 is

$$P_r(n) = \frac{\rho c}{A_M} dU_r(n) \tag{6.46}$$

Equation 6.44 is

$$V_M(n) = -\left(\frac{1-d}{1+d}\right)U_M(n) \tag{6.47}$$

and Equation 6.45 is

$$U_r(n) = \frac{2d}{1+d} U_M(n) \tag{6.48}$$

Analysis procedures such as linear prediction are unaffected by a scaling of the waveform before analysis. Since the magnitude term, d, of the radiation model is a constant for a particular acoustic tube shape, the scaling terms $\frac{\rho c}{A_M} d$ in Equation 6.46 and $\frac{2d}{1+d}$ in Equation 6.48 need not be considered in the analysis process. Hence, $P_r(n)$, $U_r(n)$ and $U_M(n)$ can all be equated, i.e.

$$P_r(n) = U_r(n) = U_r(n) \tag{6.49}$$

without causing errors in the linear predictive analysis. Hence, following a conventional -6 dB per octave pre-emphasis, as applied by $H_r(z)$, the resultant waveform may be equated to the forward travelling volume velocity at the termination of the acoustic tube model, i.e. $U_M(n)$.

The backward travelling volume velocity at the termination of the acoustic tube model, i.e. $V_M(n)$, must be known for a linear predictive analysis to be performed. Once $U_M(n)$ and the magnitude term, d, of the radiation impedance are known, then $V_M(n)$ is found from Equation 6.47. The terminating equation for a lattice structure has the form

$$V_M(n) = -\mu_M U_M(n) \tag{6.50}$$

where μ_M is the termination reflection coefficient. On comparison of Equations 6.47 and 6.50, μ_M may be defined in terms of d as

$$\mu_M = \frac{1-d}{1+d} \tag{6.51}$$

After the radiated pressure is pre-emphasized by $H_r(z)$, the load impedance, Z_L, at the termination of the acoustic tubes is defined by

$$Z_L = \frac{\rho c}{A_M} d \tag{6.52}$$

Therefore, Equation 6.51 may be rewritten as

$$\mu_M = \frac{\frac{\rho c}{A_M} - Z_L}{\frac{\rho c}{A_M} + Z_L} \tag{6.53}$$

which is the familiar form for the termination reflection coefficient, μ_M (e.g. see RABINER and SCHAFER [1978]).

If the magnitude term, d, of the radiation impedance is zero, the termination reflection coefficient, μ_M, is unity (from Equation 6.51). Therefore, the waveform incident on the termination of acoustic tubes, i.e. $U_M(n)$, is reflected back into the acoustic tubes as $V_M(n)$ with only a phase reversal. Hence the termination is lossless, and satisfies the ideal assumptions of the linear prediction acoustic tube model. For a non-zero magnitude term, d, the modulus of μ_M is less than unity, so that a portion of $U_M(n)$ is reflected back into the acoustic tubes as $V_M(n)$, and a portion of $U_M(n)$ is lost at the termination. The loss at the termination of the acoustic tubes is in the form of finite radiated volume velocity and pressure waveforms which, from Equation 6.46, 6.48 and 6.51, are

$$U_r(n) = (1-\mu_M)U_M(n) \tag{6.54}$$

and

$$P_r(n) = \frac{\rho c}{A_M} \cdot \frac{(1-\mu_M)^2}{(1+\mu_M)} U_M(n) \qquad (6.55)$$

respectively.

In the previous section, it was established that for accurate acoustic tube shape recovery when radiation has occurred, the magnitude term, d, of the radiation model must be included in the analysis process. This section has shown that a non-zero magnitude term, d, results in a non-unity and, therefore, non-ideal termination reflection coefficient, μ_M. Conventional analysis procedures such as linear prediction do not provide accurate acoustic tube shape recovery when the termination reflection coefficient has a modulus of less than unity. Therefore, when radiation is present, there exists a need for an analysis procedure which can provide accurate acoustic tube shape recovery for a non-unity termination reflection coefficient. The following sections work towards the development of such a procedure.

6.5 LOSSY TERMINATION ANALYSIS VIA AUTOCORRELATIONS

To remove the effects of radiation from the recovered acoustic tube shape, it has been shown that, in the first instance, a conventional -6 dB per octave pre-emphasis needs to be applied to the radiated waveform. The pre-emphasized waveform must then be analysed by a procedure which permits a non-unity termination reflection coefficient, μ_M. Conventional linear predictive and other analysis procedures require a unity termination reflection coefficient and, therefore, a new analysis procedure is necessary. This section investigates the development of a new analysis procedure for situations where the termination reflection coefficient, μ_M, is not unity.

The development of a new analysis procedure which is suitable for real time applications must consider simplicity and computational efficiency. A procedure based on linear predictive lattice type analyses has the advantage of simplicity and ease of implementation on digital computers. Conventionally, linear predictive analyses are performed via waveform, autocorrelation or covariance approaches. The waveform formulation is computationally inefficient when compared with either the autocorrelation or covariance formulation, and the autocorrelation formulation requires less computations than the covariance formulation. Therefore, the analysis procedure developed in this section is based on the autocorrelation implementation of the lattice formulation of linear prediction.

6.5.1 DERIVATION OF AN ANALYSIS PROCESS

This section develops the basis for a new analysis procedure which permits a termination reflection coefficient, μ_M, to have a modulus of less than unity. The new analysis procedure is based on the simple and computationally efficient autocorrelation implementation of the lattice formulation of linear prediction (see Chapter 2).

The forward autocorrelation function of the forward travelling volume velocity in the ith acoustic tube is defined as

$$A_i(r) = \sum_{n=-\infty}^{\infty} U_i(n)U_i(n-r) \tag{6.56}$$

and the backward autocorrelation function of the backward travelling volume velocity in the ith acoustic tube is defined as

$$B_i(r) = \sum_{n=-\infty}^{\infty} V_i(n)V_i(n-r) \tag{6.57}$$

where n and r are integers. A cross tube correlation function between the forward and backward travelling volume velocities at the junction of the ith and $(i+1)$th acoustic tubes is defined as

$$S_i(r) = \sum_{n=-\infty}^{\infty} U_i(n)V_i(n-r) \tag{6.58}$$

where n and r are integers. Both the forward and backward autocorrelation functions are always symmetric, i.e.

$$A_i(-r) = A_i(r) \tag{6.59}$$

and

$$B_i(-r) = B_i(r) \tag{6.60}$$

for all r but, in general, the cross tube correlation function, $S_i(r)$, is not symmetric.

The junction equations describing the relationship between the forward and backward travelling volume velocities at the junction of the ith and $(i+1)$th acoustic tubes are Equations 2.25 and 2.26, i.e.

$$U_{i+1}(n) = (1+\mu_i)U_i(n) + \mu_i V_{i+1}(n-1) \tag{6.61}$$

and

$$V_i(n) = -\mu_i U_i(n) + (1-\mu_i)V_{i+1}(n-1) \tag{6.62}$$

Solving for $U_i(n)$ and $V_i(n)$ produces

$$U_i(n) = \frac{U_{i+1}(n) - \mu_i V_{i+}\ (n-1)}{(1+\mu_i)} \tag{6.63}$$

and

$$V_i(n) = \frac{V_{i+1}(n-1) - \mu_i U_{i+1}(n)}{(1+\mu_i)} \tag{6.64}$$

Equations 6.63 and 6.64 allow the volume velocities in the ith acoustic tube to be determined from the volume velocities in the $(i+1)$th acoustic tube if the reflection coefficient at the junction of these acoustic tubes, i.e. μ_i, is known.

Substituting Equation 6.63 into the definition of $A_i(r)$ (i.e. Equation 6.56), and simplifying the result by using Equations 6.56 to 6.60, produces, for integer $r \geqslant 0$

$$A_i(\pm r) = \frac{A_{i+1}(r) + \mu_i^2 B_{i+1}(r) - \mu_i\left(S_{i+1}(r+1) + S_{i+1}(-r+1)\right)}{(1+\mu_i)^2} \tag{6.65}$$

A similar substitution of Equation 6.64 into Equation 6.5, and simplifying the result by using Equations 6.56 to 6.60 produces, for integer $r \geqslant 0$

$$B_i(\pm r) = \frac{B_{i+1}(r) + \mu_i^2 A_{i+1}(r) - \mu_i\left(S_{i+1}(r+1) + S_{i+1}(-r+1)\right)}{(1+\mu_i)^2} \tag{6.66}$$

Multiplication of Equations 6.63 and 6.64, and simplying the result by using Equations 6.56 to 6.60 produces

$$S_i(\pm r) = \frac{S_{i+1}(\pm r+1) + \mu_i^2 S_{i+1}(\mp r+1) - \mu_i\left(A_{i+1}(r) + B_{i+1}(r)\right)}{(1+\mu_i)^2} \qquad (6.67)$$

for $r \geqslant 0$. The set of Equations 6.65 to 6.67 is important in the analysis procedure developed in this section, as the forward and backward autocorrelations and cross tube correlations in the ith acoustic tube can be determined from the forward and backward autocorrelations and cross tube correlations in the (i+1)th acoustic tube, provided μ_i is known.

In the situation where μ_M is unity, then μ_i is determined from a simple relationship with the known autocorrelation functions $A_{i+1}(r)$ and $B_{i+1}(r)$ and the cross tube correlation $S_{i+1}(r)$ in the (i+1)th acoustic tube (see Chapter 2). When the modulus of μ_M is less than unity, then the knowledge of the forward and backward autocorrelations $A_i(r)$ and $B_i(r)$ and the cross tube correlations $S_i(r)$ in the ith acoustic tube must also be known, in order to determine μ_i. However, during an analysis process, only the autocorrelations and cross tube correlations in the (i+1)th acoustic tube are known, and so μ_i must be estimated from these correlations. This estimation of μ_i is presented in the following section, but the definition of a basic procedure for the analysis of a situation where μ_M is less than unity does not require the knowledge of how μ_i is calculated, and is presented below.

An analysis procedure based on Equations 6.65 to 6.67 initially requires the knowledge of the forward and backward autocorrelation functions and the cross tube correlation function in the Mth acoustic tube, i.e. at the termination. The forward travelling volume velocity in the Mth acoustic tube, i.e. $U_M(n)$, is determine from the pre-emphasized radiated pressure waveform, as discussed in Section 6.4 and described by Equation 6.49. Using the definition of $A_i(r)$, i.e. Equation 6.56, the forward auto-correlation function at the termination, $A_M(r)$, is found from the known $U_M(r)$.

Multiplication of Equation 6.50 by Equation 6.50, with $n = n-r$, summing from $-\infty$ to $+\infty$, and using Equations 6.56 and 6.57 produces

$$B_M(r) = \mu_M^2 A_M(r) \tag{6.68}$$

for integer r. Mulitplication of Equation 6.50 by $U_M(n-r)$, summing from $-\infty$ to $+\infty$, and using Equations 6.56 and 6.58 produces

$$S_M(r) = -\mu_M A_M(r) \tag{6.69}$$

for integer r. Therefore, provided μ_M is known, $B_M(r)$ and $S_M(r)$ can be determined from $A_M(r)$. Equation 6.69 shows that $S_M(r)$ is symmetric but, in general, $S_i(r)$ is not symmetric for $i < M$.

An analysis procedure can now be defined, and the first step is to calculate $A_M(r)$, $B_M(r)$ and $S_M(r)$ from the pre-emphasized radiated pressure waveform via Equations 6.56, 6.68 and 6.69 and a known μ_M. The value of μ_{M-1} is estimated from $A_M(r)$, $B_M(r)$ and $S_M(r)$, or calculated in some other manner, and used to calculate $A_{M-1}(r)$, $B_{M-1}(r)$ and $S_{M-1}(r)$ from Equations 6.65 to 6.67. Therefore, μ_{M-2} can be calculated or estimated from $A_{M-1}(r)$, $B_{M-1}(r)$ and $S_{M-1}(r)$ which, in turn, permits $A_{M-2}(r)$, $B_{M-2}(r)$ and $S_{M-2}(r)$ to be determined via Equations 6.65 to 6.67. Repetition of this process enables μ_i to be identified for $1 \leqslant i \leqslant M-1$ from the measured radiated pressure waveform from a set of acoustic tubes. A recovered acoustic tube shape is then identified from the reflection coefficients, μ_i, via Equation 2.16. Figure 6.13 presents a summary of the above procedure in flow chart format.

To ensure the analysis procedure developed above and described in Figure 6.13 is computationally efficient, only those autocorrelations and cross tube correlations that are necessary for the analysis procedure are calculated. Examination of Equations 6.65 to 6.67 shows that the determination of $A_i(r)$, $B_i(r)$ and $S_i(r)$ for $0 \leqslant r \leqslant q$ requires the knowledge of $A_{i+1}(r)$, $B_{i+1}(r)$ for $0 \leqslant r \leqslant q$ and $S_{i+1}(r)$ for $0 \leqslant r \leqslant q+1$. Although the value of q is fixed to the manner in which μ_i is calculated, it is seen from the above statement that the number of autocorrelation and cross tube correlations that need to be known in the $(i+1)$th acoustic tube is greater than in the ith acoustic tube. This is consistent with other similar autocorrelation analysis procedures, e.g. MARKEL and GRAY [1976], LEROUX and GUEGUEN [1977] and WIGGINS [1978].

A computationally efficient implementation of Equations 6.65 to 6.67 is defined for a particular value of i by initially determining

$$K_1 = \mu_i^2 \qquad (6.70)$$

and

$$K_2 = (1-\mu_i)^2 \qquad (6.71)$$

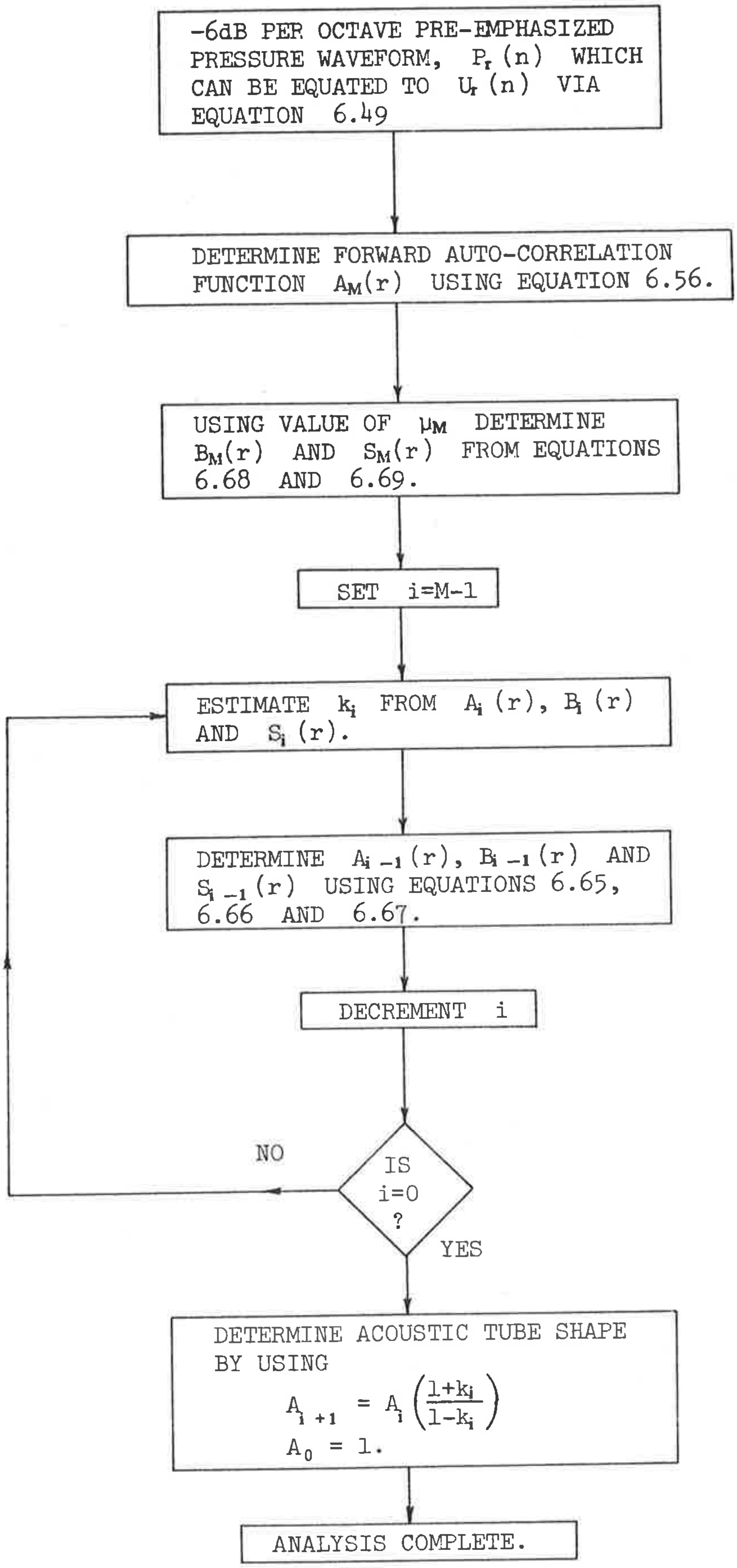

-6dB PER OCTAVE PRE-EMPHASIZED PRESSURE WAVEFORM, $P_r(n)$ WHICH CAN BE EQUATED TO $U_r(n)$ VIA EQUATION 6.49
DETERMINE FORWARD AUTO-CORRELATION FUNCTION $A_M(r)$ USING EQUATION 6.56.
USING VALUE OF μ_M DETERMINE $B_M(r)$ AND $S_M(r)$ FROM EQUATIONS 6.68 AND 6.69.
SET i=M-1
ESTIMATE k_i FROM $A_i(r)$, $B_i(r)$ AND $S_i(r)$.
DETERMINE $A_{i-1}(r)$, $B_{i-1}(r)$ AND $S_{i-1}(r)$ USING EQUATIONS 6.65, 6.66 AND 6.67.
DECREMENT i
IS i=0 ?
NO
YES
DETERMINE ACOUSTIC TUBE SHAPE BY USING
$A_{i+1} = A_i\left(\frac{1+k_i}{1-k_i}\right)$
$A_0 = 1.$
ANALYSIS COMPLETE.

FIGURE 6.13: Procedure for a new speech analysis which allows a non-unity termination reflection coefficient.

For each value of r

$$K_3 = -\mu_i\left(S_{i+1}(r+1) + S_{i+1}(-r+1)\right) \tag{6.72}$$

and

$$K_4 = -\mu_i\left(A_{i+1}(r) + B_{i+1}(r)\right) \tag{6.73}$$

Using Equations 6.70 to 6.73 permits Equations 6.65 to 6.67 to be rewritten as

$$A_i(r) = \frac{A_{i+1}(r) + K_1 B_{i+1}(r) + K_3}{K_2} \tag{6.74}$$

$$B_i(r) = \frac{B_{i+1}(r) + K_1 A_{i+1}(r) + K_3}{K_2} \tag{6.75}$$

and

$$S_i(\pm r) = \frac{S_{i+1}(\pm r + 1) + K_1 S_{i+1}(\mp r + 1) + K_3}{K_2} \tag{6.76}$$

respectively.

Additional reductions in the number of arithmetic computations required to calculate $A_i(r)$, $B_i(r)$ and $S_i(r)$ from $A_{i+1}(r)$, $B_{i+1}(r)$ $S_{i+1}(r)$ may be possible once the manner in which μ_i is calculated is known. The amount of storage space required to implement Equations 6.70 to 6.76 is slightly larger than an implementation of Equations 6.65 to 6.67 but, in practice, this disadvantage does not outweigh the advantage of increased computational efficiency.

6.5.2 CALCULATION OF REFLECTION COEFFICIENTS

The basis for an analysis procedure which may be used in situations where the termination reflection coefficient has a modulus of less than unity was presented in the previous section. The manner in which the reflection coefficient μ_i is calculated was not presented or discussed, and this section considers how μ_i might be determined so that accurate acoustic tube shape recovery may be achieved.

A general expression for μ_i in terms of autocorrelations and cross tube correlations is found by multiplying Equation 6.63 by $V_{i+1}(n-1)$ and summing from $-\infty$ to $+\infty$ to produce

$$\sum_{n=-\infty}^{\infty} U_i(n) V_{i+1}(n-1) = \frac{S_{i+1}(1) - \mu_i B_{i+1}(0)}{(1+\mu_i)} \qquad (6.77)$$

A cross junction correlation function between the forward travelling volume velocity in the ith acoustic tube and the backward travelling volume velocity in the $(i+1)$th acoustic tube is defined as

$$T_{i+1}^{UV}(r) = \sum_{n=-\infty}^{\infty} U_i(n) V_{i+1}(n-r) \qquad (6.78)$$

Using Equation 6.78, it is possible to rearrange Equation 6.77 to produce an expression for μ_i as

$$\mu_i = \frac{S_{i+1}(1) - T_{i-1}^{UV}(1)}{B_{i+1}(0) + T_{i+1}^{UV}(1)} \qquad (6.79)$$

Multiplication of Equation 6.64 by $U_{i+1}(n)$ and summing from $-\infty$ to $+\infty$ produces

$$\sum_{n=-\infty}^{\infty} V_i(n)U_{i+1}(n) = \frac{S_{i+1}(1) - \mu_i A_{i+1}(0)}{(1+\mu_i)} \qquad (6.80)$$

A cross junction correlation function between the backward travelling volume velocity in the ith acoustic tube and the forward travelling volume velocity in the $(i+1)$th acoustic tube is defined as

$$T_{i+1}^{UV}(r) = \sum_{n=-\infty}^{\infty} V_i(n)U_{i+1}(n-r) \qquad (6.81)$$

Using Equation 6.81 above, Equation 6.80 can be rewritten to provide an expression for μ_i as

$$\mu_i = \frac{S_{i+1}(1) - T_{i+1}^{VU}(0)}{A_{i+1}(0) + T_{i+1}^{VU}(0)} \qquad (6.82)$$

Many other expressions for μ_i are possible apart from the two presented above, i.e. Equations 6.79 and 6.82, but all have at least one autocorrelation or crosscorrelation which involves a volume velocity waveform from the ith acoustic tube. During an analysis process, the volume velocities in the ith acoustic tube are not known, or able to be calculated, unless μ_i is known. Therefore, Equations 6.79 and 6.82 cannot be used to determine the value of μ_i in an analysis procedure.

When the termination of the acoustic tube is lossless, then the termination reflection coefficient, μ_M, is unity, and Equation 6.68 reduces to

$$B_M(r) = A_M(r) \tag{6.83}$$

for all integer r. Substitution of Equation 6.83 into Equations 6.65 and 6.66 with $i = M-1$ produces

$$A_{M-1}(r) = \frac{(1+\mu_{M-1})^2 A_M(r) - \mu_{M-1}\big(S_M(r+1) + S_M(-r+1)\big)}{(1+\mu_{M-1})^2} \tag{6.84}$$

and

$$B_{M-1}(r) = \frac{(1+\mu_{M-1})^2 A_M(r) - \mu_{M-1}\big(S_M(r+1) + S_M(-r+1)\big)}{(1+\mu_{M-1})^2} \tag{6.85}$$

respectively, for all integer r. Comparison of Equations 6.84 and and 6.85 shows that

$$B_{M-1}(r) = A_{M-1}(r) \tag{6.86}$$

for all integer r. Repeating this process for $0 \leqslant i \leqslant M-2$ results in

$$B_i(r) = A_i(r) \tag{6.87}$$

for $0 \leqslant i \leqslant M$ and all integer r.

Substitution of Equation 6.87 into Equations 6.79 and 6.82, and comparing the results with Equation 6.82, reveals that

$$T_{i+1}^{vu}(0) = T_{i+1}^{uv}(1) \tag{6.88}$$

for a lossless termination of the acoustic tubes. If the acoustic tubes are in thermal equilibrium, and the termination is lossless, an equipartion of energy occurs among the acoustic tubes (BOGNER [1977], and BOGNER and DAVIS [1980]), i.e.

$$\frac{A_i(0)}{Z_i} = \frac{A_{i+1}(0)}{Z_{i+1}} \tag{6.89}$$

for $0 \leqslant i \leqslant M-1$, where Z_i is the acoustic impedance of the ith acoustic tube.

Squaring both sides of Equation 6.61 and summing the time index from $-\infty$ to $+\infty$ produces

$$A_{i+1}(0) = (1+\mu_i)^2 A_i(0) + 2\mu_i T^{UV}_{i+1}(1) + \mu_i^2 B_{i+1}(0) \tag{6.90}$$

Substitution of Equation 6.87 into Equation 6.90 produces

$$2\mu_i T^{UV}_{i+1}(1) = (1-\mu_i^2)A_{i+1}(0) - (1+\mu_i)^2 A_i(0) \tag{6.91}$$

Using

$$Z_i = \frac{\rho c}{A_i} \tag{6.92a}$$

and

$$Z_{i+1} = \frac{\rho c}{A_{i+1}} \tag{6.92b}$$

Equation 6.89 is rewritten as

$$(1+\mu_i)A_i(0) = (1-\mu_i)A_{i+1}(0) \tag{6.93}$$

and multiplication by $(1+\mu_i)$ produces

$$(1+\mu_i)^2 A_i(0) = (1-\mu_i^2)A_{i+1}(0) \tag{6.94}$$

Substitution of Equation 6.94 into Equation 6.91 produces

$$T_{i+1}^{uv}(1) = 0 \tag{6.95}$$

and, hence, from Equation 6.88

$$T_{i+1}^{vu}(0) = 0 \tag{6.96}$$

Equations 6.95 and 6.96 are the Cross Junction Correlation Theorems of BOGNER [1977], and BOGNER and DAVIS [1980].

The Cross Junction Correlation Theorems (Equations 6.95 and 6.96) reduce the expressions derived for μ_i, i.e. Equations 6.79 and 6.82, to

$$\mu_i = \frac{S_{i+1}(1)}{A_{i+1}(0)} \tag{6.97}$$

Equation 6.97 expresses μ_i in terms of known autocorrelation and cross tube correlations in the $(i+1)$th acoustic tube; therefore, μ_i is uniquely determined from known quantities wherever the termination is lossless. Using Equation 6.97 in conjunction with the analysis process described in Section 6.5.1 defines an analysis procedure to determine a unique set of μ_i for $0 \leqslant i \leqslant M-1$ from an acoustic waveform.

For a lossless termination of acoustic tubes, Equation 6.87 permits further simplification of the equations describing the analysis process of Section 6.5.1, i.e. Equations 6.70 and 6.76. Since the forward and reverse autocorrelation functions are equal (see Equation 6.87), only one needs to be calculated. As μ_i is determined from Equation 6.97 as a ratio of $S_{i+1}(1)$ to $A_{i+1}(0)$, a scaling of $A_{i+1}(r)$ and $S_{i+1}(r)$ by the same constant does not change the value of μ_i.

Hence, for a termination reflection coefficient of unity, the set of Equations 6.70 to 6.76 reduces to the following set of equations. For known $A_{i+1}(r)$ and $S_{i+1}(\pm r)$, the reflection coefficient μ_i is calculated as

$$\mu_i = \frac{S_{i+1}(1)}{A_{i+1}(0)} \tag{6.98}$$

which then allows the computation of the constants

$$K_1 = \mu_i^2 \tag{6.99}$$

and

$$K_2 = (1+\mu_i^2) \tag{6.100}$$

The autocorrelations and cross tube correlations in the ith acoustic tube are then calculated as

$$A_i(r) = K_2 A_{i+1}(r) - \mu_i\big(S_{i+1}(r-1) + S_{i+1}(-r+1)\big) \tag{6.101}$$

and

$$S_i(\pm r) = S_{i+1}(\pm r+1) + \mu_i S_{i+1}(\mp r+1) - 2\mu_i A_{i+1}(r) \tag{6.102}$$

for $r \geqslant 0$. A flow chart is presented in Figure 6.14 to illustrate the analysis process based on Equations 6.98 to 6.102.

The analysis procedure described in Figure 6.14 and by Equations 6.98 to 6.102 does not present the most computationally efficient process for recovering a set of reflection coefficients

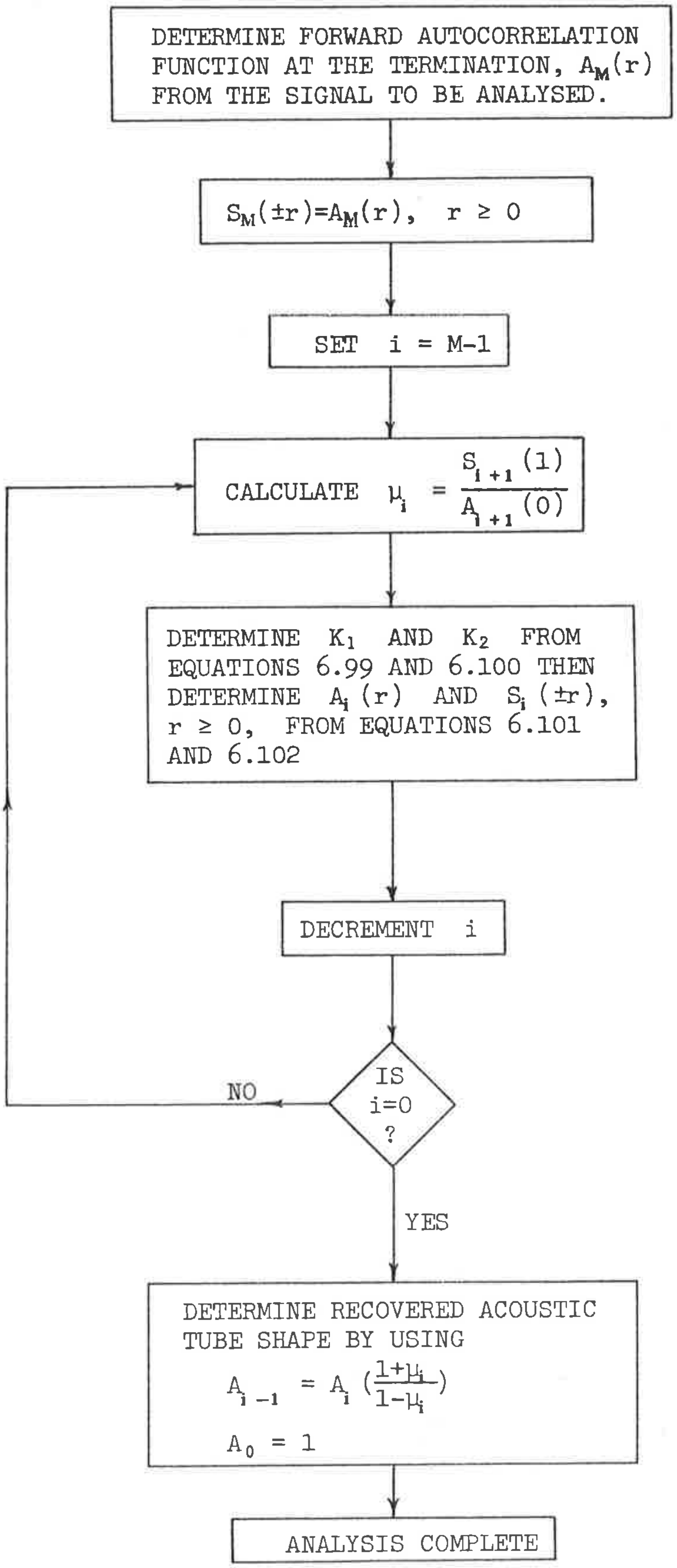

DETERMINE FORWARD AUTOCORRELATION FUNCTION AT THE TERMINATION, $A_M(r)$ FROM THE SIGNAL TO BE ANALYSED.
$S_M(\pm r)=A_M(r), \quad r \geq 0$
SET $i = M-1$
CALCULATE $\mu_i = \frac{S_{i+1}(1)}{A_{i+1}(0)}$
DETERMINE K_1 AND K_2 FROM EQUATIONS 6.99 AND 6.100 THEN DETERMINE $A_i(r)$ AND $S_i(\pm r)$, $r \geq 0$, FROM EQUATIONS 6.101 AND 6.102
DECREMENT i
IS i=0 ?
NO
YES
DETERMINE RECOVERED ACOUSTIC TUBE SHAPE BY USING
$A_{i-1} = A_i\left(\frac{1+\mu_i}{1-\mu_i}\right)$
$A_0 = 1$
ANALYSIS COMPLETE

FIGURE 6.14: Procedure for new analysis process when μ_M is equal to unity.

μ_i, $0 \leq i \leq M-1$, from an acoustic waveform. The analysis procedures of LEVINSON [1947], DURBIN [1960], and LEROUX and GUEGUEN [1977], for instance, require approximately one third the number of multiplications. The reason for the inefficiency of the analysis process defined by Figure 6.14 is the unnecessary calculation of the cross tube correlations, $S_i(r)$. Elimination of the need to calculate crosscorrelations is achieved by the analysis process of DURBIN [1960], and LEROUX and GUEGUEN [1977] by using abstract quantities which have no relation to the acoustic tube model. Hence, the analysis process described in Figure 6.14 is useful when the analysis process needs to be related to quantities on the acoustic tube model being identified.

For a lossy termination of the acoustic tube model, the termination reflection coefficient has been shown to have a modulus of less than unity; therefore, Equation 6.68 implies that

$$A_M(r) \neq B_M(r) \tag{6.103}$$

for any r. Using Equation 6.103 in conjunction with Equations 6.65 and 6.66 shows that

$$A_i(r) \neq B_i(r) \tag{6.104}$$

for $0 \leq i \leq M$ and all r. When Equation 6.104 is true, then it is easily shown that

$$T_{i+1}^{uv}(1) \neq T_{i+1}^{vu}(0) \tag{6.105}$$

from Equations 6.79 and 6.82 and, in general, neither is zero. Therefore, in the situation where the termination of acoustic tubes is lossy, the values of $T_{i+1}^{uv}(1)$ and $T_{i+1}^{vu}(0)$ are unknown and, hence, μ_i cannot be calculated from Equation 6.79 or Equation 6.82.

To use Equations 6.79 and/or 6.82 in an analysis process, when the termination of the acoustic tubes is lossy, it is necessary to estimate the value of either $T_{i+1}^{uv}(1)$ or $T_{i+1}^{vu}(0)$. Empirical investigations have shown that the magnitudes of $T_{i+1}^{uv}(1)$ and $T_{i+1}^{vu}(0)$ are highly dependent on the value of μ_M, i.e. the termination reflection coefficient. When the modulus of μ_M is close to unity, then both $T_{i+1}^{uv}(1)$ and $T_{i+1}^{vu}(0)$ are close to zero (and equal to zero when $\mu_M = 1$), and their magnitudes increase as the modulus of μ_M decreases from unity. The empirical investigations performed showed that, unless the estimates of $T_{i+1}^{uv}(1)$ and/or $T_{i+1}^{vu}(0)$ are accurate, unstable acoustic tube shapes, i.e. negative cross-sectional areas, may be recovered.

The results of empirical investigations suggest that a reasonable estimate of $T_{i+1}^{uv}(1)$ and $T_{i+1}^{vu}(0)$, which may provide improved acoustic tube shape recovery for a lossy termination of the acoustic tubes, is to assume that the magnitudes of $T_{i+1}^{uv}(1)$ and $T_{i+1}^{vu}(0)$ are small in comparison with the autocorrelations and cross tube correlations in the $(i+1)$th acoustic tube. This hypothesis is true when the modulus of μ_M is close to unity, when $T_{i+1}^{uv}(1)$ and $T_{i+1}^{vu}(0)$ approach zero. The rest of this section defines a number of analysis processes which can be derived from this hypothesis, and the following section evaluates the effectiveness of each of these analysis processes to provide improved acoustic tube shape recovery, when the termination of acoustic tubes is lossy.

From the hypothesis that $T^{UV}_{i+1}(1)$ and $T^{VU}_{i+1}(0)$ are small in comparison with the autocorrelations and cross tube correlations in the $(i+1)$th acoustic tube, Equation 6.79 permits μ_i to be estimated as

$$\mu_i = \frac{S_{i+1}(1)}{B_{i+1}(0)} \tag{6.106}$$

and Equation 6.82 permits μ_i to be estimated as

$$\mu_i = \frac{S_{i+1}(1)}{A_{i+1}(0)} \tag{6.107}$$

Equation 6.106 is recognised as the expression for μ_i which minimizes the forward predictor error of the lattice formulation of linear prediction, and Equation 6.107 minimizes the backward predictor error of the lattice formulation of linear prediction.

For a lossy termination of a set of acoustic tubes, it has been shown that, in general, $A_{i+1}(0)$ is not equal to $B_{i+1}(0)$ (see Equation 6.104) and, therefore, the μ_i calculated from Equation 6.106 is different from that calculated from Equation 6.107. Therefore, two new analysis processes can be defined by using either Equation 6.106 or Equation 6.107, in conjunction with the analysis procedure defined in Figure 6.13. The analysis process defined in Figure 6.13 used with Equation 6.106 is referred to as the Forward Lossy Termination analysis procedure, and the analysis process defined in Figure 6.13 used with Equation 6.107 is referred to as the Backward Lossy Termination analysis procedure.

There exist many other expressions for μ_i (see MAKHOUL [1977]) which can be shown to be various combinations of Equations 6.106 and 6.107. The most well-known are the Parcor form

$$\mu_i = \frac{S_{i+1}(1)}{\sqrt{A_{i+1}(0)B_{i+1}(0)}} \tag{6.108}$$

the Burg form

$$\mu_i = \frac{2S_{i+1}(1)}{A_{i+1}(0) + B_{i+1}(0)} \tag{6.109}$$

and the Minimum form

$$\mu_i = \text{SIGN} \cdot \min\left(\left|\frac{S_{i+1}(1)}{A_{i+1}(0)}\right|, \left|\frac{S_{i+1}(1)}{B_{i+1}(0)}\right|\right) \tag{6.110}$$

where SIGN is the sign of $S_{i+1}(1)/A_{i+1}(0)$ and $S_{i+1}(1)/B_{i+1}(0)$, both have the same sign since $A_{i+1}(0)$ and $B_{i+1}(0)$ are always positive. It is shown in Appendix F that the Minimum form of μ_i is equivalent to the Backward form of μ_i.

All the expressions for μ_i presented by MAKHOUL [1977] ensure that the hypothesis of $T^{UV}_{i+1}(1)$ and $T^{VU}_{i+1}(0)$ being small in comparison with the autocorrelations and cross tube correlations in the (i+1)th acoustic tube is true. Therefore, numerous Lossy Termination analysis procedures can be defined using the numerous expressions for μ_i presented by MAKHOUL [1977]. However, the four forms of μ_i presented in Equation 6.106, 6.107, 6.108 and 6.109 form a representive set of the numerous expressions for μ_i, and are the only ones considered here. The analysis process defined in Figure 6.13 used with Equation 6.108 is referred to as the

Parcor Lossy Termination analysis procedure, and the analysis process defined in Figure 6.13 used with Equation 6.109 is referred to as the Backward Lossy Termination analysis procedure.

The four Lossy Termination analysis procedures defined in this section are evaluated in the following section with synthetic speech which is generated with the termination reflection coefficient, μ_i, being the same as that used in the Lossy Termination analysis procedures. The Lossy Termination procedure which provides the best improvement in acoustic tube shape recovery is found from these evaluations. When real speech is being analysed, the termination reflection coefficient, μ_M, at the lips is unknown. Therefore, Section 6.5.2.2 evaluates a Lossy Termination analysis process when the value of μ_M is different in the synthetic speech generation process from that used in the Lossy Termination analysis.

6.5.2.1 Evaluations with Correct Lossy Termination

This section evaluates the four Lossy Termination analyses, defined in Section 6.5.2, with synthetic speech generated by the procedure detailed in Appendix E, for a sampling frequency of 10 kilohertz. The excitation used in the generation of the synthetic speech is an impulse (i.e. ideal white excitation), and the termination reflection coefficient, μ_M, of the acoustic tube model is varied from zero to unity. The value of μ_M used in the Lossy Termination analyses is the same as that used to generate the synthetic speech being analysed. The acoustic tube shapes used in the generation of the synthetic speech approximate the real vocal tract shapes measured by FANT [1970] for five vowel sounds (see Appendix C). All the assumptions of the linear prediction/acoustic tube model are satisifed, except for a loss at the termination, so that effects such as glottal pulse excitation, acoustic tube losses, etc., do not cloud the evaluations presented in this section.

The analysis procedure used throughout this section is to, firstly, apply a -6 dB per octave pre-emphasis to the synthetic speech waveform to account for the zero of the radiation impedance (see Section 6.2). The pre-emphasized waveform is then analysed by a conventional Parcor linear predictive analysis and the four Lossy Termination analyses to determine recovered acoustic tube shapes. Area distances between recovered and original acoustic tube shapes are calculated, and plotted against the value of μ_M used in the Lossy Termination analyses.

Figure 6.15 presents the area distances for the four Lossy Termination analyses and a conventional Parcor linear predictive analysis (which assumes a lossless termination) of synthetic speech for the vowel |a| versus the value of μ_M used in the Lossy Termination analyses and in generating the synthetic speech. As expected, all four Lossy Termination analyses produce the same area distances as the conventional Parcor analysis when μ_M is unity, i.e. a lossless termination. Both the Backward and Burg Lossy Termination analyses are shown to produce larger area distance than a conventional Parcor analysis, in Figure 6.15, for μ_M ranging from approximately zero to unity. The Burg Lossy Termination analysis only has a significantly greater area distance than a conventional Parcor analysis for μ_M less than 0.6. Hence, for synthetic speech of the vowel |a| the Backward and Burg Lossy Termination analyses do not provide an improvement in acoustic tube shape recovery when compared with

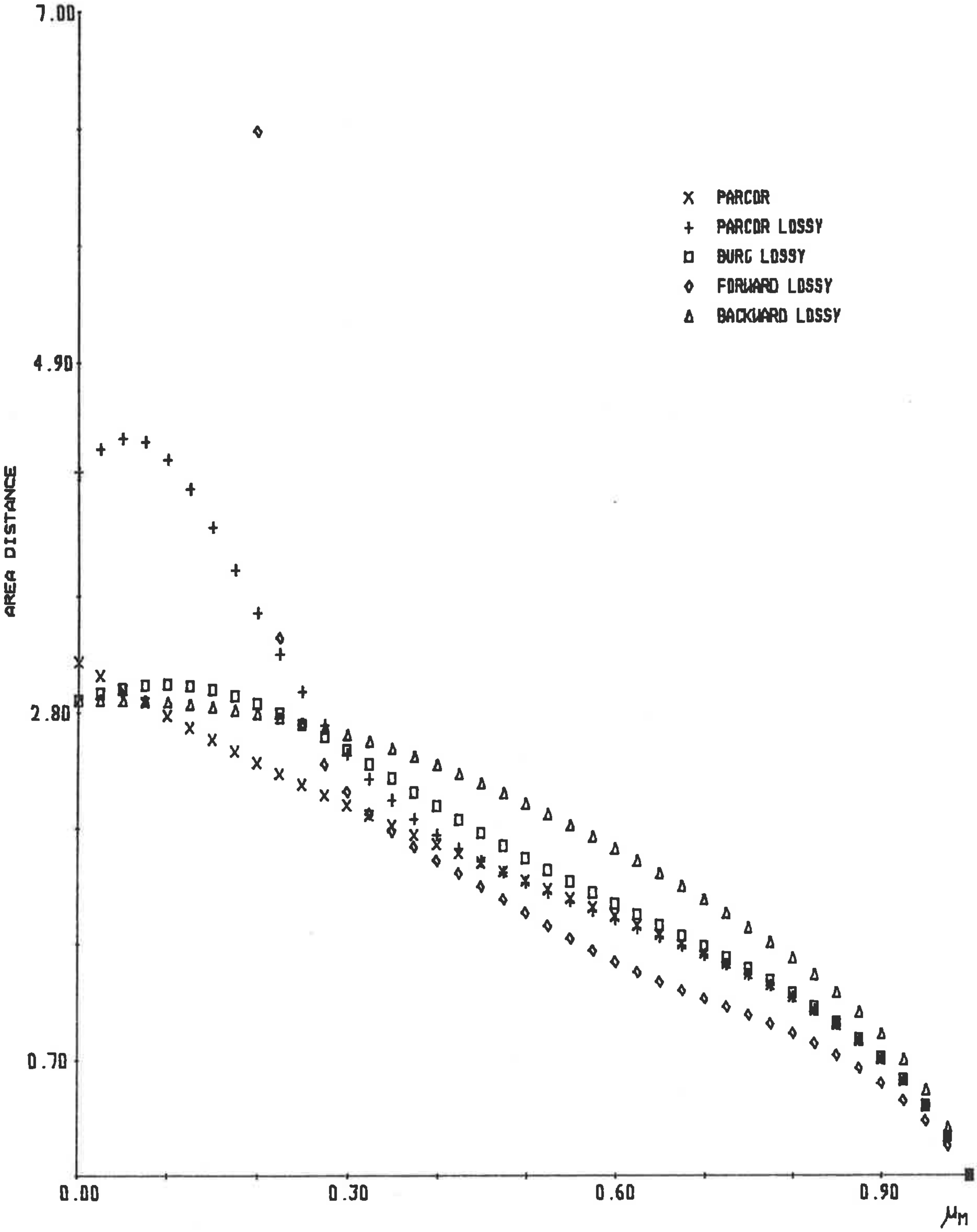


FIGURE 6.15: Area distances for a conventional parcor analysis and four lossly termination analyses of synthetic speech for the vowel /a/.

a conventional Parcor linear predictive analysis which assumes a lossless termination.

The Parcor and Forward Lossy Termination analyses produce a reduction in area distances when compared with those for a conventional Parcor analysis for certain ranges of μ_M. The Parcor Lossy Termination analysis only provides a small reduction in area distance for μ_M between 0.5 and unity, with much larger area distances for μ_M less than 0.5, when compared with a conventional Parcor analysis. In contrast, the Forward Lossy Termination analysis produces a large reduction in area distances in comparison with a conventional Parcor analysis for μ_M between 0.3 and unity, but much larger area distances when μ_M is between 0.2 and 0.3. Unrealizable acoustic tube shapes (i.e. some reflection coefficients have a modulus of greater than unity) are recovered by the Forward Lossy Termination analysis when μ_M is between zero and 0.2.

Hence, for synthetic speech of the vowel |a|, only two of the four Lossy Termination analyses, namely the Forward and Parcor Lossy Termination analyses, provide improved acoustic tube shape recovery in comparison with a conventional Parcor analysis. The improved acoustic tube shape recovery occurs for a wide range of μ_M, i.e. from 0.5 to unity and 0.3 to unity for the Parcor and Forward Lossy Termination analyses, respectively. All the Lossy Termination analyses, except for the Forward Lossy Termination analysis when μ_M is between zero and 0.2, recover realizable acoustic tube shapes, i.e. all reflection coefficients have a modulus of less than unity.

The area distances between recovered and original acoustic tube shapes versus μ_M for the four Lossy Termination analyses and a conventional Parcor analysis of synthetic speech for the vowel |e|, are presented in Figure 6.16. A comparison of Figures 6.15 and 6.16 shows similar trends for the analysis of synthetic speech of the vowels |a| and |e| by the four Lossy Termination analyses and a conventional Parcor analysis.

Figure 6.16 shows that the Backward and Burg Lossy Termination analyses produce larger area distances than a conventional Parcor analysis for the range of μ_M from zero to unity. This observation is consistent with that for the synthetic speech of the vowel |a| and, hence, for both the vowels |a| and |e|, the Backward and Burg Lossy Termination analyses do not provide improved acoustic tube shape recovery in comparison with a conventional Parcor analysis.

Figure 6.16 shows that the Parcor Lossy Termination analysis produces smaller area distances than a conventional Parcor analysis of synthetic speech for the vowel |e|, when μ_M is between unity and and approximately zero. The Forward Lossy Termination analysis produces much smaller area distances than a conventional Parcor analysis, and the other Lossy Termination analyses, when μ_M is between 0.55 and unity. For μ_M less than 0.55, the Forward Lossy Termina-analysis produces larger area distances in comparison with a conventional Parcor analysis, and unrealizable acoustic tube shapes when μ_M is between zero and 0.2.

Hence, only two of the four Lossy Termination analyses, namely the Parcor and Forward Lossy Termination analyses, produce improved acoustic tube shape recovery when compared with a conventional

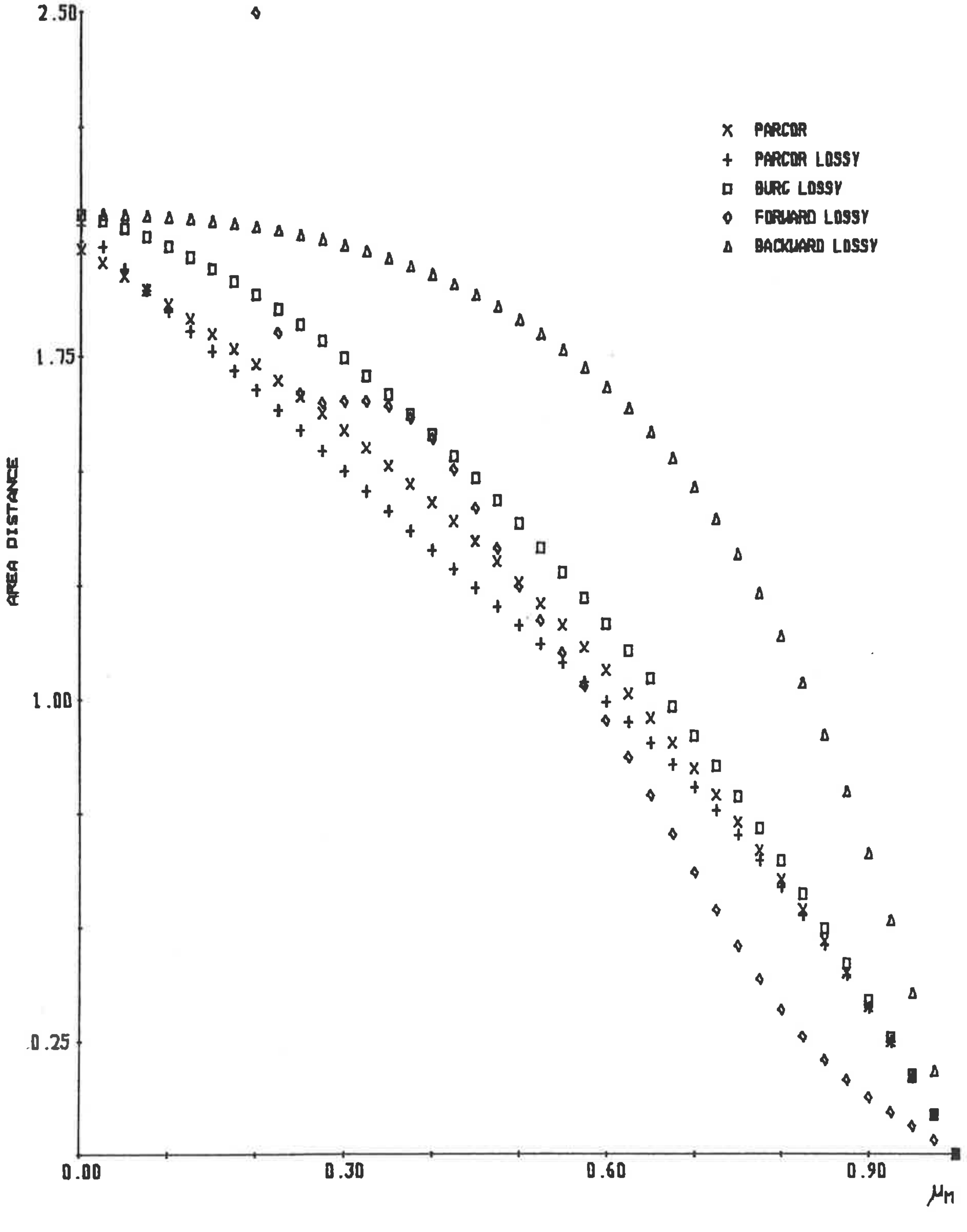


FIGURE 6.16: Area distances for a conventional parcor analysis and four lossy termination analyses of synthetic speech for the vowel /e/.

Parcor analysis of synthetic speech for the vowel |e|. All the Lossy Termination analyses recover realizable acoustic tube shapes, with the exception of the Forward Lossy Termination analysis for μ_M between zero and 0.2. In general, the observations for the performance of the four Lossy Termination analyses of synthetic speech for the vowel |e| are consistent with the observations for the performance of the same analyses of synthetic speech for the vowel |a|.

Figure 6.17 presents the area distances between recovered and original acoustic tube shapes versus μ_M for the four Lossy Termination analyses and a conventional Parcor analysis of synthetic speech for the vowel |i|. A comparison of Figure 6.17 with Figures 6.15 and 6.16, which present evaluation results for the vowels |a| and |e|, respectively, shows marked differences in the general trends for the vowel |i| from those for both the vowels |a| and |e|.

The Backward Lossy Termination analysis is shown in Figure 6.17 to produce much larger area distances than a conventional Parcor analysis. The Forward Lossy Termination analysis has slightly larger area distances than a conventional Parcor analysis for μ_M greater than 0.7, and slightly smaller area distances when μ_M is between 0.55 and 0.7. Unrealizable acoustic tube shapes are recovered by the Forward Lossy Termination analysis when μ_M is less than 0.5.

The Burg Lossy Termination analysis is shown in Figure 6.17 to produce slightly smaller area distances than a conventional Parcor analysis when μ_M is between 0.5 and unity, but much larger

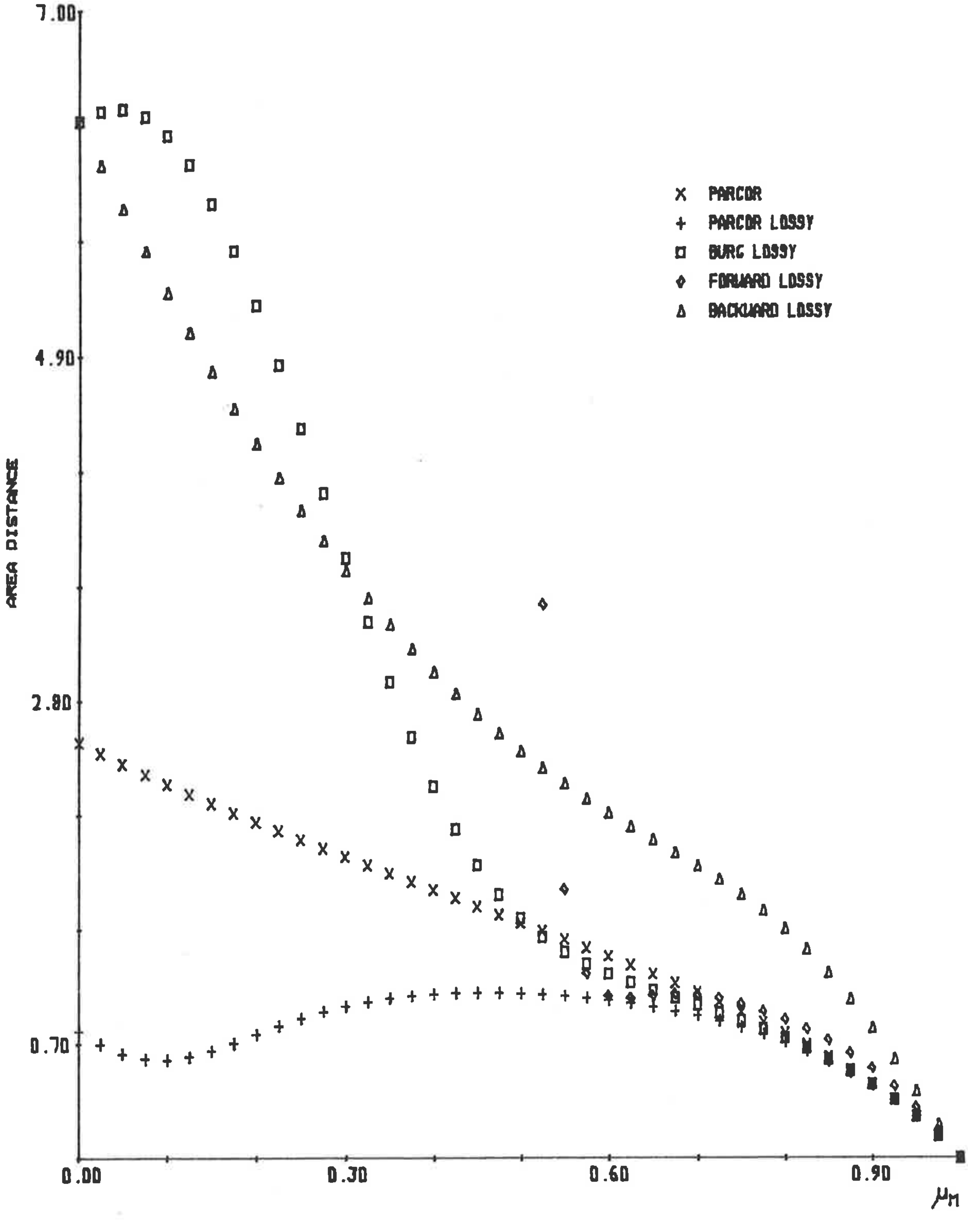


FIGURE 6.17. Area distances for a conventional parcor analysis and four lossy termination analyses of synthetic speech for the vowel /i/.

area distances when μ_M is less than 0.5. A large reduction in area distances is observed in Figure 6.17 for the Parcor Lossy Termination analysis in comparison with a conventional Parcor analysis of synthetic speech for the vowel |i|. The reduction in area distances by the Parcor Lossy Termination analysis and, hence, an improvement in acoustic tube shape recovery, increases as the value of μ_M decreases from unity to zero.

Therefore, only two of the four Lossy Termination analyses, namely the Burg and Parcor Lossy Termination analyses, provide a reduction in area distances for a wide range of μ_M, when compared with a conventional Parcor analysis of synthetic speech for the vowel |i|. The Forward Lossy Termination analysis is the only Lossy Termination analysis procedure that recovers unrealizable acoustic tube shapes from synthetic speech of the vowel |i|. Of the evaluation results presented for the four Lossy Termination analyses of synthetic speech for the vowels |a|, |e| and |i|, only the Parcor Lossy Termination analysis produces a consistent reduction in area distances when compared with those for a conventional Parcor analysis, for a wide range of μ_M.

A plot of area distances between recovered and original acoustic tube shapes for the four Lossy Termination analyses and a conventional Parcor analysis of synthetic speech for the vowel |o| versus μ_M is presented in Figure 6.18. The Backward Lossy Termination analysis produces larger area distances than a conventional Parcor analysis when μ_M is greater than 0.5, but smaller area distances when μ_M is less than 0.5. The Burg Lossy Termination analysis produces similar area distances for μ_M between 0.7 and unity, larger area distances when μ_M is between 0.2 and

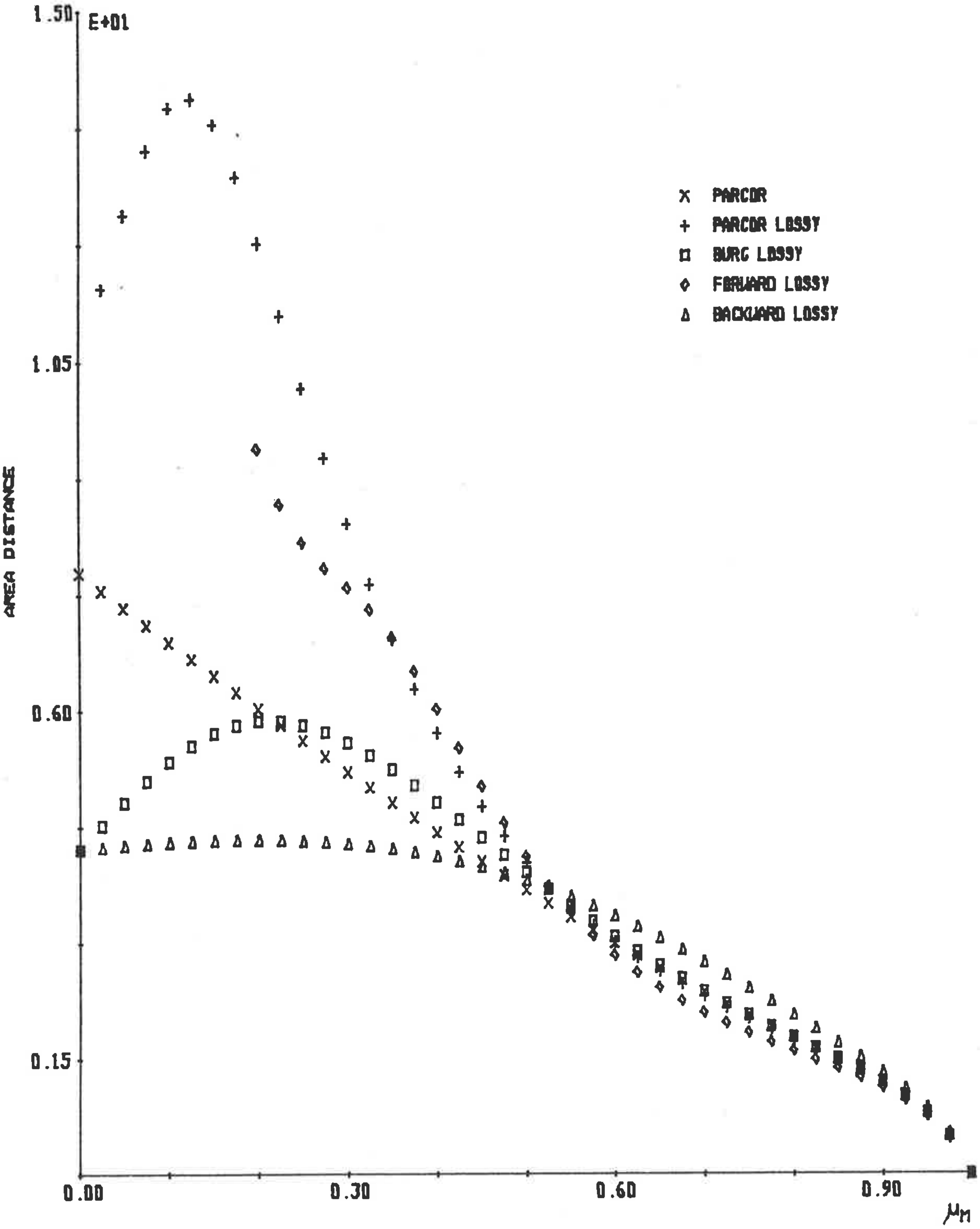


FIGURE 6.18: Area distances for a conventional parcor analysis and four lossy termination analyses of synthetic speech for the vowel /o/.

0.7, and smaller area distances when μ_M is less than 0.2, in comparison with a conventional Parcor analysis.

The area distances for a Forward Lossy Termination analysis, in comparison with a conventional Parcor analysis, are shown in Figure 6.18 to be slightly smaller when μ_M is greater than 0.5, but much larger when μ_M is less than 0.5. Unrealizable acoustic tube shapes are recovered by the Forward Lossy Termination analysis when μ_M is less than 0.2. For the Parcor Lossy Termination analysis, a small reduction in area distances, in comparison with a conventional Parcor analysis, occurs for μ_M greater than 0.5, and much larger area distances occur for μ_M less than 0.5.

For synthetic speech of the vowel |o|, only a small reduction in area distances, in comparison with a conventional Parcor analysis, occurs for two of the Lossy Termination analyses, namely the Parcor and Forward Lossy Termination analyses, and then only when μ_M is greater than 0.5. Only the Forward Lossy Termination analysis recovers unrealizable acoustic tube shapes, which occur when μ_M is less than 0.2. Of the evaluation results presented for the four Lossy Termination analyses of synthetic speech for the vowels |a|, |e|, |i| and |o|, only the Parcor Lossy Termination analysis produces consistently smaller area distances than a conventional Parcor analysis, for a consistently wide range of μ_M.

The area distances between recovered and original acoustic tube shapes for the four Lossy Termination analyses and a conventional Parcor analysis of synthetic speech for the vowel |u| are presented versus μ_M in Figure 6.19. In contrast to the evaluation results presented previously, all the Lossy Termination analyses

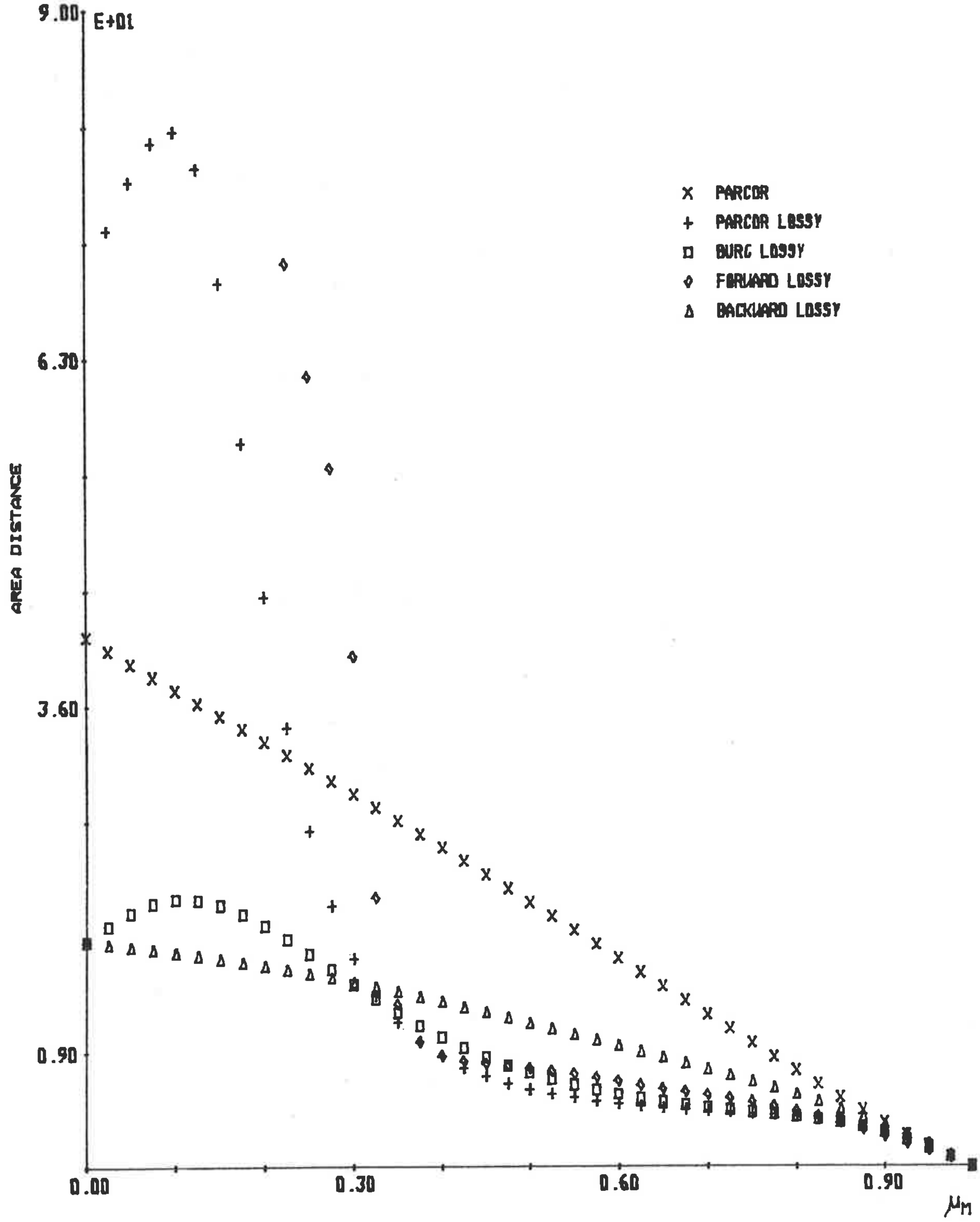


FIGURE 6.19: Area distances for a conventional parcor analysis and four lossy termination analyses of synthetic speech for the vowel /u/.

are shown in Figure 6.19 to produce smaller area distances than a conventional Parcor anlaysis, for large ranges of μ_M. The Backward and Burg Lossy Termination analyses produce much smaller area distances than a conventional Parcor analysis for μ_M between zero and unity. The Forward and Parcor Lossy Termination analyses produce smaller area distances when μ_M is greater than 0.3 and 0.2, respectively, but much larger area distances otherwise, in comparison with the area distances for a conventional Parcor analysis. Unrealizable acoustic tube shapes are recovered by the Forward Lossy Termination analysis when μ_M is less than 0.2.

The evaluations with synthetic speech of five vowels have shown that the Backward Lossy Termination analysis consistently produces the largest area distances of the four Lossy Termination analyses when μ_M is greater than 0.3, and these area distances are also larger than those for a conventional Parcor analysis for all but the vowels |o| and |u|. The Burg Lossy Termination analysis was shown to produce larger area distances than a conventional Parcor analysis over large ranges of μ_M for all the vowels except the vowel |u|. Hence, the Backward and Burg Lossy Termination analyses do not, in general, provide improved acoustic tube shape recovery.

A consistent reduction in area distances in comparison with those for a conventional Parcor analysis is observed for the Forward Lossy Termination analysis for all the vowels except the vowel |i|. The reduction in area distances always occurs for μ_M greater than 0.5 and, in some cases, for μ_M greater than 0.3. A major disadvantage of the Forward Lossy Termination analysis is the recovery of unrealizable acoustic tube shapes when μ_M is small.

Of all the four Lossy Termination analyses evaluated in this section, only the Parcor Lossy Termination analysis consistently produces a reduction in area distances in comparison with a conventional Parcor analysis. This reduction in area distances occurs over a wide range of μ_M, at least from 0.5 to unity and, in some cases, from zero to unity.

Hence, it is concluded that the use of a Parcor Lossy Termination analysis in situations where the acoustic tubes have a lossy termination, in general, produces improved acoustic tube shape recovery in comparison with a conventional Parcor analysis. This improvement in acoustic tube shape recovery is expected when μ_M is greater than 0.5, but may occur for a larger range of μ_M.

6.5.2.2 Evaluations with Incorrect Lossy Termination

The evaluation of the four Lossy Termination analyses presented in the previous section used the same value of μ_M in the analysis procedure and to generate the synthetic speech. In general, the value of μ_M used to generate speech waveforms is unknown; therefore, it is necessary to estimate a value of μ_M for a Lossy Termination analysis. This section evaluates the performance of the Parcor Lossy Termination analysis when the termination reflection coefficient used in the analysis procedure, denoted as μ_M', is different from that used to generate the synthetic speech. Only the Parcor Lossy Termination analysis is considered in this section, since it was the only Lossy Termination analysis found in the previous section to consistently produce improved acoustic tube shape recovery for a wide range of μ_M.

The synthetic speech waveforms used in the evaluations presented in this section are generated by the procedure detailed in Appendix E for a sampling frequency of 10 kilohertz, and an impulse excitation, i.e. an ideal white excitation, is used. The acoustic tube shapes used to generate the synthetic speech approximate real vocal tract shapes as measured by FANT [1960] for five vowel sounds (see Appendix C). All the assumptions of the linear prediction/ acoustic tube model, except for a loss at the termination, are satisfied so that effects caused by glottal pulse excitation, losses in acoustic tubes, etc., do not cloud the evaluations presented in this section.

The analysis procedure used throughout this section is to, firstly, apply a -6 dB per octave pre-emphasis to the synthetic speech waveform, to account for the zero of the radiation impedance (see Section 6.2). The pre-emphasized waveform is then analysed by a conventional Parcor analysis and a Parcor Lossy Termination analysis to produce a recovered acoustic tube shape. Area distances between recovered and original acoustic tube shapes are calculated, and plotted against the value of μ_M' used in the Parcor Lossy Termination analysis.

The area distances between recovered and original acoustic tube shapes for a Parcor Lossy Termination analysis and a conventional Parcor analysis of synthetic speech for the vowels |a|, |e|, |i|, |o| and |u| versus μ_M' are presented in Figure 6.20, 6.21, 6.22, 6.23 and 6.24, respectively. The synthetic speech waveforms are generated with a termination reflection coefficient, μ_M, of either 0.9, 0.8, 0.7, 0.6 or 0.5. Since the conventional Parcor analaysis uses a termination reflection coefficient of unity, no change in

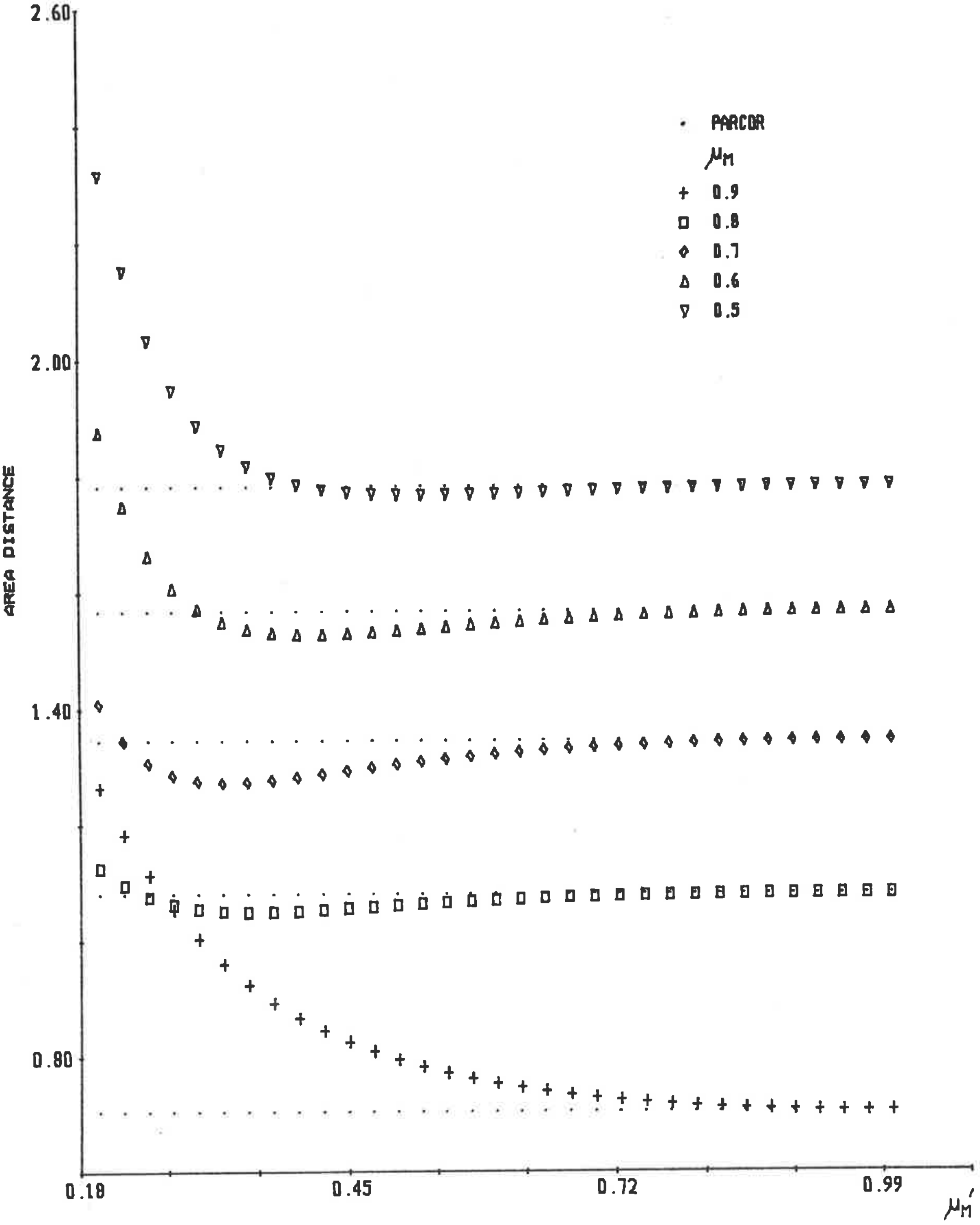


FIGURE 6.20: Area distances for a conventional parcor analysis and a parcor lossy termination analysis of synthetic speech for the vowel /a/.

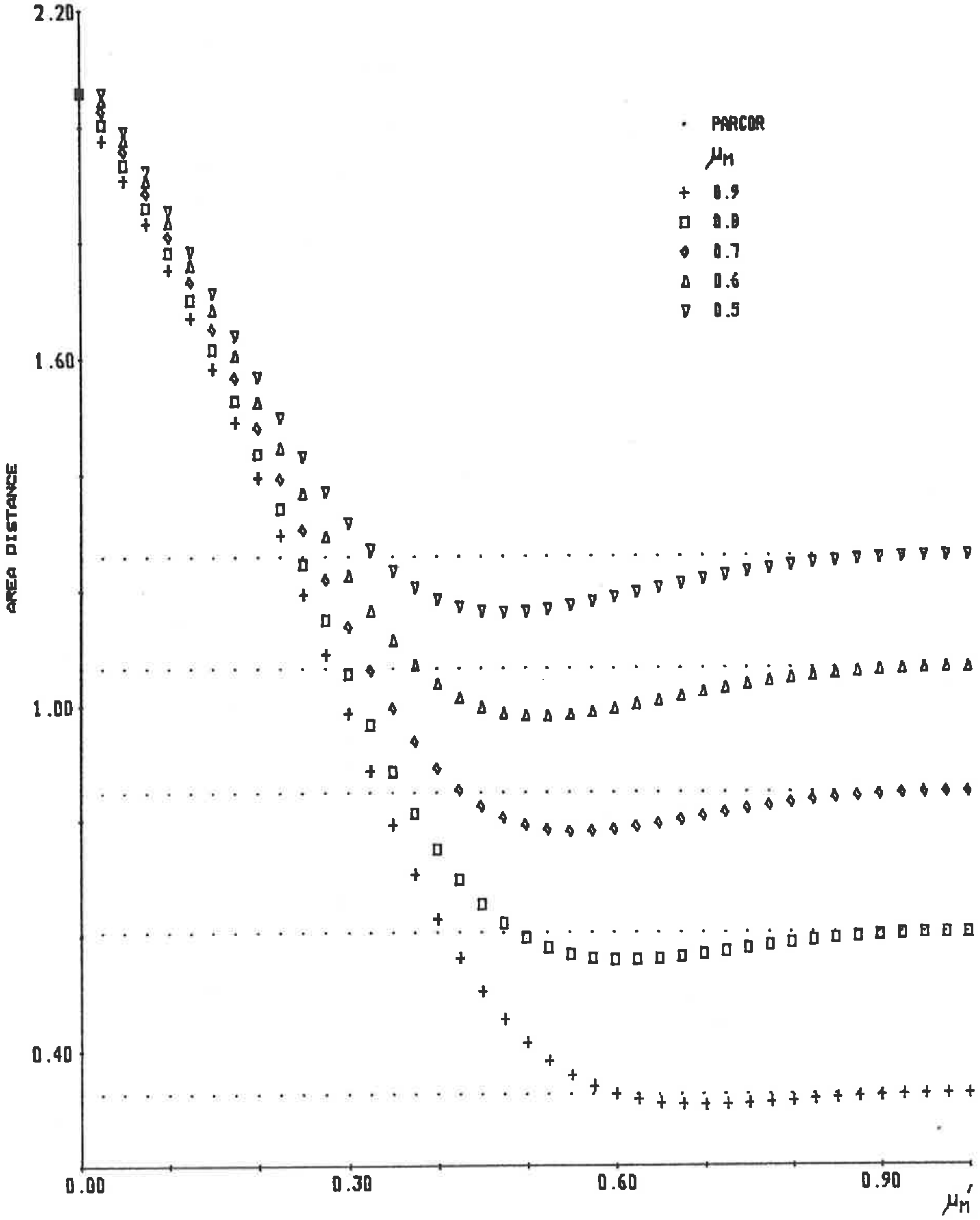


FIGURE 6.21: Area distances for a conventional parcor analysis and a parcor lossy termination analysis of synthetic speech for the vowel /e/.

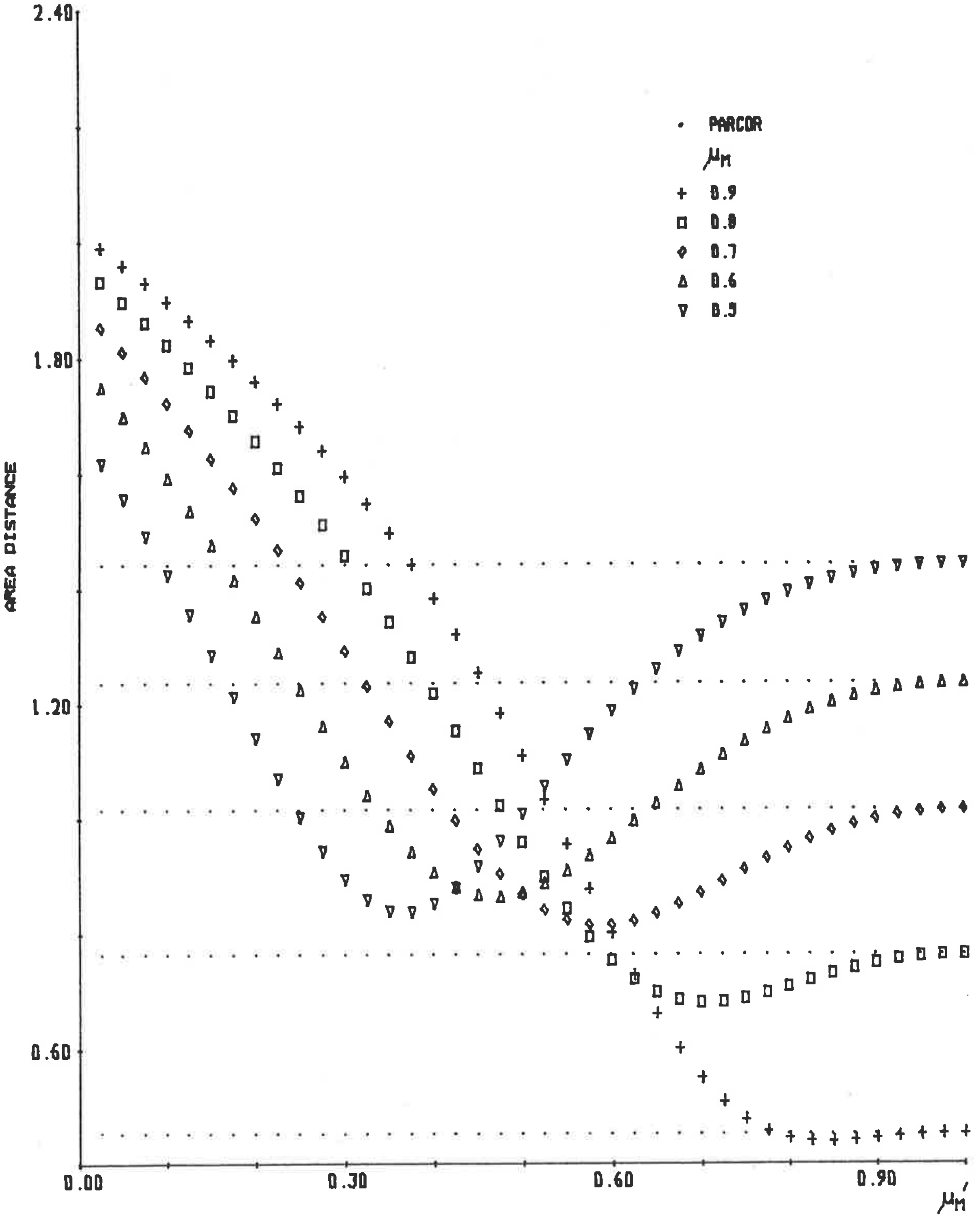


FIGURE 6.22: Area distances for a conventional parcor analysis and a parcor lossy termination analysis of synthetic speech for the vowel /i/.

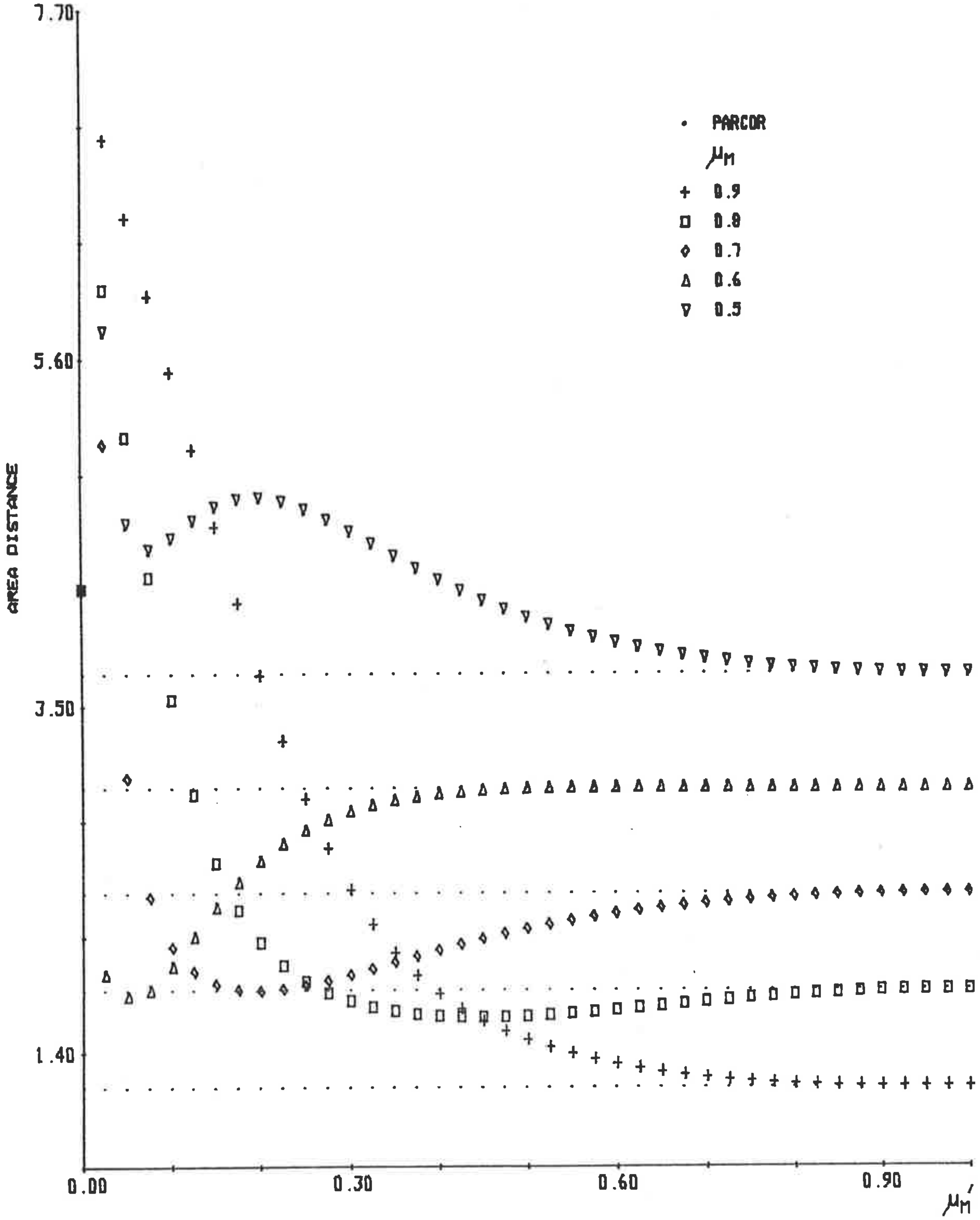


FIGURE 6.23: Area distances for a conventional parcor analysis and a parcor lossy termination analysis of synthetic speech for the vowel /o/.

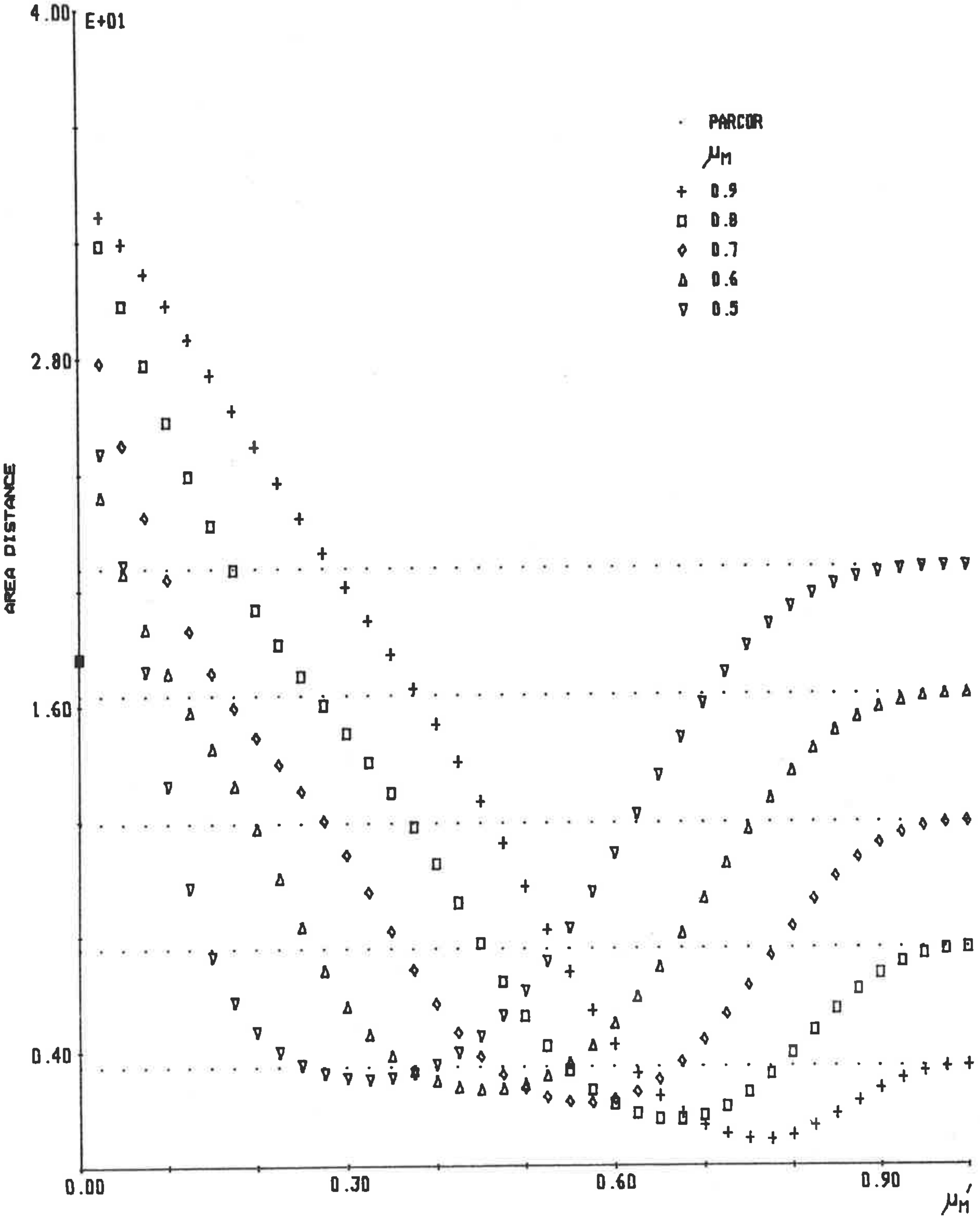


FIGURE 6.24: Area distances for a conventional parcor analysis and a parcor lossy termination analysis of synthetic speech for the vowel /u/.

area distance occurs for the conventional Parcor analysis as a function of μ_M'.

When μ_M is less than 0.9, a small reduction in area distances is observed in Figure 6.20 for the Parcor Lossy Termination analysis in comparison with a conventional Parcor analysis of synthetic speech for the vowel |a|, for μ_M and μ_M differing widely. When μ_M is equal to 0.9, then the small reduction in area distances achieved by a Parcor Lossy Termination analysis only occurs for a small range of μ_M near unity. In general, the reduction in area distances by the Parcor Lossy Termination analysis in comparison with a conventional Parcor analysis is small, which is consistent with the evaluation results presented in Section 6.5.2.1 for the synthetic speech of the vowel |a| (see Figure 6.15).

For all the values of μ_M considered in Figure 6.21, a reduction in area distances occurs for the Parcor Lossy Termination analysis, in comparison with a conventional Parcor analysis, of synthetic speech for the vowel |e| and for μ_M' and μ_M differing widely. In general, the range of μ_M' for which a reduction in area distances by the Parcor Lossy Termination analysis occurs increases as μ_M decreases, and the amount of area distance reduction increases as μ_M decreases. The larger reduction in area distances by the Parcor Lossy Termination analysis, in comparison with a conventional Parcor analysis, for synthetic speech of the vowel |e| than for the vowel |a| is consistent with the evaluation results presented in Section 6.5.2.1 (see Figures 6.15 and 6.16).

Figure 6.22 shows a large reduction in area distances for a Parcor Lossy Termination analysis, in comparison with a conventional Parcor analysis of synthetic speech for the vowel |i|. The range of μ_M' over which the reduction in area distance occurs increases markedly as the value of μ_M decreases and, therefore, μ_M' and μ_M may differ by large amounts, and a decrease in area distances still occurs. As the value of μ_M decreases, the amount by which an area distance reduction occurs increases markedly, and is much larger than that observed for synthetic speech of the vowels |a| and |e| as presented in Figures 6.20 and 6.21, respectively. This latter result is consistent with the much larger reduction in area distances that occurs for the vowel |i| than for the vowels |a| and |e| in the evaluation results presented in Section 6.5.2.1 (compare Figures 6.15, 6.16 and 6.17).

For μ_M equal to 0.9, Figure 6.23 shows that only a small reduction in area distances over a small range of μ_M' near unity occurs when the area distances of a Parcor Lossy Termination analysis are compared with those of a conventional Parcor analysis of synthetic speech for the vowel |o|. A much larger reduction in area distances occurs when μ_M is between 0.6 and 0.8, and for a very large range of μ_M'. However, for μ_M equal to 0.5, no reduction in area distances occurs for the Parcor Lossy Termination analysis in comparison with a conventional Parcor analysis. The above observations are consistent with the evaluation results presented in Section 6.5.2.1, i.e. Figure 6.18, where only a small reduction in area distances occurs for the Parcor Lossy Termination analysis when μ_M is greater than 0.5.

The evaluation results presented in Figure 6.24 for synthetic speech of the vowel |u| show that a large reduction of area distances occurs for the Parcor Lossy Termination analysis in comparison with a conventional Parcor analysis, for a wide range of μ_M'. Therefore, a reduction in area distances can be expected, although μ_M' and μ_M may differ markedly. As observed for all the evaluation results presented in this section, the range of μ_M' for which a reduction in area distances occurs for the Parcor Lossy Termination analysis of synthetic speech for the vowel |u| increases as μ_M decreases. The above observations are found to be consistent with the evaluation results presented in Section 6.5.2.1, by comparison of Figure 6.19 and 6.24.

The evaluation results presented in this section have shown that a reduction in area distances is achieved by the Parcor Lossy Termination analysis in comparison with a conventional Parcor analysis for μ_M' and μ_M differing from each other. The amount by which the value of μ_M' can differ from μ_M and still provide a reduction in area distances varies from one vowel sound to another, and depends on the value of μ_M but, in most cases, μ_M' and μ_M may differ widely. Notable exceptions are for μ_M equal to or greater than 0.9, where a reduction in area distances occurs only for a small range of μ_M' when μ_M' and μ_M are similar. In general, the smaller the value of μ_M, the larger the reduction in area distances, and the more widely μ_M' can differ from μ_M and still provide a reduction in area distances. Hence, it is concluded that the general reduction in area distances by the Parcor Lossy Termination analysis in comparison with a conventional Parcor analysis shown in Section 6.5.2.1 also occurs when μ_M' and μ_M are not exactly the same and do not differ by large amounts.

6.5.3 PARCOR LOSSY TERMINATION ANALYSIS OF REAL SPEECH

Of the four Lossy Termination analysis procedures defined in Section 6.5.2, the Parcor Lossy Termination analysis was found to consistently provide a reduction of area distance in comparison with the area distances of a conventional Parcor analysis when applied to synthetic speech for five vowel sounds. This section evaluates the Parcor Lossy Termination analysis with real speech waveforms.

Real speech waveforms were obtained for the five vowels |a|, |e|, |i|, |o| and |u| from seven Australian male speakers phonating the vowels in a |h-d| frame. A number of recording sessions, separated by at least one day, and not more than thirty days, were used, and at each session the vowels were spoken in a random order, which changed from session to session. The speech pressure waveform was measured with a condenser microphone, and the resulting electrical analog signal passed through a low pass filter, the cut-off frequency of which is 4.5 kilohertz. The filtered analog signal was then sampled at a frequency of 10 kilohertz, and stored on magnetic tape under the control of a minicomputer. A full description of the procedure and conditions under which the real speech waveforms are digitally recorded is presented in Appendix H.

To determine whether an improvement in acoustic tube/vocal tract shape recovery is achieved when using a Parcor Lossy Termination analysis in comparison with a conventional Parcor linear predictive analysis, the correct or original acoustic tube/vocal tract shape must be known. The measurement of vocal tract shapes is a very difficult and complex process, and the necessary equipment to perform such measurements was not available; therefore, the vocal tract shape is not known for any of the speech sounds. In order to perform an evaluation of the Parcor Lossy Termination analysis, the vocal tract shape for each of the five vowel sounds considered was approximated by the vocal tract shapes measured by FANT [1960] with X-ray photography.

Only the lip radiation or loss at the termination of the vocal tract is considered by the analysis process used to produce the results presented in this section, and so the other non-ideal properties of the vocal tract such as non-white excitation and vocal tract losses are ignored. Therefore, even if lip radiation is correctly accounted for, accurate vocal tract shape recovery cannot be expected. The non-ideal properties of the vocal tract not taken into account cloud the evaluation presented in this section, and so a careful interpretation of the evaluation results is necessary.

All analyses of the real speech waveforms are performed by firstly choosing 300 samples, i.e. a 30 msec time interval, of the particular vowel sound, and applying a Hamming window (BLACKMAN and TUKEY [1958], MARKEL [1971], MAKHOUL and WOLF [1972]). The windowed data is then pre-emphasized by -6 dB per octave to account for the zero of the radiation impedance (see Section 6.2), and then analysed by a Parcor Lossy Termination analysis and a conventional Parcor anlaysis, to determine recovered acoustic tube shapes. The evaluation results are presented as area distances between recovered acoustic tube shapes and the vocal tract shape measured by FANT [1960] for the corresponding vowel sound. These area distances are then plotted against the termination reflection coefficient, μ_M', used in the Parcor Lossy Termination analysis.

Figure 6.25 presents the area distances versus μ_M' for a Parcor Lossy Termination analysis of real speech for the vowel |a|. The area distances presented in Figure 6.25 show a gradual increase, except for one case, as the value of μ_M' decreases from unity. Since the area distances for μ_M' equal to unity are the same as the area distances resulting from a conventional Parcor analysis, in general a reduction of area distances is not achieved by using the Parcor Lossy Termination analysis in preference to a conventional Parcor analysis of real speech waveforms for the vowel |a|.

For real speech of the vowel |e|, the area distances for a Parcor Lossy Termination analysis versus μ_M' are presented in Figure 6.26. General trends are difficult to find in Figure 6.26, but a large reduction in area distances is observed in some cases. Otherwise, there exists a mixture of small increases and small decreases in area distances. Since a large area distance generally implies poor acoustic tube/vocal tract shape recovery, the decrease of the large area distances as μ_M' decreases from unity is sufficient incentive to use the Parcor Lossy Termination analysis in preference to a conventional Parcor analysis in the case of real speech for the vowel |e|.

The area distances versus μ_M' are presented in Figure 6.27 for a Parcor Lossy Termination analysis of real speech for the vowel |i|. A large decrease in area distances is observed in Figure 6.27 for all the speech analysed as the value of μ_M' decreases from unity. Therefore, there exists a significant advantage in using the Parcor Lossy Termination analysis in preference

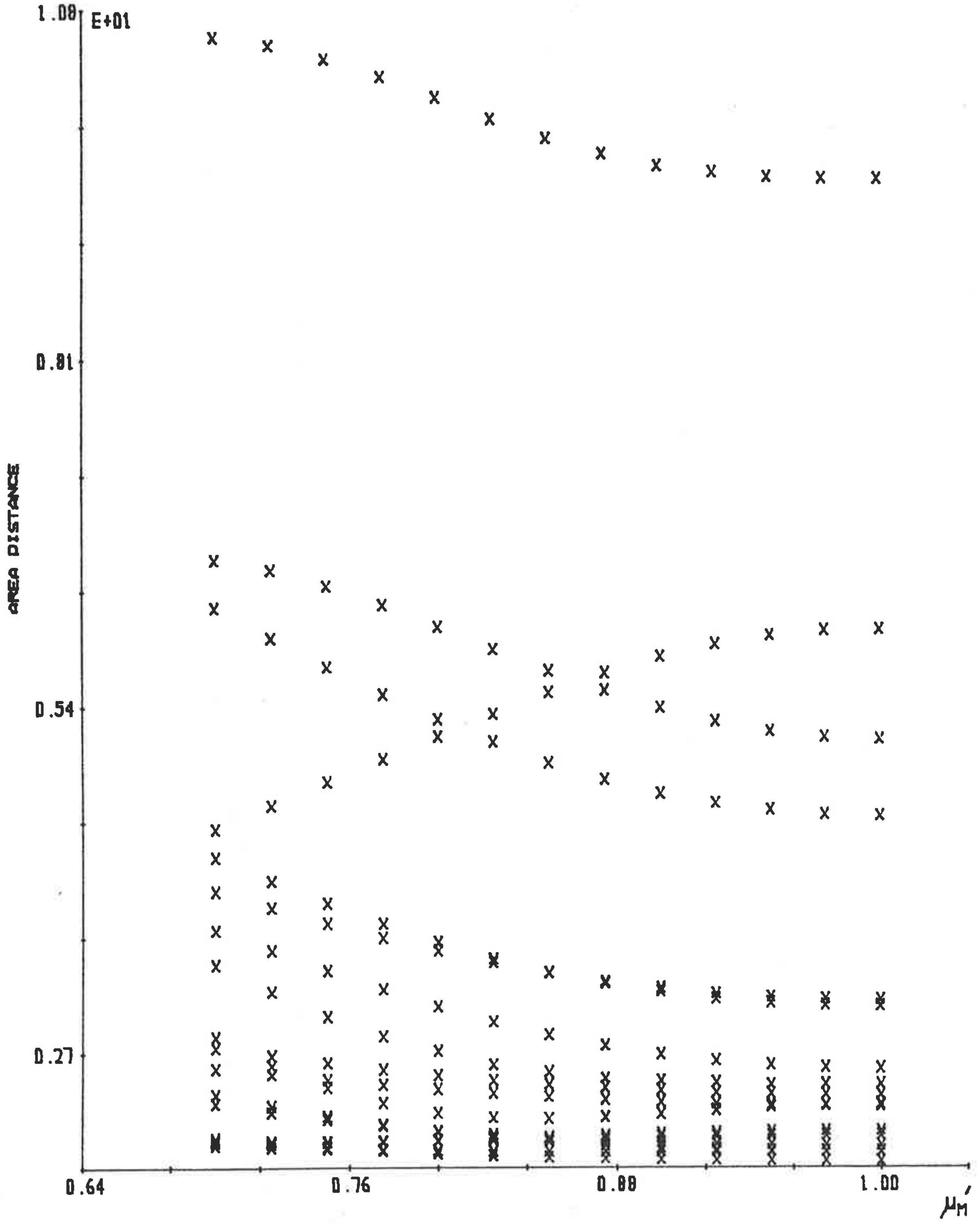


FIGURE 6.25: Area distances for a parcor lossy termination analysis of real speech for the vowel /a/.

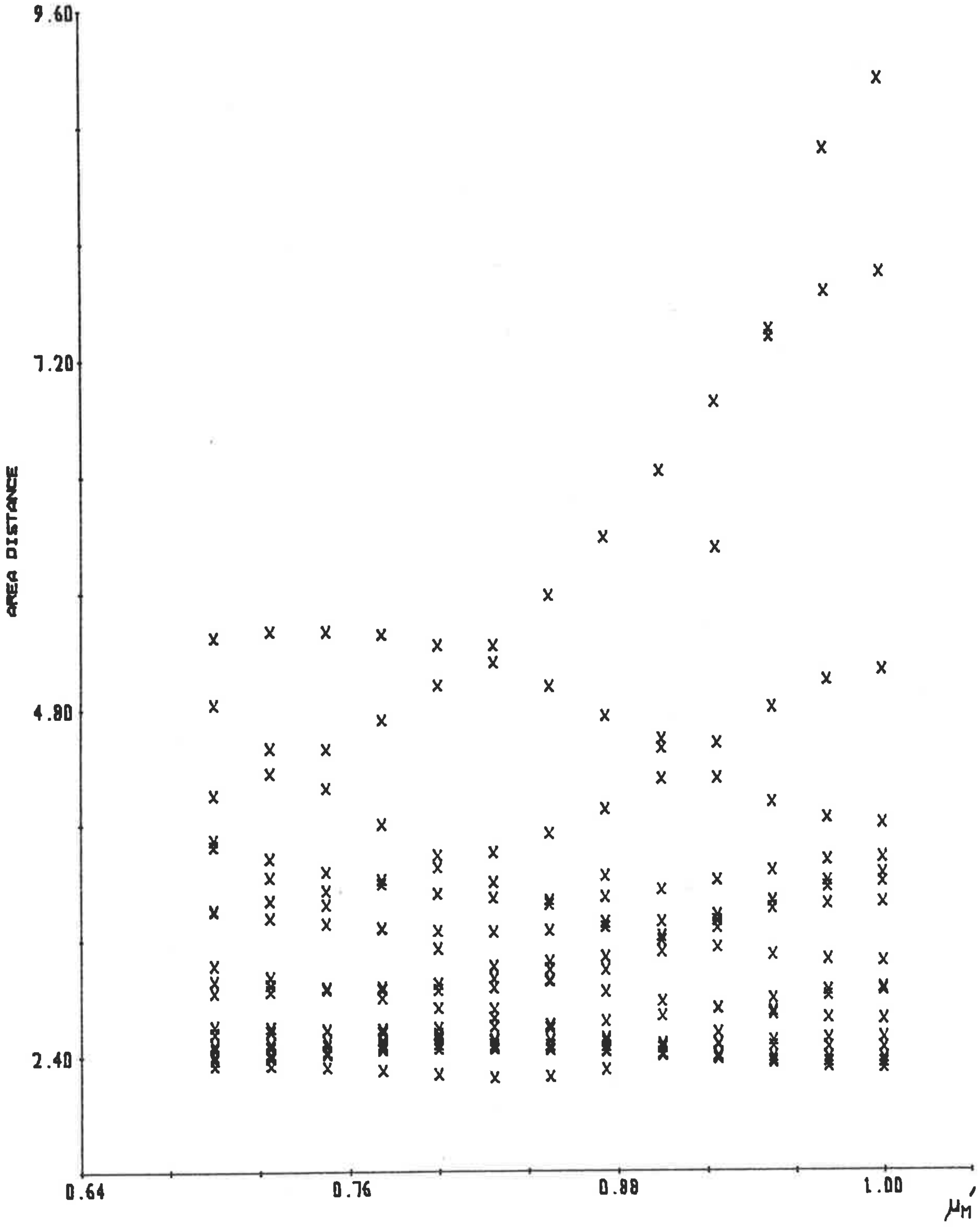


FIGURE 6.26: Area distances for a parcor lossy termination analysis of real speech for the vowel /e/.

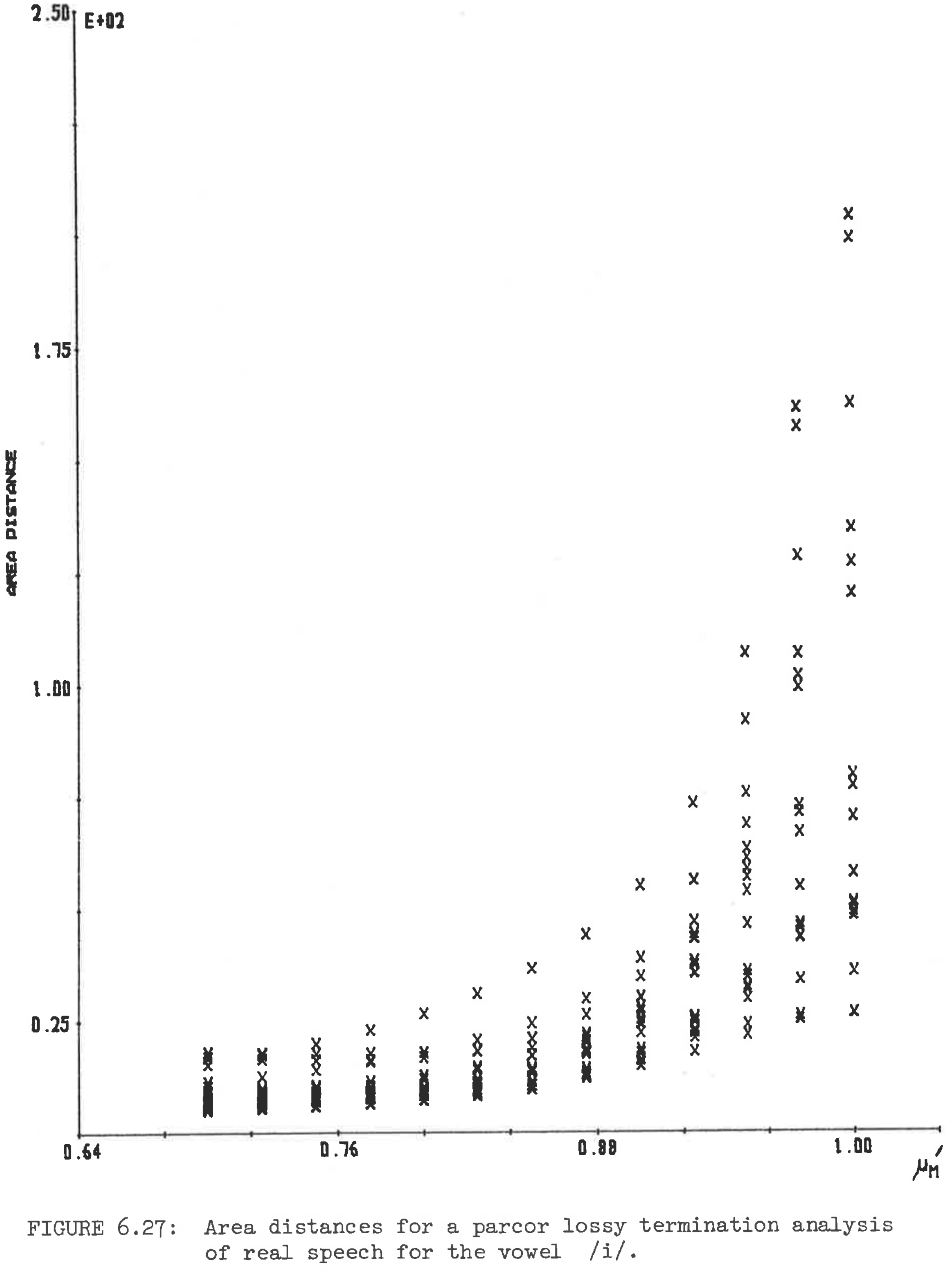


FIGURE 6.27: Area distances for a parcor lossy termination analysis of real speech for the vowel /i/.

to a conventional Parcor analysis of real speech for the vowel |i|, if vocal tract shape recovery is required.

Figure 6.28 presents the area distances versus μ_M' for a Parcor Lossy Termination analysis of real speech for the vowel |o|. Except for a few cases, a reduction in area distances is observed as μ_M' decreases from unity; therefore, in most cases, a reduction in area distances occurs for the Parcor Lossy Termination analysis in comparison with the area distances for a conventional Parcor analysis. Hence, there exists an advantage in using the Parcor Lossy Termination analysis for real speech of the vowel |o|, if vocal tract shape recovery is required.

The results of a Parcor Lossy Termination analysis of real speech for the vowel |u| are presented in Figure 6.29 as area distances versus μ_M'. As was the case for real speech of the vowel |i|, a large reduction in area distances is observed in Figure 6.29, as μ_M' decreases from unity. The larger the area distance when μ_M' is equal to unity, the larger the decrease in area distances as μ_M' decreases from unity. Hence, a significant improvement in area distance and, hence, vocal tract shapes occurs when the Parcor Lossy Termination analysis is used in preference to a conventional Parcor analysis of real speech for the vowel |u|.

The value of μ_M' that should be used in a Parcor Lossy Termination analysis of real speech to produce the smallest area distances is highly dependent on the real speech being analysed. For the vowels |i| and |u|, and for most cases of the vowel |o|, the smallest area distances occur when μ_M' is

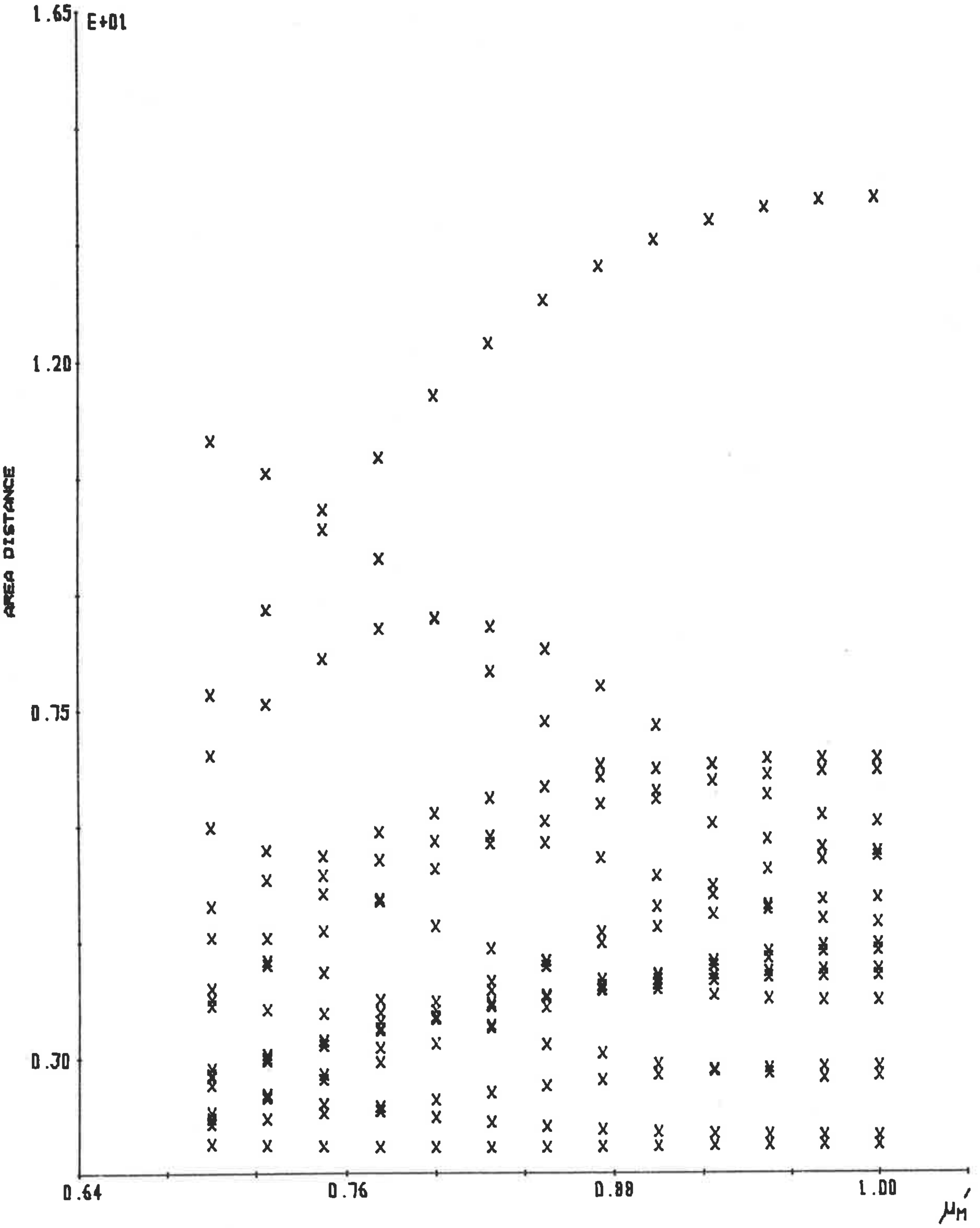


FIGURE 6.28: Area distances for a parcor lossy termination analysis of real speech for the vowel /o/.

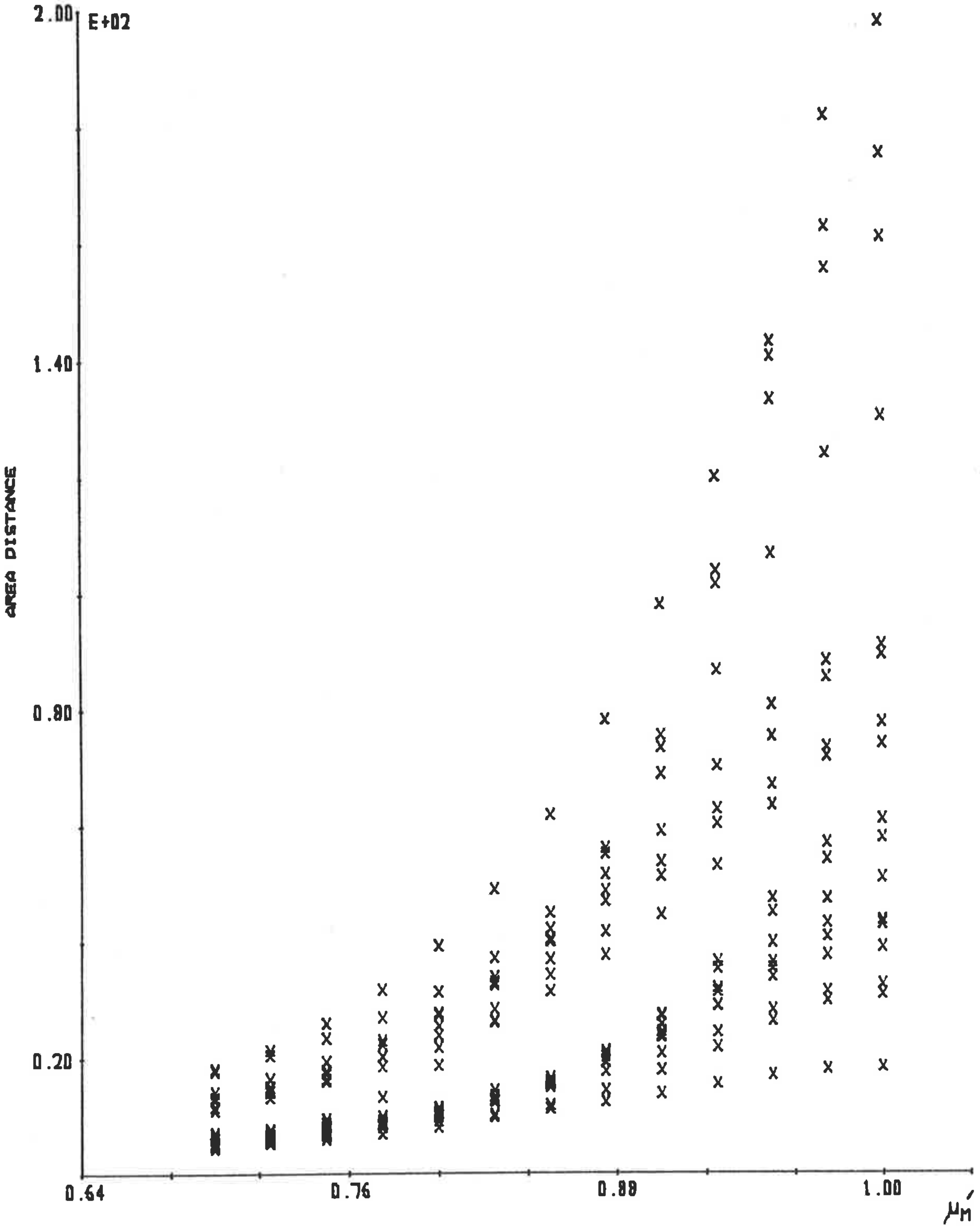


FIGURE 6.29: Area distances for a parcor lossy termination analysis of real speech for the vowel /u/.

less than 0.7. For the analysis of real speech of the vowel |e|, some of the smallest area distances occur for μ_M' less than 0.7, while others occur when μ_M' is approximately 0.85. For the vowel |a|, larger area distances occur as μ_M' decreases, and so μ_M' should be close to unity, to prevent poorer vocal tract recovery by a Parcor Lossy Termination analysis in comparison with a conventional Parcor analysis.

The significant improvement in acoustic tube/vocal tract shape recovery achieved by using a Parcor Lossy Termination analysis of real speech for the vowels |i| and |u| and, to a lesser degree, the vowel |o|, requires a compromise between the value of μ_M' that produces this large reduction in area distances and the value of μ_M' which causes increased area distances for other vowel sounds. A study of Figures 6.25 through to 6.29 results in choosing μ_M' equal to 0.8 as a suitable compromise, so that significant reductions in area distances occur for some vowels at the cost of only a small increase in area distances for other vowels.

The evaluations presented in this section for the Parcor Lossy Termination analysis of real speech waveforms have shown that there exists a significant advantage in using a Parcor Lossy Termination analysis in preference to a conventional Parcor analysis, when vocal tract shape recovery is the goal. Only one major exception was found, and that occurs for real speech of the vowel |a|, where a small increase in area distances occurs when the Parcor Lossy Termination analysis is used. The value of μ_M' which produces the best overall performance for the Parcor Lossy Termination analysis, as applied to real speech of the vowels |a|, |e|, |i|, |o| and |u|, is 0.8.

6.6 LOSSY TERMINATION ANALYSIS VIA A TRANSFER FUNCTION OF THE ACOUSTIC TUBE MODEL

Section 6.5 considered an analysis via the autocorrelation function of the output waveform from the acoustic tube model with a lossy termination. Although an improvement in acoustic tube shape recovery occurs in comparison with conventional linear predictive analysis, an accurate knowledge of the original termination reflection coefficient, μ_M, which was present for generation of the acoustic waveform is necessary for a large improvement in acoustic tube shape recovery to be achieved. This section defines and discusses an analysis procedure which uses a transfer function of the acoustic tube model and does not require the knowledge of μ_M to perform an analysis. The analysis procedure presented in this section is based on a paper presented by the author, BIELBY [1980].

A transfer function $C_i(z)$ is defined as the ratio of the forward to backward travelling volume velocities at the junction of the $(i+1)$th and ith acoustic tubes, i.e.

$$C_i(z) = \frac{U_i(z)}{z^{-1}\, V_i(z)} \qquad (6.111)$$

The transfer function $C_i(z)$ is assumed to consist of a numerator term $N_i(z)$ and a denomiator term $\mathcal{D}_i(z)$, i.e.

$$C_i(z) = \frac{N_i(z)}{\mathcal{D}_i(z)} \qquad (6.112)$$

and therefore, by combining Equations 6.111 and 6.112,

$$\mathcal{D}_i(z) U_i(z) = z^{-1}\, N_i(z) V_i(z) \qquad (6.113)$$

Using the junction equations between the $(i+1)$th and ith acoustic tubes (i.e. Equations 2.25 and 2.26), Equation 6.113 can be rewritten as

$$\mathcal{D}_{i+1}(z) = \mathcal{D}_i(z) + z^{-1}\mu_{i+1}N_i(z) \qquad (6.114a)$$

and

$$N_{i+1}(z) = \mu_{i+1}\mathcal{D}_i(z) + z^{-1}N_i(z) \qquad (6.114b)$$

which are recursive formulae for determining $\mathcal{D}_{i+1}(z)$ and $N_{i+1}(z)$ from $\mathcal{D}_i(z)$, $N_i(z)$ and μ_{i+1}. Using the definitions of the z transforms for $\mathcal{D}_i(z)$ and $N_i(z)$, i.e.

$$\mathcal{D}_i(z) = \sum_{j=0}^{\infty} d_j^{(i)} z^{-j} \qquad (6.115a)$$

and

$$N_i(z) = \sum_{j=0}^{\infty} n_j^{(i)} z^{-j} \qquad (6.115b)$$

the recursive formulae for the coefficients of $\mathcal{D}_i(z)$ and $N_i(z)$ then become

$$d_j^{(i+1)} = d_j^{(i)} + \mu_{i+1}n_{j-1}^{(i)} \qquad 0 \leqslant j \leqslant i \qquad (6.116)$$

and

$$n_j^{(i+1)} = \mu_{i+1}d_j^{(i)} + n_{j-1}^{(i)} \qquad 0 \leqslant j \leqslant i \qquad (6.117)$$

After excitation has ceased, the acoustic tube model requires that

$$\mathcal{D}_1(z) = 1 \qquad (6.118a)$$

and

$$N_1(z) = \mu_1 \tag{6.118b}$$

It has been found that the $n_j^{(i)}$ are redundant, and so the following recursive formulae can be defined for the $d_j^{(i+1)}$

$$d_j^{(i+1)} = d_j^{(i)} + \sum_{k=1}^{j} d_{j-k}^{(i-k)} \mu_{i-k+1}\mu_{i+1} \qquad 1 \leqslant j \leqslant i-1 \tag{6.119a}$$

$$d_i^{(i+1)} = \mu_1\mu_{i+1} \tag{6.119b}$$

$$d_0^{(i+1)} = d_0^{(i)} \tag{6.119c}$$

$$\text{with} \quad d_j^{(i)} \neq 0 \quad \text{for} \quad 0 \leqslant j \leqslant i-1 \tag{6.119d}$$

$$\text{and, initially,} \quad d_0^{(1)} = 1 \tag{6.119e}$$

The Equations 6.119 permit $d_j^{(i)}$ to be written in terms of previous $d_j^{(i-1)}$ and the reflection coefficients μ_i, or just in terms of the reflection coefficients as shown in Appendix G. Similar expressions are derived by ABDEL MONEN [1977] for a set of commensurate lossless transmission lines with a general termination reflection coefficient.

At the lossy termination of the acoustic tubes

$$\mathcal{D}_M(z)U_M(z) = 0 \tag{6.120}$$

which can be written as

$$\sum_{j=1}^{M-1} d_j^{(M)} U_M(n-j) = -U_M(n) \qquad n = M, M+1, \ldots, 2(M-1) \tag{6.121}$$

in the time domain. Equation 6.121 is recognised as the difference equation for an all pole filter with coefficients $d_j^{(M)}$. Equation 6.121 is a set of $(M-1)$ linear simultaneous equations in $(M-1)$ unknowns, i.e. the $d_j^{(M)}$'s, and since $U_M(n)$ is the known radiated volume velocity, then the values of $d_j^{(M)}$ are determined by solving Equation 6.121 with a conventional linear simultaneous equation solving algorithm.

Once the values of $d_j^{(M)}$ are known, the set of non-linear equations of $d_j^{(M)}$ in terms of the reflection coefficients between the acoustic tubes (see Appendix G) can be solved to determine the values of those reflection coefficients. Solving the non-linear equations produces an infinity of solutions in the present form, since there are M unknown reflection coefficients and only $M-1$ equations (the extra unknown is μ_M, which is assumed as unity in a conventional linear predictive analysis). If the number of unknowns is to be made equal to the number of equations, either the value of any one reflection coefficient must be chosen to produce $M-1$ unknowns in $M-1$ equations, or an extra equation must be introduced to produce M unknowns in M equations.

Observations of the human vocal tract show that, in regions near the glottis, large changes in cross-sectional area do not occur over short distances. Hence, in cases where the sampling frequency is relatively large, i.e. at least 10 kilohertz, then the reflection coefficients for an acoustic tube model of the vocal tract near the glottis are small. Therefore, the number of unknowns can be made equal to the number of equations by constraining the recovered acoustic tube shape near the glottis. In effect, the constraint ensures the recovered acoustic tube shape is close to the physical construction of the vocal tract near the glottis. Conventional linear predictive analyses do not take into account the physical realizability of the recovered acoustic tube shape and, hence, in many cases, no resemblance occurs between a physical vocal tract and the recovered acoustic tube shape.

The constraint applied to the recovered acoustic tube shape is referred to as the physical constraint, and the region of the vocal tract (and recovered acoustic tube shape) to which the physical constraint is applied is called the constraint region. To reduce the number of unknowns so that there are *M*-1 unknowns in *M*-1 equations, the physicial constraint is implemented by choosing the value of one reflection coefficient in the constraint region. Alternatively, the number of equations can be increased so that there are *M* unknowns in *M* equations by implementing the physical constraint as a CONSTRAINT EQUATION.

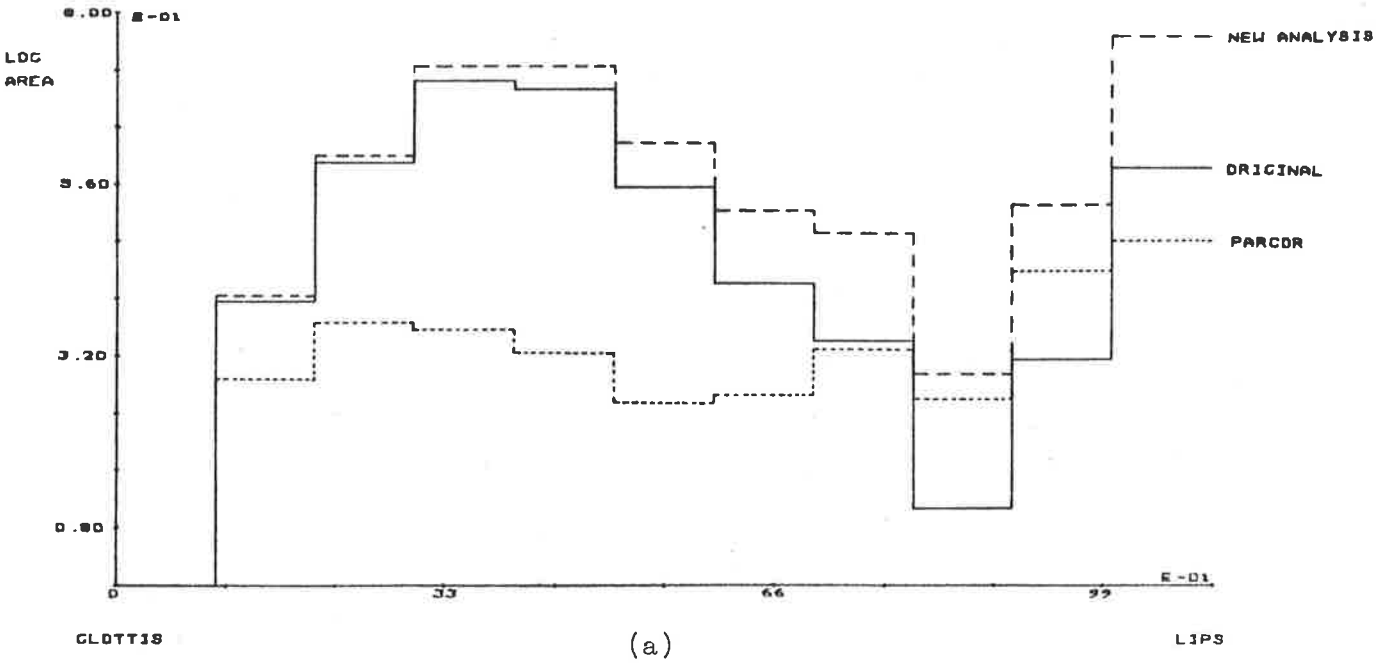
Once the number of unknowns is equal to the number of equations, a solution procedure is performed to determine the unknown reflection coefficients and, hence, define a recovered acoustic tube shape. The equations are non-linear in the unknown reflection coefficients, and so a non-linear solution algorithm must be used. A major problem with solving non-linear equations is the uncertainty of the rate of convergence to a solution. The rate of convergence is dependent on the starting point or initial guess of the solution and the flatness of the *M*-dimensional surface defined by the non-linear equations. A slow convergence to a solution is a distinct disadvantage for real time vocal tract shape recovery from speech waveforms.

For each different speech waveform, a different set of non-linear equations needs to be solved; therefore, the rate of convergence changes for analysis of different speech waveforms. There also exists the possibility that the non-linear solution algorithm may diverge and, hence, not recover an acoustic tube shape. Convergence to local solutions instead of a global solution is also possible, depending on the type of non-linear solution algorithm used and the starting point or initial guess of the solution.

One procedure for implementing the physical constraint is to choose the value of a reflection coefficient in the constraint region. The constraint region of the vocal tract was defined as that region near the glottis where only small changes in cross-sectional area occur; therefore, the reflection coefficients in the constraint region should be small. In practice, therefore, it is convenient to choose the value of a reflection coefficient in the constraint region to be zero. Using synthetic speech for the vowels |e| and |i|, an improvement in original acoustic tube shape recovery is shown in Figure 6.30 when using the above analysis procedure with the reflection coefficient μ_4 set to zero. The generation of the synthetic speech assumed a lossy termination, and the termination reflection coefficient, μ_M, is recovered accurately.

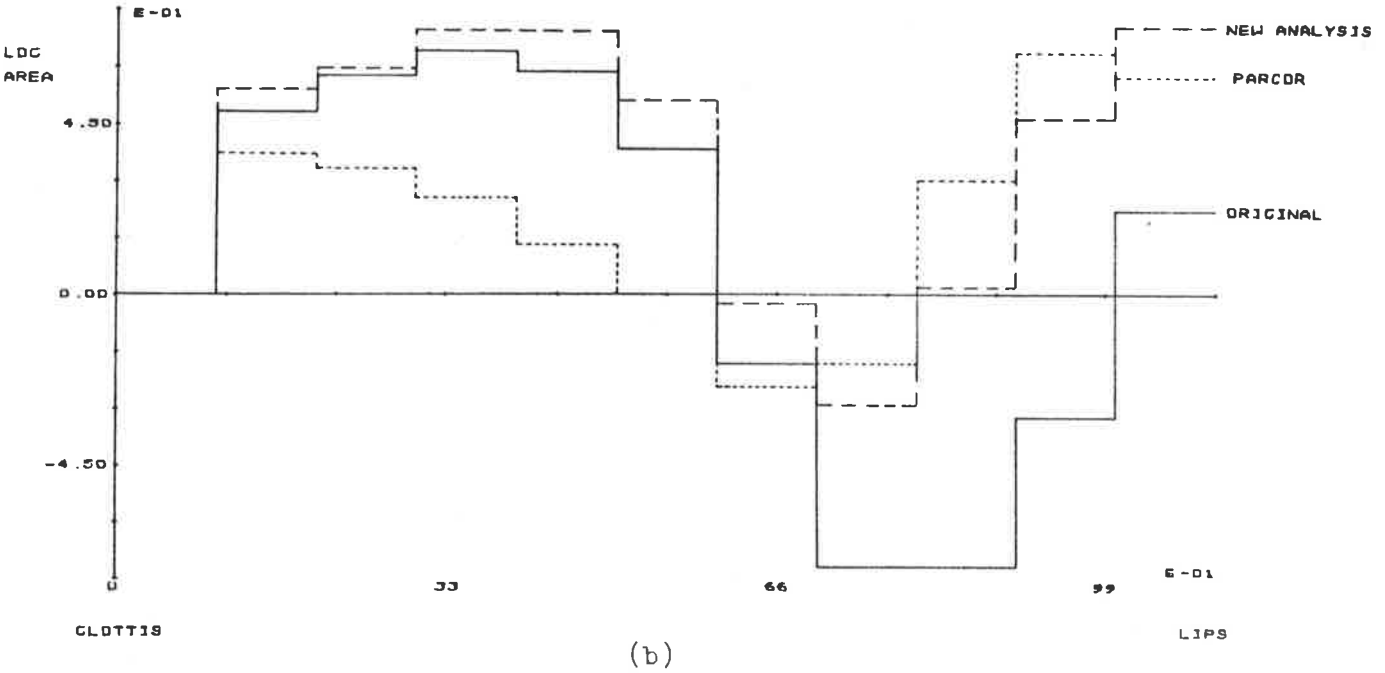
The accuracy of the recovered acoustic tube shape when assuming a reflection coefficient in the constraint region is zero is highly dependent on the choice of the constraint region and size of the cross-sectional area changes in the constraint region. A more general and realistic constraint to apply in the constraint

```
ORIGINAL REFLECTION COEFFICIENTS
    -0.425  -0.221  -0.130  +0.014  +0.156  +0.154  +0.091  +0.262  -0.235  -0.297  +0.750
CALCULATED REFLECTION COEFFICIENTS - NEW ANALYSIS
    -0.433  -0.222  -0.142  +0.000  +0.123  +0.109  +0.036  +0.221  -0.265  -0.263  +0.736
CALCULATED REFLECTION COEFFICIENTS - PARCOR
    -0.318  -0.089  +0.011  +0.037  +0.078  -0.011  -0.072  +0.077  -0.200  -0.048  +1.000
```



(a)

```
ORIGINAL REFLECTION COEFFICIENTS
    -0.503  -0.107  -0.072  +0.060  +0.228  +0.573  +0.549  +0.000  -0.425  -0.555  +0.800
CALCULATED REFLECTION COEFFICIENTS - NEW ANALYSIS
    -0.551  -0.064  -0.113  +0.000  +0.207  +0.549  +0.296  -0.341  -0.470  -0.267  +0.730
CALCULATED REFLECTION COEFFICIENTS - PARCOR
    -0.403  +0.043  +0.088  +0.142  +0.151  +0.273  -0.071  -0.304  -0.366  +0.072  +1.000
```



(b)

FIGURE 6.30: Comparison of the original acoustic tube shape with the acoustic tube shapes recovered by a parcor analysis and the new analysis with the constraint $\mu_4=0$, for (a) the vowel /e/ and (b) the vowel /i/

region is to ensure the magnitudes of the reflection coefficients are less than a certain value. Therefore, a constraint equation is defined as

$$\sum_{i} |\mu_i| \leqslant C \tag{6.122}$$

where C is a constant and i is restricted to the constraint region. Many other constraint equations may be formulated, but only the one defined above has been used to evaluate the accuracy of recovering acoustic tube shapes when a constraint equation is used.

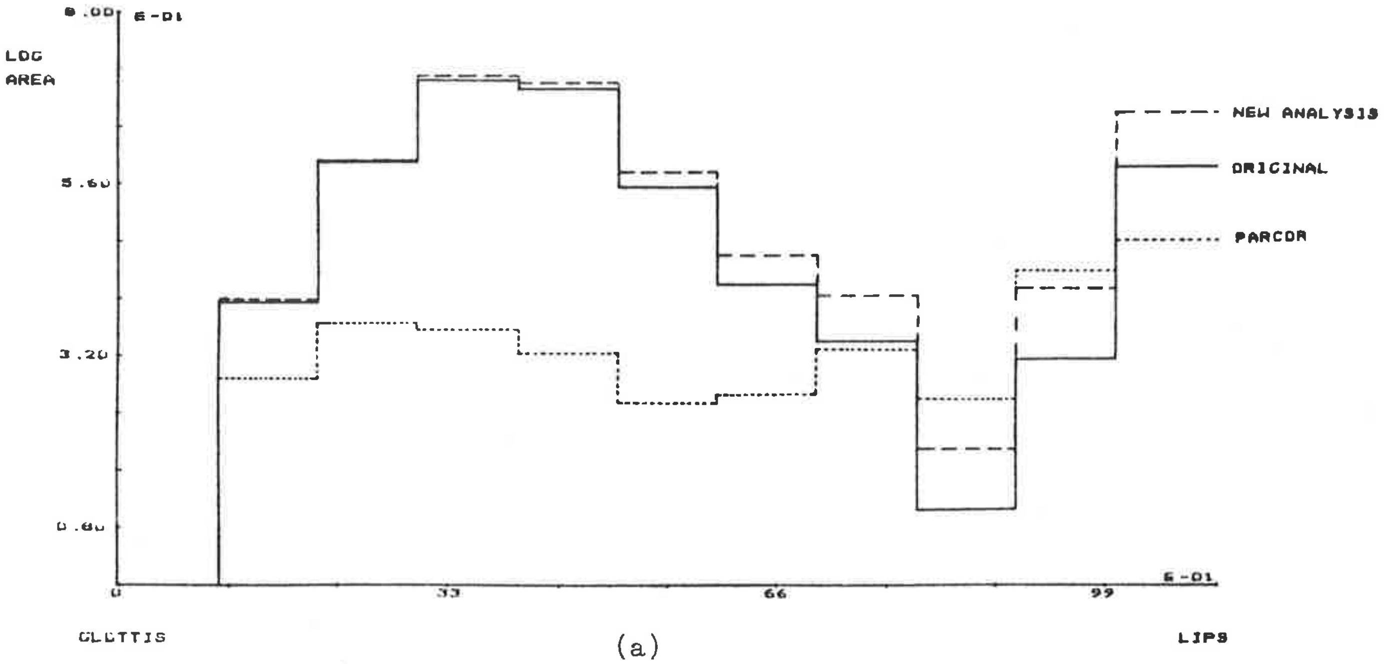
Using synthetic speech for the vowels |e| and |i|, Figure 6.31 shows that, by using a constraint equation

$$|\mu_3| + |\mu_4| \leqslant \sqrt{0.01} \tag{6.123}$$

improved acoustic tube shape recovery occurs in comparison with a conventional Parcor analysis. For each vowel, the termination reflection coefficient, μ_M, is recovered accurately. A comparison of the acoustic tube shapes recovered by choosing a reflection coefficient in the constraint region and using a constraint equation in the constraint region, i.e. comparing Figures 6.30 and 6.31, shows that the latter provides the best acoustic tube shape recovery. Investigations have shown that, in general, more accurate acoustic tube shape recovery is achieved by using a constraint equation.

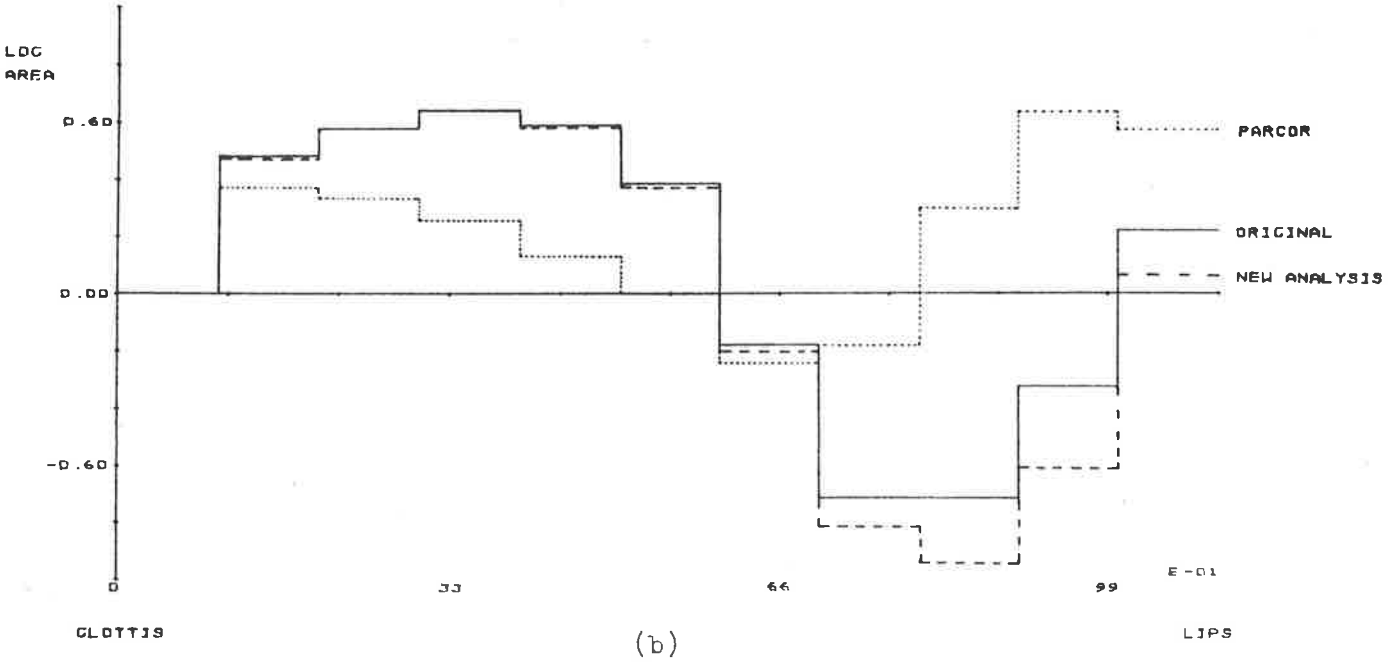
A major advantage of a constraint equation is that it may be applied to any region, and not necessarily to a constraint region. It is possible to formulate and use a constraint equation to restrict the size of any recovered reflection coefficient. An example

```
ORIGINAL REFLECTION COEFFICIENTS
    -0.425   -0.221   -0.130   +0.014   +0.156   +0.154   +0.091   +0.262   -0.235   -0.297   +0.750
CALCULATED REFLECTION COEFFICIENTS - NEW ANALYSIS
    -0.428   -0.219   -0.133   +0.010   +0.142   +0.132   +0.063   +0.241   -0.250   -0.274   +0.743
CALCULATED REFLECTION COEFFICIENTS - PARCOR
    -0.318   -0.089   +0.011   +0.037   +0.078   -0.011   -0.072   +0.077   -0.200   -0.048   +1.000
```



(a)

```
ORIGINAL REFLECTION COEFFICIENTS
    -0.503   -0.107   -0.072   +0.060   +0.228   +0.573   +0.549   +0.000   -0.475   -0.555   +0.800
CALCULATED REFLECTION COEFFICIENTS - NEW ANALYSIS
    -0.495   -0.118   -0.069   +0.067   +0.231   +0.580   +0.608   +0.145   -0.366   -0.651   +0.813
CALCULATED REFLECTION COEFFICIENTS - PARCOR
    -0.403   +0.043   +0.088   +0.142   +0.151   +0.273   -0.071   -0.504   -0.366   +0.072   +1.000
```



(b)

FIGURE 6.31: Comparison of the original acoustic tube shape with the acoustic tube shapes recovered by a parcor analysis and the new analysis with the constraint equation $|\mu_3| + |\mu_4| \le \sqrt{0.01}$, for (a) the vowel /e/ and (b) the vowel /i/

of using a constraint equation to restrict the size of a reflection coefficient is presented in Figure 6.32. The acoustic waveform used to generate Figure 6.32 was the measured output radiated waveform from three real acoustic tubes (see BIELBY [1980]). The constraint equation, in this case, has the form

$$|\mu_1| \leqslant \sqrt{0.12} \qquad (6.124)$$

an improved acoustic tube shape recovery is observed in Figure 6.32 when this constraint equation is used.

Therefore, it has been shown that a transfer function approach to analysis, where there is a lossy termination of acoustic tubes, can provide improved acoustic tube shape recovery. To balance the generalization of a lossy termination, a physical constraint is imposed on the recovered acoustic tube shape. The best improvements in acoustic tube shape recovery were obtained when the physical constraint was implemented by a constraint equation. The major advantage of the analysis process considered in this section in comparison with the autocorrelation method presented in Section 6.5 is that the termination reflection coefficient does not need to be specified or, in real situations, guessed, before an analysis takes place. Using a constraint equation based on physical vocal tract requirements also ensures realistic recovered vocal tract shapes, which is not necessarily the case for the autocorrelation analyses presented in Section 6.5.

Before the method described in this section can be used for general speech analysis, a number of its disadvantages must be overcome. A set of simultaneous non-linear equations must be solved, which requires complex solution algorithms and, hence,

```
    -0.3333                     +0.3333
CALCULATED REFLECTION COEFFICIENTS - NEW ANALYSIS
    -0.3138                     +0.3441                    -0.8136
CALCULATED REFLECTION COEFFICIENTS - PARCOR
    -0.2183                     +0.2451                    -1.0000
```

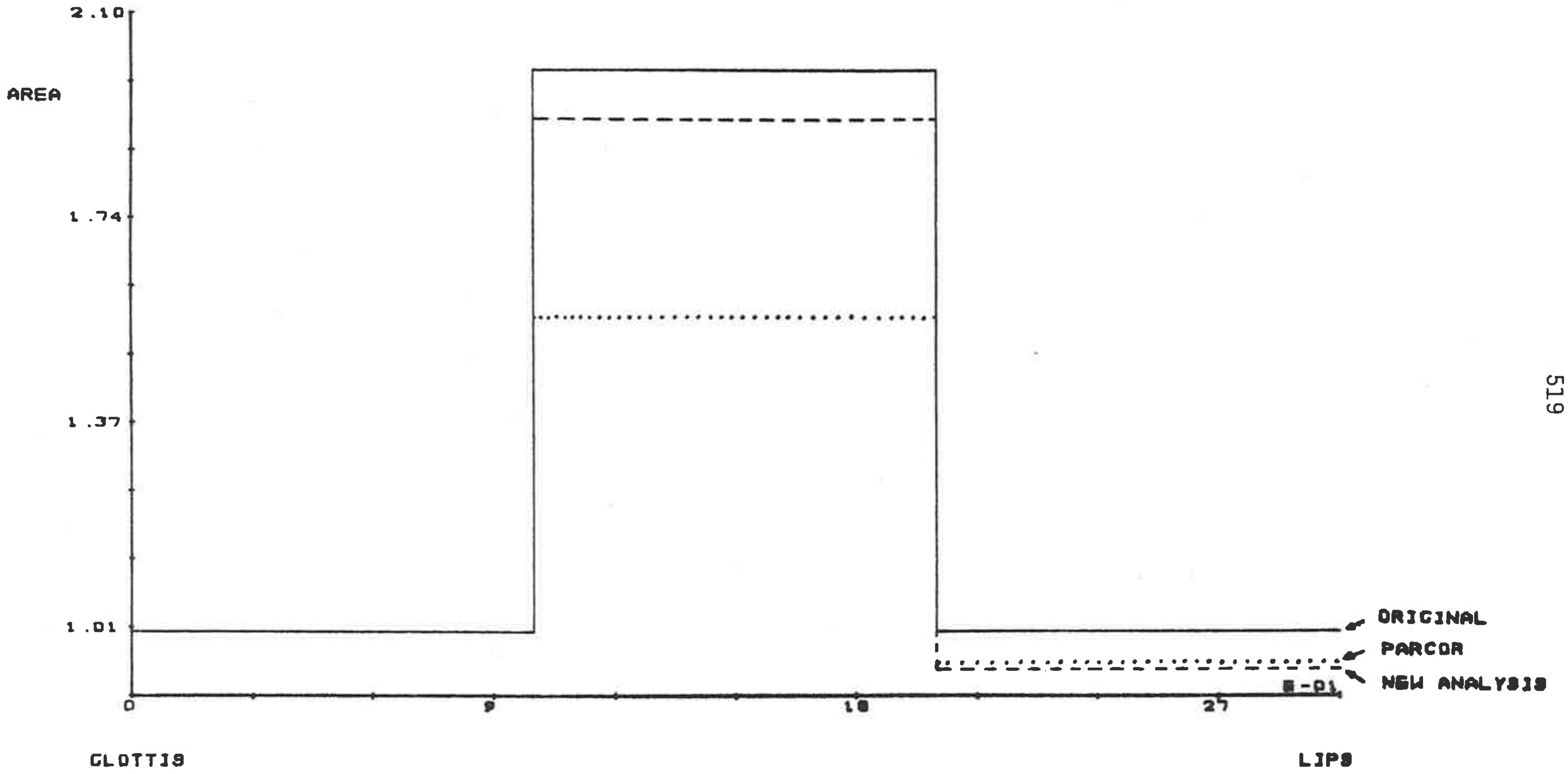


FIGURE 6.32: Comparison of the original acoustic tube shape with the acoustic tube shape recovered by a parcor analysis and the new analysis with the constraint

$$|\mu_1| \leq \sqrt{0.12}$$

large numbers of arithmetic computations. The solution algorithms are generally iterative, and the number of iterations necessary to obtain a solution varies for each speech waveform analysed. Hence, the solution or recovered acoustic tube shape may require a large amount of computer time, so that real time applications may not be possible. To overcome this difficulty, specialized solution algorithms need to be developed for the particular form of the non-linear equations.

Another problem which relates to the solution of the non-linear equations is the possibility of convergence to a local solution rather than a global solution. To a certain degree, this problem can be overcome by tightening up the constraint equation, but this conflicts with the requirement that the constraint equation should allow for the wide range of physical shapes of the real vocal tract. Therefore, new constraint equation forms which consider a restriction of the recovered acoustic tube shape to physically realizable shapes and also enhance convergence of the non-linear equation solution algorithms to global solutions are required.

In conclusion, the analysis process discussed in this section has been shown to provide improved acoustic tube/vocal tract recovery from real speech waveforms, but refinements are necessary for it to be an analysis procedure which can be used in place of existing analysis procedures.

6.7 SUMMARY

This chapter has considered the effects of a lossy termination of acoustic tubes on acoustic tube shape recovery by conventional linear predictive analysis procedures. Conventional procedures, and a number of new analysis techniques, presented in the latter half of this chapter, were evaluated to determine their effectiveness at recovering the original acoustic tube shape when a lossy termination is present.

The conventional radiation impedance for radiation of acoustic waveforms from the open end of a lossless acoustic tube was presented in detail, and then simplified by using the physical geometry of the vocal tract and the limitations of lip opening. This resulted in the radiation impedance at the open end of an acoustic tube being modelled by a single zero at $z = 1$, and a magnitude term. A comparison of the simplified radiation impedance with the more exact conventional radiation impedance showed that significant departures between the two only occurred at high frequencies, when the plane wave propagation assumption breaks down in the vocal tract.

The conventional approach to removing the radiation effects from the recovered acoustic tube shape is to apply a -6 dB per octave pre-emphasis before a linear predictive analysis is performed. From the model of the radiation impedance presented in this chapter, the -6 dB per octave pre-emphasis accounts for the zero at $z = 1$, but not the magnitude term. Evaluations of the effectiveness of a -6 dB per octave pre-emphasis to remove radiation effects from the recovered acoustic tube shape showed that, in some cases, good acoustic tube shape recovery occurs, especially when the magnitude term of the simplified radiation impedance is small. When the magnitude term is large, poor acoustic tube shape recovery occurs; therefore, the magnitude term must be considered in an analysis process if accurate acoustic tube shape recovery is to be achieved.

Further investigations permitted the magnitude term of the simplified radiation impedance to be defined in terms of the termination reflection coefficient, μ_M, of the acoustic tube model. A non-zero magnitude term, i.e. a finite radiation impedance, implies a non-unity termination reflection coefficient, μ_M, (linear predictive analyses assume μ_M is unity); therefore, improved acoustic tube shape recovery requires an analysis procedure which permits a non-unity value for μ_M.

A new autocorrelation analysis procedure was presented which permits a non-unity termination reflection coefficient, μ_M, i.e. a loss can occur at the termination of the original acoustic tubes. This new analysis procedure does not specify the manner in which a reflection coefficient between adjacent acoustic tubes is identified, but permits autocorrelation and crosscorrelation functions to be determined in an acoustic tube from autocorrelations and crosscorrelation functions in other acoustic tubes. Therefore, not only can this analysis process be used with the methods of determining reflection coefficients, μ_M, presented in this chapter, but it can also be used with any new methods of determining μ_M that may be developed in the future. The autocorrelation analysis procedure is presented in a computationally efficient form, which requires a similar number of computations to those required by the lattice formulation of linear prediction.

A study of the junction equations of the acoustic tube model showed that the determination of the reflection coefficient between two adjacent acoustic tubes requires the knowledge of the cross junction correlations between acoustic waveforms in each of the acoustic tubes. During an analysis process, only the acoustic waveforms in one of the acoustic tubes are known and, hence, the cross junction correlations are unknown; therefore, the junction between two acoustic tubes cannot be identified. If the termination of the acoustic tubes is lossless, then μ_M is unity and the cross junction correlations are always zero, i.e. the Cross Junction Correlation Theorem. Hence, for a lossless termination of acoustic tubes, as assumed by all conventional linear predictive procedures, the reflection coefficient between adjacent acoustic tubes can be determined from the acoustic waveforms in only one of the acoustic tubes; therefore, an analysis can take place.

When a non-unity termination reflection coefficient is assumed, then the value of the cross junction correlations must be estimated in order that an acoustic tube model may be identified. Empirical investigations have shown that a reasonable estimate of the cross junction correlations is that they are smaller than certain autocorrelations and cross tube correlations of the acoustic waveforms in the acoustic tube which has been identified. This estimation permits the definition of many methods to determine a value for μ_i, but only four were considered here which, when coupled with the new autocorrelation analysis procedure, resulted in the definition of four Lossy Termination autocorrelation analysis procedures.

The four Lossy Termination analyses were compared with a conventional Parcor linear predictive analysis by analysing synthetic speech for the five vowels |a|, |e|, |i|, |o| and |u|. Initially, each of the Lossy Termination analyses used the same value of μ_M as that used to generate the synthetic speech waveforms, and only one, namely the Parcor Lossy Termination analysis, was found to provide a consistent reduction in area distances, as compared with a conventional Parcor analysis, for a wide range of μ_M.

In general, the value of the termination reflection coefficient, μ_M, is not known, and so an estimate must be used when performing a Parcor Lossy Termination analysis. Evaluations were performed to determine the effect of using a different μ_M in the Parcor Lossy Termination analysis from that used to generate the synthetic speech. These evaluations showed that a reduction in area distances occurs for a Parcor Lossy Termination analysis in comparison with a conventional Parcor analysis, for the μ_M used in the Parcor Lossy Termination analysis differing widely from that used to generate the synthetic speech.

Evaluations of the Parcor Lossy Termination analysis were performed with real speech of the five vowels |a|, |e|, |i|, |o| and |u|. Reductions in area distances were found to be both small and large, with the exception of the vowel |a|, where a small increase in area distances occurs, when the Parcor Lossy Termination analysis is compared with a conventional Parcor analysis of the real speech waveforms. Therefore, in general, there exists an advantage in using a Parcor Lossy Termination analysis in preference to a conventional Parcor analysis if vocal tract shape recovery is desired. The value of μ_M which produces the smallest area distances for a Parcor Lossy Termination analysis of all the real speech for the five vowels was determined to be 0.8.

An analysis procedure permitting a non-unity value for μ_M, i.e. a loss at the termination of the acoustic tube model, and based on a transfer function of the acoustic tube model, was presented. This analysis procedure, instead of using an estimate of μ_M, places constraints, derived from physical restrictions of the vocal tract, on the recovered acoustic tube shape. The potential of the analysis procedure was presented, and accurate acoustic tube shape recovery was shown to occur. However, problems with convergence and large numbers of complex computations prevent the analysis procedure from being used to analyse speech waveforms in real time applications.

CHAPTER 7

AN ANALYSIS PROCESS FOR SPEECH

7.1 INTRODUCTION

Conventional speech analysis procedures, such as linear prediction, make assumptions about the properties of the vocal tract and speech production mechanism which cause poor vocal tract shape recovery. This thesis has defined and evaluated new analysis procedures which have been shown to provide improved acoustic tube/ vocal tract shape recovery in comparison with existing analysis procedures. This chapter combines two of these new analysis procedures into a single speech analysis procedure.

Chapter 4 considered the non-ideal glottal pulse excitation of the vocal tract and defined a new pre-emphasis filter, called the two/three adaptive pre-emphasis filter, which was shown, in Chapter 5, to produce improved acoustic tube shape recovery in comparison with existing pre-emphasis techniques. A new autocorrelation analysis procedure, called the Parcor Lossy Termination analysis, was defined and evaluated in Chapter 6 and shown to produce improved acoustic tube shape recovery in comparison with a conventional Parcor analysis when the acoustic tubes have a lossy termination. The two/three adaptive pre-emphasis filter and the Parcor Lossy Termination analysis procedure are combined into a single new speech analysis procedure in Section 7.2.

The new speech analysis procedure is evaluated with synthetic speech data in Section 7.3, and with real speech data in Section 7.4. The evaluation of the new speech analysis procedure is performed by a comparison of area distances between original acoustic tube shapes and those recovered by the new speech analysis procedure and conventional linear predictive analyses.

7.2 COMBINATION OF PROCEDURES FOR GLOTTAL PULSE EXCITATION AND A LOSSY TERMINATION

This section combines the two/three adaptive pre-emphasis filter with the Parcor Lossy Termination analysis procedure to form a new speech analysis procedure. Although the two/three adaptive pre-emphasis filter has been defined in detail in Chapter 4, and the Parcor Lossy Termination analysis in Chapter 6, a brief review of each is presented here so that this chapter is self-contained.

The two/three adaptive pre-emphasis filter consists of two separate parts, one defined by the parameter α which corrects for glottal pulse spectral slopes between 0 dB and -12 dB per octave, and another defined by the parameter β which corrects for glottal pulse spectral slopes between -12 dB and -18 dB per octave. In the z domain, the two/three adaptive pre-emphasis filter has the form

$$(1-\alpha z^{-1})^2 \qquad (7.1)$$

when defined by the parameter α, and

$$(1 - z^{-1})^2(1 - \beta z^{-1}) \tag{7.2}$$

when defined by the parameter β.

The values of α and β used in a particular analysis situation are obtained from empirical expressions between α and $R(1)/R(0)$ and β and $R(1)/R(0)$, where $R(1)/R(0)$ is evaluated from the autocorrelation function of the waveform being analysed. These empirical expressions were determined using a criterion that the smallest area distance (and, hence, the most accurate acoustic tube shape recovery) occurs after a two/three adaptive pre-emphasis is followed by a conventional linear predictive analysis. For a waveform sampling frequency between 10 and 8 kilohertz, inclusive, the value of α is determined from the empirical expression

$$\frac{R(1)}{R(0)} = 3.217\alpha - 3.552\alpha^2 + 1.339\alpha^3 \tag{7.3}$$

which is implemented as a lookup table in Appendix D. For a waveform sampling frequency of 7 kilohertz, the value of α is determined from the empirical expression

$$\frac{R(1)}{R(0)} = 3.025\alpha - 3.148\alpha^2 + 1.127\alpha^3 \tag{7.4}$$

which is implemented as a lookup table in Appendix D.

The empirical expressions between β and $R(1)/R(0)$ change significantly as the waveform sampling frequency changes between 10 and 7 kilohertz. For a waveform sampling frequency of 10 kilohertz, the empirical expression is

$$\beta = -246.791 + 247.416 \frac{R(1)}{R(0)} \tag{7.5}$$

and for a waveform sampling frequency of 9 kilohertz

$$\beta = -294.826 + 295.625 \frac{R(1)}{R(0)} \tag{7.6}$$

and for a waveform sampling frequency of 8 kilohertz

$$\beta = -341.561 + 342.549 \frac{R(1)}{R(0)} \tag{7.7}$$

and for a waveform sampling frequency of 7 kilohertz

$$\beta = -369.261 + 370.372 \frac{R(1)}{R(0)} \tag{7.8}$$

The value of $R(1)/R(0)$ for which the two/three adaptive pre-emphasis filter changes from being defined by either α or β is called the transition value or point, and for waveform sampling frequencies 10, 9, 8 and 7 kilohertz the transition values are 0.99747, 0.99730, 0.99712 and 0.99700, respectively.

The Parcor Lossy Termination analysis procedure is similar to a conventional lattice formulation of linear prediction. A -6 dB per octave pre-emphasis is first applied to the speech waveform to account for the zero of the lip radiation impedance. The autocorrelation function of the pre-emphasized speech waveform is then calculated and equated to $A_M(r)$ for $0 \leqslant r \leqslant M$. A value for the termination reflection coefficient, μ_M, is chosen and used to evaluate

$$B_M(r) = \mu_M^2 A_M(r) \tag{7.9}$$

and

$$S_M(r) = -\mu_M A_M(r) \tag{7.10}$$

for $0 \leqslant r \leqslant M$.

The Parcor Lossy Termination analysis then continues in an iterative manner by evaluating

$$\mu_i = \frac{S_{i+1}(1)}{\sqrt{A_{i+1}(0) B_{i+1}(0)}} \tag{7.11}$$

$$A_i(r) = \frac{A_{i+1}(r) + \mu_i^2 B_{i+1} - \mu_i\big(S_{i+1}(r+1) + S_{i+1}(-r+1)\big)}{(1+\mu_i)^2} \tag{7.12}$$

$$B_i(r) = \frac{B_{i+1}(r) + \mu_i^2 A_{i+1} - \mu_i\big(S_{i+1}(r+1) + S_{i+1}(-r+1)\big)}{(1+\mu_i)^2} \tag{7.13}$$

$$S_i(\pm r) = \frac{S_{i+1}(\pm r + 1) + \mu_i^2 S_{i+1}(\mp r + 1) - \mu_i\big(A_{i+1}(r) + B_{i+1}(r)\big)}{(1+\mu_i)^2} \tag{7.14}$$

for i from M-1 to one. A computationally efficient version of Equations 7.12 through to 7.14 is found in Section 6.5.1.

The Parcor Lossy Termination analysis identifies a set of reflection coefficients μ_i for $1 \leqslant i \leqslant M-1$ which identifies a set of acoustic tube cross-sectional areas A_i via the relationship

$$A_{i+1} = A_i \left(\frac{1 + \mu_i}{1 - \mu_i}\right) \tag{7.15}$$

with A_0 arbitrarily chosen as unity. The set of A_i for $0 \leqslant i \leqslant M-1$ defines the acoustic tube shape recovered by a Parcor Lossy Termination analysis.

The new speech analysis procedure applies the two/three adaptive pre-emphasis filter to the speech waveform, and then a Parcor Lossy Termination analysis is performed on the pre-emphasized speech waveform. A complete description of the new speech analysis procedure is presented in the flow chart of Figure 7.1.

7.3 EVALUATION WITH SYNTHETIC SPEECH

This section evaluates the new speech analysis procedure, which is defined in the previous section, with synthetic speech waveforms generated for a sampling frequency of 10 kilohertz. The synthetic speech waveforms are generated with glottal pulse waveform excitations and a lossy termination of the acoustic tube model, as detailed in Appendix I.

The glottal pulse excitation waveforms are derived from the glottal pulse models B, C and E of ROSENBERG [1971], the glottal pulse model of FANT [1979], or digitized glottal pulse waveforms measured by ROTHENBERG [1971], MILLER [1959], MONSEN and ENGEBRETSON [1977], SONDHI [1975], SUNDBERG and GAUFFIN [1978], and FLANAGAN and LANDGRAF [1968]. The acoustic tube shape used to generate the synthetic speech approximates the real vocal tract shapes measured by FANT [1960] for the five vowels |a|, |e|, |i|, |o| and |u| (see Appendix C). All the assumptions of the linear prediction/acoustic tube model, with the exception of a non-white excitation and a lossy termination, are satisfied.

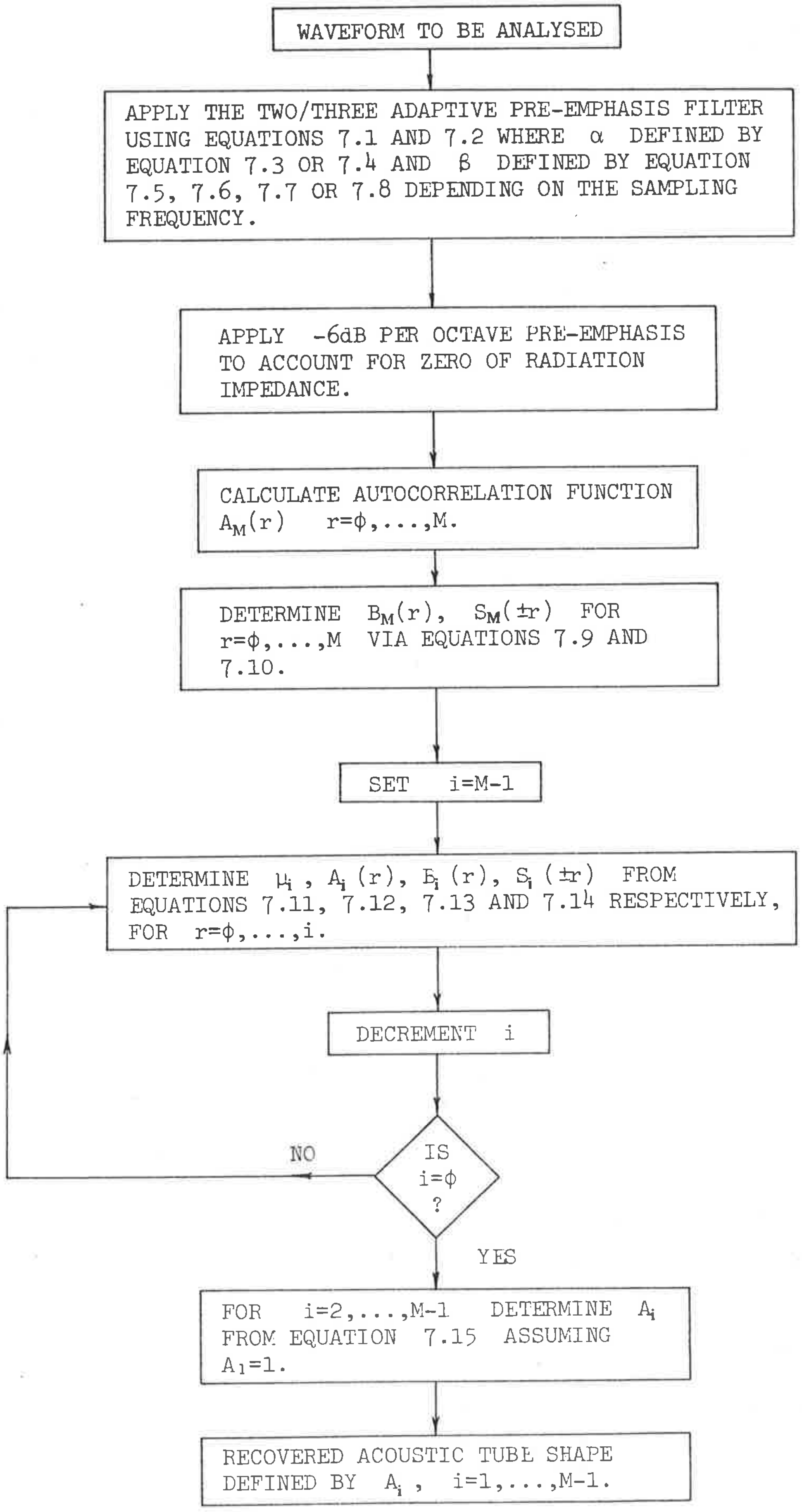

WAVEFORM TO BE ANALYSED
APPLY THE TWO/THREE ADAPTIVE PRE-EMPHASIS FILTER USING EQUATIONS 7.1 AND 7.2 WHERE α DEFINED BY EQUATION 7.3 OR 7.4 AND β DEFINED BY EQUATION 7.5, 7.6, 7.7 OR 7.8 DEPENDING ON THE SAMPLING FREQUENCY.
APPLY -6dB PER OCTAVE PRE-EMPHASIS TO ACCOUNT FOR ZERO OF RADIATION IMPEDANCE.
CALCULATE AUTOCORRELATION FUNCTION $A_M(r)$ $r=\phi,\ldots,M$.
DETERMINE $B_M(r)$, $S_M(\pm r)$ FOR $r=\phi,\ldots,M$ VIA EQUATIONS 7.9 AND 7.10.
SET $i=M-1$
DETERMINE μ_i, $A_i(r)$, $B_i(r)$, $S_i(\pm r)$ FROM EQUATIONS 7.11, 7.12, 7.13 AND 7.14 RESPECTIVELY, FOR $r=\phi,\ldots,i$.
DECREMENT i
IS $i=\phi$?
NO
YES
FOR $i=2,\ldots,M-1$ DETERMINE A_i FROM EQUATION 7.15 ASSUMING $A_1=1$.
RECOVERED ACOUSTIC TUBE SHAPE DEFINED BY A_i, $i=1,\ldots,M-1$.

FIGURE 7.1: Procedure for new speech analysis process.

The new speech analysis procedure is evaluated by comparing the area distances for the new speech analysis procedure with the area distances for a conventional Parcor analysis, and two conventional pre-emphasis techniques followed by a conventional Parcor analysis. One of the conventional pre-emphasis techniques is a +6 dB per octave pre-emphasis, which results from the simultaneous application of a +12 dB per octave pre-emphasis for glottal pulse excitation and a -6 dB per octave pre-emphasis for radiation of speech at the lips. The other conventional pre-emphasis technique considered is the adaptive pre-emphasis filter of GRAY and MARKEL [1974] and MAKHOUL and VISWANATHAN [1974], referred to as the conventional adaptive pre-emphasis filter, which has the form

$$(1 - \gamma z^{-1}) \tag{7.16}$$

where

$$\gamma = R(1)/R(0) \tag{7.17}$$

and $R(1)/R(0)$ is evaluated from the autocorrelation function of the speech waveform.

A value for the termination reflection coefficient, μ_M, must be chosen when synthetic speech is generated by the procedure detailed in Appendix I. In general, the value of μ_M used to generate speech waveforms is unknown; therefore, a value of μ_M must be estimated or chosen for the Parcor Lossy Termination analysis. Since the largest improvement in acoustic tube shape recovery for the new speech analysis procedure is expected when the value of μ_M used to generate the synthetic speech is the same as that used in the new speech analysis, Section 7.3.1 presents evaluation results for this idealized situation. Section 7.3.2 presents evaluations of the new speech analysis when the value of μ_M used in the new speech analysis procedure differs from that used to generate the synthetic speech.

7.3.1 EVALUATIONS WITH CORRECT LOSSY TERMINATION

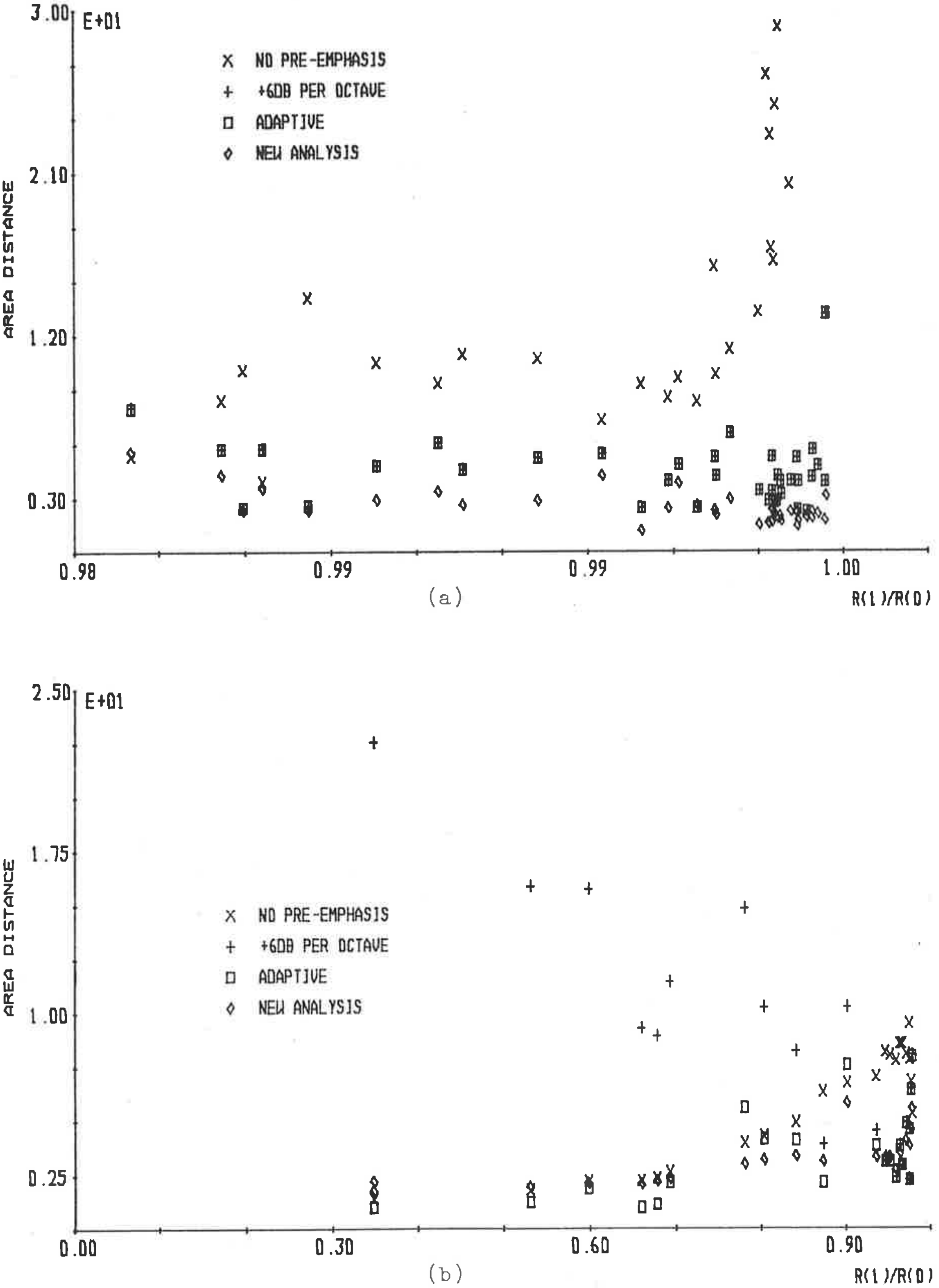
The evaluation results presented in this section use the same termination reflection coefficient, μ_M, in the Parcor Lossy Termination analysis part of the new speech analysis procedure as used to define the lossy termination for the generation of synthetic speech waveforms. Under this condition, the best acoustic tube shape recovery is most likely to occur for the new speech analysis procedure.

The synthetic speech waveforms used to perform the evaluations presented in this section are generated by the procedure detailed in Appendix I. All the glottal pulse waveforms used as excitations for the generation of synthetic speech are derived from glottal pulse models or measured glottal pulse waveforms during phonation (see Section 7.3). Evaluations for the Parcor Lossy Termination analysis presented in Chapter 6 showed that the value of μ_M which produces the best acoustic tube shape recovery depends on the synthetic speech waveform being analysed. However, it was concluded that a suitable value of μ_M which provides the best acoustic tube shape recovery for the five vowel sounds considered is 0.8. Therefore, the lossy termination used to generate the synthetic speech for the evaluation of the new speech analysis procedure is defined with μ_M equal to 0.8.

Evaluation of the new speech analysis procedure is performed by comparing area distances between original acoustic tube shapes and those recovered by no pre-emphasis, a conventional +6 dB per octave pre-emphasis, a conventional adaptive pre-emphasis (see GRAY and MARKEL [1974], MAKHOUL and VISWANATHAN [1974]), all followed by a conventional Parcor analysis, and the new speech analysis of synthetic speech waveforms for the vowels |a|, |e|, |i|, |o| and |u|. All the evaluation results are presented as plots of area distances versus $R(1)/R(0)$, which is evaluated from the autocorrelation function of the synthetic speech waveform.

Figure 7.2 presents the area distances versus $R(1)/R(0)$ for no pre-emphasis, a +6 dB per octave pre-emphasis, an adaptive pre-emphasis and the new speech analysis of synthetic speech for the vowel |a|. The area distances for no pre-emphasis are shown in Figure 7.2 to monotonically decrease as $R(1)/R(0)$ decreases from unity, which is the general trend observed for the evaluations presented in Chapter 5. No pre-emphasis of the synthetic speech for the vowel |a| produces much larger area distances than the two conventional pre-emphasis techniques and the new speech analysis when $R(1)/R(0)$ is between 0.9 and unity. When $R(1)/R(0)$ is less than 0.9, the area distances for no pre-emphasis are similar to those for an adaptive pre-emphasis and the new speech analysis procedure.

Figure 7.2 shows that the +6 dB per octave and adaptive pre-emphases produce similar area distances, which are smaller than those for no pre-emphasis when $R(1)/R(0)$ is between 0.95 and unity. When $R(1)/R(0)$ is less than 0.95, the +6 dB per octave pre-emphasis produces very large area distances, which monotonically increase as



(a)

(b)

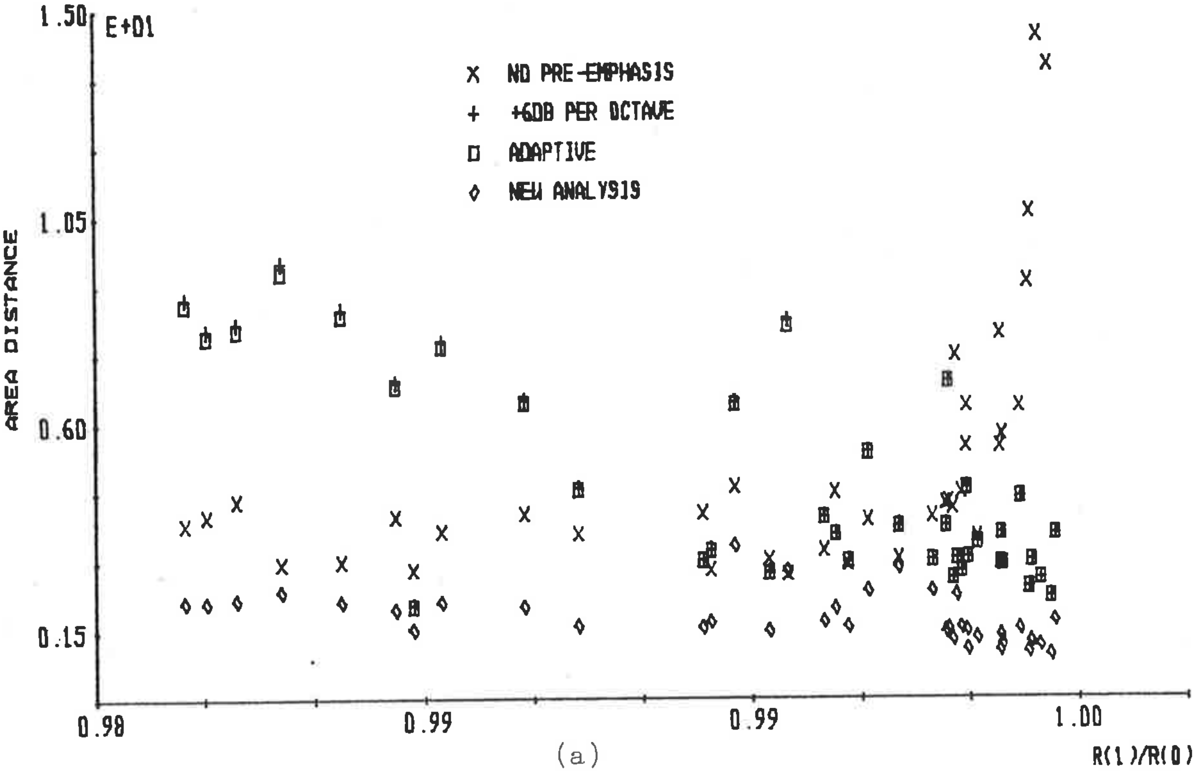
FIGURE 7.2: Analysis of synthetic speech for the vowel /a/ when R(1)/R(0) is (a) greater than 0.98 and (b) less than 0.98.

R(1)/*R*(0) decreases. Similar area distances are produced by the conventional adaptive pre-emphasis to those of no pre-emphasis and the new speech analysis procedure when *R*(1)/*R*(0) is less than 0.95.

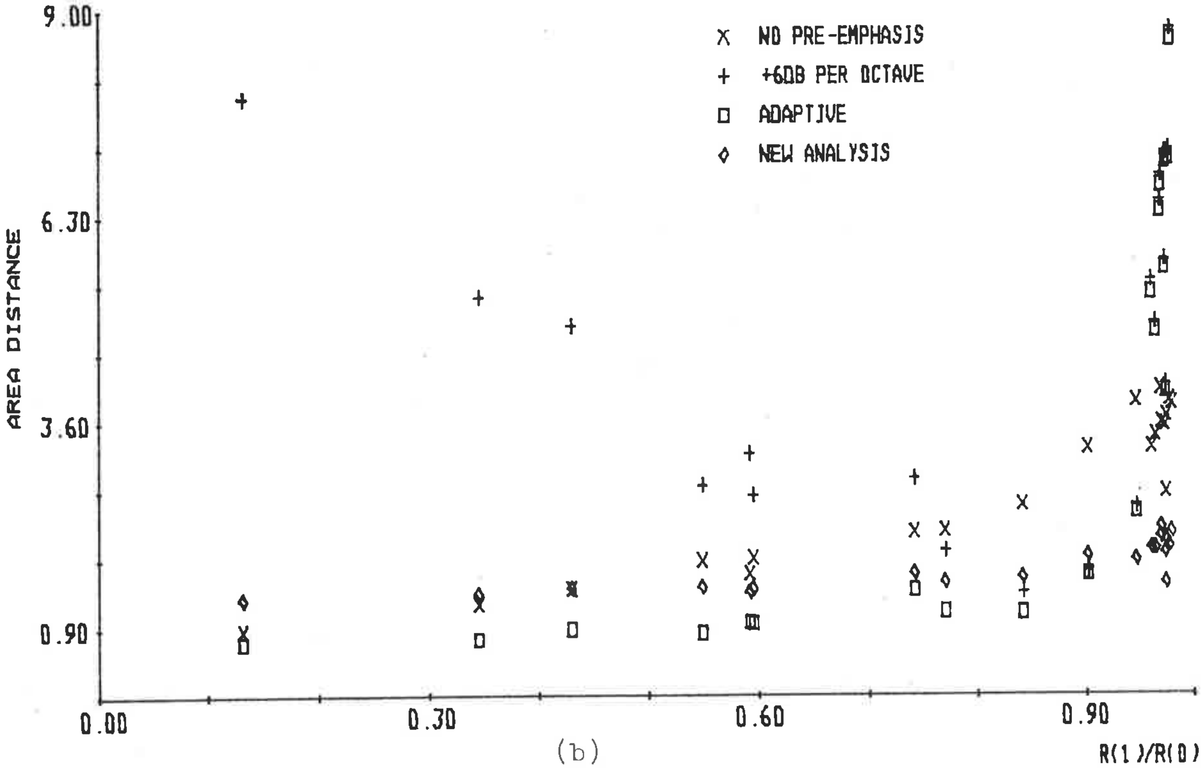
For synthetic speech of the vowel |a|, the new speech analysis procedure is shown in Figure 7.2 to produce a large reduction in area distances in comparison with no pre-emphasis, and a smaller reduction in area distances in comparison with a +6 dB per octave pre-emphasis and an adaptive pre-emphasis, whenever *R*(1)/*R*(0) is between 0.9 and unity. When *R*(1)/*R*(0) is less than 0.9, the new speech analysis procedure produces area distances which are similar to those for no pre-emphasis and an adaptive pre-emphasis, but much smaller than those for a +6 dB per octave pre-emphasis.

The variation in area distances produced by the new speech analysis procedure is smaller than for an adaptive pre-emphasis, and much smaller than for no pre-emphasis and a +6 dB per octave pre-emphasis. Therefore, the acoustic tube shape recovered by the new speech analysis procedure is less sensitive to changes in glottal pulse excitation waveforms (the only quantity that changes from one synthetic speech waveform to the next) than are conventional pre-emphases. In general, Figure 7.2 shows that the new speech analysis procedure produces a reduction in area distances and, hence, improved acoustic tube shape recovery, when compared with conventional pre-emphasis methods for analysis of synthetic speech for the vowel |a|.

Figure 7.3 presents a plot of area distances versus *R*(1)/*R*(0) for no pre-emphasis, a +6 dB per octave pre-emphasis, an adaptive pre-emphasis and the new speech analysis of synthetic speech for



(a)



(b)

FIGURE 7.3: Analysis of synthetic speech for the vowel /e/ when R(1)/R(0) is (a) greater than 0.98 and (b) less than 0.98.

the vowel |e|. No pre-emphasis is shown in Figure 7.3 to produce area distances which are much larger than those for the conventional pre-emphasis techniques and the new speech analysis procedure when $R(1)/R(0)$ is close to unity, smaller than those for the conventional pre-emphasis techniques when $R(1)/R(0)$ is between 0.95 and 0.99, and similar to those for an adaptive pre-emphasis and the new speech analysis when $R(1)/R(0)$ is less than 0.95.

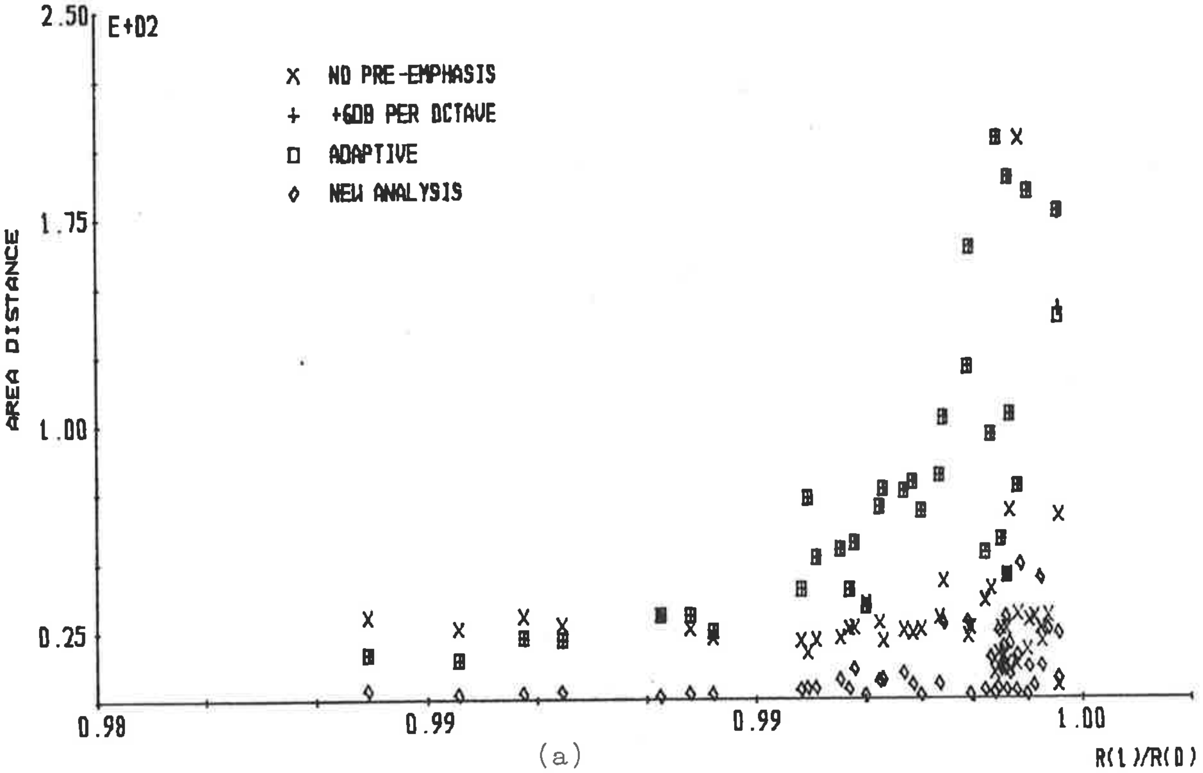
The area distances for a +6 dB per octave pre-emphasis and an adaptive pre-emphasis are similar when $R(1)/R(0)$ is greater than 0.9, and much larger than the area distances for no pre-emphasis and the new speech analysis when $R(1)/R(0)$ is between 0.9 and 0.99. When $R(1)/R(0)$ is less than 0.6, the area distances for a +6 dB per octave are very large, and an adaptive pre-emphasis produces the smallest area distances when $R(1)/R(0)$ is less than 0.9.

For synthetic speech of the vowel |e|, the new speech analysis procedure is shown in Figure 7.3 to produce a reduction in area distances in comparison with no pre-emphasis, a +6 dB per octave pre-emphasis and an adaptive pre-emphasis, when $R(1)/R(0)$ is greater than 0.9. When $R(1)/R(0)$ is less than 0.9, the area distances for the new speech analysis procedure are much smaller than the area distances for no pre-emphasis, and slightly larger than the area distances for an adaptive pre-emphasis. The variation in area distances for the new speech analysis procedure is shown in Figure 7.3 to be much smaller than the variation in area distances for the conventional pre-emphasis techniques. A similar conclusion was drawn for the analysis of synthetic speech for the vowel |a|.

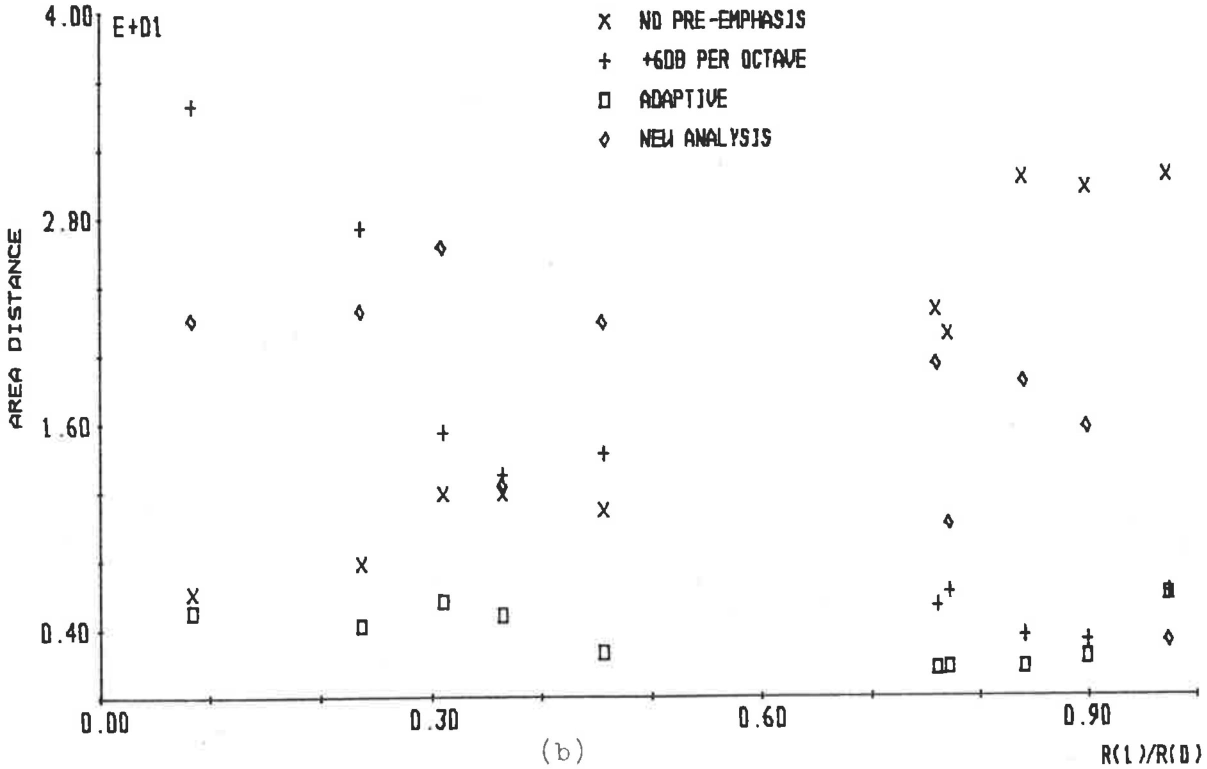
In general, Figure 7.3 shows that the new speech analysis procedure produces a reduction in area distances and, hence, an improvement in acoustic tube shape recovery, when compared with the conventional pre-emphasis techniques for $R(1)/R(0)$ greater than 0.9. When $R(1)/R(0)$ is less than 0.9, the new speech analysis procedure produces a reduction in area distance only in comparison with no pre-emphasis and a +6 dB per octave pre-emphasis.

The area distances versus $R(1)/R(0)$ presented in Figure 7.4 are for no pre-emphasis, a +6 dB per octave pre-emphasis, an adaptive pre-emphasis and the new speech analysis of synthetic speech for the vowel |i|. The area distances for no pre-emphasis of the synthetic speech are much smaller than those for a +6 dB per octave pre-emphasis and an adaptive pre-emphasis when $R(1)/R(0)$ is greater than 0.99. The area distances for no pre-emphasis are only smaller than those for a +6 dB per octave pre-emphasis when $R(1)/R(0)$ is less than 0.5, and smaller than those for the new speech analysis procedure when $R(1)/R(0)$ is less than 0.6.

Similar area distances occur for the +6 dB per octave pre-emphasis and an adaptive pre-emphasis when $R(1)/R(0)$ is greater than 0.9. The adaptive pre-emphasis produces the smallest area distances when $R(1)/R(0)$ is less than 0.9 and, therefore, the best acoustic tube shape recovery for $R(1)/R(0)$ less than 0.9. The area distances for a +6 dB per octave pre-emphasis increase as $R(1)/R(0)$ decreases from 0.9, and these area distances are the largest when $R(1)/R(0)$ is approximately zero.



(a)



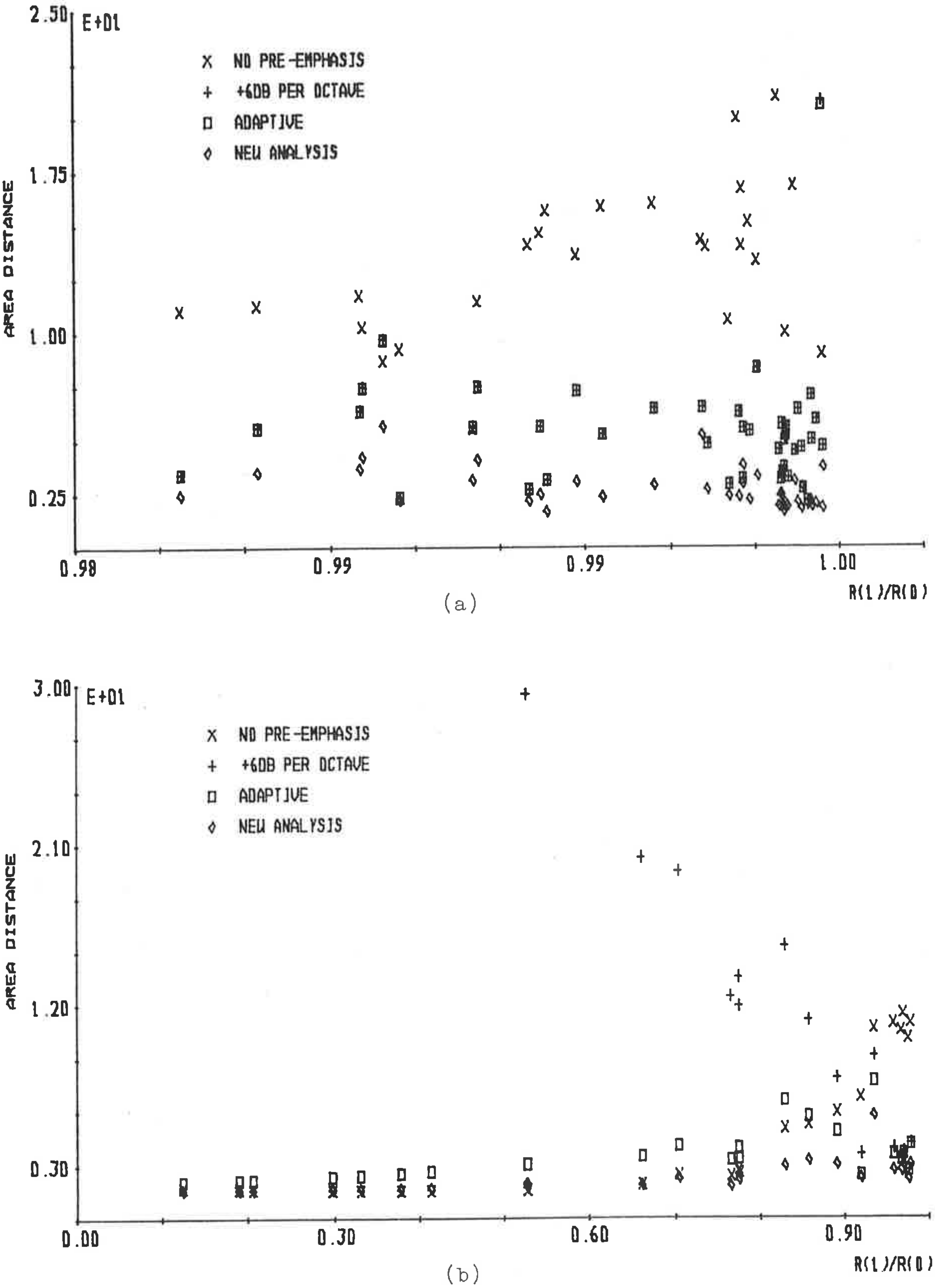
(b)

FIGURE 7.4: Analysis of synthetic speech for the vowel /i/ when R(1)/R(0) is (a) greater than 0.98 and (b) less than 0.98.

The new speech analysis procedure is shown in Figure 7.4 to produce the smallest area distances and, hence, the best acoustic tube shape recovery, when $R(1)/R(0)$ is greater than 0.95. The area distances for the new speech analysis procedure are always larger than those for an adaptive pre-emphasis when $R(1)/R(0)$ is less than 0.95. The variation in area distances for the new speech anlaysis procedure is found in Figure 7.4 to be smaller than the variation in area distances for the conventional pre-emphasis, when the range of $R(1)/R(0)$ from zero to unity is considered. This is consistent with the conclusions drawn for evaluations performed with synthetic speech for the vowels |a| and |e|.

In general, the new speech analysis procedure is shown in Figure 7.4 to produce a significant reduction in area distances and, hence, an improvement in acoustic tube shape recovery when compared with conventional pre-emphasis techniques for $R(1)/R(0)$ greater than 0.95. This is not the case when $R(1)/R(0)$ is less than 0.95, when the conventional adaptive pre-emphasis produces the smallest area distances. For all the evaluation results presented so far, the new speech analysis procedure consistently produces the smallest variation in area distances when $R(1)/R(0)$ is between zero and unity.

A plot of area distances versus $R(1)/R(0)$ is presented in Figure 7.5 for no pre-emphasis, a +6 dB per octave pre-emphasis (only for $R(1)/R(0)$ greater than 0.3), an adaptive pre-emphasis and the new speech analysis of synthetic speech for the vowel |o|. The area distances for a +6 dB per octave pre-emphasis, when $R(1)/R(0)$ is less than 0.3, are not presented in Figure 7.5, because they are very large. The area distances for no pre-



(a)

(b)

FIGURE 7.5: Analysis of synthetic speech for the vowel /o/ when R(1)/R(0) is (a) greater than 0.98 and (b) less than 0.98.

emphasis are shown in Figure 7.5 to be larger than those for the other conventional pre-emphases and the speech analysis when $R(1)/R(0)$ is greater than 0.92. When $R(1)/R(0)$ is less than 0.9, the area distances for no pre-emphasis are similar to those for an adaptive pre-emphasis and the new speech analysis procedure.

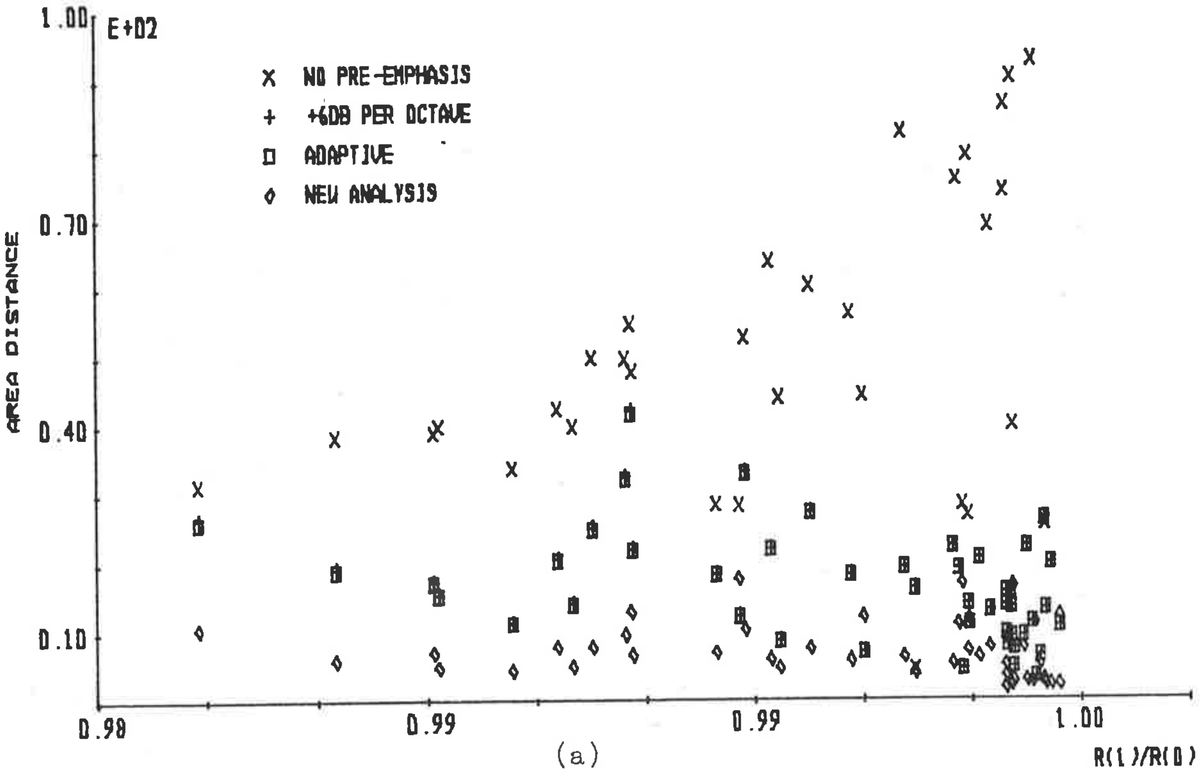
Similar area distances occur for a +6 dB per octave pre-emphasis and an adaptive pre-emphasis when $R(1)/R(0)$ is greater than 0.95, and these area distances are less than those for no pre-emphasis, but greater than those for the new speech analysis procedure. When $R(1)/R(0)$ is less than 0.9, the area distances for a +6 dB per octave pre-emphasis increase dramatically as $R(1)/R(0)$ decreases, and the area distances for an adaptive pre-emphasis are similar, but slightly larger, than those for no pre-emphasis and the new speech analysis procedure.

Figure 7.5 shows that the new speech analysis procedure produces area distances which are smaller than those for a +6 dB per octave pre-emphasis and an adaptive pre-emphasis, when $R(1)/R(0)$ is between zero and unity. A large reduction in area distances occurs for the new speech analysis procedure in comparison with no pre-emphasis when $R(1)/R(0)$ is greater than 0.85, and similar area distances occur when $R(1)/R(0)$ is less than 0.85. The variation in area distances for the new speech analysis procedure is found in Figure 7.5 to be much smaller than the variation in area distances for the conventional pre-emphases. This is consistent with the conclusions drawn for evaluations performed with synthetic speech for the vowels |a|, |e| and |i|.

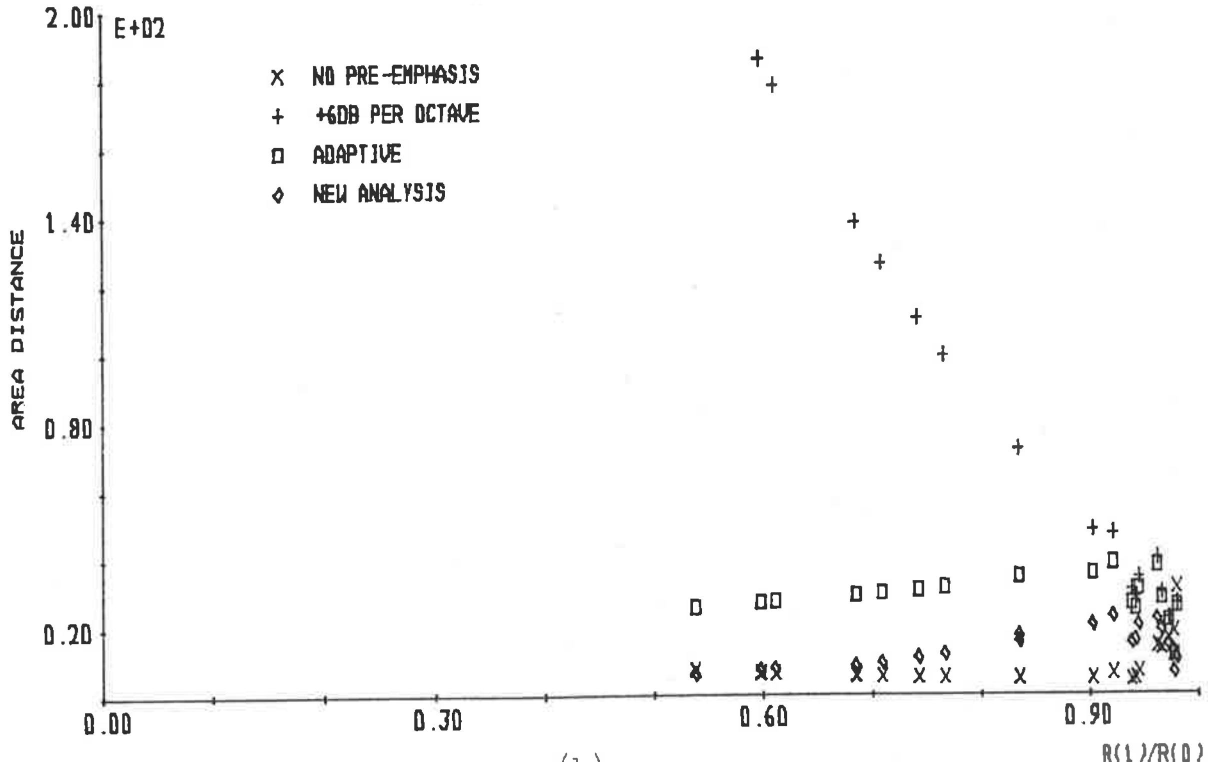
The evaluation results presented in Figure 7.5 for the analysis of synthetic speech for the vowel |o| show, in general, that the new speech analysis procedure produces reductions in area distances, ranging from very large to very small, when compared with the conventional pre-emphases. A smaller variation in area distances also occurs for the new speech analysis procedure when compared with that for the conventional pre-emphases.

For synthetic speech of the vowel |u|, the area distances versus $R(1)/R(0)$ are plotted in Figure 7.6 for no pre-emphasis, a +6 dB per octave pre-emphasis, an adaptive pre-emphasis and the new speech analysis procedure. Figure 7.6 does not contain any evaluation results for $R(1)/R(0)$ less than 0.5, as no glottal pulse waveforms could be found to produce an $R(1)/R(0)$ less than 0.5. The area distances for no pre-emphasis are shown in Figure 7.6 to be much larger than those for the other conventional pre-emphases and the new speech analysis procedure when $R(1)/R(0)$ is greater than 0.95. No pre-emphasis produces the smallest area distances when $R(1)/R(0)$ is less than 0.95.

When $R(1)/R(0)$ is greater than 0.95, similar area distances occur for the +6 dB per octave pre-emphasis and an adaptive pre-emphasis, and these area distances are less than those for no pre-emphasis, but greater than those for the new speech analysis procedure. The area distances for a +6 dB per octave pre-emphasis increase dramatically as $R(1)/R(0)$ decreases from 0.95, and the area distances for an adaptive pre-emphasis are larger than those for no pre-emphasis and the new speech analysis procedure when $R(1)/R(0)$ is less than 0.95.



(a)



(b)

FIGURE 7.6: Analysis of synthetic speech for the vowel /u/ when R(1)/R(0) is (a) greater than 0.98 and (b) less than 0.98.

Figure 7.6 shows that the new speech analysis procedure produces smaller area distances than those for a +6 dB per octave pre-emphasis and an adaptive pre-emphasis. A reduction in area distances by the new speech analysis procedure in comparison with no pre-emphasis only occurs when $R(1)/R(0)$ is greater than 0.95. The variation in area distances produced by the new speech analysis procedure is shown in Figure 7.6 to be much smaller than that for the conventional pre-emphases. This is consistent with the conclusions drawn for evalutions performed with synthetic speech for the vowels |a|, |e|, |i| and |o|.

In general, the evaluation results presented in Figure 7.6 for synthetic speech of the vowel |u| show that the new speech analysis procedure produces reductions in area distances and, hence, improved acoustic tube shape recovery, in comparison with conventional pre-emphases, when $R(1)/R(0)$ is greater than 0.95. A reduction in area distances by the new speech analysis procedure only occurs in comparison with a +6 dB per octave pre-emphasis and an adaptive pre-emphasis when $R(1)/R(0)$ is less than 0.95.

The evaluation results presented in this section for synthetic speech of the vowels |a|, |e|, |i|, |o| and |u|show that the new speech analysis procedure produces consistently smaller area distances than those for conventional pre-emphasis techniques, when $R(1)/R(0)$ is greater than 0.95. With the exception of a few isolated cases, the area distances for the new speech analysis procedure are always smaller than those for a +6 dB per octave pre-emphasis for the five vowels and $R(1)/R(0)$ between zero and unity. Similar area distances occur for no pre-emphasis, an adaptive pre-emphasis and the new speech analysis procedure, except for synthetic speech of the vowel |i|, when $R(1)/R(0)$ is less than 0.9.

The variation in area distance values for the new speech analysis procedure is much smaller than that for the conventional analysis procedures, and occurs without exception for all the synthetic speech waveforms of the five vowel sounds. Hence, the acoustic tube shape recovered by the new speech analysis procedure is less sensitive to changes in the glottal pulse excitation waveform (the only quantity that changes from one synthetic speech waveform to another) than are the acoustic tube shapes recovered by conventional analysis procedures. This is an important attribute of a speech analysis process which is required to recover slowly varying parameters from the speech waveform, e.g. low bit rate speech transmission systems.

In conclusion, the evaluations presented in this section have shown that there exists a distinct advantage in using the new speech analysis procedure in preference to conventional pre-emphasis, when analysing synthetic speech waveforms for the vowels |a|, |e|, |i|, |o| and |u|.

7.3.2 EVALUATIONS WITH INCORRECT LOSSY TERMINATION

In the application of the new speech analysis procedure to the analysis of real speech, a value for the termination reflection coefficient, μ_M, used in the Parcor Lossy Termination analysis must be estimated or chosen. This section evaluates the effects on the acoustic tube shape of different values of μ_M being used in the new speech analysis procedure and that to define the lossy termination of acoustic tubes. The value of the termination reflection coefficient used in the new speech analysis procedure is denoted as μ_M'.

All the synthetic speech waveforms used to produce the evaluation results presented in this section are generated by the procedure detailed in Appendix I. The glottal pulse excitation waveforms used to generate the synthetic speech are derived from glottal pulse models or measured glottal pulse waveforms during phonation (see Section 7.3). The lossy termination of the acoustic tubes is defined with the termination reflection coefficient, μ_M, equal to 0.8, i.e. the same as in the previous section.

The evaluation results are presented as plots of area distances between recovered and original acoustic tube shapes versus μ_M' for various values of $R(1)/R(0)$. The ratio $R(1)/R(0)$ is evaluated from the autocorrelation function of the synthetic speech waveform being analysed, and evaluation results are presented for values of $R(1)/R(0)$ as close as possible to 0.4, 0.6, 0.8, 0.9, 0.95 and 0.98. Synthetic speech waveforms for the vowel sounds |a|, |e|, |i|, |o| and |u| are used in the evaluations presented in this section.

Figure 7.7 presents the area distances versus μ_M' for various values of $R(1)/R(0)$ when the new speech analysis procedure is used to analyse synthetic speech for the vowel |a|. For any value of $R(1)/R(0)$, the area distances are observed in Figure 7.7 to increase monotonically for decreasing μ_M'. The relative increase in area distances as μ_M' changes from unity to 0.6 increases as the value of

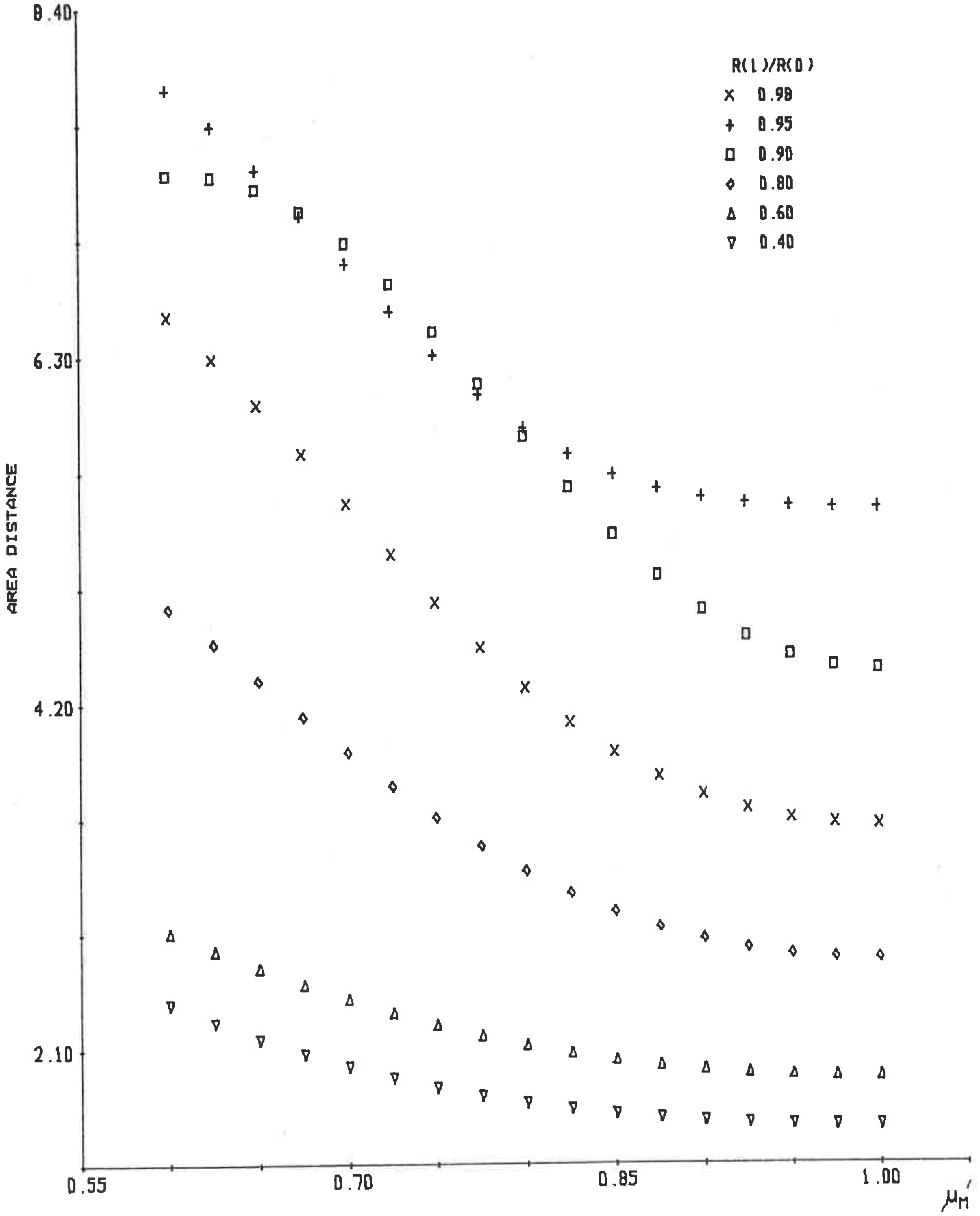


FIGURE 7.7: New speech analysis of synthetic speech for the vowel /a/.

$R(1)/R(0)$ increases. In general, Figure 7.7 indicates that, for the new speech analysis of synthetic speech for the vowel $|a|$, a decrease in area distance may be achieved if μ_M' is larger than μ_M.

Figure 7.8 presents the area distances versus $R(1)/R(0)$ for no pre-emphasis, a +6 dB per octave pre-emphasis, an adaptive pre-emphasis (obtained from Figure 7.2), and the largest area distance presented in Figure 7.8, i.e. when $\mu_M' = 0.6$. A comparison of Figures 7.8 and 7.2 shows that the performance of the new speech analysis procedure, when analysing synthetic speech for the vowel $|a|$, is not significantly affected when μ_M and μ_M' differ widely.

A plot of area distances versus μ_M' for various values of $R(1)/R(0)$ is presented in Figure 7.9, when the new speech analysis procedure is used to analyse synthetic speech of the vowel $|e|$. When $R(1)/R(0)$ is less than 0.9, Figure 7.9 shows a very small increase in area distance as μ_M' decreases from unity. Decreasing and then increasing area distances, as μ_M' decreases from unity, are observed in Figure 7.9 when $R(1)/R(0)$ is greater than or equal to 0.9. The minimum area distances when $R(1)/R(0)$ is greater than or equal to 0.9 occur at similar values of μ_M', which are slightly less than the value of μ_M used to define the lossy termination during the synthetic speech generation.

Except for $R(1)/R(0) = 0.98$, only a very small change in area distances is observed in Figure 7.9 for a relatively large change in the value of μ_M'. A comparison of the area distances presented in Figure 7.9 and those for the conventional pre-emphasis techniques presented in Figure 7.3 shows that the performance of the new speech analysis procedure, when analysing synthetic speech for

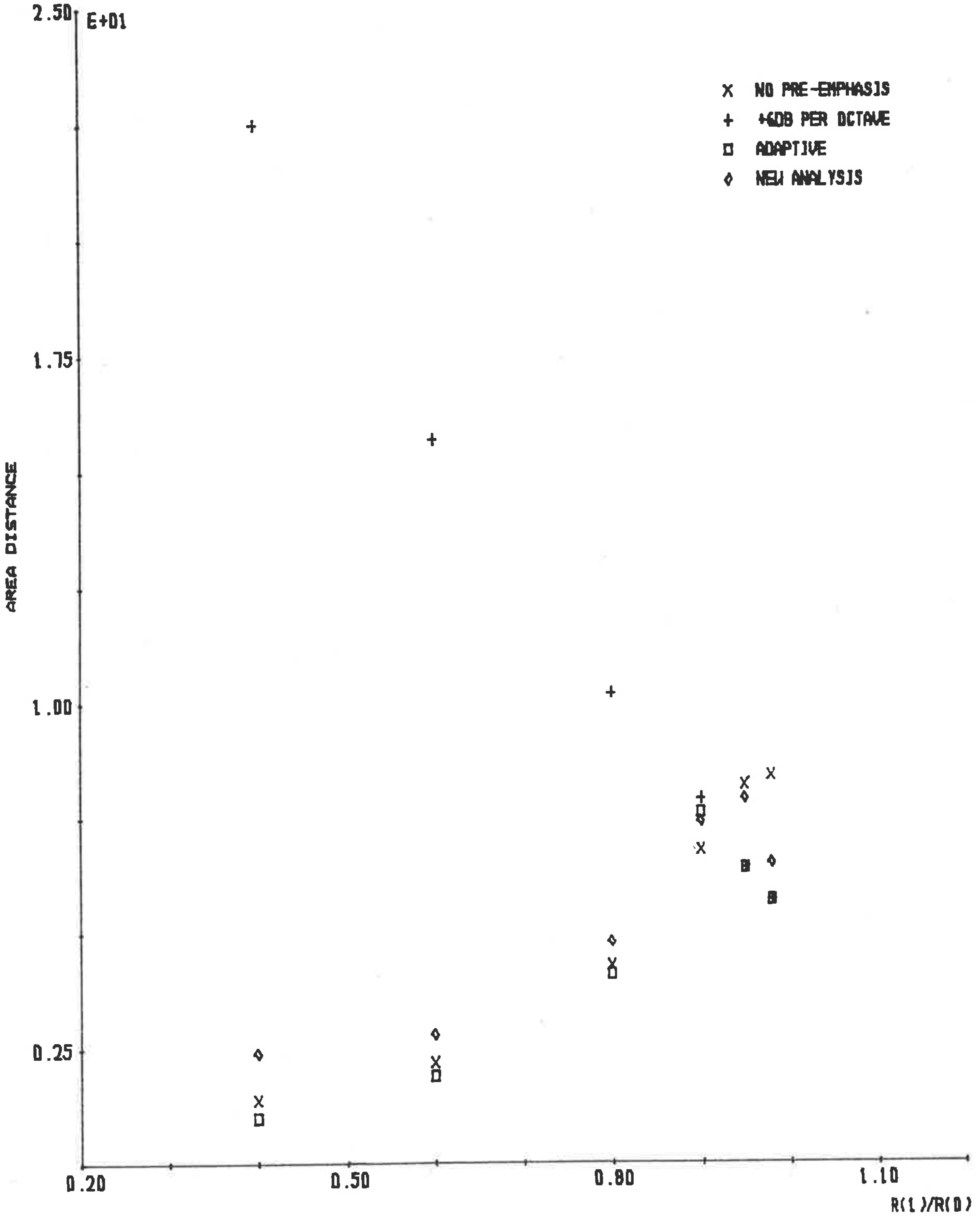


FIGURE 7.8: Analysis of synthetic speech for the vowel /a/.

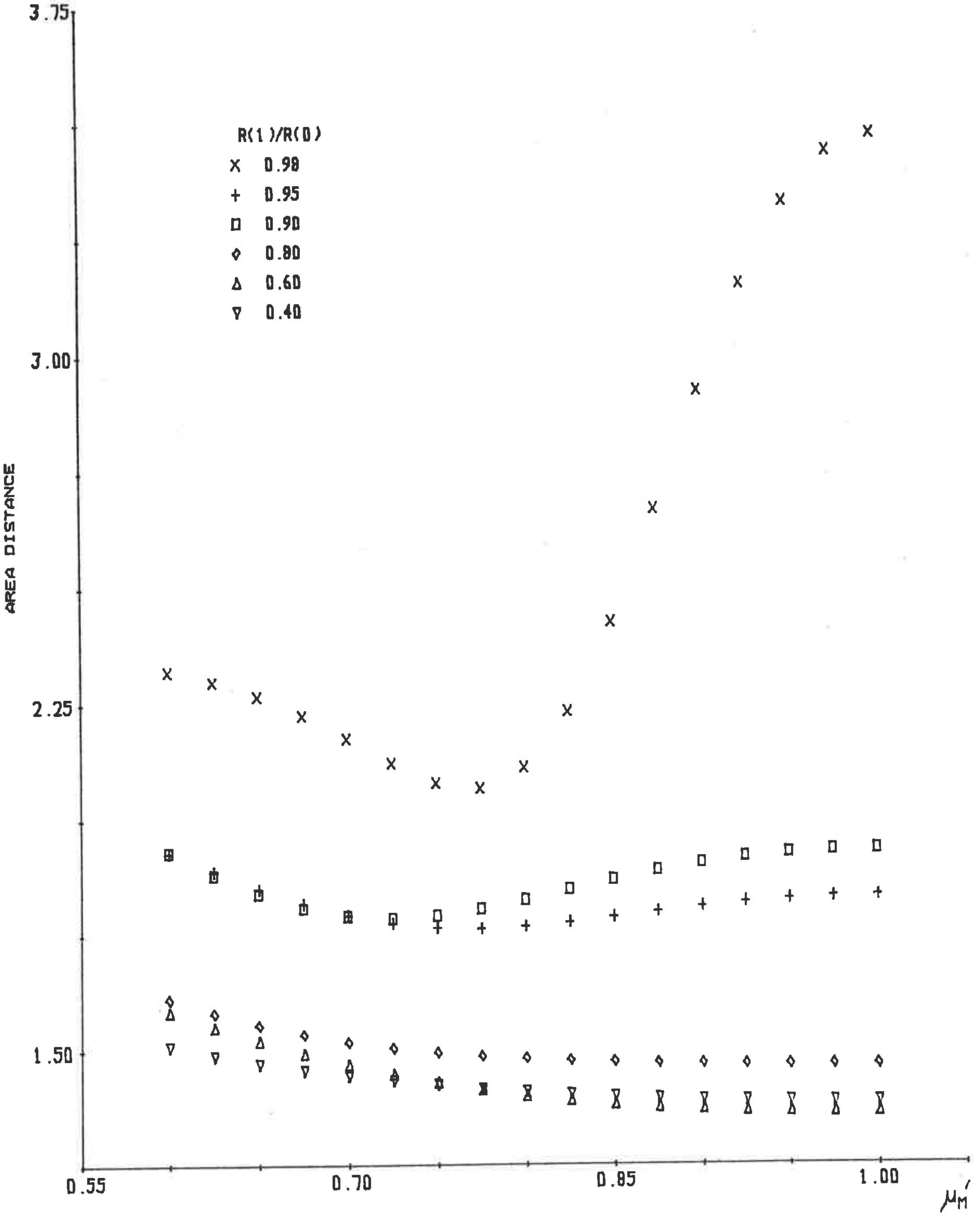


FIGURE 7.9: New speech analysis of synthetic speech for the vowel /e/.

the vowel $|e|$, is not significantly affected when μ_M and μ_M' differ widely.

Figure 7.10 presents a plot of area distances versus μ_M' for various values of $R(1)/R(0)$ when the new speech analysis procedure is used to analyse synthetic speech for the vowel $|i|$. Only a small change in area distances occurs when $R(1)/R(0)$ is equal to 0.4 and 0.98; otherwise an increase to a maximum and then a decrease in area distances is observed in Figure 7.10 as μ_M' decreases from unity. The value of μ_M' at which the maximum in area distances occurs varies over a wide range, but is less than the value of μ_M used to define the lossy termination during the generation of the synthetic speech. This is similar to the conclusion drawn for the values of μ_M at which a minimum area distance occurred for analysis of synthetic speech for the vowel $|e|$ (see Figure 7.9).

The evaluation results presented in Figure 7.10 show that, whenever μ_M' differs significantly from μ_M, which defines the lossy termination during synthetic speech production, then either a small change or a large decrease in area distances occurs. Therefore, it can be concluded that the performance of the new speech analysis procedure, when analysing synthetic speech for the vowel $|i|$, is either unaffected or improves when μ_M' and μ_M differ widely. This conclusion is not consistent with those drawn for synthetic speech of the vowels $|a|$ and $|e|$.

A plot of area distances versus μ_M' for various values of $R(1)/R(0)$ is presented in Figure 7.11 when the new speech analysis procedure is used to analyse synthetic speech for the vowel $|o|$. In general, Figure 7.11 shows that the area distances in-

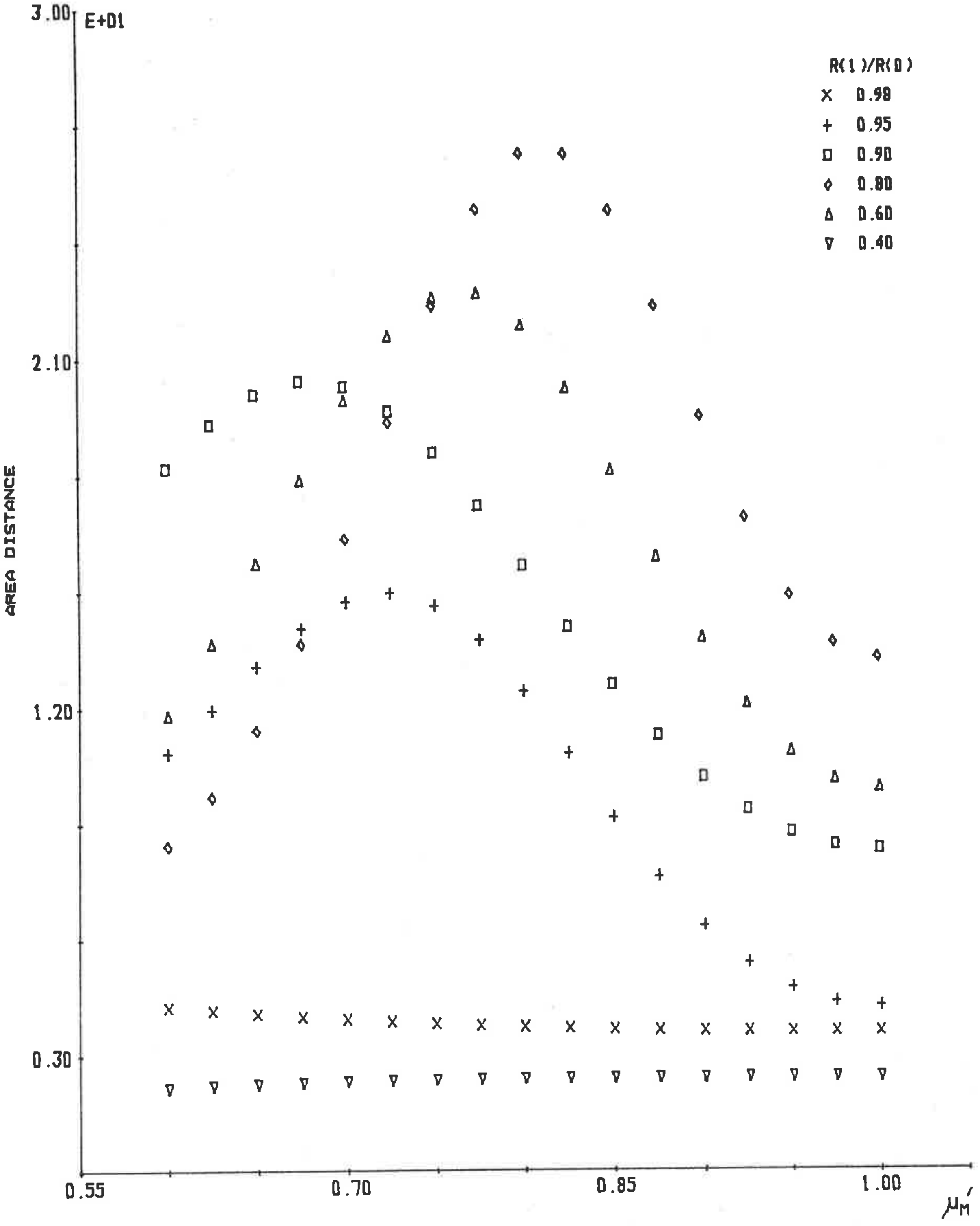


FIGURE 7.10: New speech analysis of synthetic speech for the vowel /i/.

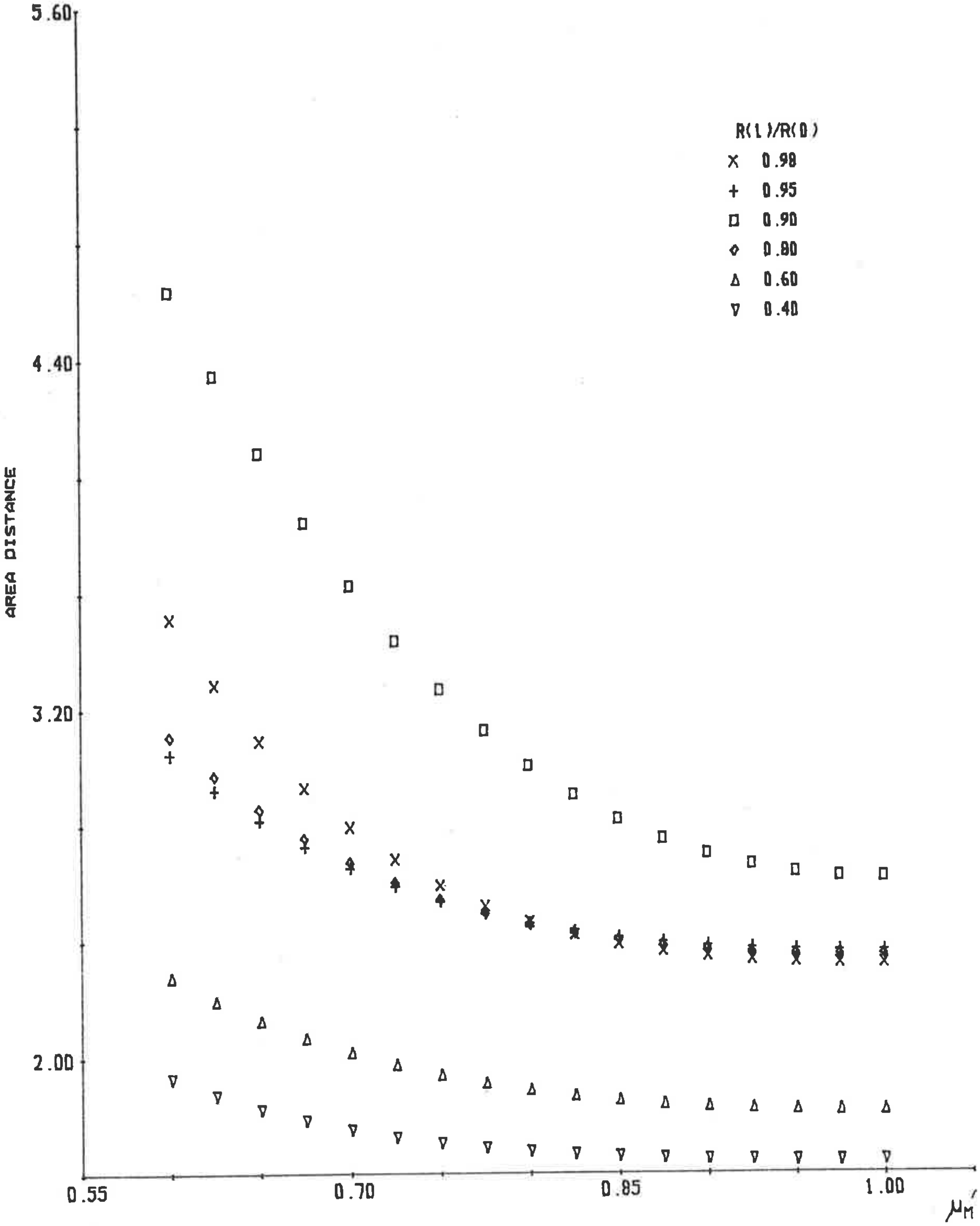


FIGURE 7.11: New speech analysis of synthetic speech for the vowel /o/.

crease as μ_M' decreases from unity. This is consistent with the evaluation results presented for the vowel |a| in Figure 7.7, with the exception that a larger change in area distances for μ_M' changing from unity to 0.6 does not necessarily occur for increasing $R(1)/R(0)$.

Figure 7.12 presents the area distances versus $R(1)/R(0)$ for no pre-emphasis, a +6 dB per octave pre-emphasis, an adaptive pre-emphasis (obtained from Figure 7.5), and the largest area distances presented in Figure 7.11, i.e. when $\mu_M' = 0.6$. The new speech analysis procedure is shown in Figure 7.12 to have a similar performance to that of a adaptive pre-emphasis in this worse case situation. A comparison of Figures 7.12 and 7.5 shows that the performance of the new speech analysis procedure, when analysing synthetic speech for the vowel |o|, is not significantly affected when μ_M and μ_M' differ widely.

Figure 7.13 presents a plot of area distances versus μ_M' for various values of $R(1)/R(0)$ when the new speech analysis procedure is used to analyse synthetic speech for the vowel |u|. The area distances for $R(1)/R(0)$ equal to 0.4 are not presented in Figure 7.13 because of the lack of glottal pulse excitation waveforms to produce this value of $R(1)/R(0)$. Except for the case when $R(1)/R(0)$ is equal to 0.98, the area distances are shown in Figure 7.13 to decrease as the value of μ_M' decreases from unity. When $R(1)/R(0)$ is equal to 0.98, the area distances decrease to a minimum, and then increase as μ_M' decreases from unity.

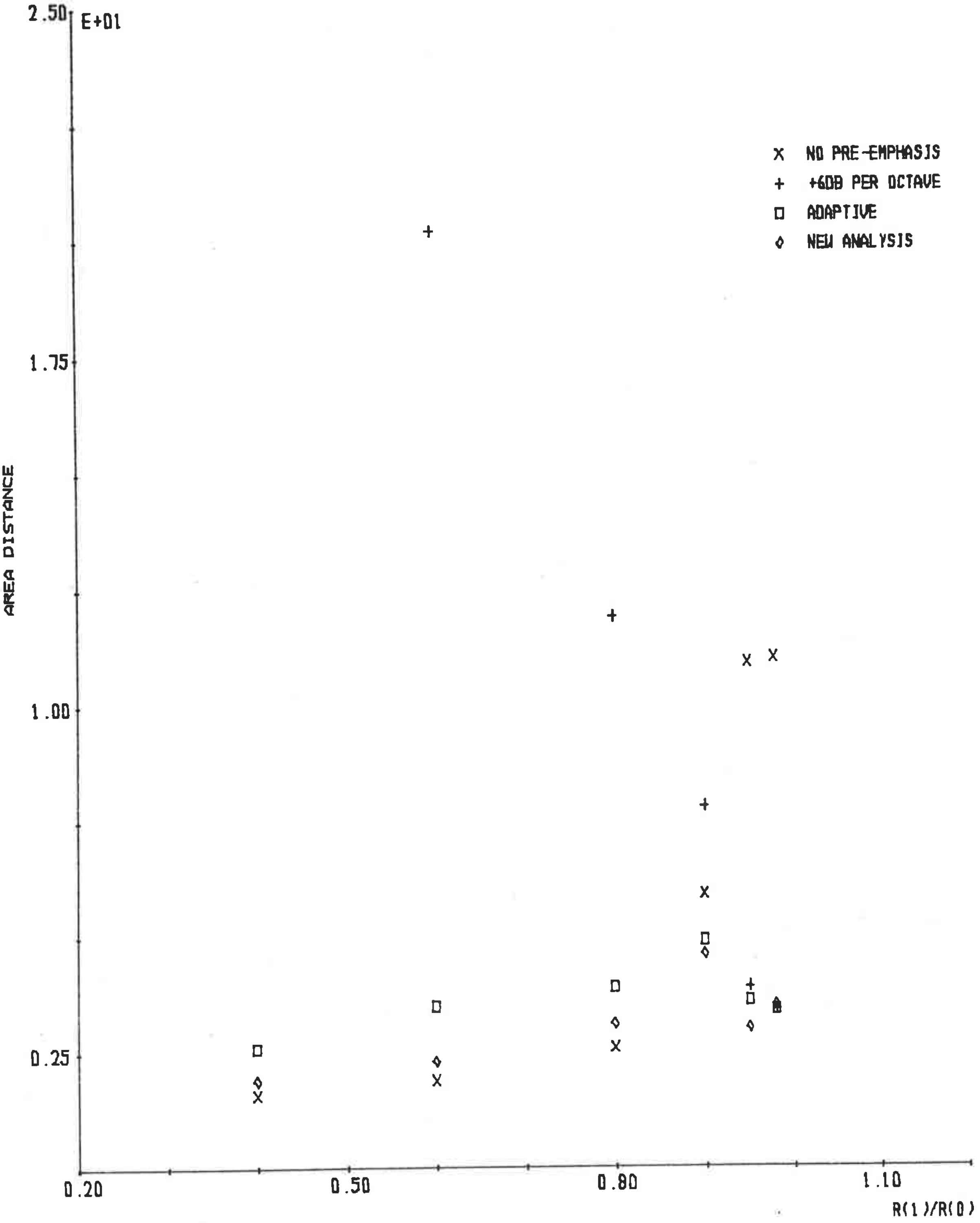


FIGURE 7.12: Analysis of synthetic speech for the vowel /o/.

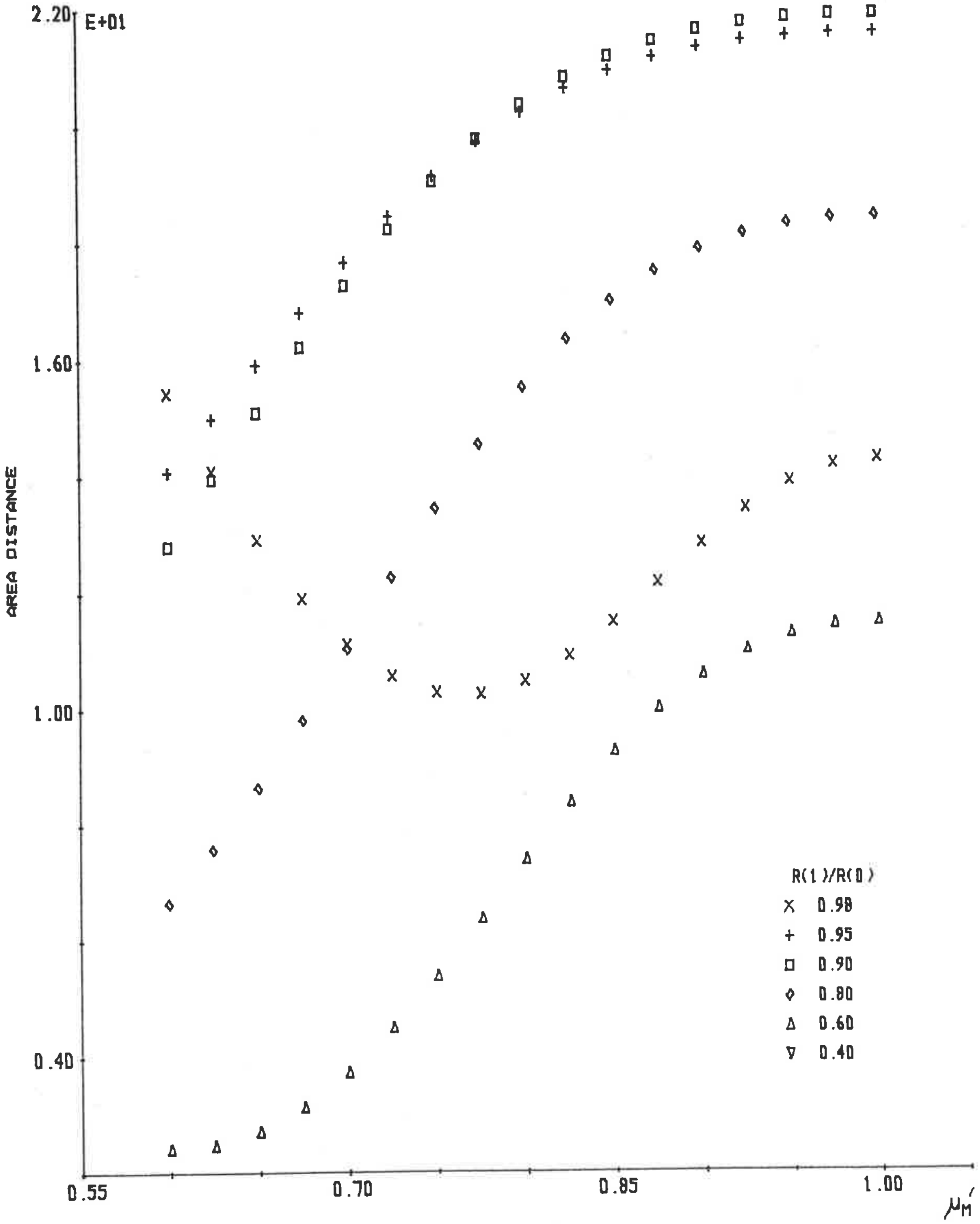


FIGURE 7.13: New speech analysis of synthetic speech for the vowel /u/.

When $R(1)/R(0)$ is less than unity, the evaluation results presented in Figure 7.13 indicate that a reduction in area distances may occur when the value of μ_M' is less than the value of μ_M, which defined the lossy termination during the generation of the synthetic speech. Hence, it is concluded that the performance of the new speech analysis procedure improves slightly when $R(1)/R(0)$ is not close to unity and μ_M' is less than μ_M.

The evaluation results presented in this section have shown that the general trends of area distances versus μ_M' change from one vowel sound to another and for changes in $R(1)/R(0)$. For the vowels |a| and |o|, the area distances resulting from a new speech analysis increase as the value of μ_M' decreases from unity and, for the vowel |u|, the area distances decrease as the value of μ_M' decreases from unity. The area distances for a new speech analysis of the vowel |e| decrease to a minimum and then increase as μ_M' decreases from unity, and the area distances for the vowel |i| increase to a maximum and then decrease as μ_M' decreases from unity.

With only a few exceptions, the changes of area distances when the new speech analysis procedure is used to analyse synthetic speech for five vowel sounds is relatively small for a relatively large change in μ_M' from unity to 0.6. Even when the largest area distances presented in this section are compared with those for no pre-emphasis, a +6 dB per octave pre-emphasis and an adaptive pre-emphasis (see Section 7.3.1), the new speech analysis procedure, in general, provides smaller area distances and, hence, improved acoustic tube shape recovery. The relatively small change in area distances that occurs for changes in μ_M' indicates that the acoustic tube shape recovered by the new speech analysis procedure is not strongly dependent on the value of μ_M'.

In general, it may be concluded that the performance of the new speech analysis procedure, in comparison with conventional pre-emphasis techniques, in only slightly affected when μ_M' and μ_M do not differ widely. When μ_M' and μ_M differ widely, then the performance of the new speech analysis procedure may improve, depending on the vowel sound being analysed, the relative sizes of μ_M' and μ_M, and the value of $R(1)/R(0)$.

7.4 EVALUATION WITH REAL SPEECH

The new speech analysis procedure has been evaluated with synthetic speech waveforms of five vowel sounds, and this section presents evaluation results for the new speech analysis procedure when used to analyse real speech waveforms of the same five vowels.

When real speech waveforms are used to perform an evaluation of the new speech analysis procedure, it is very difficult to obtain results for the wide range of $R(1)/R(0)$ values that were presented for the evaluations with synthetic speech in Section 7.3. Therefore, the evaluation results presented in this section, in general, only cover a relatively small range of $R(1)/R(0)$ values.

A major problem encountered when using real speech waveforms to evaluate an analysis procedure is the lack of knowledge of the vocal tract shape used to produce the speech sound. Measuring vocal tract shapes is a very difficult and complex process, as discussed in Chapter 1, and so the recovered vocal tract shapes are compared with the vocal tract shapes measured by FANT [1960] for corresponding vowel sounds. Hence, some differences exist between the correct vocal tract shape and those measured by FANT, which necessitates a careful interpretation of the evaluation results presented in this section, i.e. major trends should only be considered, and not the fine detail.

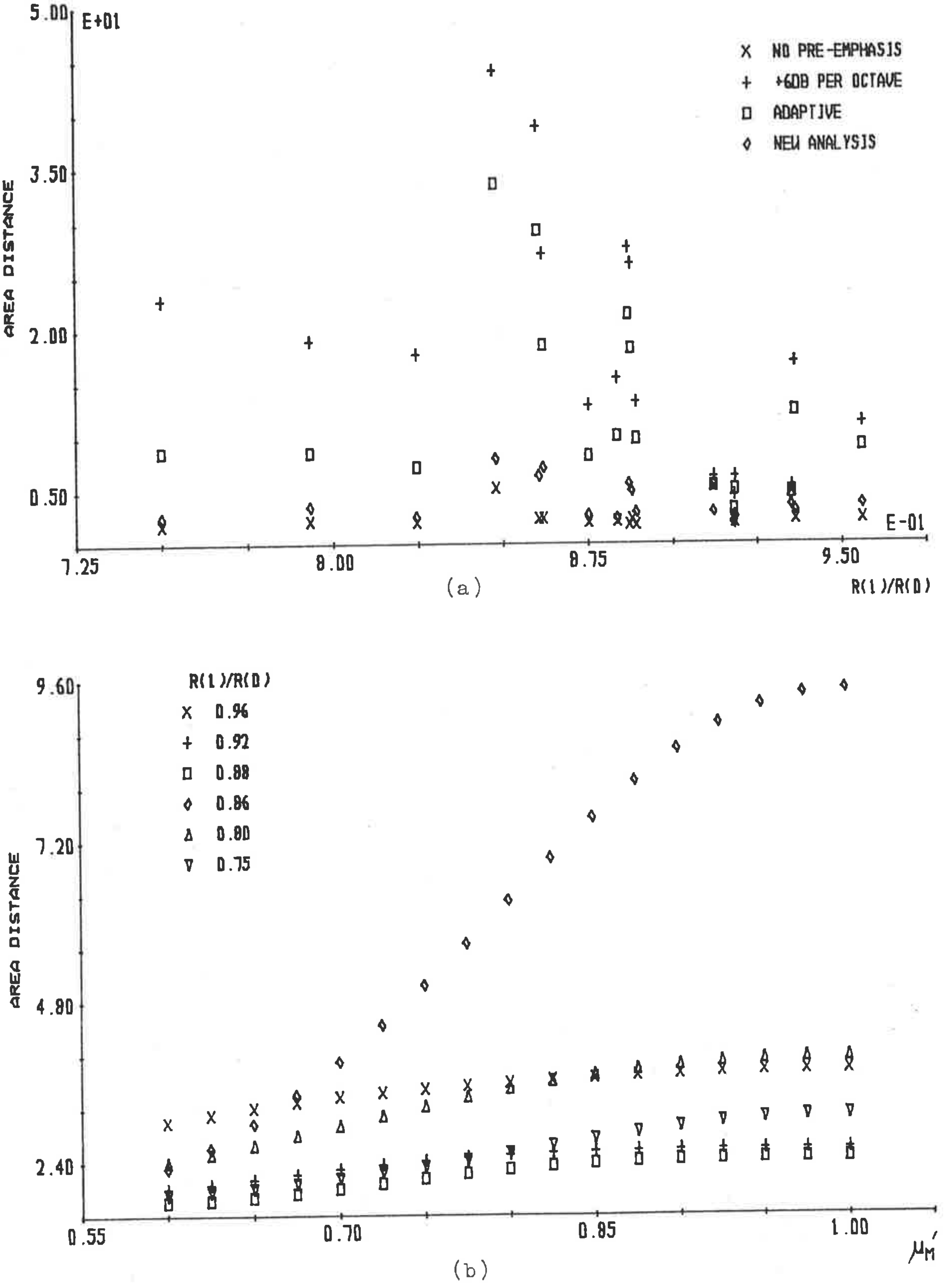
Real speech waveforms were obtained for the five vowel sounds |a|, |e|, |i|, |o| and |u| from seven Australian male speakers rephonating each of the vowels in a |h-d| frame. A number of recording sessions were used, separated by at least one day but not more than thirty days, and at each recording session the five vowels were spoken in a random order, which changed from one recording session to another. All the speech waveforms were recorded with a condensor microphone, and the resulting electrical analog signal low pass filtered to 4.5 kilohertz. The filtered analog signal was then sampled at a frequency of 10 kilohertz and stored on magnetic tape under the control of a minicomputer. A full description of the procedure and conditions under which the speech waveforms are digitally recorded is presented in Appendix H.

All the evaluation results presented in this section use 300 samples, i.e. a 30 msec time interval, of the speech waveform which is windowed by a Hamming window (BLACKMAN and TUKEY [1958], MARKEL [1971], MARKEL and WOLF [1972]). The windowed data is then analysed by the new speech analysis procedure, or the conventional pre-emphases followed by a conventional Parcor analysis (always implied, although no explicitly stated in many cases), to produce the evaluation results presented in this section.

The evaluation results are presented in two different forms, i.e. as plots of area distances versus $R(1)/R(0)$ and plots of area distances versus μ_M'. When a plot of area distances versus $R(1)/R(0)$ is presented, then the new speech analysis procedure has μ_M' equal to 0.8, and contains the results for the conventional pre-emphases followed by a conventional Parcor analysis. For plots of area distances versus μ_M', only results for the new speech analysis procedure are presented for various values of $R(1)/R(0)$. The evaluation results presented in these two plots provide information on the performance of the new speech analysis procedure to produce improved vocal tract shape recovery for a wide range of $R(1)/R(0)$ and μ_M' values.

Figure 7.14(a) presents area distances versus $R(1)/R(0)$ for the new speech analysis procedure, no pre-emphasis, a +6 dB per octave pre-emphasis and an adaptive pre-emphasis of real speech waveforms for the vowel |a|. The area distances for a +6 dB per octave pre-emphasis and an adaptive pre-emphasis are shown in Figure 7.14(a) to be much larger than those for the new speech analysis procedure and for no pre-emphasis. When $R(1)/R(0)$ is close to unity, then the area distances for a +6 dB per octave pre-emphasis and an adaptive pre-emphasis area similar, but as $R(1)/R(0)$ decreases from unity the difference between the area distances for these pre-emphases increases. This latter result is consistent with the evaluation results presented for the analysis of synthetic speech for the vowel |a|.

Except for a few cases, the area distances for the new speech analysis procedure are shown in Figure 7.14(a) to be similar to, but slightly larger than, the area distances for no pre-emphasis of real speech for the vowel |a|. Since the actual vocal tract



(a)

(b)

FIGURE 7.14: Analysis of real speech for the vowel /a/ with (a) the new speech analysis (with $\mu_M'=0.8$) and conventional pre-emphases and (b) the new speech analysis process for various values of R(1)/R(0) and μ_M'.

shape is not known, the evaluation results presented in Figure 7.14(a) do not necessarily show that no pre-emphasis provides better vocal tract shape recovery than the new speech analysis procedure. However, it can be concluded that the new speech analysis procedure produces better vocal tract shape recovery for the vowel sound $|a|$ than a +6 dB per octave pre-emphasis and an adaptive pre-emphasis. The variation in area distances that occurs for the new speech analysis procedure is similar to that for no pre-emphasis, but considerably less than that for a +6 dB per octave pre-emphasis or an adaptive pre-emphasis.

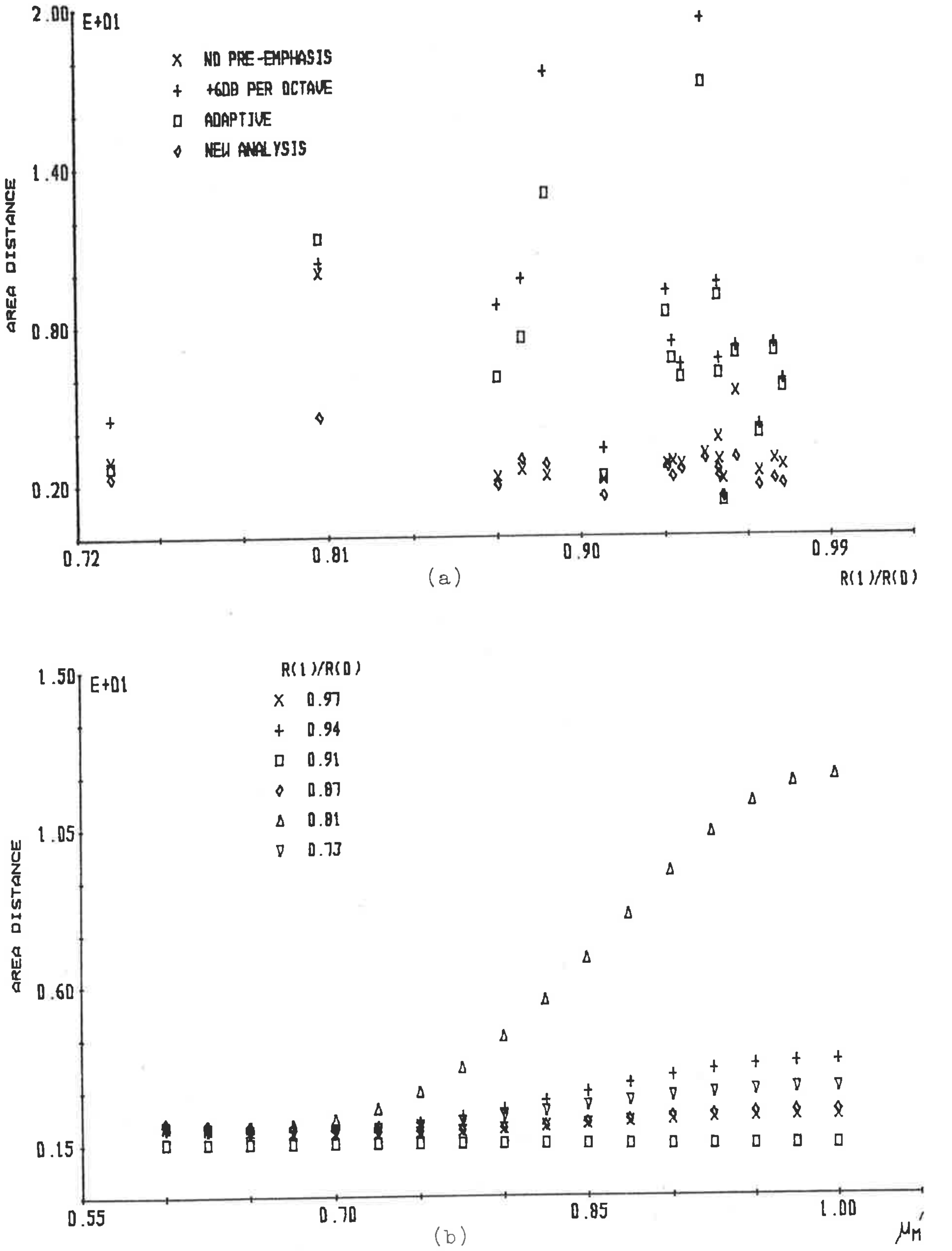
A plot of area distances versus μ_M' for various values of $R(1)/R(0)$ is presented in Figure 7.14(b) for the new speech analysis procedure when used to analyse real speech for the vowel $|a|$. The particular values of $R(1)/R(0)$ considered in Figure 7.14(b) were chosen so that the range of $R(1)/R(0)$ presented in Figure 7.14(a) is spanned at approximately equal intervals. With the exception of $R(1)/R(0) = 0.86$, only a small change in area distances is observed in Figure 7.14(b) for μ_M' changing from unity to 0.6. A decreasing area distance is observed in Figure 7.14(b) for all values of $R(1)/R(0)$ as μ_M' decreases from unity, suggesting that an improvement in vocal tract shape recovery may be possible if small values of μ_M' are used.

A comparison of Figure 7.14(a) and 7.14(b) shows that the variation in area distances produced by μ_M' changing from 0.6 to unity is small compared with the differences between the area distances for the new speech analysis procedure and those of a +6 dB per octave pre-emphasis or an adaptive pre-emphasis. Therefore, in general, the performance of the new speech analysis procedure in comparison with a +6 dB per octave pre-emphasis or an adaptive pre-emphasis is not significantly affected by μ_M' varying between 0.6 and unity.

The new speech analysis procedure has been shown to produce improved vocal tract shape recovery in comparison with the conventional +6 dB per octave and adaptive pre-emphases, but similar vocal tract shape recovery in comparison with no pre-emphasis, when analysing real speech for the vowel |a|. The variation in area distances for the new speech analysis was shown to be relatively small for changes in both $R(1)/R(0)$ and μ_M'. The evaluation results presented suggest that the new speech analysis procedure may produce improved vocal tract shape recovery from real speech of the vowel |a|, if μ_M' is less than 0.6.

Figure 7.15(a) presents the area distances versus $R(1)/R(0)$ for no pre-emphasis, a +6 dB per octave pre-emphasis, an adaptive pre-emphasis and the new speech analysis procedure used to analyse real speech of the vowel |e|. In general, Figure 7.15(a) shows the area distances for the +6 dB per octave and adaptive pre-emphasis to be much larger than the area distances for no pre-emphasis and the new speech analysis procedure. The area distances for a +6 dB per octave pre-emphasis are generally larger than those for an adaptive pre-emphasis of the real speech for the vowel |e|. These observations are consistent with those presented previously for analysis of real speech for the vowel |a|.

Except for a few cases, the area distances for the new speech analysis of real speech for the vowel |e| are shown in Figure 7.15(a) to be smaller than those for no pre-emphasis. The large



(a)

(b)

FIGURE 7.15: Analysis of real speech for the vowel /e/ with (a) the new speech analysis process (with $\mu_M'=0.8$) and conventional pre-emphases and (b) the new speech analysis process for various values of $R(1)/R(0)$ and μ_M'.

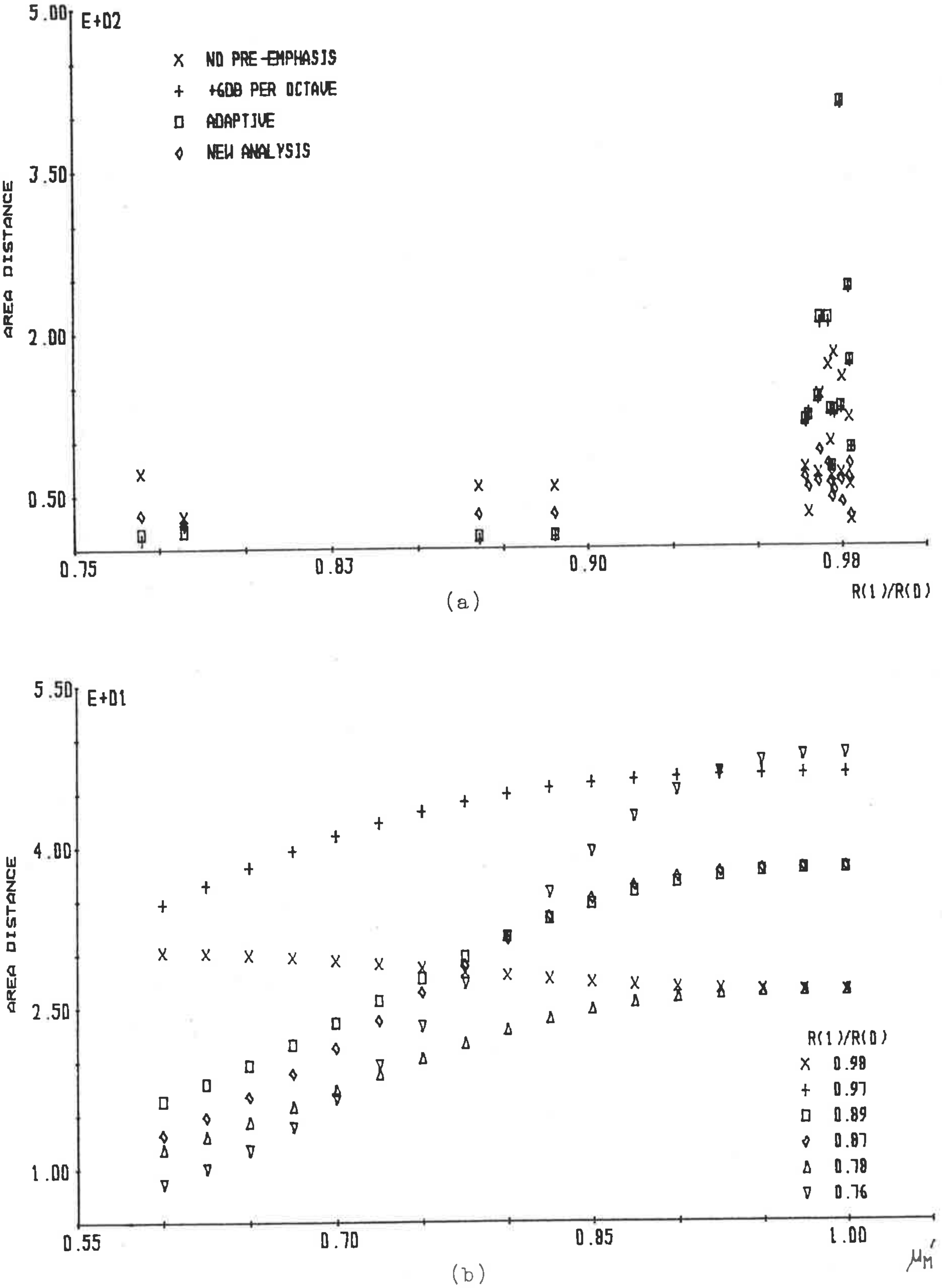
reduction in area distances when the area distances for the new speech analysis procedure are compared with those for the conventional +6 dB per octave and adaptive pre-emphasis (see Figure 7.15(a)) indicates that the new speech analysis procedure produces improved vocal tract shape recovery in comparison with those conventional pre-emphases, from real speech of the vowel |e|. These observations are consistent with those presented previously for analysis of real speech for the vowel |a|.

A plot of area distances versus μ_M' for various values of $R(1)/R(0)$ is presented in Figure 7.15(b), when the new speech analysis procedure is used to analyse real speech for the vowel |e|. The values of $R(1)/R(0)$ considered in Figure 7.15(b) were chosen so that the range of $R(1)/R(0)$ presented in Figure 7.15(a) is spanned at approximately equal intervals. Except when $R(1)/R(0) = 0.81$, only a small change in area distances occurs for μ_M' changing from 0.6 to unity. The area distances presented in Figure 7.15(b) show, in general, a decrease to a minimum and then an increase as μ_M' decreases from unity. The value of μ_M' for which the minimum area distance occurs varies between 0.65 and 0.725 and, hence, a small improvement in vocal tract shape recovery may be achieved by using $\mu_M' = 0.7$ instead of 0.8 (as used to generate Figure 7.15(a)) in the new speech analysis procedure. Except when $R(1)/R(0) = 0.81$, the performance of the new speech analysis procedure to recover vocal tract shapes, in comparison with that for conventional pre-emphasis, is only slightly affected by μ_M' changing from 0.6 to unity.

The new speech analysis process was shown to produce improved vocal tract shape recovery when compared with the conventional +6 dB per octave and adaptive pre-emphases, and similar vocal tract recovery to that for no pre-emphasis of real speech for the vowel |e|. This is consistent with the conclusion drawn for analysis of real speech for the vowel |a|. The area distances resulting from the new speech analysis procedure have a relatively small variation, in comparison with those for a conventional pre-emphasis, over the ranges of $R(1)/R(0)$ and μ_M' considered in Figure 7.15. The smallest area distances for the new speech analysis procedure are most likely to occur for μ_M' near 0.7 when analysing real speech of the vowel |e|.

For real speech of the vowel |i|, Figure 7.16(a) presents the area distances versus $R(1)/R(0)$ for the new speech analysis procedure, no pre-emphasis, a +6 dB per octave pre-emphasis and an adaptive pre-emphasis. Figure 7.16(a) shows that the +6 dB per octave and the adaptive pre-emphases produce larger area distances when $R(1)/R(0)$ is greater than 0.9, and smaller area distances when $R(1)/R(0)$ is less than 0.9, in comparison with the area distances for no pre-emphasis and the new speech analysis procedure. This observation is not consistent with that for the analysis of real speech for the vowels |a| and |e|.

With only a few exceptions, the area distances for the new speech analysis procedure are shown in Figure 7.16(a) to be smaller than those for no pre-emphasis. When $R(1)/R(0)$ is greater than 0.9, then the area distances for the new speech analysis procedure are smaller than those for the +6 dB per octave and adaptive pre-emphases, but slightly larger when $R(1)/R(0)$ is less than 0.9. The



(a)

(b)

FIGURE 7.16: Analysis of real speech for the vowel /i/ with (a) the new speech analysis process (with $\mu_M'=0.8$) and conventional pre-emphases and (b) the new speech analysis process for various values of R(1)/R(0) and μ_M'.

variation in area distances for the new speech analysis procedure is shown in Figure 7.16(a) to be smaller than the variation in area distances for the conventional pre-emphases. This observation is consistent with that for the anlaysis of real speech for the vowels |a| and |e|.

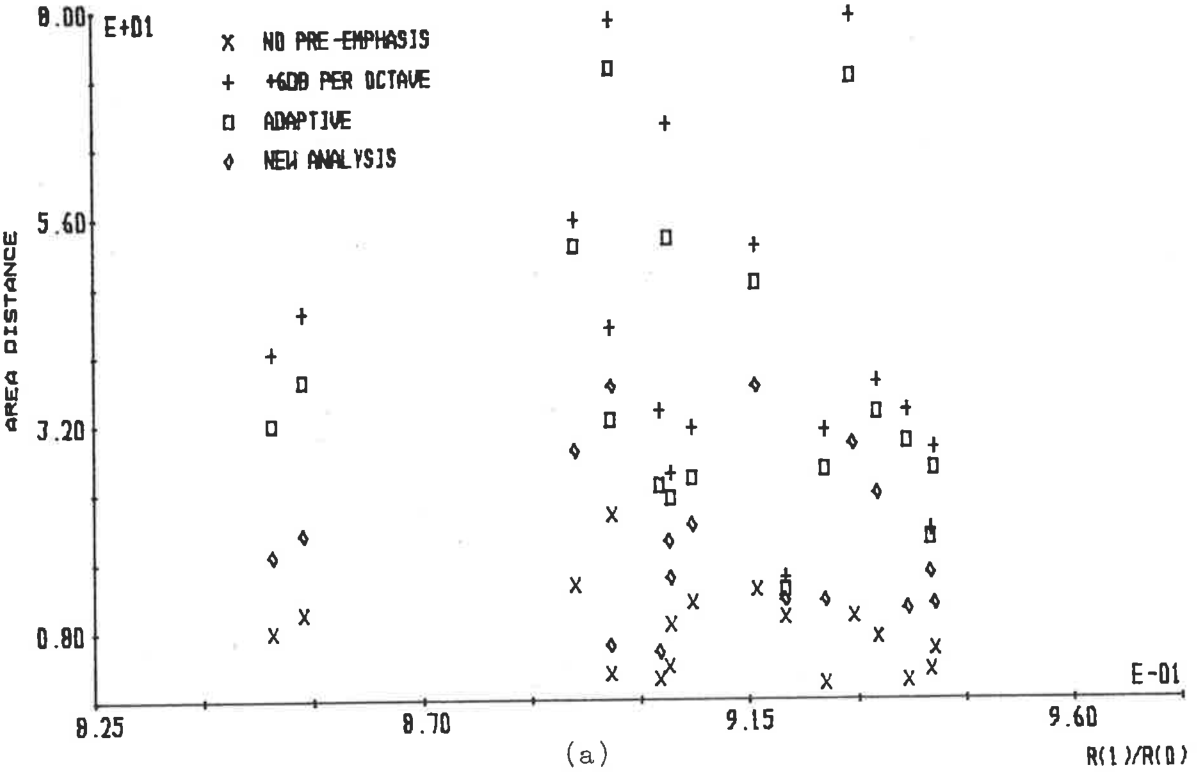
The area distances versus μ_M' for various values of $R(1)/R(0)$ are presented in Figure 7.16(b) for the new speech analysis procedure used to analyse real speech for the vowel |i|. The values of $R(1)/R(0)$ presented in Figure 7.16(b) span the range of $R(1)/R(0)$ presented in Figure 7.16(a). In contrast with the evaluation results presented in Figures 7.14(b) and 7.15(b), a relatively large change in area distances occurs when μ_M' changes from 0.6 to unity. With the exception of the area distances for $R(1)/R(0) = 0.98$, all the area distances decrease as μ_M' decrease from unity, suggesting that small area distances may occur for μ_M' less than 0.6. A comparison of the area distances for $\mu_M' = 0.6$ presented in Figure 7.16(b) with those presented in Figure 7.16(a) suggests that the area distances for the new speech analysis procedure may be the smallest if μ_M' is less than 0.6. In general, the variation in area distances for the new speech analysis procedure is smaller than that for the conventional pre-emphasis techniques, which is consistent with that for analysis of real speech for the vowels |a| and |e|.

The evaluation results presented in Figure 7.16 show that, in general, the new speech analysis procedure produces a small improvement in vocal tract shape recovery in comparison with conventional pre-emphasis techniques. The variation in area distances for the new speech analysis procedure was shown to be smaller than that for the conventional pre-emphases. Except when $R(1)/R(0)$ is close to unity, the smallest area distances for the new speech analysis procedure are most likely to occur when μ_M' is less than 0.6.

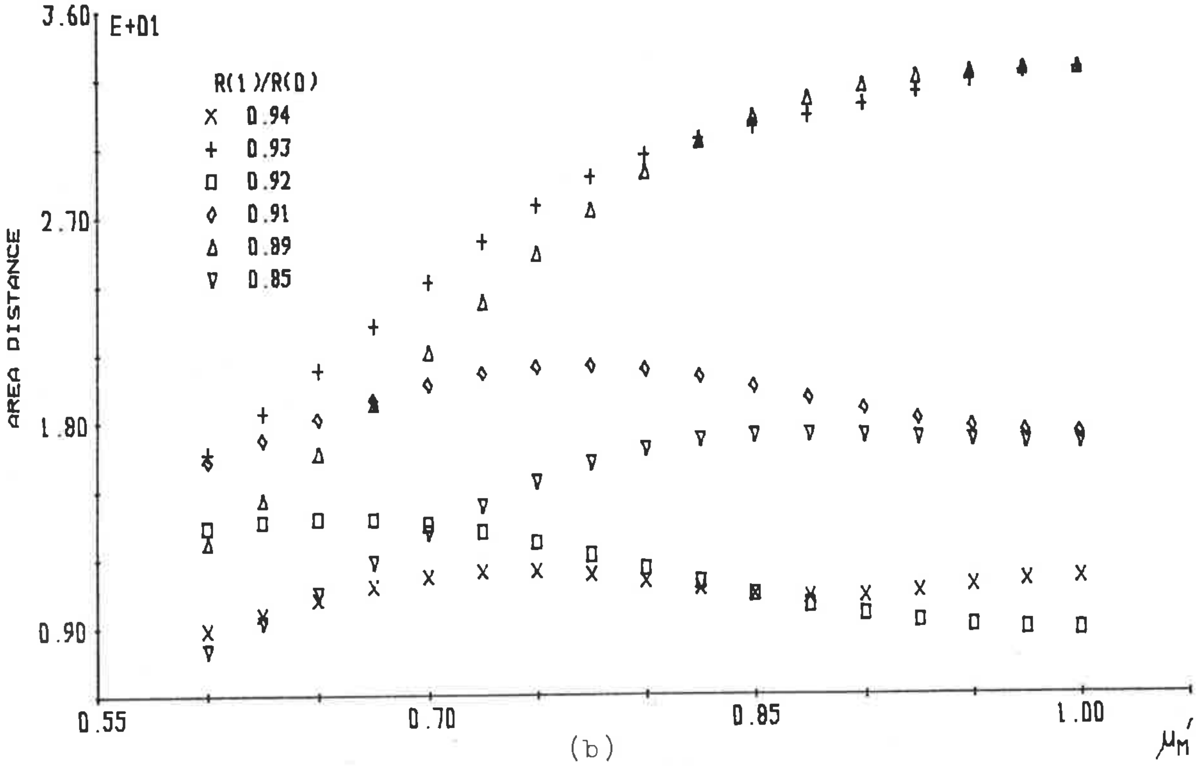
The area distances versus $R(1)/R(0)$ are presented in Figure 7.17(a) for the real speech of the vowel |o| analysed by the new speech analysis procedure, no pre-emphasis and the +6 dB per octave and adaptive pre-emphases. The area distances for the +6 dB per octave and adaptive pre-emphases are shown in Figure 7.17(a) to be larger than those of the new speech analysis procedure and no pre-emphasis. In general, the area distances for a +6 dB per octave pre-emphasis are larger than those for the other analyses considered in Figure 7.17(a). These observations are consistent with those for the analysis of real speech for the vowels |a| and |e|.

Figure 7.17(a) shows that the area distances for no pre-emphasis are smaller than those for the new speech analysis procedure (which uses $\mu_M' = 0.8$), and therefore, under the conditions used to generate the evaluation results for Figure 7.17(a), the new speech analysis procedure may not produce the best vocal tract shape recovery. A slightly larger variation in area distances is observed in Figure 7.17(a) for the new speech analysis procedure in comparison with no pre-emphasis, but a much smaller area distance variation in comparison with the +6 dB per octave and adaptive pre-emphases.

Figure 7.17(b) presents the area distances versus μ_M' for various values of $R(1)/R(0)$ when the new speech analysis procedure analyses real speech for the vowel |o|. The values of $R(1)/R(0)$ considered in Figure 7.17(b) span the values of $R(1)/R(0)$ presented



(a)



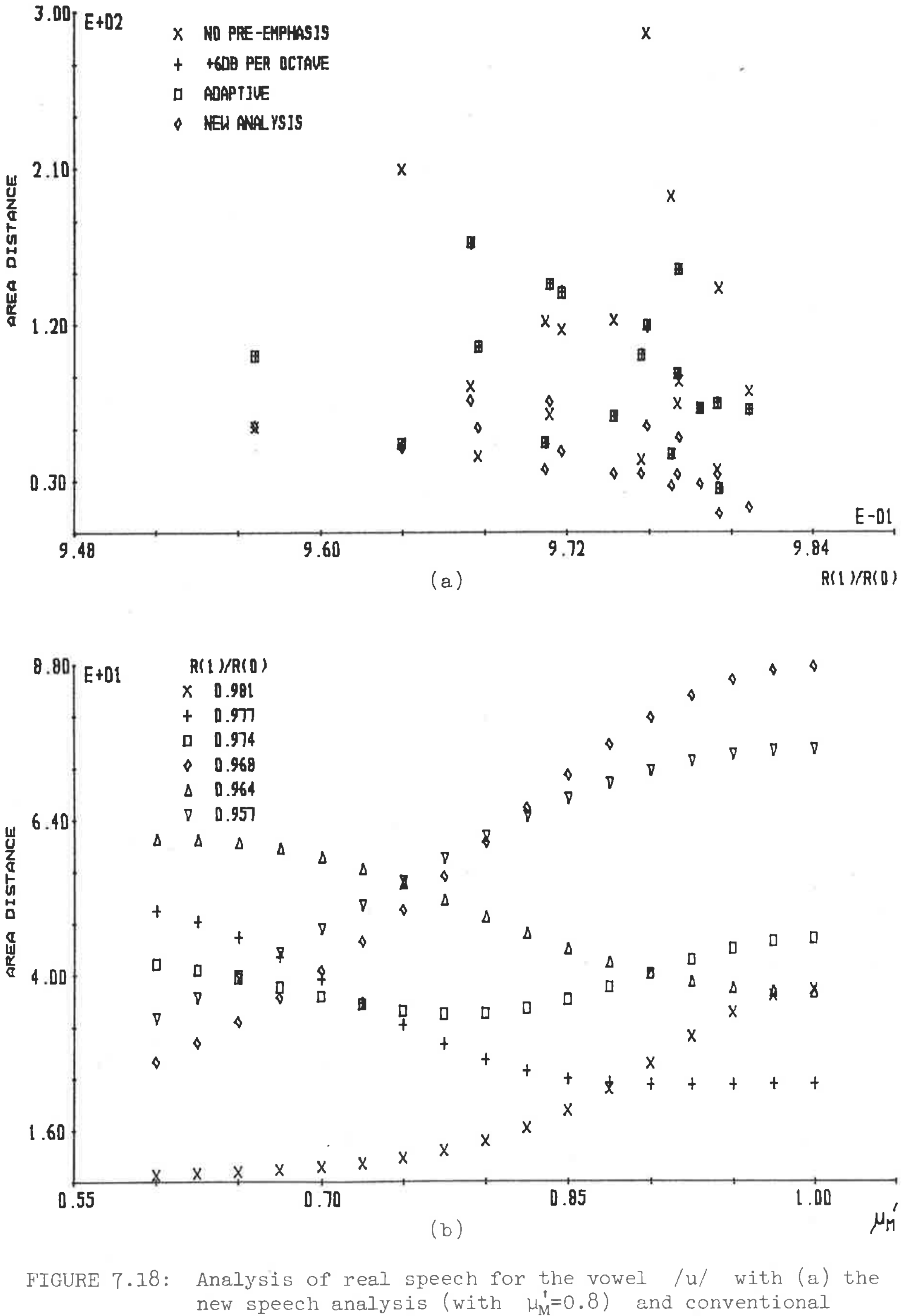
(b)

FIGURE 7.17: Analysis of real speech for the vowel /o/ with (a) the new speech analysis process (with $\mu_M'=0.8$) and conventional pre-emphases and (b) the new speech analysis process for various values of R(1)/R(0).

in Figure 7.17(a) at approximately equal increments. A number of different trends are observed in Figure 7.17(b) for area distances versus μ_M', but, in general, it may be concluded that smaller area distances occur when μ_M' is less than 0.6. When the area distances presented in Figure 7.17(b) for $\mu_M' = 0.6$ are compared with those presented in Figure 7.17(a) for the conventional pre-emphases, the new speech analysis procedure is found to produce, in general, the smallest area distances. Therefore, when μ_M' is less than 0.6, improved vocal tract shape recovery is achieved by the new speech analysis procedure in comparison with conventional pre-emphases. In general, Figure 7.17(b) shows that the variation in area distances for the new speech analysis procedure is not as large as for other conventional pre-emphases presented in Figure 7.17(a).

From the evaluation results presented in Figure 7.17, the new speech analysis procedure produces a significant improvement in vocal tract shape recovery when compared with the conventional +6 dB per octave and adaptive pre-emphases, and, in many cases, a small improvement in vocal tract shape recovery in comparison with no pre-emphasis of real speech for the vowel |o|. The variations in area distances for the new speech analysis procedure are much smaller than those for a +6 dB per octave and adaptive pre-emphasis, but similar to that for no pre-emphasis. The best vocal tract shape recovery is most likely to occur when the new speech analysis procedure is used with μ_M' less than 0.6.

Figure 7.18(a) presents the area distances versus $R(1)/R(0)$ for analysis of real speech for the vowel |u| by the new speech analysis procedure and for no pre-emphasis, the +6 dB per octave and adaptive pre-emphases. The range of $R(1)/R(0)$ presented in



(a)

(b)

FIGURE 7.18: Analysis of real speech for the vowel /u/ with (a) the new speech analysis (with $\mu_M'=0.8$) and conventional pre-emphases and (b) the new speech analysis process for various values of R(1)/R(0) and μ_M'.

Figure 7.18(a) is much smaller than the ranges of $R(1)/R(0)$ for the other vowels considered in this section. In general, Figure 7.18(a) shows the area distances for the +6 dB per octave and adaptive pre-emphases to be similar to each other, but much smaller than the area distances for no pre-emphasis. This observation contrasts with those for the other vowels considered in this section.

The new speech analysis procedure is shown in Figure 7.18(a) to produce the smallest area distances in all but a few cases. In general, a large reduction in area distances occurs when the area distances for the new speech analysis procedure are compared with those for no pre-emphasis, which differs from the observations for the other vowels when similar area distances occurred. The variation in area distances for the new speech analysis procedure is shown in Figure 7.18(a) to be significantly smaller than that for the conventional pre-emphases. This observation is consistent with that for the other vowels considered in this section.

Figure 7.18(b) presents area distances versus μ_M' for various values of $R(1)/R(0)$ for the new speech analysis procedure when used to analyse real speech for the vowel $|u|$. The values of $R(1)/R(0)$ considered in Figure 7.18(b) are chosen to span the range of $R(1)/R(0)$ presented in Figure 7.18(a) with equal increments. A number of different trends of area distance versus μ_M' are observed in Figure 7.18(b) which prevent a general conclusion being made about the value of μ_M' that produces the smallest area distances for the range of $R(1)/R(0)$ considered in Figure 7.18(b). The variation of area distances observed in Figure 7.18(b) is small when compared with those presented in Figure 7.18(b); hence, the performance of the new speech analysis procedure to produce improved vocal tract shape recovery is not significantly affected by a change in μ_M' between 0.6 and unity when analysing real speech for the vowel |u|.

The evaluation results presented in Figure 7.18 show that the new speech analysis procedure is expected to produce improved vocal tract shape recovery in comparison with no pre-emphasis and the +6 dB per octave and adaptive pre-emphases. The variation in area distances for the new speech analysis procedure is also shown in Figure 7.18 to be significantly smaller than that for no pre-emphasis and the +6 dB per octave and adaptive pre-emphases. Hence, improved vocal tract shape recovery with a relatively small variation in the recovered vocal tract shape occurs for the new speech analysis procedure when analysing real speech for the vowel |u|.

The evaluation results presented in this section have shown that the new speech analysis procedure produces, for the analysis of real speech for the vowel |a|, |e|, |o| and |u|, much smaller area distances than the +6 dB per octave and adaptive pre-emphases. For analysis of real speech for the vowel |i|, the new speech analysis procedure produces similar area distance to those for the +6 dB per octave and adaptive pre-emphases. The area distances for the new speech analysis procedure and no pre-emphasis were shown to be similar for all vowels except the vowel |u|, where the area distances for the new speech analysis procedure are much smaller than those for no pre-emphasis. Hence, it can be concluded that, for all the real speech considered in this section, the new speech analysis procedure is significantly better at producing smaller area distances and, hence, improved vocal tract shape recovery, than conventional pre-emphases followed by a conventional Parcor analysis.

In general, the variation in area distances for the new speech analysis procedure was shown to be small when μ_M' is between 0.6 and unity, hence only causing small changes in the performance of the new speech analysis procedure to provide improved vocal tract shape recovery. It was shown that a small reduction in the area distances for the new speech analysis procedure may occur when analysing real speech for the vowels |a|, |e|, |i| and |o|, with μ_M' less than 0.6. For analysis of real speech for the vowel |u|, no general reduction in area distances was observed for μ_M' between 0.6 and unity. Therefore, a slight improvement in the performance of the new speech analysis procedure may occur if μ_M' is approximately 0.6 or less.

In conclusion, the evaluation results presented in this section for analysis of real speech for five vowel sounds from seven Australian male speakers have shown a general improvement in vocal tract shape recovery when the new speech analysis procedure is used instead of conventional pre-emphases followed by a conventional Parcor analysis. In general, the variation in area distances for the new speech analysis procedure is smaller than that for conventional pre-emphases followed by a conventional Parcor analysis. This is an important property of the new speech analysis procedure for application to areas where slowly varying parameters are required to be extracted from the speech waveform, e.g. low bit rate transmission systems. This section has, therefore, shown a distinct advantage for using the new speech analysis procedure in preference to conventional pre-emphases followed by a conventional Parcor analysis if vocal tract shape recovery is required.

7.5 SUMMARY

This chapter combined the two/three adaptive pre-emphasis filter, developed in Chapter 4, with the Parcor Lossy Termination analysis, developed in Chapter 6, to form a new speech analysis procedure. The two/three adaptive pre-emphasis filter was defined to remove glottal pulse effects from the recovered acoustic tube shape, and the Parcor Lossy Termination analysis was defined to remove radiation effects from the recovered acoustic tube shape.

The new speech analysis procedure was first evaluated with synthetic speech generated from the ideal acoustic tube model, with glottal pulse waveforms as excitations and a lossy termination, for five acoustic tube shapes which approximate vocal tract shapes for the vowels |a|, |e|, |i|, |o| and |u|. The value of the termination reflection coefficient used in the Parcor Lossy Termination analysis was denoted as μ_M', and the initial evaluations were performed with μ_M' equal to the μ_M used to generate the synthetic speech. The glottal pulse waveforms used as excitations were chosen such that the value of $R(1)/R(0)$ varied from zero to unity.

In general, the new speech analysis procedure was shown to produce smaller area distances than no pre-emphasis and the +6 dB per octave and adaptive pre-emphases followed by a conventional Parcor analysis. Except for a few isolated cases, the new speech analysis procedure always produces smaller area distances than a +6 dB per octave pre-emphasis. Whenever $R(1)/R(0)$ is less than 0.9, similar area distances occur for the new speech analysis procedure, no pre-emphasis and an adaptive pre-emphasis, with the exception of the vowel |i|.

The variation in area distances for the new speech analysis procedure was shown to be considerably smaller than that for the conventional pre-emphases, for all the five vowel sounds. Therefore, the acoustic tube shape recovered by the new speech analysis procedure is less sensitive to changes in glottal pulse excitation than the acoustic tube shapes recovered by conventional pre-emphases followed by a conventional Parcor analysis.

Evaluation results were presented for the new speech analysis procedure used to analyse synthetic speech of five vowels for the situation where μ_M' is not equal to the termination reflection coefficient used to generate the synthetic speech. In general, the trends of area distances as μ_M' changes were shown to differ for different vowel sounds. Area distances were shown to decrease as μ_M' decreases from unity for the vowels |a| and |o|, but to increase as μ_M' decreases from unity for the vowel |u|. For the vowel sound |e|, the area distances were shown to decrease to a minimum and then increase as μ_M' decreases, and for the vowel sound |i|, increasing to a maximum and then decreasing area distances were shown as μ_M' decreases from unity.

In general, the variation in area distances for the new speech analysis procedure was shown to be relatively small when μ_M' and the termination reflection coefficient used to generate synthetic speech do not differ widely. Hence, the performance of the new speech analysis procedure to produce a reduction of area distances when compared with the area distances for conventional pre-emphases is not significantly affected in that situation. If μ_M' and the termination reflection coefficient used to generate synthetic speech differ widely, poor acoustic tube shape recovery, i.e. large area distances, may result.

The new speech analysis procedure was evaluated with real speech of the five vowel sounds |a|, |e|, |i|, |o| and |u| from seven Australian male speakers. These evaluations show that the new speech analysis procedure produces much smaller area distances than those for the conventional +6 dB per octave and adaptive pre-emphases followed by a conventional Parcor analysis, for real speech of the vowels |a|, |e|, |o| and |u|. For the real speech of the vowel |i|, similar area distances occur for the new speech analysis procedure and the conventional +6 dB per octave and adaptive pre-emphases followed by a conventional Parcor analysis. Similar area distances occur for the new speech analysis procedure and no pre-emphasis for all vowels except the vowel sound |u|, when the area distances for the new speech analysis procedure are much smaller.

A relatively small change in area distances was shown when μ_M' varies from 0.6 to unity, which, in general, does not significantly affect the performance of the new speech analysis procedure in comparison with conventional pre-emphasis methods. The evaluation results presented showed that the smallest area distances for the new speech analysis procedure are most likely to occur when μ_M' is approximately 0.6 or, in some situations, less than 0.6.

A comparison of the evaluation results for synthetic speech and real speech waveforms shows a consistency of the evaluation results for all five vowel sounds. This suggests that the general trends occurring for the evaluations with synthetic speech for $R(1)/R(0)$ between zero and unity may also occur for real speech if $R(1)/R(0)$ were extended to cover that range. Although the general trends of area distance versus μ_M' are different for real and synthetic speech, the relatively small variation in area distances when μ_M' changes from 0.6 to unity shows that the performance of the new speech analysis procedure, in comparison with conventional pre-emphasis methods, is not significantly affected by changes in μ_M'.

In general, the evaluation results presented in this chapter have shown that the new speech analysis procedure produces similar, or significantly smaller area distances than conventional pre-emphases followed by a conventional Parcor analysis when analysing real and synthetic speech of the vowel sounds |a|, |e|, |i|, |o| and |u|. The smaller area distances produced by the new speech anaylsis procedure generally imply improved vocal tract/acoustic tube shape recovery. The variation in area distances for the new speech analysis procedure is much smaller than that for conventional pre-emphases followed by a conventional Parcor analysis, when analysing real and synthetic speech of five vowel sounds. This latter conclusion is important when the new speech analysis procedure is applied to areas where slowly varying parameters are required to be extracted from the speech waveform, e.g. low bit rate speech transmission systems.

In conclusion, this chapter has shown that there exists a distinct advantage in using the new speech analysis procedure in preference to no pre-emphasis, a +6 dB per octave pre-emphasis and an adaptive pre-emphasis followed by a conventional Parcor analysis if vocal tract/acoustic tube shape recovery is required from real or synthetic speech waveforms.

CHAPTER 8

CONCLUSIONS

This thesis has examined the linear prediction/acoustic tube model and the linear prediction analysis procedures to develop new analysis procedures which produce improved acoustic tube shape recovery from the waveform radiated from a set of acoustic tubes. These new analysis procedures consider some of the properties of the human vocal tract and speech production mechanism, so that improved vocal tract shape recovery is possible from the speech waveform.

A lossless acoustic tube model and linear predictive analysis techniques were reviewed, and the relationship between the results of a linear predictive analysis and the lossless acoustic tube model were defined. The necessary conditions that a set of acoustic tubes must satisfy for a linear predictive analysis to recover its shape were defined. Of these necessary conditions, many are not satisfied in the human vocal tract or during the production of speech, so a linear predictive analysis of the speech waveform does not identify an acoustic tube shape which is similar to the vocal tract shape. Therefore, if vocal tract shape recovery is required from the speech waveform, new analysis procedures must be developed, and those presented in this thesis achieve improved vocal tract shape recovery.

To determine whether a new analysis procedure produces an improvement in acoustic tube/vocal tract shape recovery, it is necessary to have a quantitative measure of the similarity or dissimilarity of two acoustic tube shapes. Many of the existing distance measures used in the speech field, e.g. those for speech recognition, were found to be unsuitable, and so a new distance measure was defined. The new distance measure is basically a normalized Euclidean distance, and is called an area distance because it produces a scalar quantity which is the measure of similarity or dissimilarity between the cross-sectional areas of two acoustic tube shapes. The normalization of the area distance is performed with respect to the original acoustic tube shape, so that the area distances for a number of acoustic tube shapes recovered by different analysis procedures can be compared, to determine which analysis procedure produces the best acoustic tube shape recovery.

Linear prediction requires that a set of acoustic tubes be excited by a white excitation waveform for it to recover the acoustic tube shape but, in general, and especially for the production of speech, the excitation is non-white; therefore, investigations were performed to determine the effects of non-white excitation on the acoustic tube shape recovered by linear prediction. The results of these investigations showed that, in general, poor acoustic tube shape recovery occurs when the duration of the excitation is much larger than the round trip propagation delay within a single acoustic tube.

Investigations performed with simple acoustic tube shapes, i.e. ones with relatively few cross-sectional area changes, led to an understanding of the effects of certain non-white excitations on the acoustic tube shape recovered by linear prediction. As a result of these investigations, special acoustic tube shapes were defined, and two analysis procedures were presented which produce good special acoustic tube shape recovery for certain non-white excitations. Special acoustic tube shapes and their recovery procedures do not have any direct application to the recovery of vocal tract shapes from the speech waveform. However, the investigations performed provide an understanding of the manner by which linear prediction, a basic speech analysis tool, identifies acoustic tube shapes from an autocorrelation function.

In general, it was shown that very poor acoustic tube shape recovery occurs when conventional linear prediction analyses waveforms generated from sets of acoustic tubes excited by glottal pulse waveforms which are similar to those used to excite the vocal tract during the production of voiced speech. A number of procedures have been proposed to overcome this poor acoustic tube shape recovery, and the effectiveness of these existing procedures to provide accurate acoustic tube shape recovery was evaluated. These evaluations showed, in general, that accurate acoustic tube shape recovery was not achieved when the wide range of glottal pulse waveforms that occur for the production of the wide range of voiced sounds and from different speakers is considered.

A new adaptive pre-emphasis filter, called the two/three adaptive pre-emphasis filter, was designed specifically to provide improved acoustic tube shape recovery when glottal pulse excitation is used, or is present. The two/three adaptive pre-emphasis filter consists of two parts, one part defined by a parameter α, corrects for glottal pulse spectral slopes between zero and -12 dB per octave, and the other part, defined by the parameter β, corrects for glottal pulse spectral slopes between -12 dB and -18 dB per octave. The values of α or β used in a particular analysis situation are obtained from empirical expressions between the parameters α and β, and the value of $R(1)/R(0)$, where the latter is evaluated from the autocorrelation function of the waveform being pre-emphasized. These empirical expressions are determined as the best fit curves to the data points $(\alpha', R(1)/R(0))$ and $(\beta', R(1)/R(0))$, where α' and β' define the two/three adaptive pre-emphasis filter such that the glottal pulse waveform when pre-emphasized by the two/three adaptive pre-emphasis filter and followed by a linear predictive analysis results in the smallest area distance between recovered and original acoustic tube shapes.

A number of different waveform sampling frequencies between 10 and 7 kilohertz, inclusive, were considered, and were shown to affect the empirical relationships between α' and $R(1)/R(0)$, and β' and $R(1)/R(0)$, to various degrees. The empirical relationship between α' and $R(1)/R(0)$ was found to change significantly when a sampling frequency near 7 kilohertz is used. Therefore, one relationship defines α' in terms of $R(1)/R(0)$ when the waveform sampling frequency is between 10 and 8 kilohertz, inclusive, and another when the waveform sampling frequency is 7 kilohertz. The empirical relationship between β' and $R(1)/R(0)$ was found to change significantly as the waveform sampling frequency changes, and so empirical relationships were defined for waveform sampling frequencies of 10, 9, 8 and 7 kilohertz.

The two/three adaptive pre-emphasis filter was evaluated by comparing the area distances for no pre-emphasis, a +12 dB per octave pre-emphasis, an unvoiced/voiced adaptive pre-emphasis and the two/three adaptive pre-emphasis followed by a conventional linear predictive analysis. Evaluations performed with glottal pulse waveforms derived from glottal pulse models and from measured glottal pulse waveforms, for a sampling frequency of 10 kilohertz, showed that the two/three adaptive pre-emphasis filter is significantly better at accounting for glottal pulse waveforms than existing pre-emphasis methods. Evaluations were presented for glottal pulse waveforms sampled at 9, 8 and 7 kilohertz, which showed that the two/three adaptive pre-emphasis filter has been designed such that changes in sampling frequencies between 10 and 7 kilohertz, inclusive, do not significantly affect the performance of the two/three adaptive pre-emphasis filter to account for glottal pulse excitation waveforms.

Evaluations were performed with synthetic speech waveforms generated from an ideal acoustic tube model, the shape of which approixmates the vocal tract shapes for the vowels |a|, |e|, |i|, |o| and |u|, and which are excited with glottal pulse waveforms. These evaluations showed a significant reduction in area distances for the two/three adaptive pre-emphasis filter in comparison with the area distances for the +12 dB per octave and unvoiced/voiced adaptive pre-emphases, for all five vowel sounds. The two/three adaptive pre-emphasis, when compared with no pre-emphasis, was shown to provide a reduction in area distances for the vowel sounds |a|, |e| and |i|. Slightly larger area distances occur for the two/three adaptive pre-emphasis in comparison with no pre-emphasis for the vowel sounds |o| and |u|.

Evaluation results were presented for the two/three adaptive pre-emphasis of synthetic speech sampled at frequencies of 10, 9, 8 and 7 kilohertz, which showed no significant change in the performance of the two/three adaptive pre-emphasis filter for vowel sounds |a|, |e|, |i| and |o|. For synthetic speech of the vowel |u|, a significant change in the performance of the two/three adaptive pre-emphasis filter was observed for changes in sampling frequency between 10 and 7 kilohertz, inclusive.

In general, the evaluations presented for the analysis of synthetic speech for five vowels showed that the variations in area distances for the two/three adaptive pre-emphasis filter are much smaller than those for existing pre-emphasis methods. Therefore, the acoustic tube shape recovered after using a two/three adaptive pre-emphasis filter is less sensitive to changes in glottal pulse excitation waveforms than those for existing pre-emphasis methods.

Evaluation results were also presented for analysis of real speech for the five vowels |a|, |e|, |i|, |o| and |u| spoken by seven Australian males. In general, the two/three adaptive pre-emphasis filter was shown to produce a large reduction in area distances when compared with existing pre-emphases for the vowel sounds |a|, |e| and |o|, and similar area distances for the vowel sounds |i| and |u|. The variations in area distances for the two/three adaptive pre-emphasis were shown to be smaller than those for existing pre-emphases. A comparison of the evaluation results presented for real speech and synthetic speech waveforms showed a similarity of general trends for the area distances as a function of $R(1)/R(0)$ for each pre-emphasis method.

It can be concluded, from the evaluation results presented, that there exists a significant advantage in using the two/three adaptive pre-emphasis filter in preference to existing pre-emphasis techniques, if acoustic tube/vocal tract shape recovery is required from synthetic/real speech waveforms.

The radiation of acoustic waveforms from the open end of an acoustic tube was examined, and a model of the loss from the acoustic tube due to radiation described. A radiation impedance defined by a single zero at $z = 1$ and a magnitude term was found to be a good approximation to the correct radiation impedance, at all but high frequencies. Evaluation results were presented which showed that a conventional -6 dB per octave pre-emphasis followed by a conventional linear predictive analysis does not accurately recover the acoustic tube shape from synthetic speech waveforms generated with a lossy termination as described by the simplified radiation impedance. The conventional -6 dB per octave pre-emphasis was shown to accurately correct for the zero, but not the magnitude term of the simplified radiation impedance. When the magnitude term is defined in terms of the acoustic tube model termination reflection coefficient, μ_M, then a non-zero magnitude term, i.e. finite loss due to radiation, implies a non-unity termination reflection coefficient.

The basis for a new autocorrelation analysis procedure was presented, which permits a non-unity termination reflection coefficient, i.e. a loss at the termination of the acoustic tube model, and requires a comparable number of computations to that of a lattice formulation of linear prediction. When the termination reflection coefficient, μ_M, is non-unity, then the cross junction correlations are unknown (when $\mu_M = 1$ the cross junction correlations are zero), and therefore the junction between two acoustic tubes cannot be determined from the knowledge of acoustic waveforms in only one acoustic tube. Empirical investigations showed that an approximation of the cross junction correlations, which generally produce stable acoustic tube shape recovery, is to assume the cross junction correlations are much smaller than certain autocorrelations and cross tube correlations in one acoustic tube. This approximation resulted in the definition of four Lossy Termination autocorrelation analysis procedures.

The four Lossy Termination analyses were evaluated by a comparison of area distances with those for a conventional Parcor analysis, using synthetic speech for the vowel sounds |a|, |e|, |i|, |o| and |u|. The value of the termination reflection coefficient used in the Lossy Termination analyses was the same as that used to generate the synthetic speech. These evaluations showed that only one Lossy Termination analysis, called the Parcor Lossy Termination analysis, consistently produces smaller area distances than those for a conventional Parcor analysis for a wide range of termination reflection coefficients.

The Parcor Lossy Termination analysis was shown to produce smaller area distances than a conventional Parcor analysis, provided the value of the termination reflection coefficients used in the Parcor Lossy Termination analysis and to generate the synthetic speech do not differ widely. Evaluation results were presented for analysis of real speech waveforms for the vowel sounds |a|, |e|, |i|, |o| and |u|, and showed that the Parcor Lossy Termination analysis produces smaller area distances than a conventional Parcor analysis for the vowel sounds |e|, |i|, |o| and |u|, but slightly larger area distances than a conventional Parcor analysis for the vowel sound |a|. It was therefore concluded that there exists a significant advantage in using the Parcor Lossy Termination analysis in preference to a conventional Parcor linear predictive analysis when a loss occurs at the termination of the acoustic tubes or vocal tract.

An analysis procedure permitting a non-unity termination reflection coefficient, and based on a transfer function of the acoustic tube model, was presented. This analysis procedure places constraints, which are derived from physical restrictions of the vocal tract, on the recovered acoustic tube shape. In many cases, accurate acoustic tube shape recovery was obtained when using this analysis procedure, but problems with convergence and large numbers of complex computations prevent the analysis procedure from being used to analyse speech waveforms in real time applications.

The two/three adaptive pre-emphasis filter and the Parcor Lossy Termination analysis were combined to produce a new speech analysis procedure. The new speech analysis procedure was evaluated by comparing the area distances for the new speech analysis procedure with those for conventional pre-emphases, i.e. no pre-emphasis, a +6 dB per octave pre-emphasis and a conventional adaptive pre-emphasis of synthetic speech waveforms. In general, the new speech analysis procedure was shown to produce either smaller, or similar area distances to the conventional pre-emphases when the termination reflection coefficient used in the new speech analysis procedure is the same as that used to generate the synthetic speech. Evaluations also showed that the new speech analysis procedure produces, in general, smaller, or similar, area distances than those for conventional pre-emphasis, provided the termination reflection coefficient used in the new speech analysis procedure does not differ widely from that used to generate the synthetic speech.

The new speech analysis was evaluated with real speech waveforms for the vowel sounds |a|, |e|, |i|, |o| and |u| spoken by seven Australian males. These evaluations show that the new speech analysis procedure, in general, produces smaller area distances than the +6 dB per octave and adaptive pre-emphases. The area distances for the new speech analysis procedure are similar to those for no pre-emphasis, except for the vowel |u|, where they are much smaller than those for no pre-emphasis. The area distances for the new speech analysis procedure were shown to change by relatively small amounts for changes in the termination reflection coefficient used in the new speech analysis procedure.

The variation in area distances for the new speech analysis procedure was shown to be consistently smaller than the variation in area distances for the conventional pre-emphases when analysing either real or synthetic speech of the five vowel sounds |a|, |e|, |i|, |o| and |u|. This is an important property of the new speech analysis procedure when applied to areas where slowly varying parameters are required to be extracted from the speech waveform, e.g. low bit rate transmission systems.

In conclusion, it has been shown that there exists a significant advantage in using the two/three adaptive pre-emphasis filter, the Parcor Lossy Termination analysis, or a combination of both in the form of the new speech analysis procedure, in preference to conventional pre-emphases and linear predictive analyses, if acoustic tube/vocal tract shape recovery is required from synthetic or real speech waveforms.

APPENDIX A

IMPULSE RESPONSE OF TWO ACOUSTIC TUBES

The signal flow diagram for two commensurate acoustic tubes satisfying the assumptions of the acoustic tube model (see Chapter 2) is presented in Figure A.1. From Figure A.1, the following set of equations (i.e. A.1 to A.6) completely defines all the acoustic volume velocities within the acoustic tubes:

$$W_1^+(t) = W_0^+(t) \tag{A.1}$$

$$W_0^-(t) = W_1^-\left(t - \frac{\ell}{c}\right) \tag{A.2}$$

$$W_2^+(t) = (1 + \mu_1)W_1^+\left(t - \frac{\ell}{c}\right) + \mu_1 W_2^-\left(t - \frac{\ell}{c}\right) \tag{A.3}$$

$$W_1^-(t) = -\mu_1 W_1^+\left(t - \frac{\ell}{c}\right) + (1 - \mu_1)W_2^-\left(t - \frac{\ell}{c}\right) \tag{A.4}$$

$$W_3^+(t) = 2W_2^+\left(t - \frac{\ell}{c}\right) \tag{A.5}$$

$$W_2^-(t) = -W_2^+\left(t - \frac{\ell}{c}\right) \tag{A.6}$$

where μ_1 is the reflection coefficient between the acoustic tubes, and is defined as

$$\mu_1 = \frac{A_2 - A_1}{A_2 + A_1} \tag{A.7}$$

with A_1 and A_2 the cross-sectional areas of the first and second acoustic tubes, respectively.

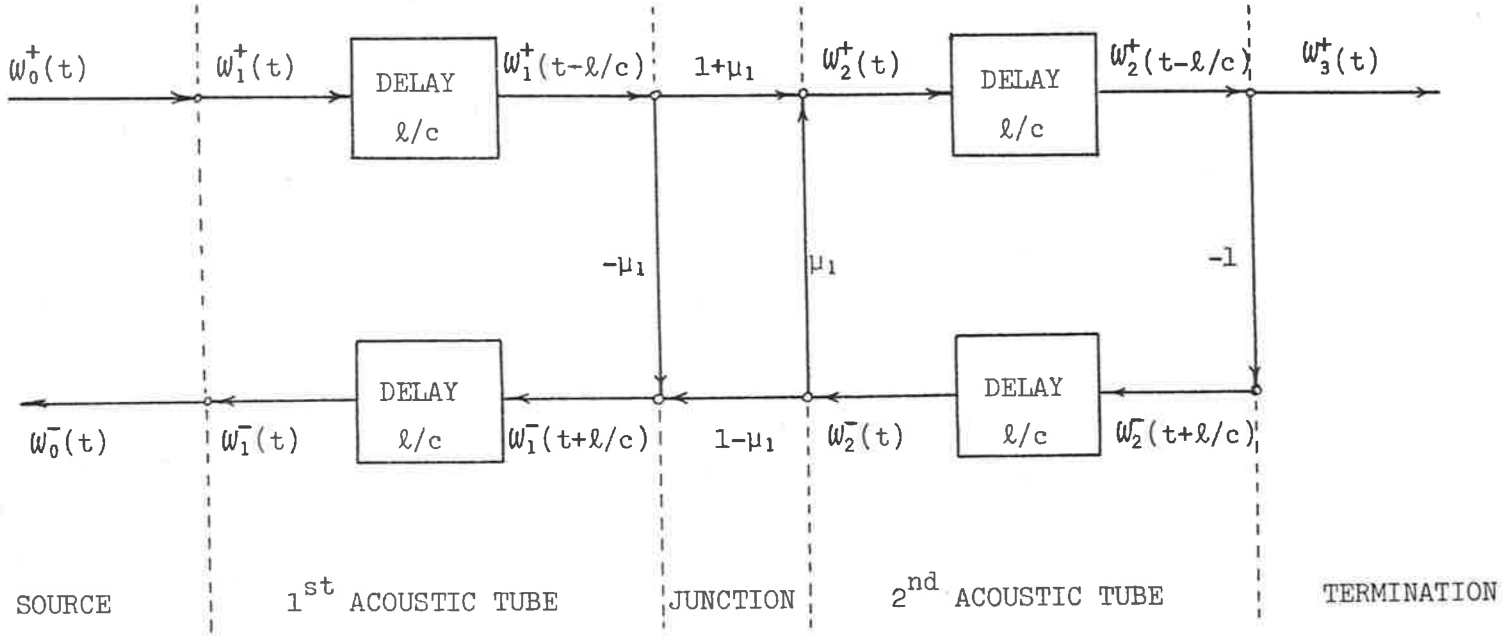


FIGURE A.1: Flow diagram for two commensurate acoustic tubes.

Substitution of Equation A.3 into Equation A.5 produces

$$W_3^+(t) = 2(1+\mu_1)W_1^+\left(t-\frac{2\ell}{c}\right) + 2\mu_1 W_2^-\left(t-\frac{2\ell}{c}\right) \qquad \text{(A.8)}$$

Similarly, a substitution of Equation A.1 into Equation A.7 produces

$$W_3^+(t) = 2(1+\mu_1)W_0^+\left(t-\frac{2\ell}{c}\right) + 2\mu_1 W_2^-\left(t-\frac{2\ell}{c}\right) \qquad \text{(A.9)}$$

Equation A.6 is rewritten as

$$W_2^-(t) = -(1+\mu_1)W_1^+\left(t-\frac{2\ell}{c}\right) - \mu_1 W_2^-\left(t-\frac{2\ell}{c}\right) \qquad \text{(A.10)}$$

on substitution of Equation A.3, and Equation A.10 becomes

$$W_2^-(t) = -(1+\mu_1)W_0^+\left(t-\frac{2\ell}{c}\right) - \mu_1 W_2^-\left(t-\frac{2\ell}{c}\right) \qquad \text{(A.11)}$$

on substitution of Equation A.1.

A substitution of Equation A.11 into Equation A.9 produces

$$W_3^+(t) = 2(1+\mu_1)\sum_{k=0}^{1}(-\mu_1)^k W_0^+\left(t-\frac{2(k+1)\ell}{c}\right) - 2\mu_1^2 W_2^-\left(t-\frac{4\ell}{c}\right) \qquad \text{(A.12)}$$

Another substitution of A.11 into A.12 this time produces

$$W_3^+(t) = 2(1+\mu_1)\sum_{k=0}^{2}(-\mu_1)^k W_0^+\left(t-\frac{2(k+1)\ell}{c}\right) - 2\mu_1^3 W_2^-\left(t-\frac{6\ell}{c}\right) \qquad \text{(A.13)}$$

Repeated substitution in this manner enables the output acoustic volume velocity, $W_3^+(t)$, to be written as

$$W_3^+(t) = 2(1+\mu_1)\sum_{k=0}^{\infty}(-\mu_1)^k W_0^+\left(t-\frac{2(k+1)\ell}{c}\right) \qquad \text{(A.14)}$$

The excitation of the two commensurate acoustic tubes is the volume veloctiy waveform $W_0^+(t)$. Hence, the impulse response of the commensurate acoustic tubes is obtained from Equation A.14 by setting $W_0^+(t)$ equal to an impulse, i.e.

$$W_0^+(t) = \delta(t) \qquad \text{(A.15)}$$

where

$$\delta(t) = \begin{cases} 1 & t = 0 \\ 0 & t \neq 0 \end{cases} \qquad \text{(A.16)}$$

Substituting Equation A.15 into Equation A.14 produces the impulse response as

$$W_3^+(t) = 2(1+\mu_1)\sum_{k=0}^{\infty}(-\mu_1)^k\delta\left[t - \frac{2(k+1)\ell}{c}\right] \qquad \text{(A.17)}$$

APPENDIX B

GENERATION OF SYNTHETIC SPEECH DATA WITH NON-WHITE EXCITATION

Synthetic speech data is used in preference to real speech when the effects or properties of the speech production system need to be controlled. Therefore, the evaluation of an analysis process to account for a certain feature of the speech production system can be performed without other features of the speech production system clouding the results. Another advantage of synthetic speech data is that all the properties of the generating system are known and, therefore, an accurate comparison of recovered and original features is possible.

Synthetic speech data is generated using the lossless acoustic tube model described in Chapter 2. The synthetic speech waveforms generated are volume velocities, so the basic equations used in the generation process are the junction equations of the acoustic tube model described in Chapter 2, i.e. Equations 2.25 and 2.26. Only sampled waveforms are generated with a sampling period equal to T where $T = \ell/c$ for ℓ being the length of each acoustic tube (i.e. commensurate acoustic tubes are assumed), and c is the velocity of sound.

A non-white excitation of the acoustic tubes is assumed and, in general, the non-white excitation is defined by the glottal pulse models of ROSENBERG [1971] and FANT [1979], or derived from published glottal pulse waveforms measured during phonation. Hence, the non-white excitation is defined by a set of parameters for an acoustic tube model, or as a waveform derived from published waveforms.

The non-white excitation of the acoustic tube model is the only violation of the assumptions of the linear predictive/acoustic tube model; hence, the termination of the acoustic tubes is lossless, i.e.

$$V_M(n) = -U_M(n) \qquad \text{(B.1)}$$

The acoustic tube model contains M lossless acoustic tubes, and N discrete synthetic speech data values are generated. An original acoustic tube shape used in the generation process is either defined by their cross-sectional areas, A_i,* or indirectly by the reflection coefficients, μ_i, between adjacent acoustic tubes.

Once the original acoustic tube shape and the non-white excitation have been defined, then the procedure detailed in Figure B.1 is used to generate the synthetic speech waveform.

*The cross-sectional area of the acoustic tube at the source is always assumed to be unity, i.e. the acoustic tube cross-sectional areas are normalized to the acoustic tube at the source.

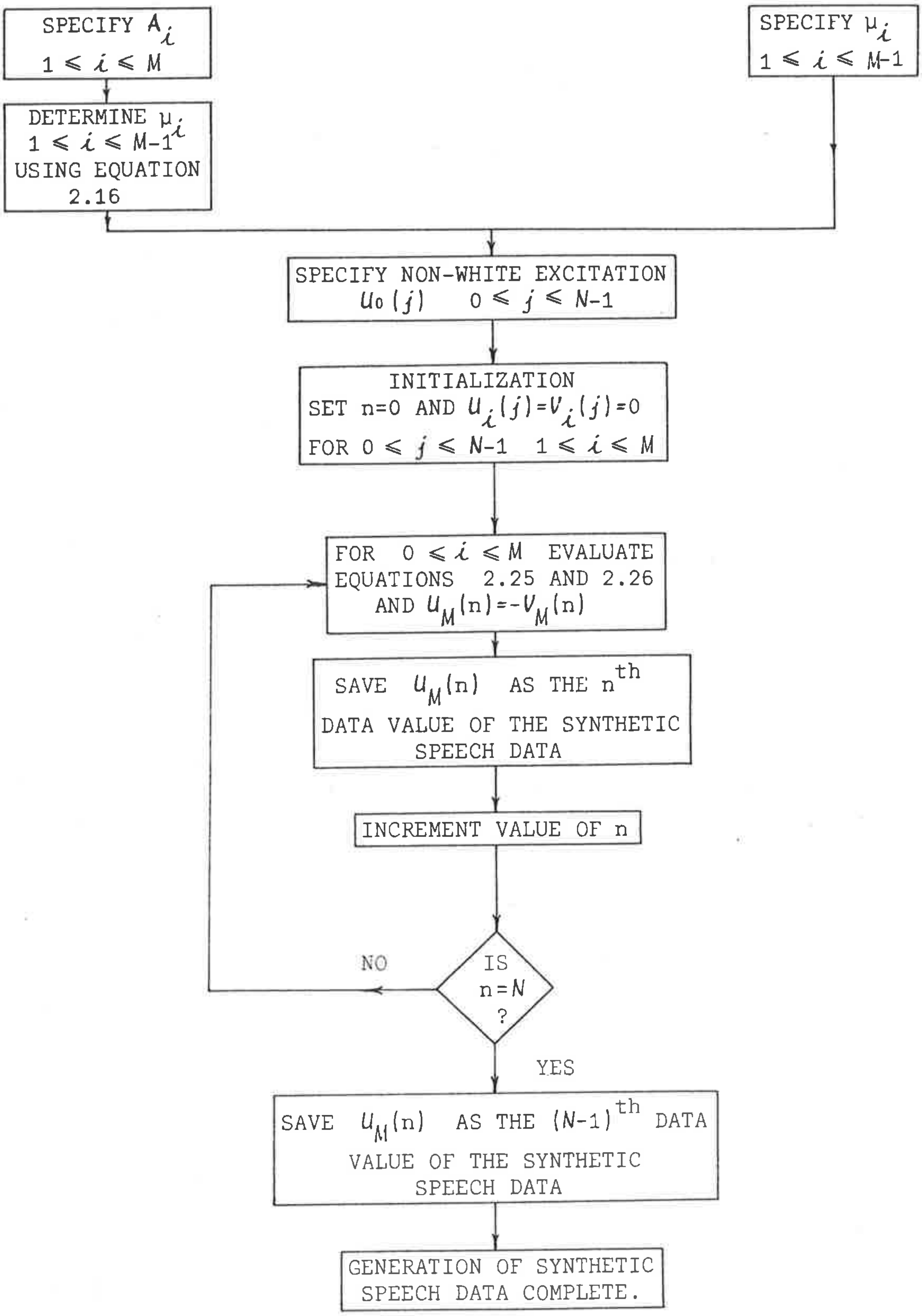

SPECIFY A_i $1 \leqslant i \leqslant M$
DETERMINE μ_i $1 \leqslant i \leqslant M-1$ USING EQUATION 2.16
SPECIFY μ_i $1 \leqslant i \leqslant M-1$
SPECIFY NON-WHITE EXCITATION $U_0(j)$ $0 \leqslant j \leqslant N-1$
INITIALIZATION SET n=0 AND $U_i(j)=V_i(j)=0$ FOR $0 \leqslant j \leqslant N-1$ $1 \leqslant i \leqslant M$
FOR $0 \leqslant i \leqslant M$ EVALUATE EQUATIONS 2.25 AND 2.26 AND $U_M(n)=-V_M(n)$
SAVE $U_M(n)$ AS THE n^{th} DATA VALUE OF THE SYNTHETIC SPEECH DATA
INCREMENT VALUE OF n
IS $n=N$?
NO
YES
SAVE $U_M(n)$ AS THE $(N-1)^{th}$ DATA VALUE OF THE SYNTHETIC SPEECH DATA
GENERATION OF SYNTHETIC SPEECH DATA COMPLETE.

FIGURE B.1: Procedure for generating synthetic speech with non-white excitation.

APPENDIX C

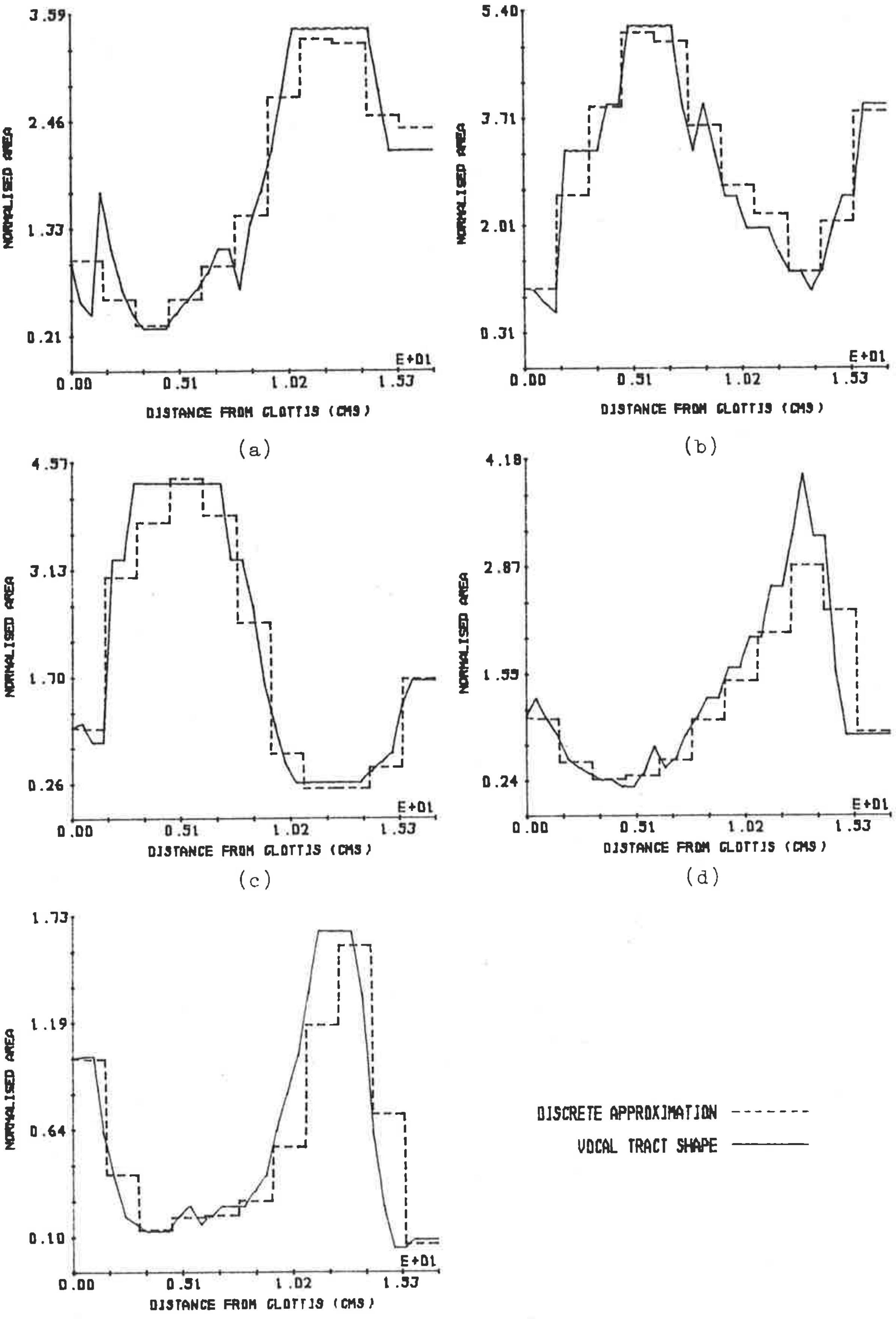
ACOUSTIC TUBE SHAPES AND REFLECTION COEFFICIENTS FOR FIVE VOWELS

One of the reasons for using synthetic speech data to evaluate speech analysis procedures is that all of the properties of the generating system are known and, hence, an accurate comparison of the recovered and original features is possible. For such evaluations and comparisons to be consistent with the results of analysing real speech, it is essential that the features used in the synthetic speech generation process be as close to the real speech situation as possible. This appendix defines a number of acoustic tube shapes (used in Appendix B to generate synthetic speech data) which are derived directly from measurements of real vocal tract shapes.

A number of researchers have measured vocal tract shapes for various speech sounds, with the early measurements being performed with X-rays. One of the first studies presenting a large number of vocal tract shapes for a wide range of speech sounds was performed by FANT [1960]. Although these vocal tract shapes were measured for Russian vowels, nasals, fricatives, etc., many researchers have used them as standards for comparing the results of their vocal tract shape recovery procedures. Because of their wide use, the discrete acoustic tube shapes presented in this appendix are derived from the Russian vowel shapes measured by FANT [1960].

It is well known that absolute cross-sectional areas of the vocal tract cannot be obtained directly from the speech waveform. Therefore, when an acoustic tube shape is defined to approximate a vocal tract shape, then only relative changes in acoustic tube cross-sectional area are important. In general, the acoustic tube shape is normalized with respect to the cross-sectional area of the acoustic tube at the source. A justification for choosing the acoustic tube at the source to normalize the acoustic tube shape is that the cross-sectional area of the corresponding section in the vocal tract, i.e. just above the glottis, is relatively constant.

The real vocal tract shapes used to generate discrete acoustic tube shapes are the five Russian vowel shapes |a|, |e|, |i|, |o| and |u| measured by FANT [1960]. Assuming the velocity of sound is 34 cms/sec, and a vocal tract length of 17 cms, then a sampling frequency of 10 kilohertz implies that the real vocal tract shape is modelled by 10 commensurate acoustic tubes. A comparison of the real vocal tract shapes and the discrete approximations used throughout this thesis for five Russian vowels is presented in Figure C.1.



(a)

(b)

(c)

(d)

(e)

FIGURE C.1: Fant vocal tract shapes for the vowels (a) /a/, (b) /e/, (c) /i/, (d) /o/ and (e) /u/ and the discrete approximation for a sampling frequency of 10 kilohertz.

APPENDIX D

LOOKUP TABLES
FOR TWO/THREE ADAPTIVE PRE-EMPHASIS FILTER

The two/three adaptive pre-emphasis filter developed and defined in Chapter 4 is described by two parameters, α and β. The value of α or β which produces a minimum area distance after a two/three adaptive pre-emphasis of a glottal pulse waveform is denoted by α' or β', respectively. The investigations of Chapter 4 discovered that α' and β' can be found approximately from a relationship with $R(1)/R(0)$ of the waveform being analysed. The relationship between β' and $R(1)/R(0)$ is a simple one, and β' is easily determined for a particular value of $R(1)/R(0)$.

A relationship defining α' in terms of $R(1)/R(0)$ was found to require excessive mathematical complexity and, hence, complex mathematical computations, to determine α' from a particular $R(1)/R(0)$. The inverse relationship, i.e. $R(1)/R(0)$ in terms of α', is found to have a relatively simple form, but the computational cost necessary for finding α' from this inverse relationship is excessive for real time applications. Alleviation of the necessity to solve for α' from a relationship of $R(1)/R(0)$ in terms α' is achieved by using a lookup table. This appendix defines the form of the required lookup tables for implementation of the two/three adaptive pre-emphasis filter.

Investigations performed in Chapter 4 showed that sampling frequencies between 10 and 8 kilohertz, inclusive, have little effect on the relationship between α' and $R(1)/R(0)$. However, a significant change in the relationship between α' and $R(1)/R(0)$ occurs for sampling frequencies of less than 8 kilohertz. Although it was concluded in Chapter 4 that the variations in α' for sampling frequencies of less than 8 kilohertz are small when the relationship for the 10 to 8 kilohertz case was used, a lookup table for a sampling frequency of 7 kilohertz is included here for completeness.

A major disadvantage with lookup tables is the necessity to allocate a fixed part of computer memory to store the table. If the lookup table contains a large number of elements, e.g. to provide an accurate representation of an equation, then a large storage space in computer memory is required, which may create major problems in small memory systems. Hence, a contradictory set of requirements may occur as, on one hand, an acurate representation of an equation requires large memory space for the lookup table and, on the other hand, small memory systems require a coarse representation of the equations.

The relationships between α' and $R(1)/R(0)$ determined in Chapter 4 are obtained by finding a best fit polynomial to a set of data points $(\alpha', R(1)/R(0))$ generated from glottal pulse waveforms. It was concluded in Chapter 4 that, because of the significant spread of data points around the best fit polynomial, the value of α computed from the best fit polynomial is not necessarily the one producing the required minimum area distance, i.e. α'. Therefore, there is no advantage in producing a lookup table which represents the relationship between α' and $R(1)/R(0)$ to a high degree of accuracy. Thus, the contradictory requirements discussed above do not necessarily apply here, as a relatively coarse representation of the relationship between α' and $R(1)/R(0)$ is acceptable.

Taking into account the spread of data points $\left(\alpha', R(1)/R(0)\right)$ about any of the best fit polynomials presented in Chapter 4, specifying the value of α to within 2 significant figures provides suitable accuracy. Therefore, the lookup table which enables α to be determined from a given $R(1)/R(0)$ contains approximately one hundered elements, which represents a relatively small memory allocation requirement for most computing systems.

The relationship between α' and $R(1)/R(0)$ determined in Chapter 4 for sampling frequencies between 10 and 8 kilohertz, inclusive, is Equation 4.45, i.e.

$$\frac{R(1)}{R(0)} = 3.217(\alpha') - 3.552(\alpha')^2 + 1.339(\alpha')^3 \qquad \text{(D.1)}$$

Since α' is only required to two significant figures, only the values of α' from 0.00 to 1.00 at intervals of 0.01 are present in the lookup table. For these values of α', the $R(1)/R(0)$ values are determined via Equation D.1, and the resultant data pairs are presented in Table D.1.

There are numerous procedures that can be used to determine the value of α from a lookup table, given a value of $R(1)/R(0)$. The procedure which is most efficient is highly dependent on the computing system in which the lookup table is implemented, e.g.

ALPHA'	R(1)/R(0)	ALPHA'	R(1)/R(0)	ALPHA'	R(1)/R(0)	ALPHA'	R(1)/R(0)
0.01	0.0000	0.26	0.5943	0.51	0.8838	0.76	0.9776
0.02	0.0160	0.27	0.6112	0.52	0.8904	0.77	0.9790
0.03	0.0474	0.28	0.6276	0.53	0.8968	0.78	0.9804
0.04	0.0782	0.29	0.6436	0.54	0.9029	0.79	0.9817
0.05	0.1083	0.30	0.6590	0.55	0.9086	0.80	0.9828
0.06	0.1376	0.31	0.6739	0.56	0.9141	0.81	0.9839
0.07	0.1664	0.32	0.6884	0.57	0.9194	0.82	0.9850
0.08	0.1944	0.33	0.7023	0.58	0.9243	0.83	0.9859
0.09	0.2218	0.34	0.7159	0.59	0.9290	0.84	0.9868
0.10	0.2485	0.35	0.7290	0.60	0.9335	0.85	0.9877
0.11	0.2746	0.36	0.7416	0.61	0.9377	0.86	0.9885
0.12	0.3001	0.37	0.7538	0.62	0.9417	0.87	0.9893
0.13	0.3249	0.38	0.7656	0.63	0.9455	0.88	0.9900
0.14	0.3491	0.39	0.7770	0.64	0.9490	0.89	0.9908
0.15	0.3727	0.40	0.7879	0.65	0.9524	0.90	0.9915
0.16	0.3957	0.41	0.7985	0.66	0.9555	0.91	0.9922
0.17	0.4181	0.42	0.8086	0.67	0.9584	0.92	0.9930
0.18	0.4399	0.43	0.8184	0.68	0.9612	0.93	0.9937
0.19	0.4612	0.44	0.8278	0.69	0.9638	0.94	0.9945
0.20	0.4818	0.45	0.8369	0.70	0.9662	0.95	0.9953
0.21	0.5019	0.46	0.8455	0.71	0.9684	0.96	0.9962
0.22	0.5215	0.47	0.8538	0.72	0.9705	0.97	0.9971
0.23	0.5405	0.48	0.8618	0.73	0.9725	0.98	0.9981
0.24	0.5590	0.49	0.8695	0.74	0.9743	0.99	0.9991
0.25	0.5769	0.50	0.8768	0.75	0.9760	1.00	1.0000

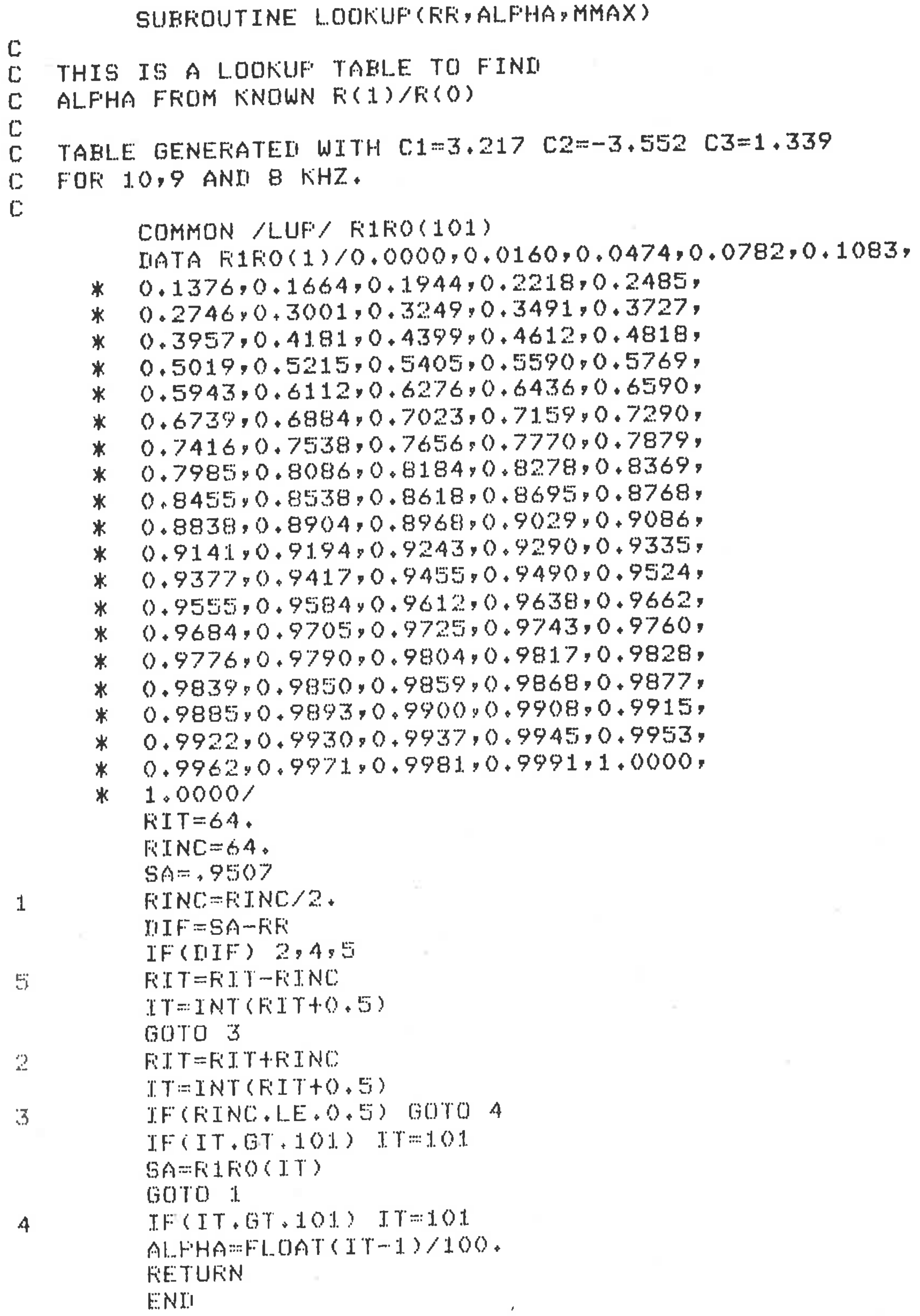
TABLE D.1: (α', R(1)/R(0)) data points for lookup table when the sampling frequency is between 10 and 8 kilohertz, inclusive.

the computational attributes, memory access times, etc. The efficiency of a lookup process is also dependent on the data, especially when the probability of certain data values occurring is much higher than for other data values. This situation appears to be the case for the determination of α' from $R(1)/R(0)$, as the investigations performed in Chapter 4 show a large proportion of the data points $(\alpha', R(1)/R(0))$ concentrated near an α' of unity. The procedure used throughout this thesis to perform the lookup process from a table is a successive approximation process, and the implementation of the successive approximation process as a FORTRAN subroutine is presented in Figure D.1.

The data presented in Table D.1 and Figure D.1 are for waveform sampling frequencies of between 10 and 8 kilohertz, inclusive. Investigations performed in Chapter 4 showed that, for a waveform sampling frequency of 7 kilohertz, a different expression between α' and $R(1)/R(0)$ than defined by Equation D.1 should be used. This expression is

$$\frac{R(1)}{R(0)} = 3.025(\alpha') - 3.148(\alpha')^2 + 1.127(\alpha')^3 \qquad \text{(D.2)}$$

The lookup table for Equation D.2 so that α' can be determined for a given value of $R(1)/R(0)$ has the same requirements as the lookup table for Equation D.1, and is implemented in the same manner. The values of α' are required to two significant figures, and so the values of α' are presented from 0.00 to 1.00 at intervals of 0.01. For these values of α', the $R(1)/R(0)$ values are determined from Equation D.2, and the data points $(\alpha', R(1)/R(0))$ are presented in Table D.2.



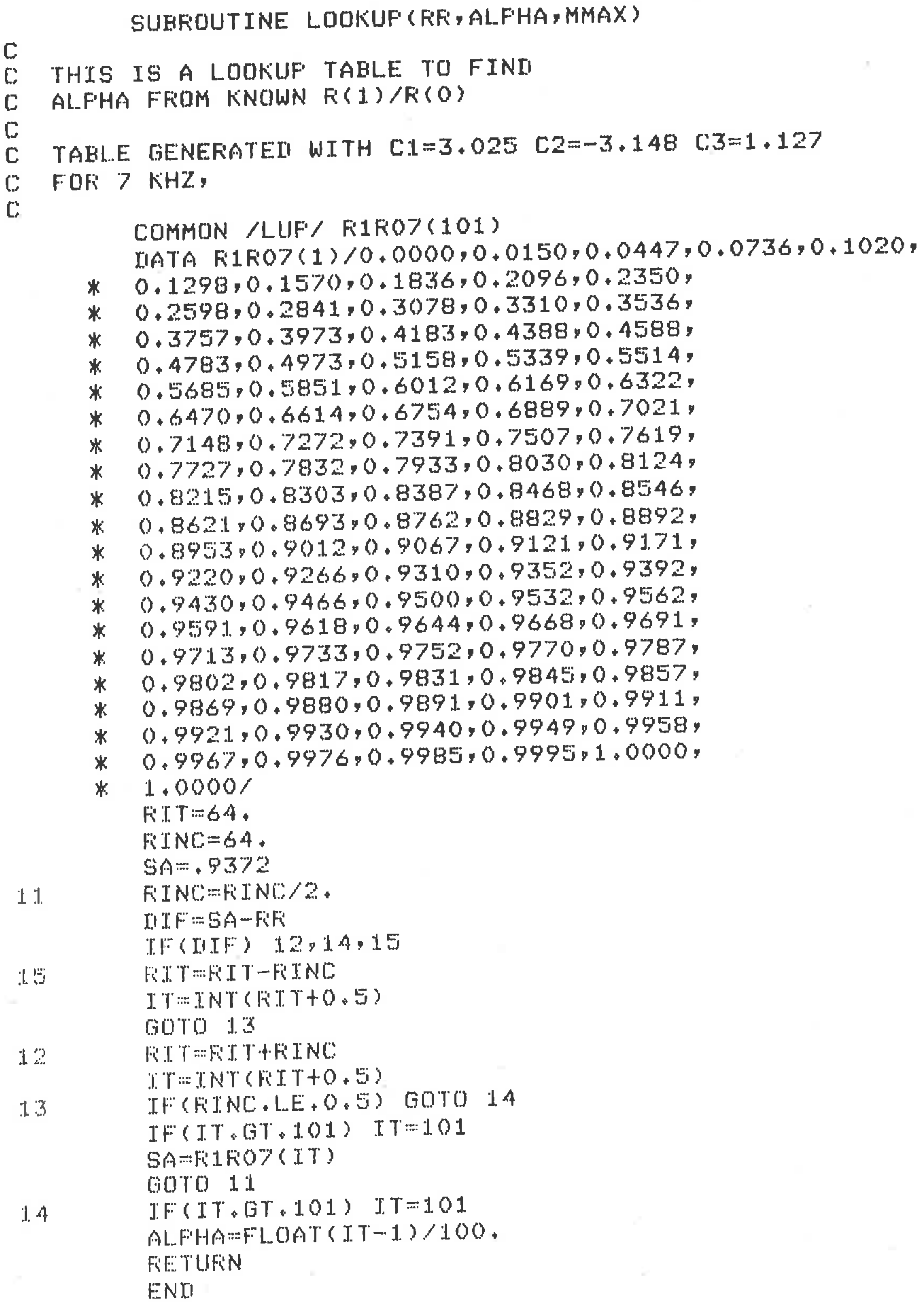
```
      SUBROUTINE LOOKUP(RR,ALPHA,MMAX)
C
C     THIS IS A LOOKUP TABLE TO FIND
C     ALPHA FROM KNOWN R(1)/R(0)
C
C     TABLE GENERATED WITH C1=3.217 C2=-3.552 C3=1.339
C     FOR 10,9 AND 8 KHZ.
C
      COMMON /LUP/ R1R0(101)
      DATA R1R0(1)/0.0000,0.0160,0.0474,0.0782,0.1083,
     * 0.1376,0.1664,0.1944,0.2218,0.2485,
     * 0.2746,0.3001,0.3249,0.3491,0.3727,
     * 0.3957,0.4181,0.4399,0.4612,0.4818,
     * 0.5019,0.5215,0.5405,0.5590,0.5769,
     * 0.5943,0.6112,0.6276,0.6436,0.6590,
     * 0.6739,0.6884,0.7023,0.7159,0.7290,
     * 0.7416,0.7538,0.7656,0.7770,0.7879,
     * 0.7985,0.8086,0.8184,0.8278,0.8369,
     * 0.8455,0.8538,0.8618,0.8695,0.8768,
     * 0.8838,0.8904,0.8968,0.9029,0.9086,
     * 0.9141,0.9194,0.9243,0.9290,0.9335,
     * 0.9377,0.9417,0.9455,0.9490,0.9524,
     * 0.9555,0.9584,0.9612,0.9638,0.9662,
     * 0.9684,0.9705,0.9725,0.9743,0.9760,
     * 0.9776,0.9790,0.9804,0.9817,0.9828,
     * 0.9839,0.9850,0.9859,0.9868,0.9877,
     * 0.9885,0.9893,0.9900,0.9908,0.9915,
     * 0.9922,0.9930,0.9937,0.9945,0.9953,
     * 0.9962,0.9971,0.9981,0.9991,1.0000,
     * 1.0000/
      RIT=64.
      RINC=64.
      SA=.9507
1     RINC=RINC/2.
      DIF=SA-RR
      IF(DIF) 2,4,5
5     RIT=RIT-RINC
      IT=INT(RIT+0.5)
      GOTO 3
2     RIT=RIT+RINC
      IT=INT(RIT+0.5)
3     IF(RINC.LE.0.5) GOTO 4
      IF(IT.GT.101) IT=101
      SA=R1R0(IT)
      GOTO 1
4     IF(IT.GT.101) IT=101
      ALPHA=FLOAT(IT-1)/100.
      RETURN
      END
```

FIGURE D.1: FORTRAN subroutine to perform a lookup for α' given R(1)/R(0) when the sampling frequency is between 10 and 8 kilohertz, inclusive.

ALPHA'	R(1)/R(0)	ALPHA'	R(1)/R(0)	ALPHA'	R(1)/R(0)	ALPHA'	R(1)/R(0)
0.01	0.0000	0.26	0.5685	0.51	0.8621	0.76	0.9713
0.02	0.0150	0.27	0.5851	0.52	0.8693	0.77	0.9733
0.03	0.0447	0.28	0.6012	0.53	0.8762	0.78	0.9752
0.04	0.0736	0.29	0.6169	0.54	0.8829	0.79	0.9770
0.05	0.1020	0.30	0.6322	0.55	0.8892	0.80	0.9787
0.06	0.1298	0.31	0.6470	0.56	0.8953	0.81	0.9802
0.07	0.1570	0.32	0.6614	0.57	0.9012	0.82	0.9817
0.08	0.1836	0.33	0.6754	0.58	0.9067	0.83	0.9831
0.09	0.2096	0.34	0.6889	0.59	0.9121	0.84	0.9845
0.10	0.2350	0.35	0.7021	0.60	0.9171	0.85	0.9857
0.11	0.2598	0.36	0.7148	0.61	0.9220	0.86	0.9869
0.12	0.2841	0.37	0.7272	0.62	0.9266	0.87	0.9880
0.13	0.3078	0.38	0.7391	0.63	0.9310	0.88	0.9891
0.14	0.3310	0.39	0.7507	0.64	0.9352	0.89	0.9901
0.15	0.3536	0.40	0.7619	0.65	0.9392	0.90	0.9911
0.16	0.3757	0.41	0.7727	0.66	0.9430	0.91	0.9921
0.17	0.3973	0.42	0.7832	0.67	0.9466	0.92	0.9930
0.18	0.4183	0.43	0.7933	0.68	0.9500	0.93	0.9940
0.19	0.4388	0.44	0.8030	0.69	0.9532	0.94	0.9949
0.20	0.4588	0.45	0.8124	0.70	0.9562	0.95	0.9958
0.21	0.4783	0.46	0.8215	0.71	0.9591	0.96	0.9967
0.22	0.4973	0.47	0.8303	0.72	0.9618	0.97	0.9976
0.23	0.5158	0.48	0.8387	0.73	0.9644	0.98	0.9985
0.24	0.5339	0.49	0.8468	0.74	0.9668	0.99	0.9995
0.25	0.5514	0.50	0.8546	0.75	0.9691	1.00	1.0000

TABLE D.2: (α', R(1)/R(0)) data points for lookup table when the sampling frequency is 7 kilohertz.

The lookup process used throughout this thesis when the sampling frequency is 7 kilohertz is the successive approximation procedure. The implementation of the successive approximation process with the data presented in Table D.2 is presented as a FORTRAN subroutine in Figure D.2.



```
          SUBROUTINE LOOKUP(RR,ALPHA,MMAX)
C
C   THIS IS A LOOKUP TABLE TO FIND
C   ALPHA FROM KNOWN R(1)/R(0)
C
C   TABLE GENERATED WITH C1=3.025 C2=-3.148 C3=1.127
C   FOR 7 KHZ,
C
          COMMON /LUP/ R1R07(101)
          DATA R1R07(1)/0.0000,0.0150,0.0447,0.0736,0.1020,
     *    0.1298,0.1570,0.1836,0.2096,0.2350,
     *    0.2598,0.2841,0.3078,0.3310,0.3536,
     *    0.3757,0.3973,0.4183,0.4388,0.4588,
     *    0.4783,0.4973,0.5158,0.5339,0.5514,
     *    0.5685,0.5851,0.6012,0.6169,0.6322,
     *    0.6470,0.6614,0.6754,0.6889,0.7021,
     *    0.7148,0.7272,0.7391,0.7507,0.7619,
     *    0.7727,0.7832,0.7933,0.8030,0.8124,
     *    0.8215,0.8303,0.8387,0.8468,0.8546,
     *    0.8621,0.8693,0.8762,0.8829,0.8892,
     *    0.8953,0.9012,0.9067,0.9121,0.9171,
     *    0.9220,0.9266,0.9310,0.9352,0.9392,
     *    0.9430,0.9466,0.9500,0.9532,0.9562,
     *    0.9591,0.9618,0.9644,0.9668,0.9691,
     *    0.9713,0.9733,0.9752,0.9770,0.9787,
     *    0.9802,0.9817,0.9831,0.9845,0.9857,
     *    0.9869,0.9880,0.9891,0.9901,0.9911,
     *    0.9921,0.9930,0.9940,0.9949,0.9958,
     *    0.9967,0.9976,0.9985,0.9995,1.0000,
     *    1.0000/
          RIT=64.
          RINC=64.
          SA=.9372
 11       RINC=RINC/2.
          DIF=SA-RR
          IF(DIF) 12,14,15
 15       RIT=RIT-RINC
          IT=INT(RIT+0.5)
          GOTO 13
 12       RIT=RIT+RINC
          IT=INT(RIT+0.5)
 13       IF(RINC.LE.0.5) GOTO 14
          IF(IT.GT.101) IT=101
          SA=R1R07(IT)
          GOTO 11
 14       IF(IT.GT.101) IT=101
          ALPHA=FLOAT(IT-1)/100.
          RETURN
          END
```

FIGURE D.2: FORTRAN subroutine to perform a lookup for α' given R(1)/R(0) when the sampling frequency is 7 kilohertz.

APPENDIX E

GENERATION OF SYNTHETIC SPEECH DATA WITH A LOSSY TERMINATION

Synthetic speech data is used in preference to real speech when the effects or properties of the speech production system need to be controlled. Using synthetic speech permits the evaluation of an analysis process to account for a certain feature of the speech production system to be performed without other features of the speech production system clouding the results. Another advantage of synthetic speech data is that all the properties of the generating system are known and, therefore, an accurate comparison of recovered and original features is possible.

The synthetic speech data is generated using the lossless acoustic tube model described in Chapter 2. The synthetic speech waveforms are sampled waveforms with a sampling period of T, and $T = \ell/c$, where ℓ is the length of each acoustic tube (i.e. all the acoustic tubes are commensurate) and c is the velocity of sound in the acoustic tubes. The synthetic speech waveforms generated are volume velocities, so the basic equations used in the generation process are the junction equations of the acoustic tube model of Chapter 2, i.e. Equations 2.25 and 2.26.

A lossy termination of the acoustic tube model is assumed, and the form of the lossy termination is defined in Chapter 6. In terms of the volume velocities at the termination, i.e. $U_M(n)$ and $V_M(n)$, the lossy termination is defined as

$$V_M(n) = -\mu_M U_M(n) + \left(1-\mu_M\right)\left(V_M(n-1) - U_M(n-1)\right) \qquad \text{(E.1)}$$

where μ_M is the termination reflection coefficient defined by Equations 6.51 and 6.53 of Chapter 6. Since μ_M is a reflection coefficient, its modulus must be less than or equal to unity, and for a loss to occur at the termination, μ_M must have a modulus of less than unity, i.e. $|\mu_M| < 1.0$. The radiated volume velocity is denoted as $U_r(n)$, and defined by

$$U_r(n) = U_M(n) - V_M(n) \qquad \text{(E.2)}$$

The lossy termination is the only violation of the assumptions of the linear prediction/acoustic tube model; hence, the excitation for the acoustic tube model is an impulse. The acoustic tube model is assumed to contain M lossless acoustic tubes, and the number of discrete synthetic speech data values generated is N. The original acoustic tube shape is either defined by their cross-sectional areas, A_i,* or indirectly by the reflection coefficients, μ_i, between adjacent acoustic tubes.

Once the original acoustic tube shape and the termination reflection coefficient are defined, then the procedure detailed in Figure E.1 is used to generate the synthetic speech waveforms.

* The cross-sectional area of the acoustic tube at the source is always assumed to be unity, i.e. the acoustic tube cross-sectional areas are normalized to the acoustic tube at the source.

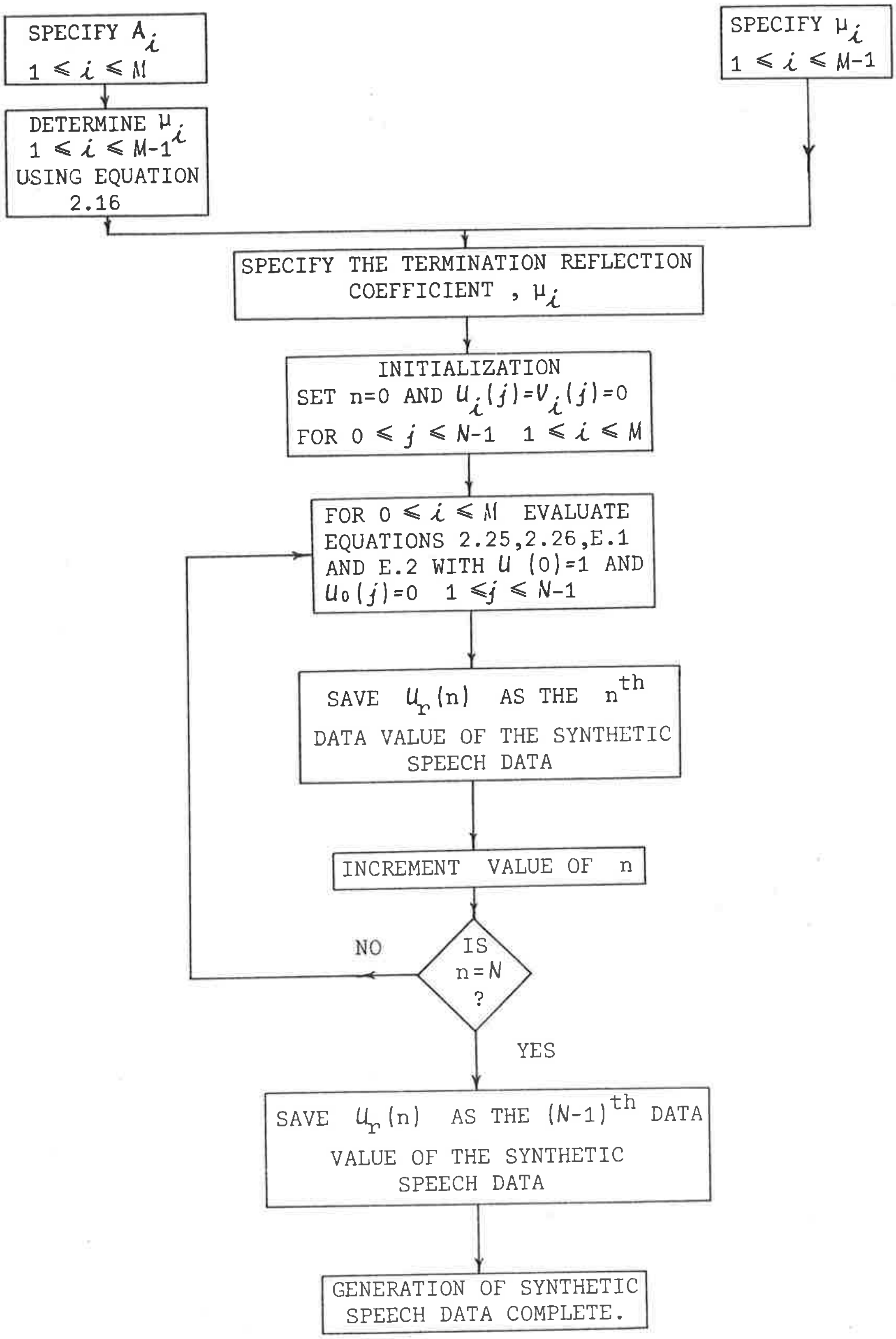

SPECIFY A_i $1 \leqslant i \leqslant M$
SPECIFY μ_i $1 \leqslant i \leqslant M-1$
DETERMINE μ_i $1 \leqslant i \leqslant M-1$ USING EQUATION 2.16
SPECIFY THE TERMINATION REFLECTION COEFFICIENT , μ_i
INITIALIZATION SET n=0 AND $U_i(j)=V_i(j)=0$ FOR $0 \leqslant j \leqslant N-1$ $1 \leqslant i \leqslant M$
FOR $0 \leqslant i \leqslant M$ EVALUATE EQUATIONS 2.25,2.26,E.1 AND E.2 WITH $U\ (0)=1$ AND $U_0(j)=0$ $1 \leqslant j \leqslant N-1$
SAVE $U_r(n)$ AS THE n^{th} DATA VALUE OF THE SYNTHETIC SPEECH DATA
INCREMENT VALUE OF n
IS $n=N$?
NO
YES
SAVE $U_r(n)$ AS THE $(N-1)^{th}$ DATA VALUE OF THE SYNTHETIC SPEECH DATA
GENERATION OF SYNTHETIC SPEECH DATA COMPLETE.

FIGURE E.1: Procedure for generating synthetic speech with lossy termination.

APPENDIX F

EQUIVALENCE BETWEEN BACKWARD AND MINIMUM FORM OF μ_i

MAKHOUL [1977] presented a large number of different expressions for calculating the reflection coefficient between the ith and (i+1)th acoustic tubes, μ_i, from the autocorrelation and cross-correlation functions in the (i+1)th acoustic tube. For a lossless termination of the acoustic tubes, all the expressions for μ_i presented by MAKHOUL [1977] are equivalent. However, for a lossy termination (as defined in Chapter 6), different values of μ_i are determined, depending on which expression for μ_i is used. This appendix shows the equivalence of the BACKWARD and MINIMUM forms for μ_i as defined by MAKHOUL [1977] for both a lossy and lossless termination of acoustic tubes.

The junction equations defining the forward and backward volume velocities at the junction of the ith and (i+1)th acoustic tubes is given by Equations 2.25 and 2.26 as

$$u_{i+1}(n) = (1+\mu_i)u_i(n) + \mu_i v_{i+1}(n-1) \tag{F.1}$$

and

$$v_i(n) = -\mu_i u_i(n) + (1-\mu_i)v_{i+1}(n-1) \tag{F.2}$$

respectively. Equations F.1 and F.2, in conjunction with the definition of the forward autocorrelation function, $A_i(r)$, (i.e. Equation 6.56) produces

$$A_i(\pm r) = \frac{A_{i+1}(r) + \mu_i^2 B_{i+1}(r) - \mu_i\big(S_{i+1}(r+1) + S_{i+1}(-r+1)\big)}{(1+\mu_i)^2} \qquad \text{(F.3)}$$

Similarly, Equations F.1 and F.2, in conjunction with the definition of the backward autocorrelation function, $B_i(r)$, (i.e. Equation 6.57) produces

$$B_i(\pm r) = \frac{B_{i+1}(r) + \mu_i^2 A_{i+1}(r) - \mu_i\big(S_{i+1}(r+1) + S_{i+1}(-r+1)\big)}{(1+\mu_i)^2} \qquad \text{(F.4)}$$

The termination of the acoustic tubes by the model described in Chapter 6 has

$$V_M(n) = -\mu_M U_M(n) \qquad \text{(F.5)}$$

Squaring Equation F.5, summing n from $-\infty$ to $+\infty$ and using the definition of $B_M(r)$ and $A_M(r)$, i.e. Equations 6.56 and 6.57, respectively, produces

$$B_M(0) = \mu_M^2 A_M(0) \qquad \text{(F.6)}$$

Equation F.6 is a special case of Equation 6.68. The termination reflection coefficient μ_M satisfies

$$0 \leqslant |\mu_M| \leqslant 1 \qquad \text{(F.7)}$$

and so

$$0 \leqslant \mu_M^2 \leqslant 1 \qquad \text{(F.8)}$$

hence, Equation F.6 can be written as

$$B_M(0) \leqslant A_M(0) \qquad \text{(F.9)}$$

Both $A_M(0)$ and $B_M(0)$ must be positive quantities, and so Equation F.9 becomes

$$0 \leqslant B_M(0) \leqslant A_M(0) \tag{F.10}$$

At this point, it is assumed that

$$B_{i+1}(0) \leqslant A_{i+1}(0) \tag{F.11}$$

and, since $A_{i+1}(0)$ and $B_{i+1}(0)$ must be positive, then

$$0 \leqslant B_{i+1}(0) \leqslant A_{i+1}(0) \tag{F.12}$$

The reflection coefficient between the ith and $(i+1)$th acoustic tubes is μ_i, and has the property

$$0 \leqslant |\mu_i| \leqslant 1 \tag{F.13}$$

therefore

$$0 \leqslant (1-\mu_i^2) \leqslant 1 \tag{F.14}$$

Multiplying Equation F.11 by Equation F.14 produces

$$(1-\mu_i^2)B_{i+1}(0) \leqslant (1-\mu_i^2)A_{i+1}(0) \tag{F.15}$$

which can rearranged to

$$B_{i+1}(0) + \mu_i^2 A_{i+1}(0) \leqslant A_{i+1}(0) + \mu_i^2 B_{i+1}(0) \tag{F.16}$$

Addition of the factor

$$-2\mu_i S_{i+1}(1) \tag{F.17}$$

and division by the factor

$$(1+\mu_i)^2 \tag{F.18}$$

which is greater than zero, to Equation F.16 produces

$$\frac{B_{i+1}(0)+\mu_i^2 A_{i+1}(0)-2\mu_i S_{i+1}(1)}{(1+\mu_i)^2} \leqslant \frac{A_{i+1}(0)+\mu_i^2 B_{i+1}(0)-2\mu_i S_{i+1}(1)}{(1+\mu_i)^2} \tag{F.19}$$

Equation F.3 with $r=0$ shows that the right-hand side of Equation F.19 is equal to $A_i(0)$, and Equation F.4 with $r=0$ shows that the left-hand side of Equation F.19 is equal to $B_i(0)$ and, hence,

$$B_i(0) \leqslant A_i(0) \tag{F.20}$$

Both $A_i(0)$ and $B_i(0)$ must be positive quantities, and so Equation F.20 becomes

$$0 \leqslant B_i(0) \leqslant A_i(0) \tag{F.21}$$

The above has proved by induction that, in general,

$$0 \leqslant B_j(0) \leqslant A_j(0) \tag{F.22}$$

for all j satisfying $0 \leqslant j \leqslant M$. The proof does not place any special conditions on the reflection coefficients other than the requirement that the acoustic tube shape be realizable, i.e. $0 \leqslant |\mu_i| \leqslant 1$ ensures that all $A_i \geqslant 0$. Hence, Equation F.22 is true for any realizable acoustic tube shape.

The MINIMUM form for μ_i as defined by MAKHOUL [1977] is

$$\mu_i = \text{SIGN} \cdot \min\left(\left|\frac{S_{i+1}(1)}{A_{i+1}(0)}\right|, \left|\frac{S_{i+1}(1)}{B_{i+1}(0)}\right|\right) \qquad \text{(F.23)}$$

where SIGN is the sign of $S_{i+1}(1)/A_{i+1}(0)$ or $S_{i+1}(1)/B_{i+1}(0)$. Equation F.22 shows that the minimum of $|S_{i+1}(1)/A_{i+1}(0)|$ and $|S_{i+1}(0)/B_{i+1}(0)|$ is $|S_{i+1}(1)/A_{i+1}(0)|$, and so Equation F.23 reduces to

$$\mu_i = \frac{S_{i+1}(1)}{A_{i+1}(0)} \qquad \text{(F.24)}$$

Equation F.24 is the BACKWARD form of μ_i as defined by MAKHOUL [1977], and so, regardless of the acoustic tube shape, the BACKWARD and MINIMUM forms of μ_i are equivalent.

APPENDIX G

NON-LINEAR SIMULTANEOUS EQUATIONS FOR TRANSFER FUNCTION ANALYSIS

An analysis method was developed in Chapter 6 which used a transfer function of the acoustic tube model to perform the analysis of acoustic waveforms. The denominator of this transfer function was shown to form the basic part of the analysis process. The denominator coefficients $d_j^{(i)}$ were shown to satisfy the following recursive set of equations:

$$d_j^{(i+1)} = d_j^{(i)} + \sum_{k=1}^{j} d_{j-k}^{(i-k)} \mu_{i-k+1}\mu_{i+1} \qquad 1 \leqslant j \leqslant i-1 \qquad \text{(G.1)}$$

$$d_i^{(i+1)} = \mu_1\mu_{i+1} \qquad \text{(G.2)}$$

$$d_0^{(i+1)} = d_0^{(i)} \qquad \text{(G.3)}$$

with $d_j^{(i)} \neq 0$ for $0 \leqslant j \leqslant i-1$ (G.4)

and, initially $d_0^{(1)} = 1$ (G.5)

For two acoustic tubes of equal length, Equations G.1 to G.5 define the denominator coefficients as

$$d_0^{(2)} = 1 \qquad \text{(G.6)}$$

$$d_1^{(2)} = \mu_1\mu_2 \qquad \text{(G.7)}$$

When three acoustic tubes are present, then Equations G.1 to G.5 define the denominator coefficients as

$$d_0^{(3)} = 1 \qquad \text{(G.8)}$$

$$d_0^{(3)} = \mu_1\mu_2 + \mu_2\mu_3 \qquad \text{(G.9)}$$

$$d_2^{(3)} = \mu_1\mu_3 \qquad \text{(G.10)}$$

For four commensurate acoustic tubes, the denominator coefficients are determined from Equations G.1 to G.5 as

$$d_0^{(4)} = 1 \qquad \text{(G.11)}$$

$$d_1^{(4)} = \mu_1\mu_2 + \mu_2\mu_3 + \mu_3\mu_4 \qquad \text{(G.12)}$$

$$d_2^{(4)} = \mu_1\mu_3 + \mu_2\mu_4 + \mu_1\mu_2\mu_3\mu_4 \qquad \text{(G.13)}$$

$$d_3^{(4)} = \mu_1\mu_4 \qquad \text{(G.14)}$$

For five commensurate acoustic tubes, the denominator coefficients are determined from Equations G.1 to G.5 as

$$d_0^{(5)} = 1 \qquad \text{(G.15)}$$

$$d_1^{(5)} = \mu_1\mu_2 + \mu_2\mu_3 + \mu_3\mu_4 + \mu_4\mu_5 \qquad \text{(G.16)}$$

$$d_2^{(5)} = \mu_1\mu_3 + \mu_2\mu_4 + \mu_3\mu_5 + \mu_1\mu_2\mu_3\mu_4 + \\ \mu_1\mu_2\mu_4\mu_5 + \mu_2\mu_3\mu_4\mu_5$$

$$d_3^{(5)} = \mu_1\mu_4 + \mu_2\mu_5 + \mu_1\mu_2\mu_3\mu_5 + \mu_1\mu_3\mu_4\mu_5 \qquad \text{(G.18)}$$

$$d_4^{(5)} = \mu_1\mu_5 \qquad \text{(G.19)}$$

As the value of i increases, i.e. the number of acoustic tubes, then the complexity of the denominator coefficients $d_j^{(i)}$ increases dramatically and, hence, the difficulty in determining the reflection coefficients from the known $d_j^{(i)}$ also increases dramatically as i increases. From the equations presented above, it is seen that, in general, the equations of μ_j in terms of $d_j^{(i)}$ are always non-linear. It is possible to define a general expression for $d_j^{(i)}$ in terms of the reflection coefficients, but this is also very complex. This general expression for $d_j^{(i)}$ is obtained in a similar manner to that used by ABDEL MONEN [1977] for a set of commensurate lossless transmission lines.

APPENDIX H

PROCEDURE AND CONDITIONS UNDER WHICH REAL SPEECH WAVEFORMS ARE DIGITALLY RECORDED

Throughout this thesis, real speech waveforms for a number of different vowel sounds are used to evaluate the performance of various pre-emphasis filters and speech analysis procedures. This appendix details the procedures and conditions under which the real speech waveforms are digitally recorded.

Seven male speakers who are native to Australia, and ranging in age from mid-twenties to mid-fifties, were used to produce the real speech waveforms. The recording of the real speech took place at a number of sessions, which were separated by at least one day, but not more than thirty days. At each session, one speaker was asked to phonate twelve vowels sounds in |h-d| frames, with the vowel sound being sustained. The |h-d| frame was used rather than CVC (consonant vowel consonant) frame since studies by STRANGE et al [1974], GOTTFRIED and STRANGE [1980], and especially MILLAR and AINSWORTH [1972], have shown that the |h-d| frame provides the least modification to the vowel sound while permitting a "speech-like" stimulus to the speaker. The initial |h-| is excited by noise and coarticulates almost completely with the following vowel, and the |-d| has a stable and confined coarticulatory pattern associated with the preceding vowel.

The twelve vowel sounds were |i|, |I|, |eI|, |e|, |æ|, |3|, |ʌ|, |u|, |U|, |o|, |ɔ|, |a|, but only the five vowel sounds |a|, |e|, |i|, |o|, |u| are used throughout this thesis. The order in which the speaker phonated the twelve vowels in the |h-d| frame was randomly organized, so that the order was different for each speaker and each time the speaker repeated the procedure on a different day.

All speech recordings were made in a quiet, but not noise-free, environment. The speech pressure waveform was measured with a BRÜEL and KJAER Precision Sound Level Meter Type 2203, with a BRÜEL and KJAER condenser microphone Type 4165 providing a frequency response from 10 hertz to greater than 15 kilohertz. The output analog signal from the sound level meter was filtered by an antialiasing filter, which is a seven section elliptic low pass filter with a cut-off frequency of 4.5 kilohertz, and is over 20 dB down at 5 kilohertz. The anialiasing filter permits a sampling frequency of greater than or equal to 10 kilohertz without significant aliasing occurring.

After the speech waveform had been filtered by the antialiasing filter, a NOVA® 2/10 minicomputer sampled the analog waveform at a frequency of 10 kilohertz, and stored the samples on magnetic tape. The actual sampling was performed by a 12 bit analog to digital converter under the control of the minicomputer.

APPENDIX I

GENERATION OF SYNTHETIC SPEECH DATA WITH A NON-WHITE EXCITATION AND A LOSSY TERMINATION

Synthetic speech data is generated using the lossless acoustic tube model described in Chapter 2. The synthetic speech waveforms are sampled waveforms where the sampling period is T, which is equal to $2\ell/c$, where ℓ is the length of each acoustic tube, i.e. commensurate acoustic tubes are used, and c is the velocity of propagation in the acoustic tubes. The basic equations used to generate the synthetic speech are the junction equations of the acoustic tube model, i.e. Equations 2.25 and 2.26, and these equations are for volume velocites, so that the synthetic speech waveforms are also volume velocities.

The non-white excitations are glottal pulse waveforms generated from the glottal pulse models of ROSENBERG [1971] and FANT [1979] or the digitized glottal pulse waveforms measured by varinous researchers. In the case of the glottal pulse models, a number of parameters are specified, and the non-white glottal pulse excitation waveform is generated from equations describing the glottal pulse model. For glottal pulse waveforms derived from published waveforms, each waveform is digitized at the appropriate frequency and stored ready for input to the acoustic tube model.

The lossy termination of the acoustic tube model is the same as that defined in Chapter 6. In terms of the volume velocities at the termination of the acoustic tube model, i.e. $U_M(n)$ and $V_M(n)$, the lossy termination is defined as

$$V_M(n) = -\mu_M U_M(n) + (1 - \mu_M)\big(V_M(n-1) - U_M(n-1)\big) \qquad \text{(I.1)}$$

where μ_M is the termination reflection coefficient defined by Equations 6.51 and 6.53 of Chapter 6. For a loss to occur at the termination of the acoustic tubes, the termination reflection coefficient must have a modulus of less than unity, i.e. $|\mu_M| < 1.0$. The radiated volume velocity is denoted by $U_r(n)$, and defined as

$$U_r(n) = U_M(n) - V_M(n) \qquad \text{(I.2)}$$

A non-white excitation and a lossy termination are the only violations of the assumptions of the linear prediction/acoustic tube model. The acoustic tube model is assumed to contain M lossless acoustic tubes of equal length, and N discrete synthetic speech data values are generated. The original acoustic tubes are defined by either their cross-sectional areas, A_i,* or indirectly by the reflection coefficient, μ_i, between adjacent acoustic tubes.

Once the original acoustic tube shape, the non-white excitation waveform and the termination reflection coefficient, μ_M, are defined, then the procedure detailed in Figure I.1 is used to generate the synthetic speech waveform.

* The cross-sectional area of the acoustic tube at the source is always assumed to be unity, i.e. the acoustic tube cross-sectional areas are normalized to the acoustic tube at the source.

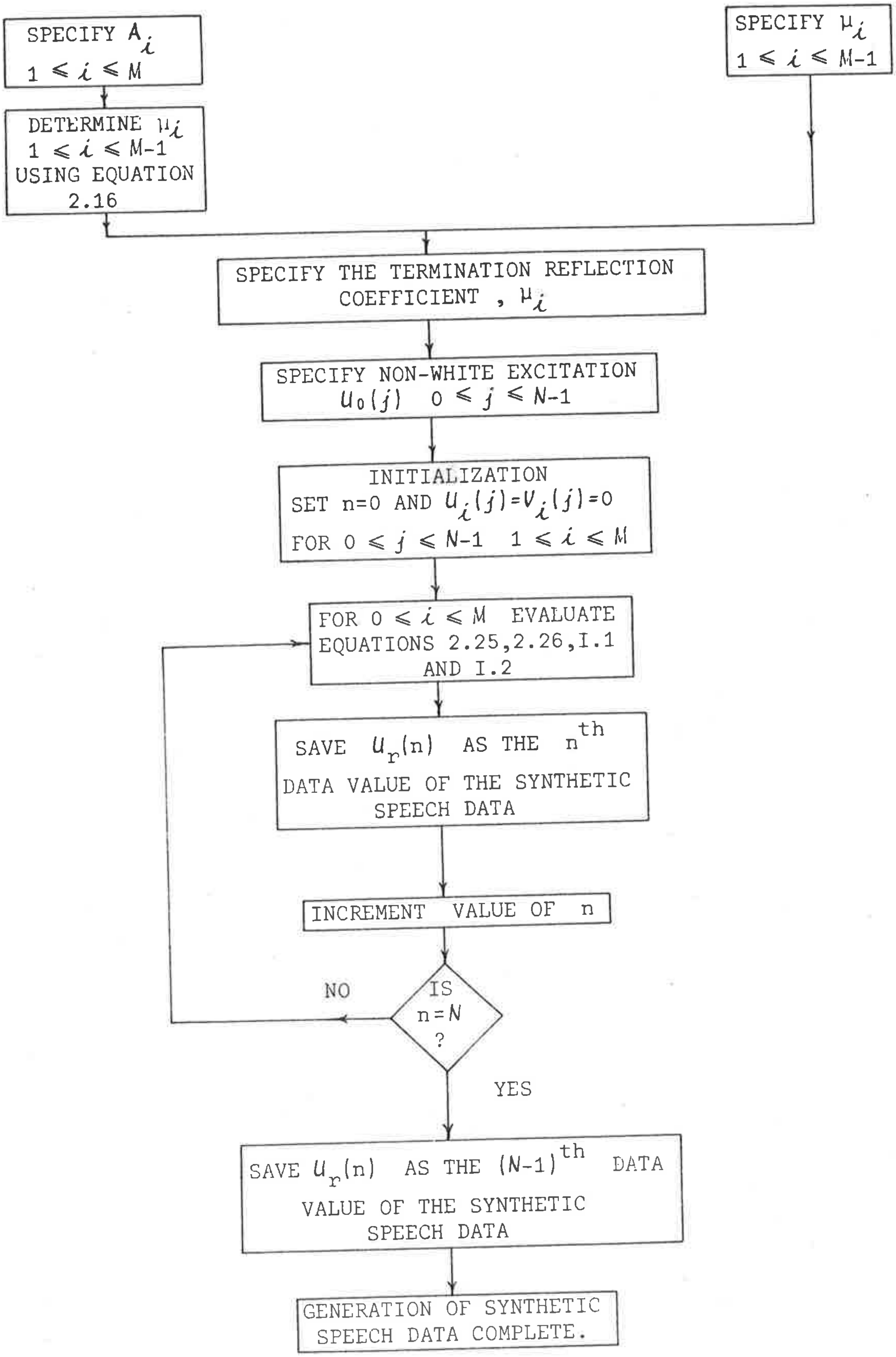

SPECIFY A_i $1 \leq i \leq M$
SPECIFY μ_i $1 \leq i \leq M-1$
DETERMINE μ_i $1 \leq i \leq M-1$ USING EQUATION 2.16
SPECIFY THE TERMINATION REFLECTION COEFFICIENT , μ_i
SPECIFY NON-WHITE EXCITATION $U_0(j)$ $0 \leq j \leq N-1$
INITIALIZATION SET n=0 AND $U_i(j)=V_i(j)=0$ FOR $0 \leq j \leq N-1$ $1 \leq i \leq M$
FOR $0 \leq i \leq M$ EVALUATE EQUATIONS 2.25,2.26,I.1 AND I.2
SAVE $U_r(n)$ AS THE n^{th} DATA VALUE OF THE SYNTHETIC SPEECH DATA
INCREMENT VALUE OF n
NO
IS $n=N$?
YES
SAVE $U_r(n)$ AS THE $(N-1)^{th}$ DATA VALUE OF THE SYNTHETIC SPEECH DATA
GENERATION OF SYNTHETIC SPEECH DATA COMPLETE.

FIGURE I.1: Procedure for generation of synthetic speech with non-white excitation and alossy termination.

REFERENCES

ABDEL-MONEN, S.S., *A Study of Transmission Line Networks with Applications to the Vocal Tract,* Ph.D. Thesis, Northeastern University, Boston, Massachusetts, August, 1977.

ANANTHAPADMANABHA, T.V., YEGANANARAYANA, B., "Epoch Extraction of Voiced Speech," *IEEE Trans. on Acoust., Speech, and Signal Proc.,* Vol. ASSP-23, pp. 562-570, 1975.

ANANTHAPADMANABHA, T.V., *Epoch Extraction and its Application to Voiced Speech Analysis,* Ph.D. Thesis, Dept. Elec. Commun. Eng., Indian Institute of Science, Bangalore, India, Sept., 1977.

ANANTHAPADMANABHA, T.V., YEGANANARAYANA, B., "Epoch Extraction from Linear Prediction Residual for Identification of Closed Glottal Intervals," *IEEE Trans. on Acoust., Speech, and Signal Proc.,* Vol. ASSP-27, No. 4, pp. 309-319, 1979.

ATAL, B.S., SCHROEDER, M.R., "Predictive Coding of Speech Signals," *Proc. 1967 Conf. Commun. and Process.,* pp. 360-361, 1967.

ATAL, B.S., SCHROEDER, M.R., "Predictive Coding of Speech Signals," *Reports of 6th Int. Cong. Acoust.,* ed. by Y. Kohasi, Tokyo, p. C-5-5, 1968a.

ATAL, B.S., SCHROEDER, M.R., "Predictive Coding of Speech Signals," *1968 Wescon Technical Papers,* Paper 8/2, 1968b.

ATAL, B.S., SCHROEDER, M.R., "Adaptive Predictive Coding of Speech Signals," *Bell System Tech. J.,* Vol. 49, pp. 1973-1976, 1970.

ATAL, B.S., "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *J. Acoust. Soc. Am.,* Vol. 47, p. 65(A), 1970a.

ATAL, B.S., "Determination of the Vocal Tract Shape Directly from the Speech Wave," *J. Acoust. Soc. Am.,* Vol. 47, p. 65(A), 1970b.

ATAL, B.S., HANAUER, S.L., "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *J. Acoust. Soc. Am.,* Vol. 50, No. 2 (Part 2), pp. 637-655, 1971.

BERANEK, L.L., *ACOUSTICS*, McGraw Hill Book Co., New York, 1954.

BEROUTI, M.G., CHILDERS, D.G., PAIGE, A., "Glottal Area versus Glottal Volume-Velocity," *Conf. Rec. IEEE Int. Conf. on Acoust., Speech, and Signal Proc.*, pp. 30-36, 1977.

BIELBY, G.J., "Vocal Tract Area Function Recovery with Resistive Lips Termination," *10th Int. Cong. on Acoustics, Sydney, July 9-16*, Paper A1-10.2, 1980.

BINGHAM, C., GODFREY, M.D., TUKEY, J.W., "Modern Techniques of Power Spectrum Estimation," *IEEE Trans. on Audio and Electroacoustics*, Vol. AU-15, No. 2, pp. 55-66, 1967.

BLACKMAN, R.B., TUKEY, J.W., *THE MEASUREMENT OF POWER SPECTRA*, Dover Publications, New York, 1958.

BOGNER, R.E., "Correlations on Transmission Lines with Applications in Signal Processing," *IREE Int. Conv., Melbourne, Aug.*, pp. 337-340, 1977.

BOGNER, R.E., DAVIS, B.R., "Correlations of Wave Variables in a Passive Network," *Joint US-Aust. Seminar-Workshop on Systems Theory and Sonic Applications, Uni. of Newcastle, March*, 1980.

BOGNER, R.E., DAVIS, B.R., "Thermodynamics and Circuit Theory," *IREECON, Sydney, Sept. 5-9*, 1983.

BURG, J., "A New Analysis Technique for Time Series Data," *Proc. NATO Advanced Study Institute on Signal Proc., Enschede, Netherlands*, 1968.

BURG, J.P., *Maximum Entropy Spectral Analysis*, Ph.D. Thesis, Stanford Uni., Stanford, Ca., May, 1975.

CARR, P.B., TRILL, D., "Long-term Larynx-Excitation Spectra," *J. Acoust. Soc. Am.*, Vol. 36, No. 11, pp. 2033-2040, 1964.

CHIBA, T., KAJIYAMA, M., *THE VOWEL: ITS NATURE AND STRUCTURE*, Tokyo Kaiseikan Pub. Co., Tokyo, 1941. *Also* Phonetic Society of Japan, 1958.

CLASEN, R.J., "Numerical Methods for Inverting Positive Definite Matrices," *The Rand Corporation, Sanata Monica, California*, AD637-930, 1966.

CRICHTON, R.G., FALLSIDE, F., "Linear Prediction Model of Speech Production with Applications to Deaf Speech Training," *Proc. IEE*, Vol. 121, pp. 865-873, 1974.

DANIEL, C., WOOD, F.S., *FITTING EQUATIONS TO DATA-COMPUTER ANALYSIS OF MULTIFACTOR DATA*, John Wiley and Sons, New York, 2nd ed., 1980.

DESCOUT, R., TOUSIGNANT, B., LECOURS, M., "Vocal Tract Area Function Measurements: Two Time Domain Methods," *Conf. Rec. IEEE Int. Conf. on Acoust., Speech, and Signal Proc.*, pp. 75-78, 1975.

DUDA, R.O., HART, P.E., *PATTERN CLASSIFICATION AND SCENE ANALYSIS*, John Wiley and Sons, New York, 1973.

DUNN, H.K., "The Calculation of Vowel Resonances, and an Electrical Vocal Tract," *J. Acoust. Soc. Am.*, Vol. 22, pp. 740-753, 1950.

DURBIN, J., "The Fitting of Time-Series Models," *Rev. Inst. Int. Statist.*, Vol. 28, No. 3, pp. 233-243, 1960.

EBERHARD, A., "An Optimal Discrete Window for the Calculation of Power Spectra," *IEEE Trans. on Audio and Electroacoustics*, Vol. AU-21, No. 1, pp. 37-43, 1973.

ENGEBRETSON, A.M., VEMULA, N.R., "Study of the Use of Linear Predictor and Related Methods of Speech Analysis for Measuring Vocal Tract Area Functions," *J. Acoust. Soc. Am.*, Vol. 56s, p. S16, 1974.

FADDEEV, D.K., FADDEEVA, V.N., *COMPUTATIONAL METHODS OF LINEAR ALGEBRA* (English translation by R.C. Williams), W.H. Freeman, San Francisco, pp. 144-147, 1963.

FANT, G.C.M., *ACOUSTIC THEORY OF SPEECH PRODUCTION*, Mouton and Co., 's-Gravenhage, The Netherlands, 1960. *Also* 2nd ed., 1970.

FANT, G., "Glottal Source and Excitation Analysis," *Speech Trans. Lab., Royal Inst. Tech., Stockholm,* QPSR 1/79, pp. 85-107, 1979.

FANT. G., "Voice Source Dynamics," *10th Int. Congr. on Acoust., Sydney, Aust.,* Paper A1-9.5, 1980.

FLANAGAN, J.L., LANDGRAF, L., "Self Oscillating Source for Vocal Tract Synthesizers," *IEEE Trans. on Audio and Electroacoustics,* Vol. AU-16, pp. 57-64, 1968.

FLANAGAN, J.L., *SPEECH ANALYSIS SYNTHESIS AND PERCEPTION*, Springer Verlag, Berlin, 2nd ed., 1972.

FU, F.S., *DIGITAL PATTERN RECOGNITION*, Springer Verlag, 2nd ed., 1980.

FUJIMURA, O., ISHIDA, H., KIRITANI, S., "Computer Controlled Dynamic Cineradiography," *Annual Bulletin (Research Inst. of Logopedics and Phonetics) Univ. of Tokyo,* No. 2, pp. 6-10, 1968.

GEVERS, M.R., WERTZ, V.J., "A D-Step Predictor in Lattice and Ladder Form," Unpublished Paper, 1980.

GIBBS, A.J., "The Design of Digital Filters," *Australian Telecommunications Research Journal,* Vol. 4, pp. 29-34, 1970.

GOLD, B., RADER, C.M., *DIGITAL PROCESSING OF SIGNALS*, McGraw-Hill Book Co., New York, 1969.

GOLDEN, R.M., KAISER, J.F., "Design of Wideband Sampled-Data Filters," *Bell System Tech. J.,* Vol. 43, No. 4 (Part 2), pp. 1533-1545, 1964.

GOPINATH, B., SONDHI, M.M., "Determination of the Shape of the Human Vocal Tract from Acoustical Measurements," *Bell System Tech. J.,* Vol. 49, pp. 1195-1214, 1970.

GOTTFRIED, T.L., STRANGE, W., "Identification of Coarticulated Vowels," *J. Acoust. Soc. Am.*, Vol. 68, pp. 1626-1635, 1980.

GRAY, A.H. JR., MARKEL, J.D., "Digital Lattice and Ladder Filter Synthesis," *IEEE Tran. on Audio and Electroacoustics*, Vol. AU-21, No. 6, pp. 491-500, 1973.

GRAY, A.H., MARKEL, J.D., "A Spectral-Flatness Measure for Studying the Autocorrelation Method of Linear Prediction of Speech Analysis," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, Vol. ASSP-22, No. 3, pp. 207-217, 1974.

GRAY, A.H. JR., MARKEL, J.D., "A Normalized Digital Filter Structure," *IEEE Trans. Acoust., Speech, and Signal Proc.*, Vol. ASSP-23, No. 3, pp. 268-277, 1975.

GRAY, A.H., MARKEL, J.D., "Distance Measures for Speech Processing," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, Vol. ASSP-24, No. 5, pp. 380-391, 1976.

GRAY, A.H. JR., MARKEL, J.D., "Linear Prediction Analysis Programs (AUTO-COVAR)," *Subroutine COVAR, from Programs for Digital Signal Processing,* ed. by Digital Signal Processing Committee, IEEE Acoustic, Speech, and Signal Processing Society.

GRENANDER, U., SZEGO, G., *TOEPLITZ FORMS AND THEIR APPLICATIONS*, University of California Press, Berkeley, California, 1958.

HEINZ, J.M., "Perturbation Functions for the Determination of Vocal Tract Area Functions from Vocal Tract Eigenvalues," *R. Inst. Tech., Stockholm Speech Trans. Lab., Q. Prog. and Status Rep.*, pp. 1-14, 1967.

HOLMES, J.N., "Formant Excitation Before and After Glottal Closure," *Conf. Rec., IEEE Int. Conf. on Acoust., Speech, and Signal Proc.*, pp. 39-42, 1976.

ITAKURA, F., SAITO, S., "Analysis Synthesis Telephony Based upon the Maximum Likelihood Method," *Reports of 6th Int. Cong. Acoust.*, ed. by Y. Kohasi, Tokyo, pp. C17-20, 1968.

ITAKURA, F., SAITO, S. "Speech Analysis - Synthesis System Based on the Partial Autocorrelation Coefficient," *Acoust. Soc. of Japan Meeting,* 1969.

ITAKURA, F., SAITO, S., "A Statistical Method for Estimation of Speech Spectral Density and Format Frequencies," *Electron. Commun. in Japan,* Vol. 53-A, No. 1, pp. 36-43, 1970.

ITAKURA, F., SAITO, S., "Digital Filtering Techniques for Speech Analysis and Synthesis," *7th International Congress on Acoustics, Budapest,* 25 C 1, 1971.

ITAKURA, F., SAITO, S., KOIKE, Y., SAWABE, H., NISHIKAWA, M., "An Audio Response Unit Based on Partial Correlation," *IEEE Trans. on Comm.,* Vol. COM-20, pp. 792-797, 1972.

ITAKURA, F., SAITO, S., "On the Optimum Quantization of Feature Parameters in the PARCOR Speech Synthesizer," *Conf. Rec., IEEE 1972 Conf. Speech Commun. and Process., New York,* Paper L4, pp. 434-437, 1972.

ITAKURA, F., "Minimum Prediction Residual Principle Applied to Speech Recognition," *IEEE Trans. Acoust., Speech, and Signal Proc.,* Vol. ASSP-23, pp. 67-72, 1975.

KAILATH, T., "A View of Three Decades of Linear Filtering Theory," *IEEE Trans. on Inform. Theory,* Vol. IT-20, pp. 148-181, 1974.

KAISER, J.F., "Design Methods for Sampled Data Filters," *Proc. 1st Allerton Conf. Circuit System Thoery,* pp. 221-236, Nov., 1963.

KAISER, J.F., *"Digital Filters,"* Chapter 7 *in SYSTEM ANALYSIS BY DIGITAL COMPUTER*, Kuo, F.F., Kaiser, J.F., John Wiley and Sons, New York, 1966.

KELLY, J.L. JR., LOCHBAUM, C., "Speech Synthesis," *Proc. 4th International Congress on Acoustic,* Vol. G42, pp. 1-4, 1962.

KINSLER, L.E., FREY, A.R., *FUNDAMENTALS OF ACOUSTICS*, John Wiley and Sons, New York, 1950.

KITAWAKI, N., ITOH, K., ITAKURA, F., "Parcor Speech Analysis - Synthesis System," *Review of the Elec. Comm. Lab., Nippon Telegraph and Telephone Public Corporation*, Vol. 26, No. 11-12, pp. 1439-1455, 1978.

LEONARD, P.J., *Least Squares Polynomial*, University of Adelaide, Computing Centre, 1965.

LEROUX, J., GUEGUEN, C., "A Fixed Point Computation of Partial Correlation Coefficients," *IEEE Trans. Acoust., Speech, and Signal Proc.*, Vol. ASSP-25, No. 3, pp. 257-259, 1977.

LEVINSON, N., "The Wiener RMS (Root Mean Square) Error Criterion in Filter Design and Prediction," *J. Math. Phys.*, Vol. 25, No. 4, pp. 261-278, 1947. (Also see Wiener, Appendix B, 1966.)

LINDQVIST, J., "Studies of the Voice Source by Means of Inverse Filtering Technique," *Proc. 5th International Congress on Acoustics*, Paper A35, 1965.

LINDQVIST, J., "Studies of the Voice Source by Means of Inverse Filtering," *Speech Trans. Lab., Royal Inst. Tech., Stockholm*, QPSR 2/65, 1965.

MAKHOUL, J., WOLF, J., "Linear Prediction and the Spectral Analysis of Speech," *NTIS No. AD-749066, BBN Report No. 2304, Bolt Beranek and Newman Inc., Cambridge, Massachusetts*, 1972.

MAKHOUL, J., "Spectral Analysis of Speech by Linear Prediction," *IEEE Trans. on Audio and Electroacoustics*, Vol. AU-21, pp. 140-148, 1973.

MAKHOUL, J., VISWANATHAN, R., "Adaptive Pre-processing for Linear Predictive Speech Compression System," *J. Acoust. Soc. Am.*, Vol. 55, No. 2, p. 475(A), 1974.

MAKHOUL, J., "Linear Prediction: a Tutorial Review," *Proc. IEEE,* Vol. 63, pp. 561-580, 1975.

MAKHOUL, J., "Stable and Efficient Lattice Methods for Linear Prediction," *IEEE Trans. Acoustic, Speech, and Signal Proc.,* Vol. ASSP-25, No. 5, pp. 423-428, 1977.

MARKEL, J.D., "Format Trajectory Estimation from a Linear Least-Squares Inverse Filter Formulation," *Monograph No. 7,* Speech Communications Research Laboratory, Santa Barbara, California, 1971a.

MARKEL, J.D., "FFT Pruning," *IEEE Trans. on Audio and Electroacoustics,* Vol. AU-19, pp. 305-311, 1971b.

MARKEL, J.D., GRAY, A.H. JR., "On Autocorrelation Equations as Applied to Speech Analysis," *IEEE Trans. on Audio and Electroacoustics,* Vol. AU-21, pp. 69-79, 1973.

MARKEL, J.D., GRAY, A.H. JR., *LINEAR PREDICTION OF SPEECH,* Springer-Verlag, New York, 1976.

MATHEWS, M.V., MILLER, J.E., DAVID, E.E. JR., "Pitch Synchronous Analysis of Voiced Sounds," *J. Acoust. Soc. Am.,* Vol. 33, No. 2, pp. 179-186, 1961.

MERMELSTEIN, P., SCHROEDER, M.R., "Determination of Smoothed Cross-Sectional Area Functions of the Vocal Tract from Formant Frequencies," *5th Int. Cong. Acoust., Liége,* Paper A24, 1965.

MERMELSTEIN, P., "Determination of the Vocal Tract Shape from Measured Formant Frequencies," *J. Acoust. Soc. Am.,* Vol. 41, No. 5, pp. 1283-1294, 1967.

MILLAR, J.B., AINSWORTH, W.A., "Identification of Synthetic Isolated Vowels in |h-d| Context," *Acustica,* Vol. 27, pp. 278-282, 1972.

MILLER, R.L., "Nature of the Vocal Cord Wave," *J. Acoust. Soc. Am.,* Vol. 31, No.6, pp. 667-677, 1959.

MONSEN, R.B., ENGEBRETSON, A.M., "Study of Variations in the Male and Female Glottal Wave," *J. Acoust. Soc. Am.*, Vol. 62, No. 4, pp. 981-993, 1977.

MORSE, P.M., *VIBRATION AND SOUND*, McGraw-Hill Book Co., New York, 1948.

MORSE, P.M., INGARD, K.V., *THEORETICAL ACOUSTICS*, McGraw-Hill Book Co., New York, 1968.

NAKAJIMA, T., OMURA, H., TANAKA, K., ISHIZAKI, S., "Estimation of Vocal Tract Area Functions by Adaptive Inverse Filtering Methods and Identification of Articulatory Model," *Proc. 1974 Stockholm Speech Comm. Seminar*, ed. by C.G.M. Fant, John Wiley and Sons, New York, 1974.

NARASIMHA, M.J., SHENOI, K., PETERSON, A.M., "A Hilbert Space Approach to Linear Predictive Analysis of Speech Signals," *Tech. Report 3606-10*, Radioscience Lab., Stanford Electronics Lab., Standofrd Uni., California, 1974.

OPPENHEIM, A.V., SCHAFER, R.W., *DIGITAL SIGNAL PROCESSING*, Prentice-Hall, Englewood Cliffs, New Jersey, 1975.

PERKELL, J.S., "Cineradiographic Studies of Speech: Implications of Certain Articulatory Movements," *Proc. 5th Int. Cong. Acoust., Liége, Belgium, Sept.*, 1965.

PINSON, E.N., "Pitch Synchronous Time-Domain Estimation of Formant Frequencies and Bandwidths," *J. Acoust. Soc. Am.*, Vol. 35, Aug., pp. 1264-1273, 1963.

RABINER, L.R., CHENG, M.J., ROSENBERG, A.E., McGONEGAL, C.A., "A Comparative Performance Study of Several Pitch Detection Algorithms," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, Vol. ASSP-24, pp. 339-418, 1976.

RABINER, L.R., SCHAFER, R.W., *DIGITAL PROCESSING OF SPEECH SIGNALS*, Prentice-Hall, Englewood Cliffs, New Jersey, 1978.

RADER, C.M., GOLD, B., "Digital Filter Design Techniques in the Frequency Domain," *Proc. IEEE,* Vol. 55, No. 2, pp. 149-171, 1967.

RAYLEIGH, Lord, *THEORY OF SOUND,* 2 volumes, MacMillan and Co., London, 1926.

ROBINSON, E.A., *STATISTICAL COMMUNICATION AND DETECTION WITH SPECIAL REFERENCE TO DIGITAL DATA PROCESSING OF RADAR AND SEISMIC SIGNALS,* Hafner Publishing Co., New York, 1967.

ROGERS, J.A.V., *Determination of Articulatory Parameters from Speech Waveforms,* Ph.D. Thesis, Imperial College of Science and Technology, University of London, 1974.

ROSENBERG, A.E., "Effects of Glottal Pulse Shape on the Quality of Natural Sounds," *J. Acoust. Soc. Am.,* Vol. 49, pp. 583-590, 1971.

ROTHENBERG, M., "A New Inverse-Filtering Technique for Deriving the Glottal Air Flow Waveform During Voicing," *J. Acoust. Soc. Am.,* Vol. 53, No. 6, pp. 1632-1645, 1973.

SAITO, S., ITAKURA, F., "The Theoretical Consideration of Statistically Optimum Methods for Speech Spectral Density," *Report No. 3107, Electrical Communication Laboratory, N.T.T., Tokyo,* 1966. (In Japanese.)

SCHROEDER, M.R., "Determination of the Geometry of the Human Vocal Tract by Acoustic Measurements," *J. Acoust. Soc. Am.,* Vol. 41, No. 4, pp. 1002-1010, 1967.

SOBAKIN, A.N., "Digital Computer Determination of Formant Parameters of the Vocal Tract from a Speech Signal," *Soviet Physic-Acoust.* (Transl. *Amer. Inst. Phys.*), Vol. 18, 1972.

SONDHI, M.M., GOPINATH, B., "Determination of the Shape of a Lossy Vocal Tract," *Proc. 7th Int. Congr. Acoust.*, Paper 23C10, 1971a.

SONDHI, M.M., GOPINATH, B., "Determination of Vocal Tract Shapes from Impulse Response at the Lips," *J. Acoust. Soc. Am.*, Vol. 49, No. 6, pp. 1867-1873, 1971b.

SONDHI, M.M., "Model for Wave Propagation in a Lossy Vocal Tract," *J. Acoust. Soc. Am.*, Vol. 55, pp. 1070-1095, 1974.

SONDHI, M.M., "Measurement of the Glottal Waveform," *J. Acoust. Soc. Am.*, Vol. 57, No. 1, pp. 228-232, 1975.

SONDHI, M.M., "Estimation of Vocal Tract Areas: The Need for Acoustical Measurements," *Proc. Articulatory Modeling and Phonetics Symposium, Grenoble, July 10-12*, pp. 77-80, 1977.

SONDHI, M.M., "Estimation of Vocal Tract Areas: The Need for Acoustical Measurements," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, Vol. ASSP-27, No. 3, pp. 268-273, 1979.

SORENSON, H.W., "Least-Squares Estimation: From Gauss to Kalman," *IEEE Spectrum*, Vol. 7, No. 7, pp. 63-68, 1970.

STEIGLITZ, K., "The Equivalent of Digital and Analog Signal Processing," *Information and Control*, Vol. 8, pp. 455-467, 1965.

STEIGLITZ, K., DIKINSON, B., "The Use of Time-Domain Selection for Improved Linear Prediction," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, Vol. ASSP-25, pp. 34-39, 1977.

STEVENS, K.N., KASOWSKI, S., FANT, G.C.M., "An Electrical Analog of the Vocal Tract," *J. Acoust. Soc. Am.*, Vol. 25, No. 4, pp. 734-742, 1953.

STEWART, G.W., LINDSAY, R.B., *ACOUSTICS*, D. Van Nostrand Co., Princeton, N.J., 1930.

STRANGE, W., et al, "Consonant Environment Specifies Vowel Identity," *Status Report on Speech Research*, Haskins Laboratories, SR-37/38, pp. 209-216, 1974.

STRUBE, H.W., "Analog Discrete-Time Filter for Speech Synthesis," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, Vol. ASSP-25, No. 1, pp. 50-55, 1977.

SUNDBERG, J., GAUFFIN, J., "Waveform and Spectrum of the Glottal Voice Source," *Speech Trans. Lab.*, Royal Inst. Tech., Stockholm, QPSR 2-3, pp. 35-50, 1978.

WAKITA, H., "Estimation of the Vocal Tract Shape by Optimal Inverse Filtering and Acoustic/Articulatory Conversion Methods," *SCRL Monograph No. 9,* Speech Communications Research Laboratory, Santa Barbara, California, 1972.

WAKITA, H., "Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms," *IEEE Trans. on Audio and Electroacoustics,* Vol. AU-21, No. 5, pp. 417-427, 1973.

WAKITA, H., GRAY, A.H. JR., "Some Theoretical Considerations for Linear Prediction of Speech and Applications," *Proc. 1974 Stockholm Speech Comm. Seminar,* ed. by G.C.M. Fant, John Wiley and Son, N.Y., 1974.

WANG, R.J., "Optimal Window Length for the Measurement of Time Varying Power Spectra," *J. Acoust. Soc. Am.*, Vol. 52, No. 1 (Part 1), pp. 33-38, 1971.

WEBSTER, A.G., "Acoustical Impedance and the Theory of Horns," *Proc. Nat. Acad. Sci., U.S.,* Vol. 5, pp. 275-282, 1919.

WELCH, P.D., "A Power Digital Method of Power Spectrum Estimation," *IBM J. of Research and Development,* Vol. 5, pp. 141-156, 1961.

WIENER, N., *ESTRAPOLATION, INTERPOLATION AND SMOOTHING OF STATIONARY TIME SERIES*, M.I.T. Press, Cambridge, Massachusetts, 1966.

WIGGINS, R., "A Lattice Filter for Determining Reflection Coefficients from Autocorrelation Coefficients," *J. Acoust. Soc. Am., Supplement No. 1, Spring Meeting,* Vol. 63, p. S79(A), 1978.