SPECTRAL MODELLING TECHNIQUES FOR SPEECH SIGNALS BASED ON LINEAR PREDICTIVE ANALYSIS

Harprit Singh Chhatwal

A Thesis submitted for the

Degree of Doctor of Philosophy of the University of London

and the

Diploma of Membership of Imperial College

January 1988

Electrical Engineering Department

Imperial College of Science and Technology

University of London

ABSTRACT

Linear Predictive (LP) analysis is employed in a wide variety of speech processing applications for the estimation of the basic speech parameters and for the transmission or storage of speech signals at low bit rates. It is in the area of spectral modelling that LP may also be used as an effective means of formant analysis or speech spectrum smoothing for speaker/speech recognition.

The main concern of this thesis is the development of spectral modelling schemes using LP methods for both vowel sounds and nasalized speech. Due to the particular energy distributions of non-nasalized voiced utterances, linear prediction is often applied to selected portions of the signal frequency response and this approach is referred to as Selective LP. However, the choice of frequencies which divide the individual regions and the predictor order to adopt in each segment is currently implemented on an intuitive basis.

A method is, therefore, investigated for the partitioning of a speech spectrum into several segments using the approach of Dynamic Programming in order to produce the optimal fit between real and model spectra on the basis of a suitably chosen error criterion. Hence, this method constitutes an accurate means for the spectral estimation of the speech signal.

This technique, when constrained in a specific manner, leads to a novel procedure for the segmentation of a frequency response into speech formant regions. Subsequent analysis employing simple frequency domain transformations generates accurate estimates of formant position which are finally incorporated into formant synthesiser configurations and produce discernible improvements compared to speech produced using standard LP modelling.

One of the major drawbacks of LP analysis is its inability to model adequately utterances having transmission zeros in their spectra — such as the nasals in normal speech. An alternative approach has been explored in this work which evaluates zero positions from segmentations of LP residual spectra into antiresonances such that each region is represented by a Selective moving average model. The zero position estimates are then used to inverse filter the original signal in order to remove the effects of these spectral zeros. The resulting data are amenable to linear predictive analysis and these LP parameters together with the zero positions form an effective pole-zero model of the nasalized utterance. The results of this procedure are found to offer significant advantages over those available using conventional LP techniques.

CONTENTS

ABSTRACT 2

CONTENTS 3

LIST OF FIGURES 10

LIST OF TABLES 14

STATEMENT OF ORIGINALITY 15

ACKNOWLEDGEMENTS 19

1. INTRODUCTION 20

1.1 PRELIMINARY 21

1.2 AN HISTORICAL PERSPECTIVE 23

1.3 THE SPEECH SIGNAL 25

1.3.1 Mechanism of Speech Production 25

1.3.2 Characteristics of Speech Sounds 26

1.4 AIMS OF THE PROJECT 29

1.5 ORGANISATION OF THE THESIS 33

1.6 REFERENCES 36

2. DISCRETE-TIME MODELS FOR THE SPEECH SIGNAL 42

2.1 INTRODUCTION 43

2.2 WAVEFORM/PARAMETRIC REPRESENTATIONS 44

2.2.1 Time-Domain Waveform Coding 45

2.2.2 Frequency-Domain Waveform Coding 47

2.2.3 Source Coding 49

2.3 ACOUSTIC TUBE MODELLING 51

2.4 FORMANT ANALYSIS 54

2.4.1 Cascade Formant Models 54

2.4.1.1 Excitation Source 56

2.4.1.2 Radiation Load 57

2.4.2 Parallel Formant Models 58

2.5 HOMOMORPHIC SYSTEMS 59

2.5.1 Generalized Homomorphic Analysis 59

2.5.2 Homomorphic Processing of Speech 62

2.6 LINEAR PREDICTIVE ANALYSIS 65

2.6.1 Formulation of the Linear Predictive Model 65

2.6.2 Linear Predictive Coding (LPC) 68

2.7 OTHER FORMS OF SPEECH MODELS 70

2.7.1 Channel Vocoder 70

2.7.2 Articulatory Speech Models 71

2.7.3 Hidden Markov Models 71

2.8 REFERENCES 75

3. LINEAR PREDICTIVE ANALYSIS OF SPEECH 95

3.1 INTRODUCTION 96

3.2 FUNDAMENTAL PRINCIPLES OF LINEAR PREDICTIVE ANALYSIS 97

3.3 FORMULATIONS OF THE LP MODEL 99

3.3.1 The Autocorrelation Method 99

3.3.1.1 Gain Computation 100

3.3.1.2 The Levinson-Durbin Recursion 103

3.3.2 The Covariance Method 106

3.3.3 Lattice Formulations and Solutions 108

3.3.4 Comparison of LP Analysis Methods 113

3.4 SPECTRAL LINEAR PREDICTION 115

3.4.1 Frequency Domain Interpretation of LP Analysis 115

3.4.2 Comparison with Other Spectrum Analysis Methods 119

3.5 APPLICATIONS OF LINEAR PREDICTIVE ANALYSIS 122

3.5.1 Pitch Detection 122

3.5.2 Formant Analysis 123

3.5.3 Acoustic Tube Modelling 124

3.6 REFERENCES 126

4. THE OPTIMAL SEGMENTATION OF SPEECH SPECTRA
FOR SPECTRAL ESTIMATION 133

4.1 INTRODUCTION 134

4.2 SELECTIVE LINEAR PREDICTION 135

4.2.1 Development of a Selective LP Modelling Scheme for Discrete Spectra 135

4.2.2 Properties of Selective LP 140

4.3 FUNDAMENTALS OF DYNAMIC PROGRAMMING 142

4.3.1 Properties of the Dynamic Programming Formulation 142

4.3.2 Dynamic Programming Nomenclature 144

4.3.3 Applications of Dynamic Programming 147

4.4 DYNAMIC PROGRAMMING APPLIED TO SPECTRAL SEGMENTATION 149

4.4.1 Evaluation of Segment Breakpoints 151

4.4.2 Optimal Assignment of Predictor Orders 154

4.5 CHOICE OF ERROR CRITERION 157

4.5.1 Spectral Distortion Measures for Speech Processing 157

4.5.2 A Squared Error Measure 159

4.5.3 The Itakura-Saito Distortion Measure 162

4.5.4 A Weighted Error Measure 166

4.6 A SUBOPTIMAL SEGMENTATION TECHNIQUE 167

4.7 CONCLUSIONS 170

4.8 REFERENCES 173

5. SPECTRAL SEGMENTATION INTO FORMANT REGIONS 186

5.1 INTRODUCTION 187

5.2 ADDITIONAL CONSTRAINTS FOR THE
SELECTION OF FORMANT REGIONS 188

5.2.1 Introduction of a Fixed Region Predictor Order 188

5.2.2 Elimination of Real Poles in the Selective LP Model 191

5.2.3 Incorporation of a Minimum Segmental Frequency Range 193

5.2.4 Implementation of Procedure for Spectral Segmentation into Formants 195

5.3 CHOICE OF NUMBER OF REGIONS IN SEGMENTED MODEL 197

5.3.1 Evaluation of Trends in the Weighted Error Curve 198

5.3.2 Development of a Criterion for the Selection of an Overall Number of Segments 201

5.3.3 Model Order Selection Criteria for Standard LP Analysis 203

5.4 FORMANT POSITION EVALUATION FROM SEGMENTED LP MODELS 207

5.4.1 Frequency Domain Inverse Mapping for the Selective LP Configuration 208

5.5 FORMANT ESTIMATION FROM STANDARD LP ANALYSIS 212

5.5.1 Root-Solving of the Predictor Polynomial 213

5.5.2 Peak-Picking from the LP Model Frequency Response 215

5.5.3 Comparative Analysis of Formant Estimation Techniques 215

5.6 CONCLUSIONS 217

5.7 REFERENCES 220

6. FORMANT SYNTHESISER MODEL REPRESENTATIONS 238

6.1 INTRODUCTION 239

6.2 IMPLEMENTATIONS OF DIGITAL FORMANT SYNTHESISERS 241

6.2.1 Comparison of the Cascade and Parallel Formant Configurations 241

6.2.2 Cascade Synthesiser Configuration for Formant Synthesis 244

6.3 EVALUATION OF FORMANT EXTRACTION ALGORITHMS BY CASCADE MODELLING 247

6.3.1 Synthesis of Formant Estimates Derived from Spectral Segmentation into Formants 247

6.3.2 Comparative Modelling for Standard LP Techniques 252

6.4 SPECTRAL SEGMENTATION BASED ON
PITCH SYNCHRONOUS ANALYSIS 256

6.4.1 Pitch Synchronous Processing of Voiced Speech 256

6.4.2 The Electroglottographic Signal 258

6.4.3 Introduction of Constraints on the Frequency Range of Formant Regions.. 260

6.4.4 Cascade Synthesis from Pitch Synchronous Analysis 262

6.5 CONCLUSIONS 264

6.6 REFERENCES 266

7. THE REPRESENTATION OF NASALIZED SPEECH USING
COMBINED POLE-ZERO MODELS 278

7.1 INTRODUCTION 279

7.2 EXISTING METHODS FOR THE ANALYSIS OF NASALIZED SPEECH 281

7.2.1 Iterative Solutions 281

7.2.2 Non-Iterative Solutions 283

7.3 OPTIMAL SEGMENTATION OF LINEAR PREDICTION
RESIDUAL SPECTRA 286

7.3.1 Pitch Synchronous Analysis of Nasalized Speech 286

7.3.2 Development of a Second Order Moving Average Scheme
for Antiresonance Modelling 289

7.3.3 Zero Estimation from LP Residual Spectra using
Selective Moving Average Models 299

7.4 SYNTHESIS OF SPEECH MODELLED BY A
COMBINED POLE-ZERO REPRESENTATION 301

7.4.1 Linear Predictive Analysis of Inverse MA Filtered Nasalized Speech 301

7.4.2 Comparison with Standard LP Modelling 303

7.5 CONCLUSIONS 304

7.6 REFERENCES 307

8. CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH 317

8.1 DISCUSSION 318

8.2 SUGGESTIONS FOR FURTHER RESEARCH 324

8.3 REFERENCES 327

APPENDIX A. THE LEVINSON-DURBIN ALGORITHM 330

APPENDIX B. EFFECT OF THE SYSTEM TRANSFER FUNCTION ON THE SPECTRUM OF A SECOND ORDER LINEAR PREDICTIVE MODEL 335

LIST OF FIGURES

Figure 1.1 Schematic Diagram of the Human Vocal Mechanism 40

Figure 1.2 Illustration of Voiced and Unvoiced Speech Waveforms 41

Figure 2.1 Source-Filter Model of Speech Production 83

Figure 2.2 Acoustic Tube Model of the Vocal Tract 83

Figure 2.3 Continuity Conditions at the Junction between Two Lossless Acoustic Tubes 84

Figure 2.4 Signal Flow Graphs for a Lossless Tube Model of the Vocal Tract 85

Figure 2.5 General Structure of a Cascade Formant Synthesizer 86

Figure 2.6 Representations of a Vocal Tract Resonance in the s-plane and the z-plane 86

Figure 2.7 Cascade Implementation of an all-pole Transfer Function 87

Figure 2.8 General Discrete-Time Model for Speech Production 88

Figure 2.9 Parallel Formant Synthesiser Model of the Vocal Tract 89

Figure 2.10 Canonic Form Representation of Homomorphic System for Convolution 90

Figure 2.11 Representation of the Characteristic System for Homomorphic Deconvolution 90

Figure 2.12 Homomorphic Analysis of Voiced Speech 91

Figure 2.13 Block Diagram of a Linear Predictive Synthesiser 92

Figure 2.14 The Channel Vocoder 93

Figure 2.15 Structure of a Constrained Left-to-Right Markov Model 94

Figure 3.1 Block Diagram of a Lattice Implementation 130

Figure 3.2 Comparison of a Signal Spectrum for Voiced Speech and a Linear Predictive Model Spectrum 131

Figure 3.3 Illustration of Spectral Estimation based on

Homomorphic Analysis for a Voiced Speech Signal 131

Figure 3.4 Linear Prediction Error Analysis for Voiced Speech 132

Figure 4.1 Representation in z-Plane of Frequency Domain Transformations for Selective LP 177

Figure 4.2 An Example of 2 Region Selective LP Modelling 178

Figure 4.3 Illustration of Dynamic Time Warping with Continuity Conditions 178

Figure 4.4 Domain of Possible Segmentation Paths for Region Endpoints 179

Figure 4.5 Domain of Possible Segmentation Paths for the Sum of Region Predictor Orders 179

Figure 4.6 Flow Diagram Illustration of Procedure for Speech Spectral Segmentation with Optimal Breakpoint Selection and Predictor Order Allocation 180

Figure 4.7 Optimal 2 and 3 Region Speech Spectral Segmentations for a Squared Error Measure 181

Figure 4.8 Optimal 2 and 3 Region Speech Spectral Segmentations for an Itakura-Saito Error Measure 182

Figure 4.9 Optimal 2 and 3 Region Speech Spectral Segmentations for a Weighted Error Measure 183

Figure 4.10 Variation of Breakpoints, Predictor Orders and Region Errors with an Increasing Number of Segments for Suboptimal Selective LP Modelling 184

Figure 4.11 Frequency Response for 6 Region Suboptimal Selective LP Model 185

Figure 5.1 Magnitude Responses for 2 Distinct Cases of a Second Order LP System Transfer Function 224

Figure 5.2 Spectral Segmentation into Formants for 4 Vowel Phonemes 225

Figure 5.3 Weighted Error Curves for Spectral Segmentation into Formants for Several Speech Sounds 227

Figure 5.4 Illustration of Averaged Normalised Weighted Error Curve 228

Figure 5.5 Comparison of Averaged Normalised Error Curve and the Corresponding Fit Based on an Exponential Decay with a Constant Offset 228

Figure 5.6 Stopping Criterion for Number of Regions to Employ in Spectral Segmentation into Formants for Various Voiced Utterances 229

Figure 5.7 Prediction Error Curves for Various Voiced Utterances 230

Figure 5.8 Behaviour of Threshold Test Criterion with Predictor Order for Various Voiced Utterances 231

Figure 5.9 Behaviour of Akaike Information Criterion with Predictor Order for Various Voiced Utterances 232

Figure 5.10 Behaviour of Final Prediction Error with Predictor Order for Various Voiced Utterances 233

Figure 5.11 Illustration of a z-Plane to s-Plane Mapping for Formant Modelling 234

Figure 5.12 Weighted Error Curve and Stopping Criterion for Spectral Segmentation into Formants for a Synthetic Speech Signal 235

Figure 5.13 Illustration of Speech Spectral Segmentation into Formants for a Synthetic Speech Signal 236

Figure 5.14 Variation of the Threshold Test Criterion with Predictor Order for a Synthetic Speech Signal 237

Figure 5.15 Comparison of Speech Spectrum and 11th Order LP Model Response for a Synthetic Speech Signal 237

Figure 6.1 Cascade Formant Synthesiser Configuration for non-Nasalized Voiced Sounds ... 270

Figure 6.2 Comparison of Signal Spectra and Cascade Model Frequency Responses for 4 Vowel Sounds 271

Figure 6.3 Illustration of Cascade Model Spectra obtained from 12th Order and 20th Order Linear Predictive Analysis 275

Figure 6.4 Illustration of Pitch Estimation Method based on Electroglottographic Analysis 276

Figure 6.5 Narrowband Spectrograms for
(a) Original Utterance
(b) Speech Synthesised from Procedure for Spectral Segmentation
(c) Speech Synthesised from Standard LP Modelling 277

Figure 7.1 Pole-Zero Analysis of Nasalized Speech Based on Inverse LP Modelling 309

Figure 7.2 Illustration of Voiced/Silence Discrimination and
Pitch Estimation using Homomorphic Processing 310

Figure 7.3 Pitch Asynchronous Spectrum of a Nasalized Speech Sound 311

Figure 7.4 Isolated Pitch Periods and Corresponding Pitch Synchronous Spectra
for 8 Consecutive Cycles of a Nasalized Utterance 312

Figure 7.5 Linear Predictive Residual Spectra based on Pitch Synchronous Analysis
of 8 Pitch Periods from a Nasalized Utterance 313

Figure 7.6 Spectral Segmentation into Antiresonance Regions for a
Pitch Synchronous LP Residual Spectrum 313

Figure 7.7 Frequency Response of Moving Average Filter obtained from
Procedure for Spectral Segmentation into Antiresonances 314

Figure 7.8 Spectra for Inverse MA Filtered Speech Signal and
Corresponding 10th Order Linear Predictive Model 314

Figure 7.9 Narrowband Spectrograms for
(a) Nasalized Speech Sound
(b) Speech Synthesised from Combined Pole-Zero Model based on
Spectral Segmentation into Antiresonances
(c) Speech Synthesised from Standard All-Pole LP Analysis 315

Figure 7.10 Pitch Synchronous Spectra for 8 Cycles of Synthetic Speech
Derived from a Combined Pole-Zero Model of Nasalized Speech 316

Figure 7.11 Pitch Synchronous Spectra for 8 Cycles of Synthetic Speech
Derived from a Conventional LP Analysis of Nasalized Speech 316

LIST OF TABLES

Table 1.1 Phonemes in the English Language 39

Table 3.1 Computational Resources for Various Linear Prediction Methods 129

Table 4.1 Performance of Optimal 2 Region Segmentations for Various Error Criteria 176

Table 4.2 Performance of Optimal 3 Region Segmentations for Various Error Criteria 176

Table 5.1 Results of Formant Extraction from a Synthetic Speech Signal for Various Formant Estimation Techniques 223

Table 6.1 Illustration of Region Endpoints for Constrained Spectral Segmentation into Formants of Pitch Synchronous Speech Spectra 269

STATEMENT OF ORIGINALITY

The following contributions reported in this thesis are believed to be original.

(1) The development of a method for the optimal segmentation of speech spectra in relation to the selection of an optimal set of region endpoints. The model spectra in the segmentation regions are Selective Linear Predictive (LP) characterisations with given predictor orders. The segmentation scheme is based on Dynamic Programming and involves appropriate constraints on the ranges of the stage and state variables and a backtracking procedure for the calculation of segment breakpoints (Chapter 4, section 4.4.1).

(2) An extension of the optimal breakpoint selection algorithm to allow an optimal allocation of predictor orders to the Selective LP models constituting the segmentation for a given overall predictor order. The complete dynamic programming algorithm represents a general breakpoint selection/predictor order allocation strategy for the optimal segmentation of speech spectra (Chapter 4, section 4.4.2).

(3) The incorporation of various spectral error measures into the segmentation procedure and the introduction of a weighted error criterion for the segmentation of voiced speech spectra (Chapter 4, section 4.5).

(4) The formulation of a suboptimal segmentation scheme based on an optimal partitioning of the Selective LP segment associated with the greatest matching error at each stage of the segmentation process (Chapter 4, section 4.6).

(5) The introduction of additional constraints into the optimal breakpoint selection/predictor

order allocation scheme in order to allow the spectral segmentation of a speech frequency response into formant regions. These constraints significantly alter the nature of the segmentation algorithm such that (i) the solution of the LP system equations does not require a general routine such as the Levinson-Durbin algorithm as the model order is fixed at a low value, (ii) the dimensionality of the state vector reduces from 2 to 1 in the segmentation process, (iii) the nature of the allowed LP model coefficients is restricted to exclude real poles and (iv) the constraints on the variation of segment endpoints are altered to permit only the selection of formant regions (Chapter 5, section 5.2).

(6) A study of the characteristics of 2nd order Selective LP models to illustrate the effect of real and complex poles on the nature of the system transfer function. This investigation determines the form of one of the constraints in the procedure for spectral segmentation into formants (Appendix B).

(7) An evaluation of trends in the weighted error curve in order to develop a model for the behaviour of this error with an increasing number of regions in the segmentation scheme (Chapter 5, section 5.3.1).

(8) The development and implementation of a criterion based on the model of the weighted error curve for the selection of the number of regions to employ in the spectral segmentation technique. This method also automatically yields an estimate for an appropriate overall model order for the Selective LP systems constituting the model spectrum (Chapter 5, section 5.3.2).

(9) The formulation of an approach for the calculation of formant estimates from the second order Selective LP models employed in the procedure for spectral segmentation into formants. This form of analysis involves an evaluation of the pole parameters from the

Selective LP model coefficients followed by an inverse mapping of these values in the Selective LP domain to the corresponding pole radii and angles in the actual frequency domain. These transformed parameters form an estimate of the required speech formant positions (Chapter 5, section 5.4).

(10) The development of a suitable cascade formant synthesiser configuration for the analysis of those formant extraction results obtained from the spectral segmentation technique for isolated data frames and complete speech utterances (Chapter 6, sections 6.2.2 and 6.4.4).

(11) The derivation of a method for assessing the influence on cascade model spectra of simplifications in the evaluation of formant bandwidths from the 2nd order Selective LP models employed in the procedure for spectral segmentation into formants (Chapter 6, section 6.3.1).

(12) The introduction of a technique based on dynamic programming for the segmentation into formant regions of pitch synchronous spectra obtained from a complete non-nasalized voiced utterance. The implementation of this procedure also involves the incorporation of constraints on the allowable ranges of the segment endpoints in order to take account of the actual spectral nature of speech formant frequencies (Chapter 6, section 6.4.3).

(13) The derivation of a second order Selective Moving Average (MA) algorithm for the modelling of antiresonances. The development of this scheme requires a transformation of the conventional non-linear 2nd order MA equations in order to obtain a set of possible solutions. A comparative evaluation with the equivalent results for a second order antiresonator yields the appropriate solution for the characterisation of spectral zeros. This technique constitutes a general procedure for the calculation of complex conjugate zero parameters directly from a chosen segment of a signal spectrum (Chapter 7, section

7.3.2).

(14) The development and implementation of a scheme for the estimation of zero positions for nasalized speech from LP residual spectra. This procedure is characterised by a segmentation of the residual spectrum into regions containing 2nd order Selective MA models which are responsible for the representation of single antiresonances. This stage is followed by an evaluation of the zero parameters from the individual Selective MA models (Chapter 7, section 7.3.3).

(15) A simple technique for the estimation of the number of antiresonances exhibited by an LP residual spectrum based on the value of zero bandwidths obtained from the procedure for spectral segmentation into antiresonances (Chapter 7, section 7.3.3).

(16) The implementation of a combined pole-zero model for nasalized speech comprised of the zeros estimated from the spectral segmentation method and the LP coefficients associated with the signal resulting from an inverse filter constituted of the spectral antiresonances acting on the nasalized speech spectrum (Chapter 7, section 7.4.1).

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor Prof. A.G. Constantinides for his thorough guidance and helpful advice during the course of this work. I am also indebted to both Dr. Geoff Kaye and Dr. Peter Alderson of the IBM UK Scientific Centre for being prepared to provide every assistance and encouragement throughout the period of my research.

I am grateful to all members of the Signal Processing and Biomedical Systems Section for their help in the realisation of the goals of this project. In particular, I would like to thank Farzad Yazdy, Makis Livadas and Ahmadreza Hosseini-Yazdi for our numerous beneficial discussions and with whom both social and technical discourse has been an enriching and stimulating experience.

The financial support provided by the IBM UK Scientific Centre, Winchester and the Science and Engineering Research Council is also gratefully acknowledged.

Finally, I wish to express my deep appreciation to my parents for their continual encouragement, guidance and forbearance throughout the course of my studies. This research would not have been possible without the wealth of support they have provided or the many sacrifices they have endured for the purpose of my education.

CHAPTER 1

INTRODUCTION

CHAPTER 1. INTRODUCTION

1.1 PRELIMINARY

The faculty of speech constitutes man's most natural, versatile and efficient means for the expression of his thoughts and ideas to other human beings. The conversion of these cerebral messages into a spoken utterance is the result of a complex physiological procedure involving specialised movements of the human respiratory and masticatory apparatus. Although other animals employ cries and calls to communicate with each other, such sounds are believed to be far removed from the conscious, careful articulations of human speech. Indeed, man's simian ancestors probably used hand signals and facial expressions as a primitive system of communication [1.33]. Gradually, as the hands became increasingly involved with the manipulation of tools, the communicative burden may have shifted towards a reliance on facial expressions and, ultimately, led to the evolutionary development of speech [1.12].

However, the spoken word is much more than a medium of conveying messages in the form of linguistic information. It also provides sociolinguistic data regarding the speaker's geographical origin and personal information concerning the speaker's identity, sex, emotional disposition etc. Furthermore, *prosodic* features such as the overall rhythm and intonation of a speech utterance also play a vital role in projecting the meaning and intentions of the speaker [1.30].

Therefore, the study of speech represents an important and challenging step towards an understanding of all forms of human intercourse. Whilst the work of phoneticians and linguists is primarily concerned with a description of the structure and semantics of language, the task of an engineer inevitably focusses on possible applications of speech systems and it is the development of these methods which is usually referred to by the term *speech processing.* This broad description can generally be subdivided into the following areas:

• *Speech Analysis Systems*

Speech recognition [1.1] is an important application which falls into this category. The input speech is, therefore, processed and a decision based on this speech analysis is employed to recognise a particular utterance. Other examples include speaker recognition/verification systems [1.2],[1.3] which involve the identification or authentification of a particular speaker from an ensemble of possible speakers.

• *Speech Synthesis Systems*

In their most general form, these techniques implement a conversion of a written or stored message into an acoustic waveform. For instance, automatic reading machines for text-to-speech synthesis [1.4] and computer voice-response [1.5] within data retrieval systems represent two possible applications.

• *Speech Analysis/Synthesis Systems*

These methods combine the properties of the two previous classes of systems as they consist of an analysis stage followed by speech synthesis. One very important illustration of this principle is provided by the coding of speech signals for data rate compression [1.6] – where the processes of speech analysis followed by appropriate coding, transmission and resynthesis at the receiver are employed to significantly reduce the bandwidth required to convey a speech signal through a transmission channel.

In the case of speech storage and retrieval systems [1.7], the analysis procedure can be performed in non-real time to provide a compact speech representation and the results of this analysis are then stored for subsequent retrieval. To generate the required verbal response, the analysis parameters are used to drive a speech synthesiser. Other speech analysis/synthesis

systems include those aimed at time compression or expansion and the enhancement of degraded speech.

A discussion of possible algorithms and implementational schemes for producing these three types of systems is provided in Chapter 2. Although the applications of these systems are often disparate, a substantial overlap does exist between their underlying methodologies and their development has also frequently been complementary. In order to gain an insight into the evolution of these techniques, an historical review of the various speech processing methods is provided in the next section.

1.2 AN HISTORICAL PERSPECTIVE

Throughout human history, man's ability to converse has always been regarded as a divine gift and has been proffered as evidence of his superiority over the lower animals. However, it was not until the eighteenth century that the nature of speech became a subject for scientific study. In 1769, the Russian Professor Kratzenstein investigated the nature and character of vowel sounds and developed resonating cavities which reproduced these sounds [1.8].

At around the same time, a Hungarian nobleman named Wolfgang von Kempelen was also engaged in an extensive programme of speech research [1.9]. His absorbing interest for almost twenty years was the construction of a talking machine consisting of a system of levers and acoustic tubes with a vibrating reed providing the vocal excitation. This ingenious device was 'played' rather like a musical instrument and was able to produce almost the entire range of speech sounds.

Subsequently, vowel production was investigated by several researchers including both Willis & Helmholtz in 1829 and Wheatstone [1.10] in 1837. This research also prompted Alexander Graham Bell to develop a mechanical speech synthesiser which was based on the

human vocal system [1.11]. However, the work of Bell was to have a much more significant impact by opening up a new dimension to the field of speech processing.

Although speech has played a dominant role in the development of human civilisations, the only practicable means that existed a little over a century ago for sending messages over long distances was to physically transport them. The research carried out by Bell which lead to the invention of the telephone in 1876 was, therefore, nothing less than revolutionary in its impact on human communication.

Sir Richard Paget's work [1.12] early in this century on vocal resonances and vocal chord vibration served to augment existing knowledge on the mechanisms of speech production and perception. However, it was not until 1936 that recently developed electronic instrumentation entered the area of speech synthesis. In this year, Dudley [1.13] proposed the *Vocoder* (Voice Coder) which automatically transmitted and resynthesised intelligible speech. This discovery is of consequence since it spawned the modern research area of Speech Bandwidth Compression systems. Due to the rapid development of subsequent speech processing techniques, a complete survey is infeasible and only a very brief description of the more significant events is presented.

The first fully digitised coding technique known as Pulse Code Modulation (PCM) was proposed in 1938 [1.14] while Shannon's major contribution to the field of information theory was announced around 1950 [1.15]. A speech recogniser with a vocabulary of isolated digits was first reported in 1952 [1.16] and later in the same decade, the eminent researchers Fant and Flanagan were independently involved in the investigation of speech resonances and their uses in speech coding and synthesis.

With the advent of digital computers and integrated circuits in the 1960's, speech research concentrated mainly on digital techniques for a variety of reasons which are discussed in Chapter 2. The first computer simulation of a speech processing system was developed in

1964 [1.17] and a method for the deconvolution of speech signals known as homomorphic analysis [1.18] was first reported in 1968. A major breakthrough which has subsequently come to dominate many areas of speech processing was the introduction of predictive techniques by Atal and Schroeder in 1967 [1.19] under the mathematical banner of Linear Predictive Analysis.

More recently, techniques such as Transform Coding [1.6], Vector Quantisation [1.20] and Hidden Markov Modelling [1.21] have also had a significant impact and these and most of the aforementioned techniques are studied in more depth in Chapter 2. The increasing power of digital computers suggests that one objective of speech processing in the future will be towards systems with which one can hold a reasonable conversation. This will, of course require speech recognisers which are capable of approaching human abilities and synthesis mechanisms which can produce original statements rather than merely vocalising prepared messages [1.22]. However, these daunting prospects will inevitably require a much greater understanding of the structure and generation of language — perhaps to the point of a reasonably complete simulation of the human thought process.

1.3 THE SPEECH SIGNAL

1.3.1 Mechanism of Speech Production

The human vocal apparatus is an extremely complex mechanism and the anatomical structures involved are illustrated in Fig. 1.1. The vocal tract is essentially a non-uniform acoustic tube (approximately 17 cm in length in the case of an adult male) which extends from the opening between the vocal chords known as the *glottis* to the lips. Its cross-sectional area varies with time and is dependent on the placement of the lips, jaw, tongue and velum. The cross-sectional area of the vocal tract varies between 0 cm^2 and 20 cm^2. During the production of nasal sounds, the vocal cavity can be coupled to the nasal cavity by the movement of the velum.

Speech sounds are produced by air being expelled from the lungs into the trachea. Depending on the speech sound to be generated, three basic mechanisms of excitation may be involved. In the case of vowel sounds (such as /I/ in p*i*t), the air flow from the lungs causes the vocal chords to vibrate which generates a sequence of quasi-periodic pulses of air pressure which serves as the excitation source to the vocal tract.

For unvoiced fricative sounds (such as /s/ in *s*ide), turbulence is created at a variety of narrow constrictions in the vocal tract — thereby producing a source of continuous noise. A third mechanism of excitation consists of a sudden release of excess air pressure after a complete closure at some point in the vocal tract and is characteristic of plosives or stops (such as /p/ in *p*ark). Although three major classes of speech sounds have been defined, these are neither rigid, complete nor mutually exclusive. Many other types of sounds may be observed which are a perturbation or combination of these basic categories — such as the group of voiced fricatives.

1.3.2 Characteristics of Speech Sounds

For the purposes of classification, natural languages are often subdivided into a number of logical units termed *phonemes* which are defined in relation to a particular language by reference to their use in discriminating different words [1.4]. Unfortunately, phonemes are often used somewhat loosely and the corresponding phonetic classifications are sometimes inconsistent — as is to be expected when completely logical statements are required concerning a field of human activity. Nevertheless, their use is widespread in a variety of forms which range from simple phonetic transcriptions to very sophisticated classifications involving features such as the stress and intonation of an utterance. Table 1.1 presents an inexhaustive list of the phonemes in use in conversational English. This classification is described in detail by Flanagan [1.8] and some of the characteristics of the speech sounds mentioned in this table are outlined in the following discussion.

- *Vowels* are produced by the excitation of the vocal tract by quasi-periodic pulses of air pressure caused by vocal chord vibration. Although their duration is dependent on the phonetic environment, these durations are relatively long and are typically in the range 170-310 ms [1.23]. Vowels are also associated with a number of spectral peaks, or *formants*, which arise as a result of the natural resonant frequencies in the sound transmission characteristic of the particular vocal tract shapes.

  Examples: /u/[1] in b*oo*t, /ɛ/ in m*e*t.

- *Glides and Semivowels* are vowel-like sounds which differ from vowels in that they exhibit a continuous formant motion.

  Examples: /j/ in *y*ou, /w/ in *w*ant (Glides) ; /r/ in *r*un, /l/ in *l*et (Semivowels)

- *Diphthongs* result from combinations of two vowel units and are also vowel-like sounds. They also present characteristics similar to glides and semivowels.

  Examples: /aI/ in f*i*ve, /eI/ in s*ay*

- *Nasals* are produced by opening the velum and coupling the nasal cavity with the vocal tract. These sounds are voiced but are radiated mainly through the nostrils. The acoustical cavity, therefore, serves as a resonant cavity which traps acoustic energy at certain natural frequencies. Thus the radiated speech contains antiresonances, or zeros, of sound transmission at these frequencies [1.7].

  Examples: /m/ in *m*outh, /n/ in *n*ostril

---

[1] *In the literature, phonemes are usually bracketed between virgules, or slanted lines of the form /·/.*

- *Fricatives* can be either voiced or unvoiced, depending on the presence or absence of vocal fold vibration. Both, however, are typified by a source of continuous noise created by turbulence in the vocal tract.

  Examples: /v/ in se*v*en, /z/ in *z*ero (Voiced) ; /ʃ/ in *sh*e, /θ/ in *th*in (Unvoiced)

- *Plosives* or stops can also be voiced or unvoiced and are characterised by an increased pressure at a point in the vocal tract followed by an abrupt release of air. The pressure buildup is observed as a period of silence in unvoiced plosives or vocal fold vibration for voiced stops.

  Examples: /b/ in *b*oy (Voiced) ; /p/ in *p*ound (Unvoiced)

- *Affricates* are combinations of stops and fricatives and exhibit characteristics similar to both types of sounds.

  Examples: /tʃ/ in *ch*ain, /dz/ in *j*ar

Broadly speaking, speech sounds can, therefore, be divided into two main categories — voiced or unvoiced. In Fig. 1.2, 100 ms segments of speech samples are shown to illustrate these two categories. It can be seen that for the voiced speech sound the waveform is fairly periodic. The interval between the major peaks gives a measure of the *pitch period* for that particular speaker, and the inverse is termed the *pitch frequency* or fundamental frequency. Typical average pitch frequencies are 120 Hz for men and 220 Hz for women.

The vocal tract geometry, on the other hand, is controlled by the speaker in order to obtain the distinction between different sounds, which is reflected in their spectra by the resonances known as formants. Below 3 kHz, three formants are usually in evidence. From Fig. 1.2, it can be seen that each pitch period consists of approximately damped oscillations, which are actually a consequence of the spectral resonances of the tract. The oscillation period

is related to the major formant position and the rate of damping to the bandwidth of the formant. It should be noted, however, that the waveform in a single pitch period results from a superposition of decaying oscillations due to the multiplicity of formants.

The unvoiced speech waveform shown in Fig. 1.2 exhibits a noise-like characteristic. This waveform is typical of this class of sounds and shows a concentration of energy at high frequencies, which is evidenced by the high number of zero crossings. In general, it can also be said that these sounds are of lower intensity than voiced sounds.

Speech signals are highly complex and the aspects which have been considered in this section relate only to certain basic characteristics. An excellent study which deals with the subject in depth is provided by Flanagan [1.8] which, in addition, contains an extensive bibliography.

1.4 AIMS OF THE PROJECT

The fundamental objective of spectral analysis is, in essence, the expression of a given function $x(t)$ as a linear combination of a prescribed set of functions $y_i(t)$ [1.24] such that:

$$x(t) = \sum_i X_i . y_i(t) \tag{1.1}$$

The set of quantities $\{X_i\}$ then indicates the importance of each member of the set $\{y_i(t)\}$ in the representation of $x(t)$ and the signal spectrum is defined on the basis of these values $\{X_i\}$. The above transformation will be of use if it yields some information concerning the structure of $x(t)$ and, where this signal is the output of a physical, chemical, biological, engineering, economic or social system, provides some knowledge of the nature of the underlying system.

A specific class of signals $\{y_i(t)\}$ is that constituted by a sequence of sinusoidal functions which leads to the ubiquitous Fourier Series representation [1.25] and it is this form

of spectrum which is generally employed in this work. For this case, the signal power spectral density (PSD) is usually termed the spectrum and is given by the squared modulus of the Fourier Transform [1.26].

The application of spectral analysis to speech signals is particularly important for two reasons. Speech spectra are firstly believed to provide a concise description of speech sounds [1.8] and are thereby used directly in speech coding applications [1.6] and, for the same reason, their deployment is also widespread in areas such as speech recognition and speaker identification [1.7],[1.27]. Indeed, it was proposed [1.28] in the 1940's that the 'reading' of spectrographic recordings (plots indicating the variation of speech spectra with time) could represent an effective aid for the severely deafened and the congenitally deaf. The second factor is a result of the consideration that clear evidence exists that the ear performs a crude frequency analysis at an early stage in its processing [1.8] and thus features salient in spectral analysis are also of importance in human speech perception.[2]

Apart from the Fourier transform approach, there are also several other methods for the spectral estimation of speech signals and one of the foremost techniques is that of Linear Prediction (LP) analysis. The advantages of this approach are manifold and are discussed along with other properties of this form of analysis in section 2.6 and Chapter 3. One property of the LP model is that it may be utilised to smooth the speech spectrum [1.7] and the degree of smoothing is controlled by the number of coefficients in the linear predictor [1.26], i.e. the LP model order.

Due to the spectral nature of non-nasalized voiced speech sounds, it is often appropriate to employ different degrees of smoothing in various portions of the spectrum [1.27] and the method of Selective Linear Prediction provides a means for achieving this requirement since it may be used to fit a specific portion of the spectrum with an LP model. The two sets

---

[2] *A detailed discussion of hearing and perception is provided by Flanagan [1.8].*

of parameters which specify a given Selective LP model are the frequencies which divide the individual regions and the predictor orders associated with the various segments. However, these values are traditionally chosen on an intuitive basis in relation to a qualitative assessment of the signal energy distribution.

In Chapter 4, this process is formalised into a procedure for the selection of break frequencies and an allocation of predictor orders such that the match between the signal spectrum and the set of Selective LP frequency responses is optimal with respect to certain error criteria. The aim of this optimal breakpoint selection/predictor order allocation scheme is, therefore, to produce accurate spectral estimates of non-nasalized voiced utterances.

Another aspect of the spectral smoothing property of LP analysis is its ability to provide estimates of formant positions from linear predictive models. Speech formants are generally defined as either the damped sinusoidal components of the vocal tract acoustic impulse response or, equivalently, as the resonances associated with the vocal tract transfer function [1.29]. The formant parameters play a vital role in many areas of speech processing such as acoustic phonetics, automatic speech/speaker recognition, speech synthesis and low bit-rate speech transmission or storage [1.29].

As discussed in Chapter 6, one of the reasons for this extensive applicability lies in the slowly varying nature of speech formants during many vocalisations [1.30] which suggests that formants are good candidates for speech coding applications. It has also been shown that [1.31] the effect of quantisation on formant values does not have a significant influence on the speech output available from formant-based speech transmission systems. The formant parameters are also closely related to sound production and propagation in the vocal tract [1.32] and have a direct interpretation in terms of physiological and physical data. In fact, the formant positions along with the speech fundamental frequency are regarded as the two most basic acoustical parameters [1.29] and thus the formants represent important information-bearing elements of speech.

However, automatic formant extraction from speech data is a major problem [1.30] and, in spite of the fact that linear predictive methods provide one of the most effective means of formant analysis, there are nevertheless several disadvantages associated with this approach. These drawbacks are discussed in Chapter 6 and include factors such as inaccuracies in the estimation of formant bandwidths and the difficulties of selecting an appropriate LP model order. An alternative formant extraction algorithm is proposed in Chapter 5 which is directed towards the segmentation of a speech spectrum into several regions — each of which contains a single formant peak. The Selective LP models associated with these segments are then employed in the calculation of accurate estimates of formant position.

Although linear predictive analysis is widely applied to the modelling of speech signals, there is one particular class of speech sounds for which this model is not altogether adequate. This is since the LP method is only well suited to the characterisation of signals which primarily consist of spectral resonances such as the speech formants. This model is, therefore, rather simplistic for utterances such as the nasal consonants and nasalized vowels which contain antiresonances or zeros of sound transmission in their spectra. Hence, in Chapter 7, a technique is developed for the calculation of the positions of these spectral zeros in a similar manner to the procedure mentioned above for spectral segmentation into formants for non-nasalized sounds. The resulting zero positions together with the coefficients of a conventional LP model thereby represent a parameterisation of the nasalized utterance.

In this thesis, various spectral modelling approaches are, therefore, proposed for the realization of three specific purposes. The main aim of Chapter 4 is to develop a procedure for the calculation of accurate spectral estimates for non-nasalized voiced sounds. The extraction of precise speech formant estimates and their assessment using synthesiser models is the main concern of Chapters 5 and 6 while Chapter 7 describes an algorithm for the characterisation of nasalized speech sounds.

1.5 ORGANISATION OF THE THESIS

This chapter has provided an introductory discussion of some of the fundamental principles of speech processing systems followed by an historical overview of the development of the techniques utilised in these systems. The general characteristics of speech sounds have also been outlined as well as a discussion of the specific aims of this project in relation to their importance in speech processing applications.

The present work is particularly concerned with parametric models for speech signals and thus Chapter 2 is largely devoted to the development of some of the principal techniques for the study of these representations with particular emphasis on those modelling schemes which are directly employed in this thesis. These parametric representations are also discussed in relation to the other main group of speech characterisations known as waveform representations.

Chapter 3 then focusses on one particular class of speech analysis methods — those based on the principles of linear predictive (LP) analysis. These techniques form the basis of many of the procedures developed in this thesis and, therefore, both the theoretical aspects and the practical applications of the various formulations of this approach are discussed in detail. As the spectral modelling properties of LP methods are also of fundamental importance in this work, these characterisations are outlined with reference to other analysis techniques which are also traditionally employed in the spectral estimation of speech signals.

A procedure for the partitioning of a speech frequency response into several regions is proposed in Chapter 4 such that each segment is characterised by a Selective LP model of a certain predictor order and associated with a particular spectral range. The speech spectral segmentation is optimal with respect to the choice of both the region endpoints and the allocation of predictor orders and a segmentation procedure is developed for the calculation of these two sets of parameters in order to minimise the discrepancy between the signal and

Selective LP spectra. This discrepancy is calculated according to a given spectral error measure and various error criteria are considered along with their respective spectral matching properties. A suboptimal segmentation scheme is finally introduced which results in a considerably lower computational cost than the optimal method.

The introduction of specific constraints into this optimal breakpoint selection/predictor order allocation technique in Chapter 5 results in a particular form of segmentation for speech spectra. This partitioning is such that the Selective LP model in each region is concerned with the representation of a single speech formant. The rationale behind this segmentation technique is the calculation of accurate formant estimates from the individual Selective LP models and a method for achieving this requirement is also developed. These formant estimates are also compared with those available from conventional LP methods for real and artificial speech data.

The material covered in Chapter 6 is directed towards an assessment of the formant extraction algorithms discussed in Chapter 5 using formant synthesiser networks in order to produce synthetic speech signals on the basis of these formant estimates. The development of a segmentation method for the set of frequency responses obtained from a complete speech utterance is achieved through pitch synchronous processing of the voiced speech signal and some of the reasons for this form of analysis being employed are also discussed.

Chapter 7 is concerned with a slightly different aspect of spectral modelling since the characterisation of nasalized speech requires an estimation of the positions of spectral zeros. A segmentation approach is again utilised, but each region is associated with a Selective moving average (MA) model rather than a Selective LP configuration. This process obviously requires a scheme for the derivation of an MA model from a given portion of a frequency response for the purpose of antiresonance modelling and such a technique is proposed in Chapter 7.

This method is then incorporated into a procedure for the segmentation of an LP residual spectrum and the zero positions obtained from this segmentation are then used to inverse filter the original speech signal in order to remove the effect of these antiresonances. The resulting signal is then amenable to standard LP analysis and these linear predictive coefficients along with the zero estimates provide a combined pole-zero model of the nasalized utterance. Synthetic speech derived from this formulation is then compared to similar signals available from all-pole LP models.

Chapter 8 finally assesses the various spectral modelling techniques developed during the course of this thesis and indicates some of the areas of interest in which further research may be beneficial.

1.6 REFERENCES

[1.1] BRISTOW, G., "Electronic Speech Recognition", *Collins, London*, 1986.

[1.2] ATAL, B.S., "Automatic Recognition of Speakers from Their Voices", *Proc. IEEE, Vol.* 64, *No.* 4, *April* 1976, *pp* 501-531.

[1.3] O' SHAUGHNESSY, D., "Speaker Recognition", *IEEE ASSP Magazine, Vol.* 3, *No.* 4, *Oct.* 1986, *pp* 4-17.

[1.4] WITTEN, I.H., "Principles of Computer Speech", *Academic Press, London*, 1982.

[1.5] RABINER, L.R., SCHAFER, R.W., "Digital Techniques for Computer Voice Response: Implementations and Applications", *Proc. IEEE, Vol.* 64, *No.* 4, *April* 1976, *pp* 416-433.

[1.6] JAYANT, N.S., NOLL, P., "Digital Coding of Waveforms", *Prentice-Hall, Englewood Cliffs, New Jersey*, 1984.

[1.7] RABINER, L.R., SCHAFER, R.W., "Digital Processing of Speech Signals", *Prentice-Hall, Englewood Cliffs, New Jersey*, 1978.

[1.8] FLANAGAN, J.L., "Speech Analysis, Synthesis and Perception", *2nd edn., Springer-Verlag, New York*, 1972.

[1.9] VON KEMPELEN, "The Mechanism of Speech and the Construction of a Speaking Machine", *Vienna*, 1791.

[1.10] WHEATSTONE, C., "Scientific Papers of Sir Charles Wheatstone", *Physical Society of London*, 1879.

[1.11] BELL, A.G, "Prehistoric Telephone Days", *Nat. Geographic Mag., Vol.* 41, *pp* 223-242.

[1.12] PAGET, R., "Human Speech", *Kegan Paul, Trench, Trubner and Co. Ltd., London*, 1930.

[1.13] DUDLEY, H., "Synthesising Speech", *Bell Labs Record, Vol.* 15, 1936, *pp* 98-102.

[1.14] REEVES, A.H., French Patent 852,183, *Oct.* 1938.

[1.15] SHANNON, C.E., "A Mathematical Theory of Communications", *Bell System Tech. Jnl., Vol.* 27, 1948, *pp* 623-656.

[1.16] DAVIS, K.H., BALASHEK, S., "Automatic Recognition of Spoken Digits", *J. Acoust. Soc. Amer., Vol.* 24, *No.* 6, 1952, *pp* 637-642.

[1.17] GOLDEN, R.M., "Digital Computer Simulation of a Sampled-Data Voice-Excited Vocoder", *J. Acoust. Soc. Amer., Vol.* 35, 1964.

[1.18] OPPENHEIM, A.V., SCHAFER, R.W., "Homomorphic Analysis of Speech", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. AU-*16, *June* 1968, *pp* 221-226.

[1.19] ATAL, B.S., SCHROEDER, M.R., "Predictive Coding of Speech Signals", *Proc.* 1967 *Conf. Commun. and Process.*, 1967, *pp* 360-361.

[1.20] ABUT, H., GRAY, R.M., REBOLLEDO, G., "Vector Quantisation of Speech and Speech-like Waveforms", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. AU-*30, 1982, *pp* 423-435.

[1.21] RABINER, L.R., JUANG, B.H., "An Introduction to Hidden Markov Models", *IEEE ASSP Magazine, January* 1986, *pp* 4-16.

[1.22] CROCHIERE, R.E., FLANAGAN, J.L., "Current Prospectives in Speech", *IEEE Commun. Magazine, Vol.* 21, *No.* 1, *Jan.* 1983, *pp* 32-40.

[1.23] HOUSE, A.S., "On Vowel Duration in English", *J. Acoust. Soc. Amer., Vol.* 33, *No.* 9, *Sept.* 1961, *pp* 1174-1178.

[1.24] YUEN, C.K., FRASER, D., "Digital Spectral Analysis", *Pitman Publishing Ltd., London*, 1979.

[1.25] JENKINS, G.M., WATTS, D.G., "Spectral Analysis and its Applications", *Holden-Day, San Francisco*, 1968.

[1.26] KAY, S.M., MARPLE, S.L. Jr., "Spectrum Analysis — A Modern Perspective", *Proc. IEEE, Vol.* 69, *No.* 11, *Nov.* 1981, *pp* 1380-1419.

[1.27] MAKHOUL, J., "Spectral Linear Prediction : Properties and Applications", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP-*23, *No.* 3, *June* 1975, *pp* 283-296.

[1.28] POTTER, R.K., KOPP, G.A., GREEN, H.C., "Visible Speech", *D. Van Nostrand Company, Inc.*, 1947.

[1.29] MARKEL, J.D., GRAY, A.H. Jr., "Linear Prediction of Speech", *Springer-Verlag, Berlin-Heidelberg-New York*, 1976.

[1.30] FALLSIDE, F., WOODS, W.A., "Computer Speech Processing", *Prentice-Hall, Englewood Cliffs, New Jersey*, 1985.

[1.31] STEAD, L.G., WESTON, R.C., "Sampling and Quantizing the Parameters of a Formant-Tracking Vocoder System", *Proc. Speech Comm. Seminar, Stockholm, Vol.* 2, *paper G*9, 1962.

[1.32] MUNSON, W.A., MONTGOMERY, H.C., "A Speech Analyzer and Synthesizer", *J. Acoust. Soc. Amer., Vol.* 22, *No.* 5, *Sept.* 1950, *p* 678.

[1.33] LINGGARD, R., "Electronic Synthesis of Speech", *Cambridge University Press*, 1985.

VOWELS

Degree of Constriction	Tongue Hump Position		
	Front	Central	Back
High	/i/ eve	/ɜ̈/ bird	/u/ boot
	/I/ it	/ə̈/ over (unstressed)	/U/ foot
Medium	/e/ hate	/ʌ/ up	/o/ obey
	/ɛ/ met	/e/ ado (unstressed)	/ɔ/ all
Low	/æ/ at		/ɑ/ father

GLIDES & SEMIVOWELS

Palatal	/j/	you
Labial	/w/	we (no final form)
Palatal	/r/	read
Alveolar	/l/	let

DIPHTHONGS

/eI/	say	/Iu/	new
/aU/	out	/oU/	go
/ɔI/	boy	/aI/	I

FRICATIVE CONSONANTS

Place of Articulation	Voiced	Unvoiced
Labio-Dental	/v/ vote	/f/ for
Dental	/ð̈/ then	/θ/ thin
Alveolar	/z/ zoo	/s/ see
Palatal	/ʒ/ azure	/ʃ/ she
Glottal		/h/ he

STOP CONSONANTS

Place of Articulation	Voiced	Unvoiced
Labial	/b/ be	/p/ pay
Alveolar	/d/ day	/t/ to
Palatal/Velar	/g/ go	/k/ key

NASALS

Labial	/m/	me
Alveolar	/n/	no
Palatal/Velar	/ŋ/	sing (no initial form)

AFFRICATES

/dʒ/	jar
/tʃ/	chew

Table 1.1 Phonemes in the English Language

(Alternative Typewritten Phoneme Symbols used in this Thesis are based on Rabiner & Schafer [1.7])

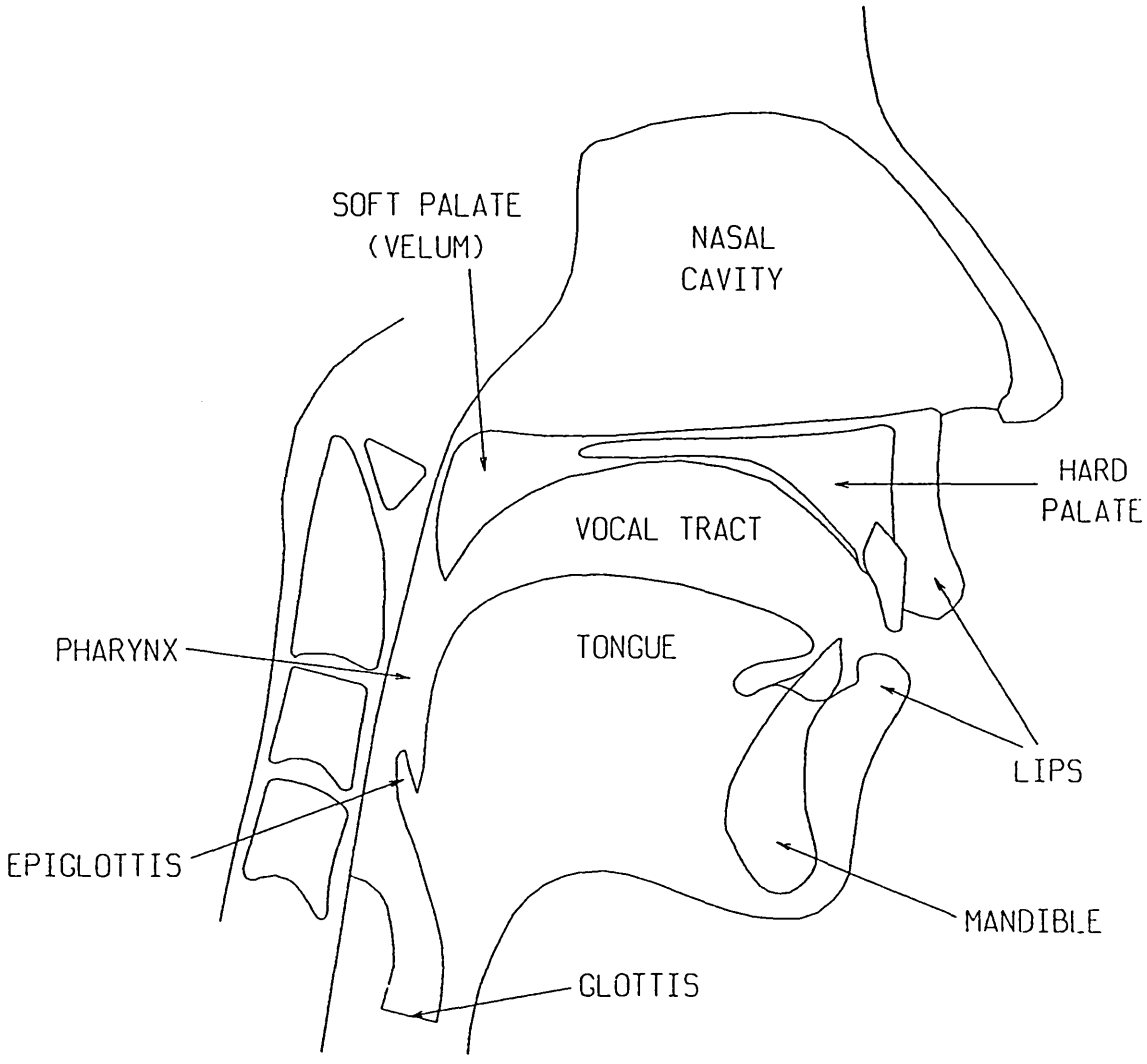


Figure 1.1 Schematic Diagram of the Human Vocal Mechanism

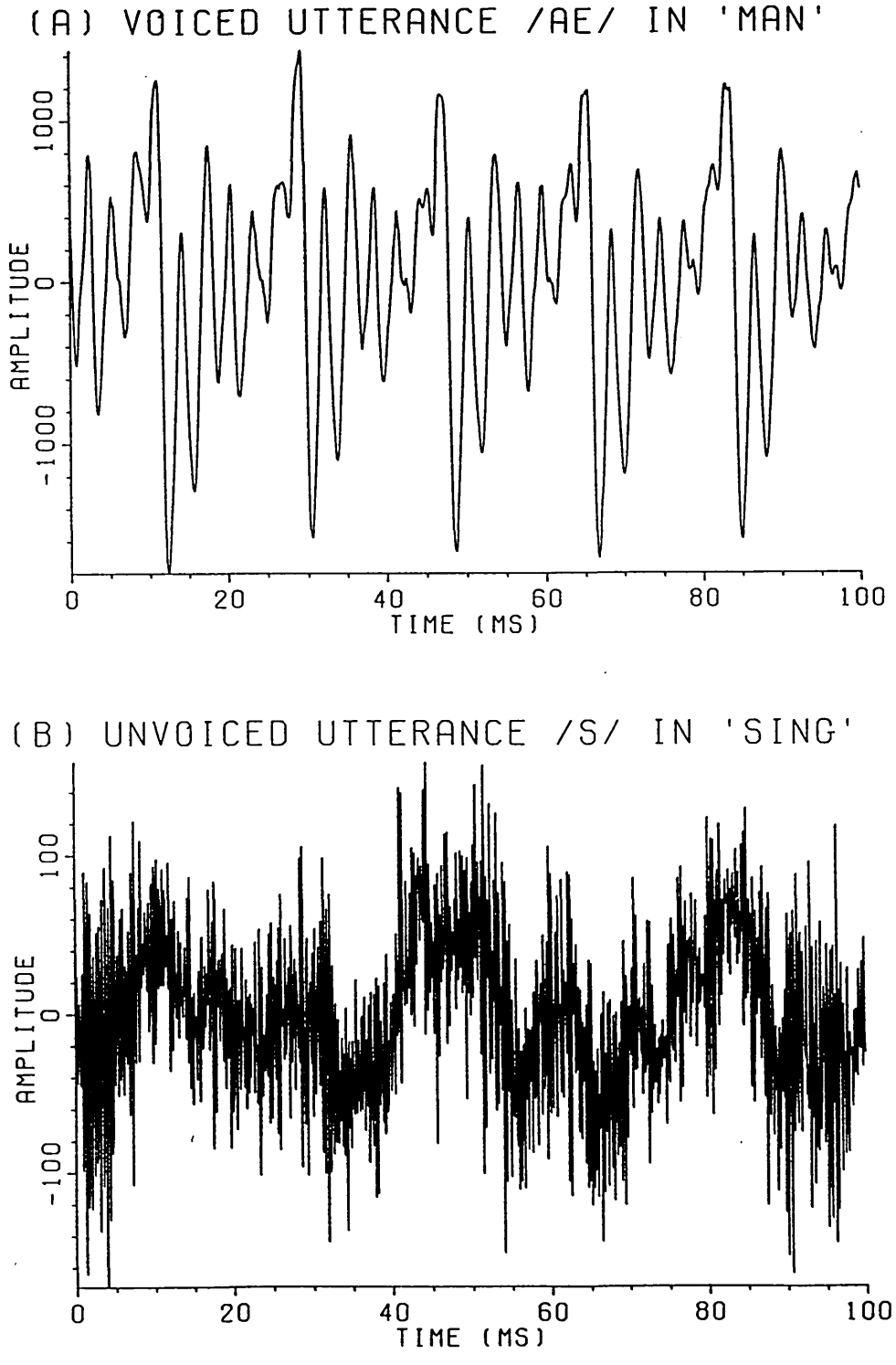


Figure 1.2 Illustration of Voiced and Unvoiced Speech Waveforms

CHAPTER 2

DISCRETE-TIME MODELS FOR THE SPEECH SIGNAL

CHAPTER 2. DISCRETE-TIME MODELS FOR THE SPEECH SIGNAL

2.1 INTRODUCTION

During the last two decades, the focus of speech analysis methods has shifted from the use of analogue techniques towards an increasing utilisation of discrete-time representations. Perhaps the most important factors to influence this development have been the advent of sampled data theory, digital computers and integrated circuits which have given new dimensions to the field of speech processing.

The earliest applications of digital techniques to speech analysis problems included simulations of complex analogue systems in order to obviate the necessity of system construction to experiment with choices of parameters and points of general system design. However, with the increasing power of digital machines and rapid advances in digital signal processing theory, it soon became clear that discrete-time systems possessed capabilities far beyond the simulation of analogue systems. Indeed, it is inappropriate to consider current speech processing systems as approximations to analogue systems as there is often no realisable analogue counterpart.

In this chapter, a few of the more significant digital techniques employed in speech analysis will, therefore, be discussed. The aim is not to present a complete study of the subject, but to give a general overview of some of the multitudinous forms of representations for speech signals. Thus, an attempt is made to unite the material of past and present techniques, with particular attention being focused upon the main characteristics and problems involved.

This review is initiated with an essential distinction being drawn between the two broad groups of speech characterisations: waveform representations and parametric

representations. Several forms of speech model are then introduced — ranging from those which rely heavily on a characterisation of the physiological nature of the human speech production system (such as acoustic tube models) to other methods whose effectiveness stems from their ability to model the statistics of the speech waveform (e.g. homomorphic analysis, linear prediction).

2.2 WAVEFORM/PARAMETRIC REPRESENTATIONS

It is known that the properties of the speech signal contain a high level of redundancy since speech is produced by relatively slow movements of the articulatory tract. For the purposes of speech coding for digital transmission or storage, it therefore seems pertinent to take account of this redundancy in order to reduce the amount of information to be transmitted or stored while accommodating a reasonable increase in system complexity.

Two approaches which may be offered as solutions to this problem include (a) the analytical reduction of redundancy using the statistical properties of the waveform or (b) the extraction of suitably chosen information-bearing parameters which form an essentially complete characterisation of the speech signal. This separation of ideas, although tentative, defines the two broad catergories into which speech coding systems can be divided:

- Waveform representations or
- Source coders or voice coders (also known as vocoders) which yield parametric representations.

Waveform coding techniques generally attempt to encode the acoustic time waveform without taking into account the production and/or perception of the signal. This is in contrast to source coding systems which extract features from the waveform which are subsequently used to drive a speech model in order to produce synthetic speech. A survey of the various waveform coding methods reveals two basic sub-categories : Time Domain and Frequency

Domain Coders.

2.2.1 Time–Domain Waveform Coding

Time-domain coders are by far the best known and currently most widely used waveform coders with advantages such as simplicity and ease of implementation characterising their use. The principal technique is Pulse Code Modulation (PCM) [2.1],[2.2] which consists of the basic operations of sampling, quantisation and representation of the discrete amplitudes by groups of binary digits.

Due to the implicit bias of the speech amplitude probability density function towards lower amplitudes [2.3], a non-uniform quantiser is often used to introduce fine quantising steps for the frequently occurring low sample values and a coarser quantisation for higher amplitudes. Another approach known as Adaptive Quantisation or APCM [2.4],[2.5] employs a dynamic quantiser (uniform or non-uniform) which is able to expand or contract in time to produce a set of time-varying or adaptive quantisation levels.

PCM produces an accurate digital representation of the analogue waveform but does not attempt to exploit the high inter-sample correlation which speech exhibits [2.6]. This fact provides the motivation for a differential encoding scheme where the input to the quantiser is the signal:

$$e(n) = x(n) - \tilde{x}(n) \tag{2.1}$$

This difference between the unquantised input sample $x(n)$ and an estimate, or prediction of the input sample $\tilde{x}(n)$ is often termed the *prediction error.* In many Differential PCM (DPCM) systems, a linear predictor may be used to obtain the estimate $\tilde{x}(n)$ as a linear combination of a number (p) of past values:

$$\tilde{x}(n) = \sum_{i=1}^{p} \alpha_i . (x(n-i) - q(n-i)) \tag{2.2}$$

where $\{\alpha_i : i=1,\ldots,p\}$ are the time-invariant coefficients of a p-th order linear predictor while $\{q(n)\}$ is the quantisation error for the difference signal $\{e(n)\}$.[1] The details of this process along with procedures for the evaluation of the predictor coefficients are discussed at considerably greater length in Section 2.6 and Chapter 3. In DPCM systems, the changes in input speech variance will be reflected in changes of prediction error variance. Since the latter is the input to the DPCM quantiser, adaptation techniques may also be employed in this quantisation process [2.5],[2.7].

Another strategy for waveform coding is exhibited by Delta Modulation (DM) systems [2.8],[2.9] — which are far less complex than their DPCM counterparts although their implementation strategy is similar. In DM, the difference signal is quantised by only 2 levels ($\pm D_n$) and the prediction function is generally first order (i.e. $p=1$) with $\alpha_1=1$. At each sample instant, the polarity of the first backward difference $\{x(n)-x(n-1)\}$ (ignoring quantisation effects) is, therefore, quantised by the step-size D_n. This step value is time-invariant for non-adaptive schemes and adapts to the slope of the input signal for Adaptive Delta Modulation [2.10]. However, this one-bit quantisation is somewhat compensated for by having a sampling rate considerably higher than the Nyquist frequency in order to increase the adjacent sample correlation.

Adaptive Predictive Coding (APC) [2.11],[2.12] is a more sophisticated technique which is designed to exploit the characteristics of the speech production mechanism to a large extent. The complete specification of an APC can be seperated into two distinct stages. The first unit employs a spectral envelope predictor which takes advantage of the correlations

---

[1] *Since the error sequence to be transmitted is a quantised error $\{e(n) - q(n)\}$, the reconstructed signal will ideally be a quantised version $\{x(n) - q(n)\}$ of the input. As shown in eqn (2.2), the input to the predictor is thus $\{x(n) - q(n)\}$ and not the input sequence $\{x(n)\}$ itself.*

between adjacent input samples by modelling the short-time spectral envelope; while the second stage exploits the quasi-periodic nature of voiced speech using a pitch periodicity predictor. Details of the structure of the predictive filters are discussed by Tribolet et al [2.13].

2.2.2 Frequency-Domain Waveform Coding

Frequency domain coders perform signal quantisation and coding in an orthogonal transform domain in order to capitalise on the less correlated samples and the non-uniform power spectral density function.

Although the sub-band coder (SBC) [2.14],[2.15] is essentially a frequency domain technique, it is actually implemented totally in the time domain. This method involves the partitioning of the speech spectrum into several sub-bands (normally 4-8) each of which is coded and transmitted separately. The sub-bands are chosen to represent those regions of the spectrum which are considered to be perceptually most important. In essence, this implies that the formant frequencies define sub-bands and, since the formant bandwidths are narrower at lower frequencies [2.16], it follows that the sub-band ranges increase with frequency. Each sub-band is translated to zero frequency where it is sampled at the Nyquist rate and coded using APCM. At the receiver, the bands are decoded and modulated back to the correct location in the spectrum.

The other important class of speech coding techniques based on frequency domain analysis is that of Transform Coding [2.15],[2.17],[2.18]. The underlying principle of a transform coder involves the operation of an NxN transform matrix T on a vector $\underline{x}$ representing the block or sequence of N speech samples $\{x(0), \ldots, x(N\text{-}1)\}$. The resultant vector $\underline{c}$ of N transform coefficients is then given by:

$$\underline{c} = \mathrm{T}\,\underline{x} \tag{2.3}$$

This set of transform coefficients is individually quantised to $\hat{\underline{c}}$ and then transmitted. In the decoding process, an inverse transformation on $\hat{\underline{c}}$ produces the vector $\hat{\underline{x}}$ as an approximation to the original segment of speech samples $\underline{x}$. Thus,

$$\hat{\underline{x}} = T^{-1}\,\hat{\underline{c}} \tag{2.4}$$

If T is chosen such that the distribution of coefficients $\{\hat{c}_0, \ldots, \hat{c}_{N-1}\}$ can be efficiently exploited by independent quantisation and/or by discarding some of these values, then this method offers the possibility of bit-rate reduction. Several transforms have been suggested for the digital coding of speech signals including the Karhunen-Loève [2.18], Walsh-Hadamard [2.19], Discrete Fourier [2.20] and Discrete Cosine [2.21] transforms. A more detailed discussion of the fundamentals of transform coding and the properties of the various transforms are given in references [2.17],[2.18],[2.20],[2.22].

A large variety of waveform coding schemes have been outlined in the last two subsections and it may be appropriate at this point to establish a common point of reference for this group. Three factors which may be cited as being indicative of the merits of a particular method include:

- Bit-rate (defined in terms of the number of bits, or binary digits per second at which information is transmitted),
- System complexity and
- Speech output quality.

As an oversimplification, it may be stated that, in order to maintain a certain output quality, a reduction in bit-rate is often associated with a corresponding increase in system complexity. Thus, the simplest waveform coding techniques such as PCM, DPCM and Delta Modulation operate successfully in the range 32-64 kbits/s while the adaptive configurations of these methods generally cover the range 16-32 kbits/s. The more complex Sub-Band Coding,

Adaptive Predictive Coding and Transform Coding schemes produce communications quality speech at transmission rates of 9.6-16 kbits/s.

In the above discussion, it has only been possible to provide a brief glimpse of some of the more significant speech waveform coding schemes which are currently available. Due to the increasing importance of efficient coding approaches, many review works have appeared in the literature — either focussing upon a few particular methods of speech coding or addressing a wider range of approaches.

A comprehensive review of the topic before 1972 is presented by Flanagan [2.23], while transform coding, particularly in the Walsh Domain, is briefly reviewed by Elliott [2.24]. Frequency domain coding techniques are dealt with by Tribolet et al [2.15] and time domain methods are covered extensively by Jayant [2.1]. A survey including a number of important approaches to speech coding can also be found in Gold [2.25] and a wider range of coders are considered at some length by Flanagan et al [2.26] along with their relative complexity and performance. More detailed expositions of waveform coding techniques can finally be found in Rabiner & Schafer [2.27] and Jayant & Noll [2.28] with the latter also providing some insights into applications for image processing.

2.2.3 Source Coding

The majority of waveform coding schemes outlined in the two previous subsections have relied upon a characterisation of the speech signal which, in general, does not make use of a rigorous mathematical or physiological model of the speech production mechanism. All of these methods involve either a direct encoding of the acoustic time waveform, or a derived output which exploits certain statistical properties of the speech signal in order to reduce the amount of transmitted or stored information.

However, another class of encoding techniques exists which draws upon a specific model of the speech signal in order to extract certain parameters from the time waveform which form a complete characterisation of the speech data. This analysis is usually performed on a frame-by-frame or *"short-time"* basis. The underlying assumption is that the properties of the speech signal vary slowly with time so that short segments (or frames) may be isolated and processed as if they were segments from a sustained sound with fixed properties.

For voiced utterances, individual cycles of the speech signal may be analysed separately — giving rise to the term *"pitch synchronous analysis"*. Alternatively *"pitch asynchronous"* methods rely on fixed size frames (with typical durations equivalent to 2 or 3 cycles of voiced speech) which may or may not be overlapped. A significant reduction in bit-rate is often possible since the number of parameters derived from a given speech segment is usually much less than the number of samples in that frame.

In Chapter 1, the mechanism of sound production was described in terms of a sound source exciting the resonances of the vocal (and possibly the nasal) tract. This description is one of the most extensively used in speech analysis/synthesis and is known as the source-filter model. This linear model of speech production was developed by Fant in the late 1950's and a detailed discussion of the properties and assumptions behind this model are given by both Fant [2.29] and Flanagan [2.23]. Fig. 2.1 illustrates such a model with the excitation generator producing either a quasi-periodic pulse train for voiced sounds or a random noise waveform for unvoiced speech. A time-varying linear system accounts for the vocal tract and radiation features which are separated from those features associated with the excitation.

In this respect, it may be noted that the source-filter model is actually an oversimplification of the actual speech production system. This is since there is inevitably some coupling between the vocal tract and the lungs during the period when the glottis is open. Although this effect implies that the characteristics of the time-varying linear system of Fig. 2.1 change during each individual cycle of the excitation, this factor is not considered to be of

great significance for practical speech output [2.30]. In the following sections, several forms of speech model are described which rely to varying extents on this source-filter approach.

2.3 ACOUSTIC TUBE MODELLING

The assumption that the vocal cavity can be regarded as a combination of a series of lossless tubes of equal length leads to the widely used acoustic tube model of the vocal tract. Acoustic tube models have been studied for several years [2.31],[2.32] as it may be argued that the resonant frequencies of the concatenated tubes will approach those of a tube with continuously varying cross-section (i.e. the vocal tract) if a large number of tubes of short length are used.

As the transverse dimension of each tube is considered to be small enough compared to a wavelength, sound propagation through an individual section is treated as a plane wave [2.33]. In addition, it is assumed that internal losses within the tubes are negligible and that the effect of the nasal cavity and any vocal tract/glottis coupling may be ignored.

A typical acoustic tube model is depicted in Fig. 2.2. Based upon the laws of conservation of mass & momentum, the equations governing pressure $p_k(x,t)$ and volume velocity $u_k(x,t)$ in the k-th acoustic tube section as a function of time t and distance x (measured from the left hand edge of the tube) are as follows [2.34]:

$$\frac{\partial p_k(x,t)}{\partial x} = -\frac{\rho}{\mathcal{A}_k} \cdot \frac{\partial u_k(x,t)}{\partial t} \quad \text{MOMENTUM EQUATION} \tag{2.5a}$$

$$\frac{\partial u_k(x,t)}{\partial x} = -\frac{\mathcal{A}_k}{\rho c^2} \cdot \frac{\partial p_k(x,t)}{\partial t} \quad \text{CONTINUITY OF MASS EQUATION} \tag{2.5b}$$

where $\mathcal{A}_k$ = Constant Cross-sectional Area of kth Section

ρ = Air Density

c = Spced of Sound in Air

Combining these two equations yields the one-dimensional wave equations for both pressure and volume velocity:

$$\frac{\partial^2 p_k(x,t)}{\partial x^2} = \frac{1}{c^2} \cdot \frac{\partial^2 p_k(x,t)}{\partial t^2} \tag{2.6a}$$

$$\frac{\partial^2 u_k(x,t)}{\partial x^2} = \frac{1}{c^2} \cdot \frac{\partial^2 u_k(x,t)}{\partial t^2} \tag{2.6b}$$

The solutions to these equations are expressed as linear combinations of forward and reverse travelling waves (indicated by the superscripts + and – respectively) of the form:

$$u_k(x,t) = u_k^+\left(t - \tfrac{x}{c}\right) - u_k^-\left(t + \tfrac{x}{c}\right) \tag{2.7a}$$

$$p_k(x,t) = \frac{\rho c}{\mathcal{A}_k}\left[u_k^+\left(t - \tfrac{x}{c}\right) + u_k^-\left(t + \tfrac{x}{c}\right)\right] \tag{2.7b}$$

In particular, continuity conditions at the junction between the k^{th} and $(k+1)^{st}$ tubes as shown in Fig. 2.3 yield:

$$p_k(\ell,t) = p_{k+1}(0,t) \quad ; \qquad u_k(\ell,t) = u_{k+1}(0,t) \tag{2.8}$$

Substituting equation (2.7) into (2.8) gives:

$$u_{k+1}^+(t) = (1+r_k)\, u_k^+(t-\tau) + r_k\, u_{k+1}^-(t) \tag{2.9a}$$

$$u_k^-(t+\tau) = -r_k\, u_k^+(t-\tau) + (1-r_k)\, u_{k+1}^-(t) \tag{2.9b}$$

where $$r_k = \frac{\mathcal{A}_{k+1} - \mathcal{A}_k}{\mathcal{A}_{k+1} + \mathcal{A}_k} \tag{2.10}$$

and $\tau = {}^{\ell}/_{c}$ = Time taken for wave to travel length (ℓ) of one tube section.

From these expressions and Fig. 2.3, it can be seen that r_k is the amount of $u_{k+1}^{-}(t)$ that is reflected at the k^{th} junction and thus r_k is known the *reflection coefficient* for this junction. It is easy to see that since $\mathcal{A}_k > 0$, $\forall k \in \{1,\ldots,N\}$, the following relationship holds for the reflection coefficients,

$$-1 \leq r_k \leq 1\,, \qquad 1 \leq k \leq N\text{–}1 \tag{2.11}$$

A signal-flow graph representation of the overall system is illustrated in Fig. 2.4 (a) and arises as a result of equation (2.9) along with boundary conditions at the two ends of the acoustic tube [2.27] - i.e. the glottis and the lips. In this diagram,

$u_G(t)$ = Source Volume Velocity at Glottis.

$u_L(t)$ = Output Volume Velocity at Lips.

r_G = Reflection Coefficient at the Glottis.

r_L = Reflection Coefficient at the Lips.

It is also possible to realise a discrete-time version of this system as the digital filter configuration shown in Fig. 2.4 (b). In this digital analogue, the sampling period (T) is such that $T = 2\tau$. Digital networks of this form can be used to produce synthetic speech samples from samples of an appropriate excitation signal [2.32]. It is also possible to employ this discrete-time representation in the evaluation of the overall transfer function of the acoustic tube model [2.35].

2.4 FORMANT ANALYSIS

The acoustic tube models described in the previous section attempt to reproduce the shape of the vocal tract in order to gain an approximation to the physical variables of air pressure and volume velocity within the tract. However, another class of speech models exists which does not rely on a direct representation of the physics of speech production. These *"Terminal-Analogue Models"* are directed towards implementing a system whose transfer function approximates the vocal tract transfer function, but their implementation bears no direct relation to the details of vocal tract structure.

2.4.1 Cascade Formant Models

One important type of terminal-analogue model is that based on formant modelling — in fact, this term is often exclusively applied to formant-based analysis. This form of representation relies on the observation that the acoustic properties of the vocal tract lead to a set of natural modes or resonant frequencies characterised by the formants. Extensive psycho-acoustical studies [2.36] have also shown that these spectral peaks are perceptually the most important features of the speech spectrum. Therefore, an efficient technique for representing the vocal tract transfer function lies in its approximation by a cascade or sequence of 2nd order resonators as shown in Fig. 2.5 — with each resonator representing one of the speech formants. A control strategy to vary the parameters of centre frequencies and bandwidths is used to simulate the continuously varying shape of the vocal tract.

Each individual 2nd order resonator may, of course, also be visualized as a two pole discrete time system so that the entire model of Fig. 2.5 is an all-pole description. Although this characterisation is adequate for the majority of speech sounds, utterances such as nasals and fricatives require both resonances and antiresonances (i.e. poles and zeros) [2.27]. However, for the moment, this discussion will be restricted to all-pole models. A typical s-plane description of a two pole model consisting of a pair of complex conjugate poles $\{s_k, s_k^*\}$

is, therefore, given by:

$$s_k , s_k^* = -\sigma_k \pm j2\pi F_k \tag{2.12}$$

This resonance is illustrated in Fig. 2.6 (a) and thus the bandwidth of this system is approximately $2\sigma_k$ and the centre frequency is around $2\pi F_k$. The corresponding z-plane model is shown in Fig. 2.6 (b) and corresponds to the following pair of complex poles:

$$z_k , z_k^* = r_k e^{\pm j\theta_k} \tag{2.13}$$

or, since $z = e^{sT}$,

$$z_k , z_k^* = e^{-\sigma_k T} . e^{\pm j2\pi F_k T} \tag{2.14}$$

Thus, $$r_k = e^{-\sigma_k T} ; \qquad \theta_k = 2\pi F_k T \tag{2.15}$$

The transfer function $\left\{ V_k(z) \right\}$ of this resonator is then given by:

$$V_k(z) = \frac{G_k}{\left(1 - r_k e^{+j\theta_k} z^{-1} \right)\left(1 - r_k e^{-j\theta_k} z^{-1} \right)} \tag{2.16}$$

$$= \frac{G_k}{\left(1 - 2 r_k \cos\theta_k z^{-1} + r_k^2 z^{-2} \right)} \tag{2.17}$$

The gain factor (G_k) is often chosen such that the gain of the 2-pole model is unity at zero frequency (i.e. $z = 1$) so that:

$$G_k = 1 - 2 r_k \cos\theta_k + r_k^2 \tag{2.18}$$

The overall transfer function $\{V(z)\}$ of the complete cascade model consisting of M resonators is:

$$V(z) = \prod_{k=1}^{M} \left[\frac{G_k}{\left(1 - 2 r_k \cos\theta_k z^{-1} + r_k^2 z^{-2} \right)} \right] \tag{2.19}$$

Such a model is illustrated as a signal-flow graph in Fig. 2.7. Although this configuration may provide a description for the vocal tract, additional filters are required to represent both the effects of radiation from the lips and of the excitation source providing an input to the vocal tract.

2.4.1.1 EXCITATION SOURCE

For voiced speech, the excitation source produces a waveform whose frequency components decay at about -12 dB/octave and so the synthetic excitation waveform must possess a similar frequency characteristic. One approach simulates the actual shape of the glottal wavefrom by a piecewise concatenation of sinusoidal or exponential functions [2.37],[2.38]. Another strategy relies on an attempt to represent the glottal frequency response using a two pole model for the transfer function of the glottal model, $G(z)$:

$$G(z) = \frac{1}{\left(1 - e^{-c} z^{-1}\right)^2} \tag{2.20}$$

The input to this filter is, therefore, an impulse train with the required pitch period and the pair of real poles at $z = e^{-c}$, $c > 0$ ensure the necessary -12 dB/octave decay. The parameter c is chosen such that the filter cut-off is around 100 Hz [2.39]. For unvoiced speech, the output speech quality is much less sensitive to the form of the excitation and broadband white noise is quite acceptable [2.30].

2.4.1.2 RADIATION LOAD

The effects of radiation from the lips account for the transformation of volume velocity at the lips (i.e. the waveform $u_L(n)$ shown in Fig. 2.5) to the acoustic pressure waveform $p_L(n)$ at some distance from the lips. This transformation is necessary since $p_L(n)$ represents the actual speech waveform which is recorded on microphone or heard by other listeners. The transfer function of the radiation model $\{R(z)\}$ is, therefore, such that

$$P_L(z) = R(z) . U_L(z) \tag{2.21}$$

A reasonable approximation to the radiation effects is given by a first backward difference operation:

$$R(z) = \left(1 - z^{-1}\right) \tag{2.22}$$

The use of such a model is justified by Flanagan [2.23], who also presents the results of some experiments on acoustic radiation to corroborate this model. The single-zero nature of this function produces a +6dB/octave lift on the lip volume velocity. Combining the effects of the glottal pulse, vocal tract and radiation components, an overall transfer function $H(z)$ for the speech model is obtained:

$$H(z) = G(z) . V(z) . R(z) \tag{2.23}$$

As the glottal component for voiced speech is associated with a −12 dB/octave decay while $R(z)$ is characterised by a +6dB/octave lift, the vocal tract transfer function $V(z)$ is essentially flat (with peaks superimposed at the formant frequencies) to produce an overall −6 dB/octave fall-off characteristic of actual speech waveforms. A diagram illustrating the complete model is shown in Fig. 2.8 with a switch implementing the change between voiced

and unvoiced excitations. It should be noted that the overall structure of such a model may be applied to a variety of speech models which provide representations of the vocal tract. It is also possible to broaden the scope of this basic model by adding other paths to account for utterances such as nasals and voiced fricatives etc. [2.40]

2.4.2 Parallel Formant Models

The formant synthesisers considered in the previous section were implemented as a cascade of 2nd order resonators in series. However, it is also possible to utilise a parallel connection of resonators with each filter being associated with an individually controllable amplitude as shown in Fig. 2.9. Under the influence of certain constraints, this configuration can produce an identical output to that for a cascade model. These constraints and some of the other advantages and disadvantages of the cascade and parallel modes of connection are discussed in Chapter 6.

It has long been recognised that the use of formant information represents a very efficient technique for encoding the major characteristics of the speech spectral envelope. For this reason, formant models are one of the oldest forms of parametric representation for speech signals and the earliest dynamically controlled formant synthesisers were first described over 30 years ago [2.44],[2.45]. Indeed, an early work by Lawrence [2.46] describes a formant synthesiser which produced good intelligibility of synthesised speech at an estimated information rate of 550 bits/s. Formant vocoders [2.23],[2.47],[2.48] have also been implemented for low bit-rate transmission of speech at rates of $600-1000$ bits/s for an output of reasonable quality. These vocoders typically transmit the first 3 formant parameters together with pitch period, voicing & overall amplitude information. These control values are usually updated approximately 50 times a second and then quantised, transmitted and finally used in a formant synthesiser at the receiver to produce output speech.

A very important topic which has not been mentioned in the above discussion concerns the formant "*analysis*" problem. This term relates to the complications involved in the extraction of accurate formant estimates from speech data. This stage often poses serious difficulties since it usually involves complex and often unreliable analysis procedures. The formant extraction problem is addressed in detail in Chapters 5,6 along with several other issues concerning formant models.

2.5 HOMOMORPHIC SYSTEMS

2.5.1 Generalized Homomorphic Analysis

A system is defined as being *homomorphic* in nature if it can be shown to satisfy a generalized principle of superposition for some operation $\otimes$ [2.50]. Therefore, if the input $x(n)$ to a homomorphic system $\mathcal{H}[\cdot]$ is such that

$$x(n) = x_1(n) \otimes x_2(n) \tag{2.24}$$

Then,
$$\begin{aligned} \mathcal{H}\big[x(n)\big] &= \mathcal{H}\big[x_1(n) \otimes x_2(n)\big] \\ &= \mathcal{H}\big[x_1(n)\big] \otimes \mathcal{H}\big[x_2(n)\big] \\ &= y_1(n) \otimes y_2(n) = y(n) \end{aligned} \tag{2.25}$$

In the specific case where $\otimes$ denotes addition, $\mathcal{H}[\cdot]$ represents a linear system and thus homomorphic analysis illustrates a generalisation of the concept of a linear system. For the analysis of speech, it is instructive to study "*homomorphic systems for convolution*". The discrete-time convolution operator $(*)$ is usually expressed as:

$$x_1(n) * x_2(n) = \sum_{k=-\infty}^{\infty} x_1(n-k) \cdot x_2(k) \tag{2.26}$$

An important property of homomorphic systems is that there exists a canonic form for the representation of these systems [2.51]. The function $\mathcal{H}[\cdot]$ can, therefore, be replaced by a

cascade of 3 homomorphic units, as shown in Fig. 2.10. The first unit $\mathcal{D}[\cdot]$ is known as the *characteristic system for homomorpic deconvolution* and satisfies the following constraint:

$$\begin{aligned}\mathcal{D}\big[x(n)\big] &= \mathcal{D}\big[x_1(n) * x_2(n)\big] \\ &= \mathcal{D}\big[x_1(n)\big] + \mathcal{D}\big[x_2(n)\big] \\ &= \hat{x}_1(n) + \hat{x}_2(n) = \hat{x}(n)\end{aligned} \tag{2.27}$$

The second system $\mathcal{L}[\cdot]$ is a conventional linear system while the third unit $\mathcal{D}^{-1}[\cdot]$ represents the inverse of $\mathcal{D}[\cdot]$. The existence of such a canonic form implies that the design of the overall system reduces to the problem of producing an appropriate linear system. The validity of this decomposition is illustrated as follows:

$$\begin{aligned}\mathcal{H}\big[x(n)\big] &= \mathcal{D}^{-1}\Big[\mathcal{L}\Big[\mathcal{D}\big[x(n)\big]\Big]\Big] \\ &= \mathcal{D}^{-1}\Big[\mathcal{L}\Big[\mathcal{D}\big[x_1(n) * x_2(n)\big]\Big]\Big] \\ &= \mathcal{D}^{-1}\Big[\mathcal{L}\Big[\mathcal{D}\big[x_1(n)\big] + \mathcal{D}\big[x_2(n)\big]\Big]\Big] \\ &= \mathcal{D}^{-1}\Big[\mathcal{L}\Big[\mathcal{D}\big[x_1(n)\big]\Big] + \mathcal{L}\Big[\mathcal{D}\big[x_2(n)\big]\Big]\Big] \\ &= \mathcal{D}^{-1}\Big[\mathcal{L}\Big[\mathcal{D}\big[x_1(n)\big]\Big]\Big] * \mathcal{D}^{-1}\Big[\mathcal{L}\Big[\mathcal{D}\big[x_2(n)\big]\Big]\Big] \\ &= \mathcal{H}\big[x_1(n)\big] * \mathcal{H}\big[x_2(n)\big]\end{aligned} \tag{2.28}$$

This final result is, therefore, as required for the homomorphic system, $\mathcal{H}[\cdot]$. From equation (2.27), it is clear that, in frequency domain terms,

$$X(z) = X_1(z).X_2(z) \tag{2.29a}$$

and $$\hat{X}(z) = \hat{X}_1(z) + \hat{X}_2(z) \tag{2.29b}$$

Thus, the characteristic system $\mathfrak{D}[\cdot]$ converts a multiplicative combination of input z-transforms into an additive sum of output z-transforms. One approach to achieve this result is the use of logarithms such that

$$\hat{X}(z) = \log\left[X(z)\right] \tag{2.30}$$

The system $\mathfrak{D}[\cdot]$ is then as shown in Fig. 2.11. In terms of Fourier representations, the operator $\mathfrak{D}[\cdot]$ is, therefore, realised as an inverse Fourier transform operation on the logarithm of the input Fourier transform:

$$X\left(e^{j\omega}\right) = \mathfrak{F}\left[x(n)\right] \tag{2.31a}$$

$$\hat{X}\left(e^{j\omega}\right) = \log\left[X\left(e^{j\omega}\right)\right] \tag{2.31b}$$

$$\hat{x}(n) = \mathfrak{F}^{-1}\left[\hat{X}\left(e^{j\omega}\right)\right] \tag{2.31c}$$

where $\mathfrak{F}[\cdot]$ represents a Fourier transform and $\mathfrak{F}^{-1}[\cdot]$ denotes its inverse. The output, $\hat{x}(n)$, of the characteristic system is known as the "*complex cepstrum*" [2.52] of the input $s(n)$ and, in fact, the term *cepstral processing* is often synonymous with homomorphic analysis in the field of speech processing. Other terms which are associated with cepstral analysis include:

(i) "*Quefrency*" – the independent variable, or index n, of the complex cepstrum [2.53].

(ii) High time/low time regions of the cepstrum – corresponding to high and low quefrencies respectively.

The complex logarithm described by equation (2.31b) is calculated according to the following expression in order to deal with the problem of uniqueness [2.54]:

$$\log\left[X\left(e^{j\omega}\right)\right] = \log\left[\left|X\left(e^{j\omega}\right)\right|\right] + j\arg\left[X\left(e^{j\omega}\right)\right] \tag{2.32}$$

Since the ear is known to be relatively insensitive to phase [2.55], the phase information retained in this equation can be discarded as follows for speech-oriented analysis:

$$X\left(e^{j\omega}\right) = \mathcal{F}\left[x(n)\right] \tag{2.33a}$$

$$C\left(e^{j\omega}\right) = \log\left[\left|X\left(e^{j\omega}\right)\right|\right] \tag{2.33b}$$

$$c(n) = \mathcal{F}^{-1}\left[C\left(e^{j\omega}\right)\right] \tag{2.33c}$$

The sequence $\{c(n)\}$ is simply referred to as the "*cepstrum*" and displays the same general properties as the complex cepstrum [2.50].

2.5.2 Homomorphic Processing of Speech

In the case of speech signals, homomorphic systems for convolution are particularly relevant since, as discussed in section 2.2.3, the speech waveform $s(n)$ is often modelled as the convolution of an excitation function $e(n)$ and the vocal tract impulse response $v(n)$:

$$s(n) = e(n) * v(n) \tag{2.34}$$

As described in the last sub-section, the system $\mathcal{D}[\cdot]$ is such that

$$\mathcal{D}[s(n)] = \mathcal{D}[e(n)] + \mathcal{D}[v(n)] \tag{2.35a}$$

$$\text{or} \qquad \hat{s}(n) = \hat{e}(n) + \hat{v}(n) \tag{2.35b}$$

where $\{\hat{s}(n)\}$, $\{\hat{e}(n)\}$, $\{\hat{v}(n)\}$ denote the complex cepstra corresponding to the sequences $\{s(n)\}$, $\{e(n)\}$, $\{v(n)\}$ respectively. Thus the characteristic system separates the excitation and vocal tract effects into two additive components. This result is particularly important in the light of the two following observations concerning the cepstral sequences $\{\hat{e}(n)\}$, $\{\hat{v}(n)\}$.

(a) Cepstrum of the Excitation:

Since the excitation signal for voiced speech in the source-filter model may be regarded as a quasi-periodic series of pulses spaced at the pitch period N_p, the waveform $e(n)$ may be expressed as:

$$e(n) = \sum_{k=-\infty}^{\infty} \alpha_k . \delta(n - kN_p) \tag{2.36}$$

where $\delta(n)$ is the unit impulse defined by

$$\delta(n) = 1 \qquad n = 0$$

$$= 0 \qquad \text{otherwise.}$$

It can be shown [2.27] that a sequence such as $e(n)$ is associated with a complex cepstrum $\hat{e}(n)$ of the form:

$$\hat{e}(n) = \sum_{k=-\infty}^{\infty} \beta_k . \delta(n - kN_p) \tag{2.37}$$

The cepstrum of the excitation signal is, therefore, also a train of impulses (of varying amplitude) spaced at the pitch period. In addition, a high frequency periodic signal will produce a more widely spaced sequence of pulses in the cepstrum than a low frequency signal.

(b) Cepstrum of the Vocal Tract Impulse Response:

As discussed in section 2.4, the transfer function of the vocal tract can be described by a rational z-transform $V(z)$ consisting of a series of poles and zeros inside the unit circle — although only poles may be necessary for most voiced speech. Thus,

$$V(z) = G \cdot \frac{\prod_{k=1}^{M} \left(1 - b_k z^{-1}\right)}{\prod_{k=1}^{N} \left(1 - a_k z^{-1}\right)}, \qquad \text{where } |a_k|, |b_k| < 1. \tag{2.38}$$

The complex cepstrum, $\hat{v}(n)$, associated with this expression is such that [2.27]:

$$\begin{aligned} \hat{v}(n) &= \log[G] && n = 0 \\ &= \sum_{k=1}^{N} \frac{a_k^n}{n} - \sum_{k=1}^{M} \frac{b_k^n}{n} && n > 0 \\ &= 0 && n < 0 \end{aligned} \tag{2.39}$$

Hence, the complex cepstrum of the vocal tract impulse response is a rapidly decaying series concentrated around n=0. These two observations lead to the result that the excitation and the vocal tract components tend to occupy disjoint intervals within the speech cepstrum. The low-time region near the origin is representative of the slow variation with frequency of the vocal tract features, while the more rapid variation of the excitation contributes to the high-time portion.

The extraction of the deconvolved vocal tract component is, therefore, achieved by simply setting to zero the high-time region of the cepstrum and performing a Fourier transform to yield the estimated vocal tract transfer function. This frequency response is a smoothed

version of the speech spectrum and can be used for formant estimation etc. [2.56],[2.57].

Alternatively, the pitch period can be estimated by searching for a peak away from the origin in the vicinity of the expected pitch period [2.53],[2.58]. Subsidiary peaks at multiples of the pitch period may also be evident. This procedure can also be employed as the basis of a voiced/unvoiced detection algorithm.

The interlinked processes of spectral envelope estimation and pitch detection are illustrated in Fig. 2.12 for an actual speech signal. This particular segment is sampled at 10kHz and originates from the voiced utterance /OW/ as in "bought". The cepstrum is calculated as above and an expanded version of the high time part of the cepstrum in Fig. 2.12 (d) clearly illustrates a peak at a quefrency corresponding to the pitch period. The estimated vocal tract transfer function in Fig. 2.12 (f) is also as expected since it provides a smoothed version of the speech spectrum of Fig. 2.12 (b).

The principles of homomorphic analysis can also be employed for vocoding purposes with the deconvolved vocal tract information and pitch period being transmitted separately. At the receiver, the characteristic system for homomorphic convolution $\left\{\mathcal{D}^{-1}[\cdot]\right\}$ is applied to the low-time region of the cepstrum, resulting in a signal which is then convolved with an appropriate excitation function based on the pitch information. Speech transmission with little degradation has been achieved with a bit-rate of 7.8 kbits/s [2.59].

2.6 LINEAR PREDICTIVE ANALYSIS

2.6.1 Formulation of the Linear Predictive Model

In section 2.4.1, the transfer function of the speech production model, $H(z)$, was described in terms of a filter combining the effects of glottal excitation, vocal tract and the radiation load such that:

$$H(z) = G(z).V(z).R(z)$$

$$= \frac{\left(1-z^{-1}\right)}{\left(1-e^{-c}z^{-1}\right)^2} \cdot \prod_{k=1}^{M}\left[\frac{G_k}{\left(1-2r_k\cos\theta_k z^{-1}+r_k{}^2 z^{-2}\right)}\right] \tag{2.40}$$

As the parameter c is such that, $0 < c \ll 1$, the numerator term $\left(1-z^{-1}\right)$ is virtually cancelled by one of the denominator terms $\left(1-e^{-c}z^{-1}\right)$ [2.39]. Thus, an expansion of the denominator yields an expression of the form:

$$H(z) = \frac{G}{A(z)} \tag{2.41}$$

where $$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \tag{2.42}$$

The system $H(z)$ consisting of the glottal, vocal tract and radiation components is, therefore, characterised by a single all-pole model $1/A(z)$. As $H(z)$ describes the transfer function of the speech model, it effectively represents the filter which converts an excitation function, $u(n)$, to the output speech, $s(n)$. Therefore,

$$H(z) = \frac{S(z)}{U(z)} \tag{2.43}$$

Combining equations (2.41) and (2.43) yields

$$s(n) = \sum_{k=1}^{p} \alpha_k s(n-k) + Gu(n) \tag{2.44}$$

The system excitation signal, $u(n)$, will either consist of unit impulses spaced at the appropriate pitch period for voiced speech, or random noise of unit variance for unvoiced sounds. This configuration is illustrated in Fig. 2.13 for a linear predictive synthesiser

network.[2] Now, a linear predictor with prediction coefficients $\{\alpha_k: k=1,\ldots,p\}$ is defined as a system where the present output $\tilde{s}(n)$ is estimated using a linear combination of the previous p speech samples. Thus,

$$\tilde{s}(n) = \sum_{k=1}^{p} \alpha_k s(n-k) \tag{2.45}$$

This type of filter was discussed in section 2.2 in relation to DPCM coders in order to reduce the redundancy of the speech signal by a minimisation of the prediction error, $e(n)$:

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \tag{2.46}$$

From equations (2.44),(2.46), it can be seen that, if the speech signal conforms exactly to the all-pole description mentioned above, then

$$e(n) = Gu(n) \tag{2.47}$$

Linear prediction (LP) analysis can, therefore, be formulated as the problem of calculating a set of predictor coefficients $\{\alpha_k: k=1,\ldots,p\}$ and gain parameter G for a given model order p in such a manner as to obtain a good estimate of the input speech sequence. The basic approach relies on minimising the mean-squared prediction error in order to obtain the required predictor parameters. Not only does this minimisation yield a model which forms a very useful and accurate representation for many speech sounds, it also leads to a set of

---

[2] *This direct form implementation is the most straightforward method for speech synthesis from the predictor parameters. However, other configurations which are more appropriate for hardware implementations are also possible [2.39].*

linear equations which can be solved efficiently to obtain the LP parameters. The various techniques which can be employed to solve these equations along with a variety of other issues concerning Linear Predictive Analysis are considered in much greater depth in Chapter 3.

2.6.2 Linear Predictive Coding (LPC)

One of the many applications of LP models (which may discussed here in relation to other techniques covered in previous sections) is the deployment of LP methods within vocoding schemes. The first linear prediction vocoder was described in 1968 by Itakura and Saito [2.60] while various issues affecting LP parameter quantisation and model order selection are discussed in a classical paper by Atal & Hanauer [2.35]. This work describes various coding schemes with bit rates down to 2.4 kbits/s being found to produce fairly good speech reproduction. Many other developments on the subject of LP vocoders have occurred during the last decade and have enabled linear predictive methods to become one of the most popular approaches to the problem of low bit-rate speech coding.

In an attempt to attain higher quality speech output from an LP vocoder, a technique which uses a non-impulsive source has been suggested [2.61] and has been found to be important in effectively masking the characteristic buzz often associated with LP synthetic speech. Another work [2.62] has suggested the Karhunen-Loève transformation of a modified set of predictor coeffecients in order to exploit the redundancy still inherent in these coefficients. Although the output quality is speaker-dependent, bit-rates as low as 600 kbits/s have been attained. However, this achievement is offset by a considerable increase in system complexity due to the introduction of eigenvector analysis for the computation of the transformed coefficients.

Other coding techniques employing LP methods include the Residual Excited Linear Predictive (RELP) Coder developed by Un & Magill [2.63]. This system transmits not only the predictor parameters, but also a quantised version of the prediction error signal. This

"residual" is used at the receiver to excite the LP filter instead of a series of impulses or random noise. High quality speech has been achieved at bit-rates around 9.6 kbits/s. A major advantage of RELP coders is their inherent robustness in most operating environments since there is no need for an explicit pitch extraction algorithm.

More recent developments include the introduction of multi-pulse [2.64],[2.65] and code-excited (CELP) [2.66],[2.67] linear predictive coders. The multi-pulse model assumes that the excitation can be represented with sufficient accuracy as a sequence of pulses (typically 4 to 8 pulses every 5ms). In code-excited coders, the excitation is selected from a random codebook of white Gaussian sequences. The optimum excitation in both coders is chosen to minimise a subjective criterion based on properties of human auditory perception. Both forms of coders produce high quality output at approx. 9.6 kbits/s. Although these coding algorithms are complex — requiring over 10 million multiply-add operations per second [2.64] — new fast digital signal processor chips offer the possibility of real-time implementation.

Vector quantisation (VQ) systems employing LP analysis offer yet another alternative for very low bit-rate transmission. Although information theory and rate distortion theory have long upheld the value of vector or block quantisation, its widespread use in speech analysis has arisen largely as a consequence of its application to the LP vocoder [2.68],[2.69].

In the VQ process, the complete set of LP coefficients (or a suitably modified version of them) is mapped onto a finite set of reference vectors. The reference vector which minimises a certain distance measure between itself and the input sequence of coefficients is that chosen to represent the input. Cluster analysis [2.70] provides a powerful tool for the design of the codebook of reference templates. As only the index of the chosen reference vector is to be transmitted, a significant reduction in overall bit-rate is possible. A comparative analysis has shown [2.71] that the subjective performance of a 1 kbits/s LPC-VQ coder is similar to a 2.4 kbits/s standard LPC vocoder. The obvious disadvantage is the considerable computation required to both produce the codebook and for a full search for the optimal reference template.

2.7 OTHER FORMS OF SPEECH MODELS

2.7.1 Channel Vocoder

The oldest form of source coding device for speech analysis is the channel vocoder, which was first postulated by Dudley [2.72],[2.73] over 40 years ago. The general structure of a typical channel vocoder is illustrated in Fig. 2.14. The vocal tract spectral envelope is, therefore, estimated by a bank of N (typically 14) band-pass filters covering the range of the speech frequency spectrum. The output of each channel is rectified and low-pass filtered to represent the time-varying average signal amplitude of that frequency band. The extraction of source information is governed by voiced/unvoiced decision and pitch detection algorithms. At the receiver, these excitation signals are generated and used as an input to a bank of N band-pass filters whose outputs are amplitude modulated by the corresponding channel signals and finally recombined to form the synthesised speech.

The representation of the vocal tract transfer function can, therefore, be regarded as a process of sampling the power spectral density function in order to reduce the bandwidth of the transmitted signal. Hence this framework affords a more informal mathematical description of the vocal tract than those presented in previous sections.

Although intelligible speech can usually be produced by channel vocoders operating at around 2.4 kbits/s [2.74], the high cost of the coding operation and a certain lack of naturalness in the decoded speech limit the uses of the channel vocoder. However, for historic reasons and due to its inherent versatility, the channel vocoder has been the subject of several intensive investigations [2.75],[2.76] leading to a fair degree of implementational sophistication.

2.7.2 Articulatory Speech Models

This form of modelling attempts to directly represent the articulatory process which generates a speech sound. As the process of speech production is a fairly complex operation, articulatory modelling entails a description of a whole array of physical parameters. These variables may include the effects of lung pressure, glottal area and mass, vocal tract wall vibration, viscous friction, the vocal tract area function, nasal coupling, frication noise etc. In fact, it is not unusual for 50 parameters to be required to adequately describe an articulatory model. The analysis problem of obtaining articulatory parameters from spoken utterances is, therefore, particularly complex.

Despite this significant drawback, articulatory speech synthesis has the potential for very natural speech output at 2.4 kbits/s [2.77]. This is largely since the synthesiser control signals have a direct interpretation in terms of physiological data so that, since these parameters are known to vary slowly in the human voice production system, they are potential candidates for efficient coding. Since it is impossible to determine the best set of parameters to control the articulatory model, a wide variety of schemes have been suggested [2.78],[2.79] but the often prohibitive computational cost of parameter estimation remains a serious obstacle to the widespread usage of articulatory models.

2.7.3 Hidden Markov Models

One of the fundamental problems in speech processing today is that of achieving a high accuracy, large vocabulary, speaker/speech independent speech recognition system. Although this daunting task has, not surprisingly, failed to receive a complete solution, a large degree of success has been achieved in recent years — with *Hidden Markov Models* (*HMM*) contributing to a great extent in this development.

The concept of these models is substantially different from any of those discussed in previous sections since they do not, in general, model specific structures within the speech production mechanism, nor do the sub-units of the HMM correspond to discernible segments of the input speech signal. Instead, they provide a mathematical framework which is analytically tractable and seems to provide an overall description of the statistical properties of the speech signal. This is in direct contrast to articulatory models which attempt to describe not only the acoustic output, but also the speech production system.

The basic theory of Markov chains has been well established for over 80 years with a Markov process being constituted of a certain number (N) of states given by $\mathbb{Q} = \{q_1,\ldots,q_N\}$. Transitions from one state to another are governed by the state transition probability distribution $\mathbb{A} = \{a_{ij} : 1 \leq i,j \leq N\}$ where a_{ij} is the probability that the system is in state j given that it was in state i at the previous sampling instant.

$$\text{i.e.,} \quad a_{ij} \triangleq \text{Prob}(\text{state } q_j \text{ at time } t{+}1 \mid \text{state } q_i \text{ at time } t)$$

Another vector $\pi = \{\pi_i : 1 \leq i \leq N\}$ represents the initial state distribution such that

$$\pi_i \triangleq \text{Prob}(\text{state } q_i \text{ at time } t{=}1)$$

In "*Hidden*" Markov Models, the states themselves are not the observed outputs. A discrete set of M possible observations given by $\mathbb{V} = \{v_1,\ldots,v_M\}$ may instead be the only observable symbols. Within each state, there will be set of probabilities for the observation of a particular output and these probabilities are denoted by $\mathbb{B} = \{b_j(k) : 1 \leq j \leq N,\ 1 \leq k \leq M\}$ where

$$b_j(k) \triangleq \text{Prob}(\text{output } v_k \text{ at time } t \mid \text{state } q_j \text{ at time } t)$$

These observable outputs may be LPC vectors, cepstral coefficients etc. [2.80] which have been reduced to a finite set of symbols by vector quantisation. It is also possible to use continuous probability density functions for the observation densities, or even mixtures of various continuous functions [2.81]. The complete Hidden Markov Model ($\mathbb{H}$) is then specified by these three sets of probabilities,

$$\mathbb{H} = (\pi, \mathbb{A}, \mathbb{B})$$

There are various iterative procedures [2.82] for calculating the best set of model parameters for a given observation sequence $\mathbf{O} = (O_1, O_2, \ldots, O_T)$ where $O_t \in \mathbb{V}$, $\forall t \in \{1, \ldots, T\}$. This operation is known as the *"training"* phase of the recognition process while the *"evaluation"* problem attempts to calculate the probability $\text{Prob}(\mathbf{O} \mid \mathbb{H})$ of obtaining a given observation sequence $\mathbf{O}$ for a particular model $\mathbb{H}$. The utterance associated with the model $\mathbb{H}^*$ which produces the highest value of $\text{Prob}(\mathbf{O} \mid \mathbb{H})$ is then chosen as the recognised output utterance. A major advantage of HMM's is that although the training stage is computationally expensive, the important evaluation stage (which is carried out during the actual recognition process) is relatively simple and thus high quality speech recognition can be implemented at a reasonable cost in computation and memory.

In theory, it is possible to have a state transition matrix $\mathbb{A}$ such that transitions are possible from any one state to any other. However, in practice, there seems to be no significant loss of recognition performance if the much simpler structure of a constrained model such as that shown in Fig. 2.15 is employed. In fact, improved recognition rates can often be achieved for a constrained model over the unconstrained case [2.83]. The configuration of Fig. 2.15 is known as a left-to-right model since the state sequence[3] must always proceed in this direction.

---

[3] *It should be noted that, in general, these states do not correspond to words, phonemes or any other such speech sub-units [2.83].*

A great deal of work has been carried out on HMM's in recent years [2.81],[2.82],[2.84],[2.85]. This activity has mainly been concerned with speech recognition although other applications such as formant tracking [2.86] have also been considered. Recent progress has led to several systems including a real time, isolated utterance speech recognizer for 20,000 words based on the IBM Personal Computer AT and IBM Signal Processors conceived by Jelinek et al [2.87]. These and other similar developments tend to suggest that Hidden Markov Models have a significant part to play in the future towards approaching the elusive goal of high quality speech recognition systems.

2.8 REFERENCES

[2.1] JAYANT, N.S., "Digital Coding of Speech Waveforms : PCM, DPCM and DM Quantisers", *Proc. IEEE, Vol* 62, *No* 5, *May* 1974, *pp* 611-632.

[2.2] OLIVER, B.M., PIERCE, J.R., SHANNON, C.E., "The Philosophy of PCM", *Proc. IRE, Vol* 36, *No* 11, *November* 1948, *pp* 1324-1331.

[2.3] DAVENPORT, W.B., Jr., "An Experimental Study of Speech Wave Probability Distributions", *J. Acoust. Soc. Amer., Vol* 24, *No* 4, *July* 1952, *pp* 390-399.

[2.4] WILKINSON, R.M., "An Adaptive Pulse Code Modulator for Speech", *Proc.* 1971 *IEEE Intl. Conf. on Comm., Montréal, Canada, June* 1971, *pp* 1-11 *to* 1-15.

[2.5] JAYANT, N.S., "Adaptive Quantization with a One-Word Memory", *Bell Syst. Tech. Jnl., Vol* 52, *No* 7, *September* 1973, *pp* 1119-1144.

[2.6] NOLL, P., "A Comparative Study of Various Schemes for Speech Encoding", *Bell Syst. Tech. Jnl., Vol* 54, *No* 9, *November* 1975, *pp* 1597-1614.

[2.7] CUMMISKEY, P., JAYANT, N.S., FLANAGAN, J.L., "Adaptive Quantization in Differential PCM Coding of Speech", *Bell Syst. Tech. Jnl., Vol* 52, *No* 7, *September* 1973, *pp* 1105-1118.

[2.8] JAGER, F. de, "Deltamodulation, a Method of PCM Transmission Using the 1-Unit Code", *Philips Res. Rep., Vol* 7, *No* 6, *December* 1952, *pp* 442-466.

[2.9] STEELE, R., "Delta Modulation Systems", *Pentech Press, London*, 1975.

[2.10] JAYANT, N.S., "Adaptive Delta Modulation with a One-Bit Memory", *Bell Syst. Tech. Jnl., March* 1970, *pp* 321-342.

[2.11] ATAL, B.S., SCHROEDER, M.R., "Adaptive Predictive Coding of Speech Signals", *Bell Syst. Tech. Jnl., Vol* 49, *No* 8, *October* 1970, *pp* 1973-1986.

[2.12] ATAL, B.S., SCHROEDER, M.R., "Predictive Coding of Speech Signals and Subjective Error Criteria", *IEEE Trans. on Acoust., Speech and Sig. Proc., Vol ASSP-27, No* 3, *June* 1979, *pp* 247-254.

[2.13] TRIBOLET, J.M., NOLL, P.,McDERMOTT, B.J., CROCHIERE, R.E., "A Comparison of the Performance of Four Low Bit Rate Speech Waveform Coders", *Bell Syst. Tech. Jnl., Vol* 58, 1979, *pp* 699-712.

[2.14] CROCHIERE, R.E., WEBBER, S.A., FLANAGAN, J.L., "Digital Coding of Speech in Sub-Bands", *Bell Syst. Tech. Jnl., Vol* 55, *No* 8, *October* 1976, *pp* 1069-1085.

[2.15] TRIBOLET, J.M., CROCHIERE, R.E., "Frequency Domain Coding of Speech", *IEEE Trans. on Acoust., Speech and Sig. Proc., Vol ASSP-27, no* 5, *October* 1979, *pp* 512-530.

[2.16] KLATT, D.H., "Software for a Cascade/Parallel Formant Synthesizer", *J. Acoust. Soc. Amer., Vol* 67, *No* 3, *March* 1980, *pp* 971-995.

[2.17] ZELINSKI, R., NOLL, P., "Adaptive Transform Coding of Speech Signals", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol ASSP-25, No* 4, *August* 1977, *pp* 299-309.

[2.18] CAMPANELLA, S.J., ROBINSON, J.S., "A Comparison of Orthogonal Transformations for Digital Speech Processing", *IEEE Trans. on Comm. Tech., Vol COM-19, No* 6, *December* 1971, *pp* 1045-1049.

[2.19] SHUM, F., ELLIOT, A., BROWN, W., "Speech Processing with Walsh-Hadamard Transforms", *IEEE Trans. Audio and Electroacoustics, Vol AU-21*, 1973, *pp* 174-179.

[2.20] AHMED, N., RAO, K.R., "Orthogonal Transforms for Digital Signal Processing", *Springer-Verlag, Berlin-Heidelberg-New York*, 1975.

[2.21] AHMED, N., NATARAJAN, T., RAO, K.R., "Discrete Cosine Transform", *IEEE Trans. on Computers, Vol C-23, No* 1, *January* 1974, *pp* 90-93.

[2.22] HARMUTH, H.F., "Transmission of Information by Orthogonal Functions", *Springer-Verlag, Berlin-Heidelberg-New York*, 1972.

[2.23] FLANAGAN, J.L., "Speech Analysis, Synthesis and Perception", *2nd ed., Springer-Verlag, Berlin-Heidelberg-New York*, 1972.

[2.24] ELLIOTT, A.R., "Voice Processing in the Sequency Domain", *Proc. Symp. Appl.*

Walsh Func., Washington D.C., April 1973, *pp* 10-14.

[2.25] GOLD, B., "Digital Speech Networks", *Proc. IEEE, Vol* 65, *No* 12, *December* 1977, *pp* 1636-1658.

[2.26] FLANAGAN, J.L., SCHROEDER, M.R., ATAL, B.S., CROCHIERE, R.E., JAYANT, N.S., TRIBOLET, J.M., "Speech Coding", *IEEE Trans. on Comm., Vol COM*-27, *No* 4, *April* 1979, *pp* 710-736.

[2.27] RABINER, L.R., SCHAFER, R.W., "Digital Processing of Speech Signals", *Prentice-Hall, Englewood Cliffs, New Jersey*, 1978.

[2.28] JAYANT, N.S., NOLL, P., "Digital Coding of Waveforms", *Prentice-Hall, Englewood Cliffs, New Jersey*, 1984.

[2.29] FANT, G., "Acoustic Theory of Speech Production", *Mouton & Co., 's-Gravenhage*, 1960.

[2.30] WITTEN, I.H., "Principles of Computer Speech", *Academic Press, London*, 1982.

[2.31] CHIBA, T., KAJIYAMA, M., "The Vowel, Its Nature and Structure", *Tokyo Kaiseikan Publishing Co., Tokyo*, 1941.

[2.32] KELLY, J.L., Jr., LOCHBAUM, C., "Speech Synthesis", *Proc. 4th Int. Cong. Acoust., G*42, 1962, *pp* 1-4.

[2.33] 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 Barbera, California*, 1972.

[2.34] MORSE, P.M., INGARD, K.V., "Theoretical Acoustics", *McGraw-Hill Book Co., New York*, 1968.

[2.35] ATAL, B.S., HANAUER, S.L., "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", *J. Acoust. Soc. Amer., Vol* 50, *No* 2, *August* 1971, *pp* 637-655.

[2.36] BERANEK, L.L., "The Design of Communications Systems", *Proc. IRE, Vol* 35, *September* 1947, *pp* 880-890.

[2.37] ROSENBERG, A.E., "Effect of Glottal Pulse Shape on the Quality of Natural

Vowels", *J. Acoust. Soc. Amer.*, *Vol* 49, *No* 2, *February* 1971, *pp* 583-590.

[2.38] HOLMES, J.N., "The Influence of Glottal Waveform on the Naturalness of Speech from a Parallel Formant Synthesizer", *IEEE Trans. Audio and Electroacoustics*, *Vol AU*-21, *No* 3, *June* 1973, *pp* 298-305.

[2.39] MARKEL, J.D., GRAY, A.H., Jr., "Linear Prediction of Speech", *Springer-Verlag, Berlin-Heidelberg-New York*, 1976.

[2.40] RABINER, L.R., "Digital-Formant Synthesizer for Speech-Synthesis Studies", *J. Acoust. Soc. Amer.*, *Vol* 43, *No* 4, *April* 1968, *pp* 822-828.

[2.41] RYE, J.M., HOLMES, J.N., "A Versatile Software Parallel-Formant Speech Synthesizer", *JSRU Research Report No.* 1016, *November* 1982, *Joint Speech Research Unit, Cheltenham, Glos., UK.*

[2.42] HOLMES, J.N., "Speech Synthesis", *Mills and Boon, London*, 1972.

[2.43] HOLMES, J.N., "Formant Synthesizers: Cascade or Parallel", *JSRU Research Report No.* 1017, 1982, *Joint Speech Research Unit, Cheltenham, Glos., UK.*

[2.44] MUNSON, W.A., MONTGOMERY, H.C., "A Speech Analyzer and Synthesizer", *J. Acoust. Soc. Amer.*, *Vol* 22, *No* 5, *September* 1950, *p* 678.

[2.45] FANT, G., "Acoustic Analysis and Synthesis of Speech with Applications to Swedish", *Ericsson Technics*, *Vol* 1, *pp* 1-106.

[2.46] LAWRENCE, W., "The Synthesis of Speech from Signals which have a Low Information Rate", *in Communication Theory, Butterworth and Co. Ltd.*, 1953, *pp* 460-469.

[2.47] ROSENBERG, A.E., SCHAFER, R.W., RABINER, L.R., "Effects of Smoothing and Quantizing the Parameters of Formant-Coded Voiced Speech", *J. Acoust. Soc. Amer.*, *Vol* 50, *No* 6, *December* 1971, *pp* 1532-1538.

[2.48] COKER, C.H., "Real-Time Formant Vocoder, Using a Filter Bank, a General Purpose Digital Computer and an Analog Synthesizer", *J. Acoust. Soc. Amer.*, *Vol* 38, 1965, *p* 940.

[2.49] FANT, G., "On the Predictability of Formant Levels and Spectrum Envelopes from Formant Frequencies", *in For Roman Jakobson, Mouton & Co., 's-Gravenhage, pp* 109-120.

[2.50] OPPENHEIM, A.V., SCHAFER, R.W., "Homomorphic Analysis of Speech", *IEEE Trans. Audio and Electroacoustics, Vol AU-*16, *No* 2, *June* 1968, *pp* 221-226.

[2.51] OPPENHEIM, A.V., SCHAFER, R.W., STOCKHAM, T.G., Jr., "Nonlinear Filtering of Multiplied and Convolved Signals", *Proc. IEEE, Vol* 56, *No* 8, *August* 1968, *pp* 1264-1291.

[2.52] BOGERT, B., HEALY, M., TUKEY, J., "The Quefrency Analysis of Time Series for Echoes", *Proc. Symp. on Time Series Analysis, J. Wiley, New York*, 1963, *Chap.* 15, *pp.* 209-243.

[2.53] HESS, W., "Pitch Determination of Speech Signals", *Springer-Verlag, Berlin-Heidelberg-New York-Tokyo*, 1983.

[2.54] SCHAFER, R.W., "Echo Removal by Discrete Generalized Linear Filtering", *Technical Report* 466, *Research Lab. of Electronics, MIT, Cambridge, Mass., February*, 1969.

[2.55] OPPENHEIM, A.V., "Applications of Digital Signal Processing", *Prentice-Hall, Englewood Cliffs, New Jersey*, 1978.

[2.56] SCHAFER, R.W., RABINER, L.R., "System for Automatic Formant Analysis of Voiced Speech", *J. Acoust. Soc. Amer., Vol* 47, *No* 2, *February*, 1970, *pp* 634-648.

[2.57] OLIVE, J., "Automatic Formant Tracking in a Newton-Raphson Technique", *J. Acoust. Soc. Amer., Vol* 50, *August* 1971, *pp* 661-670.

[2.58] NOLL, A.M., "Cepstrum Pitch Determination", *J. Acoust. Soc. Amer., Vol* 41, *February* 1967, *pp* 293-309.

[2.59] OPPENHEIM, A.V., "Speech Analysis-Synthesis Systems Based on Homomorphic Filtering", *J. Acoust. Soc. Amer., Vol* 45, *No* 2, 1969, *pp* 458-465.

[2.60] ITAKURA, F., SAITO, S., "Analysis Synthesis Telephony Based on the Maximum

Likelihood Method", *Proc. 6th Intl. Congr. on Acoust., Tokyo, August* 1968, *Paper C-5-5, pp C17-20.*

[2.61] SAMBUR, M.R., ROSENBERG, A.E., RABINER, L.R., McGONEGAL, C.A., "On Reducing the Buzz in LPC Synthesis", *J. Acoust. Soc. Amer., Vol* 63, *No* 3, *March* 1978, *pp* 918-924.

[2.62] SAMBUR, M.R., "An Efficient Linear-Prediction Vocoder", *Bell Syst. Tech. Jnl., Vol* 54, *No* 10, *December* 1975, *pp* 1693-1723.

[2.63] UN, C.K., MAGILL, D.T., "The Residual-Excited Linear Prediction Vocoder with Transmission Rate below 9.6 kbits s^{-1}", *IEEE Trans. on Comm. Tech., Vol COM-23,* 1975, *pp* 1466-1473.

[2.64] ATAL, B.S., "High-Quality Speech at Low Bit Rates: Multi-Pulse and Stochastically Excited Linear Predictive Coders", *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Proc., Tokyo, Japan, April* 1986, *Vol* 3, *pp* 1681-1684.

[2.65] ATAL, B.S., REMDE, J.R., "A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates", *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Proc., Paris, France,* 1982, *pp* 614-617.

[2.66] KROON, P., ATAL, B.S., "Quantization Procedures for the Excitation in CELP Coders", *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Proc., Dallas, Texas, April* 1987, *Vol* 3, *pp* 1649-1652.

[2.67] SCHROEDER, M.R., ATAL, B.S, "Code-Excited Linear Prediction (CELP): High Quality Speech at Very Low Bit Rates", *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Proc.,* 1985, *pp* 937-940.

[2.68] BUZO, A., GRAY, A.H., Jr., GRAY, R.M., MARKEL, J.D., "Speech Coding based on Vector Quantisation", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol ASSP-28,* 1980, *pp* 562-574.

[2.69] ABUT, H., GRAY, R.M., REBOLLEDO, G., "Vector Quantisation of Speech and Speech-Like Waveforms", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol ASSP-*

30, 1982, *pp* 423-425.

[2.70] ANDERBERG, M.R., "Cluster Analysis for Applications", *Academic Press*, 1973.

[2.71] JUANG, B.H., WONG, D.Y., GRAY, A.H., Jr., "Distortion Performance of Vector Quantisation for LPC Voice Coding", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol ASSP*-30, 1982, *pp* 294-304.

[2.72] DUDLEY, H., "Remaking Speech", *J. Acoust. Soc. Amer., Vol* 11, *No* 2, *October* 1939, *pp* 169-177.

[2.73] DUDLEY, H., "The Vocoder", *Bell Lab. Record, Vol* 18, *No* 4, *December* 1939, *pp* 122-126.

[2.74] GOLD, B., RADER, C.M., "The Channel Vocoder", *IEEE Trans. Audio and Electroacoustics, Vol AU*-15, *No* 4, *December* 1967, *pp* 148-161.

[2.75] SCHROEDER, M.R., "Vocoders: Analysis and Synthesis of Speech", *Proc. IEEE, Vol* 54, *May* 1966, *pp* 720-734.

[2.76] GOLD, B., RADER, C.M., "Systems for Compressing the Bandwidth of Speech", *IEEE Trans. Audio and Electroacoustics, Vol AU*-15, *No* 3, *September* 1967, *pp* 131-135.

[2.77] SONDHI, M.M., SCHROETER, J., "A Hybrid Time-Frequency Domain Articulatory Speech Synthesizer", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol ASSP*-35, *No* 7, *July* 1987, *pp* 955-967.

[2.78] ISHIZAKA, K., FLANAGAN, J.L., "Synthesis of Voiced Sounds from a Two-Mass Model of the Vocal Cords", *Bell Syst. Tech. Jnl., Vol* 51, *No* 6, 1972, *pp* 1233-1268.

[2.79] KABASAWA, Y., ISHIZAKA, K., ARAI, Y., "Simplified Digital Model of the Lossy Vocal Tract and Vocal Cords", *Proc. 11th Intl. Congr. Acoust., Paris, France, Vol* 4, 1983, *pp* 175-178.

[2.80] RABINER, L.R., JUANG, B.H., LEVINSON, S.E., SONDHI, M.M., "Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixture Densities", *AT&T Tech. Jnl., Vol* 64, *No* 6, *July-August* 1985, *pp* 1211-1234.

[2.81] JUANG, B.H., RABINER, L.R., "Mixture Autoregressive Hidden Markov Models for Speech Signals", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol ASSP-33, No 6, December* 1985, *pp* 1404-1413.

[2.82] LEVINSON, S.E., RABINER, L.R., SONDHI, M.M., "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition", *Bell Syst. Tech. Jnl., Vol* 62, *No* 4, *April* 1983, *pp* 1035-1074.

[2.83] RABINER, L.R., LEVINSON, S.E., SONDHI, M.M., "On the Application of Vector Quantization and Hidden Markov Models to Speaker-Independent, Isolated Word Recognition", *Bell Syst. Tech. Jnl., Vol* 62, *No* 4, *April* 1983, *pp* 1075-1105.

[2.84] RABINER, L.R., JUANG, B.H., "An Introduction to Hidden Markov Models", *IEEE ASSP Magazine, January* 1986, *pp* 4-16.

[2.85] RABINER, L.R., JUANG, B.H., LEVINSON, S.E., SONDHI, M.M., "Some Properties of Continuous Hidden Markov Model Representations", *AT&T Tech. Jnl., Vol* 64, *No* 6, *July-August* 1985, *pp* 1251-1270.

[2.86] KOPEC, G.E., "Formant Tracking Using Hidden Markov Models and Vector Quantization", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol ASSP-34, No* 4, *August* 1986, *pp* 709-729.

[2.87] JELINEK, F. et al, "Experiments with the Tangora 20,000 Word Speech Recognizer", *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Proc., Dallas, Texas, April* 1987, *Vol* 2, *pp* 701-704.

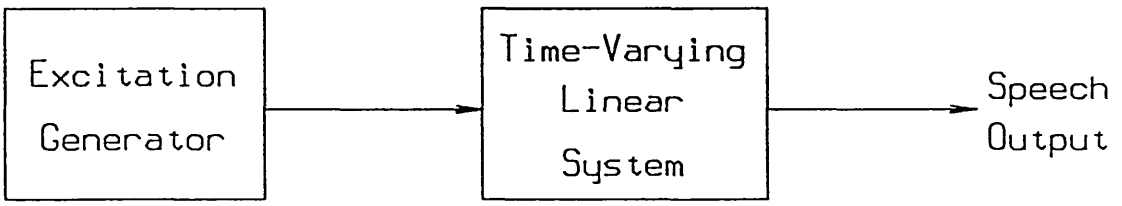


Figure 2.1 Source-Filter Model of Speech Production

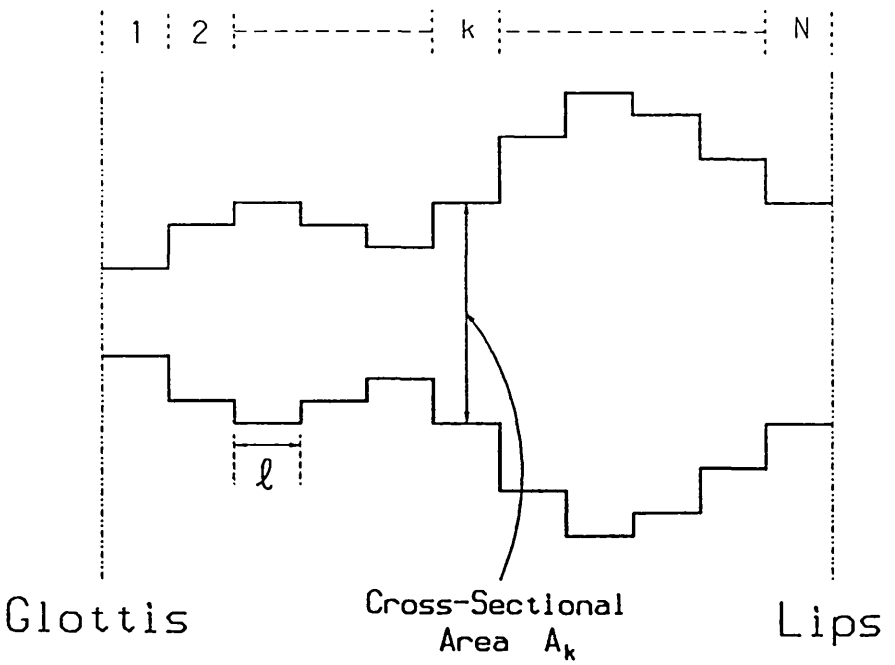


Figure 2.2 Acoustic Tube Model of the Vocal Tract

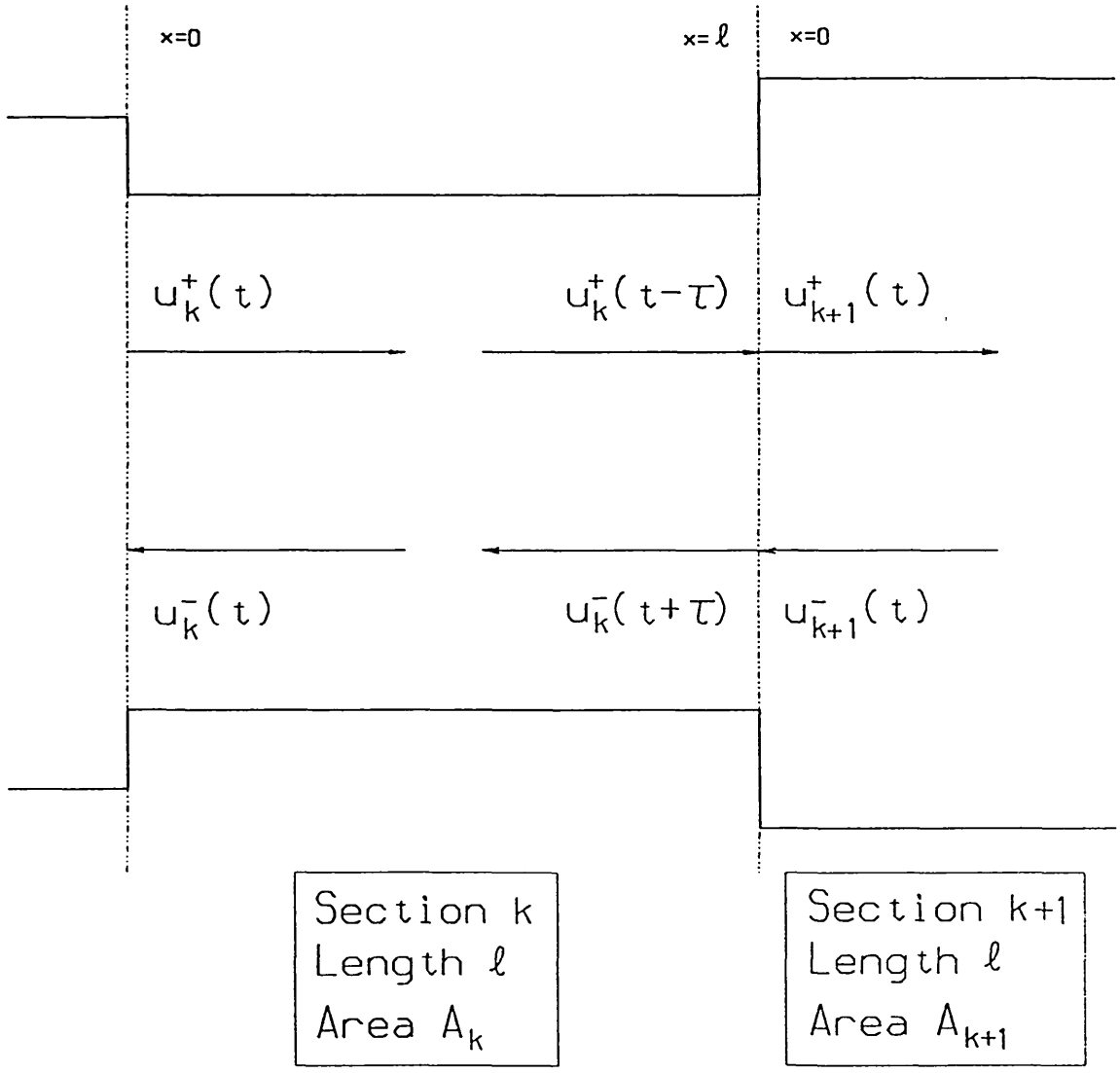


Figure 2.3 Continuity Conditions at the Junction
between Two Lossless Acoustic Tubes

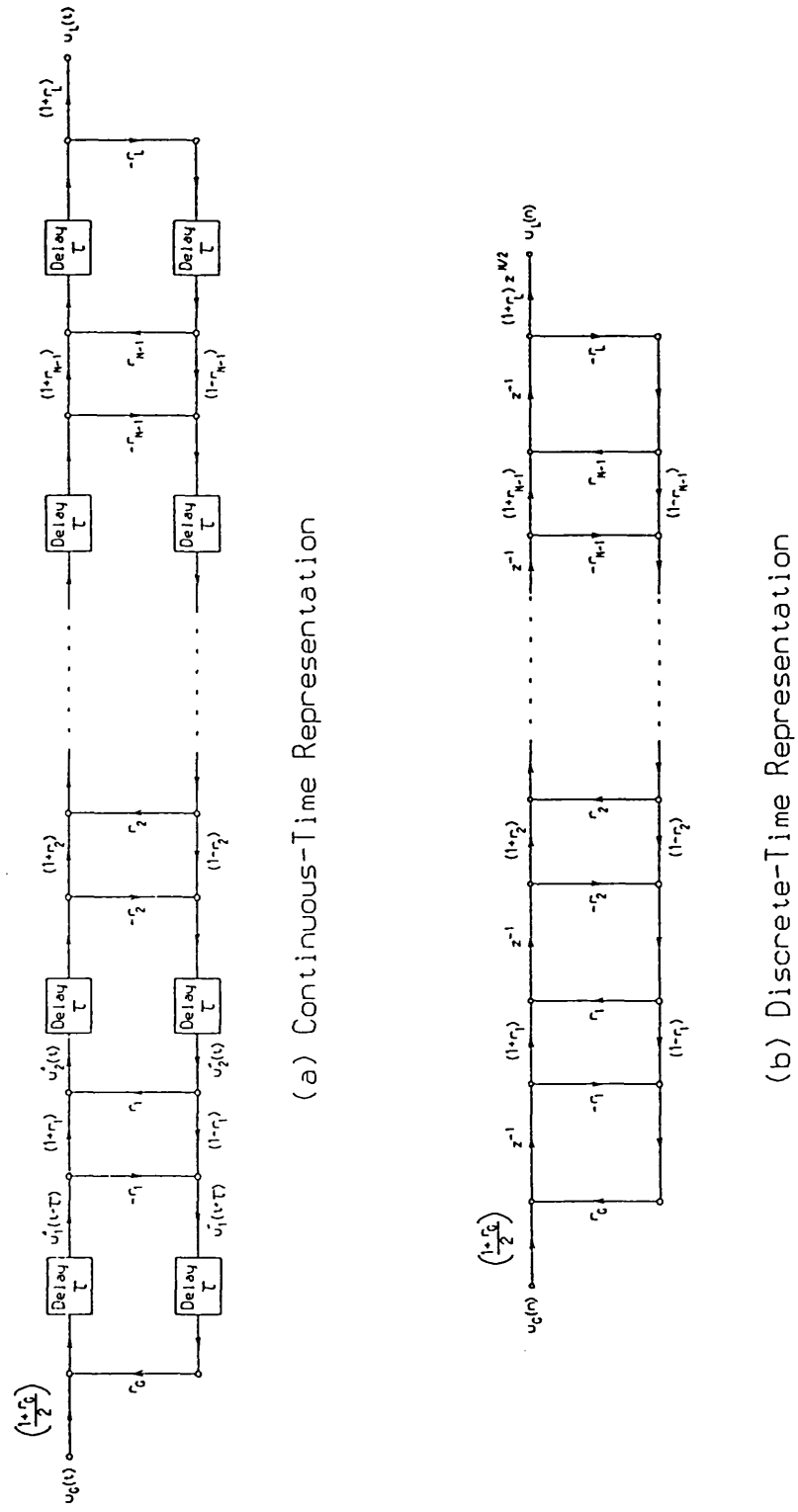


Figure 2.4 Signal Flow Graphs for a Lossless Tube Model of the Vocal Tract

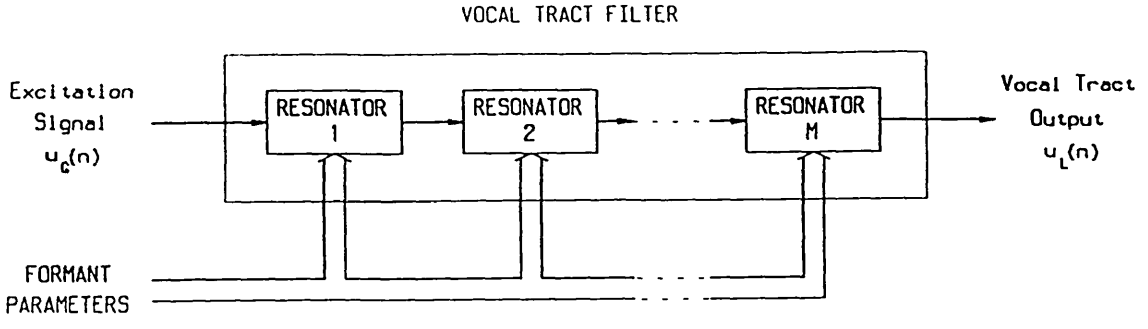


Figure 2.5 General Structure of a Cascade Formant Synthesizer

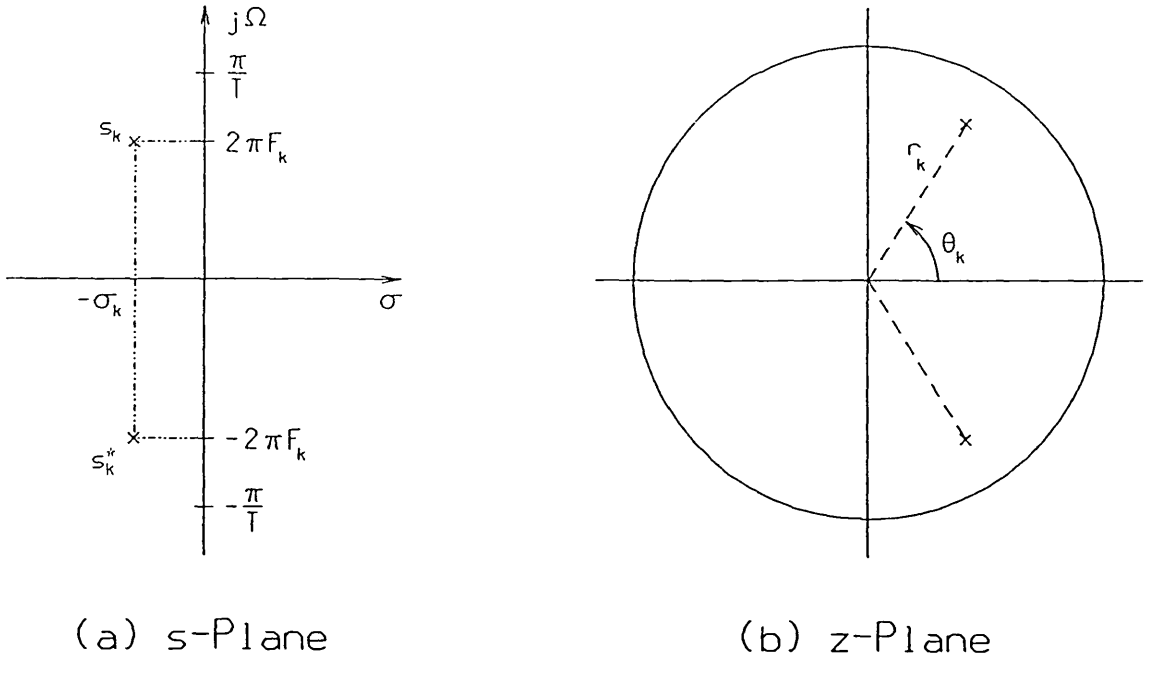


Figure 2.6 Representations of a Vocal Tract Resonance
in the s-plane and the z-plane

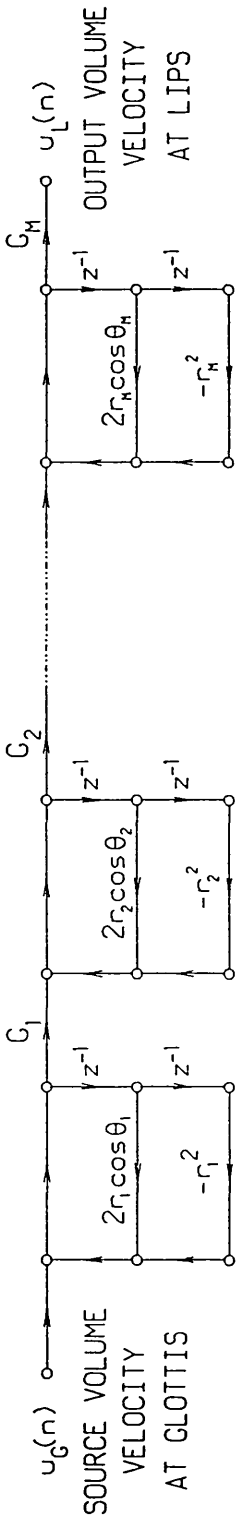


Figure 2.7 Cascade Implementation of an all-pole Transfer Function

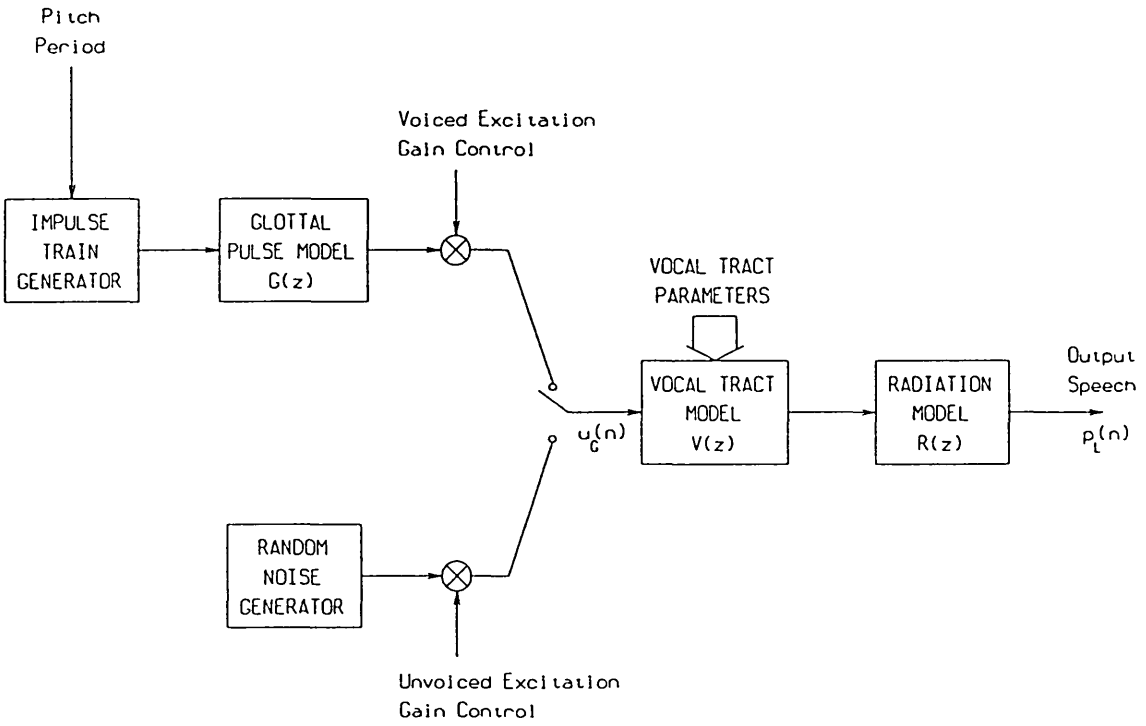


Figure 2.8 General Discrete-Time Model for Speech Production

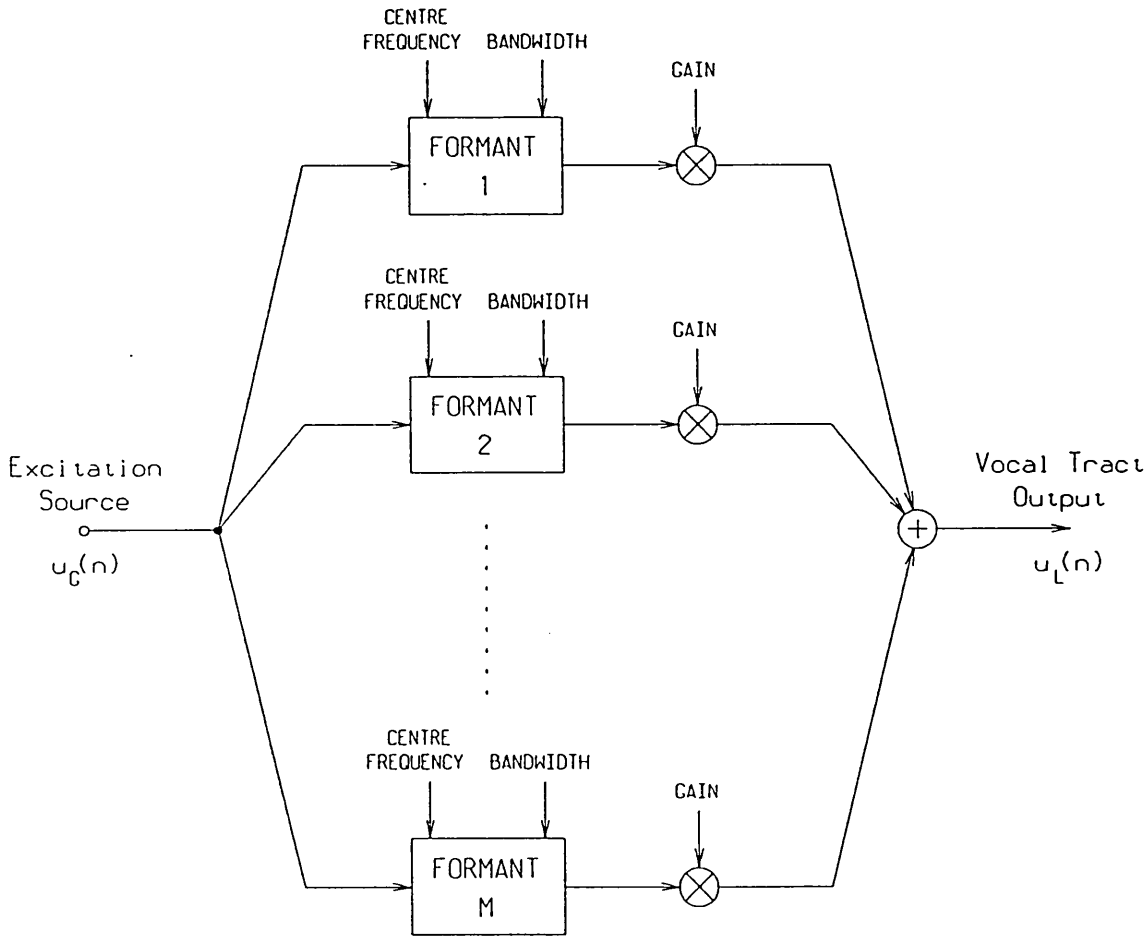


Figure 2.9 Parallel Formant Synthesiser Model of the Vocal Tract

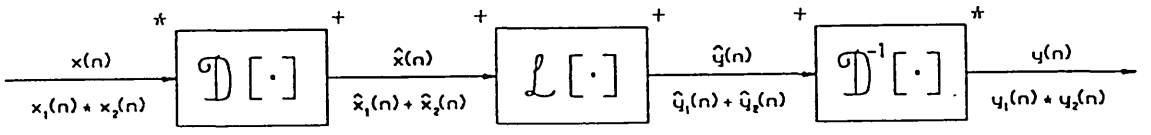


Figure 2.10 Canonic Form Representation of Homomorphic System for Convolution

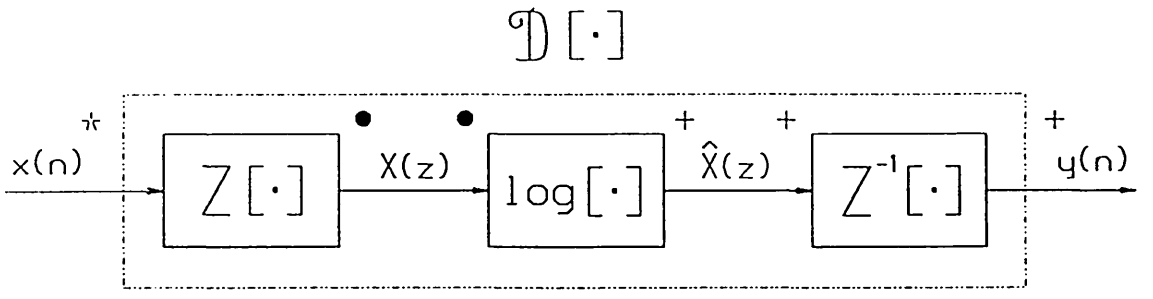


Figure 2.11 Representation of the Characteristic System for Homomorphic Deconvolution

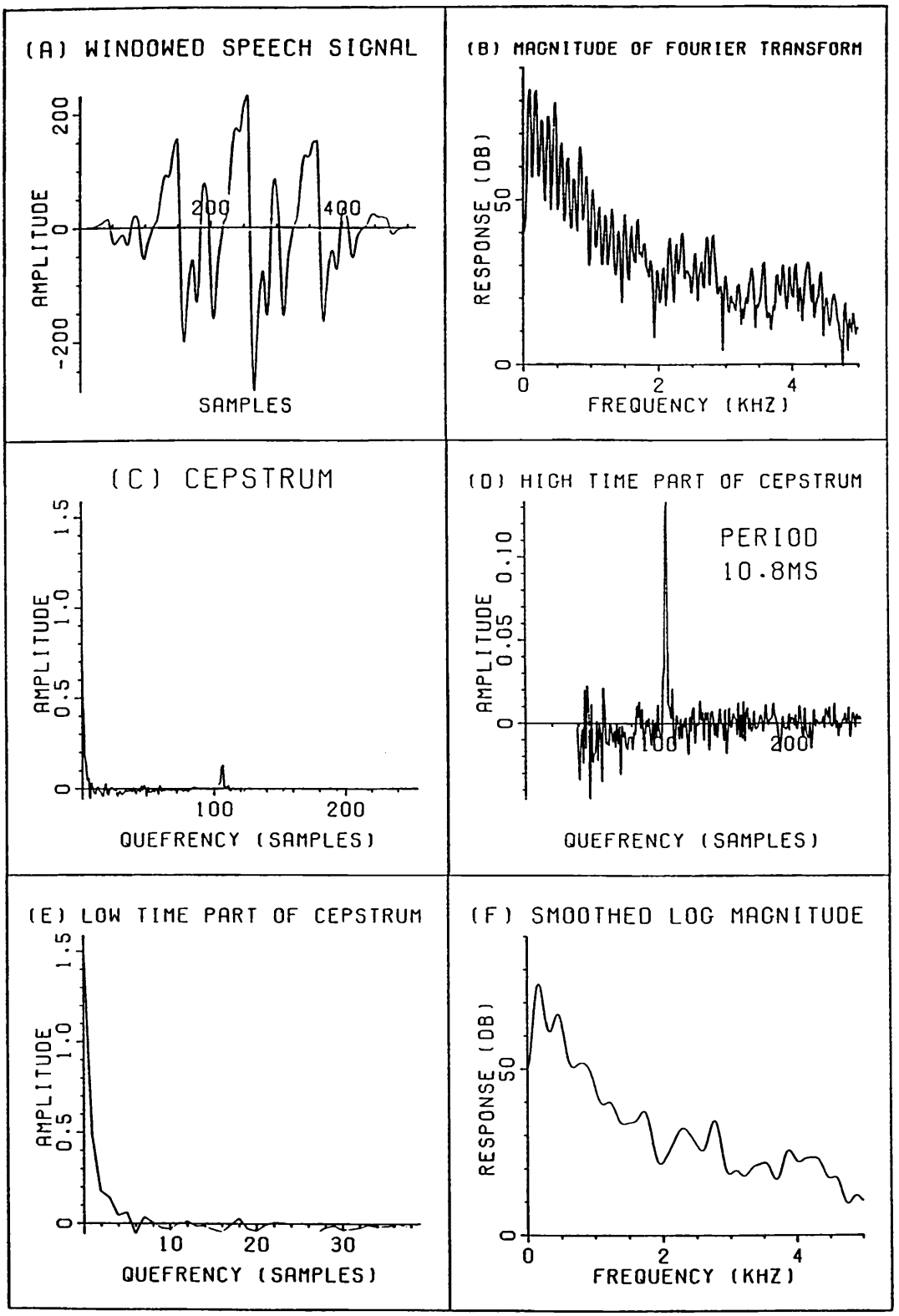


Figure 2.12 Homomorphic Analysis of Voiced Speech

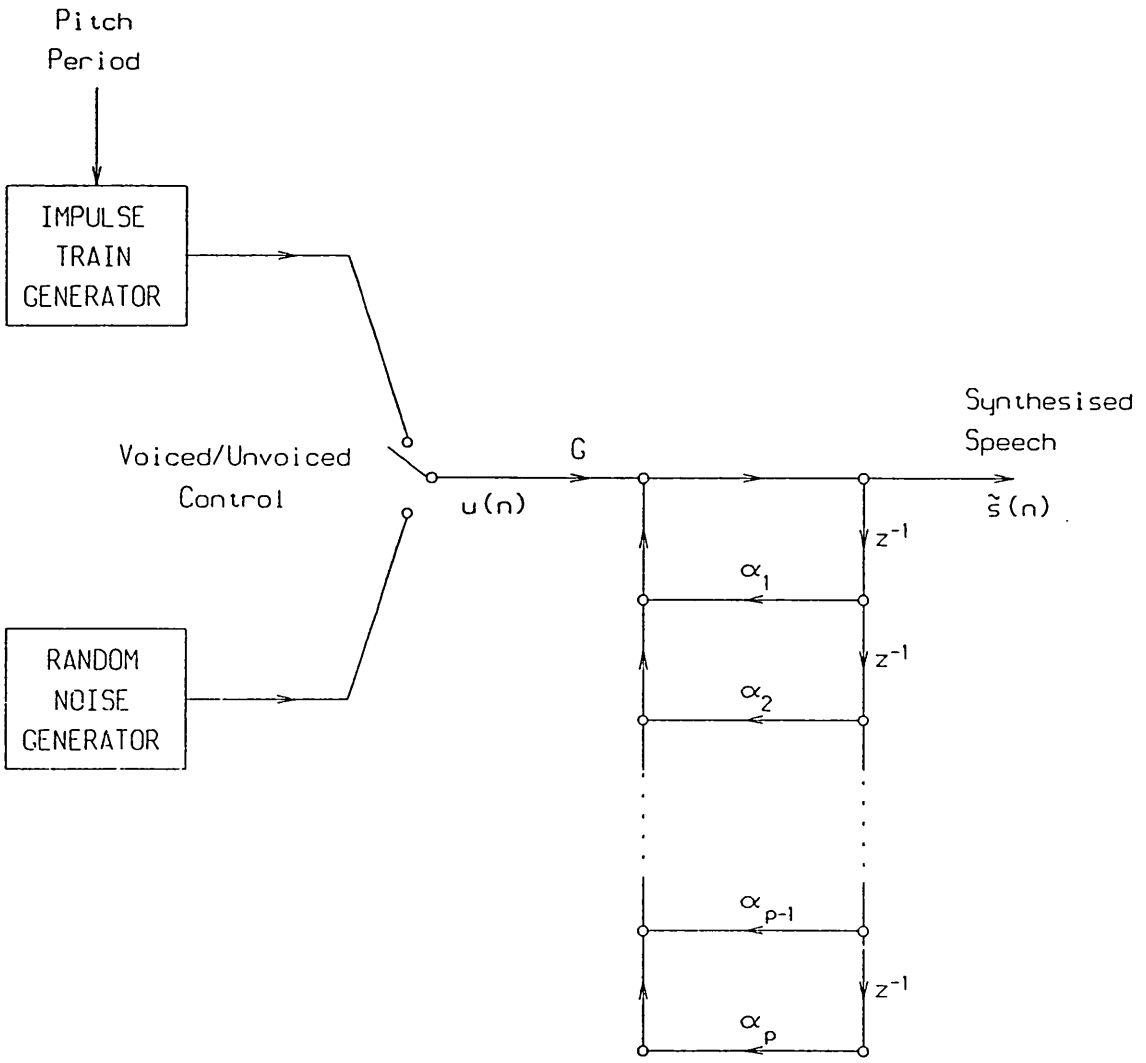


Figure 2.13 Block Diagram of a Linear Predictive Synthesiser

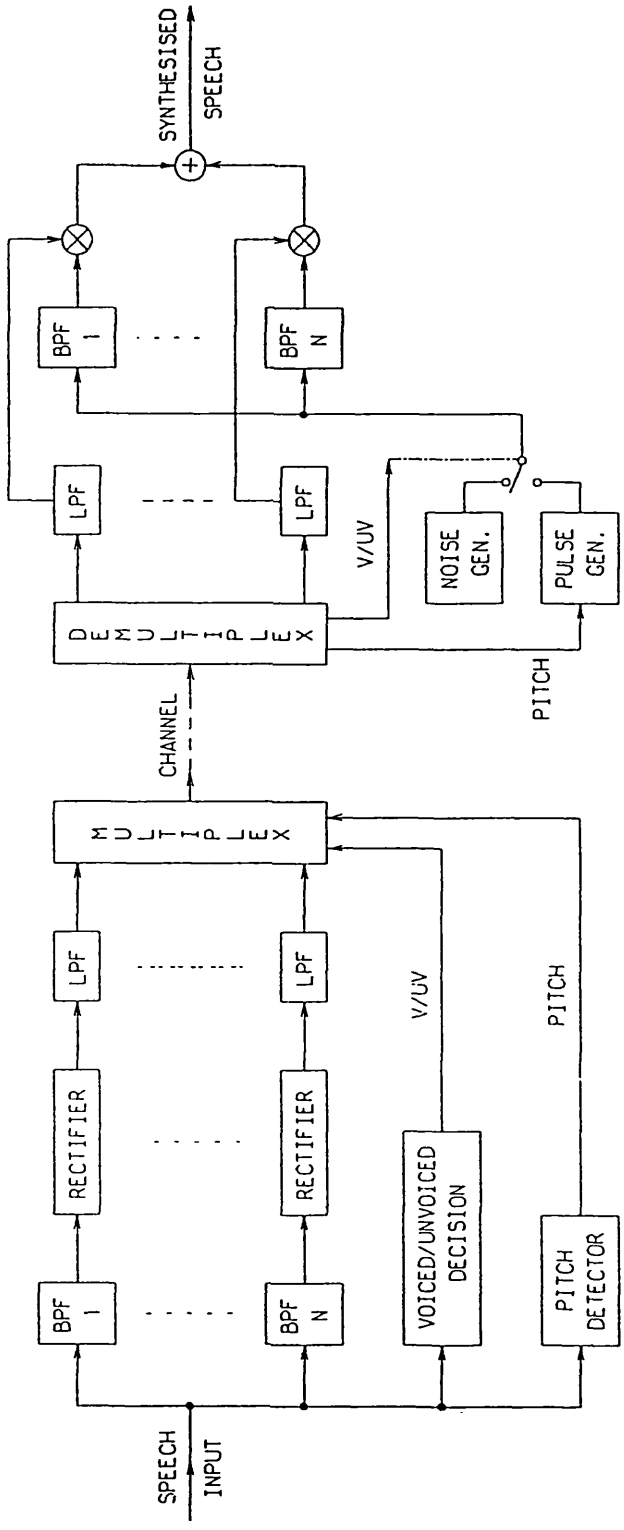


Figure 2.14 The Channel Vocoder

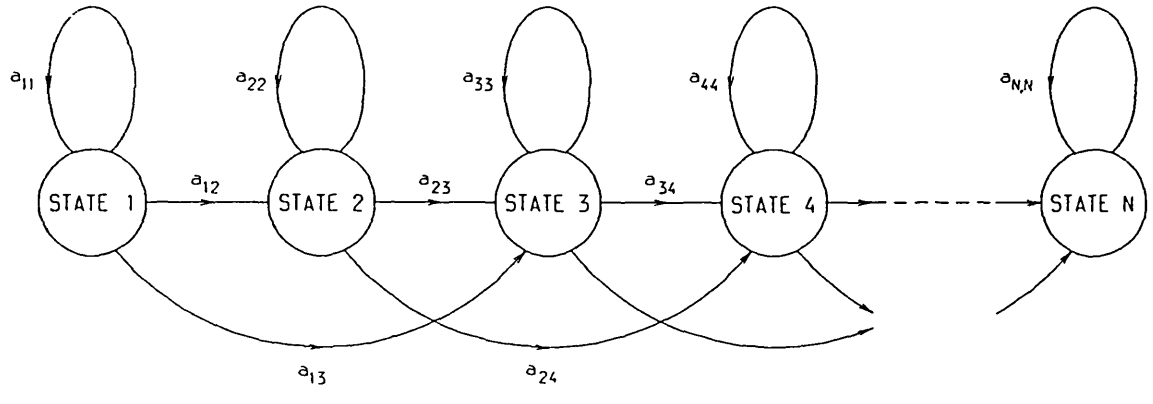


Figure 2.15 Structure of a Constrained Left-to-Right Markov Model

CHAPTER 3

LINEAR PREDICTIVE ANALYSIS OF SPEECH

CHAPTER 3. LINEAR PREDICTIVE ANALYSIS OF SPEECH

3.1 INTRODUCTION

Among the most ubiquitous forms of speech processing techniques are those based on linear prediction (LP). The scope of these methods covers most areas of speech processing – with applications ranging from acoustic tube modelling, speech synthesis, pitch & formant estimation to the low bit-rate transmission or storage of speech signals. The vocoding aspects of linear prediction have already been covered in section 2.6 and this chapter will attempt to draw attention to some of the other properties and uses of LP methods. The importance of these techniques lies both in their ability to provide accurate estimates of speech parameters and in their relative speed of computation.

Although linear least squares estimation (or prediction) was first postulated by Gauss in 1795 [3.1], the term "*linear prediction*" appears to date from its use by Wiener in 1949 [3.2]. The techniques and methods of linear prediction have also been in use for several years within the fields of control and information theory under the titles of system estimation and system identification. However, the first direct applications of LP techniques to speech analysis and synthesis were studied by Saito & Itakura [3.3] and Atal & Schroeder [3.4].

Since then, linear predictive analysis methods have appeared in a variety of guises and formulations for the modelling of the speech waveform. However, many of these formulations yield identical results and the only differences between some of the methods lie in the nature of the underlying assumptions or the manner in which the analysis problem is viewed [3.5]. In this chapter, a generalized discussion of LP methods is, therefore, followed by a description of the three basic forms of LP analysis to which all other formulations have been shown to be equivalent. This discussion focusses on both the solutions to the LP analysis problem offered by the three techniques and also their relative advantages and disadvantages.

As the importance of linear prediction lies not only in its relative speed of computation but also in its ability to provide accurate estimates of the basic speech parameters, this chapter also provides an insight into the various applications of LP analysis within the field of speech processing. In particular, the use of LP modelling in the spectral estimation of speech signals is considered in relation to other available techniques as this topic is of particular relevance to the work described in this thesis.

3.2 FUNDAMENTAL PRINCIPLES OF LINEAR PREDICTIVE ANALYSIS

As mentioned in section 2.6, a linear predictor of order p with predictor coefficients $\{\alpha_k : k=1,\ldots,p\}$ attempts to estimate a current signal sample, $s(n)$, by a linearly weighted sum of the p past samples. Let this approximation of $s(n)$ be $\tilde{s}(n)$ where

$$\tilde{s}(n) = \sum_{k=1}^{p} \alpha_k \, s(n-k) \tag{3.1}$$

The prediction error, $e(n)$, is then the error between the actual value and the predicted estimate such that

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k \, s(n-k) \tag{3.2}$$

The method of least squares used in LP analysis then attempts to minimise the total squared error over the analysis frame in order to determine the optimal set of predictor coefficients. Therefore, the total squared error, E, is given by:

$$E = \sum_{n} e^2(n) \tag{3.3}$$

$$= \sum_{n} \left[s(n) - \sum_{k=1}^{p} \alpha_k \, s(n-k) \right]^2 \tag{3.4}$$

The range of summation in this expression is temporarily left unspecified. The values of α_k which minimise E can be found by setting

$$\frac{\partial E}{\partial \alpha_i} = 0, \qquad 1 \leq i \leq p \tag{3.5}$$

From equations (3.4),(3.5), the following set of equations is obtained:

$$\sum_{k=1}^{p} \alpha_k \sum_n s(n\text{-}i).s(n\text{-}k) = \sum_n s(n\text{-}i).s(n) , \qquad 1 \leq i \leq p \tag{3.6}$$

Now, let $$\phi(i,k) = \sum_n s(n\text{-}i).s(n\text{-}k) \tag{3.7}$$

Then, $$\sum_{k=1}^{p} \alpha_k \phi(i,k) = \phi(i,0) , \qquad 1 \leq i \leq p \tag{3.8}$$

This set of equations is known in least squares terminology as the *Normal Equations.* In addition, from equation (3.4), the minimal squared error can be shown to be

$$E = \sum_n s^2(n) - \sum_{k=1}^{p} \alpha_k \sum_n s(n).s(n\text{-}k) \tag{3.9}$$

Thus, using (3.7),

$$E = \phi(0,0) - \sum_{k=1}^{p} \alpha_k \phi(0,k) \tag{3.10}$$

Therefore, to solve for the optimum predictor coefficients, the values $\{\phi(i,k) : 1 \leq i \leq p, 0 \leq k \leq p\}$ are first computed and then equation (3.8) can be solved to yield the α_k values. However, the form of the equations and the method of solution depends on the limits of summation over n in the above equations. These limits are, in fact, very significant since two

of the basic methods for LP analysis emerge from two different approaches employed to deal with both the summation range and the nature of the waveform segment $\{s(n)\}$. These two techniques are developed in detail in the next section.

3.3 FORMULATIONS OF THE LP MODEL

3.3.1 The Autocorrelation Method

In this case, the total squared error is minimised over the infinite duration $-\infty < n < \infty$. Equations (3.8),(3.10) then become

$$\sum_{k=1}^{p} \alpha_k R(i\text{-}k) = R(i) , \qquad 1 \leq i \leq p \tag{3.11}$$

and $$E = R(0) - \sum_{k=1}^{p} \alpha_k R(k) \tag{3.12}$$

where $$R(k) = \sum_{n=-\infty}^{\infty} s(n).s(n+k) \tag{3.13}$$

represents the autocorrelation function of the signal $s(n)$. However, in practice, the signal of interest will only be known over the finite interval $0 \leq n \leq N\text{-}1$. It is, therefore, necessary to apply a finite length window over the speech segment such that the signal is identically zero outside the required interval $0 \leq n \leq N\text{-}1$. This requirement implies that the autocorrelation function is then given by:

where $$R(k) = \sum_{n=0}^{N\text{-}1\text{-}k} s(n).s(n+k) \tag{3.14}$$

It is also important to note that $R(k)$ is an even function of k such that:

$$R(\text{-}k) = R(k) \tag{3.15}$$

The set of equations described by equation (3.11) can thus be expressed in matrix form as follows:

$$\begin{bmatrix} R(0) & R(1) & R(2) & \cdots & R(p\text{-}1) \\ R(1) & R(0) & R(1) & \cdots & R(p\text{-}2) \\ R(2) & R(1) & R(0) & \cdots & R(p\text{-}3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(p\text{-}1) & R(p\text{-}2) & R(p\text{-}3) & \cdots & R(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ R(3) \\ \vdots \\ R(p) \end{bmatrix} \tag{3.16}$$

The $(p \times p)$ matrix of autocorrelation values is a symmetric Toeplitz matrix, i.e. all the elements along each diagonal are equal. This particular property can be exploited to obtain an efficient algorithm for the calculation of the predictor coefficients $\{\alpha_k : k=1,\ldots,p\}$. However, a means for calculating the other LP model parameter — the model gain — must first be established.

3.3.1.1 GAIN COMPUTATION

In section 2.6, the output of a LP process was viewed as a linear system with an input $u(n)$ and output $s(n)$ such that

$$s(n) = \sum_{k=1}^{p} \alpha_k \, s(n\text{-}k) \; + \; Gu(n) \tag{3.17}$$

However, in section 3.2, the prediction error signal was defined as

$$e(n) = s(n) - \sum_{k=1}^{p} \alpha_k \, s(n\text{-}k) \tag{3.18}$$

From these two expressions it can be seen that, if the signal $s(n)$ conforms exactly to the LP model, then

$$e(n) = Gu(n) \tag{3.19}$$

However, since equation (3.19) is only valid if the signal is completely described by the LP system, it is not generally possible to solve for the model gain, G, directly from the error signal. Instead, the more reasonable assumption is made that the total energy in the excitation input is to be equal to the total energy of the prediction error, i.e.,

$$G^2 \sum_{n=0}^{N-1} u^2(n) = \sum_{n=0}^{N-1} e^2(n) = E \tag{3.20}$$

The two types of input of special interest will be considered separately.

(a) Voiced Speech

The excitation in this case is assumed to be impulsive[1] such that

$$u(n) = \delta(n) \tag{3.21}$$

The impulse response, $h(n)$, of the LP model is then given by

$$h(n) = \sum_{k=1}^{p} \alpha_k \, h(n-k) + G\delta(n) \tag{3.22}$$

It can then be shown [3.6] that the autocorrelation function of $h(n)$ defined as

---

[1] *For this assumption to be valid, the effects of glottal excitation are combined with those of the vocal tract and thus both components are modelled by the linear predictor.*

$$\tilde{R}(k) = \sum_{n=-\infty}^{\infty} h(n).h(n+k) \tag{3.23}$$

satisfies the relations

$$\tilde{R}(i) = \sum_{k=1}^{p} \alpha_k \tilde{R}(i-k) , \qquad 1 \leq i \leq p \tag{3.24}$$

& $$\tilde{R}(0) = \sum_{k=1}^{p} \alpha_k \tilde{R}(k) + G^2 \tag{3.25}$$

Using these results, equating the total energies in the signals $s(n)$ & $h(n)$ yields

$$G^2 = R(0) - \sum_{k=1}^{p} \alpha_k R(k) = E \tag{3.26}$$

This expression enables the model gain to be calculated directly from the total squared prediction error. Furthermore, equations (3.24),(3.25) are now identical to those for $R(i)$ & $R(0)$ in section 3.3.1 and, therefore, it follows that

$$\tilde{R}(i) = R(i) , \qquad 0 \leq i \leq p \tag{3.27}$$

The first $(p+1)$ coefficients of the autocorrelation function of the LP model impulse response are, therefore, identical to the first $(p+1)$ coefficients of the signal autocorrelation function.

(b) Unvoiced Speech

The input is now considered to be white noise with zero mean and unity variance,

i.e. $$\mathcal{E}\left[u(n)\right] = 0 ; \qquad \mathcal{E}\left[u(n).u(n-k)\right] = \delta(k) \tag{3.28}$$

where $\mathcal{E}[\cdot]$ denotes the expected value. For this form of input, let the output be $g(n)$ such that

$$g(n) = \sum_{k=1}^{p} \alpha_k \, g(n-k) + Gu(n) \tag{3.29}$$

The autocorrelation function $\tilde{R}(k)$ of $g(n)$ is then expressed by

$$\tilde{R}(k) = \mathcal{E}\left[g(n).g(n-k)\right] \tag{3.30}$$

and can be shown [3.7] to obey exactly the same relationships of equations (3.24),(3.25),(3.26),(3.27) above. Equation (3.26) can thus be employed to evaluate the model gain for both types of model input.

3.3.1.2 THE LEVINSON-DURBIN RECURSION

The equations necessary to solve for the coefficients of an LP model using the autocorrelation method are now available and are given by:

$$\sum_{k=1}^{p} \alpha_k R(i-k) = R(i) , \qquad 1 \leq i \leq p \tag{3.31}$$

$$\& \qquad G^2 = E = R(0) - \sum_{k=1}^{p} \alpha_k R(k) \tag{3.32}$$

As mentioned earlier, the Toeplitz nature of the autocorrelation matrix in equation (3.31) can be exploited and several recursive procedures have been devised for solving this system of equations. One of the most efficient algorithms is based on Levinson's work on predicting stationary time series from finite sets of past data [3.8]. A later modification by Durbin [3.9] established the algorithm in its present form which is twice as fast as the original algorithm [3.10]. Hence, the resulting recursive procedure is known as the *Levinson-Durbin*

solution for the autocorrelation equations. This algorithm can be specified as follows:

THE LEVINSON-DURBIN ALGORITHM

$$E^{(0)} = R(0) \tag{3.33}$$

$$k_i = \frac{\left\{ R(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} R(i-j) \right\}}{E^{(i-1)}} \qquad i = 1,\ldots,p \tag{3.34a}$$

$$\alpha_i^{(i)} = k_i \tag{3.34b}$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)}, \qquad j = 1,\ldots,i-1 \tag{3.34c}$$

$$E^{(i)} = \left(1 - k_i^2\right) E^{(i-1)} \tag{3.34d}$$

Equations (3.34a) - (3.34d) are, therefore, solved recursively for $i = 1,\ldots,p$ and the final solution is given by

$$\alpha_j = \alpha_j^{(p)}, \qquad 1 \le i \le p \tag{3.35a}$$

$$G^2 = E^{(p)} \tag{3.35b}$$

A complete derivation of this algorithm is covered in Appendix A. It should be noted that in the process of calculating the predictor coefficients for a predictor of order p, the solutions for all predictors of order less than p are also obtained. Another important point concerns the values obtained for the parameters $\{k_i\}$. As the error, $E^{(i)}$, can never be negative (since it is a squared error), it follows from equation (3.34d) that the quantities $\{k_i\}$ are such

that they lie in the interval:

$$-1 \leq k_i \leq 1 \ , \qquad 1 \leq i \leq p \tag{3.36}$$

or, $$0 \leq 1 - k_i^{\ 2} \leq 1 \ , \quad 1 \leq i \leq p \tag{3.37}$$

Substituting this result into equation (3.34d) yields

$$0 \leq E^{(i)} \leq E^{(i-1)} \tag{3.38}$$

The minimum total error $E^{(i)}$ is, therefore, a monotonically decreasing function of the order i of the predictor. The condition on the parameters $\{k_i\}$ represented by equation (3.36) is also important since it can be shown [3.5] that this provides a necessary and sufficient condition for all the roots of the predictor polynomial

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k \, z^{-k} \tag{3.39}$$

to lie within the unit circle — thereby guaranteeing the stability of an LP system derived using the autocorrelation method. It is also interesting to note that an approximate maximum likelihood approach for calculating the predictor parameters suggested by Itakura & Saito [3.11] results in a set of equations identical to those for the autocorrelation method. This maximum likelihood method is one of the earliest formulations of linear prediction and relies on the assumptions that $N \gg p$ (i.e. the number of data points greatly exceeds the predictor order) and that the input speech can be modelled as a stationary Gaussian random process.

Another technique known as the *correlation matching* method [3.12] attempts to obtain a match between the autocorrelations of the input sequence $\{s(n)\}$ and the impulse response of the LP synthesis filter $G/A(z)$. As described in section 3.3.1.1, this characteristic is

also a property of the autocorrelation method and, therefore, the correlation matching approach is also equivalent to the autocorrelation approach.

These observations illustrate the point that several contrasting techniques with different underlying assumptions can lead to an identical set of autocorrelation equations.

3.3.2 The Covariance Method

The fundamental difference between this technique and the autocorrelation method lies in the assumption that the total squared prediction error E is to be minimised over a finite interval, $0 \leq n \leq N-1$, i.e.

$$E = \sum_{n=0}^{N-1} e^2(n) \tag{3.40}$$

The definition of $\phi(i,k)$ in section 3.2 then becomes

$$\phi(i,k) = \sum_{n=0}^{N-1} s(n-i).s(n-k) \quad , \qquad \begin{matrix} 1 \leq i \leq p \\ 0 \leq k \leq p \end{matrix} \tag{3.41}$$

It can be seen from this expression that an evaluation of all the necessary values of $\phi(i,k)$ requires the use of the segment of speech in the interval $-p \leq n \leq N-1$ — a total of $p+N$ samples. Therefore, unlike the autocorrelation method, the use of a window which tapers the speech segment to zero at the limits of the interval $0 \leq n \leq N-1$ is not appropriate since some of the necessary values lie outside this particular range. As described in section 3.2, the set of equations to be solved in order to obtain the predictor parameters is given by

$$\sum_{k=1}^{p} \alpha_k \phi(i,k) = \phi(i,0) \, , \qquad 1 \leq i \leq p \tag{3.42a}$$

$$E = \phi(0,0) - \sum_{k=1}^{p} \alpha_k \phi(0,k) \tag{3.42b}$$

In matrix form, equation (3.42a) becomes

$$\begin{bmatrix} \phi(1,1) & \phi(1,2) & \cdots & \phi(1,p) \\ \phi(2,1) & \phi(2,2) & \cdots & \phi(2,p) \\ \vdots & \vdots & \ddots & \vdots \\ \phi(p,1) & \phi(p,2) & \cdots & \phi(p,p) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} \phi(1,0) \\ \phi(2,0) \\ \vdots \\ \phi(p,0) \end{bmatrix} \tag{3.43}$$

As the matrix of values $\{\phi(i,k) : 1 \leq i \leq p,\ 1 \leq k \leq p\}$ has the properties of a covariance matrix [3.12], this method is known as the "*Covariance Method*". From the definition of $\phi(i,k)$, it follows that:

$$\phi(i,k) = \phi(k,i) \tag{3.44}$$

The covariance matrix is, therefore, symmetric but <u>not</u> Toeplitz. In fact, the diagonal terms are related by the expression:[2]

$$\phi(i+1,k+1) = \phi(i,k) + s(-i-1).s(-k-1) - s(N-1-i).s(N-1-k) \tag{3.45}$$

Because of this property, the method of solution employed in the autocorrelation technique can not be applied in this case. Instead, a reasonably efficient approach is offered by the use of the "*Cholesky Decomposition*" procedure [3.6]. This technique uses an LDU decomposition of the covariance matrix (i.e. a factorisation into a product of a lower triangular, a diagonal and an upper triangular matrix). From this transformed result, it is

---

[2] *This expression may be used to greatly reduce the computation required in order to calculate the correlation matrix.*

relatively simple to solve for the linear prediction coefficients.

An important point to note is that the stability of the predictor polynomial, $A(z)$, is not guaranteed for the covariance method [3.13]. However, as the results of the covariance technique approach those produced by the autocorrelation method for long analysis frames (i.e. $N \gg p$), the LP filter will, in practice, almost always be stable. As for the autocorrelation method, several different formulations of the LP model tend to produce an identical system of equations. Therefore, the covariance equations described above also arise from the procedure known as *Prony's Method* [3.14] — which attempts to represent the signal by the summation of a set of complex exponentials.

3.3.3 Lattice Formulations and Solutions

In spite of the fact that efficient strategies exist for the solution of the normal equations for the covariance and autocorrelation methods, the computation of a matrix of correlation values is still necessary for both techniques. This calculation can, in fact, form the major part of the computational load for small values of the predictor order. However, an alternative class of systems, known as *Lattice Methods*, offers a fundamentally different approach for the calculation of the predictor parameters.

Although lattice methods do not, in general, result in a system of equations similar to those for the autocorrelation and covariance techniques, it is instructive to begin with the particular lattice formulation which is identical to the autocorrelation method. From the Levinson-Durbin algorithm described in section 3.3.1.2, the set of predictor coefficients produced at the i-th stage of the procedure $\{\alpha_k^{(i)} : j=1,\ldots,i\}$ corresponds to the i-th order predictor polynomial $A^{(i)}(z)$ as follows:

$$A^{(i)}(z) = 1 - \sum_{k=1}^{i} \alpha_k^{(i)} z^{-k} \tag{3.46}$$

For the input speech signal $\{s(n)\}$, the output prediction error $\{f^{(i)}(n)\}$ would, therefore, be of the form

$$f^{(i)}(n) = s(n) - \sum_{k=1}^{i} \alpha_k^{(i)} s(n-k) \tag{3.47}$$

or, in terms of z-transforms,

$$F^{(i)}(z) = A^{(i)}(z) . S(z) \tag{3.48}$$

However, in the Levinson-Durbin algorithm, the predictor coefficients at the i-th stage are related to those at the $(i$-1)th stage by the expressions:

$$\alpha_i^{(i)} = k_i \tag{3.49a}$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i . \alpha_{i-j}^{(i-1)}, \qquad j = 1,\ldots,i-1 \tag{3.49b}$$

Substituting equations (3.49a),(3.49b) into equation (3.46) yields the following recurrence formula for $A^{(i)}(z)$:

$$A^{(i)}(z) = A^{(i-1)}(z) - k_i z^{-i} A^{(i-1)}(z^{-1}) \tag{3.50}$$

This result together with equation (3.48) produces

$$F^{(i)}(z) = A^{(i-1)}(z) . S(z) - k_i z^{-i} A^{(i-1)}(z^{-1}) . S(z) \tag{3.51}$$

The first term in equation (3.51) is obviously $F^{(i-1)}(z)$ and if we let

$$B^{(i)}(z) = z^{-i} A^{(i)}(z^{-1}) . S(z) \tag{3.52}$$

Then, $$F^{(i)}(z) = F^{(i-1)}(z) - k_i z^{-1} B^{(i-1)}(z) \tag{3.53}$$

A physical interpretation of $B^{(i)}(z)$ arises from an inverse z-transform of equation (3.52),

$$b^{(i)}(n) = s(n-i) - \sum_{k=1}^{i} \alpha_k^{(i)} s(n+k-i) \tag{3.54}$$

This equation implies the prediction of the value $s(n-i)$ from the i input signal samples $\{s(n+k-i) : k=1,\ldots,i\}$ which follow $s(n-i)$. The sequence $\{b^{(i)}(n)\}$ is, therefore, known as the "*backward prediction error sequence*" or *backward residual.* In addition, substituting equation (3.50) into equation (3.52) produces

$$B^{(i)}(z) = z^{-i} A^{(i-1)}(z^{-1}) . S(z) - k_i A^{(i-1)}(z) . S(z) \tag{3.55}$$

$$= z^{-1} B^{(i-1)}(z) - k_i F^{(i-1)}(z) \tag{3.56}$$

Therefore, the i-th stage backward prediction error is given by

$$b^{(i)}(n) = b^{(i-1)}(n-1) - k_i f^{(i-1)}(n) \tag{3.57}$$

A similar recurrence relation can be established for the forward prediction error using equation (3.53),

$$f^{(i)}(n) = f^{(i-1)}(n) - k_i b^{(i-1)}(n-1) \tag{3.58}$$

Equations (3.57) and (3.58) provide a recursion for both the forward and backward residuals for an i-th order predictor in terms of the corresponding errors for an (i-1)th order predictor. An initial condition results from considering a zeroth order predictor (i.e. no predictor) such that

$$f^{(0)}(n) \;=\; b^{(0)}(n) \;=\; s(n) \tag{3.59}$$

A graphical description of this set of equations is illustrated by the lattice network shown in Fig. 3.1. It is this structure which provides lattice methods with their title and it should be noted that the forward error at the p-th stage, $f^{(p)}(n)$, yields the linear prediction error, $e(n)$, discussed in relation to the autocorrelation method. It has also been shown [3.5] that the coefficients $\{k_i : i=1,\ldots,p\}$ can be computed directly from the forward & backward prediction errors such that

$$k_i \;=\; \frac{\sum_{n=0}^{N-1} f^{(i-1)}(n).\, b^{(i-1)}(n-1)}{\sqrt{\left[\sum_{n=0}^{N-1}\left(f^{(i-1)}(n)\right)^2\right]\left[\sum_{n=0}^{N-1}\left(b^{(i-1)}(n-1)\right)^2\right]}} \tag{3.60}$$

$$= \frac{\mathcal{E}\left[f^{(i-1)}(n).\, b^{(i-1)}(n-1)\right]}{\sqrt{\mathcal{E}\left[\left(f^{(i-1)}(n)\right)^2\right].\,\mathcal{E}\left[\left(b^{(i-1)}(n-1)\right)^2\right]}}\;, \qquad 1 \le i \le p \tag{3.61}$$

The use of equation (3.60) in the Levinson-Durbin algorithm will yield identical predictor coefficients to those for the autocorrelation method. Therefore, this technique provides an alternative means of solving the normal equations without resorting to the calculation of a correlation matrix. As an expression of the form of equation (3.60) is

representative of the degree of correlation between the forward & backward prediction errors, the parameters $\{k_i\}$ are also known as *"partial correlation"* or *PARCOR* coefficients. Although this lattice technique does not correspond to the minimisation of some error criterion based on $e^{(i)}(n)$, $b^{(i)}(n)$, another important lattice formulation due to Burg [3.15] attempts to minimise the sum $\left(E^{(i)}\right)$ of the squared forward & backward residuals such that

$$E^{(i)} = \sum_{n=0}^{N-1} \left[\left(f^{(i)}(n)\right)^2 + \left(b^{(i)}(n)\right)^2 \right] \tag{3.62}$$

This requirement yields the following expression for the parameters $\{k_i\}$:

$$k_i = \frac{2\, \mathcal{E}\left[f^{(i-1)}(n) \cdot b^{(i-1)}(n-1)\right]}{\mathcal{E}\left[\left(f^{(i-1)}(n)\right)^2\right] + \mathcal{E}\left[\left(b^{(i-1)}(n-1)\right)^2\right]}, \qquad 1 \le i \le p \tag{3.63}$$

Although these values lie in the same range as for the autocorrelation method, i.e.

$$-1 \le k_i \le 1 \tag{3.64}$$

the parameters $\{k_i\}$ estimated using equation (3.63) will, in general, differ from those calculated using the autocorrelation method. Even though lattice methods do not involve the minimisation of any global error criterion (as do the covariance & autocorrelation methods), a stage-by-stage minimisation may take place as in Burg's method. However, this form of minimisation will generally lead to a sub-optimal solution with respect to the least-squares error minimum and will only tend to an optimal solution as the signal becomes stationary [3.16].

There are also several other available lattice formulations which are based on different forms of PARCOR coefficients. One approach known as the covariance lattice method [3.16] is slightly different in that it relies on a lattice strategy which is based on the covariance of the

signal rather than forward and backward prediction errors.

Several contrasting schemes have been described in this section for producing an LP model so that, in order to choose the most appropriate technique for a particular application, it is necessary to discuss some of the merits of each approach. This comparison is carried out briefly in the next subsection in relation to some of the issues which are pertinent in modelling signals using linear prediction.

3.3.4 Comparison of LP Analysis Methods

(a) Computational Complexity

In order to assess the complexity of the three basic LP techniques, the exact processing and storage resources required by the various methods are of considerable interest. Table 3.1 illustrates in detail the computation and storage requirements [3.17] for the autocorrelation, covariance and lattice methods in terms of the length (N) of the data sequence and the order (p) of the predictor. In terms of data storage, the lattice method clearly requires greater capacity since it is necessary to store the forward and backward prediction errors as well as the data sequence. However, the lattice method does have the advantage of not requiring a window nor a correlation matrix to be held. In spite of this fact, as $N \gg p$ for most applications, the covariance & autocorrelation methods generally require less storage than the lattice method.

Similarly, the number of multiplications required to implement each solution indicates that the Levinson-Durbin algorithm for the autocorrelation method is significantly more efficient than the other techniques.[3] However, this advantage is offset by the fact that the

---

[3] *A modified lattice method which achieves approximately the same efficiency as the covariance method has, however, been suggested by Makhoul [3.16]. This strategy is based on the calculation of signal covariances instead of forward and backward residuals.*

overall amount of processing required is often dominated by the calculation of the matrix coefficients and this procedure requires a similar number of operations as for the covariance method. In any case, if $N \gg p$, both these schemes can obtain solutions at a considerably lower computational cost than the Burg lattice method.

Another issue which may be of relevance is the ease of programming associated with each of these techniques and the autocorrelation method is considered to be simplest in this respect [3.17].

(b) System Stability

As mentioned in relation to the Levinson-Durbin recursion, the autocorrelation method is guaranteed to produce a stable LP system (i.e. all poles within the unit circle in the z-plane). This stability is, however, not guaranteed in finite wordlength implementations or vocoding applications where it may be necessary to quantise the LP coefficients. This is since small changes in predictor coefficients can lead to relatively large deviations in pole positions [3.18] — with even the possibility of an interpolation between two stable sets of predictor coefficients producing an unstable result. However, the Levinson-Durbin algorithm provides a convenient test for stability since if the parameters $\{k_i\}$ satisfy the following condition:

$$-1 \leq k_i \leq 1 \tag{3.65}$$

it is then known that the resulting LP system will be stable.

No similar condition for stability can, however, be supplied for the covariance method and the solution produced by this technique will not necessarily be stable [3.10]. Nevertheless, as the covariance method result approaches that for the autocorrelation method for a long analysis frame, the predictor polynomial will, in practice, almost always be stable. In contrast, the linear predictive filter for the lattice method is always guaranteed to be stable with or

without a finite word length implementation.

(c) Length of Frame

In the above discussion on computational complexity, it was assumed that the length of the data sequence (N) is identical for all three methods of LP solution. However, the necessity of windowing in the autocorrelation method imposes the restriction that N must usually be of the order of several pitch periods [3.5]. This is to ensure that the tapering effects of the window do not seriously affect the properties of the data sequence under consideration.

For the lattice and covariance techniques, there is no such windowing constraint and it is possible to use values of N on the order of twice the predictor order [3.13]. In using short duration frame lengths, pitch synchronous analysis is usually employed but in those cases where pitch extraction is a problem, values of N for the covariance and lattice techniques are typically comparable to those for the autocorrelation method.

3.4 SPECTRAL LINEAR PREDICTION

One of the most important properties of LP analysis is the ability to directly model the frequency response of a given signal independently of the corresponding time sequence. This facet of linear prediction as a spectral modelling technique opens up many new applications and is of particular relevance to the work described in this thesis. In this section, it is shown that this frequency domain formulation yields exactly the same results as the time domain solution described in previous sections and thus the LP technique is an analysis technique which can be approached in either the time or frequency domain.

3.4.1 Frequency Domain Interpretation of LP Analysis

In section 3.2, the minimum squared prediction error, E, was denoted by the sum

$$E = \sum_{n=0}^{N-1} e^2(n) \tag{3.66}$$

Now, the discrete Fourier transform of $e(n)$ is given by:

$$E\left(e^{j\frac{2\pi m}{N}}\right) = E\left(e^{j\omega m}\right) = \sum_{n=0}^{N-1} e(n).e^{-jn\omega m} \tag{3.67}$$

i.e. $E(\cdot)$ is evaluated at the frequencies $\left\{\omega_m = \frac{2\pi m}{N} : m=0,\ldots,N-1\right\}$. Equations (3.66) and (3.67) can be related with reference to the discrete version of Parseval's Theorem [3.20] which states that, for a real sequence $\{x(0),\ldots,x(N\text{-}1)\}$ with a corresponding Fourier transform $\left\{X\left(e^{j\omega m}\right) : m=0,\ldots,N\text{-}1\right\}$,

$$\sum_{n=0}^{N-1} x^2(n) = \frac{1}{N}\sum_{m=0}^{N-1}\left|X\left(e^{j\omega m}\right)\right|^2 \tag{3.68}$$

Thus, from equation (3.66),

$$E = \frac{1}{N}\sum_{m=0}^{N-1}\left|E\left(e^{j\omega m}\right)\right|^2 \tag{3.69}$$

In addition, the sequence $e(n)$ is the result of passing the input sequence $s(n)$ through the LP filter $A(z)$, i.e.

$$E(z) = A(z).S(z) = \left[1 - \sum_{k=1}^{p}\alpha_k z^{-k}\right].S(z) \tag{3.70}$$

$$\text{or,} \quad \left|E\left(e^{j\omega m}\right)\right|^2 = \left|A\left(e^{j\omega m}\right)\right|^2 \cdot \left|S\left(e^{j\omega m}\right)\right|^2 \triangleq \left|A\left(e^{j\omega m}\right)\right|^2 \cdot P(\omega_m) \tag{3.71}$$

where $P(\omega_m)$ denotes the signal power spectrum.

Thus,

$$E = \frac{1}{N} \sum_{m=0}^{N-1} \left| A\left(e^{j\omega_m}\right) \right|^2 \cdot P(\omega_m) \tag{3.72}$$

This result provides a mechanism for calculating the predictor coefficients by minimising the error E using

$$\frac{\partial E}{\partial \alpha_i} = 0, \qquad 1 \leq i \leq p \tag{3.73}$$

It is relatively simple to show [3.12] that this constraint results in the following expression:

$$\frac{1}{N} \sum_{m=0}^{N-1} P(\omega_m) \cos(i\omega_m) = \frac{1}{N} \sum_{k=1}^{p} \alpha_k \sum_{m=0}^{N-1} P(\omega_m) \cos(i\text{-}k)\,\omega_m \tag{3.74}$$

In section 3.3, the autocorrelation coefficients $\{R(k) : k=0,\ldots,p\}$ were calculated directly from the time sequence. However, using the Wiener-Khinchine relation [3.21], it is also possible to evaluate the same autocorrelation values by an inverse Fourier Transform of the signal power spectrum, i.e.

$$R(k) = \frac{1}{N} \sum_{m=0}^{N-1} P(\omega_m)\, e^{jk\omega_m} \tag{3.75}$$

or, since the power spectrum is real and even for real signals,

$$R(k) = \frac{1}{N} \sum_{m=0}^{N-1} P(\omega_m) \cos(k\omega_m) \tag{3.76}$$

Thus, equation (3.74) reduces to

$$R(i) = \sum_{k=1}^{p} \alpha_k R(i\text{-}k)\,, \qquad 1 \leq i \leq p \tag{3.77}$$

This set of p linear equations is identical to the normal equations described in section 3.3. Furthermore, if equations (3.76),(3.77) are substituted into equation (3.72),

$$E = R(0) - \sum_{k=1}^{p} \alpha_k R(k) \tag{3.78}$$

This result is also identical to that in section 3.3 and thus the model gain (G) can again be evaluated using $G^2 = E$. These important results illustrate the property that the predictor parameters obtained from this frequency domain approach are identical to those from the time domain formulation. The significance of the frequency domain method lies in the fact that *the predictor coefficients and model gain can be calculated solely from the signal power spectrum* if the autocorrelation coefficients are evaluated using equation (3.76).

Another important feature of linear predictive spectral modelling emerges from a consideration of the manner in which the signal spectrum is approximated by the all-pole LP model spectrum. From section 2.6, the transfer function, $\hat{S}(z)$, of the LP synthesis model is given by:

$$\hat{S}(z) = \frac{G}{A(z)} \tag{3.79}$$

$$\text{or,} \quad \hat{P}(\omega_m) \triangleq \left|\hat{S}\left(e^{j\omega m}\right)\right|^2 = \frac{G^2}{\left|A\left(e^{j\omega m}\right)\right|^2} \tag{3.80}$$

where $\{\hat{P}(\omega_m) : m=0,\ldots,N-1\}$ is the model power spectrum. Combining this expression with equation (3.72) produces the result:

$$E = \frac{G^2}{N} \sum_{m=0}^{N-1} \frac{P(\omega_m)}{\hat{P}(\omega_m)} \tag{3.81}$$

Since the summand is always positive, it follows that the LP solution which provides *the lowest value of total error also minimises the sum of the ratios of signal power to the LP model power spectrum* over all defined frequency points. As a consequence of this observation, regions where $P(\omega_m) > \hat{P}(\omega_m)$ will result in a greater contribution to the total error than areas where $P(\omega_m) < \hat{P}(\omega_m)$. The model spectrum will, therefore, provide a much closer match to the signal frequency response at spectral peaks than in regions of lower signal energy. This result is illustrated in Fig. 3.2 for a speech segment sampled at 10kHz from the voiced utterance /ER/ in the word "mercy". The underlying curve depicts the FFT of the speech signal and this is contrasted with the smoother curve corresponding to the frequency response of a 12th order LP model.

3.4.2 Comparison with Other Spectrum Analysis Methods

One spectral modelling scheme which has already been discussed previously is that based on homomorphic processing. In order to effect a comparative evaluation of this method in relation to linear prediction, Fig. 3.3 shows the Fourier transform for the same signal as employed in Fig. 3.2 along with a model spectrum obtained using homomorphic smoothing. The manner in which this smoothed spectrum can be produced is described in section 2.5.

In both of these figures, the disadvantage of simply using an FFT to obtain a speech spectral estimate are clearly evident. One major drawback is the occurrence of pitch ripples (i.e. individual harmonics of the excitation) in the spectrum which can lead to serious difficulties in attempting to resolve individual formants for applications such as formant extraction etc. Although the use of a shorter analysis frame can reduce the effects of these ripples, it is still not a simple process to reliably locate or identify formant frequencies from a straightforward Fourier transform. Furthermore, the bit-rate or memory requirements to transmit or store the Fourier transform representation are considerably greater than those required to produce a spectral estimate for other parametric representations.

The homomorphically smoothed spectrum shows no similar pitch harmonics and the calculation of formant positions is often a relatively simple process. However, the formant bandwidths obtained from these spectra are usually grossly inaccurate [3.22] due to the inherent smoothing of the short-time log spectrum associated with homomorphic processing.

The result of linear predictive analysis shown in Fig. 3.2 also appears to provide a good representation of the formant structure with an almost complete elimination of the pitch ripples. This is since linear prediction tends to model the envelope of the spectrum without being sensitive to the fine structure caused by the periodicity of the time signal [3.19]. Furthermore, the formant bandwidths are usually more accurate than those obtained from homomorphic smoothing.

The LP spectrum also seems to result in fewer extraneous peaks (which may be interpreted as formants) than the homomorphic spectrum. This consideration is very significant in formant trajectory estimation since this application not only requires precise formant estimates, but it is also imperative to be able to associate each peak with a given formant and to track the variation in position of this formant with time. The number of spectral peaks must, therefore, correspond closely to the number of speech formants with as few additional peaks as possible.

This important characteristic of LP spectral modelling arises from the fact that the LP method is an all-pole representation with a particular model of order p giving rise to total of p poles. Now, a typical speech spectrum will generally be associated with an average density of 2 poles per kiloHertz due to the vocal tract transfer function [3.13] and thus a total of f_s poles are required to model the region from d.c. to the Nyquist frequency — where f_s is the sampling rate in kHz. In addition, around 2-4 poles are required to represent the effects of the excitation source and the radiation load. Thus, for a 10kHz sampling rate, a model order of around 12-14 would be required. Although this figure is only an estimate, it provides a rough figure for the total number of parameters to employ in the LP spectral model and consequently the number

of spectral peaks in the model spectrum.

However, there is no such relationship between model parameters and the number of spectral peaks for homomorphic analysis and it is, therefore, much more difficult to estimate the required number of parameters to employ in this technique. As a result, the occurrence of extraneous peaks is much more probable.

Another approach to spectral modelling is provided by the technique known as "*Analysis-by-Synthesis (AbS)*". This general technique begins by assuming a given model for some representation of the speech signal such as the Fourier transform. The parameters of this model are then adjusted in a systematic manner in order to find the set of parameters which produces the closest model representation to the given speech representation. In spectral AbS, the chosen characteristic is the speech Fourier transform and the model typically consists of a set of poles and zeros. The fit between the real spectrum $P(\omega_m)$ and the model spectrum $\hat{P}(\omega_m)$ is optimised through the minimisation of an error criterion, $\tilde{E}$, which is often of the form [3.19]:

$$\tilde{E} = \sum_{m=0}^{N-1} \left[\log \left(\frac{P(\omega_m)}{\hat{P}(\omega_m)} \right) \right]^2 = \sum_{m=0}^{N-1} \left[\log P(\omega_m) - \log \hat{P}(\omega_m) \right]^2 \qquad (3.82)$$

This error measure consisting of the total squared error between the two log spectra is preferred to that obtained in LP analysis as described by equation (3.81) since it is thought to relate more closely to speech perception [3.18]. Another disadvantage of using equation (3.81) is that errors which occur when $P(\omega_m) > \hat{P}(\omega_m)$ cancel those when $P(\omega_m) < \hat{P}(\omega_m)$ [3.10] whereas this problem does not arise for $\tilde{E}$ since a squared difference is employed.

As an aside, it may be worthwhile to note that error criteria of the form of equation (3.82) can also be employed in a form of LP analysis where $\hat{P}(\omega_m)$ constitutes an LP model spectrum and $\tilde{E}$ is then minimised with respect to the predictor coefficients [3.10]. However,

this modification to the LP method introduces serious problems since it results in a set of non-linear equations which must be solved iteratively and thus imposes a dramatic increase on the computational load. The issue of the use of error criteria in spectral modelling is considered in more detail in Chapter 4.

Returning to the problem of spectral analysis-by-synthesis, early attempts to minimise the error $\tilde{E}$ were directed at recursively varying each individual pole or zero [3.23]. However, a Newton-Raphson technique developed by Olive [3.24] allows for an iterative computation on all poles simultaneously and in a straghtforward automatic manner. Nevertheless, the main disadvantage of spectral AbS is the computational cost associated with this technique. This factor is of such great significance that the simplicity, ease of implementation and efficiency of LP analysis outweighs the advantages of AbS in most applications.

3.5 APPLICATIONS OF LINEAR PREDICTIVE ANALYSIS

The vocoding and spectral estimation properties of LP analysis have already been discussed in previous sections and this section will attempt to draw attention to some of the other procedures and applications in which LP methods have been employed.

3.5.1 Pitch Detection

In section 3.2, the prediction error signal, $e(n)$, was defined as

$$e(n) = s(n) - \sum_{k=1}^{p} \alpha_k \, s(n-k) \tag{3.83}$$

and when the input signal $s(n)$ is modelled exactly by the LP all-pole description, the prediction error is related to the LP model excitation $u(n)$ as:

$$e(n) = Gu(n) \tag{3.84}$$

The excitation used for voiced speech consists of impulses spaced at the pitch period and thus, the prediction error for a true speech signal would be expected to exhibit similar behaviour. To illustrate this point, Fig. 3.4 shows a windowed speech signal and the corresponding prediction error signal for a 12th order LP analysis based on the autocorrelation method. As expected, the prediction error is large at the beginning of each pitch period and the simplest means of pitch estimation from LP analysis is to define the difference between peaks above a certain threshold in the prediction error as the pitch period.

A more sophisticated approach known as *SIFT* (*S*imple *I*nverse *F*iltering *T*echnique) has also been suggested by Markel [3.25]. The signal is first passed through a 900 Hz cutoff lowpass filter and then decimated by a factor of 5 in order to remove most of the higher speech formants. The resulting down-sampled signal is then processed by a 4th order LP analysis – which is deemed sufficient to represent the decimated data. The linear prediction error is then evaluated and autocorrelation analysis is carried out on this error. The largest peak within an appropriate range in this autocorrelation function is finally chosen as the pitch period.

3.5.2 Formant Analysis

The spectral estimation properties of LP analysis were described in section 3.4 and it was noted that the LP spectrum provides a smoothed version of the Fourier transform in which the spectral peaks corresponding to speech formants are clearly recognisable. In fact, one simple technique for formant extraction consists of simply peak picking the LP model frequency response and calculating formant centre frequencies and bandwidths directly from these peaks [3.26].

Another more sophisticated approach relies on a factorisation of the predictor polynomial, $A(z)$, in order to solve for the roots of this polynomial [3.27]. On the basis of these roots, a decision is made as to which roots correspond to true speech formants. These methods for formant extraction based on linear prediction are examined in considerable detail

in Chapters 5,6. The advantage of LP techniques over other methods is their ability to offer minimal complexity, minimal computation time and maximal accuracy in formant estimation [3.5].

3.5.3 Acoustic Tube Modelling

A consideration of the similarity between the lattice-type speech synthesis structures for the acoustic tube models of section 2.3 and the linear predictive lattice formulations of section 3.3.3 may suggest a relationship between these two systems. Such a relationship does indeed exist and to deduce its exact form, consider the configuration of Fig 2.4 which yields

$$U^+_{k+1}(z) = (1 + r_k)\, z^{-1}\, U^+_k(z) + r_k\, U^-_{k+1}(z) \tag{3.85a}$$

$$U^-_k(z) = -r_k\, z^{-1}\, U^+_k(z) + (1 - r_k)\, U^-_{k+1}(z) \tag{3.85b}$$

where $U^+_k(z)$, $U^-_k(z)$ are the z-transforms of the forward and backward volume velocities respectively in the k-th acoustic tube section and r_k is the reflection coefficient for this section. Equations (3.85) can be rewritten as

$$U^+_{k+1}(z) = \frac{1}{(1 - r_k)}\left[z^{-1}\, U^+_k(z) + r_k\, U^-_k(z)\right] \tag{3.86a}$$

$$U^-_{k+1}(z) = \frac{1}{(1 - r_k)}\left[U^-_k(z) + r_k\, z^{-1}\, U^+_k(z)\right] \tag{3.86b}$$

Also, for the lattice formulation of section 3.3.3,

$$B^{(i)}(z) = z^{-1} B^{(i-1)}(z) - k_i F^{(i-1)}(z) \tag{3.87a}$$

$$\& \quad F^{(i)}(z) = F^{(i-1)}(z) - k_i z^{-1} B^{(i-1)}(z) \tag{3.87b}$$

where $F^{(i)}(z)$, $B^{(i)}(z)$ are the z-transforms of the i-th stage forward and backward prediction errors respectively and k_i is the i-th PARCOR coefficient. Equations (3.86) can, therefore, be converted to a form similar to that for equation (3.87) if (a) the (k+1)st section corresponds to the i-th iteration in the lattice formulation and (b) the superscript $^+$ is equivalent to a backward error and vice versa. Now, apart from an overall gain value of $\frac{1}{(1 - r_k)}$ (which can be incorporated into the overall system gain), equations (3.86) & (3.87) are identical if the following relationship holds,

$$r_i = - k_i , \qquad 1 \leq i \leq p \tag{3.88}$$

However, the reflection coefficient r_i is also given by:

$$r_i = \frac{\mathcal{A}_{i+1} - \mathcal{A}_i}{\mathcal{A}_{i+1} + \mathcal{A}_i} \tag{3.89}$$

where $\mathcal{A}_i$ is the cross-sectional area of the i-th tube. Thus,

$$\mathcal{A}_{i+1} = \left(\frac{1 - k_i}{1 + k_i}\right) \mathcal{A}_i \tag{3.90}$$

The linear predictive PARCOR coefficients can, therefore, be employed to obtain an estimate of the area function of the vocal tract. Wakita [3.28] has shown that if the input speech is suitably preprocessed in order to remove the effects of the excitation and the radiation load, then the area functions obtained from LP analysis are often a reasonable approximation to the actual vocal tract shapes employed in human speech. These applications of the LP parameters offer an indication of the versatility of the linear predictive method and the way in which it can be used to obtain robust, reliable and accurate estimates of most of the basic speech parameters.

3.6 REFERENCES

[3.1] SORENSON, H.W., "Least-Squares Estimation: From Gauss to Kalman", *IEEE Spectrum, Vol.* 7, 1970, *pp* 63-68.

[3.2] WIENER, N., "Extrapolation, Interpolation and Smoothing of Stationary Time Series", *M.I.T. Press, Massachusetts*, 1966.

[3.3] 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.

[3.4] ATAL, B.S., SCHROEDER, M.R., "Predictive Coding of Speech Signals", *Proc.* 1967 *Conf. Commun. and Process.*, 1967, *pp* 360-361.

[3.5] MARKEL, J.D., GRAY, A.H. Jr., "Linear Prediction of Speech", *Springer-Verlag, Berlin-Heidelberg-New York*, 1976.

[3.6] ATAL, B.S., HANAUER, S.L., "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", *J. Acoust. Soc. Amer., Vol.* 50, *No.* 2, 1971, *pp* 637-655.

[3.7] BOX, G.E., JENKINS, G.M., "Time Series Analysis, Forecasting and Control", *Holden-Day, San Francisco, Calif.*, 1970.

[3.8] LEVINSON, N., "The Wiener RMS (root mean square) error criterion in filter design and prediction", *J. Math. Phys., Vol.* 25, *No.* 4, 1947, *pp* 261-278.

[3.9] DURBIN, J., "The fitting of time-series models", *Rev. Inst. Int. Statist., Vol.* 28, *No.* 3, 1960, *pp* 233-243.

[3.10] MAKHOUL, J., "Linear Prediction: A Tutorial Review", *Proc. IEEE, Vol.* 63, *April*, 1975, *pp* 561-580.

[3.11] ITAKURA, F.I., SAITO, S., "A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies", *Elec. and Comm. in Japan, Vol.* 53-*A*, *No.* 1, 1970, *pp* 36-43.

[3.12] MAKHOUL, J., WOLF, J., "Linear Prediction and the Spectral Analysis of Speech",

BBN Report No. 2304, *Bolt Beranek and Newman, Inc., Cambridge, Mass., August* 1972.

[3.13] RABINER, L.R., SCHAFER, R.W., "Digital Processing of Speech Signals", *Prentice-Hall, Englewood Cliffs, New Jersey,* 1978.

[3.14] McDONOUGH, R.N., HUGGINS, W.H., "Best least-squares representation of signals by exponentials", *IEEE Trans. Automat. Contr., Vol. AC-*13, *Aug.* 1968, *pp* 408-412.

[3.15] BURG, J., "A New Analysis Technique for Time Series Data", *Proc. NATO Advanced Study Inst. on Signal Proc., Enschede, Netherlands,* 1968.

[3.16] MAKHOUL, J., "Stable and Efficient Lattice Methods for Linear Prediction", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP-*25, *Oct.* 1977, *pp* 423-428.

[3.17] WITTEN, I.H., "Algorithms for adaptive linear prediction", *Computer Jnl., Vol.* 23, *No.* 1, *Jan.* 1980, *pp* 78-84.

[3.18] VISWANATHAN, R., MAKHOUL, J., "Quantization Properties of Transmission Parameters in Linear Predictive Systems", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP-*23, *No.* 3, *June* 1975, *pp* 309-321.

[3.19] MAKHOUL, J., "Spectral Linear Prediction: Properties and Applications", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP-*23, *No.* 3, *June* 1975, *pp* 283-296.

[3.20] ELLIOTT, D.F., RAO, K.R., "Fast Transforms, Algorithms, Analyses, Applications", *Academic Press Inc.,* 1982.

[3.21] EDWARD, J.A., FITELSON, M.M., "Notes on maximum entropy processing", *IEEE Trans. Inform. Theory, Vol. IT-*19, *March* 1973, *pp* 232-234.

[3.22] KOPEC, G.E., OPPENHEIM, A.V., TRIBOLET, J.M., "Speech Analysis by Homomorphic Prediction", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP-*25, *No.* 1, *Feb.* 1977, *pp* 40-49.

[3.23] MATHEWS, M.V., MILLER, J.E., DAVID, E.E. Jr., "Pitch Synchronous Analysis of Voiced Sounds", *J. Acoust. Soc. Amer., Vol.* 33, *No.* 2, 1961, *pp* 179-186.

[3.24] OLIVE, J., "Automatic Formant Tracking in a Newton-Raphson Technique", *J.*

Acoust. Soc. Amer., *Vol.* 50, *August* 1971, *pp* 661-670.

[3.25] MARKEL, J.D., "The SIFT Algorithm for Fundamental Frequency Estimation", *IEEE Trans. on Audio and Electroacoustics*, *Vol. AU*-20, *No.* 5, *Dec.* 1972, *pp* 367-377.

[3.26] McCANDLESS, S.S., "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra", *IEEE Trans. on Acoust. Speech and Sig. Proc.*, *Vol. ASSP*-22, *No.* 2, *April* 1974, *pp* 135-141.

[3.27] MARKEL, J.D., "Digital Inverse Filtering — A New Tool for Formant Trajectory Estimation", *IEEE Trans. on Audio and Electroacoustics*, *Vol. AU*-20, *No.* 2, *June* 1972, *pp* 129-137.

[3.28] 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, *Oct.* 1973, *pp* 417-427.

PROPERTY		Autocorrelation Method (Levinson-Durbin)	Covariance Method (Cholesky)	Lattice Method (Burg)
STORAGE	Data	N	N	$3N$
	Window	N	—	—
	Correlation Matrix	$p+1$	$\frac{p^2}{2} + \mathrm{O}(p)$	—
COMPUTATION (Multiplications)	Windowing	N	—	—
	Matrix Evaluation	$pN + \mathrm{O}(N)$	$pN + \frac{p^2}{2} + \mathrm{O}(N) + \mathrm{O}(p)$	—
	Matrix Solution	$p^2 + \mathrm{O}(p)$	$\frac{p^3}{6} + p^2 + \mathrm{O}(p)$	$5pN + \mathrm{O}(p)$

Table 3.1 Computational Resources for Various Linear Prediction Methods

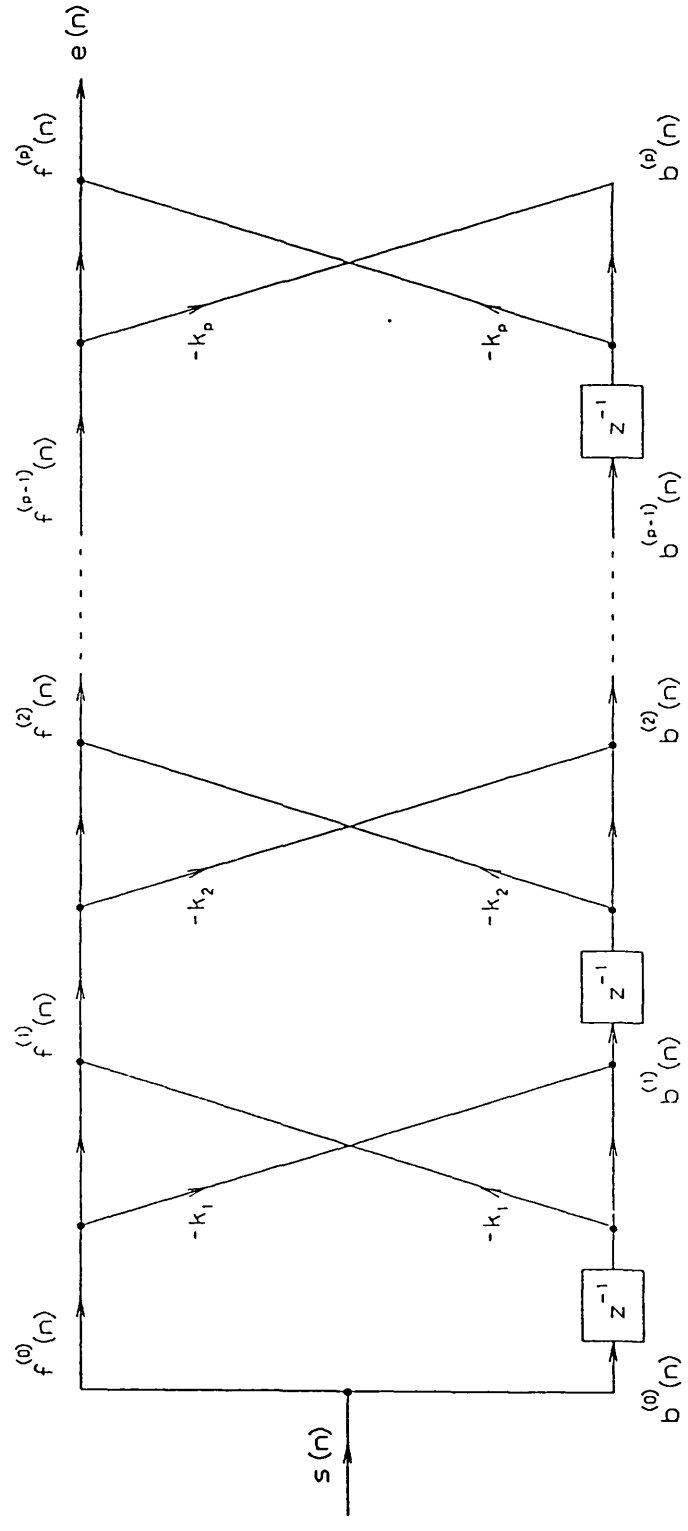


Figure 3.1 Block Diagram of a Lattice Implementation

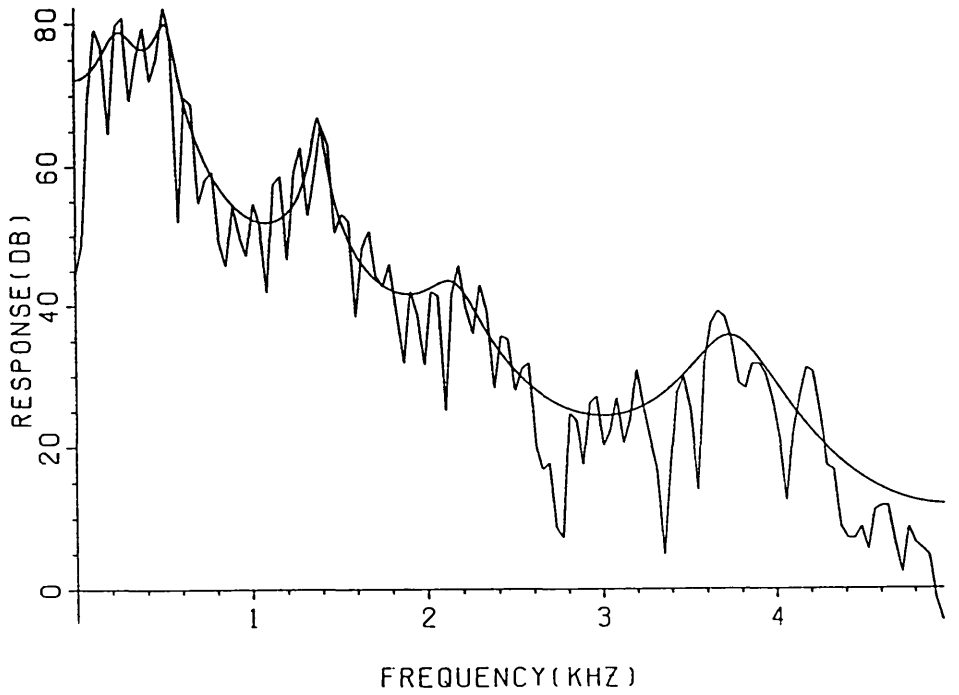


Figure 3.2 Comparison of a Signal Spectrum for Voiced Speech and a Linear Predictive Model Spectrum

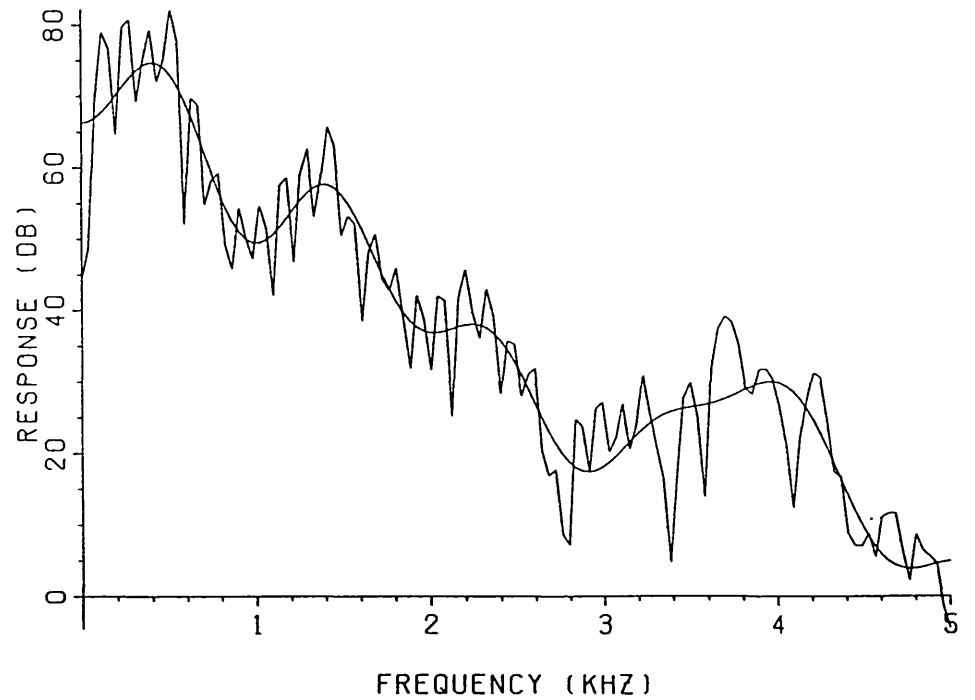


Figure 3.3 Illustration of Spectral Estimation based on Homomorphic Analysis for a Voiced Speech Signal

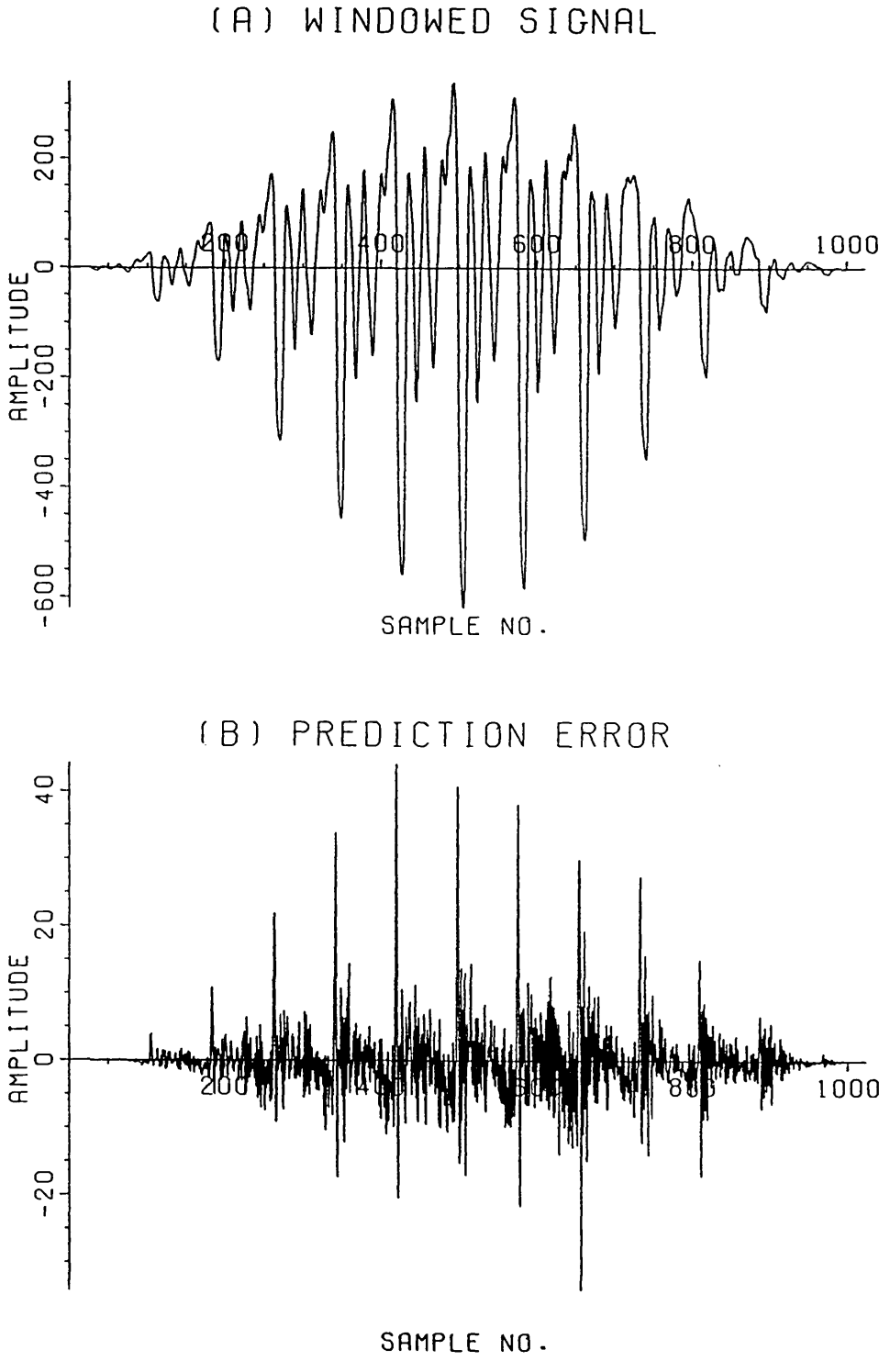


Figure 3.4 Linear Prediction Error Analysis for Voiced Speech

CHAPTER 4

THE OPTIMAL SEGMENTATION OF SPEECH SPECTRA FOR SPECTRAL ESTIMATION

CHAPTER 4. THE OPTIMAL SEGMENTATION OF SPEECH SPECTRA FOR SPECTRAL ESTIMATION

4.1 INTRODUCTION

In the last chapter, the pre-eminence of linear predictive methods was discussed in relation to a wide variety of speech processing applications. Some of the reasons for this diversity arise from the existence of efficient algorithms for the evaluation of predictor coefficients and the ability of LP parameters to provide accurate estimates of basic speech parameters such as fundamental frequency, formant positions and vocal tract area functions.

In addition, the role of LP in the area of spectral modelling was highlighted as an effective means of speech spectrum smoothing for several applications. For instance, the model spectra may be employed in speaker/speech recognition systems [4.1] as representations of the original signal. Alternatively, the smoothed signal frequency response is also of a suitable form to enhance considerably the efficacy of formant extraction processes.

It is also possible to generalise the method of Spectral Linear Prediction to the case where a particular region of the signal spectrum is matched by a model frequency response rather than the entire spectral range. This technique is known as "*Selective Linear Prediction*" and was first described by Makhoul [4.2],[4.3] for continuous signal spectra. However, since this work is primarily concerned with discrete time & frequency systems, a Selective LP procedure is developed in this chapter for a discrete signal frequency response. Although Selective LP may be employed for a variety of signal spectra, the traditional rationale behind its applicability to speech sounds is fairly specific and thus some of these reasons are outlined in section 4.2.

Two important factors which determine the accuracy of the model spectrum are (i) the break frequencies separating the Selective LP segments and (ii) the predictor order to be adopted in each individual region of the overall spectrum. However, the choice of these parameters is usually fairly arbitrary and is generally decided on the basis of a qualitative assessment of the signal energy distribution.

In this chapter, a fundamentally different approach is introduced for the Selective LP modelling of speech spectra which optimises the fit between signal and model frequency responses with respect to both the segment break frequencies and predictor orders. The technique utilised in the implementation of this optimisation process is that of Dynamic Programming (DP). Therefore, following an introductory discussion of the principles of Dynamic Programming, section 4.4 details the development of an algorithm for frequency response segmentation based on DP.

The error measure to be minimised during the optimisation procedure is an important consideration and thus several error criteria are discussed in the context of their applicability to the particular type of signal under consideration. Another segmentation scheme is finally outlined which, although suboptimal in nature, results in significant computational advantages over the optimal approach.

4.2 SELECTIVE LINEAR PREDICTION

4.2.1 Development of a Selective LP Modelling Scheme for Discrete Spectra

In the time domain formulation of the autocorrelation method of LP analysis mentioned in Chapter 3, the N windowed speech samples $\{s(n) : n=0,\ldots,N-1\}$ were employed in the evaluation of the signal autocorrelation sequence $\{R(k)\}$ as:

$$R(k) = \sum_{n=0}^{N-1-k} s(n).s(n+k) \qquad (4.1)$$

However, for the purpose of spectral estimation, the only available signal description is the discrete power spectrum $\{ P(\omega_m) : m=0,\ldots,N-1,\ \omega_m = \frac{2\pi m}{N} \}$ which is given by

$$P(\omega_m) = P\left(\frac{2\pi m}{N}\right) = \left| \sum_{n=0}^{N-1} s(n) \cdot e^{-jn\omega_m} \right|^2 \tag{4.2}$$

As discussed in section 3.4, the autocorrelation values may be obtained in this case by an inverse DFT operation,

$$R(k) = \frac{1}{N} \sum_{m=0}^{N-1} P(\omega_m)\, e^{jk\omega_m} \tag{4.3}$$

or
$$R(k) = \frac{1}{N} \left\{ P(0) \cdot e^{j\cdot 0} + \sum_{m=1}^{N/2-1} P\left(\frac{2\pi m}{N}\right) e^{j\frac{2\pi km}{N}} + P(\pi) \cdot e^{jk\pi} + \sum_{m=N/2+1}^{N-1} P\left(\frac{2\pi m}{N}\right) e^{j\frac{2\pi km}{N}} \right\} \tag{4.4}$$

However, as $P(\omega_m)$ is real and even for real signals, i.e.

$$P(2\pi-\omega_m) = P(\omega_m) \tag{4.5a}$$

or
$$P\left(\frac{2\pi(N-m)}{N}\right) = P\left(\frac{2\pi m}{N}\right) \tag{4.5b}$$

Equation (4.4) then becomes:

$$R(k) = \frac{1}{N} \left\{ P(0) + (-1)^k P(\pi) + \sum_{m=1}^{N/2-1} P\left(\frac{2\pi m}{N}\right) \left(e^{j\frac{2\pi km}{N}} + e^{-j\frac{2\pi km}{N}} \right) \right\} \tag{4.6}$$

Thus,
$$R(k) = \frac{1}{N} \left\{ P(0) + (-1)^k P(\pi) + 2 \sum_{m=1}^{N/2-1} P(\omega_m) \cos(k\omega_m) \right\} \tag{4.7}$$

This expression reduces the computational requirements considerably over a direct evaluation of the autocorrelation sequence from equation (4.3).

In the approach to Selective LP based on discrete spectra described in this section, the above technique is generalised to the case where a given portion of the speech spectrum is to be modelled by an LP estimate. The underlying concept involves a translation of the selected region in the actual frequency domain to the angular range from 0 to π in the Selective LP domain.

Thus, if the chosen segment covers the range $\{P(\omega_m) : a \leq m \leq b\}$, then the frequency range which is to be modelled is given by the semi-closed interval[1]

$$\omega_a \leq \omega_m < \omega_{b+1} \tag{4.8a}$$

or

$$\frac{2\pi a}{N} \leq \omega_m < \frac{2\pi(b+1)}{N} \tag{4.8b}$$

This situation is illustrated in Fig. 4.1 (a) where • denotes that $\omega_m = \omega_a$ is part of the chosen interval while o indicates that $\omega_m = \omega_{b+1}$ is not included. As shown in Fig. 4.1 (b), this frequency range is mapped onto the entire upper half of the unit circle in the Selective LP domain such that $\omega_a \longrightarrow 0$ and $\omega_{b+1} \longrightarrow \pi$. Similarly, the corresponding conjugate region $\omega_{N-1-b} \leq \omega_m < \omega_{N-a}$ is mapped onto the lower half of the unit circle.

In order to achieve the desired linear mapping, the transformations necessary to convert radian frequencies (ω_m) in the signal domain to frequencies (ω'_m) in the Selective LP domain are given by:

---

[1] *The validity of this expression is revealed on consideration of the case where the entire frequency response is to be modelled. Thus, for* $a=0$, $b=N/2-1$, *the chosen range is* $0 \leq \omega_m < \pi$ *as required.*

$$\omega_m' = \pi\left(\frac{m-a}{b-a+1}\right) = \pi\left(\frac{\omega_m-\omega_a}{\omega_b-\omega_a+\frac{2\pi}{N}}\right) \qquad a \le m \le b \tag{4.9a}$$

and $$\omega_m' = \pi\left(1+\frac{m-(N\text{-}1\text{-}b)}{b-a+1}\right) = \pi\left(1+\frac{\omega_m-\omega_{N-1-b}}{\omega_b-\omega_a+\frac{2\pi}{N}}\right) \qquad N\text{-}1\text{-}b \le m \le N\text{-}1\text{-}a \tag{4.9b}$$

Therefore, equation (4.9a) defines ω_m' over the range $0 \le \omega_m' < \pi$ while (4.9b) provides a similar definition over the remainder $\pi \le \omega_m' < 2\pi$ of the unit circle. A complete "frequency response" $\{P'(\omega_m')\}$ is now available in the Selective LP domain such that

$$P'(\omega_m') = \begin{cases} P(\omega_m) & a \le m \le b \\ P(\omega_m) & N\text{-}1\text{-}b \le m \le N\text{-}1\text{-}a \\ \text{undefined otherwise} \end{cases} \tag{4.10}$$

Furthermore, this translated spectrum is constituted of a total number of points (N') given by

$$N' = 2(b-a+1) \tag{4.11}$$

An expression such as equation (4.7) can, therefore, be employed to calculate the required autocorrelation values $\{R'(k)\}$,

$$R'(k) = \frac{1}{N'}\left\{P'(0) + (-1)^k P'(\pi) + 2\sum_{m=a+1}^{b} P'(\omega_m')\cos(k\omega_m')\right\} \tag{4.12}$$

or, in terms of the actual power spectrum $\{P(\omega_m)\}$,

$$R'(k) = \frac{1}{N'}\left\{P(\omega_a) + (-1)^k P(\omega_{b+1}) + 2\sum_{m=a+1}^{b} P(\omega_m)\cos(k\omega_m')\right\} \tag{4.13}$$

This equation represents a crucial step in the Selective LP modelling approach for discrete spectra and the remaining stages for the evaluation of a model spectrum are similar to

those for standard LP modelling described in Chapter 3. The entire procedure for the calculation of this spectral estimate is, therefore, as follows:

(1) On the basis of a chosen range of spectral samples $\left\{P(\omega_m) = P\left(\frac{2\pi m}{N}\right) : a \leq m \leq b\right\}$, transform actual frequency values (ω_m) into the corresponding frequencies (ω_m') within the Selective LP domain:

$$\omega_m' = \pi\left(\frac{m-a}{b-a+1}\right) = \pi\left(\frac{\omega_m - \omega_a}{\omega_b - \omega_a + \frac{2\pi}{N}}\right) \qquad a \leq m \leq b \tag{4.14}$$

(2) For a linear prediction model of order p, compute the first $(p+1)$ autocorrelation coefficients $\{R'(k) : k=0,\ldots,p\}$ using

$$R'(k) = \frac{1}{N'}\left\{P(\omega_a) + (-1)^k P(\omega_{b+1}) + 2\sum_{m=a+1}^{b} P(\omega_m)\cos(k\omega_m')\right\} \tag{4.15}$$

with the size of the transform being given by $N'=2(b-a+1)$.

(3) Use these autocorrelation values to calculate the predictor coefficients $\{\alpha_k : k=1,\ldots,p\}$ and model gain (G) with the aid of the Levinson-Durbin recursion covered in section 3.3.

(4) The discrete model spectrum $\{\hat{P}(\omega_m) : m=a,\ldots,b\}$ is finally derived from the LP parameters such that

$$\hat{P}(\omega_m) = \frac{G^2}{\left|1 - \sum_{k=1}^{p} \alpha_k\, e^{-jk\omega_m'}\right|^2} \qquad a \leq m \leq b \tag{4.16}$$

This spectrum is defined at the originally selected spectral portion of the original data $\{\omega_m : m=a,\ldots,b\}$ and forms a linear predictive approximation to the set of spectral

samples $\{P(\omega_m) : m=a,\ldots,b\}$.

An important point to note concerning the Selective LP model considered in this section is that it is based on the autocorrelation technique and not the covariance method of linear prediction. This is due to the problems associated with spectral interpretation for the covariance approach since this method does not produce a one-dimensional frequency domain representation of the data sequence [4.4].

4.2.2 Properties of Selective LP

An example of Selective Linear Prediction applied to speech signals is provided by Fig. 4.2. The speech data is taken from the utterance /OW/ as in the word "b*ough*t" sampled at 10 kHz with a frame length of $N=256$. The figure illustrates the speech spectrum along with the 12th order LP model frequency response and the fit for a 2 region Selective LP model. The predictor orders and the breakpoint are as shown in Fig. 4.2. It should be noted that, since the spectral matching for each region is carried out independently, there is no guarantee that there will not be a discontinuity at the break point between the two Selective LP regions.

One of the reasons that Selective LP has traditionally been regarded as being of potential value lies in the particular spectral energy distributions of speech sounds which suggest the desirability of splitting the speech spectrum into several regions and treating each segment as an independent spectral estimation problem.

This point is illustrated by the consideration that it is often necessary to sample speech at a 20 kHz rate to deal with the whole range of voiced and unvoiced sounds. Although the 0–5 kHz interval is of most interest for voiced speech since this region contains the first four or five formants and the majority of the signal energy, unvoiced speech may exhibit considerable signal energy outside this range.

As an accurate spectral characterisation of unvoiced sounds does not appear to require as many coefficients as voiced utterances [4.4], Selective LP provides a means of analysing low and high frequency regions separately with fewer total coefficients while retaining an accurate spectral representation. For example, a high predictor order (e.g. $p_1=12$) may be allocated to the 0–5 kHz region for voiced speech and a lower order (e.g. $p_2=5$) for the 5–10 kHz segment. This is in contrast to the case where a predictor of order 24-28 would be required to match the entire 0–10 kHz range.

In this manner, not only is the computational load reduced, but a lower number of poles will need to be interpreted if formant analysis is involved. The linear prediction parameters can also be stored for subsequent use in speaker/speech recognition systems where the Selective LP spectrum is used as a template against which the actual power spectral density of a given utterance can be compared.

Even in the 0–5 kHz interval, it would seem worthwhile to segment this range into smaller intervals since the majority of signal energy is concentrated for most voiced sounds below 2–3 kHz and so it would be more important to represent these areas accurately than higher frequencies. To implement the same process in the time domain would, of course, be a much more complicated and inflexible operation (particularly if the sampling frequency is not a direct multiple of the break frequency).

It should be noted that this Selective LP approach is not to be confused with the Sample-Selective Linear Prediction method discussed by Mizoguchi [4.5] and Miyoshi [4.6]. This latter technique is performed in the time domain and is based on the selective use of the speech signal in order to reject data disturbed by the excitation source. This is in contrast to the procedure introduced in the preceding subsection which employs values selected from a frequency domain signal representation.

4.3 FUNDAMENTALS OF DYNAMIC PROGRAMMING

In the previous section, the break frequencies between Selective LP segments and the predictor order within each region could be selected totally arbitrarily. However, this work is concerned with the development of a procedure whereby these two sets of parameters can be evaluated such that the fit between the original speech spectrum and the model frequency response consisting of several LP regions is 'best' in relation to a suitable error criterion. Whilst the exact nature of this error measure will be discussed in detail in section 4.5, this section is devoted to the introduction of an approach which can be employed to calculate the optimal sets of breakpoints and predictor orders for a particular speech spectrum.

With 2 regions and given values for the two predictor orders, the optimal break frequency can be selected fairly simply as that which yields the minimum total error. This problem will, however, become incredibly complex if 3 or more regions are considered — and even more so if the number of the predictor coefficients to adopt in each of the various regions is also to be chosen optimally in some preconstrained manner.

It is at this point that we must resort to a more sophisticated technique than a simple enumeration of all possible combinations of break frequencies and predictor orders to determine that which is associated with the minimum error. The approach adopted in this thesis is that of Dynamic Programming (DP) and, therefore, it is instructive to discuss some of the general principles and properties of DP which result in its suitability to the particular problem under investigation.

4.3.1 Properties of the Dynamic Programming Formulation

Dynamic Programming is a technique for the solution of a large class of problems which either are or can be transformed into "*multi-stage decision processes*". A problem is considered to be a multi-stage decision process if it can be formulated in such a manner that

the parameter values which define the solution of the problem can be determined through the analysis of a sequence of simpler problems, each of which requires the calculation of only one (or a few) variables.

Since Dynamic Programming can not be completely described by any particular algorithm, it is more accurately visualised in terms of a general approach to optimisation which has been applied to a diverse variety of problem areas and types [4.7],[4.8],[4.9] which have been found to be amenable to the dynamic programming methodology. Some of the reasons for this widespread popularity of DP techniques are outlined below:

- One attractive feature of dynamic programming is that it involves the substitution of one N-variable problem by N-single variable problems. This substitution will usually require very much less computational effort whenever it is possible [4.10]. It should be noted that classical extremum methods can not generally be employed to achieve this form of simplification.

- Another important advantage of DP over almost all other extant computational methods, and especially classical optimisation techniques, is the ability of dynamic programming to determine global extrema rather than local optima.

- A characteristic which is of particular interest in this work is the efficiency of DP in producing solutions for constrained optimisation problems. In virtually all other optimisation techniques, the introduction of certain types of constraints can cause significant difficulties. For example, the imposition of integrality cannot be handled by classical methods while this requirement actually simplifies the DP computational process to a great extent. Furthermore, similar considerations apply to the nonnegativity of certain variables and other such restrictions.

- A further important property is the "embedded" nature of the DP formulation. Thus, in the course of solving a particular problem, all the other subproblems which are embedded in the larger problem are also solved.

4.3.2 Dynamic Programming Nomenclature

(a) The physical, operational or conceptual system under investigation in a DP problem is considered to progress through a sequence of consecutive *stages*, i.e. $\{n=1,\ldots,N\}$ for an N-stage process where n is termed the *stage variable.*

(b) Given the stage variable, all the remaining parameters required to describe the system are known as *state variables* $\{\underline{\lambda_n} : n=1,\ldots N\}$. The multi-stage process associated with a multi-stage system is then defined as a sequence of *states* through which the dynamical system evolves. Each vector $\underline{\lambda_n}$ is a member of the state space Λ which is a set containing all possible system states. One of the limitations sometimes associated with a DP solution is the high dimensionality of the state space. Thus, if the number of state variables is large, then problems can arise with an excessively large information storage requirement and a high computational load.

(c) At each stage of the multi-stage process, a set of alternative actions or decisions may be taken in a manner which affects the system evolution. The sequence of decisions $\{\underline{x_n}(\underline{\lambda_n}) : n=1,\ldots,N\}$ is referred to as a *policy* and the vector $\underline{x_n}(\underline{\lambda_n})$[2] which is a member of this decision-making process is called a *decision variable.*

---

[2] *It should be noted that the decision* $\underline{x_n}(\underline{\lambda_n})$ *must depend on only the current state vector* $\underline{\lambda_n}$ *at any stage. This characteristic is known as the Markovian property.*

(d) A *stage transformation operator* $T(\cdot)$ expresses the output state $\underline{\lambda}_{n-1}$ not just as a function of the input state $\underline{\lambda}_n$[3] but also as a function of the decision variable $\underline{x}_n(\underline{\lambda}_n)$, i.e.

$$\underline{\lambda}_{n-1} = T(\underline{\lambda}_n, \underline{x}_n) \tag{4.17}$$

(e) A stage criterion function $c_n(\underline{\lambda}_n, \underline{x}_n)$ — a scalar function which measures the utility of a particular state vector $\underline{\lambda}_n$ and decision vector $\underline{x}_n(\underline{\lambda}_n)$.

(f) The aim of a decision-making optimisation process is to determine the optimal sequence of decisions to be taken at successive stages along the evolution of the system. The sequential decision process can be assessed according to a multi-variable criterion function[4] $h_n(\underline{\lambda}_1, \ldots, \underline{\lambda}_n \,;\, \underline{x}_1, \ldots, \underline{x}_n)$.

The task of a dynamic programming formulation is to determine the optimal value $g_n(\underline{\lambda}_n)$ of the criterion function which can be achieved for the state $\underline{\lambda}_n$. Thus,

$$g_n(\underline{\lambda}_n) = \underset{\underline{x}_n, \ldots, \underline{x}_1}{\text{opt}} \left\{ h_n(\underline{\lambda}_1, \ldots, \underline{\lambda}_n \,;\, \underline{x}_1, \ldots, \underline{x}_n) \right\} \tag{4.18}$$

where "opt" may denote maximisation or minimisation depending on the particular application. Now the total criterion function can also be represented as a combination of the individual stage criterion functions such as

$$h_n(\underline{\lambda}_1, \ldots, \underline{\lambda}_n \,;\, \underline{x}_1, \ldots, \underline{x}_n) = \sum_{i=1}^{n} c_i(\underline{\lambda}_i, \underline{x}_i) \tag{4.19}$$

---

[3] *Although this situation describes a backward dynamic programming formulation, it is also possible to express a problem in the reverse form where* $T(\cdot)$ *relates the current state and decision to the* <u>*next*</u> *state.*

[4] *This function* $h_n(\cdot)$ *is also known synonymously as a* <u>*cost function*</u>, <u>*return function*</u>, *or* <u>*objective function*</u>.

In this case, it is therefore assumed that the stage criterion functions are additive — although other means of combining the values $\{c_i(\underline{\lambda}_i,\underline{x}_i)\}$ are also possible. Thus, from equation (4.18),

$$g_n(\underline{\lambda}_n) = \underset{\underline{x}_n,\cdots,\underline{x}_1}{\text{opt}} \left\{ c_n(\underline{\lambda}_n,\underline{x}_n) + \sum_{i=1}^{n-1} c_i(\underline{\lambda}_i,\underline{x}_i) \right\} \tag{4.20}$$

It can then be shown [4.11] that this result is equivalent to

$$g_n(\underline{\lambda}_n) = \underset{\underline{x}_n}{\text{opt}} \left\{ c_n(\underline{\lambda}_n,\underline{x}_n) + \underset{\underline{x}_{n-1},\cdots,\underline{x}_1}{\text{opt}} \left[\sum_{i=1}^{n-1} c_i(\underline{\lambda}_i,\underline{x}_i) \right] \right\} \tag{4.21}$$

$$g_n(\underline{\lambda}_n) = \underset{\underline{x}_n}{\text{opt}} \left\{ c_n(\underline{\lambda}_n,\underline{x}_n) + g_{n-1}(\underline{\lambda}_{n-1}) \right\} \tag{4.22}$$

or, using equation (4.17),

$$g_n(\underline{\lambda}_n) = \underset{\underline{x}_n}{\text{opt}} \left\{ c_n(\underline{\lambda}_n,\underline{x}_n) + g_{n-1}\left[\mathrm{T}(\underline{\lambda}_n,\underline{x}_n) \right] \right\} \tag{4.23a}$$

For n=1,

$$g_1(\underline{\lambda}_1) = \underset{\underline{x}_1}{\text{opt}} \left\{ c_1(\underline{\lambda}_1,\underline{x}_1) \right\} \tag{4.23b}$$

The recurrence relation of equation (4.23) represents a fundamental result of Dynamic Programming. It elucidates the essential procedure required to transform a multi-dimensional optimisation problem into a sequence of simpler optimisation problems. Indeed, for each stage, equation (4.23) involves only a one dimensional optimisation.

This result is a direct consequence of *Bellman's Principle of Optimality.* According to this principle, it is sufficient to consider only the stage criterion function $c_n(\cdot)$ of the current stage and the optimal total criterion function $g_{n-1}(\cdot)$ of the previous stage in order to

determine the optimal total criterion function $g_n(\cdot)$ of the n-th stage. This is since $g_{n-1}(\cdot)$ reflects all past history of the process up to the (n-1)th stage. Specifically, Bellman states [4.12]

> *Principle of Optimality*
>
> *An optimal sequence of decisions has the property that whatever the initial state and decision are, the remaining decisions must constitute an optimal sequence of decisions with regard to the states resulting from the first decision.*

In this subsection, it has only been possible to deliver a very brief outline of the formulation of a dynamic programming problem and no attempt has been made to provide a rigorous mathematical treatment of the subject. More detailed discussions of the conditions to be met in order for the principle of optimality to be invoked and other issues of interest can be found in references [4.13],[4.10],[4.14].

4.3.3 Applications of Dynamic Programming

There are several areas of speech processing in which the Dynamic Programming formulation has found widespread application. One particular method which is of some relevance to the issue of speech spectral segmentation discussed in the next section is that of "*non-linear time alignment*" or "*dynamic time warping*".

In speech recognition, it is often necessary to compare reference templates against test vocalisations in order to recognise an utterance or verify a particular speaker. However, it is not a straightforward process to compare temporal patterns such as pitch, intensity and formant variations since even a single speaker is unlikely to be able to speak at precisely the

same rate for different repetitions of a particular phrase. Furthermore, time distortion is often non-uniform with vowels appearing to exhibit greater durational variations than consonants with changes in speaking rate [4.15].

One solution to this problem has been to non-linearly warp the time scale of the input pattern in order to obtain the best possible registration between the stored reference pattern and the measured pattern for a sample utterance. For instance, an input sequence may consist of N frames labelled $\{n=1,\ldots,N\}$ while the corresponding sequence for the reference pattern is $\{m=1,\ldots,M\}$ where M is not necessarily equal to N. A warping function $w(\cdot)$ of the form

$$m = w(n) \tag{4.24}$$

is then required which maps the sequence $\{n=1,\ldots,N\}$ onto $\{m=1,\ldots,M\}$ with a minimum total error. This error is calculated on the basis of a suitably chosen distance measure between input and reference frames. Dynamic Programming can, therefore, be employed to determine the time registration path which minimises the sum of frame-to-frame distances along its length.

To limit the degree of non-linearity of the warping function, continuity conditions can be imposed which are often of the form

$$\begin{aligned} w(n+1) - w(n) &= 0,1,2 \qquad \text{if } w(n) \neq w(n\text{-}1) \\ &= 1,2 \qquad \text{if } w(n) = w(n\text{-}1) \end{aligned} \tag{4.25}$$

These conditions imply that $w(\cdot)$ is monotonic increasing with a maximum slope of 2 and a minimum slope of $\frac{1}{2}$. Fig 4.3 illustrates a typical warping function for N=30, M=20. The parallelogram denotes the possible domain of the warping function introduced by the continuity constraints.

Dynamic time warping has been applied both for speaker verification [4.16] and for speech recognition [4.17]. It is also possible to extend this approach beyond isolated words to connected utterances using two levels of dynamic programming — one at a word level and one at the phrase level to find the optimal means of connecting the partial matches resulting from the first level [4.18]. One major advantage of the DP approach is that the three operations of segmentation for word boundary detection, non-linear time alignment and recognition are performed simultaneously [4.19]. Another possible application of dynamic programming lies in its use in Hidden Markov Modelling where DP can be employed to solve the "*evaluation*" problem mentioned in Chapter 2.

4.4 DYNAMIC PROGRAMMING APPLIED TO SPECTRAL SEGMENTATION

In order to obtain a set of equations describing the dynamic programming solution to the optimal segmentation of speech spectra, it is first necessary to establish an appropriate terminology.

Let

Total Number of Selective LP Regions	$= R$
Length of Region n	$= r_n$
Endpoint of Region n	$= t_n = \sum_{i=1}^{n} r_i$
Order of Predictor in Region n	$= p_n$
Sum of First n Predictor Orders	$= u_n = \sum_{i=1}^{n} p_n$
Error in Region n	$= E_n(t_n, u_n, r_n, p_n)$

As mentioned in section 4.3, the overall optimisation procedure will be a two stage process — once with respect to the segment lengths $\{r_i : i=1,\ldots,R\}$ and once with respect to the region predictor orders $\{p_i : i=1,\ldots,R\}$. In terms of the nomenclature discussed in the previous section, it is clear that the state vector is two-dimensional such that

$$\underline{\lambda_n} = \begin{pmatrix} t_n \\ u_n \end{pmatrix} \tag{4.26}$$

with $\{n=1,\ldots,R\}$ denoting the stage variable. There are also two decision variables r_n,p_n such that

$$\underline{x_n} = \begin{pmatrix} r_n \\ p_n \end{pmatrix} \tag{4.27}$$

From the above definitions, it is also possible to relate the $(n$-1)th stage state vector to the n-th stage state and decision vectors as follows:

$$\begin{pmatrix} t_{n-1} \\ u_{n-1} \end{pmatrix} = \begin{pmatrix} t_n \\ u_n \end{pmatrix} - \begin{pmatrix} r_n \\ p_n \end{pmatrix} \tag{4.28}$$

The stage transformation operator T(·) is, therefore, of the form

$$\mathrm{T}(\underline{\lambda_n},\underline{x_n}) \;=\; \underline{\lambda_{n-1}} \;=\; \underline{\lambda_n} - \underline{x_n} \tag{4.29}$$

Furthermore, the function $E_n(t_n,u_n,r_n,p_n)$ represents a certain form of error criterion which is a measure of the discrepancy between the Selective LP model in the nth region and the portion of the spectrum which lies within this range. The exact form of this error measure will be described in section 4.5. The error $E_n(t_n,u_n,r_n,p_n)$ actually corresponds to the additive stage criterion function $c_n(\underline{\lambda_n},\underline{x_n})$ of section 4.3.

In order to simplify the development of the segmentation equations, it will first be assumed that the R predictor orders are given and the optimisation process is, therefore, simply concerned with the calculation of the segment endpoints. This constraint can be removed subsequently to yield the overall specification of the solution.

4.4.1 Evaluation of Segment Breakpoints

For a speech data frame consisting of N samples, the length of the spectrum to the folding frequency (i.e. half the sampling frequency) will be $N/2$ and thus the requirement is to calculate the set of values $\{t_1,\ldots,t_R{=}N/2\}$ which produces the best match of the n Selective LP spectra to the $N/2$ samples in the signal spectrum. In order to realise this principle as an implementable DP algorithm, it is essential to define an objective function which can then be optimised with respect to the required segmentation parameters. Therefore, using the above terminology,

$g_n(t_n)$ = LEAST COST APPROXIMATION OF FIRST t_n SPEECH SPECTRAL SAMPLES BY SAMPLES OF THE FIRST n SELECTIVE LP SPECTRA.

This function can then be expressed in terms of a DP recurrence relation of the form discussed in section 4.3,

$$g_n(t_n) = \min_{r_n} \left\{ E_n(t_n,u_n,r_n,p_n) + g_{n\text{-}1}(t_n - r_n) \right\} \tag{4.30}$$

This equation may be justified as follows:

As r_n denotes the length of the nth segment, it is known that the first $t_n{-}r_n$ spectral samples are to be represented by $(n{-}1)$ Selective LP regions and $g_{n\text{-}1}(t_n{-}r_n)$ yields the lowest possible error for this condition. The error in the nth segment itself is then given by $E_n(t_n,u_n,r_n,p_n)$. The length r_n is then varied over its allowable range in order to determine the length which minimises the total error up to and including the nth segment. This value is finally assigned to $g_n(t_n)$ and can be used to iteratively compute $g_{n+1}(t_{n+1})$ etc.

Note

It is important to remember that another minimisation is also inherent in equation (4.30) since, as discussed in Chapter 3, the linear prediction process is itself such that the samples constituting the LP spectrum are obtained in such a manner as to minimise the sum over all frequencies of the ratio of the signal power to the model power spectrum.

The possible range of values for r_n, t_n at each stage must now be considered. The most stringent constraint on their variation is that, from reference [4.20] and in the light of experimental results, reasonable spectral matches are only obtained if the number of samples in a particular segment exceeds at least twice the predictor order in that segment. In accordance with this observation, the following constraints are obtained for the possible ranges of t_n, r_n:

1 SEGMENT:	$2p_1 \leq t_1 \leq N/2 - 2\sum_{i=2}^{R} p_i$	$r_1 = t_1$
2 SEGMENTS:	$2(p_1 + p_2) \leq t_2 \leq N/2 - 2\sum_{i=3}^{R} p_i$	$2p_2 \leq r_2 \leq t_2 - 2p_1$
$\vdots$	$\vdots$	$\vdots$
n SEGMENTS:	$2\sum_{i=1}^{n} p_i \leq t_n \leq N/2 - 2\sum_{i=n+1}^{R} p_i$	$2p_n \leq r_n \leq t_n - 2\sum_{i=1}^{n-1} p_i$

The Dynamic Time Warping procedure was outlined in the previous subsection and the manner in which this technique can be used for non-linear time alignment was described. This contrasts with the above method for the spectral segmentation of voiced speech. However, as in the case of Dynamic Time Warping, it is possible to depict the trajectory of system states in the form of a grid for the aforementioned spectral segmentation procedure.

This representation is shown in Fig. 4.4 which also illustrates a typical path constituted of the optimal breakpoints. The parallelogram denotes the way in which the above

conditions restrict the possible set of values $\{t_i : i=1,\ldots,R\}$. The boundary illustrated in Fig. 4.4 for the domain of possible paths assumes that the predictor orders are all equal (i.e. $p_1=p_2=\ldots=p_R$) and, if this is not the case, then this domain will simply be characterised by a more complicated boundary. It should also be noted that these restrictions on r_n,t_n imply that any trajectory must be monotonic increasing with a minimum slope of $2p_n$ at any stage n.

With the introduction of the above constraints, equation (4.30) becomes:

$$g_n(t_n) = \min_{2p_n \le r_n \le t_n - 2\sum_{i=1}^{n-1} p_i} \left\{ E_n(t_n,u_n,r_n,p_n) + g_{n-1}(t_n - r_n) \right\} \qquad (4.31a)$$

$$2\sum_{i=1}^{n} p_i \le t_n \le N/2 - 2\sum_{i=n+1}^{R} p_i$$

Initial condition:

$$g_1(t_1) = E_1(t_1,u_1,t_1,u_1) \qquad 2u_1 \le t_1 \le N/2 - 2\sum_{i=2}^{R} p_i \qquad (4.31b)$$

Starting with equation (4.31b) and subsequently using (4.31a), the number of regions is increased such that $n=1,\ldots,R$ where R is the required number of regions. The final result is then given by $g_R(N/2)$ which denotes the minimum error when all $N/2$ samples of the signal spectrum are represented by R Selective LP regions.

However, this process does <u>not</u> automatically yield the optimal endpoints for each segment and these values must be evaluated by the process of "*Backtracking*". For this to be possible, the value of r_n at each stage n which produces the minimum error $g_n(t_n)$ (i.e. the optimising value r_n^*) as in the following expression must be evaluated at each stage.

$$r_n^* = \operatorname*{argmin}_{2p_n \le r_n \le t_n - 2\sum_{i=1}^{n-1} p_i} \left\{ E_n(t_n,u_n,r_n,p_n) + g_{n-1}(t_n - r_n) \right\} \qquad (4.32)$$

$$2\sum_{i=1}^{n} p_i \le t_n \le N/2 - 2\sum_{i=n+1}^{R} p_i$$

At the end of R stages, the R optimal segment lengths are, therefore, available and can be employed to calculate the optimal segment endpoints $\{t_n^* : n=1,\ldots,R\}$ recursively as follows starting at the last sample:

Initial Condition: $t_R^* = N/2$

$$t_{n-1}^* = t_n^* - r_n^* \qquad n=R, R-1,\ldots,2 \tag{4.33}$$

At this point, the set of optimal break frequencies is available and the sequence of predictor orders are known a priori and thus the Selective LP frequency response consisting of R regions can be reconstructed as in section 4.2 and compared to the real speech spectrum.

4.4.2 Optimal Assignment of Predictor Orders

The concept of optimality in relation to the choice of segment endpoints was discussed in the last subsection, but the implication of this term in the calculation of region predictor orders is less apparent. This is due to the fact that it is meaningless to discuss the 'optimal' number of predictor coefficients to employ within any one LP region since it was found in Chapter 3 that the fit of the model improves indefinitely as the predictor order increases.

However, it is possible to determine the optimal allocation of coefficients between the various Selective LP segments in order to achieve the best match between the real and model spectra such that the sum of the individual predictor orders over all the regions is a constant. It is this definition of optimality which shall be considered in this section.

Therefore, using the terminology outlined above, a total predictor order u_R is to be distributed amongst the R Selective LP models such that

$$u_R = \sum_{i=1}^{R} p_i \tag{4.34}$$

It is, therefore, necessary to incorporate this requirement into the procedure described in the previous section for the calculation of the segment endpoints. In order to achieve this result, a new objective function $f_n(t_n,u_n)$ is defined as follows:

$f_n(t_n,u_n)$ = LEAST COST APPROXIMATION OF FIRST t_n SPEECH SPECTRAL SAMPLES BY FIRST n SELECTIVE LP REGIONS WITH u_n AS THE SUM OF THE FIRST n PREDICTOR ORDERS.

The constraints on the variation of t_n,r_n in the optimisation process were outlined in the last section. The corresponding limitations for u_n,p_n arise from the consideration that, in Selective LP, the minimum number of coefficients normally assigned to any one region is usually 2. This requirement ensures that there is likely to be at least 1 complex pole pair in the model spectrum. Any other value can, of course, also be specified and, in general, there will be a minimum possible number of coefficients p_{min} as follows:

$$p_i \geq p_{min}\,, \qquad \forall i \in \{1,\ldots,R\} \tag{4.35}$$

The following constraints are, therefore, obtained for the parameters u_n,p_n:

1 SEGMENT:	$p_{min} \leq u_1 \leq u_R - (R-1)p_{min}$	$p_1 = u_1$
2 SEGMENTS:	$2p_{min} \leq u_2 \leq u_R - (R-2)p_{min}$	$p_{min} \leq p_2 \leq u_2 - p_{min}$
$\vdots$	$\vdots$	$\vdots$
n SEGMENTS:	$np_{min} \leq u_n \leq u_R - (R-n)p_{min}$	$p_{min} \leq p_n \leq u_n - (n-1)p_{min}$

These restrictions can again be illustrated on the domain of possible values for $\{u_i : i=1,\ldots,R\}$ as shown in Fig. 4.5. A typical path is also shown and again any trajectory must be a monotonic increasing function of the stage variable n with a minimum slope of p_{min} at any particular stage.

The next step involves the incorporation of these results into the recurrence relations developed for the calculation of segment breakpoints in the previous section. However, this procedure involved fixed values for the predictor values whereas the parameters $\{p_n,u_n : n=1,\ldots,R\}$ are now treated as variables and this modification produces the following changes in the constraints on r_n,t_n:

$$2u_n \leq t_n \leq N/2 - 2(u_R-u_n) \tag{4.36a}$$

$$2p_n \leq r_n \leq t_n - 2(u_n-p_n) \tag{4.36b}$$

Thus the DP formulation for the optimal segmentation of speech spectra is as follows:

For 1 region:

$$f_1(t_1,u_1) = E_1(t_1,u_1,t_1,u_1) \qquad \text{for} \quad p_{min} \leq u_1 \leq u_R - (R-1)p_{min} \qquad 2u_1 \leq t_1 \leq N/2 - 2(u_R-u_1) \tag{4.37}$$

For subsequent regions:

$$f_n(t_n,u_n) = \min_{p_{min} \leq p_n \leq u_n-(n-1)p_{min}} \; \min_{2p_n \leq r_n \leq t_n-2(u_n-p_n)} \left\{ E_n(t_n,u_n,r_n,p_n) + f_{n-1}(t_n - r_n,\, u_n - p_n) \right\} \tag{4.38}$$

subject to the constraints:

$$np_{min} \leq u_n \leq u_R - (R-n)p_{min}$$

$$\& \qquad 2u_n \leq t_n \leq N/2 - 2(u_R-u_n)$$

As before, these equations are employed recursively to produce the final result $f_R(N/2,u_R)$. The value of p_n which produces the minimum error for each u_n can again be stored to yield the optimal set of predictor orders $\{p_n^* : n=1,\ldots,R\}$. Therefore, equations (4.37),(4.38) specify the complete solution to the problem of choosing the breakpoints and the order of the predictor in each Selective LP region optimally to minimise the error between the real and model spectra. The entire procedure for speech spectral segmentation is illustrated in the form of a flow diagram in Fig. 4.6.

One point of interest which can now be covered concerns the Principle of Optimality associated with dynamic programming solutions to the optimisation of multi-stage decision processes. Although the general principle was outlined in section 4.3.2, it is now possible to describe the manner in which this principle manifests itself in the particular application of DP to the segmentation of frequency spectra in the following terms:

> *Let the optimal segmentation policy for a given speech spectrum into R regions with an overall predictor order u_R be such that the sets of optimal region lengths and individual segment predictor orders are given by $\{r_1^*, r_2^*, \ldots, r_R^*\}$ and $\{p_1^*, p_2^*, \ldots, p_R^*\}$ respectively. This segmentation policy then includes, as a portion of itself, the optimal policy for the segmentation of the first t_n points into n regions using a total predictor order u_n and is explicitly given by the subsets $\{r_1^*, r_2^*, \ldots, r_n^*\}$ and $\{p_1^*, p_2^*, \ldots, p_n^*\}$.*

4.5 CHOICE OF ERROR CRITERION

4.5.1 Spectral Distortion Measures for Speech Processing

A distortion measure is an assignment of a non-negative number to an input/output pair of a system. The distortion between an input (or original) and an output (or reproduction) represents the cost or distortion resulting when that input is represented by that

output. Such measures have a wide variety of applications in the design and comparison of systems [4.21].

Certain goals must be met by a prospective distortion measure if it is to be of use in practical systems. One requirement is that the measure must be subjectively meaningful in the sense that small and large distortions correspond respectively to 'good' and 'bad' subjective quality. While subjective testing must provide the ultimate means of differentiating between the terms good and bad, analytical tools can be used to gain an insight into the differences and similarities between various types of measures. Other necessary properties include mathematical tractability and ease of computation in the sense that it must be possible to efficiently compute actual distortions resulting in a real system.

The most common form of distortion measure is the traditional squared error or error power or error energy. However, this criterion suffers from the disadvantage that it does not always appear to be subjectively meaningful in the sense that a large squared error distortion does not necessarily imply poor speech quality. For example, the fricative /ʃ/ as in *shut* is essentially a white noise process and thus any representative waveform will sound the same even though it may exhibit a large squared error distortion.

One solution to this problem lies in the use of a *spectral* distortion measure which is a function of two spectral densities, $P(\omega_m)$ and $\hat{P}(\omega_m)$ for example, and assigns a non-negative number $d(P,\hat{P})$ to represent the distortion in using $\hat{P}(\omega_m)$ to represent $P(\omega_m)$. The following subsections outline some of these measures which have been employed in the DP solution to the optimal speech spectral segmentation problem developed in the last section.

As the calculation of an error function forms an integral part of the dynamic programming algorithm, one of the primary considerations in the choice of an error measure is its simplicity and ease of implementation. It should also be noted that these distortion measures do not constitute an exhaustive list and a more detailed background discussion and

reference information is provided by Gray and Markel [4.22].

4.5.2 A Squared Error Measure

Many of the distortion measures developed in the subsequent subsections are most concisely expressed in terms of L_k norms. For $k \geq 1$, the L_k norm of a complex-valued function g is defined as

$$|g|_k = \left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} |g(\theta)|^k d\theta \right\}^{\frac{1}{k}} \tag{4.39}$$

In the case of a spectral distortion measure, $g(\theta)$ may denote a spectral density over the range $-\pi$ to π of the normalized frequencies θ. Furthermore, for a discrete power spectral density $\left\{ P(\omega_m) = P\left(\frac{2\pi m}{N}\right) : m=0,\ldots,N\text{-}1 \right\}$, this norm may be expressed as:

$$|P|_k = \left\{ \frac{1}{N} \sum_{m=0}^{N-1} |P(\omega_m)|^k \right\}^{\frac{1}{k}} \tag{4.40}$$

One of the oldest distortion measures proposed for speech is formed by the L_k norm of the difference of log spectra:

$$d_{\ln k}(P,\hat{P}) = |\ln P - \ln \hat{P}|_k = \left| \ln \left(\frac{P}{\hat{P}} \right) \right|_k \tag{4.41}$$

The most common choices for k are 1,2 and ∞ — yielding mean absolute, root mean square and maximum deviation measures respectively. It should also be noted that the functions $d_{\ln k}(P,\hat{P})$ represent *metrics* or *distances* in the sense that they satisfy the symmetry requirement

$$d_{\ln k}(P,\hat{P}) \;=\; d_{\ln k}(\hat{P},P) \tag{4.42}$$

and a triangle inequality:

$$d_{\ln k}(f,g) \;\leq\; d_{\ln k}(f,h) \;+\; d_{\ln k}(h,g) \tag{4.43}$$

Gray and Markel [4.22] have shown that the choice of the value of k in the L_k norm is not critical for speech data as there is a strong correlation between all of the L_k measures for widely varying k values. However, largely due to its analytical tractability, the L_2 norm is the most popular in speech processing techniques [4.23],[4.24]. One particular application which has already been mentioned is the method of Spectral Analysis-by-Synthesis in section 3.4. This technique uses the L_2 norm in order to derive model parameters which minimise this distortion measure between real and model spectra.

It was, therefore, decided to utilise the $d_{\ln 2}(\cdot,\cdot)$ measure in the DP procedure outlined in section 4.4. This step requires the error $E_n(t_n,u_n,r_n,p_n)$ in the n-th Selective LP region to be expressed in terms of the signal and Selective LP model spectral densities within that region. Thus, the range of frequencies which are of interest are those associated with the n-th region and are explicitly given by $\left\{\omega_m=\frac{2\pi(m\text{-}1)}{N} : m=t_n-r_n+1,\ldots,t_n\right\}$.

The extent of the signal power spectrum which is to be considered is then $\left\{P(\omega_m) : m=t_n-r_n+1,\ldots,t_n\right\}$ and that for the Selective LP model will be represented by $\left\{\hat{P}(\omega_m\,;t_n,u_n,r_n,p_n) : m=t_n-r_n+1,\ldots,t_n\right\}$[5]. Hence, the entire Selective LP spectrum is characterised by a combination of these responses within the individual regions, i.e.

---

[5] *This notation reflects the fact that the model power spectrum within the n-th region depends on each of the parameters* t_n,u_n,r_n,p_n.

$$\left\{\hat{P}(\omega_m; t_1,u_1,t_1,u_1) : m=1,\ldots,t_1\right\},$$
$$\left\{\hat{P}(\omega_m; t_2,u_2,r_2,p_2) : m=t_2-r_2+1,\ldots,t_2\right\},$$
$$\vdots$$
$$\left\{\hat{P}(\omega_m; t_R,u_R,r_R,p_R) : m=t_R-r_R+1,\ldots,t_R=N/2\right\}.$$

According to these definitions, an error measure based on the L_2 norm in one particular Selective LP region is then expressed as:

$$d'_{\ln 2}(P,\hat{P}) = \sqrt{\sum_{m=t_n-r_n+1}^{t_n} \left[\ln P(\omega_m) - \ln \hat{P}(\omega_m; t_n,u_n,r_n,p_n)\right]^2} \qquad (4.44)$$

The actual definition used for $E_n(t_n,u_n,r_n,p_n)$ is a slightly modified (but essentially equivalent) version of this result and is based on spectral densities calculated in dB as follows:

$$E_n^{sq}(t_n,u_n,r_n,p_n) = \sum_{m=t_n-r_n+1}^{t_n} \left[10 \log P(\omega_m) - 10 \log \hat{P}(\omega_m; t_n,u_n,r_n,p_n)\right]^2 \qquad (4.45)$$

This form of error criterion was introduced into the procedure for optimal speech spectral segmentation developed in the last section. The vocalisation under investigation was the sound /OW/ as in *bought* with the length (N) of the particular frame being 256 samples and thus the length of the corresponding spectrum to the folding frequency was given by $N/2 = 128$. Optimal 2 and 3 region segmentations were evaluated with an overall predictor order (u_R) of 12 and, as discussed in the previous section, the minimum allowable predictor order (p_{min}) in any region was taken as 2.

These results are shown in Fig. 4.7 (a) and (b) which shows the underlying speech spectrum with the optimal Selective LP models superimposed upon it. A 12th order standard (single region) LP model spectrum is also shown for comparison since the order of this predictor is identical to the overall order of the Selective LP models. In addition to these

curves, the optimal region end frequencies and predictor orders are also illustrated along with the corresponding region errors and an overall squared error which is simply the sum of the individual errors. Hence, this overall error is precisely the optimal objective function $f_R(N/2,u_R)$ calculated in the DP procedure.

Both figures indicate a considerably improved fit between signal and model spectra over the single region LP model but a more quantitative assessment of this improvement is obtained by comparing the total squared errors indicated in Fig. 4.7 with the equivalent value for the single LP region given below:

SQUARED ERROR FOR SINGLE 12th ORDER LP REGION = 10 075 units

The reductions are, therefore, 35.7% and 40.2% on this figure while employing the same overall predictor order.

4.5.3 The Itakura-Saito Distortion Measure

In their study of the maximum likelihood formulation of linear prediction mentioned in section 3.3.1, Itakura and Saito [4.25] made use of the following measure:

$$d_{is}(P,\hat{P}) = \left| \frac{P}{\hat{P}} - \ln\left(\frac{P}{\hat{P}}\right) - 1 \right|_1 \tag{4.46}$$

$$= \frac{1}{N}\sum_{m=0}^{N-1}\left[\frac{P(\omega_m)}{\hat{P}(\omega_m)} - \ln\left(\frac{P(\omega_m)}{\hat{P}(\omega_m)}\right) - 1\right] \tag{4.47}$$

It should be noted that the absolute value of the summand in this expression is not required since this particular summand is always positive. There are several important properties associated with this criterion. For instance, from the nature of the summand in equation (4.47), it is clear that

$$\text{if} \qquad R\left[P(\omega_m),\hat{P}(\omega_m)\right] = \frac{P(\omega_m)}{\hat{P}(\omega_m)} - \ln\left(\frac{P(\omega_m)}{\hat{P}(\omega_m)}\right) - 1 \tag{4.48}$$

$$\text{then} \qquad R\left[P(\omega_m),\hat{P}(\omega_m)\right] \simeq \frac{P(\omega_m)}{\hat{P}(\omega_m)} \qquad \text{when } P(\omega_m) \gg \hat{P}(\omega_m) \tag{4.49a}$$

$$\text{and} \qquad R\left[P(\omega_m),\hat{P}(\omega_m)\right] \simeq -\ln\left(\frac{P(\omega_m)}{\hat{P}(\omega_m)}\right) \qquad \text{when } P(\omega_m) \ll \hat{P}(\omega_m) \tag{4.49b}$$

Therefore, positive errors where the signal spectrum $P(\omega_m)$ is greater than the LP model spectrum $\hat{P}(\omega_m)$ play a much more important role in this distance measure than negative values. For instance, when $P(\omega_m)$ is 10 dB higher than $\hat{P}(\omega_m)$ at any one frequency, then $R\left[P(\omega_m),\hat{P}(\omega_m)\right] = 6.70$ while this value drops to $R\left[P(\omega_m),\hat{P}(\omega_m)\right] = 1.40$ when $P(\omega_m)$ is 10 dB lower than $\hat{P}(\omega_m)$.

This result is characteristic of the linear predictive spectrum as discussed in Chapter 3 since it will tend to weight poles more heavily than zeros in the speech spectral density [4.4]. In fact, the problem of choosing the predictor coefficients $\{\alpha_i : i=1,\ldots,p\}$ and model gain (G) in order to minimise the Itakura-Saito distance between real and LP model frequency responses results in exactly the same Toeplitz minimisation problem as the autocorrelation method discussed in Chapter 3 [4.26].

Another characteristic of the Itakura-Saito measure arises from the expansion

$$\begin{aligned}\frac{P(\omega_m)}{\hat{P}(\omega_m)} &= \exp\left[\ln\left(\frac{P(\omega_m)}{\hat{P}(\omega_m)}\right)\right] \\ &= 1 + \ln\left(\frac{P(\omega_m)}{\hat{P}(\omega_m)}\right) + \tfrac{1}{2}\left[\ln\left(\frac{P(\omega_m)}{\hat{P}(\omega_m)}\right)\right]^2 + \ldots\end{aligned} \tag{4.50}$$

$$\text{Thus,} \qquad R\left[P(\omega_m),\hat{P}(\omega_m)\right] \simeq \tfrac{1}{2}\left[\ln\left(\frac{P(\omega_m)}{\hat{P}(\omega_m)}\right)\right]^2 \qquad \text{if } \frac{P(\omega_m)}{\hat{P}(\omega_m)} \simeq 1, \ \forall\omega_m \tag{4.51}$$

and $$d_{is}(P,\hat{P}) \simeq \frac{1}{2} \cdot \frac{1}{N} \sum_{m=0}^{N-1} \left[\ln P(\omega_m) - \ln \hat{P}(\omega_m) \right]^2 \tag{4.52}$$

or, from the discussion of the measure based on the L_2 norm developed in the previous section,

$$d_{is}(P,\hat{P}) \simeq \frac{1}{2} \left(d_{\ln 2}(P,\hat{P}) \right)^2 \tag{4.53}$$

Thus, for small distortions, d_{is} becomes approximately proportional to the mean-square log spectral distortion.

In order to employ the Itakura-Saito distance measure in the segmentation algorithm, an error measure of the form shown below was defined:

$$E_n^{is}(t_n,u_n,r_n,p_n) = \sum_{m=t_n-r_n+1}^{t_n} R\left[P(\omega_m), \hat{P}(\omega_m; t_n,u_n,r_n,p_n) \right] \tag{4.54}$$

Optimal 2 and 3 region segmentations were again computed for the same speech frame as used for the squared error measure. An overall predictor order of 12 and a value of $p_{min}=2$ were also chosen. The results are shown in Fig. 4.8 (a),(b) with the same notation as for Fig. 4.7. Qualitatively, there is again a great improvement in the match between the Selective LP models and the original speech spectrum over that for the 12th order standard LP frequency response. The actual values for the total errors are 33.5% and 37.2% lower than the equivalent error for the single region LP model presented below.

ERROR BASED ON THE ITAKURA-SAITO
MEASURE FOR SINGLE 12th ORDER LP REGION = 134.40 units

There are a few disadvantages associated with the use of this form of error measure and one of these drawbacks can be illustrated by noting that it can be shown [4.4] that the total squared prediction error ($E^{(p)}$) resulting from a p-th order LP model is related to the model power spectral density as follows:

$$\ln (E^{(p)}) = \frac{1}{N} \sum_{m=0}^{N-1} \ln \hat{P}(\omega_m) \tag{4.55}$$

In addition, as the predictor order p approaches infinity, the model spectrum approaches the signal spectrum and thus

$$\ln (E^{(\infty)}) = \frac{1}{N} \sum_{m=0}^{N-1} \ln P(\omega_m) \tag{4.56}$$

Therefore, using equations (4.55) and (4.56) in the definition of $d_{is}(P,\hat{P})$,

$$d_{is}(P,\hat{P}) = \frac{1}{N} \sum_{m=0}^{N-1} \left[\frac{P(\omega_m)}{\hat{P}(\omega_m)} - 1 \right] - \ln \left(\frac{E^{(\infty)}}{E^{(p)}} \right) \tag{4.57}$$

This result illustrates the fact that, for an Itakura-Saito distortion measure, discrepancies which occur when $P(\omega_m) > \hat{P}(\omega_m)$ tend to cancel those when $P(\omega_m) < \hat{P}(\omega_m)$ [4.3]. This is in contrast to the squared error measure where all errors are additive. Along with this property of the Itakura-Saito measure, Viswanathan and Makhoul [4.24] also suggest that an error measure based on the L_2 norm relates more closely to actual speech perception since it was found to produce better results in terms of speech quality in their experiments on the quantisation of reflection coefficients.

4.5.4 A Weighted Error Measure

The use of a squared error measure in the optimal spectral segmentation scheme was found to produce good spectral estimates of the original speech signal. However, one of the primary motivations of employing a Selective LP model was the use of segments with a higher predictor order in spectral regions where the signal energy was greatest and a lower number of coefficients in areas of lower signal energy. Neither the squared error nor the Itakura-Saito measures fulfil this requirement since the error at any particular frequency is dependent on only the discrepancy between the real and model spectra — irrespective of the actual signal energy.

One solution to this problem lies in the use of a "*Weighted (squared) Error Criterion*" of the form

$$E_n^{wt}(t_n,u_n,r_n,p_n) = \sum_{m=t_n-r_n+1}^{t_n} B\Big(P(\omega_m)\Big)\Big[B\Big(P(\omega_m)\Big) - B\Big(\hat{P}(\omega_m;t_n,u_n,r_n,p_n)\Big)\Big]^2 \qquad (4.58)$$

where $B(r) = 10\log(r)$ represents the log spectrum in dB. In this case, the squared difference is actually weighted by multiplication with the actual level (in dB) of the signal spectrum. This error measure was incorporated in the DP procedure and yielded results as shown in Fig. 4.9 (a),(b) for a total number of regions given by $R=2,3$. The same speech data and overall predictor order were employed as for the two other error criteria. There is again a significant improvement over the 12th order linear predictive model spectrum which produces the following value for the total weighted error.

WEIGHTED ERROR FOR SINGLE 12th ORDER LP REGION = 196 775 units

Therefore, the 2 and 3 region segmentations represent 41.8% and 43.5% improvements on this figure while employing the same overall predictor order. The various results for the 3 types of error criterion considered in this section are summarised in Tables 4.1 and 4.2.

4.6 A SUBOPTIMAL SEGMENTATION TECHNIQUE

The segmentation scheme of the previous section has resulted in partitions of the speech spectrum which are optimal with respect to the choice of segment breakpoints and predictor orders. However, the entire process is computationally intensive and the load required for the calculation of segmentations above 4 regions for an overall predictor order of around 12 can be excessive in some cases. A compromise solution is, therefore, proposed in this section which results in a suboptimal procedure for speech spectral segmentation. This process will not necessarily yield the global minimum error calculated by the optimal algorithm, but does fulfil the requirement of relaxing the computational requirements considerably.

The approach basically relies on determining which region at a particular stage in the algorithm is associated with the largest error and splitting only this segment into two regions with respect to an optimal solution of both the breakpoint and the two individual predictor orders. Thus, the overall error reduces monotonically to the point where the required number of regions is reached. These requirements are incorporated in the following specification of the suboptimal speech spectral segmentation procedure.

Consider the case where m segments and their corresponding parameters have been evaluated and let the superscript (m) denote the m-th iteration in the process. In particular, $E_n^{\mathrm{wt}}(t_n,u_n,r_n,p_n)^{(m)}$ represents the weighted[6] error in the n-th region at the m-th iteration.

STEP 1

Find region q which is associated with largest error. As this region will be split into two new Selective LP regions, the predictor order in the original segment must be at least $2p_{min}$.

---

[6] *The weighted error will be used in this procedure although the other error criteria can also be employed to produce similar results.*

$$\text{Thus,} \qquad q = \underset{1 \le n \le m \text{ s.t. } p_n^{(m)} \ge 2p_{min}}{\text{argmax}} \left\{ E_n^{wt}(t_n, u_n, r_n, p_n)^{(m)} \right\} \tag{4.59}$$

STEP 2

Determine the optimal breakpoint and predictor orders for the partition of the q-th segment. If the length of the first portion of the partition is r' and the predictor order is p' then the corresponding values for the second portion are $r_q^{(m)} - r'$, $p_q^{(m)} - p'$. This first region also covers the semi-closed interval $(t_q^{(m)} - r_q^{(m)},\ t_q^{(m)} - r_q^{(m)} + r']$ while the second region is given by $(t_q^{(m)} - r_q^{(m)} + r',\ t_q^{(m)}]$.

Hence, the errors in the two segments can be represented as:

$$E^{wt}(t_q^{(m)} - r_q^{(m)} + r',\ u_q^{(m)} - p_q^{(m)} + p',\ r',\ p') \quad \& \quad E^{wt}(t_q^{(m)},\ u_q^{(m)},\ r_q^{(m)} - r',\ p_q^{(m)} - p').$$

The task is then to allow each of r', p' to vary across their allowable ranges and evaluate the combination which produces the lowest sum (E^*) of these two segment errors. The constraints on the variation of r', p' are identical to those for the optimal segmentation, viz. the minimum number of predictor coefficients in either of the two segments is p_{min} and the length of either region must be at least twice the corresponding predictor order. In accordance with these observations, the following expression is used to calculate the minimum error E^*:

$$E^* = \min_{p_{min} \le p' \le p_q^{(m)} - p_{min}} \ \min_{2p' \le r' \le r_q^{(m)} - 2(p_q^{(m)} - p')} \left\{ E^{wt}(t_q^{(m)} - r_q^{(m)} + r',\ u_q^{(m)} - p_q^{(m)} + p',\ r',\ p') \ + \ E^{wt}(t_q^{(m)},\ u_q^{(m)},\ r_q^{(m)} - r',\ p_q^{(m)} - p') \right\} \tag{4.60}$$

Let r^*, p^* be the values of r', p' which correspond to the minimum error E^*.

STEP 3

Insert the two new regions in the correct position within the sequence of Selective LP regions.

Errors:

$$E_n^{wt}(t_n,u_n,r_n,p_n)^{(m+1)} = E_n^{wt}(t_n,u_n,r_n,p_n)^{(m)} \qquad n=1,\ldots,q-1$$

$$E_{n+1}^{wt}(t_{n+1},u_{n+1},r_{n+1},p_{n+1})^{(m+1)} = E_n^{wt}(t_n,u_n,r_n,p_n)^{(m)} \qquad n=q+1,\ldots,m$$

$$E_q^{wt}(t_q,u_q,r_q,p_q)^{(m+1)} = E^{wt}(t_q^{(m)}-r_q^{(m)}+r^*,\ u_q^{(m)}-p_q^{(m)}+p^*,\ r^*,\ p^*)$$

$$E_{q+1}^{wt}(t_{q+1},u_{q+1},r_{q+1},p_{q+1})^{(m+1)} = E^{wt}(t_q^{(m)},\ u_q^{(m)},\ r_q^{(m)}-r^*,\ p_q^{(m)}-p^*)$$

Other Parameters:

$$t_n^{(m+1)}=t_n^{(m)} \quad u_n^{(m+1)}=u_n^{(m)} \quad r_n^{(m+1)}=r_n^{(m)} \quad p_n^{(m+1)}=p_n^{(m)} \qquad n=1,\ldots,q-1$$

$$t_{n+1}^{(m+1)}=t_n^{(m)} \quad u_{n+1}^{(m+1)}=u_n^{(m)} \quad r_{n+1}^{(m+1)}=r_n^{(m)} \quad p_{n+1}^{(m+1)}=p_n^{(m)} \qquad n=q+1,\ldots,m$$

$$t_q^{(m+1)}=t_q^{(m)}-r_q^{(m)}+r^* \quad u_q^{(m+1)}=u_q^{(m)}-p_q^{(m)}+p^* \quad r_q^{(m+1)}=r^* \quad p_n^{(m+1)}=p^*$$

$$t_{q+1}^{(m+1)}=t_q^{(m)} \quad u_{q+1}^{(m+1)}=u_q^{(m)} \quad r_{q+1}^{(m+1)}=r_q^{(m)}-r^* \quad p_{q+1}^{(m+1)}=p_q^{(m)}-p^*$$

STEP 4

Let $m \longleftarrow m+1$. If $m=R$ then stop.

Otherwise return to Step 1.

This procedure was used to segment the same speech utterance as employed in the previous section for an initial order of 12 and the manner in which the breakpoints and predictor orders in each region varied with an increasing number of segments up to six is shown in Fig. 4.10.

Three cases from this set are particularly significant. The first two concern the 2 and 3 region suboptimal segmentations for which the parameters are exactly the same as for the 2 and 3 region optimal segmentations in section 4.5 with a weighted error criterion. Thus, for this particular case, the suboptimal case does, in fact, produce the optimal solution and was

also found not to deviate significantly from the optimal results in other cases.

In addition, the result of the 6 region suboptimal segmentation is shown as a frequency response in Fig. 4.11. As well as providing a good spectral estimate, this Selective LP model consists solely of second order regions and seems to segment the speech spectrum into its individual formants. This issue is developed in considerable detail in the next chapter with the use of a modification of the optimal segmentation scheme rather than the suboptimal procedure described in this section.

4.7 CONCLUSIONS

This chapter has largely been devoted to the development of a technique for the segmentation of a speech spectral density into several regions which are modelled by Selective Linear Prediction spectra. The procedure employs Dynamic Programming to optimise the choice of break frequencies between regions and the allocation of predictor coefficients to each of the segments given an overall predictor order. This optimisation is carried out such that the discrepancy between the signal spectrum and the combined Selective LP responses is minimised in relation to a variety of error measures. Details of the segmentation procedure are also provided by references [4.27],[4.28][4.29].

The first step towards achieving the required goal was the introduction of a Selective LP modelling scheme for discrete spectra whereby an arbitrary range of speech spectral samples could be represented by a Selective LP model with a given order. The dynamic programming procedure for spectral segmentation was then set within the wider framework of a general background of the DP technique. The final stage then involved the incorporation of an error measure which was able to quantify the mismatch between the signal and model spectra within the given Selective LP region.

Several error criteria were considered and each of them was found to result in considerably improved spectral estimates of the original signal when compared to standard LP modelling. However, as pointed out by Gray et al [4.26], it is particularly difficult to comment on the complex and open problem of characterising which distortion measures best represent the subjective quality of particular psychophysical characteristics and features of speech. The choice of a particular error measure will largely depend on the prospective application and whether the mathematical properties of this measure are thought to suit the procedure in which it is employed.

In the case of optimal speech spectral segmentation, all three of the error criteria under consideration produced fairly similar results and indeed, this outcome may have been expected since it was found in section 4.5 that several of the error criteria are actually interrelated — as exemplified by the proportional relationship between the Itakura-Saito distance measure and the squared error criterion for small distortions. In any case, one major advantage of the Dynamic Programming formulation is the ease with which a change in the required form of the error criterion can be carried out in the segmentation algorithm. This is in contrast to classical methods where such a modification could involve a complete reappraisal of the optimisation methodology.

It is also important to bear in mind that the error measure employed in the optimal segmentation procedure is only responsible for calculating the optimal breakpoints and predictor orders. The actual model spectrum within each region is still a linear predictive model which, as described in Chapter 3, minimises the ratio of the signal power to the LP model power spectrum. Another point of interest concerning the segmentation results is that the greatest improvement was usually detected in the 2 region segmentation, with little reduction in the overall error for a larger number of regions. This suggests that it may be adequate to provide an optimal spectral partition into 2 segments as a signal spectral estimate for the majority of applications. One other means of decreasing the computational

requirements of the optimal spectral segmentation scheme lies in the use of the suboptimal procedure developed in section 4.6. This process was found to produce similar results to the optimal technique at a considerably lower computational cost.

4.8 REFERENCES

[4.1] BRISTOW, G., "Electronic Speech Recognition", *Collins, London*, 1986.

[4.2] MAKHOUL, J., "Selective Linear Predictive Spectral Matching", *Presented at the 86th Meeting of the Acoustical Society of America, Los Angeles, California, Oct.* 1973. (*Abstract included in the J. Acoust. Soc. Amer., Vol.* 55, *No.* 2, *Feb.* 1974, *pg* 397).

[4.3] MAKHOUL, J., "Spectral Linear Prediction : Properties and Applications", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP-*23, *No.* 3, *June* 1975, *pp* 283-296.

[4.4] MARKEL, J.D., GRAY, A.H., Jr., "Linear Prediction of Speech", *Springer-Verlag, Berlin-Heidelberg-New York*, 1976.

[4.5] MIZOGUCHI, R., YANAGIDA, M., KAKUSHO, O., "Speech Analysis by Selective Linear Prediction in the Time Domain", *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Proc., Paris, France, May* 1982, *Vol.* 3, *pp* 1573-1576.

[4.6] MIYOSHI, Y., et al, "Analysis of Speech Signals of Short Pitch Period by a Sample-Selective Linear Prediction", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP-*35, *No.* 9, *Sept.* 1987, *pp* 1233-1240.

[4.7] DREYFUS, S.E., "Dynamic Programming and the Calculus of Variations", *Academic Press*, 1965.

[4.8] SCHARF, L.L., ELLIOTT, H. "Aspects of Dynamic Programming in Signal and Image Processing", *IEEE Trans. Automatic Control, Vol.* 26, *No.* 5, *Oct.* 1981, *pp* 1018-1029.

[4.9] HOWARD, R.A., "Dynamic Programming and Markov Processes", *Wiley, New York*, 1960.

[4.10] COOPER, L., COOPER, M.W., "Introduction to Dynamic Programming", *Pergamon Press, Oxford-New York*, 1981.

[4.11] MITTEN, L.G., "Composition Principles for Synthesis of Optimal Multistage Processes", *Operations Research, Vol.* 12, 1964, *pp* 610-619.

[4.12] BELLMAN, R.E., "Dynamic Programming", *Princeton University Press*, 1957.

[4.13] BELLMAN, R.E., DREYFUS, S.E., "Applied Dynamic Programming", *Princeton University Press*, 1962.

[4.14] DREYFUS, S.E., LAW, A.M., "The Art and Theory of Dynamic Programming", *Academic Press, New York-London*, 1977.

[4.15] FALLSIDE, F., WOODS, W.A., "Computer Speech Processing", *Prentice-Hall, Englewood Cliffs, New Jersey*, 1985.

[4.16] ROSENBERG, A.E., "Automatic Speaker Verification : A Review", *Proc. IEEE, Vol.* 64, *No.* 4, *April* 1976, *pp* 475-487.

[4.17] ITAKURA, F., "Minimum Prediction Residual Principle Applied to Speech Recognition", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP*-23, *No.* 1, *Feb.* 1975, *pp* 67-72.

[4.18] SAKOE, H., "Two-level DP Matching – a Dynamic Programming based Pattern Matching Algorithm for Connected-Word Recognition", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP*-27, 1979, *pp* 288-295.

[4.19] NEY, H., "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP*-32, *No.* 2, *April* 1984, *pp* 263-271.

[4.20] RABINER, L.R., SCHAFER, R.W., "Digital Processing of Speech Signals", *Prentice-Hall, Englewood Cliffs, New Jersey*, 1978.

[4.21] BERGER, T., "Rate Distortion Theory – A Mathematical Basis for Data Compression", *Prentice-Hall, Englewood Cliffs, New Jersey*, 1971.

[4.22] GRAY, A.H., Jr., MARKEL, J.D., "Distance Measures for Speech Processing", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP*-24, *No.* 5, *Oct.* 1976, *pp* 380-391.

[4.23] GRAY, A.H., Jr., GRAY, R.M., MARKEL, J.D., "Comparison of Optimal Quantizations of Speech Reflection Coefficients", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP*-25, *No.* 1, *Feb.* 1977, *pp* 9-23.

[4.24] VISWANATHAN, R., MAKHOUL, J., "Quantization Properties of Transmission Parameters in Linear Predictive Systems", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP-23, No. 3, June* 1975, *pp* 309-321.

[4.25] ITAKURA, F., SAITO, S., "Analysis Synthesis Telephony based on the Maximum Likelihood Method", *Proc. 6th Intl. Congr. Acoust., Tokyo, Japan,* 1968, *pp C-17 to C-20.*

[4.26] GRAY, R.M., BUZO, A., GRAY, A.H., Jr., "Distortion Measures for Speech Processing", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP-28, No. 4, August* 1980, *pp* 367-376.

[4.27] CHHATWAL, H.S., CONSTANTINIDES, A.G., "Application of Dynamic Programming to the Optimal Segmentation of Speech Spectra for Selective Linear Prediction Modelling", *Proc. IEE Intl. Conf. on Speech Input/Output: Techniques and Applications, March* 1986, *pp* 43-48.

[4.28] CHHATWAL, H.S., CONSTANTINIDES, A.G., "Optimally Segmented Autoregressive Models", *Proc. Second Intl. Workshop on Spectral Techniques, Montrèal, Canada, October* 1986.

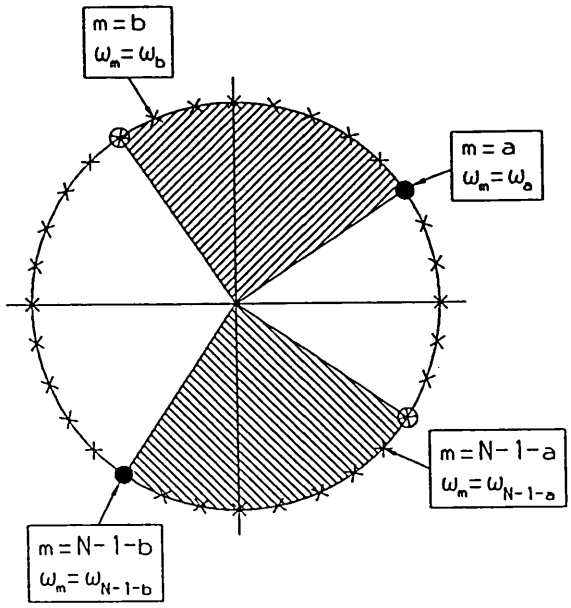
[4.29] CHHATWAL, H.S., CONSTANTINIDES, A.G., "Spectral Estimation of Speech Signals Using Optimal Spectral Segmentation", *Proc. Inst. of Acoustics Autumn Conf., Nov.* 1986, *Vol.* 8, *Part* 7, *pp* 127-144.

Type of Error Measure		Squared	Itakura-Saito	Weighted
Optimal End Frequencies (kHz)	f_1 f_2	1.21 5.00	3.59 5.00	3.59 5.00
Optimal Predictor Orders	p_1 p_2	2 10	10 2	10 2
Overall Error		6474	89.36	114450
Standard LP Model Error		10075	134.40	196775
Improvement		35.7%	33.5%	41.8%

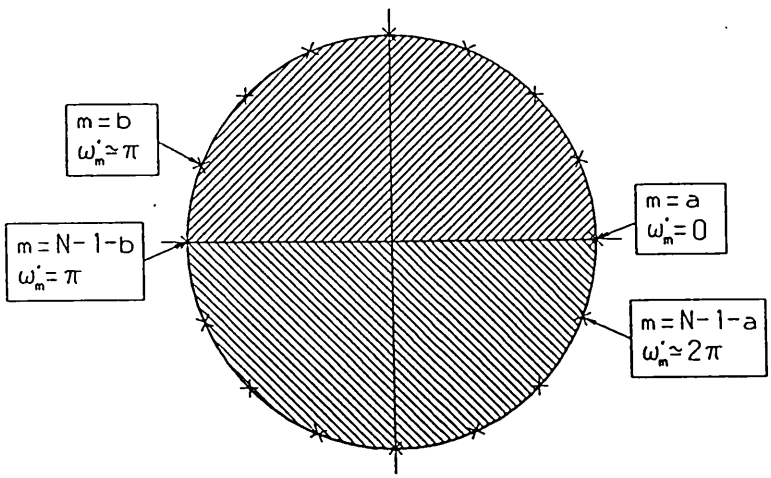
Table 4.1 Performance of Optimal 2 Region Segmentations for Various Error Criteria

Type of Error Measure		Squared	Itakura-Saito	Weighted
Optimal End Frequencies (kHz)	f_1 f_2 f_3	1.09 3.13 5.00	1.91 3.59 5.00	1.91 3.59 5.00
Optimal Predictor Orders	p_1 p_2 p_3	4 4 4	6 4 2	6 4 2
Overall Error		6024	84.43	111090
Standard LP Model Error		10075	134.40	196775
Improvement		40.2%	37.2%	43.5%

Table 4.2 Performance of Optimal 3 Region Segmentations for Various Error Criteria



(a) ACTUAL FREQUENCY DOMAIN



(b) SELECTIVE LP FREQUENCY DOMAIN

Figure 4.1 Representation in z-Plane of Frequency Domain Transformations for Selective LP

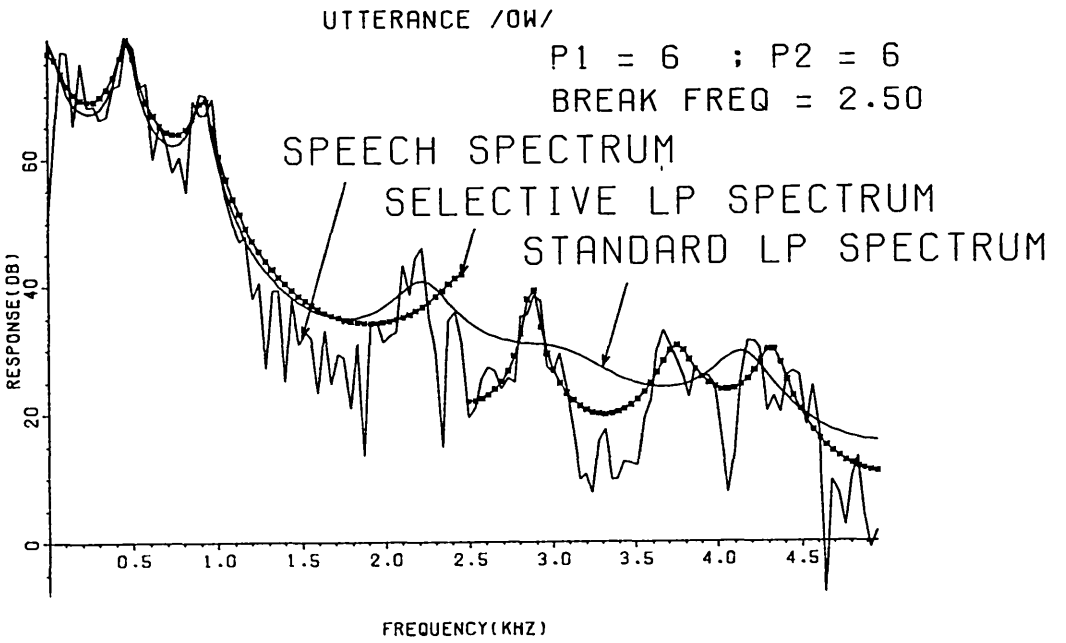


Figure 4.2 An Example of 2 Region Selective LP Modelling

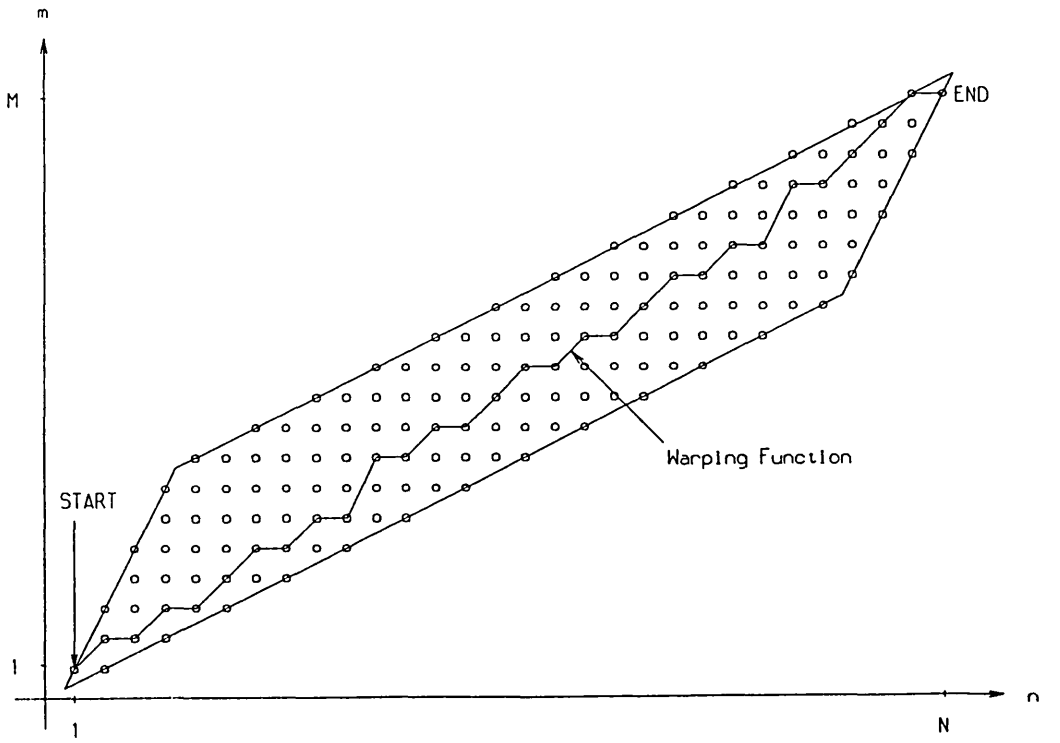


Figure 4.3 Illustration of Dynamic Time Warping with Continuity Conditions

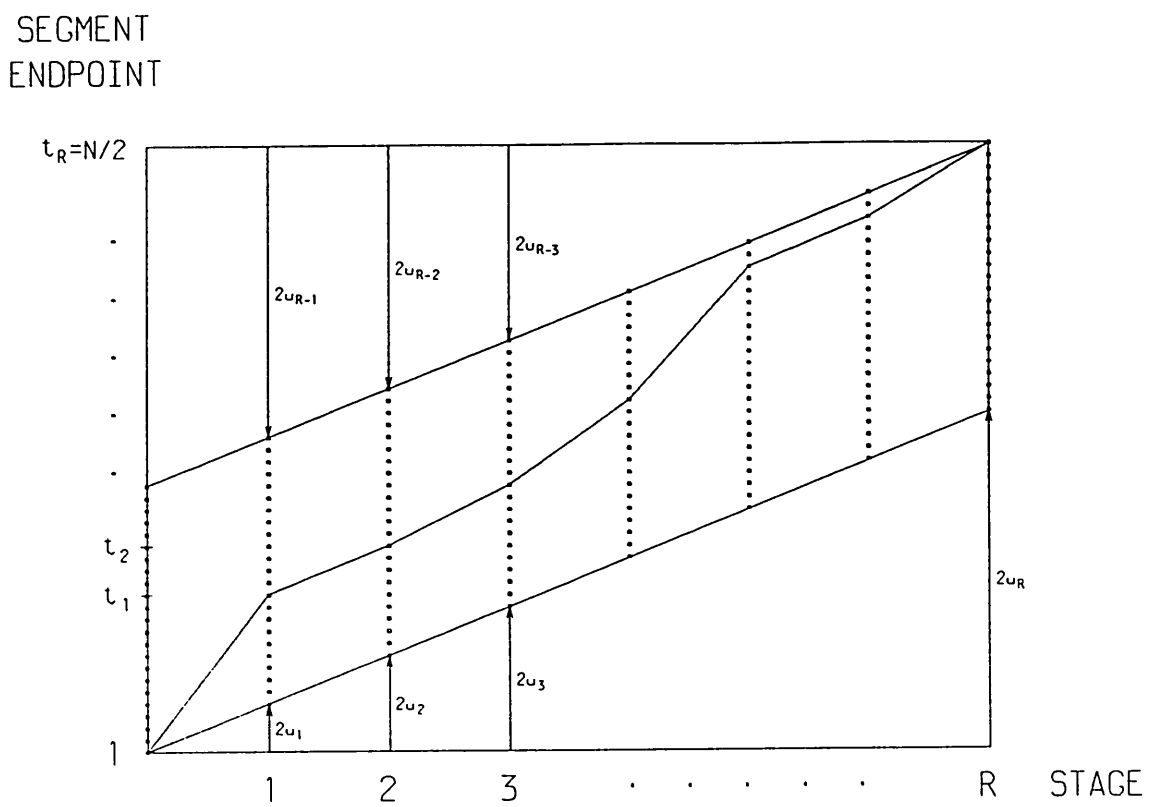


Figure 4.4 Domain of Possible Segmentation Paths for Region Endpoints

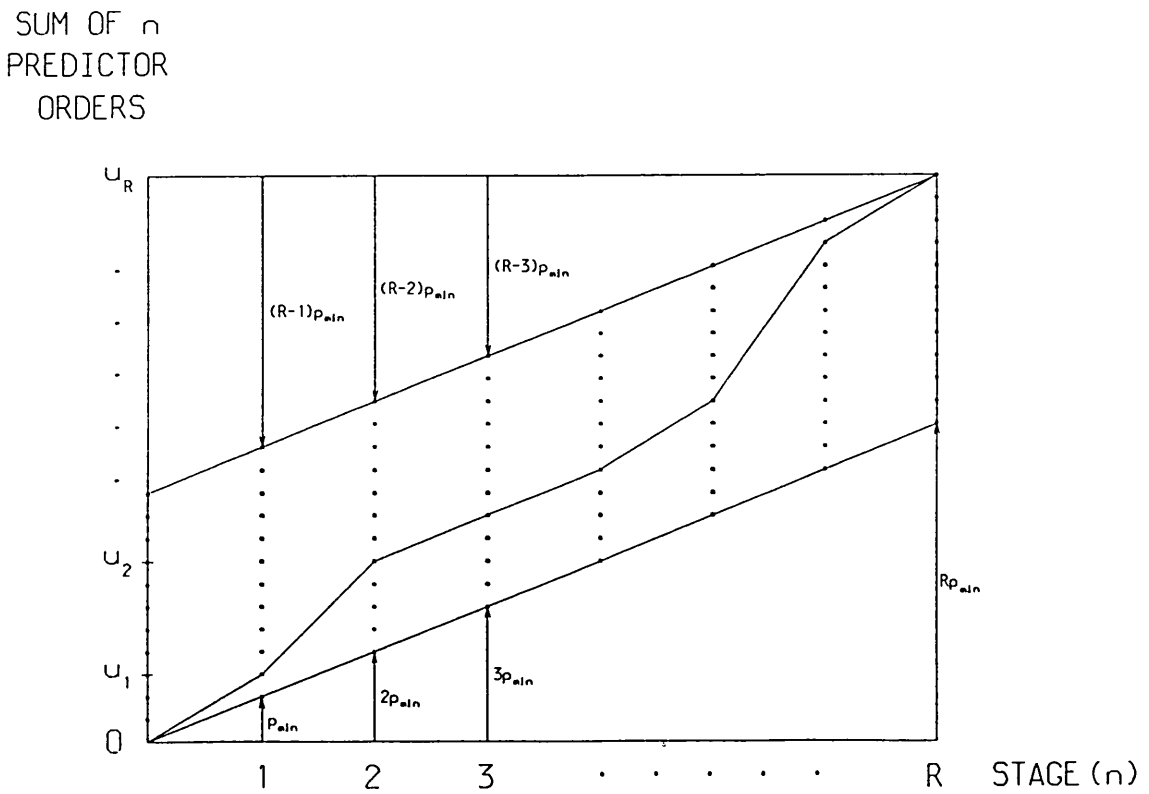


Figure 4.5 Domain of Possible Segmentation Paths for the Sum of Region Predictor Orders

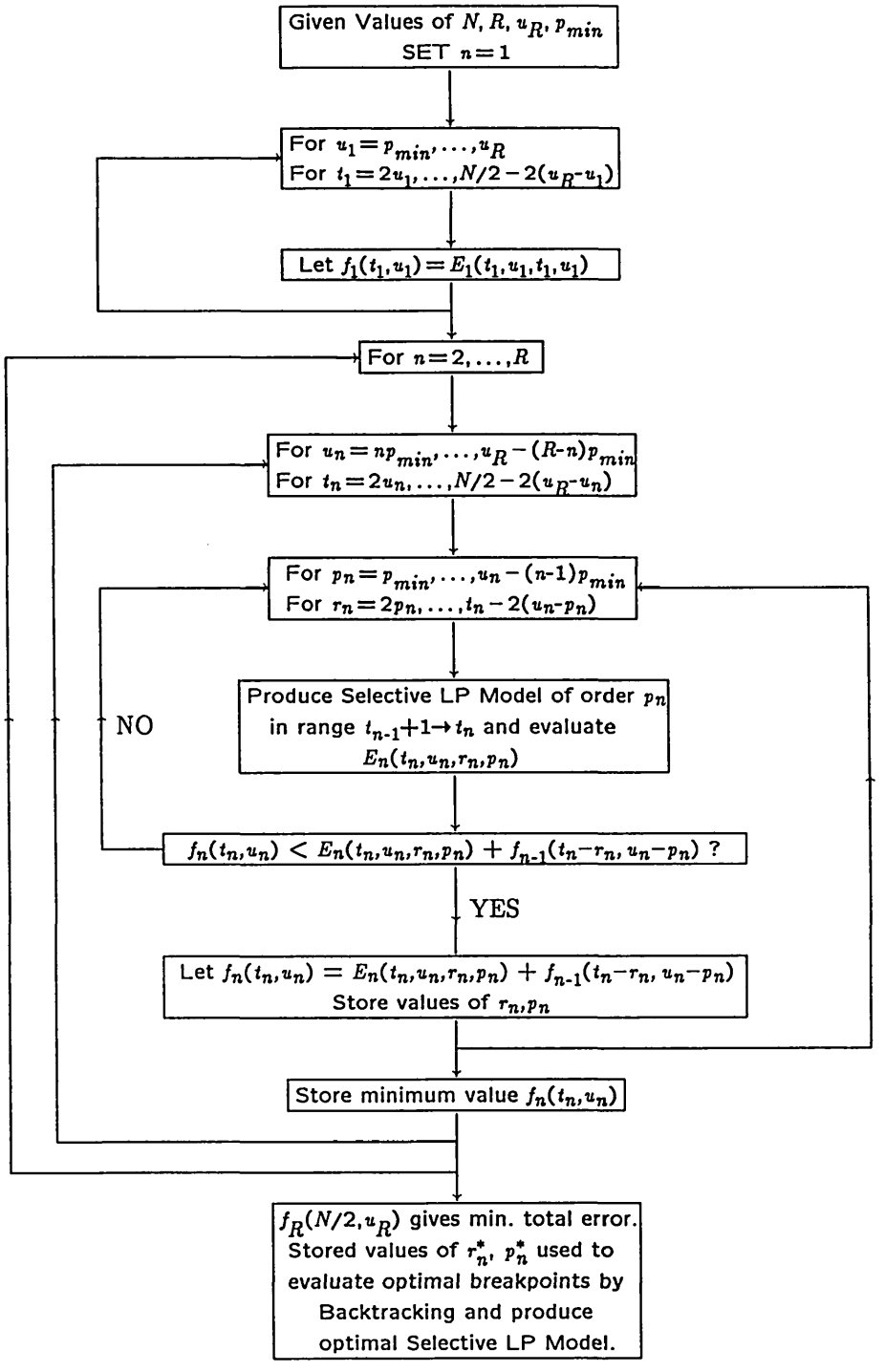
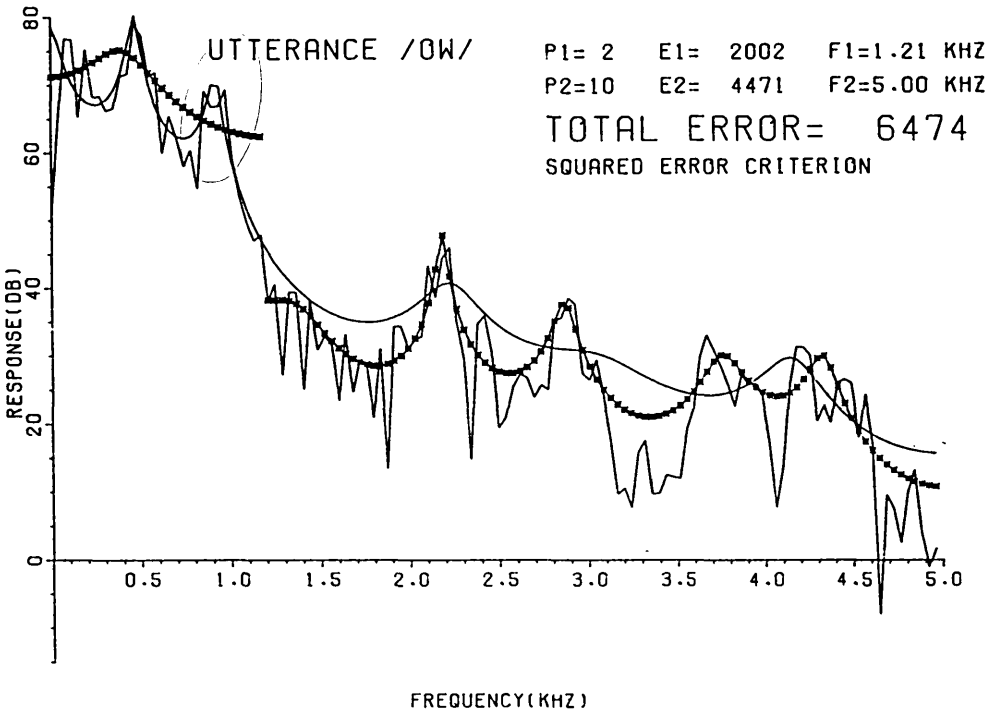
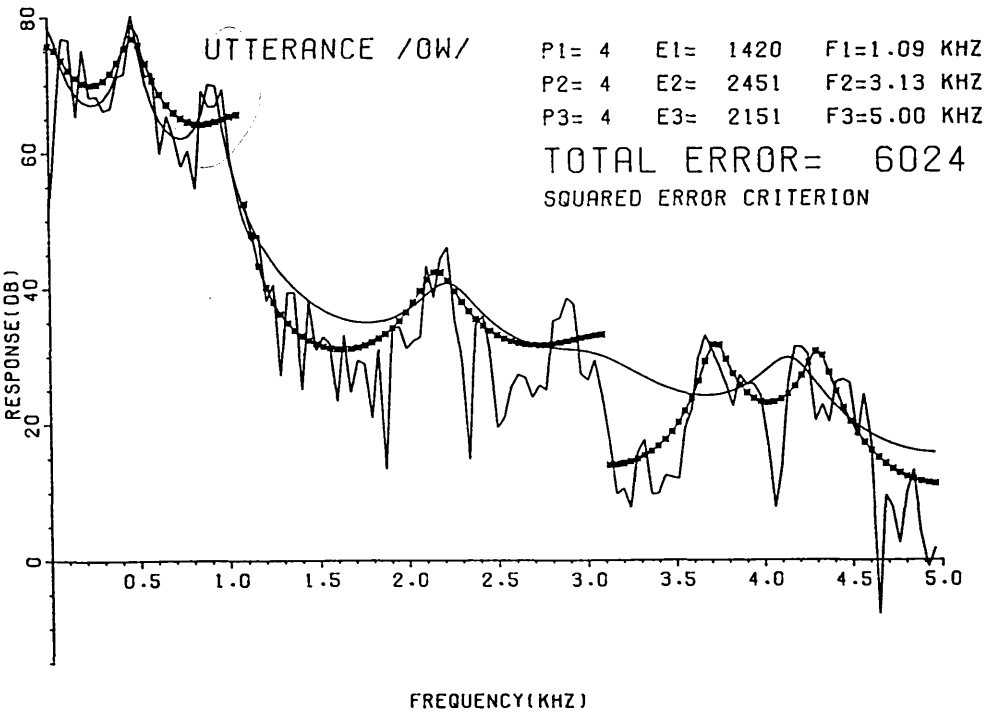


Figure 4.6 Flow Diagram Illustration of Procedure for Speech Spectral Segmentation with Optimal Breakpoint Selection and Predictor Order Allocation



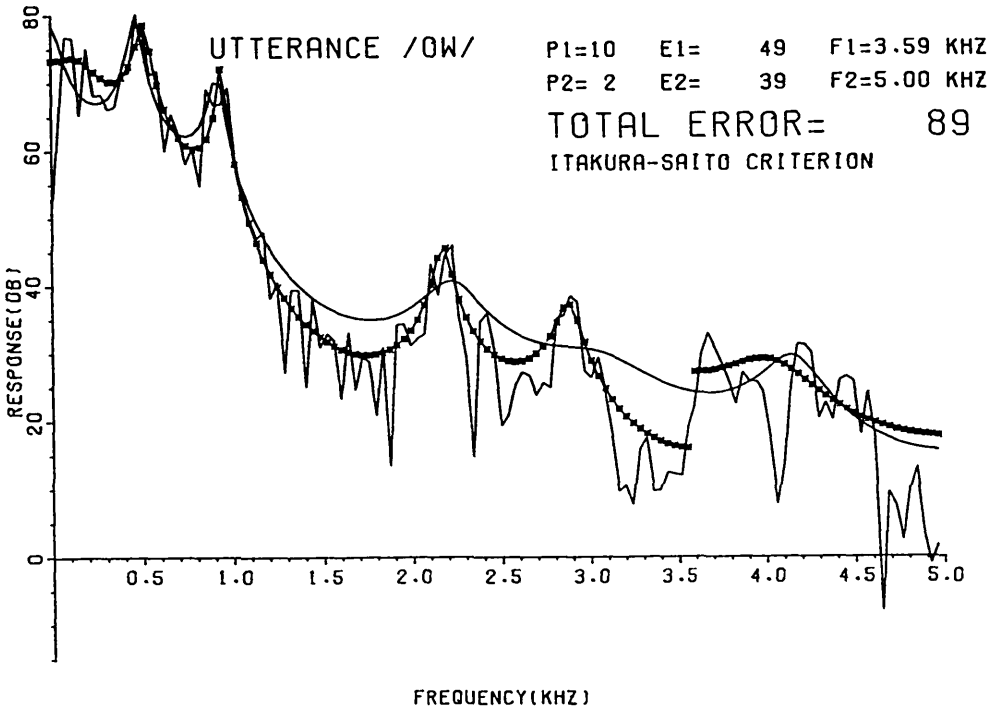
(a)



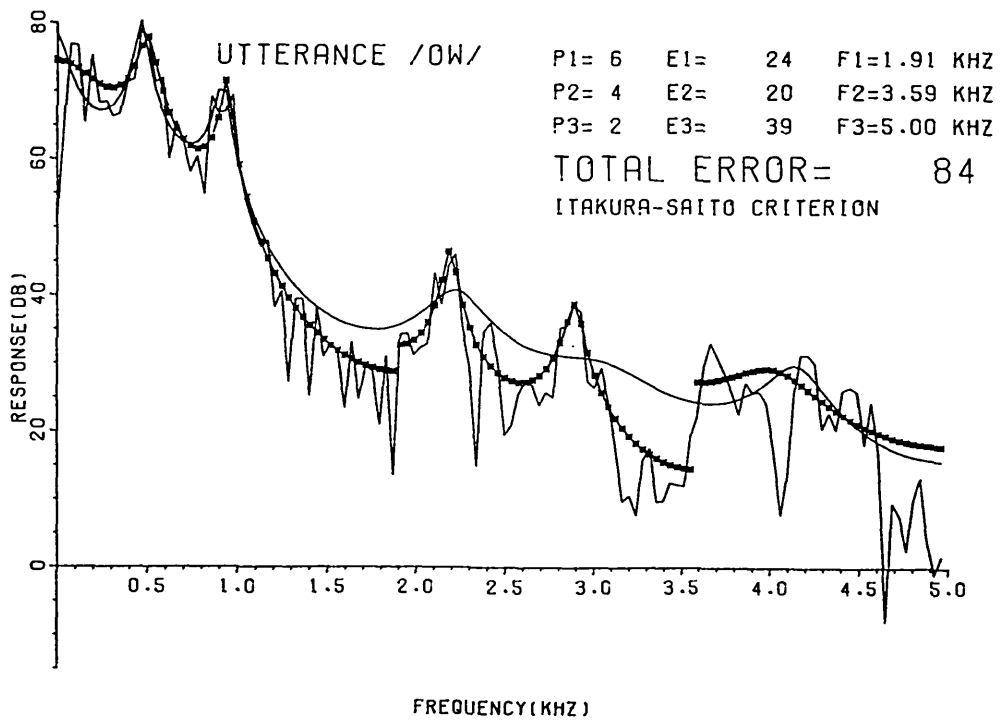
(b)

Figure 4.7 Optimal 2 and 3 Region Speech Spectral Segmentations for a Squared Error Measure

——— : Standard LP fit to Speech Spectrum —•—•— : Selective LP fit

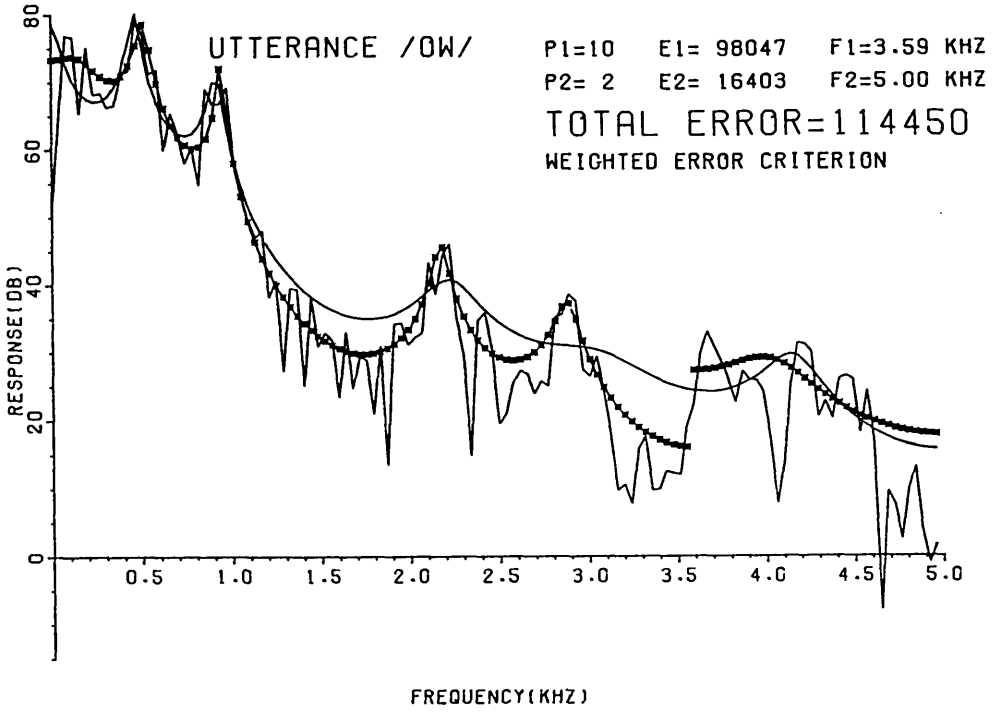


(a)

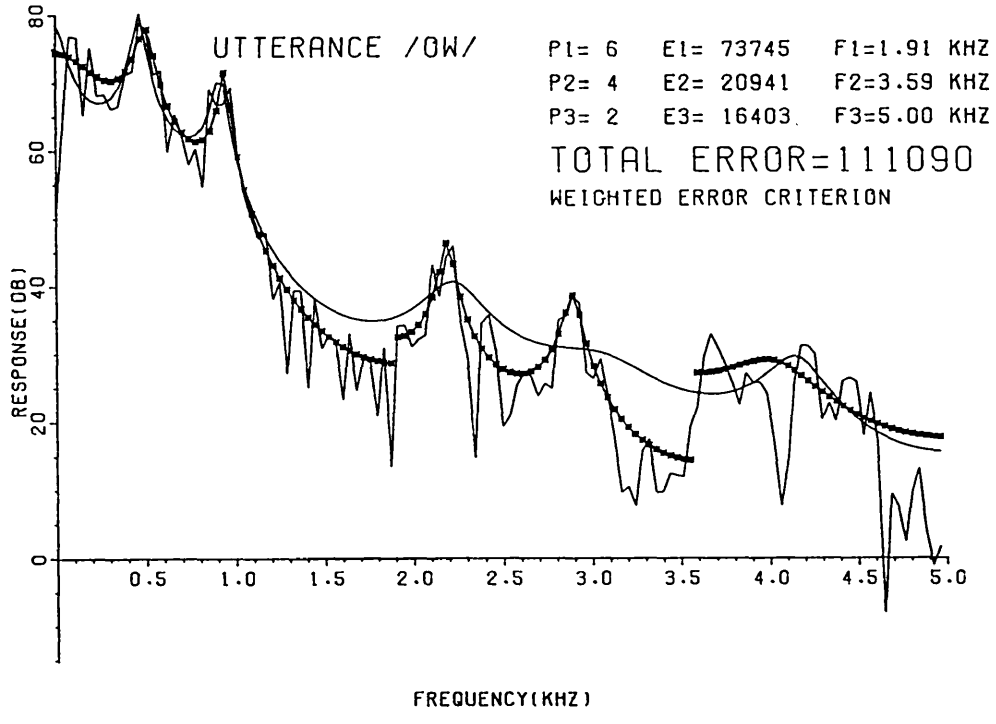


(b)

Figure 4.8 Optimal 2 and 3 Region Speech Spectral Segmentations for an Itakura-Saito Error Measure

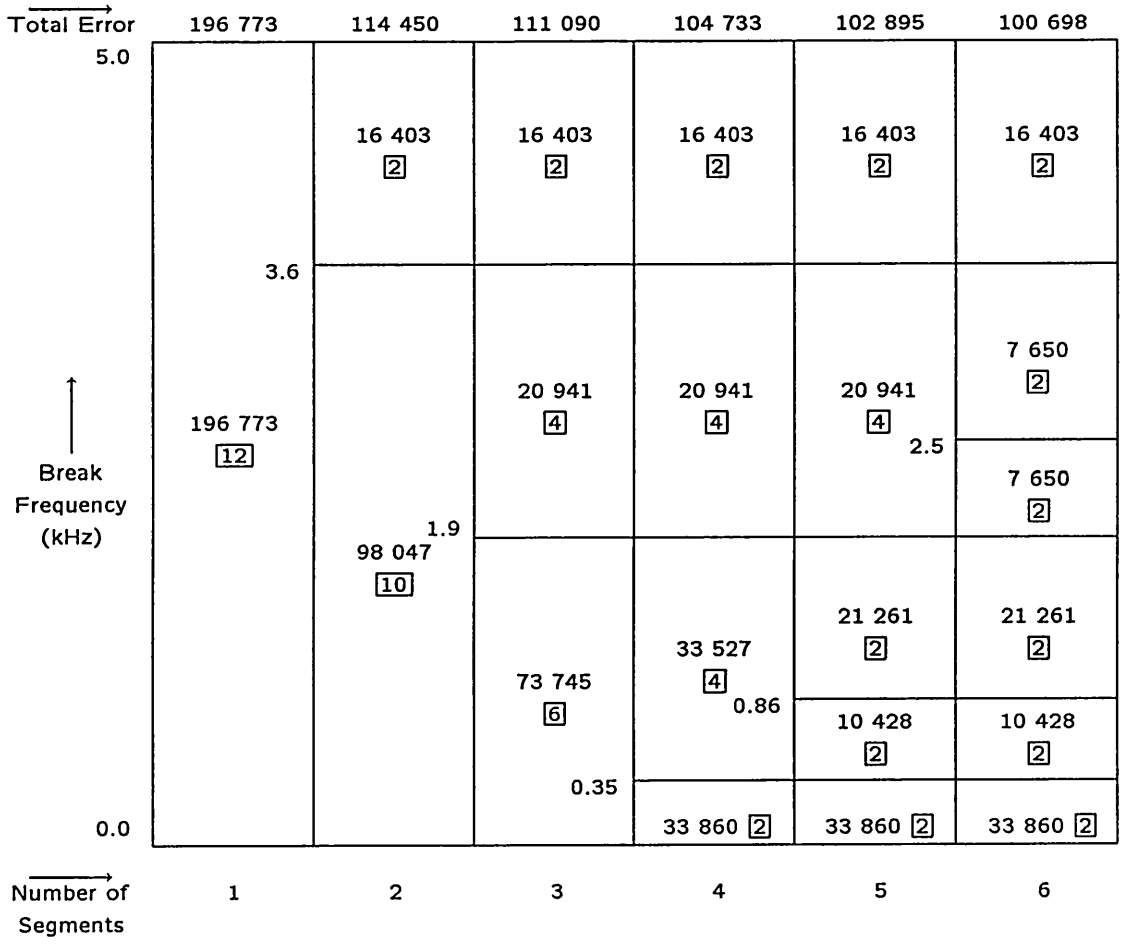


(a)



(b)

Figure 4.9 Optimal 2 and 3 Region Speech Spectral Segmentations for a Weighted Error Measure



(Predictor Orders shown in Boxes)

Figure 4.10 Variation of Breakpoints, Predictor Orders and Region Errors with an Increasing Number of Segments for Suboptimal Selective LP Modelling

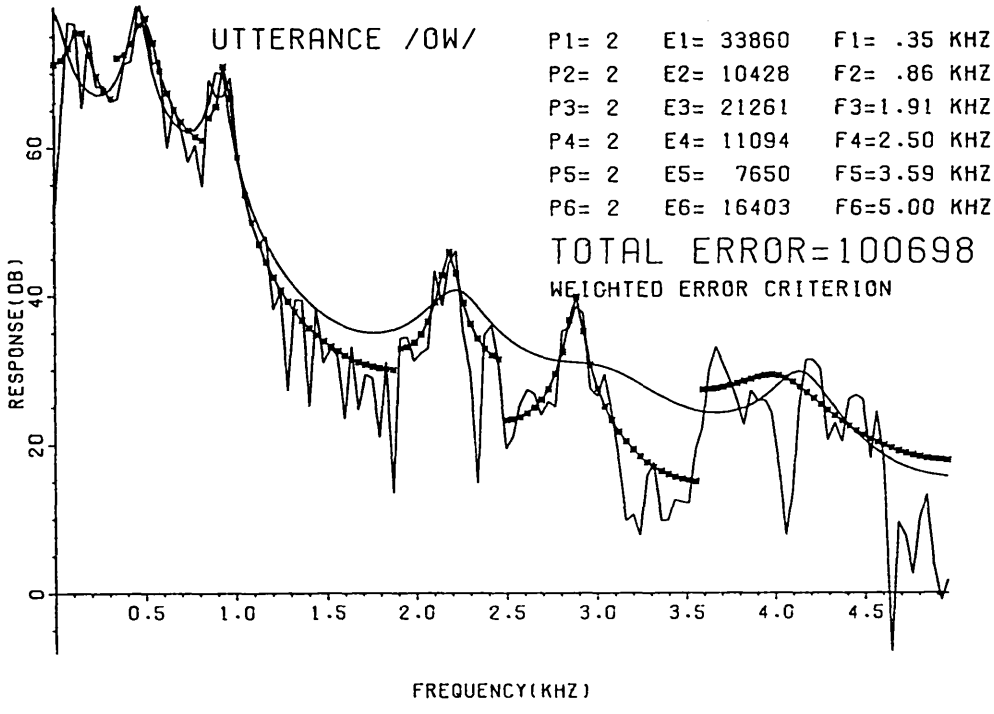


Figure 4.11 Frequency Response for 6 Region Suboptimal Selective LP Model

CHAPTER 5

SPECTRAL SEGMENTATION INTO FORMANT REGIONS

CHAPTER 5. SPECTRAL SEGMENTATION INTO FORMANT REGIONS

5.1 INTRODUCTION

The suboptimal segmentation technique developed in the last chapter produced a partition of a voiced speech spectrum which appeared to subdivide the frequency response in such a manner that each region contained a single formant peak. The importance of this form of segmentation stems from the proposition that it should be possible to derive accurate estimates of formant frequencies and bandwidths from the resulting 2nd order Selective LP models which are associated with each speech formant.

However, the suboptimal scheme as outlined in section 4.6 may not always result in a segmentation into 2nd order regions nor produce the global minimum error for a 2nd order segmentation. To overcome this difficulty, a special case of the optimal segmentation scheme is developed in this chapter to achieve precisely the above requirements. Certain constraints are, therefore, incorporated into the general breakpoint selection/predictor order allocation technique to allow the partitioning of a spectrum into segments containing single formant peaks. The nature of these restrictions is entirely due to the special requirements of selecting formant-type regions.

Another very important consideration in the choice of an appropriate linear prediction model is the overall predictor order to be employed in the analysis scheme. In the proposed technique for spectral segmentation into formants, the total number of predictor coefficients is closely related to the overall number of segments. Thus, in this chapter, the provision of an effective procedure for the calculation of the number of Selective LP regions automatically yields a possible solution to the complex problem of choosing an overall predictor order for a given speech data frame. The effectiveness of this method is also compared with those techniques which are traditionally used in estimating an appropriate model order for standard

LP analysis.

Once the speech frequency response has been segmented into 2nd order Selective LP regions, a vital requirement is the availability of a technique for the conversion of the Selective LP model parameters into formant centre frequencies and bandwidths. Such a technique is proposed in section 5.4 which involves an inverse mapping of the model coefficients from the Selective LP domain to the actual frequency domain.

5.2 ADDITIONAL CONSTRAINTS FOR THE SELECTION OF FORMANT REGIONS

5.2.1 Introduction of a Fixed Region Predictor Order

From the analysis of section 2.6 and Chapter 3, the transfer function, $H(z)$, of a linear prediction model of order p is given by

$$H(z) = \frac{G}{A(z)} = \frac{G}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}} \tag{5.1}$$

Now, the roots $\{z_k : k=1,\ldots,p\}$ of the predictor polynomial $A(z)$ can be used to express $A(z)$ as a factored polynomial

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} = \prod_{k=1}^{p} (1 - z_k z^{-1}) \tag{5.2}$$

As the coefficients $\{\alpha_k : k=1,\ldots,p\}$ of the predictor are real, the roots $\{z_k : k=1,\ldots,p\}$ may either exist as complex conjugate pole pairs in the LP system response or lie along the real axis. Those poles $\{f_k, f_k^* : k=1,\ldots,N_p\}$ which are complex conjugate may be combined to produce a system function $F(z)$ while $S(z)$ represents the corresponding expression for the combined effect of the real poles $\{s_k : k=1,\ldots,N_s\}$. Therefore,

$$A(z) = F(z).S(z) = \left(\prod_{k=1}^{N_f} (1 - f_k z^{-1})(1 - f_k^* z^{-1}) \right) \left(\prod_{k=1}^{N_s} (1 - s_k z^{-1}) \right) \tag{5.3}$$

From the discussion of formant analysis in Chapter 2, the expression $(1 - f_k z^{-1})(1 - f_k^* z^{-1})$ will be recognised as being representative of the response of a second-order digital resonator. Therefore, it is these poles which generally give rise to formants in the linear predictive model spectrum and the total number of formants in this frequency response will be given by N_f. The role of $S(z)$ will be explored in more detail in subsequent sections. Also, from the above factorisation of the predictor polynomial, it is clear that the predictor order (p) is equivalent to

$$p = 2N_f + N_s \tag{5.4}$$

Thus, for a given speech data frame, the maximum number of formants which can arise in the model spectral density will be $p/2$ (i.e. if there are no real poles).

For the particular application of spectral segmentation into formants, a Selective LP model is required in each region which characterises a single speech formant. The model spectrum should, therefore, exhibit a single resonant peak and thus, the model order in each region must be set to a value of 2. Hence,

CONSTRAINT 1

Employ 2nd Order Predictor in each Selective LP Region.

Alternatively, using the notation of Chapter 4,

$$\text{Set} \qquad p_n = 2 , \qquad \forall n \in \{1,\ldots,R\} \tag{5.5}$$

This condition also results in two other important computational considerations. The first is a considerable reduction in the computational requirements over the optimal breakpoint selection/predictor order allocation technique of Chapter 4 since the segmentation procedure now only involves a choice of the optimal break frequencies and not the predictor orders as well. The subsequent computational saving is very significant and is mainly responsible for the optimal segmentation into formants being tractable for a large number of Selective LP regions.

The other simplification which arises from the above restriction is due to the particular nature of a 2nd order linear predictive model. From Chapter 3, the use of the autocorrelation method results in the following relationship between the predictor coefficients $\{\alpha_1,\alpha_2\}$ of a 2nd order LP system and the corresponding autocorrelation values $\{R(0), R(1), R(2)\}$:

$$\sum_{k=1}^{2} \alpha_k \, R|i\text{-}k| \;=\; R(i)\,, \qquad\qquad i=1,2 \tag{5.6}$$

Therefore, the equivalent matrix relation is given by

$$\begin{bmatrix} R(0) & R(1) \\ R(1) & R(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \end{bmatrix} \tag{5.7}$$

Direct inversion of the 2nd order Toeplitz matrix produces the result

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \frac{1}{[R(0)]^2 - [R(1)]^2} \begin{bmatrix} R(0) & -R(1) \\ -R(1) & R(0) \end{bmatrix} \begin{bmatrix} R(1) \\ R(2) \end{bmatrix} \tag{5.8}$$

Thus,
$$\alpha_1 \;=\; \frac{R(0)\,R(1) - R(1)\,R(2)}{[R(0)]^2 - [R(1)]^2} \tag{5.9a}$$

$$\alpha_2 \;=\; \frac{R(0)\,R(2) - [R(1)]^2}{[R(0)]^2 - [R(1)]^2} \tag{5.9b}$$

In addition, the total squared prediction error (E) is related to the model gain (G) and the autocorrelation sequence as follows:

$$G^2 = E = R(0) - \sum_{k=1}^{2} \alpha_k R(k) \tag{5.10}$$

$$\text{So,} \quad G = \sqrt{R(0) - \alpha_1 R(1) - \alpha_2 R(2)} \tag{5.11}$$

As the autocorrelation values can be derived from the signal spectrum using the Selective LP procedure, it is clear that equations (5.9a),(5.9b) and (5.11) form a complete solution to the problem of calculating the 2nd order LP model parameters $\{\alpha_1,\alpha_2\}$, G directly from the signal frequency response. This result eliminates the need for a general routine such as the Levinson-Durbin algorithm.

5.2.2 Elimination of Real Poles in the Selective LP Model

The 2nd order LP model system function $H(z)$ will be of the form:

$$H(z) = \frac{G}{1 - \alpha_1 z^{-1} - \alpha_2 z^{-2}} \tag{5.12}$$

Thus the corresponding poles (z_1, z_2) in the z-domain will be at the following positions:

$$z_1, z_2 = \frac{\alpha_1 \pm \sqrt{\alpha_1^2 + 4\alpha_2}}{2} \tag{5.13}$$

Therefore, the above transfer function can exhibit two distinct forms of behaviour. The first case occurs when $\alpha_1^2 + 4\alpha_2 > 0$ and the pair (z_1, z_2) will then both lie along the real axis of the z-domain. However, if $\alpha_1^2 + 4\alpha_2 < 0$, then (z_1, z_2) will be complex conjugate (i.e. $z_2 = z_1^*$) and the roots will be given by:

$$z_1, z_1^* = \frac{\alpha_1}{2} \pm j\sqrt{-\left(\frac{\alpha_1}{2}\right)^2 - \alpha_2} \tag{5.14}$$

In order to illustrate the effect of these two cases on the system frequency response, Fig. 5.1 (a) shows the magnitude response arising from a pair of real poles corresponding to the transfer function

$$H(z) = \frac{1}{1 - 0.2z^{-1} - 0.48z^{-2}} \tag{5.15}$$

The equivalent response is shown in Fig. 5.1 (b) for the complex conjugate pole pair $\left(z_1 = 0.9\,e^{j\pi/3},\ z_1^* = 0.9\,e^{-j\pi/3}\right)$ resulting from the transfer function

$$H(z) = \frac{1}{1 - 0.9z^{-1} + 0.81z^{-2}} \tag{5.16}$$

These results show that only the complex poles produce a clear resonance peak in the magnitude response. In fact, it is shown in Appendix B that a system with real poles will never yield a maximum in the frequency response while complex conjugate poles modelling speech formants will usually produce such a peak. In any case, a pole description such as equation (5.14) is of the form necessary for a second order digital resonator for which characteristics such as centre frequency and bandwidth can be defined. This form of equivalence must be established since it is necessary to relate these resonator parameters to those of the formant being modelled in order to estimate the formant position. However, the response of a real-pole model can not be compared in this way to a second order resonator and thus the terms resonant frequency and bandwidth lose their inherent meaning in the context of formant position estimation.

The occurrence of real poles must, therefore, be prevented in any of the Selective LP models constituting the spectral segmentation into formant regions and this condition represents the second restriction on the segmentation procedure [5.1],[5.2].

CONSTRAINT 2

Prevent Occurrence of Real Poles in the 2nd Order Selective LP Model.

The manner in which this objective is achieved is as follows:

Set error in the n-th region $E_n(t_n, u_n=2n, r_n, p_n=2)$ to a very large value if, for a certain t_n, r_n, the predictor coefficients $\{\alpha_1,\alpha_2\}$ are such that real poles occur in that region, i.e.

$$\text{if} \qquad \alpha_1{}^2 + 4\alpha_2 > 0 \tag{5.17}$$

This prevents this combination of t_n, r_n leading to the minimum cost function $f_n(t_n, u_n=2n)$.

5.2.3 Incorporation of a Minimum Segmental Frequency Range

The speech signal was described in section 2.5 as the convolution of an excitation function $e(n)$ and the vocal tract impulse response $v(n)$. As $e(n)$ represents a periodic signal, its effect on the speech spectrum is to produce ripples which correspond to individual harmonics of the excitation and are spaced at the fundamental frequency of the speech signal [5.3]. These harmonics are clearly in evidence in the speech spectrum of Fig. 4.2 and those of previous chapters. The sequence $v(n)$ is responsible for producing the overall formant structure of voiced speech.

The task of the procedure for spectral segmentation into formants is to select those peaks in the speech spectrum which correspond to formants but, as discussed in relation to

homomorphic processing in Chapter 2, these peaks can often be confused with those corresponding to the harmonics of the excitation. The following constraint is, therefore, introduced into the segmentation scheme in order to prevent this occurrence [5.4].

CONSTRAINT 3

Establish a Minimum Possible Frequency Range for each Selective LP Segment.

A minimum length (r_{min}) is, therefore, fixed such that all the segments cover a frequency range of at least 250 Hz (approx.) which allows formant-type regions to be selected but precludes the possible isolation of harmonics of the pitch period in the original frequency response — which are generally spaced at around 100 Hz for male speakers.

This condition replaces the similar restriction in the optimal breakpoint selection/predictor order allocation procedure of Chapter 4 where the length (in samples of the speech spectrum) of any one region was constrained to be at least twice the predictor order in that region. Another point of interest concerning the technique for speech spectral segmentation into formants is the choice of error criterion. Of the three error measures discussed in the previous chapter, the weighted error measure was preferred since the other two forms of criterion were found experimentally to produce partitions with an excessive number of segments in the high frequency region of the signal spectrum due to the greater impact of noise on this low signal energy area of the speech frequency response.

As in the optimal speech spectral segmentation method, the form of error criterion only affects the choice of region lengths and endpoints $\{r_n, t_n : n=1,\ldots,R\}$ while the model in each region is still a standard Selective LP model. One major advantage of this approach is that it takes advantage of both the local spectral matching properties of the linear predictive methodology and the suitability of the weighted error measure over a broader spectral range. This is since, as noted in Chapter 3, LP tends to match the signal response much more effectively in the regions of spectral peaks than in areas of lower signal energy [5.5]. The

perceptually important regions corresponding to formant peaks are, therefore, well represented. This property is particularly important for the work described in this chapter since it enables precise estimates of formant position to be evaluated from these effective formant characterisations.

However, the LP spectral matching process operates uniformly over the entire frequency range and results in an estimate which treats regions of both high and low signal energy in the same manner. As mentioned above, the Weighted error criterion has certain advantages in this respect and so it is this form of error measure which is implemented in the optimisation procedure discussed in this chapter and is also used in all subsequent results.

5.2.4 Implementation of Procedure for Spectral Segmentation into Formants

Using the nomenclature introduced in Chapter 4, the three constraints developed over the course of the three previous subsections can now be incorporated into the optimal speech spectral segmentation algorithm to yield the following specification of the dynamic programming equations for speech spectral segmentation into formants:

For 1 region:

$$f_1(t_1, u_1=2) = E_1(t_1, u_1=2, r_1=t_1, p_1=2) \qquad r_{min} \leq t_1 \leq N/2 \tag{5.18}$$

For subsequent regions:

$$f_n(t_n, u_n=2n) = \min_{r_{min} \leq r_n \leq t_n-(n\text{-}1)r_{min}} \left\{ E_n(t_n, u_n=2n, r_n, p_n=2) + f_{n\text{-}1}(t_n - r_n,\ u_n - p_n=2(n\text{-}1)) \right\} \tag{5.19}$$

for $nr_{min} \leq t_n \leq N/2$

Both equations are subject to the constraint that t_n,r_n are such that the following condition holds for the n-th Selective LP region.

$$\alpha_1^2 + 4\alpha_2 < 0 \tag{5.20}$$

Backtracking can again be employed as in section 4.4 in order to calculate the optimal breakpoints $\{t_n^* : n=1,\ldots,R\}$. The two equations (5.18) and (5.19) along with constraint (5.20) specify the solution utilised for the segmentation of speech spectra into their constituent formants.

There are three important points to note from these expressions. The first is that the entire process is a series of one variable optimisations with respect to the region lengths and not in relation to the predictor orders as well. Another point is that for $n < R$, the range of region endpoints $\{t_n\}$ in the above equations spans the interval up to $N/2$, even though the values

$$N/2 - (R-n)\, r_{min} < t_n \leq N/2 \tag{5.21}$$

are not required in order to compute $f_{n+1}(t_{n+1}, u_{n+1})$ etc. The reason for this overspecification is that the optimal segmentations for the entire spectrum at each stage $\{f_n(N/2, u_n=2n) : n=1,\ldots,R\}$ are used to compute the required number of Selective LP regions as described in the next section and thus it is required that t_n should reach $N/2$ at each stage.

It can also be seen that this property illustrates the embedded nature of the DP formulation in that, during the calculation of $f_R(N/2, u_R=2R)$, all other optimal segmentations for $n < R$ are also obtained. Another computational factor is that, during the final iteration of the DP equations (i.e. when $n=R$), it is only necessary to compute $f_n(t_n, u_n)$ for the final point $t_n = N/2$ since this is the only value which is required during the final stage.

The results of spectral segmentation with the above constraints are shown in Fig. 5.2 for 4 different voiced phonemes with $N/2=128$ frequency samples and a sampling rate of 10 kHz in all cases. The actual vocalisations under investigation are (a) the utterance /OW/ as in "b*ough*t", (b) /IY/ as in "m*ea*n", (c) /E/ as in "m*e*n" and (d) /AE/ as in "s*a*ng". The notation for the various curves in Fig. 5.2 is the same as that for the optimal segmentations of Chapter 4 and a single region LP spectrum of the same overall order as the Selective LP model is also shown for comparison.

One property of these results is that the Selective LP models seem to provide a considerably improved spectral estimate of the original signal than the standard LP models. The other observation is that these plots indicate that this form of segmentation does indeed tend to produce divisions which separate the individual formants as required into distinct Selective LP regions.

It may also have been noted that the number of regions in the various segmentations ranges from 4 to 6 depending on the utterance under investigation and it is the choice of the total number of segments that is discussed in the next section. This issue is of particular relevance since, in this case, the overall predictor order is given by twice the number of segments and thus a specification of the number of regions in the segmentation will also provide an estimate of the predictor order required to adequately model a speech spectrum for the purposes of formant estimation.

5.3 CHOICE OF NUMBER OF REGIONS IN SEGMENTED MODEL

The choice of an overall predictor order for a linear prediction model always represents a complex and problematic decision. As noted in section 4.4, the task of calculating an '*optimal*' number of predictor coefficients is a non-trivial exercise since the fit of an LP model improves indefinitely as the order p increases. The requirement is, therefore, to determine a value of p which is adequate for the problem at hand in order to incur the minimum

computational expense.

Although this computational factor is the main consideration in applications such as speech coding, the choice of model order is even more critical in the case of procedures aimed at speech formant estimation. This is since, as described in section 5.2, the predictor order is closely related to the number of poles responsible for formant modelling in the corresponding LP system function. Therefore, if a value for p is chosen which is too low, then there will be an insufficient number of poles to characterise all the formants in the speech spectrum.

However, the introduction of an excessive number of coefficients can also cause errors since peaks may be selected in the signal frequency response which do not actually correspond to true speech formants. In the development of effective formant extraction algorithms, it is, therefore, vital that a good estimate of the overall model order be available for an LP system to represent a given speech spectrum. A means of calculating the total number of regions (and thereby the predictor order) to employ in the scheme for spectral segmentation into formants is developed in the following subsections [5.6].

5.3.1 Evaluation of Trends in the Weighted Error Curve

Many of the criteria which are traditionally employed in the estimation of an overall LP model order are based on the examination of the manner in which the prediction error associated with a linear prediction model varies with the number of predictor coefficients. These methods are discussed in more detail in section 5.3.3. A first step towards the development of a similar criterion for the method of spectral segmentation into formants introduced in this chapter is the analysis of the behaviour of the overall weighted error ($\mathfrak{E}^k$) for the segmented model in relation to the corresponding number of segments (k). In terms of the nomenclature developed in the previous chapter, this error $\mathfrak{E}^k$ is, therefore, defined as:

$$\mathfrak{E}^k = \sum_{n=1}^{k} E_n(t_n^*, u_n = 2n, r_n^*, p_n = 2) \tag{5.22}$$

and represents the sum of the individual segment errors in the k-segment model where $\{r_n^*, t_n^* : n=1,\ldots,k\}$ denote the optimal region lengths and endpoints respectively for this model. During the Dynamic Programming procedure, the complete set of errors $\{\mathfrak{E}^k : k=1,\ldots,R\}$ is, therefore, evaluated and its variation with k will be described as the *weighted error curve* in the following discussion.

These curves are illustrated in Fig. 5.3 for 22 different speech frames with a number of segments ranging between 2 and 8 in each case. The underlying speech data which was employed to produce these results was derived from a wide variety of vowel and vowel-like speech sounds. Although there are individual variations between the various curves[1], there seems to be a general consistency amongst these results and in the manner in which the total weighted error ($\mathfrak{E}^k$) reduces as the number of regions increases.

In order to obtain a fair estimate of the general nature of the weighted error curve irrespective of these local perturbations, these curves were each normalised with respect to the 2 region model error ($\mathfrak{E}^2$) and then averaged over the 22 individual frames in order to produce the result shown in Fig. 5.4. From the asymptotic nature of this weighted error curve, it may be conjectured that a good hypothesis for the variation of $\mathfrak{E}^k$ with k seems to be that the weighted error exhibits an exponential decay with a constant offset, i.e.

$$\mathfrak{E}^k \simeq \tilde{\mathfrak{E}}^k = c_1 e^{-c_2 k} + c_3 \tag{5.23}$$

---

[1] *It is, in fact, vital that there are some discrepancies amongst the weighted error curves since it is exactly these variations which form the basis of the method introduced in the next subsection for the selection of one particular number of regions (k) to employ in the segmented model in preference to all other values of k.*

where c_1, c_2, c_3 are constants. The validity of this supposition can then be tested in the following manner.

One convenient form in which the above hypothesis-testing problem can be formulated is as an unconstrained minimisation problem where the unknown parameters c_1, c_2, c_3 are to chosen in order to minimise the discrepancy between the actual weighted errors $\{\mathfrak{E}^k : k=2,\ldots,8\}$ and the estimated values $\{\tilde{\mathfrak{E}}^k : k=2,\ldots,8\}$. This discrepancy may take the form of an L_2 norm where the squared differences between $\mathfrak{E}^k, \tilde{\mathfrak{E}}^k$ are minimised with the result that the problem can be stated as:

$$\min_{c_1, c_2, c_3} \left\{ \sum_{k=1}^{M} \left[\mathfrak{E}^k - (c_1 e^{-c_2 k} + c_3) \right]^2 \right\} \tag{5.24}$$

A technique for solving this problem is provided by the FORTRAN routine E04FDF from the NAG (Numerical Algorithms Group) Library of subroutines [5.7]. This program employs a Newton-type method for the unconstrained minimisation of a sum of squares of M nonlinear functions in N variables [5.8] (i.e. $M=7$, $N=3$ for the above problem). Therefore, from a starting point supplied by the user, a sequence of points is generated which is intended to converge to a local minimum of the sum of squares. These points are generated using estimates of the curvature of the summand in equation (5.24).

The result of using this method on the data of Fig. 5.4 is shown as a continuous function in Fig. 5.5 alongwith the test values $\{\mathfrak{E}^k : k=1,\ldots,R\}$. As the fit between the actual weighted error curve and its estimate is very good, this suggests that the hypothesis of equation (5.23) is an appropriate model for the behaviour of the weighted error as the number of segments is increased.

5.3.2 Development of a Criterion for the Selection of an Overall Number of Segments

Although the model developed in the previous subsection for the weighted error curve represents a description of the general nature of the error $\mathfrak{E}^k$, there will be individual variations and discontinuities for specific speech frames which will be superimposed on the overall pattern. However, the detection of these changes may be problematic against the background of the exponential decay of the error curve and thus this process will be aided if a relationship of the type of equation (5.23) can be converted to a simpler form.

The ideal mapping would be into a linear graph with zero gradient such that any deviations from this flat response will be easily discernible. This objective implies that the transformed function should be completely independent of the number of segments (k) — which can be achieved by eliminating this parameter from the following expressions for $\tilde{\mathfrak{E}}^k, \tilde{\mathfrak{E}}^{k-1}$:

$$\tilde{\mathfrak{E}}^k = c_1 e^{-c_2 k} + c_3 \tag{5.25a}$$

$$\tilde{\mathfrak{E}}^{k-1} = c_1 e^{-c_2(k-1)} + c_3 \tag{5.25b}$$

or
$$k = -\frac{1}{c_2} \ln\left[\frac{\tilde{\mathfrak{E}}^k - c_3}{c_1}\right] = 1 - \frac{1}{c_2} \ln\left[\frac{\tilde{\mathfrak{E}}^{k-1} - c_3}{c_1}\right] \tag{5.26}$$

$$\Longrightarrow \quad \frac{\tilde{\mathfrak{E}}^{k-1} - c_3}{\tilde{\mathfrak{E}}^k - c_3} = e^{c_2} \tag{5.27}$$

Although this expression is independent of k, an evaluation of the LHS does involve obtaining an estimate of the parameter c_3. As this value will change for each particular signal frame, it may be suggested that a more convenient transformation would be more appropriate.

Consider the use of three estimated errors:

$$\tilde{\mathcal{E}}^{k-1} = c_1 e^{-c_2(k-1)} + c_3 \tag{5.28a}$$

$$\tilde{\mathcal{E}}^{k} = c_1 e^{-c_2 k} + c_3 \tag{5.28b}$$

$$\tilde{\mathcal{E}}^{k+1} = c_1 e^{-c_2(k+1)} + c_3 \tag{5.28c}$$

Thus, $$\frac{\tilde{\mathcal{E}}^{k-1} - \tilde{\mathcal{E}}^{k}}{\tilde{\mathcal{E}}^{k} - \tilde{\mathcal{E}}^{k+1}} = \frac{e^{-c_2 k}(e^{c_2} - 1)}{e^{-c_2 k}(1 - e^{-c_2})} \tag{5.29}$$

or $$\frac{\mathcal{E}^{k-1} - \mathcal{E}^{k}}{\mathcal{E}^{k} - \mathcal{E}^{k+1}} \simeq \frac{\tilde{\mathcal{E}}^{k-1} - \tilde{\mathcal{E}}^{k}}{\tilde{\mathcal{E}}^{k} - \tilde{\mathcal{E}}^{k+1}} = e^{c_2} \tag{5.30}$$

This result is also independent of k and, in addition, the LHS may be evaluated without any a priori knowledge of the parameters c_1, c_2, c_3. The function $T(k)$ on which the choice of an overall number of segments will be based is, therefore, given by

$$T(k) = \frac{\mathcal{E}^{k-1} - \mathcal{E}^{k}}{\mathcal{E}^{k} - \mathcal{E}^{k+1}} \tag{5.31}$$

The nature of this mapping within the framework of the procedure for spectral segmentation into formants must now be investigated. The overall pattern of the asymptotically decreasing weighted error curve would be expected to exhibit a sharp drop when the required number of regions is reached. This is since the segmentation technique introduced in this chapter has been precisely geared towards the selection of formant regions in the speech frequency response. The weighted error should, therefore, fall steadily until the number of segments is adequate to model the number of formants in the signal spectrum and then decrease less rapidly. This observation suggests that, for the chosen number of regions k^*,

$$\mathcal{E}^{k^*-1} - \mathcal{E}^{k^*} \gg \mathcal{E}^{k^*} - \mathcal{E}^{k^*+1} \tag{5.32}$$

Therefore, with reference to equation (5.31), $T(k)$ should reach a maximum for the value $k=k^*$. The variation of $T(k)$ for the same four utterances as investigated in section 5.2 is shown in Fig. 5.6. The numbers of segments at which the maximum value of $T(k)$ occurs are those used in the segmentations of Fig. 5.2. Therefore, equation (5.31) provides a reliable stopping criterion with an easily discernible maximum for a variety of voiced speech sounds.

5.3.3 Model Order Selection Criteria for Standard LP Analysis

Several criteria have been suggested as objective bases for the selection of an appropriate LP model order. Most of these tests are derived from the total squared prediction error $\left(E^{(p)}\right)$ associated with a p-th order LP system. The first of these criteria to be investigated in this subsection is discussed by Makhoul [5.9] and is based on the premise that, if the signal spectrum is precisely described by an all-pole spectrum consisting of p_0 poles, then it follows that

$$E^{(p)} = E^{(p_0)}, \qquad \forall p \geq p_0 \tag{5.33}$$

Therefore, the error curve of the variation of $E^{(p)}$ with p will remain flat for $p > p_0$ and the occurrence of this property should provide a simple test for the best value of p to adopt in the linear predictive analysis. However, the actual speech signal is unlikely to conform exactly to the all-pole description and thus the error curve will only be expected to remain substantially flat for $p > p_0$. A threshold test of the form illustrated in the following expression should, therefore, provide a reasonable stopping criterion for the selection of a model order,

$$\mathrm{TT}(p) = 1 - \frac{E^{(p+1)}}{E^{(p)}} < \delta \tag{5.34}$$

The chosen value of p is that for which TT(p) falls below the threshold value δ. This form of test is actually based on hypothesis testing procedures using maximum likelihood ratios [5.9].

Another order selection criterion suggested by Akaike [5.10] relies on the minimisation of an information theoretic function and is thereby termed the Akaike Information Criterion (AIC(p)). This measure can be defined as:

$$\mathrm{AIC}(p) \;=\; \ln\left(E^{(p)}\right) \;+\; \frac{2(p+1)}{N} \tag{5.35}$$

where N is the number of data samples. The value of p for which AIC(p) is a minimum is taken to be the optimal value. Although the first term in this expression will always decrease with an increasing model order, the second term represents a penalty for the use of extra LP coefficients which do not result in a substantial reduction in the prediction error and thus there should be a point at which AIC(p) achieves a global minimum.

The final form of measure considered in this subsection was also developed by Akaike [5.11] and is known as the final prediction error (FPE(p)). This criterion selects the order of the LP process in order that the average error for the system where $p \longrightarrow \infty$ is minimised and is explicitly given by

$$\mathrm{FPE}(p) \;=\; E^{(p)}\left(\frac{N+p+1}{N-p-1}\right) \tag{5.36}$$

As the term in parantheses increases as p approaches N, it again provides an opposing influence to the decreasing nature of the values of $E^{(p)}$ (although the effect is multiplicative in this case rather than additive as in the Akaike Information Criterion). The order which is

selected is again that for which FPE(p) is a minimum.

The performance of the three criteria which have been introduced above was assessed in relation to the four voiced utterances used earlier in this chapter for the illustration of both the segmentation into formants procedure and the test for calculating an appropriate number of segments for this method. Fig. 5.7 shows the behaviour of the prediction errors $\left(E^{(p)}\right)$ for these four speech frames up to a total predictor order of 30.

The corresponding values of TT(p) evaluated according to equation (5.34) are depicted in Fig. 5.8 and in considering these results, it must be remembered that a sharp drop in level for TT(p) and a subsequently flat response is expected at the chosen value of the predictor order. However, the actual nature of these results is very erratic and does not give a clear indication of the value of p which is to be selected, especially in the cases of Figs. 5.8 (a) and (c). Another difficulty in the use of the threshold test is the choice of an actual threshold (δ). The selection of this value can represent a critical decision affecting the performance of this criterion — as is evident from the results of Fig. 5.8.

The behaviour of the Akaike Information Criterion (AIC) and the final prediction error (FPE) is shown in Fig. 5.9 and Fig. 5.10 respectively and, since the results for both of these measures are similar in most respects, their performance can be assessed together. In both cases, the actual error curve is modified by the effect of a gently upwardly sloping function of p which is intended to produce a global minimum for AIC(p) or FPE(p).

However, as illustrated in Fig. 5.9 and Fig. 5.10, the slope of these functions is often very small and, therefore, the actual absolute minimum is relatively sensitive to the nature of these additional terms. For instance, the effectiveness of the AIC will depend heavily on the gradient of the linear term in equation (5.35). Hence, it is suggested by Makhoul [5.9] that the criterion of equation (5.35) should not be regarded as immutable and may be adjusted to suit a particular application.

Another uncertainty regards the choice of N in equations (5.35) and (5.36) since this value is often replaced by the "*effective*" number of data points in the signal frame. This somewhat imprecise term is introduced to take account of any possible windowing of the speech signal. One possible definition of the effective width of a particular window is given by the ratio of the energy under the window relative to that for a rectangular window. For example, this rationale yields an effective length of approximately $0.4N$ for a Hamming window [5.9].

A further difficulty with the use of these criteria is the nature of the underlying assumptions since the AIC assumes that the signal has a Gaussian probability distribution [5.12] and is also associated with an uncorrelated noise excitation for the LP system. These assumptions are obviously inaccurate for a voiced speech signal and thus the deployment of this criterion may often be inappropriate. It has also been claimed by Kashyap [5.13] that the AIC is statistically inconsistent in that the error in choosing the correct order does not tend to zero as $N \longrightarrow \infty$.

Each of the three criteria described above are, therefore, associated with several disadvantages which account for the difficulties often ascribed to the problem of model order selection. This is in contrast to the function of the weighted error developed in section 5.3.2 which was found to produce a clear indication of the number of regions[2] to employ in the segmented model. One of the reasons for this difference in performance may lie in the fact that the entire procedure developed earlier in this chapter is aimed towards spectral segmentations which result in formant regions and thus the test based on the weighted error is particularly effective in estimating the number of speech formants within a given spectrum.

---

[2] *In comparing the results of Figs. 5.7, 5.8, 5.9 and 5.10 with the test of Fig. 5.6, it is important to remember that the overall predictor order for the model derived from the procedure for segmentation into formants is precisely equal to twice the number of segments in that model.*

However, the threshold test, AIC and FPE criteria are intended to deal with the problem of model order selection in a more general sense. Their use is not, therefore, specifically associated with speech analysis nor formant estimation and thus their success may not be totally satisfactory for the particular problem of selecting an appropriate model order for formant extraction.

Due to these problems, a fixed model order which is independent of the speech signal is often used in a variety of LP systems. In these cases, the predictor order is usually a function of the sampling frequency (f_s in Hz). Thus, as mentioned briefly in section 3.4.2, the vocal tract transfer function will produce an average of one resonant peak per 1000 Hz and hence a total of approximately $f_s/2000$ formants up to folding frequency. As discussed in section 5.2, the number of predictor coefficients must be at least twice the number of speech formants in order to characterise the formant resonances adequately and thus the predictor order is often chosen such that

$$p = f_s/1000 + \zeta \tag{5.37}$$

where ζ is a constant in the range 2 - 4 [5.14] required to represent the spectral properties of the source excitation and the radiation load.

5.4 FORMANT POSITION EVALUATION FROM SEGMENTED LP MODELS

During the course of this chapter, a procedure has been developed for the segmentation of a speech frequency response into formant-type regions which are characterised by 2nd order Selective LP models. The next stage in the analysis of this form of segmentation involves an examination of the feasability of obtaining estimates of formant position from the parameters associated with the Selective LP segments [5.2],[5.4],[5.6].

5.4.1 Frequency Domain Inverse Mapping for the Selective LP Configuration

The characteristics of a speech formant are traditionally represented in terms of its centre frequency and bandwidth and thus a conversion will be necessary to transform the position of the complex pole pair (z, z^*) associated with the 2nd order Selective LP model into the required formant parameters. If this set of poles in the z-domain is given by:

$$z,\ z^* \ = \ re^{\pm j\theta} \tag{5.38}$$

then the z-plane to s-plane transformation $z = e^{sT}$ (where T is the system sampling period) can be employed [5.15] to convert these poles into the corresponding set $(s,\ s^*)$[3]:

$$s,\ s^* \ = \ \frac{1}{T}\ (\ \ln r \pm j\theta\ ) \tag{5.39}$$

This situation is illustrated in Fig. 5.11 and, ignoring the effect of one pole upon the other in the 3 dB region of the peak associated with the latter, it is clear that the separation between the 3 dB points of the system frequency response — i.e. the Bandwidth (BW) — will be given by the distance between the two frequencies ω_1, ω_2 in Fig. 5.11. Therefore, the pole bandwidth is given by:

$$\text{BW} \ = \ -\frac{2}{T} \ln r \qquad (\text{rad s}^{-1}) \tag{5.40}$$

or

$$\text{BW} \ = \ -\frac{f_s}{\pi} \ln r \qquad (\text{Hz}) \tag{5.41}$$

where $f_s \ = \ \frac{1}{T} \ =$ SYSTEM SAMPLING FREQUENCY (Hz)

---

[3] *The complex logarithm employed in this expression is defined in the same terms as used for the analysis of homomorphic systems in section 2.5.*

For the particular case of a 2nd order Selective LP model, the pole positions (z, $z^* = r' e^{\pm j\theta'}$) can be evaluated from the predictor parameters $\{\alpha_1, \alpha_2\}$ as described in section 5.2.2 such that

$$r' e^{\pm j\theta'} = \frac{\alpha_1}{2} \pm j\sqrt{-\left(\frac{\alpha_1}{2}\right)^2 - \alpha_2} = p \pm jq \tag{5.42}$$

Thus, $$r' = \sqrt{p^2 + q^2} = \sqrt{\left(\frac{\alpha_1}{2}\right)^2 - \left(\frac{\alpha_1}{2}\right)^2 - \alpha_2}$$

$$\Longrightarrow \qquad r' = \sqrt{-\alpha_2} \tag{5.43}$$

and $$\theta' = \tan^{-1}\left(\frac{q}{p}\right) = \tan^{-1}\left[\frac{\sqrt{-\left(\frac{\alpha_1}{2}\right)^2 - \alpha_2}}{\frac{\alpha_1}{2}}\right]$$

$$\Longrightarrow \qquad \theta' = \tan^{-1}\left(\frac{\sqrt{-\alpha_1^2 - 4\alpha_2}}{\alpha_1}\right) \tag{5.44}$$

Therefore, the pole position in the Selective LP domain can be evaluated directly from the linear prediction coefficients. However, this set of parameters (r', θ') does not specify the true position of the formant pole (r, θ) in the actual frequency domain. This property can be demonstrated by considering a particular Selective LP segment covering the spectral range $\left\{ P(\omega_m) : m = t_n - r_n + 1, \ldots, t_n, \ \omega_m = \frac{2\pi(m-1)}{N} \right\}$. From the analysis of section 4.2, this set of spectral samples are responsible for modelling the frequency range

$$\frac{2\pi(t_n - r_n)}{N} \leq \omega_m < \frac{2\pi t_n}{N} \tag{5.45}$$

or $$f_s \cdot \frac{(t_n - r_n)}{N} \leq f_m < f_s \cdot \frac{t_n}{N} \tag{5.46}$$

where f_m is expressed in Hz. The effective '*sampling frequency*' (i.e. twice the folding frequency) of the Selective LP domain (f_s') is then given by:

$$f_s' = 2 f_s \frac{r_n}{N} \qquad \text{(Hz)} \tag{5.47}$$

The formant bandwidth within the Selective LP domain can then be expressed as in equation (5.41) with r', f_s' replacing r, f_s. As $f_s' < f_s$, it is clear that the requirement of equal pole bandwidths in both the Selective LP and actual frequency domains will lead to different values of pole radii in the two domains. Since the process of formant modelling requires that the spectral shape of the model frequency response should be unaltered through any inverse mapping of the Selective LP domain to the actual frequency domain, a convenient means of transformation is such that pole bandwidths are preserved through the mapping procedure. Therefore, to maintain the same bandwidth in both frequency domains,

$$\text{BW} = -\frac{f_s}{\pi} \ln r = -\frac{f_s'}{\pi} \ln r' \tag{5.48}$$

or
$$r = (r')^{f_s'/f_s} \tag{5.49}$$

This expression indicates that $r > r'$ since $f_s > f_s'$.

A corresponding transformation is now required to convert the pole centre frequency (θ') in the Selective LP domain to the true frequency (θ) in the actual frequency domain. As a linear mapping was employed in Chapter 4 to convert signal frequencies (ω_m) to Selective LP frequencies (ω_m'), the inverse mapping must also be linear. From equation (5.45), the radian frequencies 0, π in the Selective LP domain correspond to the actual frequencies $\frac{2\pi(t_n - r_n)}{N}$, $\frac{2\pi t_n}{N}$ respectively. Therefore, by linear interpolation between these two points,

$$\theta = \frac{2\pi (t_n - r_n)}{N} + \frac{\theta'}{\pi} \cdot \frac{2\pi r_n}{N}$$

or

$$\theta = \frac{2}{N} \left[\pi (t_n - r_n) + \theta' r_n \right] \tag{5.50}$$

Therefore, equations (5.49),(5.50) provide a means for the conversion of a pole (r', θ') in the Selective LP domain to the formant estimate (r, θ) in the actual frequency domain. To recapitulate, the entire procedure for the estimation of a formant position (r, θ) from the segmented speech spectrum is as follows:

(1) Calculate the autocorrelation sequence $\{R(0), R(1), R(2)\}$ from the frequency response samples in the region under consideration as described for the Selective LP procedure outlined in section 4.2.

(2) Use the following equations developed in section 5.2 to derive the 2nd order model predictor coefficients

$$\alpha_1 = \frac{R(0)\,R(1) - R(1)\,R(2)}{\left[R(0)\right]^2 - \left[R(1)\right]^2}$$

$$\alpha_2 = \frac{R(0)\,R(2) - \left[R(1)\right]^2}{\left[R(0)\right]^2 - \left[R(1)\right]^2}$$

(3) Evaluate the position of the pole in the Selective LP domain (r', θ') such that:

$$r' = \sqrt{-\alpha_2}$$

$$\theta' = \tan^{-1}\left(\frac{\sqrt{-\alpha_1^2 - 4\alpha_2}}{\alpha_1}\right)$$

(4) Transform this set of parameters to the true pole position (r, θ) in the actual frequency domain:

$$r = (r')^{f_s'/f_s}$$

$$\theta = \frac{2}{N}\left[\pi(t_n - r_n) + \theta' r_n\right]$$

where $f_s' = 2f_s\frac{r_n}{N}$ and r_n, t_n are the length (in samples of the speech spectrum) and endpoint respectively of the segment under investigation.

In order to assess the performance of this system of equations, it is necessary to compare their effectiveness with those techniques which are traditionally employed to calculate formant estimates from standard LP systems.

5.5 FORMANT ESTIMATION FROM STANDARD LP ANALYSIS

From the discussion of section 2.4, it is clear that the formant parameters are amongst the most fundamental acoustic features of the speech signal as they can be regarded as complex pole-pairs of the vocal tract transfer function. However, automatic formant extraction from speech data represents a major problem [5.15] due to the fact that the vocal tract impulse response is not a directly observable quantity. Since the actual output speech signal can be modelled as a convolution of this impulse response with the quasi-periodic glottal driving function, a deconvolution process is required to separate these two processes. This requirement can be fulfilled by the spectral smoothing properties of the linear predictive formulation.

Some of the first applications of LP analysis to formant estimation were described by Itakura and Saito [5.16], Atal and Hanauer [5.17] and Markel [5.18]. McCandless [5.19] also presented an algorithm based upon LP spectra for dealing with several problem areas. The

two main procedures which are usually employed in the calculation of formant estimates from linear predictive models are outlined in the following subsections.

5.5.1 Root-Solving of the Predictor Polynomial

The analysis of section 5.2 was used to illustrate the fact that the roots of the predictor polynomial, $A(z)$, can be divided into two specific types. Those roots which are complex conjugate are generally associated with formant description while the real roots provide the correct overall spectral shape for the LP model frequency response. Therefore, an evaluation of the complex poles of the LP transfer function should yield good estimates of the speech formant positions. Thus, the following equation is solved:

$$A(z) \;=\; 1 - \sum_{k=1}^{p} \alpha_k \, z^{-k} \;=\; 0 \tag{5.51}$$

to produce a set of roots $\{z_i : i=1,\ldots,p\}$ which may be either real or complex conjugate. However, this process is not as simple as for the Selective LP 2nd order models since the polynomial order is considerably larger. A convenient means for solving this problem lies in the use of FORTRAN routine C02AEF from the NAG (Numerical Algorithms Group) Library of subroutines [5.20]. This algorithm uses the method of Grant and Hutchins [5.21] to determine the roots of an $(N-1)$th order real polynomial, $f(z)$, such that

$$f(z) \;=\; A(1)\, z^{N-1} + A(2)\, z^{N-2} + \;\ldots\; + A(N) \;=\; 0 \tag{5.52}$$

The roots are located by a method based on the simultaneous use of Newton's method to solve

$$\Re(f) \;=\; \Im(f) \;=\; 0 \tag{5.53a}$$

and a steepest descent method to minimise

$$\left[\Re(f)\right]^2 + \left[\Im(f)\right]^2 \tag{5.53b}$$

where $\Re(f)$ is the real part of $f(\cdot)$ and $\Im(f)$ the imaginary part. As successive roots are evaluated, they are divided out of the polynomial. In the particular case of the predictor polynomial, from equation (5.51),

$$1.z^p - \alpha_1 z^{p-1} - \alpha_2 z^{p-2} - \ldots - \alpha_p = 0 \tag{5.54}$$

Comparing this expression with equation (5.52), it is clear that

$$\begin{aligned} N &= p+1 \\ A(1) &= 1 \\ A(i) &= -\alpha_{i-1}, \qquad 2 \leq i \leq N \end{aligned} \tag{5.55}$$

Using these parameter values, the roots of the predictor polynomial may be evaluated.

For the purpose of formant estimation, not only are the real roots ignored, but also those complex roots which are associated with a very large bandwidth [5.22]. This is since true speech formants rarely have bandwidths greater than 700 Hz (approx.) [5.23] and thus any complex poles with larger bandwidths are generated by the LP procedure in order to provide general spectral shaping and can, therefore, be rejected in the formant extraction analysis. Therefore, the bandwidths of the complex roots are calculated according to equation (5.41) and only those roots which have a bandwidth less than a value of around 700 Hz are considered to be formant estimates.

5.5.2 Peak-Picking from the LP Model Frequency Response

This method is considerably less involved than the root-solving approach and, in its simplest terms, searches for peaks in the LP model spectrum and evaluates formant centre frequencies and bandwidths directly from these peaks [5.15]. Although ease of implementation characterises this particular approach, there are several problems associated with the peak-picking method. The first arises in the case when two formants are closely spaced such that a 3 dB fall-off does not occur on each side of a resonance and it is, therefore, difficult to obtain an accurate estimate of formant bandwidth.

In these cases, it may be possible to employ the discrete peak value in the spectrum and the samples on each side of this peak in the evaluation of formant parameters using a form of parabolic interpolation [5.15]. However, in addition to the possible inaccuracies associated with basing formant measurements on merely three spectral samples, this technique still can not be used in the case where either a particular formant has such a large bandwidth or is so near an adjacent formant that there is no clear peak in the model spectrum. For these types of frequency response, it will be problematic to evaluate both formant bandwidths and centre frequencies with the peak-picking technique.

Although the use of the root-solving method will alleviate some of these difficulties, it introduces the extra problem of choosing which complex roots correspond to speech formants and eliminating those associated with the shaping characteristic. In addition, the issue discussed in subsection 5.3.3 of the choice of an overall predictor order remains a problem for both types of model.

5.5.3 Comparative Analysis of Formant Estimation Techniques

Three methods have now been developed for the estimation of formant positions from linear predictive models. The first is based on the scheme for spectral segmentation into

formants introduced in this chapter while the other two are the root-solving and peak-picking approaches generally employed in standard LP analysis. To compare the results of these various techniques, a speech signal was synthesised from a set of 6 formants[4] which are based on Klatt [5.23] as being appropriate for the sound /ʌ/ as in "but". These original parameters are described in terms of the formant centre frequencies $\{F_0,\ldots,F_5\}$ and bandwidths $\{B_0,\ldots,B_5\}$ in Table 5.1.

The behaviour of the criterion of subsection 5.3.2 for selecting the number of regions to employ in the segmentation technique is shown in Fig. 5.12 and it is clearly evident that this test suggests the use of 6 regions representing the 6 formants in the segmented model. The nature of the 6 region segmentation is illustrated in Fig. 5.13 for the same graphical description as for Fig. 4.2. This result shows that the segmentation divides the speech spectrum into its individual formants and the analysis of section 5.4 can then be used to derive the formant estimates indicated for the segmented model in Table 5.1. The estimated formant parameters are, therefore, in reasonably close agreement with the actual synthesis values — especially in the case of centre frequencies.

In order to carry out the same calculations for the root-solving and peak-picking schemes, it was first necessary to select a suitable model order using one of the techniques investigated in subsection 5.3.3. For example, the threshold test criterion behaves as shown in Fig. 5.14 for a number of predictor coefficients up to 30. Although the test value is very erratic, a suitable choice appears to be $p=11$ as the criterion value is reasonably low after this point.

With this predictor order, the LP model spectrum is as shown in Fig. 5.15 and the two different methods produce the results shown in Table 5.1. It is clear that the performance of

---

[4] *The manner in which a synthetic speech signal can be produced from a series of formants is described in Chapter 6.*

both techniques is significantly inferior to that of the segmented model for those values which could be calculated. However, a much more serious problem lies in the fact that several parameters were unobtainable (indicated by * in Table 5.1) which, in the case of $B0$ & $B5$ for the peak-picking method, implies that a 3 dB fall-off on either side of a resonance was not exhibited in the model frequency response. For the root-solving technique, only 5 complex poles were produced and thus no information was available on $F0, B0$.

Another point to note is that the bandwidth estimates for the two standard methods are significantly larger than the original values used to produce the signal — especially for the higher formants. This property of LP analysis to overestimate the bandwidths of spectral peaks has also been noted by other researchers [5.24]. The bandwidth estimation process for the segmented model is particularly successful in view of the fact that formant bandwidths are known to be difficult to deduce from analyses of natural speech [5.23]. Other comparisons of the standard and segmented modelling techniques are discussed in Chapter 6 in relation to formant modelling.

5.6 CONCLUSIONS

This chapter has been concerned with the introduction of a technique for the partitioning of a speech spectrum into several regions — each of which contains a single speech formant. This form of segmentation was developed from the optimal breakpoint selection/predictor order allocation technique of Chapter 4 with the introduction of some constraints. One of these restrictions involves the use of 2nd order Selective LP models in all the various regions constituting the segmentation. This constraint was found to produce a considerable reduction in the complexity of both the Dynamic Programming area of the segmentation procedure and the process of model spectrum evaluation since it eliminates the need for a routine such as the Levinson-Durbin algorithm to calculate the LP coefficients.

For the purposes of formant modelling, the selection of an appropriate linear predictive model order is usually critical to the performance of the formant extraction process. In the case of the procedure for spectral segmentation into formants, this problem is equivalent to choosing a suitable number of segments in the model since the predictor order is precisely given by twice the number of regions.

On the basis of the behaviour of the weighted error curve, a criterion was, therefore, developed for the selection of an overall number of segments. The segmentations produced by models based on this criterion were found to be quite effective in the partitioning of speech spectra into their constituent formant regions for several vowel sounds. This performance was then compared to those techniques which are applied in standard LP analysis for model order selection and it was found that these latter methods are generally associated with serious theoretical and implementational drawbacks which impair their effectiveness for the specific purpose of formant modelling. For this reason, in most applications of LP analysis, a fixed predictor order is usually employed — irrespective of the particular speech data under consideration [5.22].

The final stages of this chapter have been concerned with the analysis of the 2nd order Selective LP models to obtain estimates of formant positions. As the predictor order is only 2 in these models, it is relatively straightforward to obtain the pole parameters in the Selective LP domain and then a simple inverse mapping can be employed to convert these values to formant centre frequency and bandwidth estimates in the actual frequency domain.

This method was applied to a synthetic speech signal and found to produce considerably more accurate estimates than techniques such as solving for the roots of the predictor polynomial or selecting peaks from the model frequency response. These latter techniques are those traditionally used to derive formant data from LP models and it was also observed that these procedures can often lead to the serious problem of specific speech formants being totally absent in the formant extraction results.

Another means of testing the efficacy of various formant estimation algorithms is to deploy the results in formant synthesis models and then compare the synthesised speech with the input data. This issue is explored in considerable detail in the following chapter.

5.7 REFERENCES

[5.1] CHHATWAL, H.S., CONSTANTINIDES, A.G., "Application of Dynamic Programming to the Optimal Segmentation of Speech Spectra for Selective Linear Prediction Modelling", *Proc. IEE Intl. Conf. on Speech Input/Output: Techniques and Applications, March* 1986, *pp* 43-48.

[5.2] CHHATWAL, H.S., CONSTANTINIDES, A.G., "Optimally Segmented Autoregressive Models", *Proc. Second Intl. Workshop on Spectral Techniques, Montréal, Canada, October* 1986.

[5.3] OPPENHEIM, A.V., "Applications of Digital Signal Processing", *Prentice-Hall, Englewood Cliffs, New Jersey,* 1978.

[5.4] CHHATWAL, H.S., CONSTANTINIDES, A.G., "Spectral Estimation of Speech Signals Using Optimal Spectral Segmentation", *Proc. Inst. of Acoustics Autumn Conf., Nov.* 1986, *Vol.* 8, *Part* 7, *pp* 127-144.

[5.5] MAKHOUL, J., "Spectral Linear Prediction : Properties and Applications", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP-*23, *No.* 3, *June* 1975, *pp* 283-296.

[5.6] CHHATWAL, H.S., CONSTANTINIDES, A.G, "Speech Spectral Segmentation for Spectral Estimation and Formant Modelling", *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Proc., Dallas, Texas, April* 1987, *Vol.* 1, *pp* 316-319.

[5.7] NUMERICAL ALGORITHMS GROUP, "NAG Fortran Library Manual", *NAG Central Office, Oxford, Mark* 10, *Vol.* 3, 1983.

[5.8] GILL, P.E., MURRAY, W., "Algorithms for the Solution of the Non-Linear Least Squares Problem", *SIAM Journal on Numerical Analysis, Vol.* 15, 1978, *pp* 977-992.

[5.9] MAKHOUL, J., "Linear Prediction : A Tutorial Review", *Proc. IEEE, Vol.* 63, *April* 1975, *pp* 561-580.

[5.10] AKAIKE, H., "A New Look at Statistical Model Identification", *IEEE Trans. Automat. Contr., Vol. AC-*19, *Dec.* 1974, *pp* 716-723.

[5.11] AKAIKE, H., "Power Spectrum Estimation through Autoregressive Model Fitting", *Ann. Inst. Statist. Math.*, *Vol.* 21, 1969, *pp* 407-419.

[5.12] KAY, S.M., MARPLE, S.L., Jr., "Spectrum Analysis — A Modern Perspective", *Proc. IEEE*, *Vol.* 69, *No.* 11, *Nov.* 1981, *pp* 1380-1419.

[5.13] KASHYAP, R.L., "Inconsistency of the AIC Rule for Estimating the Order of Autoregressive Models", *IEEE Trans. Automat. Contr.*, *Vol. AC*-25, *Oct.* 1980, *pp* 996-998.

[5.14] RABINER, L.R., SCHAFER, R.W., "Digital Processing of Speech Signals", *Prentice-Hall, Englewood Cliffs, New Jersey*, 1978.

[5.15] MARKEL, J.D., GRAY, A.H., Jr., "Linear Prediction of Speech", *Springer-Verlag, Berlin-Heidelberg-New York*, 1976.

[5.16] ITAKURA, F., SAITO, S., "A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies", *Electron. Commun. Japan*, *Vol.* 53-*A*, 1970, *pp* 36-43.

[5.17] ATAL, B.S., HANAUER, S.L., "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", *J. Acoust. Soc. Amer.*, *Vol.* 50, 1971, *pp* 637-655.

[5.18] MARKEL, J.D., "Automatic Formant and Fundamental Frequency from a Digital Inverse Filter Formulation", *Proc. Intl. Conf. Speech Commun. Process.*, *Boston, Massachusetts*, 1972, *pp* 81-84.

[5.19] McCANDLESS, S.S., "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra", *IEEE Trans. on Acoust. Speech and Sig. Proc.*, *Vol. ASSP*-22, *No.* 2, *April* 1974, *pp* 135-141.

[5.20] NUMERICAL ALGORITHMS GROUP, "NAG Fortran Library Manual", *NAG Central Office, Oxford, Mark* 10, *Vol.* 1, 1983.

[5.21] GRANT, J.A., HITCHINS, G.D., "Two Algorithms for the Solution of Polynomial Equations to Limiting Machine Precision", *Computer Jnl.*, *Vol.* 18, 1975, *pp* 258-264.

[5.22] MARKEL, J.D., "Digital Inverse Filtering — A New Tool for Formant Trajectory

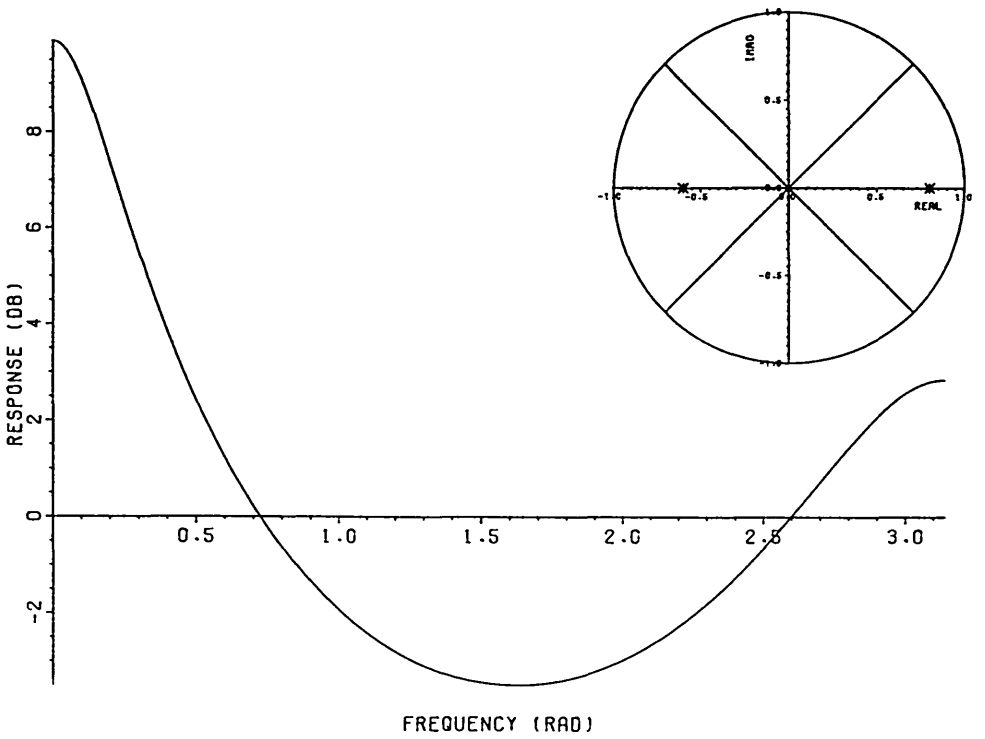
Estimation", *IEEE Trans. Audio and Electroacoustics, Vol. AU-20, No. 2, June* 1972, *pp* 129-137.

[5.23] KLATT, D.H., "Software for a Cascade/Parallel Formant Synthesizer", *J. Acoust. Soc. Amer., Vol.* 67, *No.* 3, *March* 1980, *pp* 971–995.

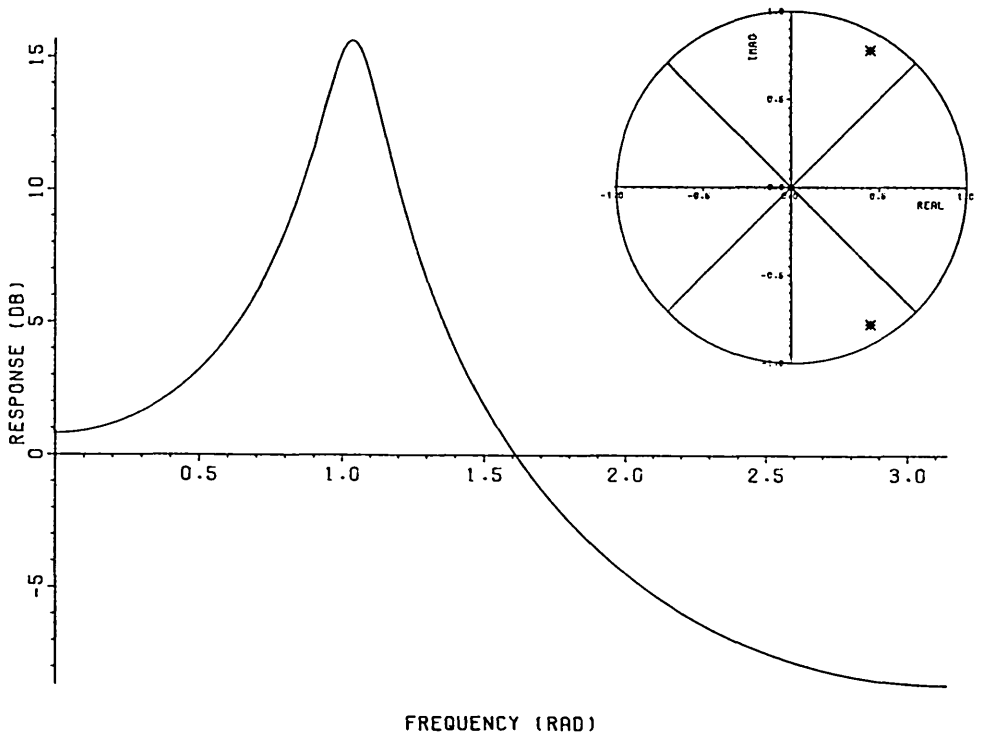
[5.24] BRISTOW, G., "Electronic Speech Recognition", *Collins, London*, 1986.

	ORIGINAL	SEGMENTED MODEL	ROOT SOLVING	PEAK PICKING
F_0	120	133	*	0
F_1	620	629	602	610
F_2	1220	1220	1214	1205
F_3	2550	2544	2557	2580
F_4	3300	3296	3325	3340
F_5	3750	3759	3765	*
B_0	100	103	*	*
B_1	80	86	98	100
B_2	40	38	45	68
B_3	140	136	162	173
B_4	250	258	299	339
B_5	200	218	261	*

Table 5.1 Results of Formant Extraction from a Synthetic Speech Signal for Various Formant Estimation Techniques

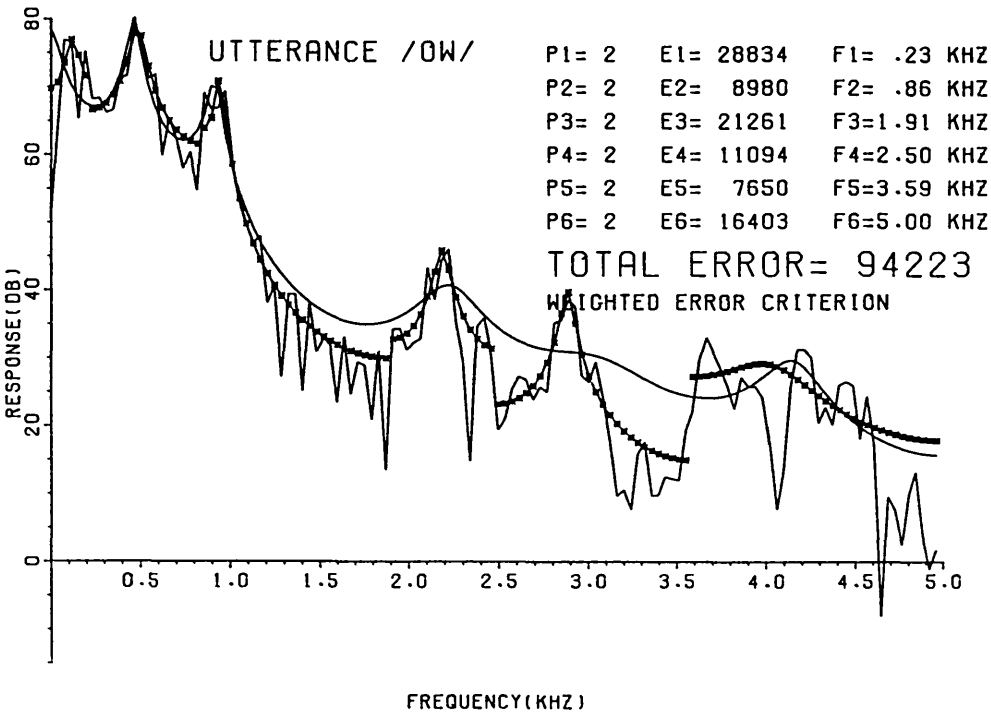


(a) Real Poles

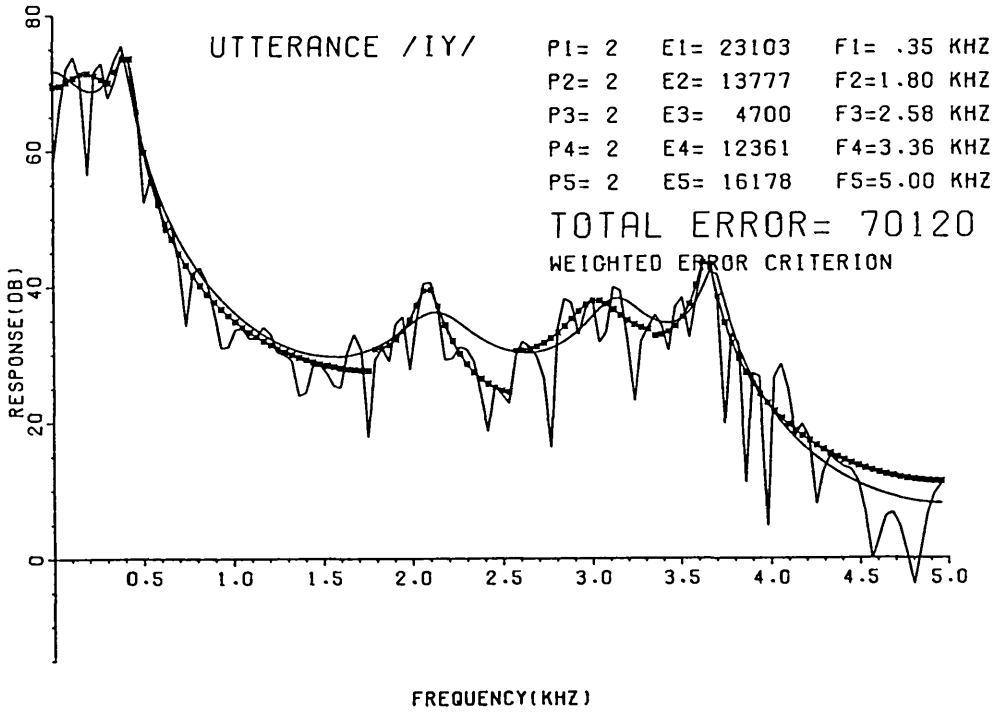


(b) Complex Poles

Figure 5.1 Magnitude Responses for 2 Distinct Cases of a Second Order LP System Transfer Function

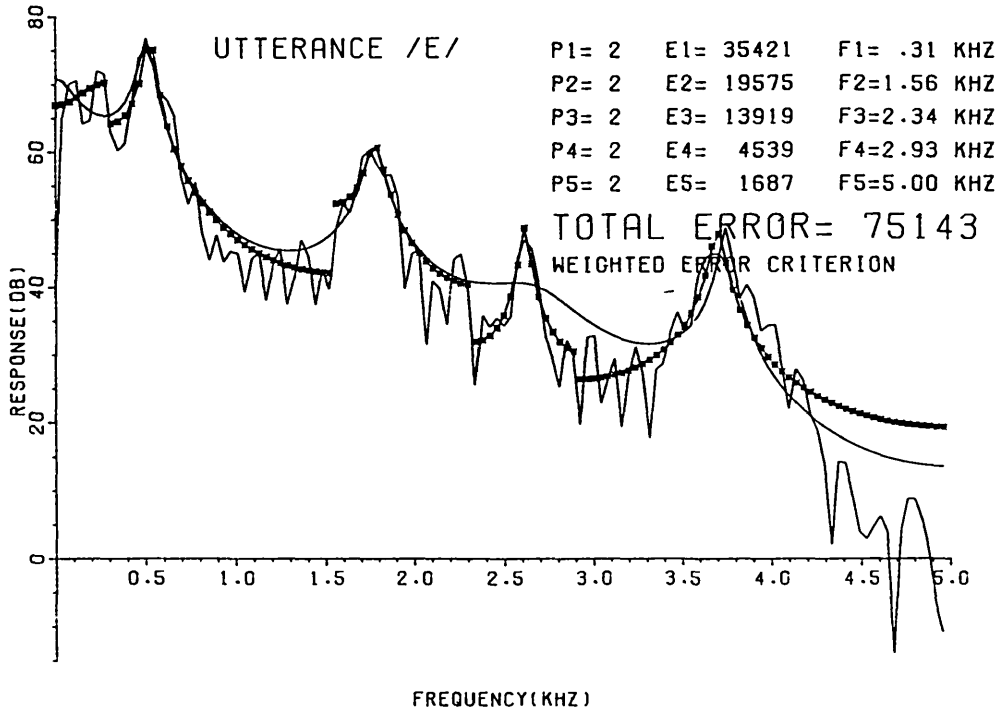


(a)

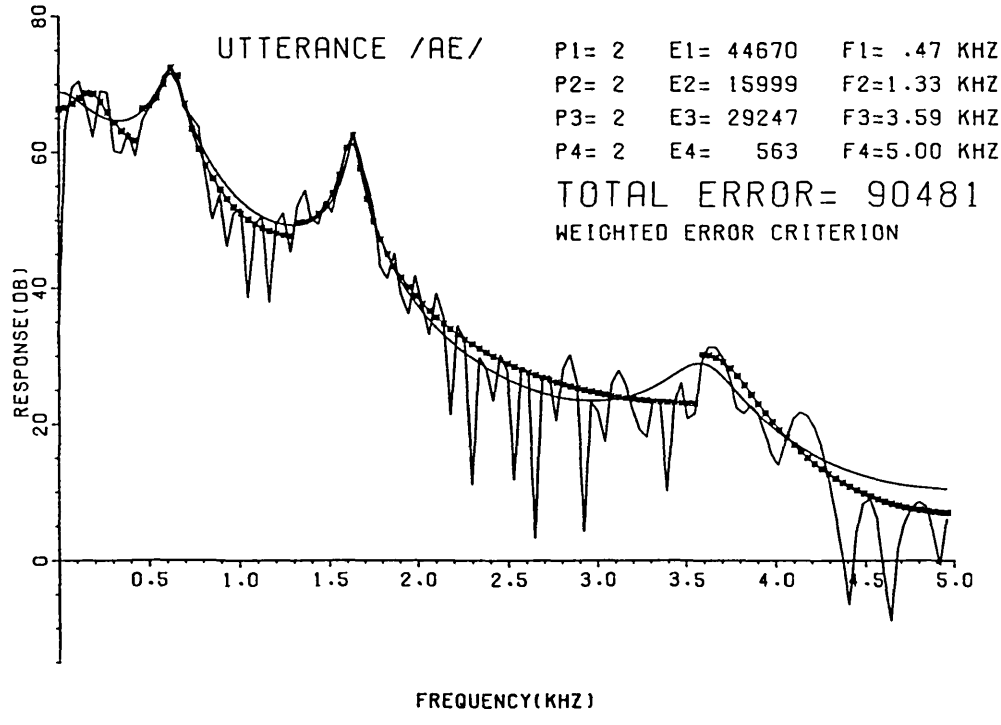


(b)

Figure 5.2 Spectral Segmentation into Formants for 4 Vowel Phonemes



(c)



(d)

Figure 5.2 (Contd.)

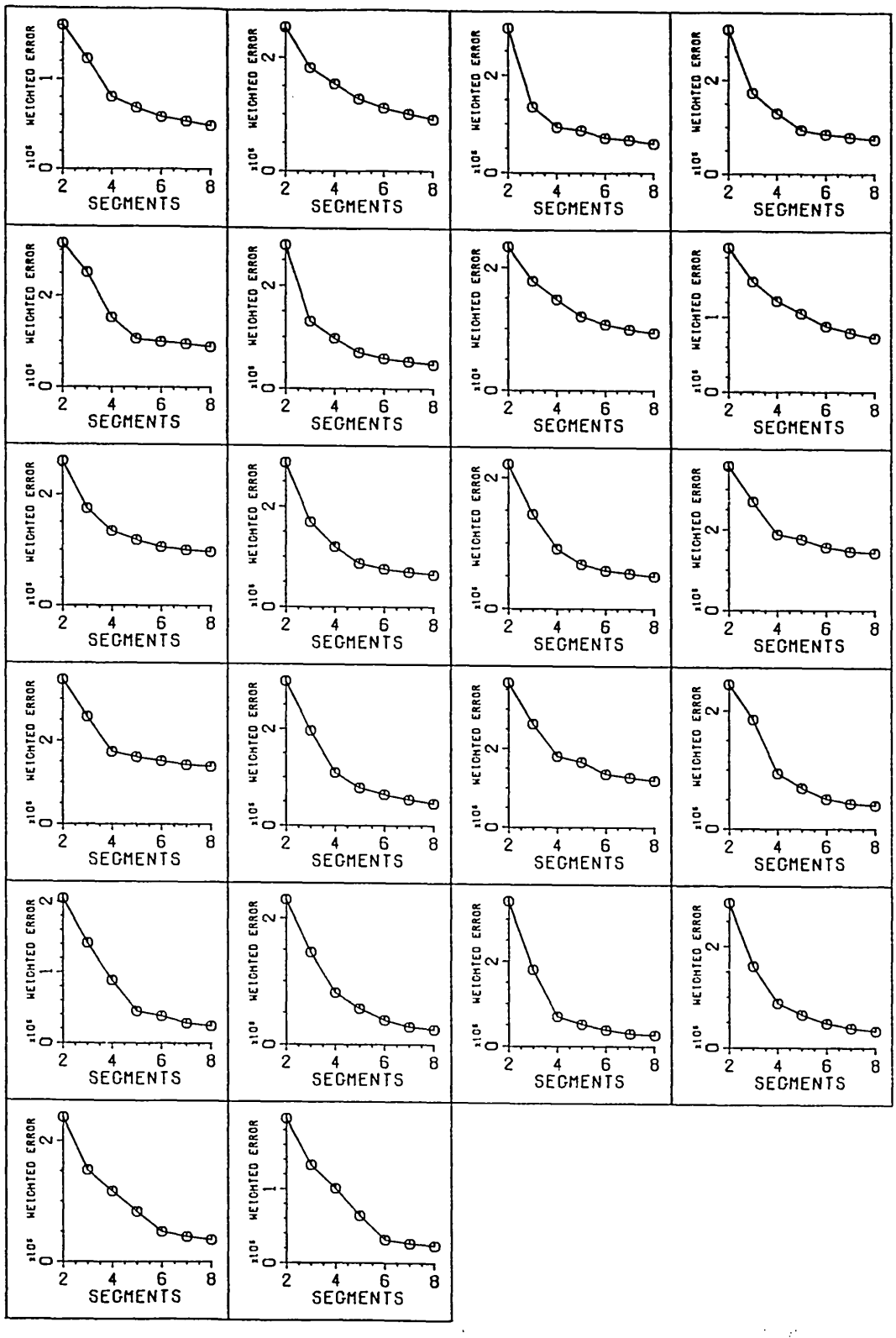


Figure 5.3 Weighted Error Curves for Spectral Segmentation into Formants for Several Speech Sounds

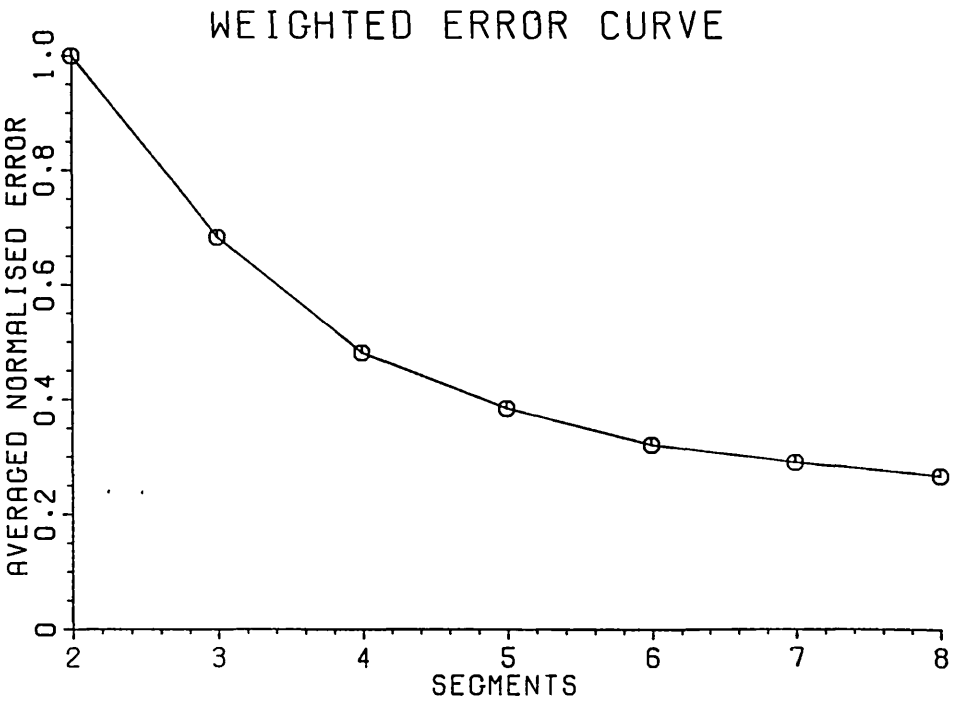


Figure 5.4 Illustration of Averaged Normalised Weighted Error Curve

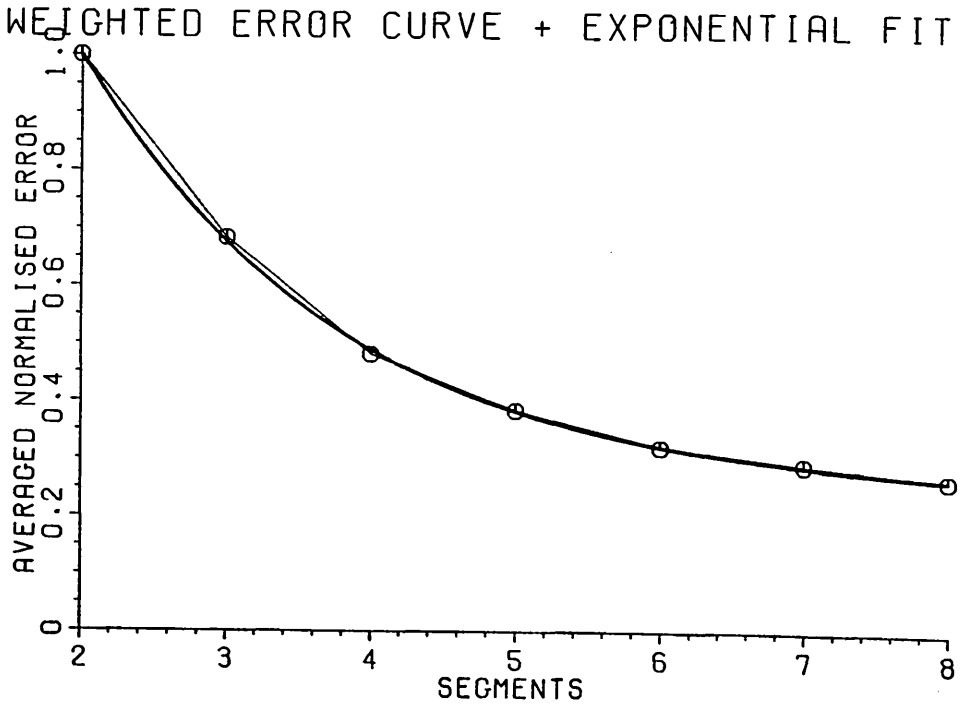


Figure 5.5 Comparison of Averaged Normalised Error Curve and the Corresponding Fit Based on an Exponential Decay with a Constant Offset

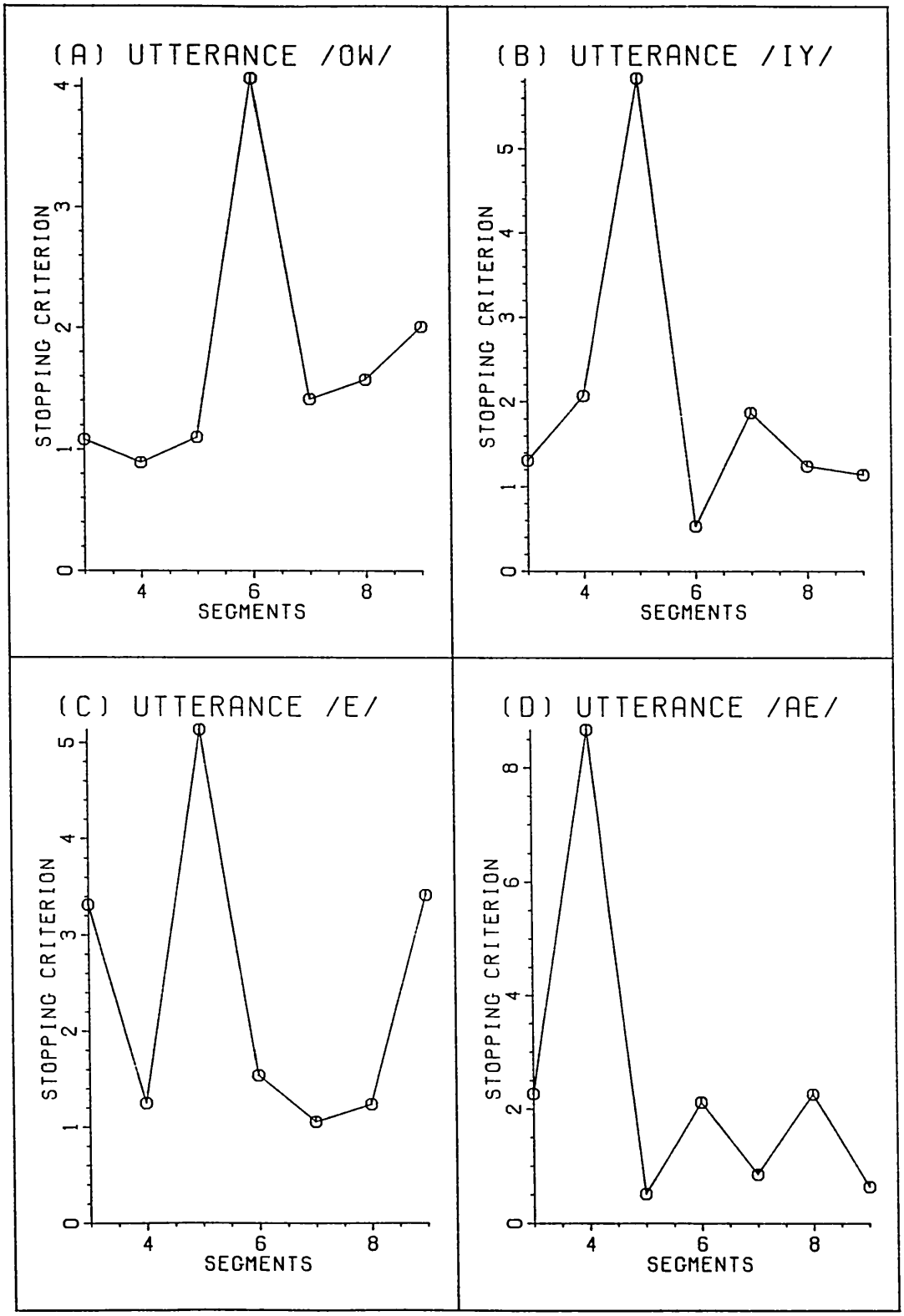


Figure 5.6 Stopping Criterion for Number of Regions to Employ in Spectral Segmentation into Formants for Various Voiced Utterances

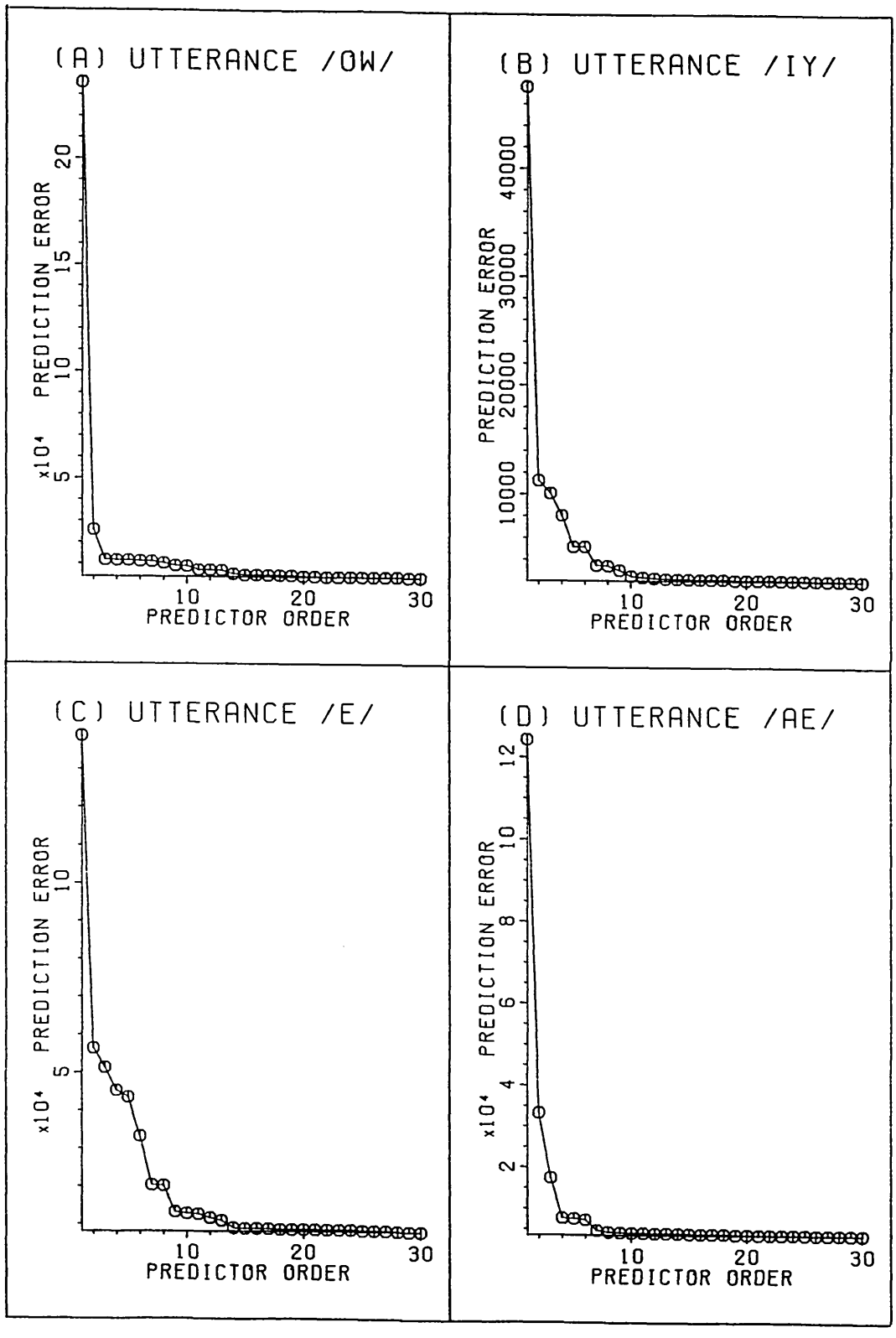


Figure 5.7 Prediction Error Curves for Various Voiced Utterances

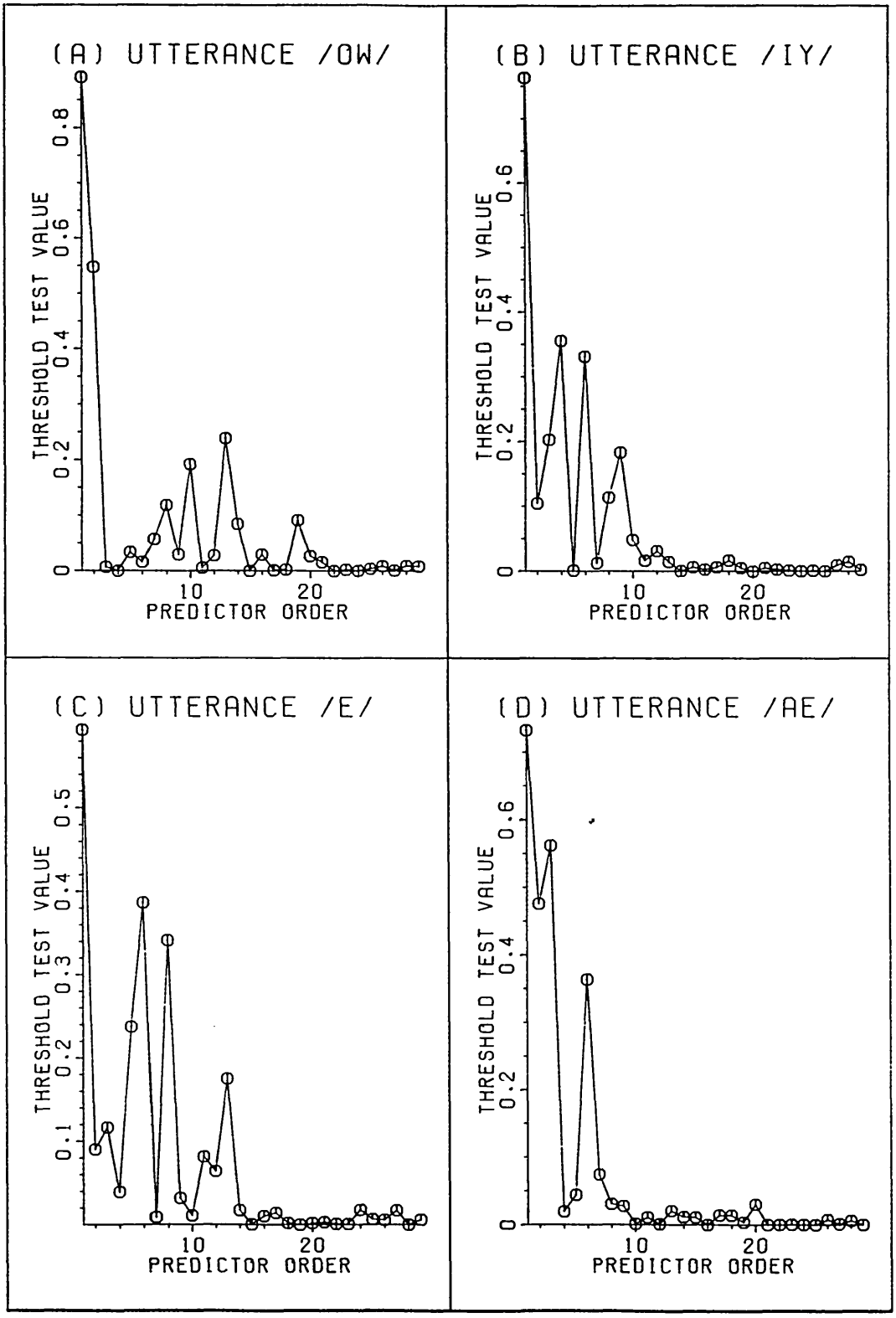


Figure 5.8 Behaviour of Threshold Test Criterion with Predictor Order for Various Voiced Utterances

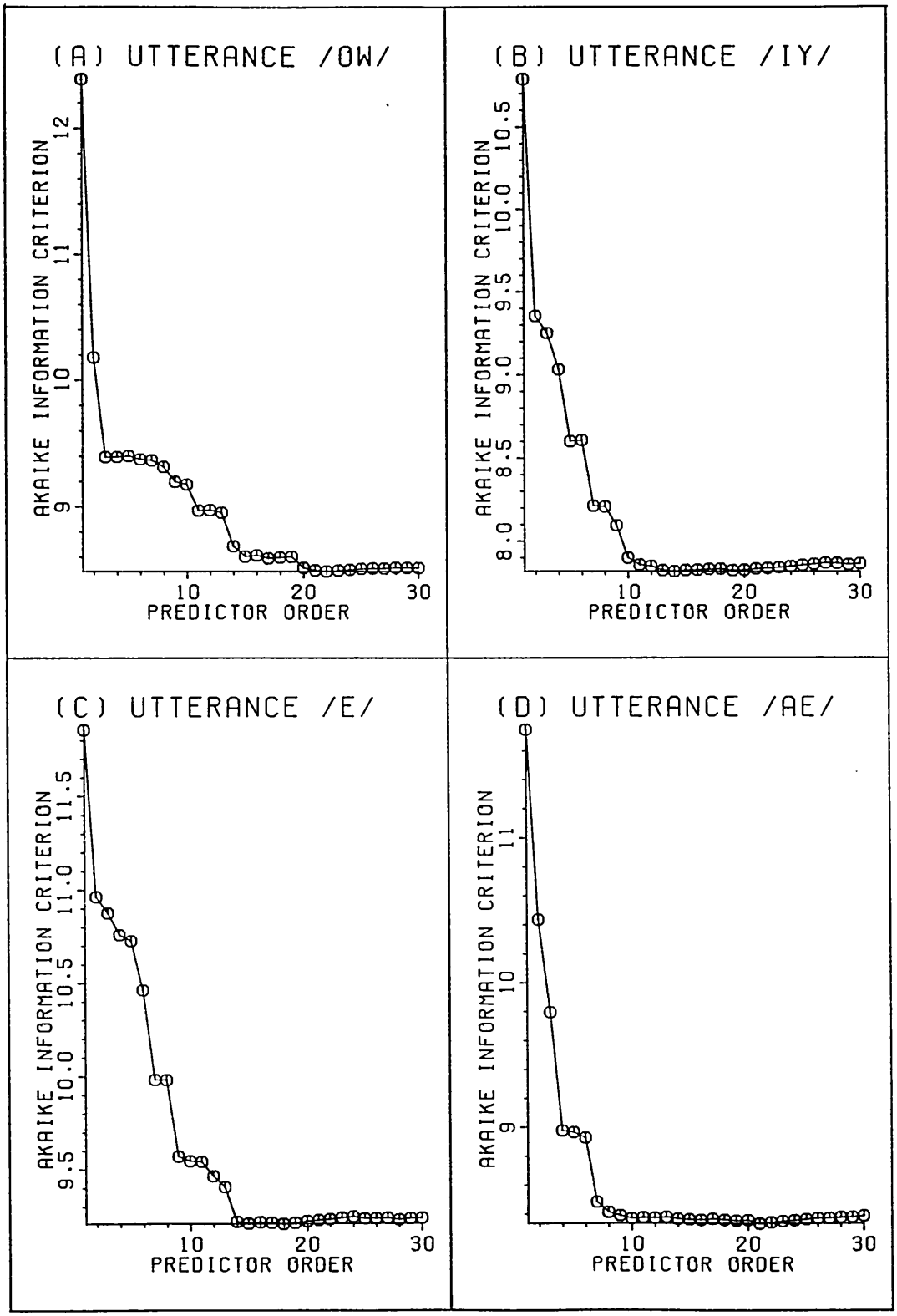


Figure 5.9 Behaviour of Akaike Information Criterion with Predictor Order for Various Voiced Utterances

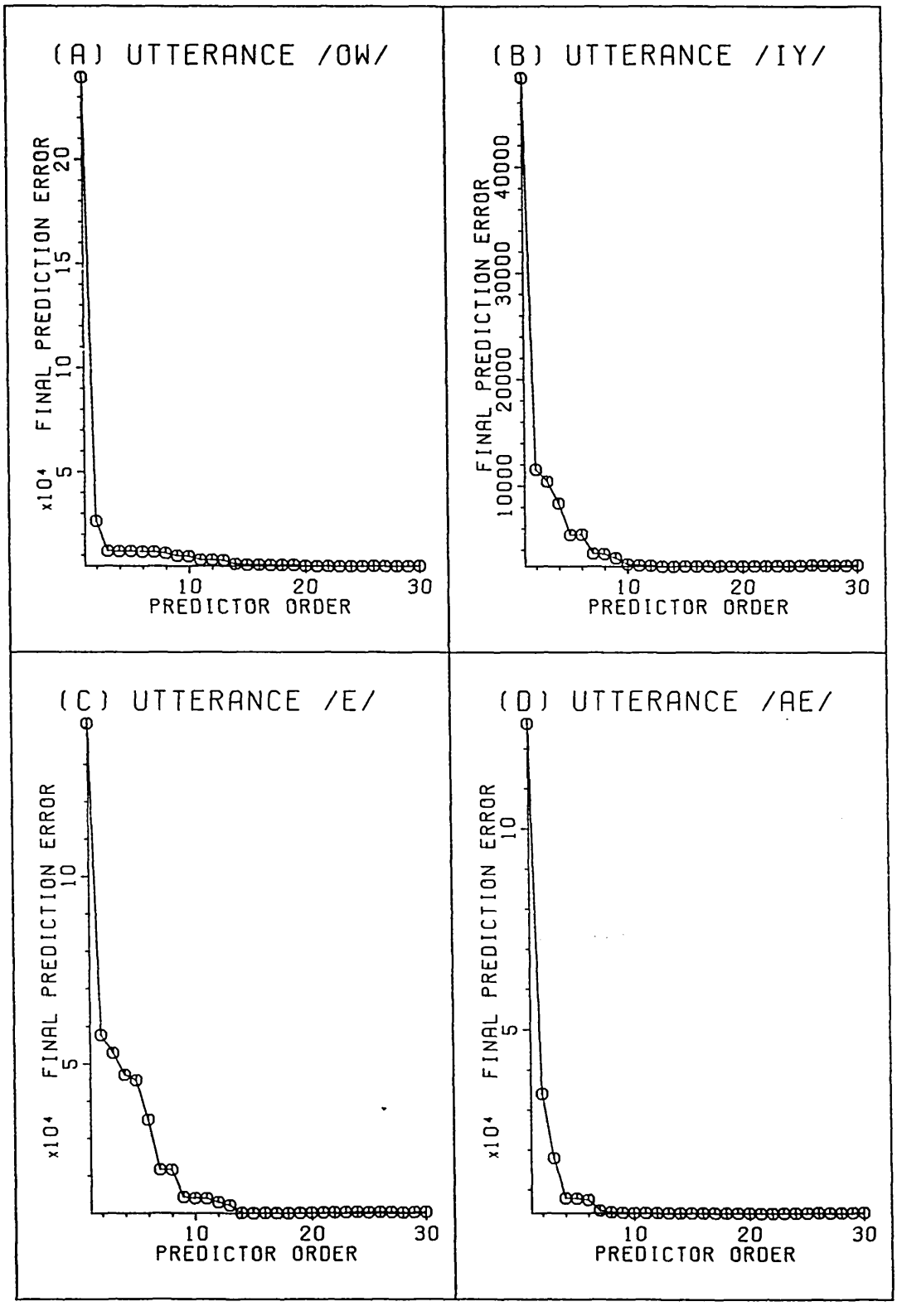


Figure 5.10 Behaviour of Final Prediction Error with Predictor Order for Various Voiced Utterances

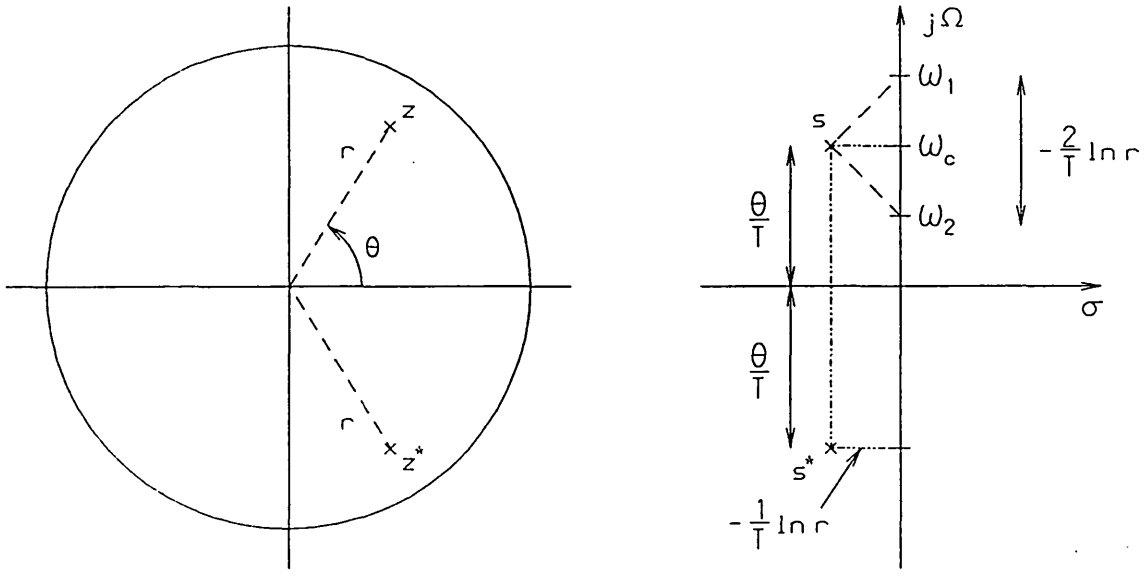


Figure 5.11 Illustration of a z-Plane to s-Plane Mapping for Formant Modelling

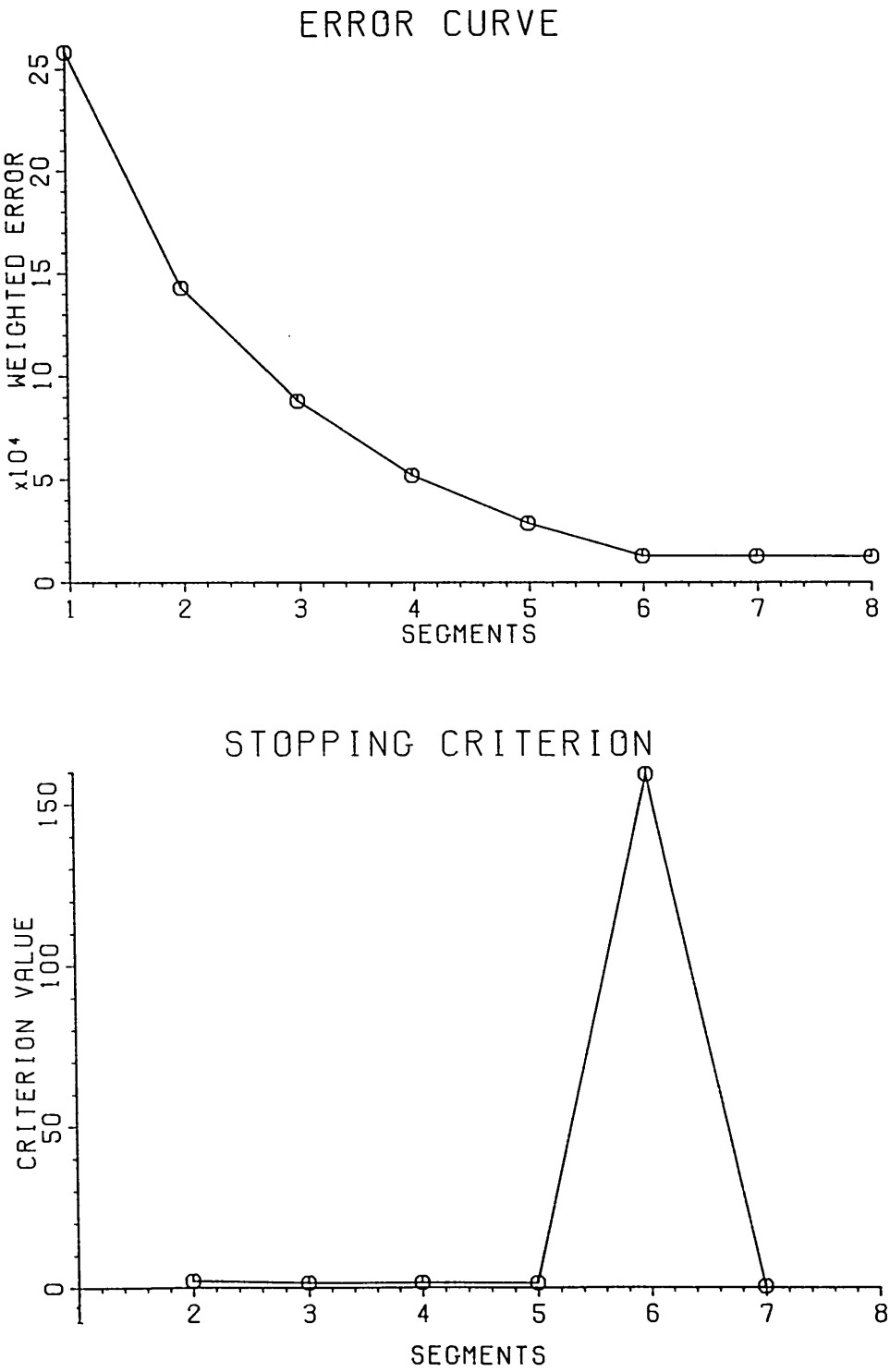


Figure 5.12 Weighted Error Curve and Stopping Criterion for Spectral Segmentation into Formants for a Synthetic Speech Signal

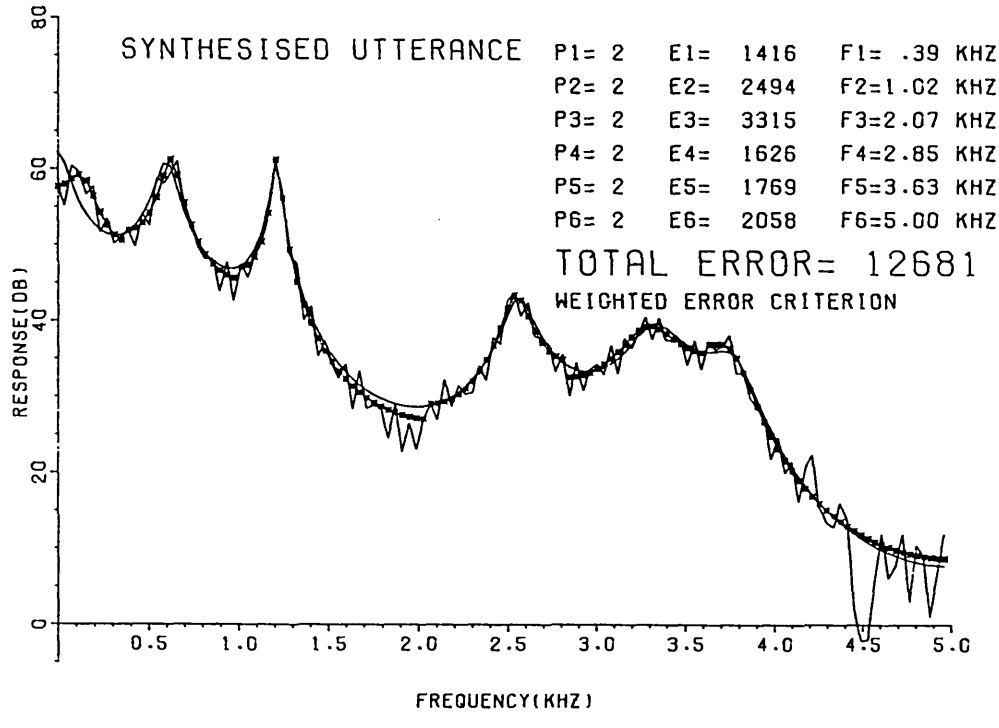


Figure 5.13 Illustration of Speech Spectral Segmentation into Formants for a Synthetic Speech Signal

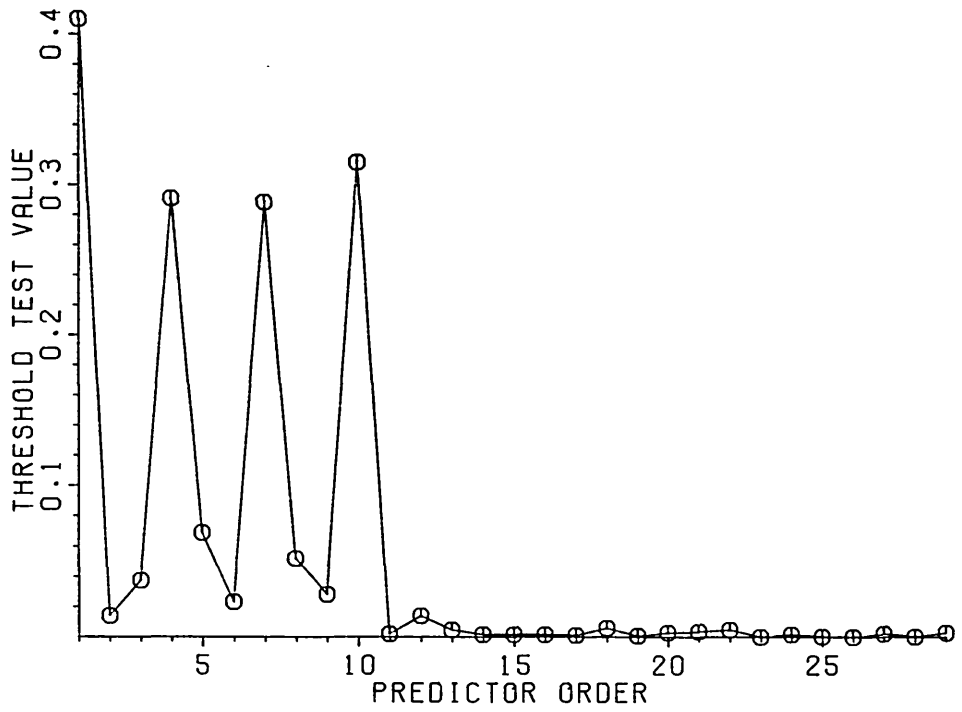


Figure 5.14 Variation of the Threshold Test Criterion with Predictor Order for a Synthetic Speech Signal

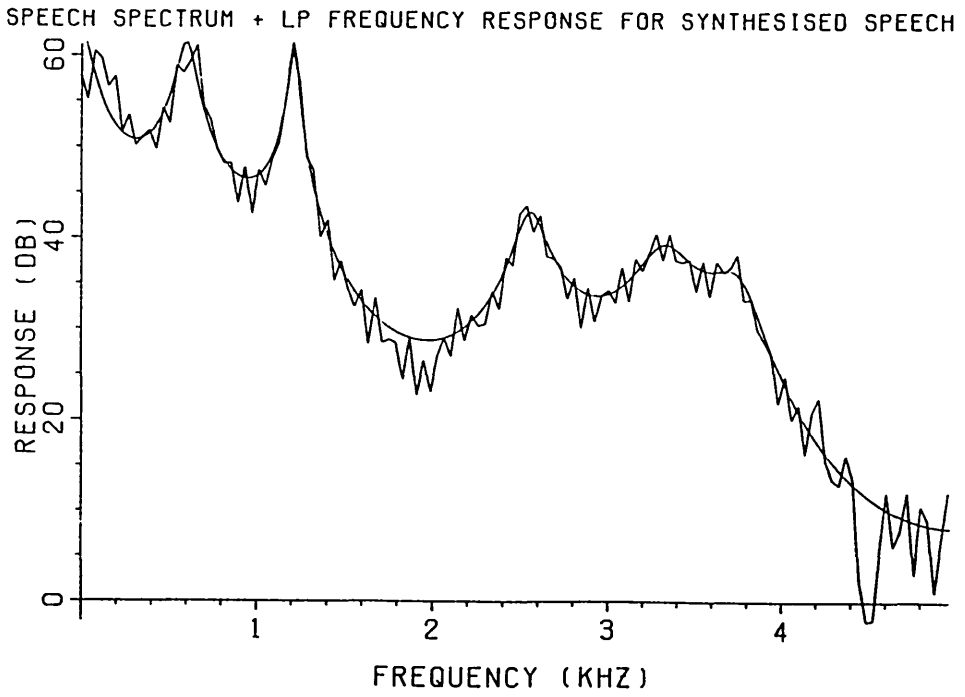


Figure 5.15 Comparison of Speech Spectrum and 11th Order LP Model Response for a Synthetic Speech Signal

CHAPTER 6

FORMANT SYNTHESISER MODEL REPRESENTATIONS

CHAPTER 6. FORMANT SYNTHESISER MODEL REPRESENTATIONS

6.1 INTRODUCTION

One of the fundamental goals of speech processing systems is the generation of synthetic speech from low information-rate control signals. This ability may be employed in a range of applications such as machine voice output, synthesis at the receiver of a vocoding system or simply to aid research into human speech production or perception [6.1].

Although several methods have been proposed for achieving this objective, one of the foremost techniques is the class of formant synthesiser models. These representations have the advantage that formant parameters are closely related to speech production and sound propagation in the vocal tract and that the acoustic output during vowel production may be specified accurately from a knowledge of formant positions and the fundamental frequency [6.2]. The speech formants are, therefore, important information-bearing elements of speech.

An associated property is that it may be argued that formants are good candidates for efficient speech coding [6.3]. This is since the human articulators, i.e. the tongue, lips, jaw etc. are massive elements and the associated muscles are obviously capable of exerting only a finite force so that the velocity of these articulators must be finite. As the formant frequencies are a direct function of the positions of these articulators, the speech formants will only vary slowly[1] during natural speech production and can, therefore, be updated relatively infrequently in a synthesiser configuration [6.4]. Another vital factor which accounts for the widespread use of formant synthesisers is their ability to produce good quality, natural-sounding speech [6.5].

---

[1] *One obvious exception to this slowly varying formant structure occurs when there is a complete closure of the vocal tract — for example, during the production of a stop consonant. However, the general principle does hold for the majority of vowel and vowel-like vocalisations.*

In their most basic form, formant synthesisers can generate speech from formant frequencies and bandwidths and fundamental frequency information. However, even at this stage, there is a fundamental dichotomy between two different types of synthesiser configurations — cascade and parallel. The first part of this chapter is thus devoted to a discussion of the merits and deficiencies of these two modes of connection.

The majority of this chapter is concerned with an assessment of the performance of the formant extraction algorithms considered in the previous chapter by employing the estimates obtained from these methods in a formant synthesiser network. The type of synthesiser configuration which will be utilised in this analysis is, therefore, the next subject of discussion. The spectra produced for the synthesised output of this configuration for formant estimates calculated using the procedure for spectral segmentation into formants are then compared with those available using the techniques based on standard LP analysis considered in Chapter 5.

For the sake of simplicity, the various methods of analysis which have been developed in this thesis to this point have been primarily concerned with data consisting of isolated speech frames. However, if these techniques are to be of more widespread use, their applicability must be extended to complete speech utterances and it is this problem which is addressed in section 6.4. The approach adopted for the study of continuous speech is that of pitch synchronous analysis (as described in section 2.2) and thus a method is first introduced for the estimation of the speech fundamental frequency. In this chapter, this technique is based on the electroglottographic signal which is used to isolate the individual pitch periods constituting the speech waveform.

The spectra corresponding to these distinct cycles of the speech signal are then employed to compute formant estimates for that particular time interval. It is also necessary to slightly modify the constraints associated with the procedure for spectral segmentation into formants for the pitch synchronous data and these modifications are also detailed in section 6.4. The sets of formant estimates for the consecutive pitch periods are finally incorporated

into the formant synthesiser along with the fundamental frequency information available from the electroglottographic analysis. The output speech can then be compared with the synthetic vocalisations produced by formant estimation from standard LP analysis.

6.2 IMPLEMENTATIONS OF DIGITAL FORMANT SYNTHESISERS

6.2.1 Comparison of the Cascade and Parallel Formant Configurations

The general structures of the cascade and parallel modes of connection were described in section 2.4 and it was demonstrated that the transfer function, $V_c(z)$, of a cascade model for the vocal tract consisting of M second order resonators is given by:

$$V_c(z) = \prod_{k=1}^{M} \left[\frac{G_k}{\left(1 - 2 r_k \cos\theta_k \, z^{-1} + r_k^{\,2} \, z^{-2} \right)} \right] \tag{6.1}$$

where $\{G_k : k=1,\ldots,M\}$ is the set of gains associated with the various resonators[2] and $\{r_k \, e^{+j\theta_k}, \; r_k \, e^{-j\theta_k} : k=1,\ldots,M\}$ denotes the set of complex conjugate poles which constitutes the cascade model. It should be noted that the above expression does not take account of the spectral effects of the excitation source and the radiation load and the manner in which these factors may be incorporated into the model is described later in this section. In the parallel model, the corresponding transfer function, $V_p(z)$ can be expressed as:

$$V_p(z) = \sum_{k=1}^{M} \left[\frac{A_k}{\left(1 - 2 r_k \cos\theta_k \, z^{-1} + r_k^{\,2} \, z^{-2} \right)} \right] \tag{6.2}$$

---

[2] *As mentioned in Chapter 2, these gain factors are often chosen such that the gain of each resonator is unity at zero frequency [6.6]. They are, therefore, completely specified by the choice of* $\{r_k, \theta_k\}$ *and are thus not usually classed as variables associated with the model.*

where $\{A_k : k=1,\ldots,M\}$ are the resonator gains. Unlike the cascade model, these amplitudes are individually controllable and can be used to alter the spectral characteristics of the complete synthesiser.

However, the choice of $\{A_k : k=1,\ldots,M\}$ is fairly critical since this configuration will only produce an identical output to that for a cascade model with the same resonator parameters under the influence of certain constraints. These restrictions are set out below and are chosen such that the partial fraction expansion of the vocal tract response shown in equation (6.1) only produces constant numerator terms [6.6] which can then be replaced by the gain values $\{A_1,A_2,\ldots,A_M\}$.

(1) The formant bandwidths are all equal.

(2) The formant gain parameters $\{A_1,A_2,\ldots,A_M\}$ correspond to the correct partial fraction coefficients.

Even small deviations from these conditions may lead to the introduction of zeros in the system transfer function which can appreciably upset the amplitudes of spectral components between formants. Larger changes in the formant gains to achieve intended formant amplitude variations are likely to have disastrous effects on the overall spectral shape [6.1].

A variety of techniques may be used to alleviate these problems. Holmes [6.7],[6.8] has advocated an approach whereby the output of each formant resonator is passed through a spectrum weighting filter which is designed to substantially reduce any components outside the frequency range permitted to the associated formant. This reduces the possibility of the response in the neighbourhood of a formant being affected by the skirt responses of adjacent formants.

Another strategy adopted by Klatt [6.9] employs a set of strict interrelationships between formant centre frequencies, bandwidths and amplitudes to achieve the same result. These rules include constraints such as:

- The bandwidth of a given formant peak may only be varied such that the amplitude of this resonator is inversely proportional to its bandwidth.

- Formant frequencies are adjusted such that the resonator gain is proportional to the centre frequency.

- A change in one formant frequency will also affect the amplitudes of higher formant peaks by a factor proportional to frequency squared.

Therefore, it will be difficult to use formant information extracted directly from actual speech data in such a formant model since it is unlikely to conform to these restrictions. In any case, it may be noted that transfer function zeros may still occur in the model spectrum even if these rules are observed [6.9].

In contrast, cascade models possess the inherent advantage that the relative amplitudes of the formant peaks appear to be correct [6.10] without the need for independent amplitude controls for each formant. As noted above, not only does the introduction of formant gain parameters in parallel configurations require additional control information, incorrect values for these parameters may also lead to the presence of extraneous zeros or an inappropriate spectral distribution in the model frequency response [6.1]. However, this property of parallel synthesisers may also be used to positive effect in the case of nasals, fricatives and plosives where required spectral features such as zeros can be introduced without resorting to additional filter paths — as is necessary for cascade models.

Another advantage of the cascade configuration is its ability to represent the vocal tract transfer function during non-nasalized speech more accurately than a parallel model. For this reason, it is more difficult to match the spectra of certain vowels using a parallel formant synthesiser [6.9]. In fact, for the production of non-nasalized vowel or vowel-like sounds, the cascade model seems to offer no significant disadvantages while resulting in a significantly reduced system complexity. As it only these vocalisations which are considered in the procedure for spectral segmentation into formants, the cascade configuration will be employed in the remainder of this chapter to assess the various formant extraction algorithms.

6.2.2 Cascade Synthesiser Configuration for Formant Synthesis

The operation of the cascade formant synthesiser considered in this subsection is based on the conventional model of speech production discussed in Chapter 2. The speech signal is, therefore, envisaged as the output of a time-varying filter — modelling the response of the vocal tract — which is driven by a substantially separable excitation function. In this case, the time-varying filter consists of a cascade connection of second order resonators.

The type of source-filter configuration to be investigated [6.11],[6.12] is shown in Fig. 6.1 along with the corresponding spectral trends at the intermediate points. This simple network constitutes only that part of a typical digital cascade synthesiser which is concerned with the generation of non-nasalized voiced utterances. The functions of each of the separate units in this synthesiser will first be described individually.

(a) Impulse Train Generator

This device produces a sequence of scaled impulses spaced at the required fundamental period. The output spectrum of this block will, therefore, be substantially flat. This form of excitation is as described for the source-filter model of Chapter 2 and some justification for this type of synthesiser input is also provided by the discussion of the nature of the excitation source later

in this subsection.

(b) Glottal Pulse Model

This unit accounts for the frequency characteristic of the excitation source or glottal waveform which, as mentioned in section 2.4, is associated with a −12 dB/octave decay in the frequency domain.

(c) Vocal Tract Filter

The natural resonances of the vocal tract are represented by this chain of second order resonators in series. The overall frequency profile of this filter is substantially flat [6.6] with peaks superimposed at the formant frequencies.

(d) Radiation Compensation

This network accounts for the effect of radiation of speech from the lips which is characterised by a '*differentiating*' action [6.6] and thereby produces a +6 dB/octave lift in the frequency response. This spectral trend combines with the −12 dB/octave fall-off of the excitation source to yield the expected −6 dB/octave slope of the output speech.

It should be noted that the cascade of formants shown in Fig. 6.1 does not numerically match the sequence of regions resulting from the procedure for spectral segmentation into formants as discussed in Chapter 5. This is since the first region in the partitions of Fig. 5.2? actually corresponds to the glottal formant (also known as the fundamental or Formant 0) [6.13] and, therefore, region 2 descibes the first formant, etc.

As suggested by the term Formant 0 for the glottal formant, an extra resonator, $R0$, is often added to the vocal tract model instead of a separate glottal pulse model. Since Formant 0 is effectively a single complex pole near d.c., its frequency characteristic will fall at −12 dB/octave and so produce the required response to model the excitation source discussed

above. Hence, in this case, a combined vocal tract and excitation source filter is produced with a net −12 dB/octave frequency response — as illustrated in Fig. 6.1. Therefore, as the segmentation approach yields information on the position of Formant 0, this is the approach which is adopted for the cascade modelling of the formant information derived in Chapter 5 [6.12].

It can also be demonstrated as follows that the impulse train is an appropriate excitation for voiced speech sounds. Let $V(z)$ be the z-transform of the vocal-tract impulse response while $R(z)$ is that of the radiation characteristic. Let $g(n)$ be the glottal waveform exciting $V(z)$, $R(z)$ to produce the output speech $s(n)$. According to the source-filter model, the z-transform of $s(n)$ is then given by

$$S(z) = G(z) . V(z) . R(z) \tag{6.3}$$

It is instructive to use an alternative, but mathematically equivalent, description. Let $g(n)$ be the output of a double integrator system excited by the second derivative $g^{(2)}(n)$ of $g(n)$. The cascade of the double integrator and $V(z)$ is denoted by $\hat{V}(z)$. Thus,

$$S(z) = G^{(2)}(z) . \hat{V}(z) . R(z) \tag{6.4}$$

where $G^{(2)}(z)$ is the z-transform of $g^{(2)}(n)$. The waveform $g^{(2)}(n)$ is known to be well represented by a sequence of impulses spaced at the pitch period [6.14] and to possess a generally flat spectral characteristic [6.15]. This waveform can, therefore, be approximated by the output of the impulse train generator. The double integrator is associated with a −12 dB/octave decay and, therefore, $\hat{V}(z)$ represents the combined effect of the vocal tract and the excitation source. Thus, the overall model described by $G^{(2)}(z)$, $\hat{V}(z)$, $R(z)$ is equivalent to the cascade formant model described above which consists of the impulse train generator, a combined glottal pulse and vocal tract model and a radiation compensation filter.

Note

In analogue cascade models, it is necessary to implement a fixed "*higher-pole correction network*" to account for the effect of all those poles which are positioned at a higher frequency than the folding frequency (half the sampling frequency) of the synthesised output. However, owing to the frequency domain periodicity of sampled data systems, any one pole also represents an infinity of other poles with a spacing equal to the sampling frequency. Therefore, the use of a digital system automatically guarantees the presence of higher order poles and obviates the necessity of an external higher-pole correction network [6.5].

6.3 EVALUATION OF FORMANT EXTRACTION ALGORITHMS BY CASCADE MODELLING

6.3.1 Synthesis of Formant Estimates Derived from Spectral Segmentation into Formants

The procedure outlined in Chapter 5 for partitioning a speech spectrum into its constituent formants and the subsequent analysis of the various regions to estimate formant centre frequencies and bandwidths was applied to the same 4 vowel sounds shown in Fig. 5.2?. The results of this formant extraction process were then incorporated into a cascade model of the form described in the previous section. Using the pole positions corresponding to the various formants and taking into account the spectral trends required for the excitation source and radiation load, it is possible [6.11] to produce a cascade formant model frequency response associated with the particular utterance under investigation.

These frequency characteristics are illustrated in Fig. 6.2 (a)–(d) along with the corresponding speech spectra. The gain of the synthesiser network was chosen to match the total energy in the real and model frequency responses in each case and thus facilitate a qualitative comparison between the two types of response. These plots indicate a close similarity between the signal and cascade model spectra for most cases. It may be suggested

that one of the reasons for any discrepancies lies in possible inaccuracies arising from the calculation of a formant bandwidth (BW) from the associated pole radius (r) in Chapter 5. The expression used to relate these two quantities was described as follows:

$$\text{BW} = -\frac{f_s}{\pi} \ln r \quad \text{(Hz)} \tag{6.5}$$

$$\text{or} \quad b = -2 \ln r \quad \text{(rad)} \tag{6.6}$$

where f_s is the signal sampling rate and b is expressed as a normalised frequency. However, the analysis leading to these results was based on the premise that the system response in the region of a particular pole was not affected significantly by any other poles. Since this assumption may not be an accurate description for some formants, it is possible that the bandwidth estimates employed to produce the results of Fig. 6.2 may be in error.

To determine the importance of the above simplification on the model spectra, it is necessary to determine an expression for formant bandwidth which takes into account the effect of both complex poles in the Selective LP model. This requirement can be met as outlined in the following stages.

Step 1 Establish Magnitude Response associated with the Complex Pole Pair.

For a system consisting of a pair of complex conjugate poles of the form:

$$z, z^* = r\, e^{\pm j\theta} \tag{6.7}$$

The power spectrum $P(\omega)$ will be given by

$$P(\omega) = \frac{1}{\left[(\cos\omega - r\cos\theta)^2 + (\sin\omega - r\sin\theta)^2\right]} \times$$

$$\frac{1}{\left[(\cos\omega - r\cos\theta)^2 + (\sin\omega + r\sin\theta)^2\right]}$$

$$= \frac{1}{\left[1 + r^2 - 2r\cos(\omega-\theta)\right]\left[1 + r^2 - 2r\cos(\omega+\theta)\right]} \tag{6.8}$$

Step 2 Determine frequency (ω_c) corresponding to the peak response.

Consider

$$G(\omega) = \left[1 + r^2 - 2r\cos(\omega-\theta)\right]\left[1 + r^2 - 2r\cos(\omega+\theta)\right] \tag{6.9}$$

As $f(x) = \frac{1}{x}$, $x > 0$ is a strictly monotonic decreasing function, $G(\omega)$ will have a minimum at the same frequency (ω_c) as the peak in $P(\omega)$.

Thus, $$\frac{dG}{d\omega} = 2r\sin(\omega-\theta).\left[1 + r^2 - 2r\cos(\omega+\theta)\right] + 2r\sin(\omega+\theta).\left[1 + r^2 - 2r\cos(\omega-\theta)\right]$$

$$= 2r(1+r^2)\left[\sin(\omega-\theta) + \sin(\omega+\theta)\right] - 4r^2\sin 2\omega$$

$$= 4r(1+r^2)\sin\omega\cos\theta - 4r^2\sin 2\omega \tag{6.10}$$

The function $G(\omega)$ will, therefore, have an extremum at the frequency ω_c such that

$$4r(1+r^2)\sin\omega_c\cos\theta - 4r^2.2\sin\omega_c\cos\omega_c = 0 \tag{6.11}$$

or $$\cos\omega_c = \left(\frac{1+r^2}{2r}\right)\cos\theta \tag{6.12}$$

Step 3 *Show that this extremum represents a minimum for $G(\omega)$.*

From equation (6.10),

$$\left.\frac{d^2G}{d\omega^2}\right|_{\omega=\omega_c} = 4r(1+r^2)\cos\omega_c\cos\theta - 8r^2\cos 2\omega_c \tag{6.13}$$

Using equation (6.12),

$$\left.\frac{d^2G}{d\omega^2}\right|_{\omega=\omega_c} = 4r.2r\cos\omega_c.\cos\omega_c - 8r^2(2\cos^2\omega_c - 1)$$

$$= 8r^2(1 - \cos^2\omega_c) = 8r^2\sin^2\omega_c \tag{6.14}$$

This result will, therefore, always be positive for all (r,θ) for which there is an extremum and thus $G(\omega)$ has a minimum at the frequency ω_c.

Step 4 *Establish inverse relation to calculate ω for given value of $G(\omega)$.*

From equation (6.9),

$$\begin{aligned} G(\omega) &= (1+r^2)^2 - 2r(1+r^2).2\cos\omega\cos\theta + 4r^2\cos(\omega-\theta).\cos(\omega+\theta) \\ &= (1+r^2)^2 - 4r(1+r^2)\cos\theta.\cos\omega + 4r^2(\cos^2\omega - \sin^2\theta) \end{aligned} \tag{6.15}$$

or $$4r^2\cos^2\omega - 4r(1+r^2)\cos\theta.\cos\omega + \left[(1+r^2)^2 - G(\omega) - 4r^2\sin^2\theta\right] = 0$$

Thus, $$\cos\omega = \frac{4r(1+r^2)\cos\theta}{8r^2} \pm \frac{\sqrt{16r^2(1+r^2)^2\cos^2\theta - 16r^2\left[(1+r^2)^2 - G(\omega) - 4r^2\sin^2\theta\right]}}{8r^2} \tag{6.16}$$

$$= \left(\frac{1+r^2}{2r}\right)\cos\theta \pm \frac{\sqrt{-(1+r^2)^2\sin^2\theta + G(\omega) + 4r^2\sin^2\theta}}{2r}$$

$$\cos\omega = \left(\frac{1+r^2}{2r}\right)\cos\theta \pm \frac{\sqrt{G(\omega) - (1-r^2)^2\sin^2\theta}}{2r} \tag{6.17}$$

Therefore, as expected, each particular value of $G(\omega)$ around the minimum occurs at two different frequencies (except for the lowest value $G(\omega_c)$).

Step 5 *Obtain the 3 dB frequencies ω_1, ω_2.*

At the frequency ω_c, from equations (6.12) and (6.17),

$$G(\omega_c) = (1-r^2)^2\sin^2\theta$$

$$\text{or} \quad P(\omega_c) = \frac{1}{G(\omega_c)} = \frac{1}{(1-r^2)^2\sin^2\theta} \tag{6.18}$$

Now, at the 3 dB frequencies ω_1, ω_2 either side of the peak in $P(\omega)$,

$$P(\omega_1) = P(\omega_2) = \frac{P(\omega_c)}{2}$$

$$\text{Hence,} \quad G(\omega_1) = G(\omega_2) = 2\,G(\omega_c) = 2\,(1-r^2)^2\sin^2\theta \tag{6.19}$$

So, according to equation (6.17), the frequencies ω_1, ω_2 occur when

$$\cos\omega = \left(\frac{1+r^2}{2r}\right)\cos\theta \pm \left(\frac{1-r^2}{2r}\right)\sin\theta \tag{6.20}$$

Thus,

$$\text{Pole Bandwidth } (b) = |\omega_1 - \omega_2| = \cos^{-1}\left[\left(\frac{1+r^2}{2r}\right)\cos\theta - \left(\frac{1-r^2}{2r}\right)\sin\theta\right] - \cos^{-1}\left[\left(\frac{1+r^2}{2r}\right)\cos\theta + \left(\frac{1-r^2}{2r}\right)\sin\theta\right] \tag{6.21}$$

The formant positions for the results of Fig. 6.2 were, therefore, recalculated on the basis of this relationship between the z-plane position of the complex conjugate pole pair

($re^{\pm j\theta}$) and the corresponding pole bandwidth (b). However, this alteration was found to produce changes of less than 5 % in the various formant bandwidths and resulted in no visually evident deviations from the model spectra of Fig. 6.2.

Perhaps a more important factor to bear in mind when considering the causes of discrepancies between the signal and model frequency responses of Fig. 6.2 is that there is no direct amplitude control for individual formants in a cascade model. As discussed in the previous section, the model spectra will nevertheless be almost correct in most cases and this problem is not usually a severe drawback. However, slight errors in levels for vowel formants can occur — as illustrated in Fig. 6.2.

6.3.2 Comparative Modelling for Standard LP Techniques

Since the method of spectral segmentation into formants is computationally expensive when compared to standard LP modelling, it is important to assess the advantages this method offers over the conventional approach of formant position estimation from linear predictive analysis [6.12]. Two different approaches to formant extraction were outlined in section 5.5 — root-solving and peak-picking. As both techniques were found to produce very similar results, the following discussion will first centre on the root-solving procedure — which is more amenable to automatic analysis. This method was used to determine formant frequencies and bandwidths from complex pole positions for the utterance employed in Fig. 6.2 (a) using two different predictor orders.

(i) $p=12.$ This value is the same as the overall order of the 6 region Selective LP model shown in Fig. 6.2 (a).

(ii) $p=20.$ The weighted error between the signal spectrum and LP model frequency response in this case is approximately equal to that for the segmented Selective LP model of Fig. 6.2 (a).

The cascade model frequency responses resulting from these two cases are shown in Fig. 6.3 (a),(b) and a comparison with the spectra of Fig. 6.2 (a) illustrates that the standard models based on single region LP responses are less accurate representations of the original spectrum than that resulting from a Selective LP model. Some of the reasons for these differences are explained as follows:

- The method described in section 5.4 for the evaluation of formant bandwidths from complex pole radii assumes that the remaining poles have a negligible influence in the 3 dB region of a particular pole. Although this supposition was demonstrated above to be adequate, in general, for a 2nd order Selective LP model where there is only complex pole in the z-domain, it may well be inaccurate in, for example, a 12th order LP model where it is more likely that two poles will be closely spaced and thereby result in erroneous pole estimates.

- The other difficulty with the root-solving method is that it involves a choice between those complex poles which correspond to speech formants and others which are associated with the shaping characteristic of the LP system response. A threshold is, therefore, required for formant bandwidths above which all complex poles are rejected in the formant extraction process. An inappropriate level will thus either result in the preclusion of some true speech formants or the acceptance of those poles which are actually responsible for spectral shaping.

The alternative to the root-solving approach is the peak-picking procedure. Although the latter technique is generally less complex and requires a lower computational load, it is also associated with a major disadvantage.

- A calculation of formant bandwidth directly from LP spectra requires a 3 dB fall-off on each side of the spectral peak and this situation is unlikely to occur either when two formants are close together or when a particular pole is associated with a low radius.

There are also three other general points [6.11] which apply to both the root-solving and peak-picking methods.

- *Absence of Spectral Shaping Characteristic*

As mentioned in section 5.2, the predictor polynomial,

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \tag{6.22}$$

can be partitioned into two separate system functions, $F(z)$, $S(z)$ such that the LP system response is given by:

$$H(z) = \frac{G}{A(z)} = \frac{G}{F(z).S(z)} \tag{6.23}$$

The function $F(z)$ represents the combined effect of those complex conjugate roots which are associated with formant modelling while $S(z)$ is comprised of those real roots and high bandwidth complex roots which together produce the correct spectral shaping for the LP spectrum.

When either of the above procedures for formant estimation are implemented, only the roots of $F(z)$ are, therefore, selected and then employed to produce the cascade model. Hence, the roots of $S(z)$ are missing in this representation with the result that the model spectrum is totally distorted. This contrasts with the spectral segmentation technique where the Selective LP model is concerned with matching only a single formant peak rather than the entire

spectral range and the overall shape of the signal frequency response is of less importance.

• *Difficulties associated with the Choice of Predictor Order*

The problem of selecting an appropriate LP model order for formant analysis was addressed in section 5.3. It was noted that this choice is often critical with a value which is too low leading to an inadequate number of poles to characterise all of the formants in the speech spectrum. This situation is probably prevalent in the case of $p=12$ in Fig. 6.3 (a) since some of the 12 predictor coefficients are represented in the shaping characteristic which leaves insufficient poles to model the 6 formants which, according to the spectral segmentation technique, are exhibited by the signal spectrum.

However, introducing extra coefficients into the LP model can also cause errors since extraneous peaks may be selected in the signal spectrum which do not actually correspond to true speech formants. Although techniques exist for the estimation of a suitable LP model order, these methods were found to be not altogether satisfactory for their assigned purpose in section 5.3.3. This is in contrast to the scheme for spectral segmentation into formants for which an effective procedure aimed at model order selection for vowel sounds was developed in section 5.3.2.

• *Overestimation of High Frequency Formant Bandwidths*

Another point to note from the result of Fig. 6.3 (a) is that the linear predictive analysis procedure tends to overestimate the bandwidths of the speech formants — especially those at the higher frequencies. This characteristic was also observed in the results of section 5.5.3 which were concerned with the analysis of a synthetic speech signal and has also been noted previously in the literature [6.16],[6.17].

Formant bandwidths are a function of several factors including energy losses due to heat conduction, viscosity, cavity-wall motions and radiation of sound from the lips [6.9]. The importance of appropriate bandwidths in a formant synthesiser network is that they influence the effective intensity of a formant peak and thus must be selected properly to achieve natural-sounding speech [6.9],[6.18].

However, formant bandwidths are known to be difficult to deduce accurately from natural speech [6.4],[6.19]. Although linear predictive methods are commonly used to estimate these bandwidths [6.4], the above discussion indicates that this approach does have some drawbacks. The results of Fig. 6.2 show that the Selective LP approach has certain advantages in this respect as it differs from those spectra available using standard LP modelling in that the formant bandwidths appear to be relatively well preserved. This observation is also borne out by the analysis of section 5.5.3 where the Selective LP technique yielded much closer formant bandwidth approximations than either the root-solving or peak-picking methods.

6.4 SPECTRAL SEGMENTATION BASED ON PITCH SYNCHRONOUS ANALYSIS

6.4.1 Pitch Synchronous Processing of Voiced Speech

The methods of analysis employed during the course of this work for the segmentation of speech spectra and formant modelling have been largely applied to isolated speech frames in order to illustrate the general principles involved in the various techniques. In this section, these methods are extended to include input data consisting of complete utterances [6.12] so that the performance of the approaches developed in this thesis can be ultimately assessed by comparing the synthetic speech output available from these methods with the actual vocalisation.

As discussed in Chapter 2, one means for the analysis of a continuous speech waveform is to subdivide the signal into several fixed-size frames which are then processed individually. The formant parameters available from each of these frames can then be concatenated in a

synthesiser network to produce the output speech. The frame lengths for the results described in previous chapters and earlier in this chapter have usually been set at 25.6 ms. However, this interval may encapsulate 2–3 cycles of voiced speech and it was found that this frame length tended to produce a disjointed speech output which failed to track rapidly changing formant structures. Although shorter intervals may be employed, the frame lengths will then approach the pitch period and thus there may be a considerable variation in the nature of the speech data between consecutive frames [6.20] — depending on which part of the speech cycle falls within a particular frame.

One elegant solution to this problem lies in the use of "*pitch synchronous analysis*" [6.21]. In this technique, the speech waveform is partitioned into its constituent cycles and then each of these segments are analysed separately. Another consequence of pitch synchronous methods is also of interest for the work developed in this thesis.

In Chapter 5, the speech signal was described as a convolution of the excitation function and the vocal tract impulse response such that a speech spectrum computed on a pitch asynchronous basis will consist of ripples corresponding to harmonics of the excitation as well as the transfer function of the vocal tract. The occurrence of these "*pitch harmonics*" [6.19] tends to impede the process of formant extraction such that it is difficult to obtain accurate estimates of the formant parameters — especially bandwidth values [6:19]. However, as only a single cycle of the speech waveform is processed in pitch synchronous analysis, the pitch harmonics will not be evident in the corresponding spectra. This topic and other advantages of pitch synchronous methods will be developed further in Chapter 7.

A fundamental requirement for a pitch synchronous approach is obviously an accurate algorithm for the calculation of pitch period estimates. Several pitch determination techniques have been proposed in the past including the inverse filter approach [6.22] and the SIFT algorithm [6.23] which are both based on inverse filtering the speech signal (or a suitably filtered version of it) using an LP model. Autocorrelation methods, e.g. [6.24], generally rely

on searching for peaks in the speech short-time autocorrelation function while the Gold & Rabiner technique [6.25] is basically a feature extraction method which produces a pitch estimate from a variety of signal features relating to waveform peaks and troughs.

Another technique was described in section 2.5 and is based on homomorphic analysis which can be used to determine fundamental frequency directly from the speech cepstrum. This method will also be discussed in more detail in Chapter 7. A comprehensive discussion and bibliography related to pitch determination algorithms is provided by Hess [6.26].

6.4.2 The Electroglottographic Signal

The pitch estimation methods discussed above all rely directly on the acoustic waveform to calculate pitch periods. However, there is one other procedure which employs a slightly different approach to achieve the same end. Electroglottographic (EGG) analysis uses a signal known as the EGG waveform as a second synchronised channel of information alongside the speech signal. The EGG data is then used to perform a division of the speech waveform into individual pitch periods. In spite of the fact that this method has the distinct disadvantage that it requires the acquisition and storage of 2 channels of information, it can produce very accurate and reliable estimates of fundamental frequency if this drawback can be tolerated.

The free surface of the vocal folds constitutes the vibrators within the larynx. Although observations of the vibratory motion of the folds have revealed that these folds undergo complex three dimensional movements [6.27], a single vibratory cycle can usually be separated into 3 distinct phases: (a) an opening phase during which the folds pull apart increasing the glottal opening; (b) a closing phase when the folds come together reducing the glottis; and (c) a closed phase during which the glottal opening is closed or the glottal area is zero [6.27].

The EGG waveform is used to describe laryngeal behaviour in the vicinity of the glottis. Although the EGG signal does not directly measure the opening between the vocal folds, it does tend to covary with the glottal opening — in fact, most researchers believe that the EGG waveform reflects the amount of vocal fold contact [6.28]. However, for the purposes of pitch extraction, this observation is sufficient since the EGG signal represents a simple waveform whose fundamental frequency can easily be measured and is the same as that of the speech signal.

The device used to register the EGG waveform is known as the electroglottograph. This instrument measures the electrical impedance variations of the larynx using a pair of plate electrodes held in contact with the skin on both sides of the thyroid cartilage [6.29]. This signal is recorded synchronously with the acoustic waveform. Therefore, electroglottography represents a simple and non-invasive measurement technique which is unaffected by supraglottal influences such as excitation source/vocal tract coupling and the ambient noise level during the recording process.

If the EGG signal corresponds to the sequence $\{e(n)\}$, then the first backward difference $d(n)$ of the EGG is given by

$$d(n) = e(n) - e(n\text{-}1) \tag{6.24}$$

This difference signal is of consequence since it provides peaks which are useful in the identification of the instants of both glottal opening and closure [6.28]. A typical EGG signal is shown in Fig. 6.4 along with the corresponding speech data and the first difference of the EGG. The clear minima in the difference signal correspond to points of first contact during vocal fold closure [6.29] and thus the time interval between successive minima provides a very simple and effective means of estimating the speech pitch period. This method was employed in the pitch synchronous analysis techniques described in the next subsection.

6.4.3 Introduction of Constraints on the Frequency Range of Formant Regions

The procedure outlined above was used to isolate individual cycles of a speech utterance given by the word "*rely*" which was found to consist of 59 identifiable speech cycles. These pitch periods were then zero-padded to the same length ($N=256$) before calculating the corresponding speech spectra. This process was implemented for two specific reasons.

The first motivating factor is that zero-padding provides a means of increasing the frequency resolution [6.30] of the signal frequency response and thereby obtaining more accurate formant estimates. A frame length larger than $N=256$ was not considered since this would greatly increase the amount of computation required to compute optimal segmentations using the Dynamic Programming formulation.

The other reason is also related to the computational process as zero-padding to the same length for each data frame simplifies the segmentation approach since exactly the same procedure can be deployed for all of the various speech spectra. If this process was not carried out, each of the spectra would be of a different length since the pitch periods themselves vary across the utterance and the segmentation algorithm would then have to account of these variations.

The pitch synchronous spectra were then employed in a procedure for spectral segmentation into formants similar to that described in section 5.2 with two slight modifications. One of these changes concerns the third constraint incorporated in the segmentation scheme in section 5.2.3. This restriction related to the establishment of a minimum frequency range for each Selective LP region and was introduced so that pitch harmonics would not be incorrectly classified as formant regions.

In the pitch synchronous method, these harmonics of the excitation should not be evident and thus the imposition of this constraint is not necessary. Thus the minimum length of any region reverts to the value ascribed by the optimal breakpoint selection/predictor order

allocation scheme outlined in section 4.4. The number of spectral samples in any particular segment must, therefore, exceed at least twice the predictor order in that segment.

Although this restriction has been relaxed to suit the particular properties of pitch synchronous analysis, it is also possible to introduce a set of additional constraints which should provide for a more appropriate segmentation into formants of speech spectra and also result in a reduction of the required computational load. These constraints relate to the limits of variation for the segment endpoints $\{t_n : n=1,\ldots,R\}$. From the characteristic nature of speech formants, it is known that these resonances always tend to lie within certain intervals of the frequency response and it should, therefore, be possible to restrict the ranges for $\{t_n\}$ to cover only these regions.

On the basis of typical limits for the variation of formant frequencies [6.9],[6.31], the minimum and maximum possible frequency values shown in Table 6.1 can be proposed for the set $\{t_n\}$ assuming a signal sampling rate of 10 kHz. In considering these values, it must be remembered that they represent segment end frequencies rather than expected formant frequencies and thus a frequency range of 0.15 – 0.50 kHz does not imply that the lowest frequency resonance will lie in this range — but that the first segment can cover any interval from 0 – 0.15 kHz to 0 – 0.50 kHz. Thus the ranges are, by necessity, broader than would be expected for formant frequencies.

The second point to note is that the maximum end frequencies are all given by the folding frequency (5 kHz) for 3 regions or more. This is since the optimal segmentations for the entire spectrum at each stage — i.e. $\{f_n(t_n=N/2,\, u_n=2n) : n=1,\ldots,R\}$ in the nomenclature of Chapters 4 and 5 — are required to compute the criterion for choosing the number of Selective LP regions as in section 5.3. It is, therefore, necessary that t_n should reach $N/2$ at each stage and thus produce a maximum end frequency of 5 kHz.

It is also assumed in Table 6.1 that 8 Selective LP regions are sufficient to model the speech formants in the 0–5 kHz region and thus for $n=8$, it is only necessary to compute $f_n(t_n=N/2,\ u_n=2n)$ since only this value is required during the final stage. These modifications approximately halve the number of possible combinations of segment lengths and endpoints $\{r_n,\ t_n\}$ and thereby also reduce the computational load by around the same factor.

6.4.4 Cascade Synthesis from Pitch Synchronous Analysis

Apart from the constraints mentioned above, the dynamic programming formulation of the procedure for spectral segmentation into formants is identical to that described in section 5.2 and the remaining stages of analysis required to produce cascade models are outlined below.

(a) The overall errors between signal and Selective LP spectra at each stage of the dynamic programming process $\{\mathfrak{E}^k : k=1,\ldots,R\}$ are employed in the criterion developed in section 5.3 to produce appropriate choices for the numbers of regions to use in the segmented model.

(b) The formant estimates associated with each Selective LP region are then derived as in section 5.4 from the linear prediction parameters for the second order Selective LP models.

(c) The formant centre frequencies and bandwidths calculated for each cycle of the speech waveform are finally incorporated into a cascade formant synthesiser of the form described in section 6.2. Therefore, these parameters are used to characterise the spectral properties of each particular pitch period and this information, together with a fundamental frequency estimate derived from the EGG signal, is used to produce a synthesised speech output. The gain value of the synthesiser during each cycle is chosen

as in section 6.3 such that the total energy is matched between the signal spectrum and the cascade model frequency response.

This procedure was used to generate a synthesised speech signal corresponding to the non-nasalized, voiced utterance "*rely*". Both the original acoustic waveform and the synthesiser output were then subjected to spectrographic[3] analysis to produce the results of Fig. 6.5 (a),(b). The narrowband spectrograms shown were computed on the basis of a 25.6 ms analysis window with an overlap of 22.4 ms between successive windows. As the original speech is sampled at 10 kHz, the spectra computed from each analysis frame are based on 256 data points and thus produce 128 spectral samples covering the range 0–5 kHz.

Fig. 6.5 (c) shows an equivalent spectrogram for speech synthesised from formant estimates obtained using standard LP analysis. The speech spectra associated with each of the 59 pitch periods were, therefore, employed to derive a 12th order LP model and the roots of the corresponding predictor polynomial were then evaluated. As mentioned in section 5.5, all real roots and those complex roots corresponding to a high pole bandwidth were eliminated and the remaining poles were used to calculate formant estimates. These parameters were finally incorporated into the same synthesiser configuration as utilised for the Selective LP model formant estimates to produce an output with a spectrogram as shown in Fig. 6.5 (c).

It is evident from a comparison of these three figures and from listening to the three speech signals that the segmented model produces results which are a much closer approximation to the original then the conventional LP approach with the following two discrepancies being particularly discernable. There appears to be a consistent overestimation of formant bandwidths in Fig. 6.5 (c) resulting in a greater degree of 'smearing' of the model

---

[3] *Speech spectrograms represent time-frequency-intensity displays of the short-time spectra of speech [6.32]. They are often used to illustrate the spectral properties of speech signals.*

spectra when compared to both Fig. 6.5 (b) and the original. Fig. 6.5 (c) also illustrates the same form of spectral distortion in the model frequency response described in section 6.3.2 in relation to the omission of the spectral shaping characteristic in the cascade model with the result that, in Fig. 6.5 (c), the lower frequency formants are almost totally absent.

6.5 CONCLUSIONS

The formant estimates obtained in Chapter 5 using both the procedure for spectral segmentation into formants and standard methods of formant extraction based on LP analysis were assessed in this chapter by the process of cascade synthesiser modelling. An appropriate synthesiser configuration was first developed in relation to the requirements of modelling non-nasalized voiced sounds and it was suggested that a simple cascade formant synthesiser is the most suitable representation for this class of speech sounds.

A system of this form was used to produce cascade model spectra for formant estimates derived from both the segmented models and standard LP methods. It was found that there are several disadvantages associated with both the peak-picking and root-solving techniques which can lead to severe aberrations in the cascade model outputs. There are also other factors which are common to both approaches and can lead to problems such as the overestimation of formant bandwidths and an incorrect spectral emphasis for the synthesised speech frequency response. Some discrepancies were also evident for the results obtained using spectral segmentation but these appear to be mainly associated with the limitations of the cascade formant model.

The segmentation and formant modelling techniques were finally extended to complete utterances rather than isolated speech frames. This process was achieved through the use of an electroglottographic signal recorded synchronously with the speech waveform. The fundamental frequency estimates obtained from this method were then employed in the evaluation of pitch synchronous speech spectra which were used as the basis of the formant

extraction algorithms. Some slight modifications were also incorporated into the segmentation procedure to take account of the actual nature of speech formants. The synthesised speech signals available from the cascade model configurations also displayed the same characteristics as the results of the isolated speech frames with the procedure for spectral segmentation into formants producing a synthesised output which appears to be a closer approximation to the original vocalisation than that provided by standard LP methods.

6.6 REFERENCES

[6.1] RYE, J.M., HOLMES, J.N., "A Versatile Software Parallel-Formant Speech Synthesiser", *JSRU Research Report No.* 1016, *November* 1982, *Joint Speech Research Unit, Cheltenham, Glos., UK.*

[6.2] MUNSON, W.A., MONTGOMERY, H.C., "A Speech Analyzer and Synthesizer", *J. Acoust. Soc. Amer., Vol* 22, *No* 5, *September* 1950, *p* 678.

[6.3] FLANAGAN, J.L., "Automatic Extraction of Formant Frequencies from Continuous Speech", *J. Acoust. Soc. Amer., Vol.* 28, *Jan.* 1956, *pp* 110-118.

[6.4] FALLSIDE, F., WOODS, W.A., "Computer Speech Processing", *Prentice-Hall, Englewood Cliffs, New Jersey,* 1985.

[6.5] RABINER, L.R., "Digital-Formant Synthesizer for Speech-Synthesis Studies", *J. Acoust. Soc. Amer., Vol* 43, *No* 4, *April* 1968, *pp* 822-828.

[6.6] WITTEN, I.H., "Principles of Computer Speech", *Academic Press, London,* 1982.

[6.7] HOLMES, J.N., "Speech Synthesis", *Mills and Boon, London,* 1972.

[6.8] HOLMES, J.N., "Formant Synthesizers: Cascade or Parallel", *JSRU Research Report No.* 1017, 1982, *Joint Speech Research Unit, Cheltenham, Glos., UK.*

[6.9] KLATT, D.H., "Software for a Cascade/Parallel Formant Synthesizer", *J. Acoust. Soc. Amer., Vol.* 67, *No.* 3, *March* 1980, *pp* 971-955.

[6.10] FANT, G., "On the Predictability of Formant Levels and Spectrum Envelopes from Formant Frequencies", *in For Roman Jakobson, Mouton & Co., 's-Gravenhage,* 1956, *pp.* 109-120.

[6.11] CHHATWAL, H.S., CONSTANTINIDES, A.G., "Spectral Estimation of Speech Signals Using Optimal Spectral Segmentation", *Proc. Inst. of Acoustics Autumn Conf., Nov.* 1986, *Vol.* 8, *Part* 7, *pp* 127-144.

[6.12] CHHATWAL, H.S., CONSTANTINIDES, A.G, "Speech Spectral Segmentation for Spectral Estimation and Formant Modelling", *Proc. IEEE Intl. Conf. on Acoustics,*

Speech and Signal Proc., Dallas, Texas, April 1987, *Vol.* 1, *pp* 316-319.

[6.13] SCHAFER, R.W., MARKEL, J.D., "Speech Analysis", *IEEE Press, New York*, 1979.

[6.14] ANANTHAPADMANABHA, T.V., YEGNANARAYANA, B., "Epoch Extraction from Linear Prediction Residual for Identification of Closed Glottis Interval", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP-27, No.* 4, *August* 1979, *pp* 309-319.

[6.15] HOLMES, J.N., "Formant Excitation before and after Glottal Closure", *Conf. Rec., IEEE Intl. Conf. Acoust. Speech and Sig. Proc.*, 1975, *pp* 39-42.

[6.16] LONGUET-HIGGINS, H.C., MANLEY, J.C., "A Simple 4-Formant Speech Encoder", *Proc. Inst. of Acoustics Autumn Conf., Nov.* 1986, *Vol.* 8, *Part* 7, *pp* 145-149.

[6.17] BRISTOW, G., "Electronic Speech Recognition", *Collins, London*, 1986.

[6.18] DUNN, H.K., "Methods of Measuring Vowel Formant Bandwidths", *J. Acoust. Soc. Amer., Vol* 33, *Dec.* 1961, *pp* 241-250.

[6.19] PINSON, E.N., "Pitch-Synchronous Time-Domain Estimation of Formant Frequencies and Bandwidths", *J. Acoust. Soc. Amer., Vol* 35, *August* 1963, *pp* 1264-1273.

[6.20] RABINER, L.R., SCHAFER, R.W., "Digital Processing of Speech Signals", *Prentice-Hall, Englewood Cliffs, New Jersey*, 1978.

[6.21] MATHEWS, M.V., MILLER, J.E., DAVID, E.E. Jr., "Pitch Synchronous Analysis of Voiced Sounds", *J. Acoust. Soc. Amer., Vol* 33, *Feb.* 1961, *pp* 179-196.

[6.22] ATAL, B.S., HANAUER, S.L., "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", *J. Acoust. Soc. Amer., Vol.* 50, *August* 1971, *pp* 637-655.

[6.23] MARKEL, J.D., "The SIFT Algorithm for Fundamental Frequency Estimation", *IEEE Trans. Audio and Electroacoustics, Vol. AU-*20, 1972, *pp* 367-377.

[6.24] SONDHI, M.M., "New Methods of Pitch Extraction", *IEEE Trans. Audio and Electroacoustics, Vol. AU-*16, *No.* 2, *June* 1968, *pp* 262-266.

[6.25] GOLD, B., RABINER, L.R., "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain", *J. Acoust. Soc. Amer., Vol.* 46, *No.* 2, *August*

1969, *pp* 442-448.

[6.26] HESS, W., "Pitch Determination of Speech Signals", *Springer-Verlag, Berlin-Heidelberg-New York-Tokyo*, 1983.

[6.27] CHILDERS, D.G., KRISHNAMURTHY, A.K., "A Critical Review of Electroglottography", *CRC Critical Reviews in Biomedical Engineering, Vol.* 12, *Issue* 2, 1985, *pp* 131-161.

[6.28] KRISHNAMURTHY, A.K., "Two Channel (Speech and EGG) Analysis for Formant Tracking and Inverse Filtering", *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Proc., Vol.* 3, 1984, *pp* 36.6.1-36.6.4.

[6.29] CHILDERS, D.G., LARAR, J.N., "Electroglottography for Laryngeal Function Assessment and Speech Analysis", *IEEE Trans. on Biomedical Eng., Vol. BME*-31, *No.* 12, *Dec.* 1984, *pp* 807-816.

[6.30] RABINER, L.R., GOLD, B., "Theory and Application of Digital Signal Processing", *Prentice-Hall, Englewood Cliffs, New Jersey*, 1975.

[6.31] SINCLAIR, D.A., "Selection of a Formant Synthesiser Model for Text-to-Speech Synthesis", *Proc. Inst. of Acoustics Autumn Conf., Nov.* 1986, *Vol.* 8, *Part* 7, *pp* 363-369.

[6.32] POTTER, R.K., KOPP, G.A., GREEN, H.C., "Visible Speech", *D. Van Nostrand Company, Inc.*, 1947.

Region Endpoint	*Minimum Frequency (kHz)*	*Maximum Frequency (kHz)*
t_1	0.15	0.50
t_2	0.50	2.00
t_3	0.70	5.00
t_4	1.00	5.00
t_5	2.00	5.00
t_6	3.00	5.00
t_7	3.20	5.00
t_8	5.00	5.00

Table 6.1 Illustration of Region Endpoints for Constrained Spectral Segmentation into Formants of Pitch Synchronous Speech Spectra

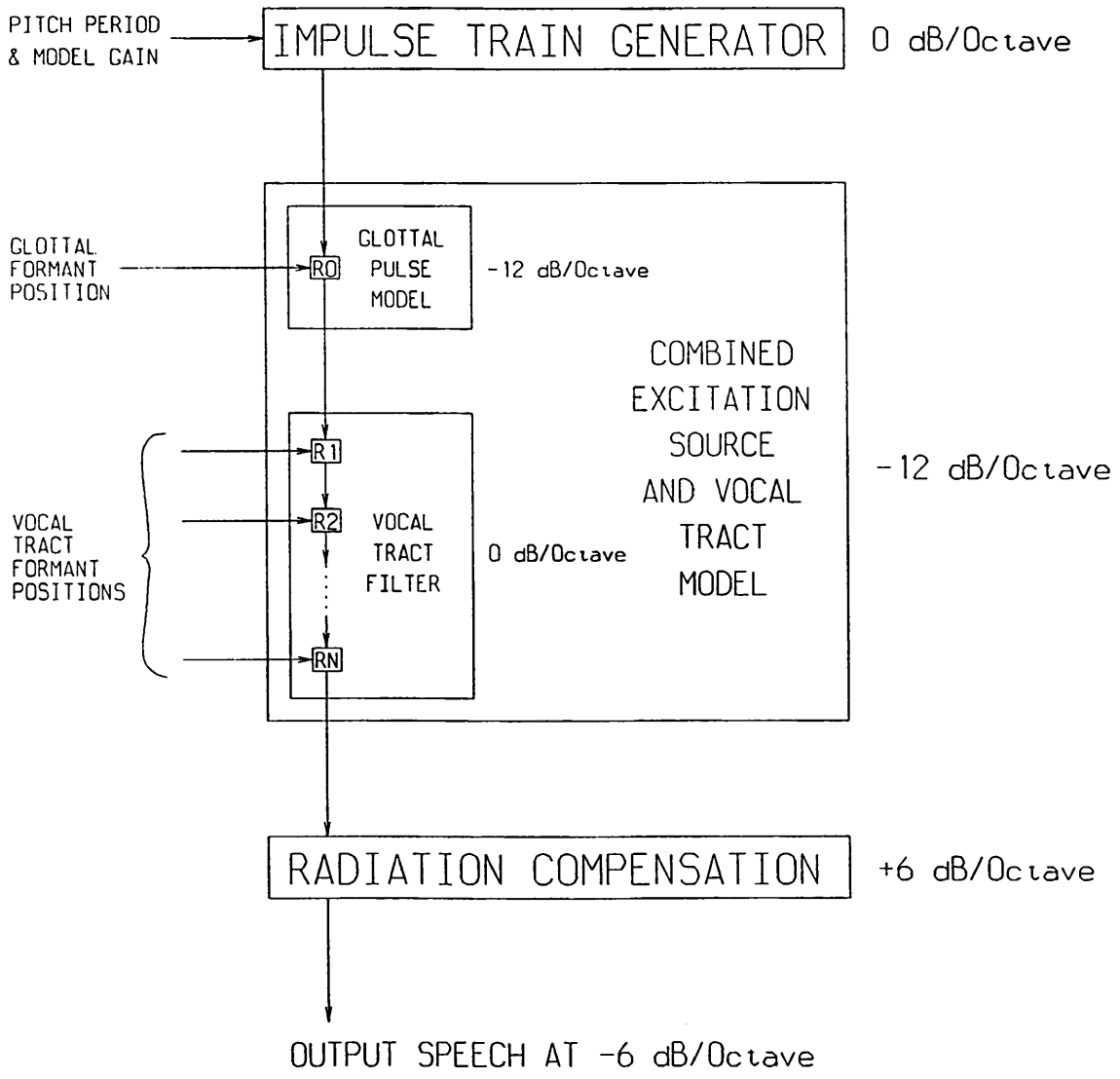
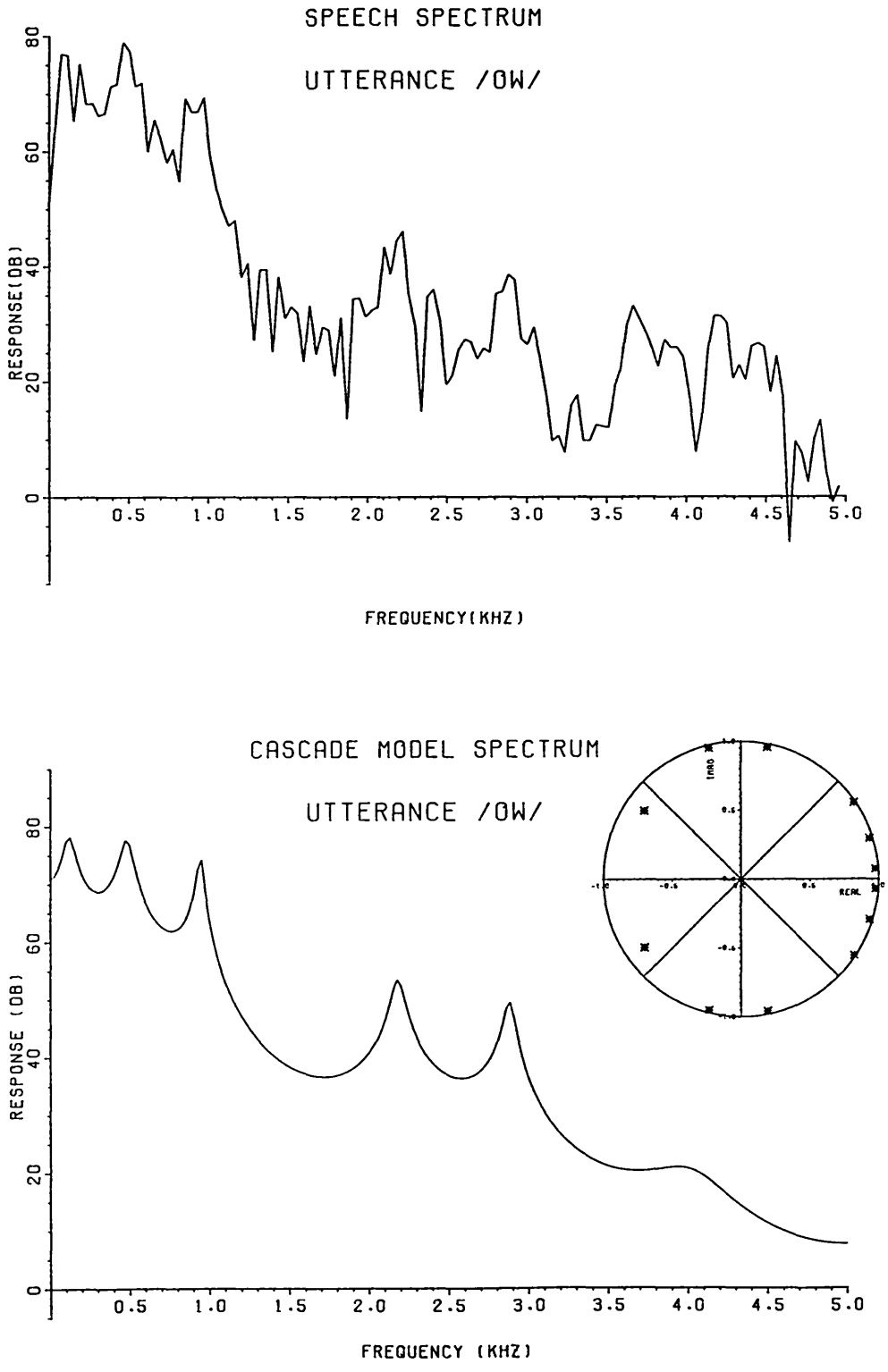
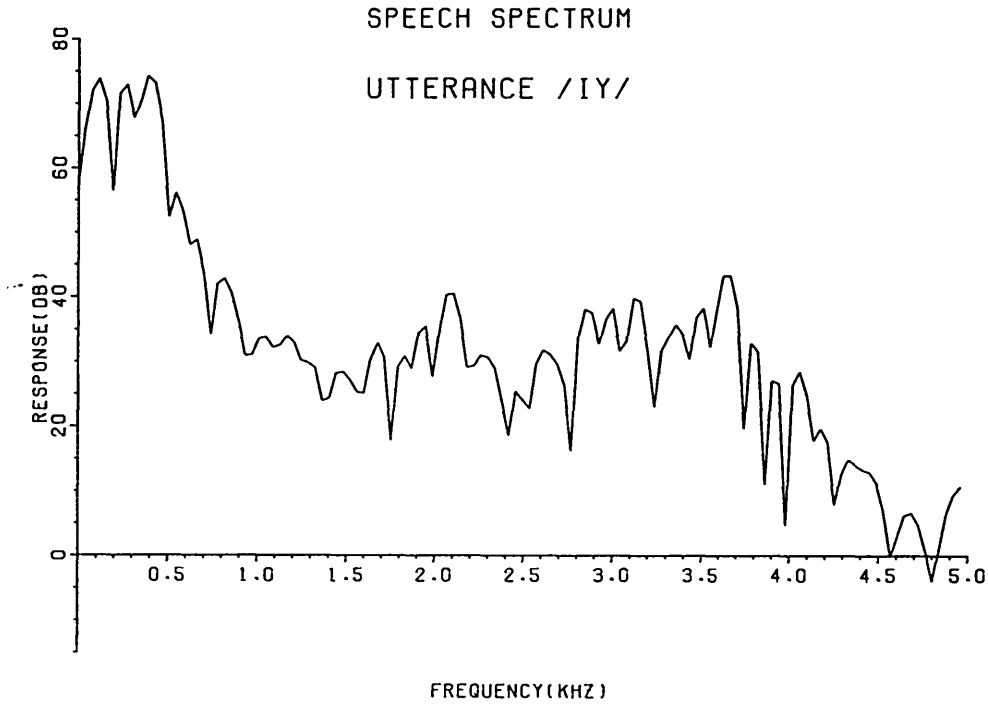


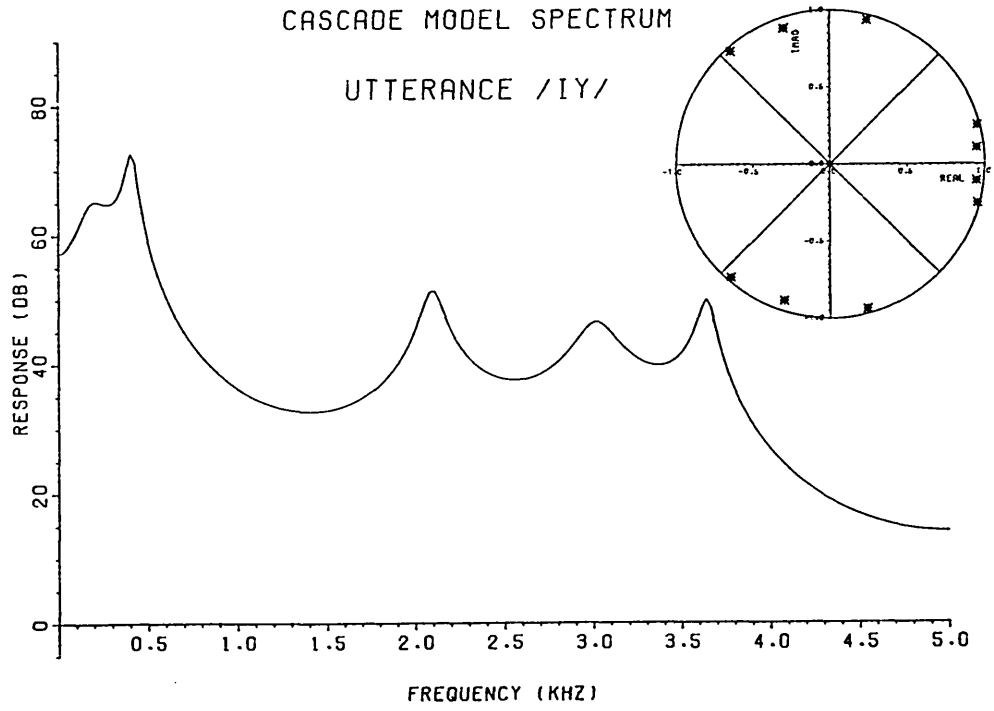
Figure 6.1 Cascade Formant Synthesiser Configuration
for non-Nasalized Voiced Sounds



(a)

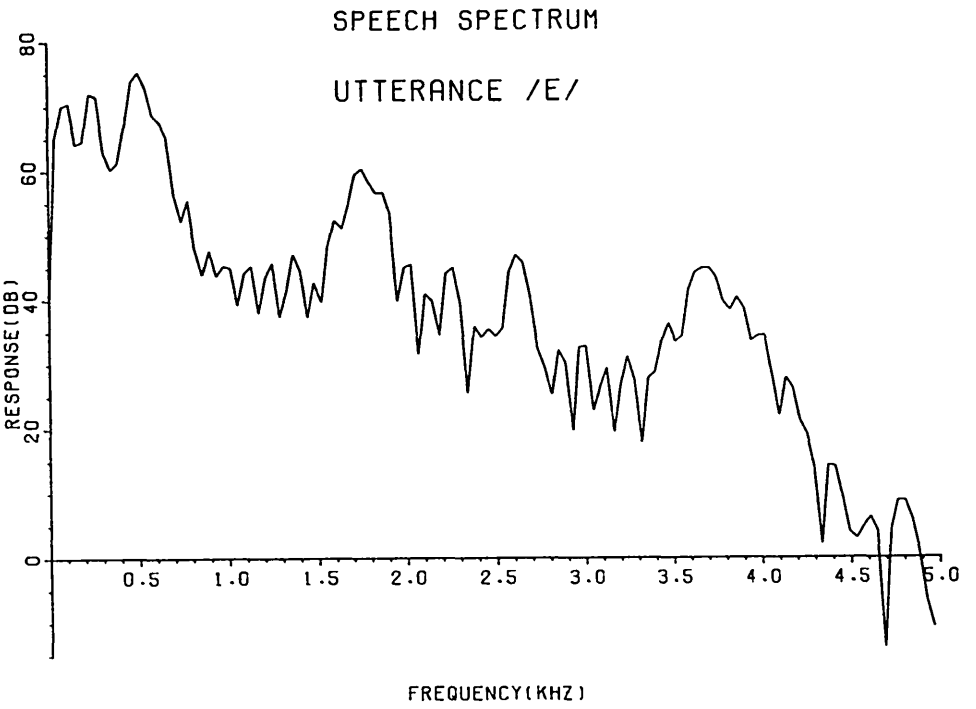
Figure 6.2 Comparison of Signal Spectra and Cascade Model Frequency Responses for 4 Vowel Sounds

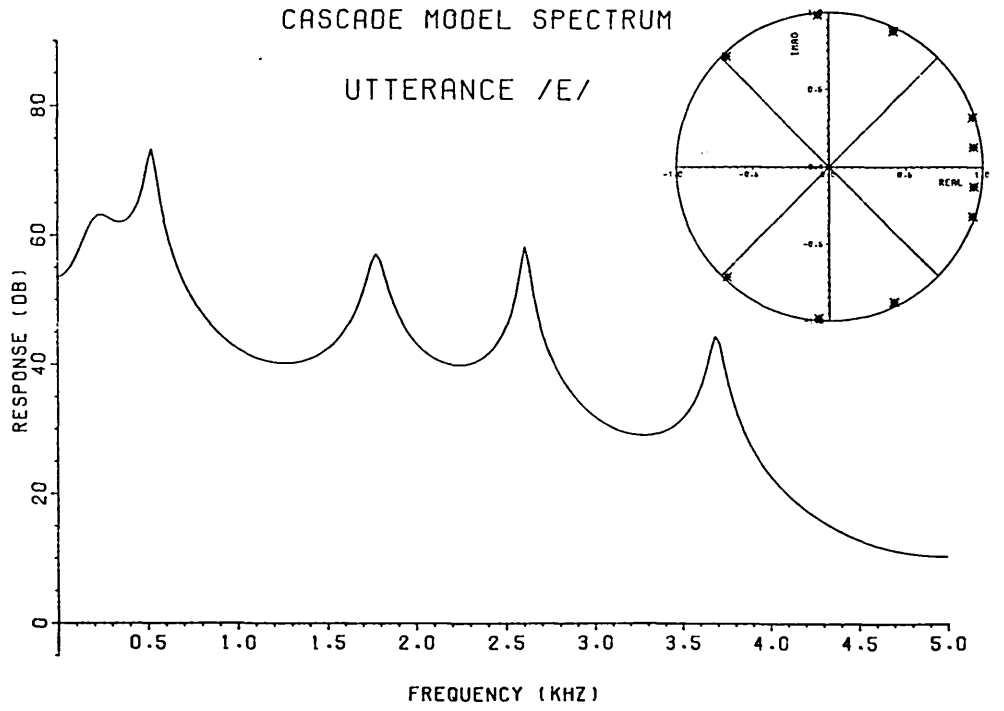




(b)

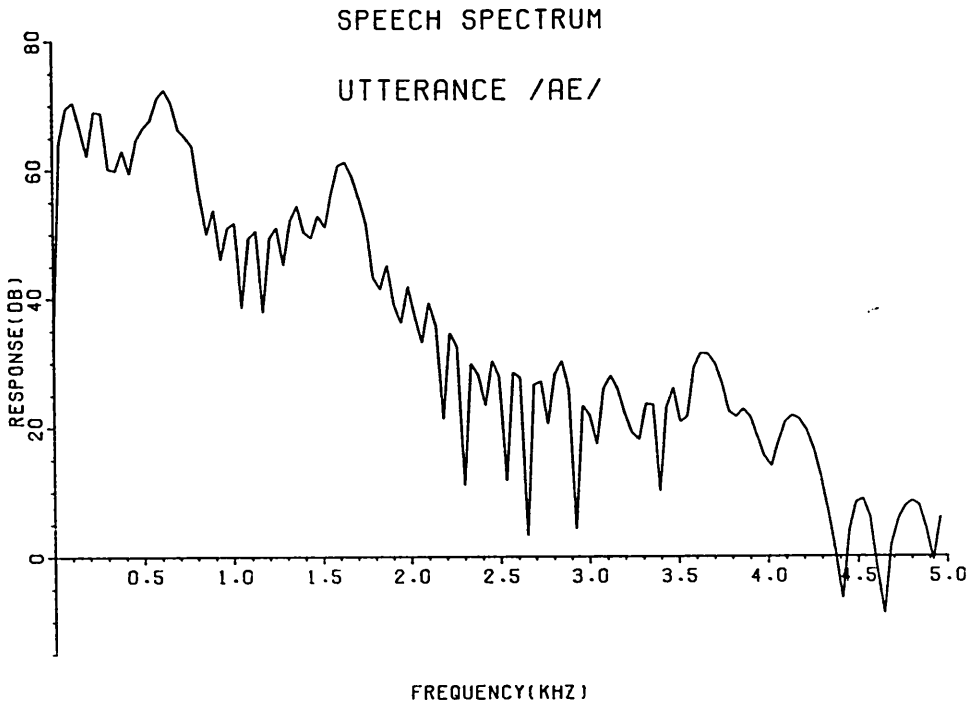
Figure 6.2 (Contd.)

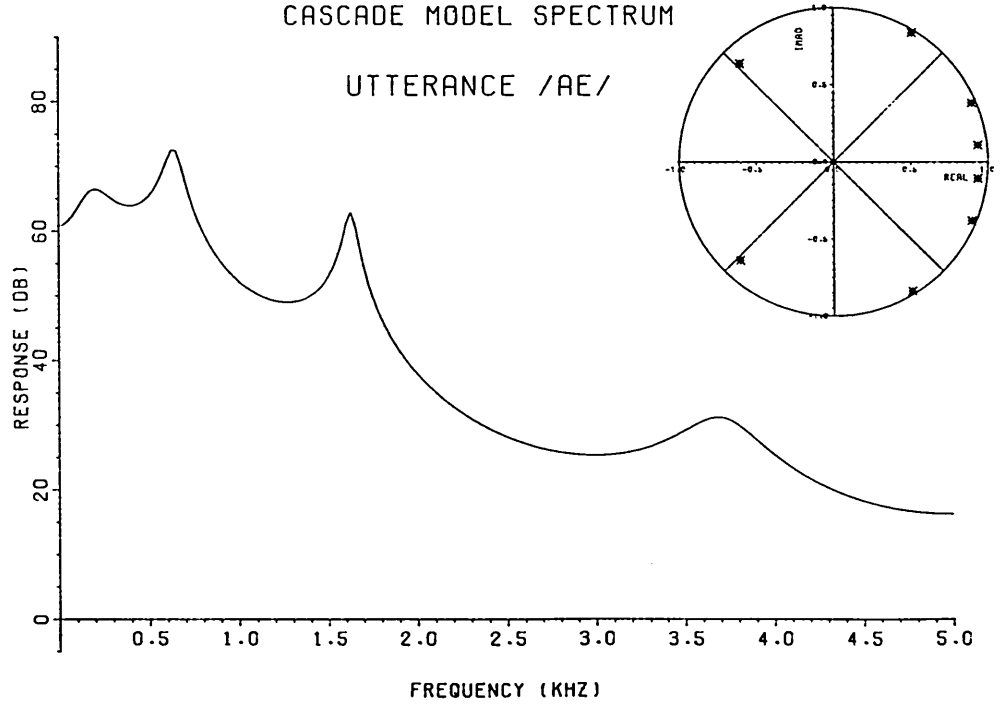




(c)

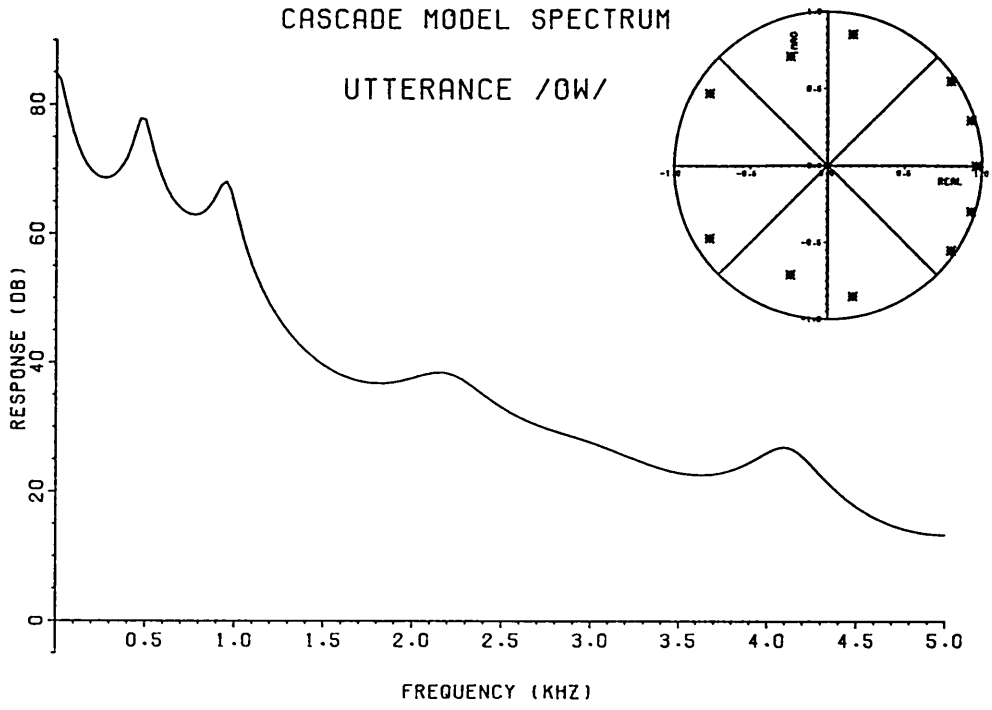
Figure 6.2 (Contd.)



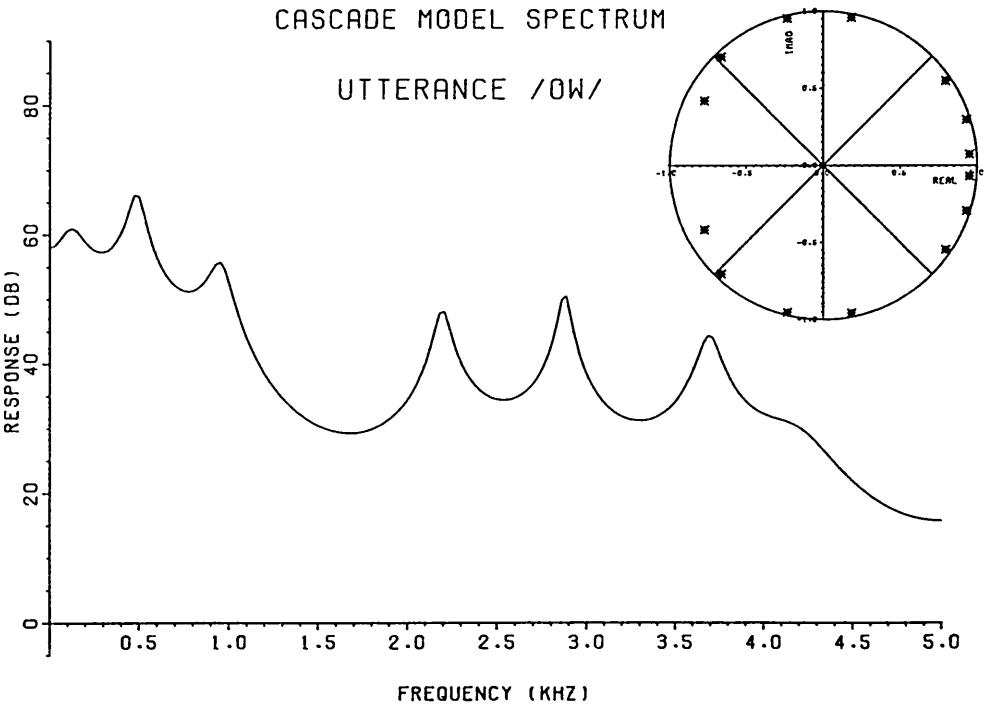


(d)

Figure 6.2 (Contd.)



(a) 12th Order Predictor



(b) 20th Order Predictor

Figure 6.3 Illustration of Cascade Model Spectra obtained from 12th Order and 20th Order Linear Predictive Analysis

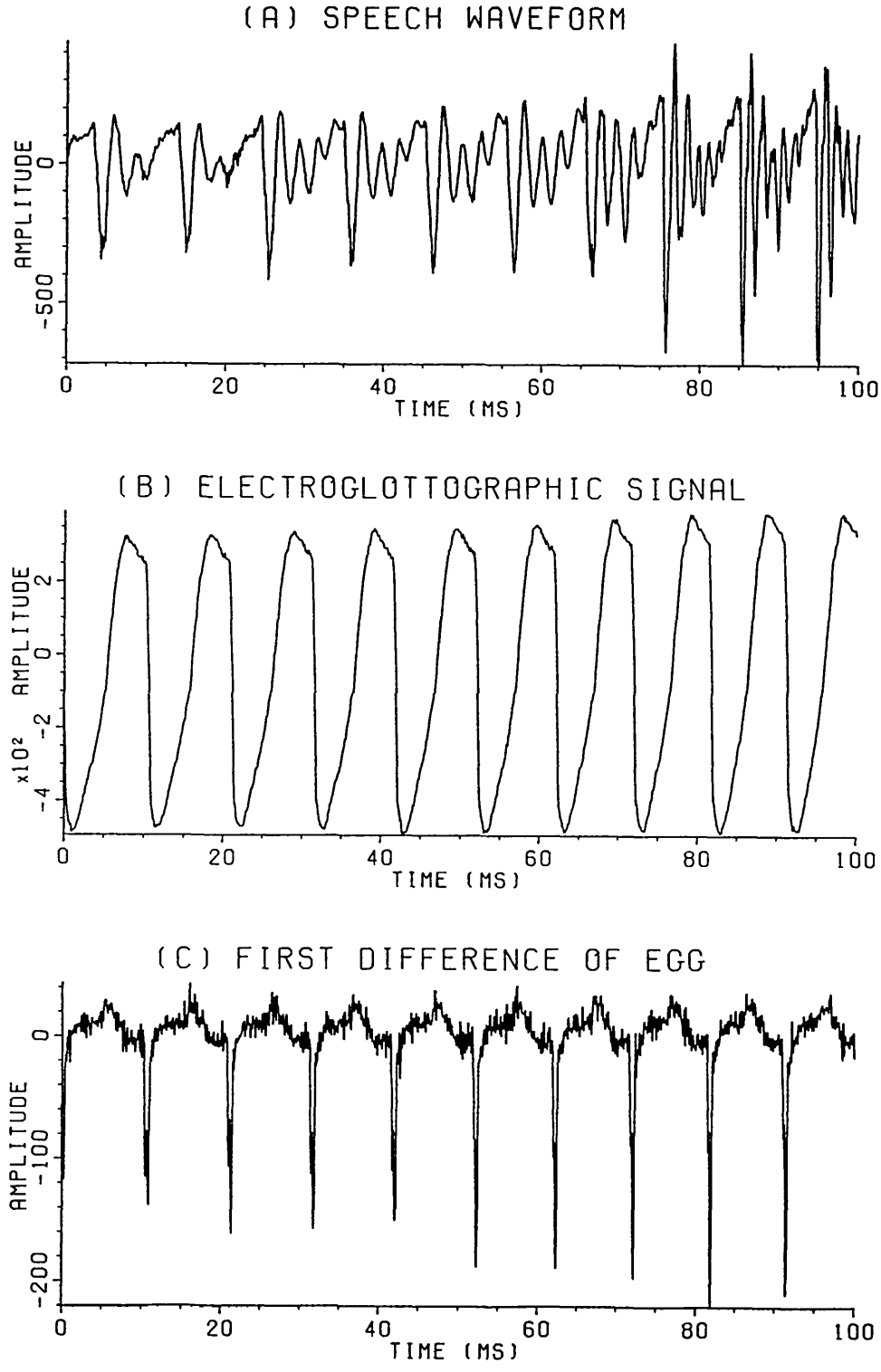


Figure 6.4 Illustration of Pitch Estimation Method based on Electroglottographic Analysis

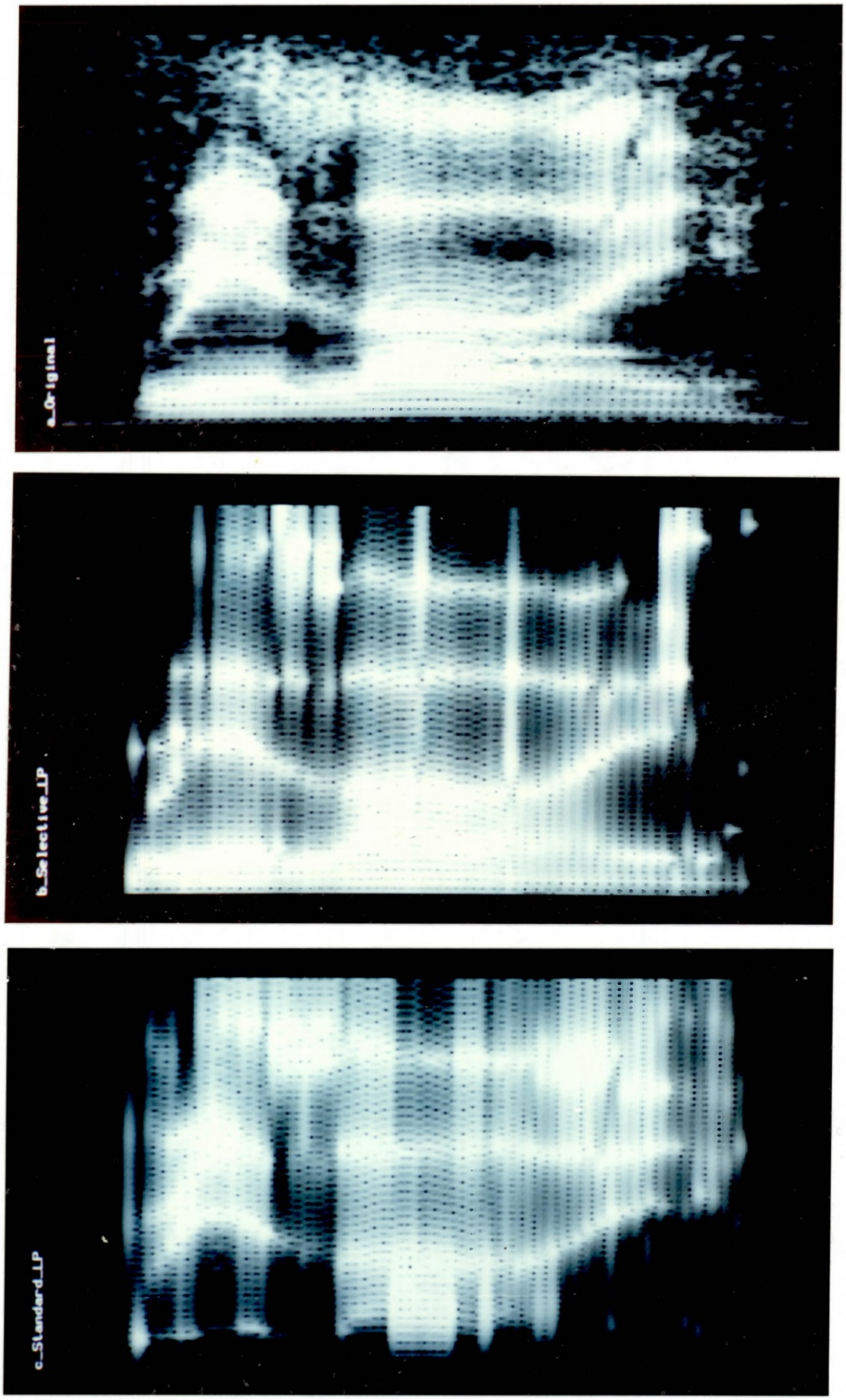


Figure 6.5 Narrowband Spectrograms for
(a) Original Utterance
(b) Speech Synthesised from Procedure for Spectral Segmentation
(c) Speech Synthesised from Standard LP Modelling

CHAPTER 7

THE REPRESENTATION OF NASALIZED SPEECH USING COMBINED POLE-ZERO MODELS

CHAPTER 7. THE REPRESENTATION OF NASALIZED SPEECH USING COMBINED POLE-ZERO MODELS

7.1 INTRODUCTION

The previous three chapters were largely concerned with one particular application of linear predictive analysis — that of speech formant extraction. The work described in this chapter is directed towards a slightly different aspect of LP modelling. One of the major disadvantages of the linear predictive formulation is the all-pole nature of the model which is ill-suited to the modelling of sounds such as nasals and nasalized vowels. This is since a side-branch resonator is coupled to the main acoustic cavity during the production of these vocalisations. The oral cavity forms the side-branch resonator in the case of nasal consonants while the nose should be considered as a cavity during a nasalized vowel [7.1]. This side-branch resonator will then tend to trap acoustic energy at certain natural frequencies which appear as antiresonances, or zeros of sound transmission in the radiated speech [7.2].

It may be argued that the linear predictive spectrum can nevertheless be used to represent these zeros since a spectral zero can, in fact, be modelled by a large number of poles [7.3] and it was also found in Chapter 3 that the LP model spectrum can approximate any signal frequency response arbitrarily closely by simply increasing the number of predictor coefficients. However, if there are a number of influential zeros in the signal spectrum, the required model order can become very large and this situation is obviously undesirable for data compression applications.

In addition, the physical significance of the roots of the predictor polynomial then becomes unclear since it will not be apparent if a particular root corresponds to the nasal zeros or additional nasal poles, or if it is related to the expected resonances of the vocal tract [7.4]. This problem will be of importance in applications such as the differentiation of various

nasalized sounds where the identification of the spectral zeros is important.

Various methods have been proposed to cope with this inadequacy of LP analysis and the first part of this chapter is devoted to a discussion of some of the more important modelling techniques for nasalized speech. Some of these approaches actually use an LP modelling stage as the first step of the analysis in order to inverse filter the nasalized speech and thereby remove the effect of the signal poles. The zeros are then estimated from the spectrum of the linear prediction residual. A similar approach is employed in this chapter with the LP residual spectrum acting as the basis of a segmentation process similar to those of Chapters 4 and 5.

The first stage in this process involves the calculation of the required spectrum using pitch synchronous analysis and the reasons for using this particular method are also discussed. In Chapter 5, the speech spectrum was segmented into several regions — each of which was modelled by a second order Selective LP model. In order to apply a similar segmentation scheme to the LP residual spectrum, an algorithm is first developed for modelling a single spectral zero by a system which consists entirely of 2 zeros in the z-domain and is known as a second order moving average (MA) model.

A Selective MA technique which can be used to characterise a given region of a signal frequency response is then incorporated into a procedure for segmenting the LP residual spectrum into regions containing zeros. The positions of the zeros associated with each of these segments can then be calculated from the 2nd order MA models.

A set of zeros is now available for the original nasalized speech and these antiresonances can then be removed from the speech waveform to yield a signal which should be characterised entirely by spectral poles. The resulting data is, therefore, amenable to standard LP analysis and these linear predictive parameters together with the zero positions are used to form a model of the nasalized utterance. The final analysis stage involves the

deployment of this combined pole-zero model in a synthesiser network to produce an output which can be compared to the original speech signal. These results are also appraised in relation to those available from a direct LP model of the nasalized vocalisation.

7.2 EXISTING METHODS FOR THE ANALYSIS OF NASALIZED SPEECH

Unlike the method of all-pole modelling, the pole-zero problem is inherently non-linear [7.5] so that either iterative optimisation techniques must be utilised to calculate the pole and zero positions or non-iterative schemes have to be employed to obtain reasonable estimates of these parameters. A large variety of pole-zero techniques are available for each of these classes and several of these methods are outlined in the following two subsections.

7.2.1 Iterative Solutions

The linear prediction model discussed in Chapter 3 was used to describe signals $s(n)$ which conform to the following description for some input signal $u(n)$:

$$s(n) = \sum_{k=1}^{p} \alpha_k \, s(n-k) + G u(n) \tag{7.1}$$

where $\{\alpha_k : k=1,\ldots,p\}$ and G are the predictor coefficients and gain respectively of the p-th order model. The transfer function, $H(z)$, of such a system is given by:

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} \alpha_k \, z^{-k}} \tag{7.2}$$

This description consists entirely of poles in the z-domain and is often termed an "*autoregressive (AR) model*". However, it is also possible to envisage another system function which is composed of both poles and zeros such that

$$H(z) = \frac{S(z)}{U(z)} = G \frac{1 - \sum_{l=1}^{q} \beta_l z^{-l}}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}} \tag{7.3}$$

and, therefore,

$$s(n) = \sum_{k=1}^{p} \alpha_k s(n-k) + G\left[u(n) - \sum_{l=1}^{q} \beta_l u(n-l)\right] \tag{7.4}$$

Since an all-zero description (i.e. $\{\alpha_k = 0 : k = 1,\ldots,p\}$ in equation (7.3)) is often termed a "*moving average (MA) model*", the overall pole-zero system of equation (7.3) is known as an "*autoregressive moving average (ARMA) model*". The model spectrum, $\left\{ \hat{P}(\omega_m) : m = 0,\ldots,N-1, \ \omega_m = \frac{2\pi m}{N} \right\}$ corresponding to $H(z)$ is then given by

$$\hat{P}(\omega_m) = \left|H\left(e^{j\omega_m}\right)\right|^2 = G^2 \frac{\left|B\left(e^{j\omega_m}\right)\right|^2}{\left|A\left(e^{j\omega_m}\right)\right|^2} \tag{7.5}$$

where $B(z)$, $A(z)$ are the numerator and denominator polynomials respectively in equation (7.3). The linear prediction error, E, was related in Chapter 3 to the match between the signal spectrum $P(\omega_m)$ and the model spectrum $\hat{P}(\omega_m)$ as follows:

$$E = \frac{G^2}{N} \sum_{m=0}^{N-1} \frac{P(\omega_m)}{\hat{P}(\omega_m)} \tag{7.6}$$

Similarly, in this case, a matching error between the real and model spectra is such that [7.5]

$$E = \frac{1}{N} \sum_{m=0}^{N-1} P(\omega_m) \cdot \frac{\left|A\left(e^{j\omega_m}\right)\right|^2}{\left|B\left(e^{j\omega_m}\right)\right|^2} \tag{7.7}$$

This error E can also be interpreted as the residual energy obtained by passing a signal through the filter $A(z)/B(z)$. The pole-zero estimation problem then reduces to a calculation

of $\{\alpha_k\}, \{\beta_l\}$ such that E in equation (7.7) is minimised. However, this requirement leads to a set of $(p+q)$ nonlinear equations which must then be solved iteratively.

A particular class of techniques which can be used in this case is that of gradient methods and one such scheme was employed by Tretter and Steiglitz [7.6] in pole-zero modelling. Other approaches, such as the Newton-Raphson method [7.5], can also be utilised but both these schemes involve complex matrix operations and require an initial estimate of the solution which is close to the optimum. If this is not the case, then these approaches cannot be guaranteed to converge to a global error minimum.

7.2.2 Non-Iterative Solutions

Unlike the iterative methods described above which attempt to solve for both the system poles and zeros simultaneously, most of the non-iterative solutions generally estimate the pole positions first, and then involve a stage aimed at zero estimation. This process can be accomplished since it may be shown [7.5] that the signal autocorrelation function $\{R(i)\}$ for the output of a pole-zero model obeys the following relationship for $i > q$ — where q is the number of system zeros,

$$R(i) = \sum_{k=1}^{p} \alpha_k R(i-k) \tag{7.8}$$

This expression is identical to that developed in Chapter 3 for an AR process and, therefore, the coefficients $\{\alpha_k : k=1,\ldots,p\}$ can be estimated by solving this equation as in Chapter 3 for $q+1 \leq i \leq q+p$. The effect of the poles can then be removed by applying the all-pole predictor polynomial, $A(z)$, as an inverse filter to the ARMA process. This can, of course, also be achieved in the spectral domain to yield the LP residual spectrum, $P_e(\omega_m)$, such that

$$P_e(\omega_m) = P(\omega_m) \cdot \left|A\left(e^{j\omega_m}\right)\right|^2 \tag{7.9}$$

The pole-zero estimation problem now reduces to the calculation of the zeros in $P_e(\omega_m)$. One particular method, proposed by Shanks [7.7], determines the coefficients β_l by minimising the mean-square error between $s(n)$ in equation (7.4) and the impulse response of the pole-zero filter. This produces a set of linear equations which may be solved to obtain $\{\beta_l : l=1,\ldots,q\}$. This approach has been used successfully on pure synthetic speech [7.8] but often produces unsatisfactory results for real speech signals with sharp cusps and other undesirable features which render the model spectra meaningless [7.9].

A second possible procedure for the estimation of zero positions from the residual spectrum is known as *inverse LP modelling* [7.10]. In this method, the spectrum $P_e(\omega_m)$ is inverted in order to transform the zeros of the residual into poles of the inverted spectrum. These 'poles' can then be estimated by applying a q-th order LP analysis on the inverted residual spectrum to obtain estimates of $\{\beta_l\}$. This method only provides good results for zero frequencies if $P(\omega_m)$ is smooth relative to the model spectrum since inaccuracies can arise if the variation of the signal spectrum about the model frequency response is large [7.5].

One means of reducing this variation is to smooth the error spectrum $P_e(\omega_m)$ before inversion using a method such as homomorphic smoothing as in Chapter 2. However, it is known that this form of spectral smoothing based on the speech cepstrum leads to poor estimates of pole and zero bandwidths [7.4] and thus the bandwidths of the zeros in the combined pole-zero model will be in error. The degree of smoothing required is also unclear and thus it is difficult to determine the length of the low-time portion of the cepstrum required to perform the necessary smoothing operation.

The result of using inverse LP modelling on the spectrum of the nasalized consonant /m/ is shown in Fig. 7.1. The curves shown in Fig. 7.1 (a),(b) depict the speech spectrum and

the homomorpically smoothed frequency response while Fig. 7.1 (c) shows the 12th order LP model spectrum calculated from Fig. 7.1 (b). The result shown in Fig. 7.1 (d) is the spectrum obtained using a 12–pole, 10–zero system for the inverse LP modelling procedure.

Another method known as *cepstral prediction* [7.9],[7.11] stems from the fact that the poles of the linearly weighted complex cepstrum of a signal comprise the poles and zeros of that signal. This is since a sequence $s(n)$ characterised by an ARMA process, i.e.

$$S(z) = \frac{B(z)}{A(z)} \tag{7.10}$$

will have a complex cepstrum $\hat{s}(n)$ with a z-transform given by

$$\hat{S}(z) = \log S(z) = \log\left[B(z)\right] - \log\left[A(z)\right] \tag{7.11}$$

Therefore, the z-transform of the linearly weighted cepstrum $n\hat{s}(n)$ is [7.9]:

$$-z\hat{S}'(z) = \frac{zA'(z).B(z) - zB'(z).A(z)}{A(z).B(z)} \tag{7.12}$$

where prime denotes differentiation with respect to z. Thus the sequence $\{n\hat{s}(n)\}$ has a z-transform whose poles are the poles and zeros of $s(n)$. The zeros of an ARMA process will, therefore, appear as poles of the sequence resulting from this method being applied to the residual spectrum. These poles can finally be derived using a standard LP model. Although this method appears to perform adequately for synthetic speech, it has not been particularly successful for real speech signals [7.9].

Other more speculative approaches include that based on the use of Padé approximants [7.12]. This notion determines a rational approximation to the system function with specified orders μ, ν for the numerator and denominator respectively by matching the

power series expansions of the given function and the approximant for terms of order less than $\mu+\nu+1$. Another possible method relies on the calculation of a high-order LP model for the pole-zero signal and the subsequent decomposition of this predictor polynomial into a pole predictor and a zero predictor [7.13].

However, all of the non-iterative methods discussed in this section suffer from a common drawback which relates to the choice of the two model orders p, q. This is since it appears to be necessary that a good a priori estimate of the number of system poles and zeros is available for a reasonable solution [7.5]. This situation is obviously a disadvantage for speech signals since it is often difficult to assess beforehand the number of resonances and antiresonances likely to be exhibited by the signal spectrum for the whole range of speech sounds.

7.3 OPTIMAL SEGMENTATION OF LINEAR PREDICTION RESIDUAL SPECTRA

7.3.1 Pitch Synchronous Analysis of Nasalized Speech

There is one major disadvantage in applying any pole-zero modelling technique directly to pitch asynchronous speech spectra. This drawback is a result of the excitation source periodicity which causes pitch harmonics spaced at the fundamental frequency and thus the spectral amplitude approaches zero between these harmonics [7.14]. Any pole-zero estimation method may, therefore, misinterpret these troughs as being due to the presence of zeros of the vocal tract/nasal tract combination and thereby produce grossly inaccurate estimates for these antiresonances. Hence, with any of the potential techniques for the measurement of spectral zeros, it appears to be important to precede such analysis by a deconvolution of the speech signal into its excitation source and vocal tract components [7.8].

There are two main methods which are generally considered for such a deconvolution. The first technique has already been discussed in the previous subsection with respect to inverse LP modelling and is characterised by the use of homomorphic smoothing to obtain an

approximation of the vocal tract impulse response to which a pole-zero analysis can be applied. However, as mentioned in section 7.2.2, this approach does have the distinct disadvantage that cepstral analysis often fails to accurately model the bandwidths of spectral poles and zeros [7.4] and may, therefore, result in inaccurate estimates for these values.

Another widely used method of source-filter deconvolution is that of pitch synchronous analysis. As described in section 6.4.1, this approach consists of the extraction of individual pitch periods from the speech waveform and considering a single pitch period to correspond to the impulse response of the vocal tract. It is, therefore, assumed that this impulse response is shorter in duration than a single pitch period. The spectra associated with each of the speech cycles are then employed as the basis of the pole-zero analysis.

One other property of the speech waveform which may cause the occurrence of zeros that are not due to the combined vocal tract/nasal tract transfer function is the form of analysis window which is applied to the speech frame under consideration. This is since the pitch synchronous speech spectrum is a result of the convolution of the Fourier transforms of the vocal tract/nasal tract impulse response and that of the analysis window [7.4].

The window used in the particular case of pitch synchronous analysis is rectangular — i.e. a particular period is isolated directly from the speech waveform — and it can then be shown [7.4] that this window has no effect at the particular spectral analysis points of the Fourier Transform if the pitch period of the signal is identical to the length of the analysis frame. The window will not, therefore, produce any zeros in the pitch synchronous spectrum and, in the absence of extraneous noise, any antiresonances in this frequency response should be purely due to the characteristics of the vocal and nasal tracts.

It is, therefore, apparent that pitch synchronous analysis offers several advantages in the pole-zero estimation of nasalized speech. One issue which has not been covered to this point is the actual analysis technique which is to be employed to perform the pitch estimation

process. In the last chapter, a synchronised electroglottographic signal was used for this operation, but the nasalized utterances available for analysis in this case were not recorded with an EGG waveform and thus this particular approach could not be implemented.

Of the other pitch determination algorithms mentioned in section 6.4.1, several were based on the inverse filtering of the speech waveform using an LP model as described in section 3.5. However, the effect of nasalization is to cancel a significant amount of the spectral periodicity which is necessary to obtain the sharp spikes in the LP error signal which forms the basis of this type of pitch detection process [7.9].

An alternative method has also been discussed previously in this thesis since the procedure for pitch estimation using homomorphic analysis was developed in section 2.5. This method relies on the identification of sharp peaks corresponding to the pitch period in the high-time portion of the speech cepstrum. By setting an appropriate threshold for the minimum height of a peak as being indicative of a pitch period estimate, a voiced/silence classification can also be achieved. These two processes are depicted in Fig. 7.2 for 4 consecutive frames from the start of the utterance "*rely*".

The points developed earlier in this subsection can now be illustrated with reference to an utterance sampled at 10 kHz consisting of a nasalized vowel which corresponds to the voiced phoneme in the French word "s*on*". Fig. 7.3 shows a pitch asynchronous spectrum evaluated from 256 data points while Fig. 7.4 depicts 8 consecutive pitch periods taken from the same utterance. These cycles were, therefore, isolated on the basis of the fundamental frequency estimate available from cepstral processing and then searching for zero crossings in the speech waveform spaced at or near the estimated pitch period. As in Chapter 6, these individual frames were then each zero-padded to 256 points in order to increase the frequency resolution of the resulting spectra.

The 8 pitch synchronous spectra overlaid on each other are also illustrated in Fig. 7.4. A comparison of these results with Fig. 7.3 shows the improvement available using pitch synchronous analysis in the identification of zero positions as there are fewer, but much more readily discernable, antiresonances in the pitch synchronous result.

7.3.2 Development of a Second Order Moving Average Scheme for Antiresonance Modelling

In Chapter 5, a technique was developed for the segmentation of a speech spectrum into several regions such that each segment contained a single speech formant. These individual regions were themselves characterised by 2nd order Selective LP models which could be related to complex pole pairs and thereby to the formant parameters.

The requirement in this section is to achieve the same objective for a spectrum with several antiresonances which can be modelled by complex conjugate zero pairs. The overall aim is, therefore, to develop a procedure for calculating the parameters of a 2nd order moving average (MA) model corresponding to a particular range of a given signal spectrum.

It was shown in section 4.2 that the autocorrelation values could be calculated from a frequency response by an inverse Fourier Transform and thus the first stage in the analysis of the 2nd order MA models (as in the case of LP analysis in Chapter 3) is to calculate the model parameters in terms of the autocorrelation coefficients.

Step 1 Illustration of Conventional Approach to 2nd Order MA Modelling

A moving average process, $x(n)$, is obtained from the output of a filter with a transfer function, $H(z)$, which consists only of zeros [7.15] such that:

$$H(z) = G\left[1 - \sum_{l=1}^{q} \beta_l z^{-l}\right] \tag{7.13}$$

Thus, $$x(n) = G\left[u(n) - \sum_{l=1}^{q} \beta_l\, u(n\text{-}l)\right] \tag{7.14}$$

where the system input, $u(n)$, is such that:

$$\mathcal{E}\left[u(n)\right] = 0 \quad \text{and} \quad \mathcal{E}\left[u(n\text{-}l).\,u(n)\right] = 1 \qquad \text{for } l=0$$

$$= 0 \qquad \text{otherwise}$$

and $\mathcal{E}[\cdot]$ denotes the expected value. Thus, for a second order model,

$$x(n) = G\left[u(n) - \beta_1\, u(n\text{-}1) - \beta_2\, u(n\text{-}2)\right] \tag{7.15}$$

The first three coefficients of the autocorrelation function $\{R(k)\}$ for this process are then given by [7.15]:

$$R(0) = G^2\left[1 + \beta_1{}^2 + \beta_2{}^2\right] \tag{7.16a}$$

$$R(1) = G^2\left[-\beta_1 + \beta_1\beta_2\right] \tag{7.16b}$$

$$R(2) = G^2.\,-\beta_2 \tag{7.16c}$$

These equations can, therefore, also be represented as the following matrix expression:

$$\begin{bmatrix} R(0) \\ R(1) \\ R(2) \end{bmatrix} = G^2 \begin{bmatrix} 1 & -\beta_1 & -\beta_2 \\ 0 & 1 & -\beta_1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -\beta_1 \\ -\beta_2 \end{bmatrix} \tag{7.17}$$

From equations (7.16a)–(7.16c), there are 3 unknown parameters: the model gain, G, and the coefficients $\{\beta_1, \beta_2\}$. With these 3 equations, it should be possible to solve for these 3 values.

From (7.16b),(7.16c),

$$R(1) \;=\; \frac{R(2)}{\beta_2}.\beta_1.(1-\beta_2)$$

$$\Longrightarrow \qquad \beta_1 \;=\; \frac{R(1)\,\beta_2}{R(2)\,(1-\beta_2)} \tag{7.18}$$

Substitute into (7.16a),

$$-R(0)\,\beta_2 \;=\; R(2)\left\{1 \;+\; \left[\frac{R(1)\,\beta_2}{R(2)\,(1-\beta_2)}\right]^2 \;+\; {\beta_2}^2\right\} \tag{7.19}$$

This expression will yield a quartic equation in β_2. In fact, it is known that the equations for an MA process of order greater than or equal to 2 are non-linear and must be solved iteratively [7.16]. An alternative approach must, therefore, be employed in this particular case.

Step 2 Alternative Description of 2nd Order MA Model Equations

Consider rewriting equation (7.15) in the following equivalent form

$$x(n) \;=\; -G\beta_1\left[-\frac{1}{\beta_1}.u(n) \;+\; u(n\text{-}1) \;+\; \frac{\beta_2}{\beta_1}.u(n\text{-}2)\right] \tag{7.20}$$

then let $$x(n) \;=\; H\left[\gamma_0\,u(n) \;+\; u(n\text{-}1) \;+\; \gamma_2\,u(n\text{-}2)\right] \tag{7.21}$$

Thus, $$H \equiv -G\beta_1\,; \qquad \gamma_0 \equiv -\frac{1}{\beta_1}\,; \qquad \gamma_2 \equiv \frac{\beta_2}{\beta_1} \tag{7.22}$$

Now, the autocorrelation values may be expressed as:

$$R(0) = H^2\left[\gamma_0{}^2 + 1 + \gamma_2{}^2\right] \tag{7.23a}$$

$$R(1) = H^2\left[\gamma_0 + \gamma_2\right] \tag{7.23b}$$

$$R(2) = H^2 . \gamma_0\gamma_2 \tag{7.23c}$$

Step 3 Calculate MA Parameters from Autocorrelation Values

From equation (7.23b),

$$\left(\gamma_0 + \gamma_2\right)^2 = \frac{\left[R(1)\right]^2}{H^4}$$

$$\Longrightarrow \qquad \gamma_0{}^2 + 1 + \gamma_2{}^2 = \frac{\left[R(1)\right]^2}{H^4} - 2\gamma_0\gamma_2 + 1$$

$$= \frac{\left[R(1)\right]^2}{H^4} - 2\frac{R(2)}{H^2} + 1 \tag{7.24}$$

This result can then be substituted into equation (7.23a) to give

$$R(0) = \frac{\left[R(1)\right]^2}{H^2} - 2R(2) + H^2$$

or $$H^4 - \left[R(0) + 2R(2)\right]H^2 + \left[R(1)\right]^2 = 0$$

Thus, $$H^2 = \frac{\left[R(0) + 2R(2)\right] \pm \sqrt{\left[R(0) + 2R(2)\right]^2 - 4\left[R(1)\right]^2}}{2} \tag{7.25}$$

This expression will, therefore, produce two roots for H^2 depending on whether the plus or minus sign is used in the numerator of equation (7.25). Then, from equations (7.23b), (7.23c),

$$R(1) = H^2\left(\gamma_0 + \frac{R(2)}{H^2\gamma_0}\right)$$

$$\Longrightarrow \qquad H^2\gamma_0{}^2 - R(1)\gamma_0 + R(2) = 0 \tag{7.26}$$

Hence, $$\gamma_0 = \frac{R(1) \pm \sqrt{[R(1)]^2 - 4H^2 R(2)}}{2H^2} \tag{7.27}$$

and $$\gamma_2 = \frac{R(2)}{H^2 \gamma_0} \tag{7.28}$$

Therefore, equations (7.25), (7.27), (7.28) represent the solution of the MA parameter estimation problem. However, as both the expressions for H^2 and γ_0 can produce two possible values for these parameters, there will be a total of 4 possible combinations for H^2, γ_0, γ_2. To determine the appropriate root for the modelling of antiresonances, the above expressions are compared with those expected for a system consisting of a pair of complex zeros.

Step 4 Relate MA Model Coefficients to Antiresonance Parameters

A complex zero pair transfer function is given by

$$H(z) = \frac{X(z)}{U(z)} = K\left[\left(1 - re^{j\theta}z^{-1}\right)\left(1 - re^{-j\theta}z^{-1}\right)\right]$$

$$= K\left[1 - 2r\cos\theta \,.\, z^{-1} + r^2 z^{-2}\right] \tag{7.29}$$

where (r, θ) are the zero radius and angle respectively and K is the antiresonance gain. This system will thus be associated with a sequence, $x(n)$, of the form

$$x(n) = K\left[u(n) - 2r\cos\theta \,.\, u(n\text{-}1) + r^2 . u(n\text{-}2)\right] \tag{7.30}$$

Relating this expression to that for the MA model, it is apparent that

$$G \equiv K\,; \qquad \beta_1 \equiv 2r\cos\theta\,; \qquad \beta_2 \equiv -r^2 \tag{7.31}$$

In addition, the autocorrelation values can be calculated as above to yield

$$R(0) = K^2\left[1 + 4r^2\cos^2\theta + r^4\right] \tag{7.32a}$$

$$R(1) = K^2\left[-2r\cos\theta.(1+r^2)\right] \tag{7.32b}$$

$$R(2) = K^2.r^2 \tag{7.32c}$$

Substituting these results into equation (7.25) produces

$$H^2 = \frac{K^2\left[(1+r^2)^2 + 4r^2\cos^2\theta\right]}{2} \pm \frac{\sqrt{K^4\left[(1+r^2)^2 + 4r^2\cos^2\theta\right]^2 - 16K^4r^2.\cos^2\theta.(1+r^2)^2}}{2}$$

$$= \frac{K^2\left[(1+r^2)^2 + 4r^2\cos^2\theta\right]}{2} \pm \frac{K^2\left[(1+r^2)^2 - 4r^2\cos^2\theta\right]}{2} \tag{7.33}$$

$$H^2 = \begin{cases} K^2.(1+r^2)^2 & \text{CASE } \boxed{\text{A}} \\ K^2.4r^2\cos^2\theta & \text{CASE } \boxed{\text{B}} \end{cases} \tag{7.34}$$

Now, from above,

$$\gamma_0 = \frac{R(1) \pm \sqrt{\left[R(1)\right]^2 - 4H^2R(2)}}{2H^2} \tag{7.35}$$

Thus, for case $\boxed{\text{A}}$ using equation (7.32),

$$\gamma_0 = \frac{-2K^2r\cos\theta\,(1+r^2)}{2K^2\,(1+r^2)^2} \pm \frac{\sqrt{4K^4r^2\cos^2\theta.(1+r^2)^2 - 4K^4r^2\,(1+r^2)^2}}{2K^2\,(1+r^2)^2}$$

$$\gamma_0 = -\frac{r\cos\theta}{(1+r^2)} \pm \frac{\sqrt{4K^4r^2.(1+r^2)^2.-\sin^2\theta}}{2K^2(1+r^2)^2}$$

$$= -\frac{r\cos\theta}{(1+r^2)} \pm j\frac{r\sin\theta}{(1+r^2)} \tag{7.36}$$

However, it is necessary that γ_0, γ_2 are both real in order that the pair of zeros are complex conjugate and, therefore, the correct value of H^2 must be that provided by case [B].

Thus, $$H^2 = 4K^2r^2\cos^2\theta \tag{7.37}$$

Now, for case [B], using equations (7.32), (7.35), (7.37),

$$\gamma_0 = \frac{-2K^2r\cos\theta\,(1+r^2)}{8K^2r^2\cos^2\theta} \pm \frac{\sqrt{4K^4r^2\cos^2\theta.(1+r^2)^2 - 16K^4r^4\cos^2\theta}}{8K^2r^2\cos^2\theta}$$

$$= -\frac{(1+r^2)}{4r\cos\theta} \pm \frac{\sqrt{4K^4r^2.\cos^2\theta.(1-r^2)^2}}{8K^2r^2\cos^2\theta}$$

$$= -\frac{(1+r^2)}{4r\cos\theta} \pm \frac{(1-r^2)}{4r\cos\theta} \tag{7.38}$$

$$\gamma_0 = \begin{cases} -\frac{r}{2\cos\theta} & \text{CASE [C]} \\ -\frac{1}{2r\cos\theta} & \text{CASE [D]} \end{cases} \tag{7.39}$$

Consider the two cases separately. For case [C], using equations (7.32), (7.37),

$$\gamma_2 = \frac{R(2)}{H^2\gamma_0} = \frac{K^2r^2}{4K^2r^2\cos^2\theta} \cdot -\frac{2\cos\theta}{r}$$

$$\Longrightarrow \qquad \gamma_2 \;=\; -\frac{1}{2r\cos\theta} \qquad \text{for case } \boxed{C} \tag{7.40a}$$

Similarly, for the other case,

$$\gamma_2 \;=\; -\frac{r}{2\cos\theta} \qquad \text{for case } \boxed{D} \tag{7.40b}$$

Comparing these results with equation (7.39) shows that the values of γ_0, γ_2 simply interchange for the 2 cases. It can also be seen that,

$$r \;=\; \sqrt{\frac{\gamma_0}{\gamma_2}} \qquad \text{for case } \boxed{C} \tag{7.41a}$$

and $$r \;=\; \sqrt{\frac{\gamma_2}{\gamma_0}} \qquad \text{for case } \boxed{D} \tag{7.41b}$$

Therefore, a simple means of testing for which case is appropriate is by calculating γ_0, γ_2 for both cases $\boxed{C}$, $\boxed{D}$ and then determining which case fulfils the requirement that $r < 1$ from equation (7.41). Given the value of r, the other two parameters required for characterising the antiresonances, (i.e. θ, K) can also be evaluated from equations (7.39), (7.40) as follows:

$$\cos\theta \;=\; -\frac{r}{2\gamma_0} \qquad \text{for case } \boxed{C} \tag{7.42a}$$

$$\cos\theta \;=\; -\frac{r}{2\gamma_2} \qquad \text{for case } \boxed{D} \tag{7.42b}$$

and, from the equivalences established above for the various types of model:

$$H \;=\; -G\beta_1 \;=\; -2Kr\cos\theta$$

Thus, $$K \;=\; -\frac{H}{2r\cos\theta} \qquad \text{for both cases } \boxed{C}, \boxed{D} \tag{7.43}$$

Step 5 Summarise Algorithm for Antiresonance Estimation from Signal Spectrum

(1) Calculate first three autocorrelation values $\{R(0),\ R(1),\ R(2)\}$ from the signal frequency response in the spectral range under consideration as described for the Selective LP procedure in section 4.2.

(2) Evaluate two possible values for

$$H^2 = \frac{[R(0) + 2R(2)] + \sqrt{[R(0) + 2R(2)]^2 - 4[R(1)]^2}}{2} \qquad \text{CASE } \boxed{A}$$

$$\text{and} \quad H^2 = \frac{[R(0) + 2R(2)] - \sqrt{[R(0) + 2R(2)]^2 - 4[R(1)]^2}}{2} \qquad \text{CASE } \boxed{B}$$

(3) Calculate two values of γ_0 for each of the two cases [A] and [B],

$$\gamma_0 = \frac{R(1) \pm \sqrt{[R(1)]^2 - 4H^2 R(2)}}{2H^2}$$

Only one of the cases [A], [B] will produce two real roots for γ_0. Choose that case and thereby determine the correct value of H^2.

(4) The two values of γ_0 corresponding to this one value of H^2 represent two more cases for consideration.

$$\gamma_0 = \frac{R(1) + \sqrt{[R(1)]^2 - 4H^2 R(2)}}{2H^2} \qquad \text{CASE } \boxed{C}$$

$$\gamma_0 = \frac{R(1) - \sqrt{[R(1)]^2 - 4H^2 R(2)}}{2H^2} \qquad \text{CASE } \boxed{D}$$

(5) Determine two possible values of γ_2 given by:

$$\gamma_2 = \frac{R(2)}{H^2\gamma_0}$$

(6) Evaluate two values for the complex zero radius (r) according to

$$r = \sqrt{\frac{\gamma_0}{\gamma_2}}$$

Choose either case [C] or [D] depending on which produces the result $r < 1$ and thus obtain estimates of γ_0, γ_2, r.

(7) Calculate the zero angle (θ) and antiresonance gain (K) according to:

$$\theta = \cos^{-1}\left(-\frac{r}{2\gamma_0}\right); \qquad K = -\frac{H}{2r\cos\theta}$$

(8) The antiresonance parameters (r, θ, K) are now available and may be used to calculate the model frequency response, $H\left(e^{j\omega_m}\right)$, as follows:

$$H\left(e^{j\omega_m}\right) = K\left(1 - 2r\cos\theta . e^{-j\omega_m} + r^2 e^{-2j\omega_m}\right)$$

An analysis procedure is now available for the calculation of complex zero positions from segments of a signal spectrum. The next step involves obtaining a suitable frequency response for the application of this method.

7.3.3 Zero Estimation from LP Residual Spectra using Selective Moving Average Models

In section 7.2.2, it was mentioned that an LP inverse filter is often applied to ARMA signals in order to remove the effect of signal poles prior to an estimation of the zero positions. A similar analysis is used in this section to provide a signal spectrum which can be segmented into regions containing antiresonances so that the 2nd order MA model developed in the previous subsection may be employed to model these segments and thereby obtain zero estimates for nasalized speech.

A spectral approach as described in Chapter 4 was used to derive a 12th order model spectrum for the pitch periods of the utterance employed as the basis of Fig. 7.4 and then used to inverse filter the pitch synchronous spectra. The resulting LP residual spectra are shown in Fig. 7.5 superimposed on each other for the same 8 pitch periods described by Fig. 7.4. The positions of the various spectral antiresonances are clearly in evidence in this result.

A dynamic programming procedure similar to that developed in section 5.2 for the partitioning of a speech frequency response into formant regions was then employed on the LP residual spectra with 3 basic modifications. The constraint concerning a minimum frequency range for each segment in order to prevent pitch harmonics wrongly being classified as formants is obviously not applicable and was, therefore, relaxed (as in section 6.4.3) to the limit given in Chapter 4 such that the number of spectral samples in any region is at least twice the predictor order in that segment.

Another change relates to the choice of error criterion in the segmentation procedure. The weighted error measure was used in Chapter 5 because of the particular nature of voiced speech sounds [7.17] and these considerations obviously do not apply to LP residual spectra and thus the simple squared error measure discussed in section 4.5.2 was employed in this analysis. The final alteration relates to the model associated with each region since the 2nd order Selective LP configuration is obviously replaced by the 2nd order MA model detailed in

the previous subsection.

The optimal segmentation of one particular pitch synchronous spectrum into 8 regions is illustrated in Fig. 7.6. There is a major difference between this form of segmentation for residual spectra and that for actual speech frequency responses. It was demonstrated in Chapter 6 that the cascade model consisting of a sequence of 2nd order resonators tuned to different frequencies was a reasonable model for voiced speech. The spectra corresponding to this class of speech sounds could, therefore, be segmented into a series of adjoining Selective LP segments in the frequency domain.

However, no such guarantee exists for the LP residual spectrum which is usually characterised by a set of sharp minima located at the zero frequencies and a generally flat response between these antiresonances [7.9]. It is, therefore, apparent that the Selective MA model is only considered to be a good model of the residual spectrum in the vicinity of the antiresonances and is not a suitable representation for the entire spectrum. Matching this spectrum with a set of Selective MA regions will, therefore, result in some of the segments accurately representing the minima and the remaining regions corresponding to the flat portions of the response and thus being associated with very large bandwidth values.

This is the form of segmentation depicted in Fig. 7.6 and so only those regions which correspond to a reasonably low bandwidth value are chosen as being representative of the spectral zeros. Antiresonances associated with nasalized speech are known to have bandwidths of less than 500 Hz [7.1],[7.18] and so this value was chosen as a threshold for the choice of antiresonance regions. In the case of Fig. 7.6, this results in regions $2, 3, 5, 6$ being selected as those concerned with the modelling of spectral zeros.

The next analysis stage involves an estimation of the zero centre frequencies and bandwidths and these estimates may be calculated in a similar manner to that described in section 5.4 for the Selective LP regions. The zero radius and angle in the Selective MA domain

(r', θ') are available as part of the modelling procedure outlined in the last subsection and, in order to account for the disparity between the frequency ranges of the Selective MA and signal spectra, an inverse mapping is required to convert these values to (r, θ) in the actual frequency domain. As in section 5.4, the transformation of the radius ensures the same zero bandwidth in both domains [7.19] and is given by

$$r = (r')^{f_s'/f_s} \tag{7.44}$$

where f_s' is the effective 'sampling' frequency associated with the Selective MA region as described in section 5.4.1 and f_s is the signal sampling frequency. Also,

$$\theta = \frac{2}{N}\left[\pi(t_n - r_n) + \theta' r_n\right] \tag{7.45}$$

where N is the number of samples in the complete spectrum and r_n, t_n are the length and endpoint respectively of the segment under investigation. This subsection has, therefore, outlined a procedure for deriving zero estimates directly from an LP residual spectrum using a dynamic programming formulation for the segmentation of this frequency response into regions characterised by second order moving average models.

7.4 SYNTHESIS OF SPEECH MODELLED BY A COMBINED POLE-ZERO REPRESENTATION

7.4.1 Linear Predictive Analysis of Inverse MA Filtered Nasalized Speech

If a certain number (N_z) of zero positions $\left\{r_k e^{j\theta_k},\ r_k e^{-j\theta_k} : k=1,\ldots,N_z\right\}$ have been estimated for an ARMA signal, the combined effect of these zeros may be represented as

$$B\left(e^{j\omega_m}\right) = \prod_{k=1}^{N_z}\left(1 - r_k e^{j\theta_k}.e^{-j\omega_m}\right)\left(1 - r_k e^{-j\theta_k}.e^{-j\omega_m}\right)$$

$$= \prod_{k=1}^{N_z}\left(1 - 2r_k\cos\theta_k.e^{-j\omega_m} + r_k{}^2.e^{-2j\omega_m}\right) \tag{7.46}$$

An expansion of these N_z product terms will then yield a response of the form:

$$B\left(e^{j\omega_m}\right) = 1 - \sum_{l=1}^{2N_z}\beta_l\, e^{-jl\omega_m} \tag{7.47}$$

However, from section 7.2, the frequency response, $H\left(e^{j\omega_m}\right)$, of the complete ARMA process may be defined as:

$$H\left(e^{j\omega_m}\right) = G\,\frac{1 - \sum_{l=1}^{q}\beta_l\, e^{-jl\omega_m}}{1 - \sum_{k=1}^{p}\alpha_k\, e^{-jk\omega_m}} \tag{7.48}$$

The function, $B\left(e^{j\omega_m}\right)$, is thus equivalent to the numerator polynomial in this expression with $q=2N_z$. The response $H\left(e^{j\omega_m}\right)$ can now be divided by the moving average filter characteristic $B\left(e^{j\omega_m}\right)$ to produce a new frequency response $H'\left(e^{j\omega_m}\right)$,

$$H'\left(e^{j\omega_m}\right) = \frac{H\left(e^{j\omega_m}\right)}{B\left(e^{j\omega_m}\right)} = \frac{G}{1 - \sum_{k=1}^{p}\alpha_k\, e^{-jk\omega_m}} \tag{7.49}$$

This result is obviously equivalent to the response of a standard LP model and thus, if a pole-zero signal is inverse filtered by a moving average filter constituted of all the system zeros, the resulting signal should consist entirely of poles and be directly amenable to LP analysis.

This process was implemented with the results available from the zero estimation algorithm outlined in the previous subsection. For the pitch synchronous LP residual spectrum used as the basis of the spectral segmentation shown in Fig. 7.6, the 4 zeros derived from the

antiresonance estimation procedure produce a combined frequency response of the form shown in Fig. 7.7.

Based on this result, the inverse MA filtered spectrum is shown as the underlying curve in Fig. 7.8 and it is apparent that the sharp minima evident in Fig. 7.4 have now been effectively removed to yield a response which is modelled as shown in Fig. 7.8 by a 10th order LP spectrum. These linear predictive coefficients together with the zero parameters form a combined pole-zero model of the original nasalized speech. It is also clear from section 7.2 that a synthesised speech signal, $s(n)$, can now be produced using the following expression:

$$s(n) = \sum_{k=1}^{p} \alpha_k \, s(n\text{-}k) + G\left[u(n) - \sum_{l=1}^{q} \beta_l \, u(n\text{-}l)\right] \tag{7.50}$$

where $u(n)$ is a sequence of unit impulses spaced at the required pitch period and G is the gain of the LP model of Fig. 7.8. A narrowband spectrogram of speech synthesised on this basis for the nasalized utterance corresponding to the voiced phoneme in the French word "*son*" is shown in Fig. 7.9 (b) along with the equivalent spectrogram of the speech signal in Fig. 7.9 (a). In the next subsection, these results are compared with the synthetic speech produced by standard LP analysis.

7.4.2 Comparison with Standard LP Modelling

As the combined pole-zero model described above consists of a total of 4 complex conjugate zeros and a 10th order LP model, the overall number of parameters required to specify the model is 18 since each complex zero is constituted of 2 parameters — the zero radius (r) and angle (θ). In order to effect a reasonable comparison, an 18th order LP analysis was, therefore, performed on the same nasalized utterance. This procedure was again carried out on a pitch synchronous basis and the spectrogram of the speech produced by this approach is shown in Fig. 7.9 (c).

Another means of comparison is also provided by Fig. 7.10 and Fig. 7.11 which illustrate the pitch synchronous spectra available from speech synthesised by the combined pole-zero model and the all-pole configuration respectively. These figures depict the superimposed frequency responses of 8 consecutive pitch periods from these two signals and are, therefore, comparable to the similar result for the original utterance shown in Fig. 7.4.

Both the spectrograms and the pitch synchronous spectra show that the combined pole-zero model output is a better representation of the speech signal in terms of preserving the signal antiresonances. Listening tests also tend to support the same conclusion and it is interesting to note that the synthetic speech produced by the standard LP analysis is more akin to the utterance /OW/ as in the word "b*ough*t" – which is the non-nasalized equivalent of the speech sound under investigation and thus illustrates the inadequacy of conventional linear predictive methods being applied to the representation of nasalized speech.

7.5 CONCLUSIONS

The properties of nasal consonants and nasalized vowels differ in one fundamental respect from the vowel-like sounds studied in the rest of this thesis. Nasalized speech contains antiresonances which result in the requirement of pole-zero analysis for this class of speech sounds. The traditional pole-zero estimation techniques discussed in the first part of this chapter could loosely be classified into two specific forms of analysis approaches – iterative and non-iterative methods.

The iterative schemes are generally characterised by the need to perform complex mathematical operations and have not achieved widespread use in speech processing systems [7.3]. Most of the non-iterative procedures rely on an initial all-pole modelling stage (which is intended to remove the effect of the system poles) followed by an estimation of the zero positions. However, several of these methods are speculative and have not been found to produce acceptable results with real speech sounds while those others which operate on a

homomorphically smoothed spectral estimate tend to yield poor estimates of antiresonance bandwidths [7.20]. All of these methods also require a priori estimates of the number of poles and zeros to be derived during the analysis — which may not correspond to the actual number of formants and antiresonances in the speech spectrum.

The primary aim of this chapter has been to calculate zero estimates using spectral segmentations consisting of 2nd order Selective moving average models. In order to achieve this objective, a fundamental requirement is the availability of spectral estimates for which the antiresonance minima are clearly in evidence. The discussion of section 7.3.1 illustrates the importance of pitch synchronous analysis in this respect since it eliminates the effect of two factors other than the vocal tract/nasal tract combination which may also give rise to spectral zeros. These are the excitation source, which produces pitch harmonics in pitch asynchronous spectra, and the effect of the analysis window at the particular analysis points associated with the pitch synchronous Fourier Transform. The method used in this chapter to actually achieve the necessary pitch determination was that of homomorphic processing.

It was then shown that the formulation of a conventional second order moving average model will lead to a set of equations which must be solved iteratively in order to obtain the model coefficients. However, a suitable transformation of these equations was found to produce an exact result with 4 possible solutions for the transformed parameters. A technique was then described for choosing that solution which is suitable for the modelling of antiresonances and thereby obtain a general procedure for the calculation of complex conjugate zero positions directly from segments of a signal spectrum.

This analysis scheme was then incorporated into a dynamic programming method for the segmentation of an LP residual spectrum into regions containing isolated antiresonances. The residual frequency response was chosen because of the spectral 'whitening' character [7.9] of the LP inverse filter which tends to remove the effect of system poles in producing the LP residual.

Zero estimates derived from the segmentation process were then used to inverse filter the original speech waveform and thereby yield a signal which is directly amenable to standard LP analysis. The linear predictive coefficients together with the zero positions constitute a combined pole-zero model which was employed in the production of synthetic speech and finally compared to a similar signal available from a conventional LP approach. The output of the combined pole-zero model was found to represent a closer approximation to the nasalized speech with the nature of the spectral zeros being well preserved in comparison to the synthetic speech derived from an all-pole characterisation.

7.6 REFERENCES

[7.1] KLATT, D.H., "Software for a Cascade/Parallel Formant Synthesizer", *J. Acoust. Soc. Amer., Vol.* 67, *No.* 3, *March* 1980, *pp* 971-955.

[7.2] FLANAGAN, J.L., "Speech Analysis, Synthesis and Perception", *2nd ed., Springer-Verlag, Berlin-Heidelberg-New York*, 1972.

[7.3] FALLSIDE, F., WOODS, W.A., "Computer Speech Processing", *Prentice-Hall, Englewood Cliffs, New Jersey*, 1985.

[7.4] RABINER, L.R., SCHAFER, R.W., "Digital Processing of Speech Signals", *Prentice-Hall, Englewood Cliffs, New Jersey*, 1978.

[7.5] MAKHOUL, J., "Linear Prediction: A Tutorial Review", *Proc. IEEE, Vol.* 63, *April* 1975, *pp* 561-580.

[7.6] TRETTER, S.A., STEIGLITZ, K., "Power-Spectrum Identification in terms of Rational Models", *IEEE Trans. Automat. Control, Vol. AC*-12, *April* 1967, *pp* 185-188.

[7.7] SHANKS, J.L., "Recursion Filters for Digital Processing", *Geophysics, Vol.* 32, *No.* 1, 1967, *pp* 33-51.

[7.8] OPPENHEIM, A.V., "Applications of Digital Signal Processing", *Prentice-Hall, Englewood Cliffs, New Jersey*, 1978.

[7.9] MARKEL, J.D., GRAY, A.H. Jr., "Linear Prediction of Speech", *Springer-Verlag, Berlin-Heidelberg-New York*, 1976.

[7.10] MAKHOUL, J., "Selective Linear Prediction and Analysis-by-Synthesis in Speech Analysis", *Bolt Beranek and Newman Inc., Cambridge, Mass., Rep.* 2578, *April* 1974.

[7.11] OPPENHEIM, A., TRIBOLET, J.M., "Pole-Zero Modeling using Cepstral Prediction", *Res. Lab. Electronics, M.I.T., Cambridge, Mass., QPR* 111, 1973, *pp* 157-159.

[7.12] MORF, M., KAILATH, T., DICKINSON, B., "General Speech Models and Linear Estimation Theory", *IEEE Speech Proc. Conf., Pittsburgh, Pa., April* 1974.

[7.13] SONG, K.H., UN, C.K., "Pole-Zero Modeling of Speech Based on High-Order Pole Model Fitting and Decomposition Method", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP-31, No. 6, Dec.* 1983, *pp* 1556-1565.

[7.14] PINSON, E.N., "Pitch-Synchronous Time-Domain Estimation of Formant Frequencies and Bandwidths", *J. Acoust. Soc. Amer., Vol* 35, *August* 1963, *pp* 1264-1273.

[7.15] KAY, S.M., MARPLE, S.L. Jr., "Spectrum Analysis – A Modern Perspective", *Proc. IEEE, Vol.* 69, *No.* 11, *Nov.* 1981, *pp* 1380-1419.

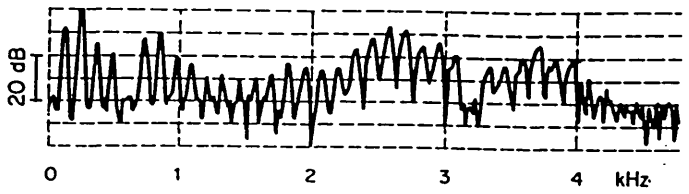
[7.16] BOX, G.E.P., JENKINS, G.M., "Time Series Analysis: Forecasting and Control", *Holden-Day, San Francisco, CA*, 1970.

[7.17] CHHATWAL, H.S., CONSTANTINIDES, A.G., "Application of Dynamic Programming to the Optimal Segmentation of Speech Spectra for Selective Linear Prediction Modelling", *Proc. IEE Intl. Conf. on Speech Input/Output: Techniques and Applications, March* 1986, *pp* 43-48.

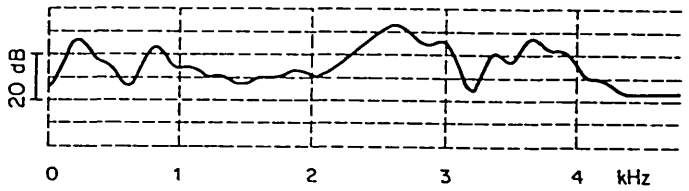
[7.18] FUJIMURA, O., "Analysis of Nasal Consonants", *J. Acoust. Soc. Amer., Vol.* 34, *No.* 12, *Dec.* 1962, *pp* 1865-1875.

[7.19] CHHATWAL, H.S., CONSTANTINIDES, A.G, "Speech Spectral Segmentation for Spectral Estimation and Formant Modelling", *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Proc., Dallas, Texas, April* 1987, *Vol.* 1, *pp* 316-319.

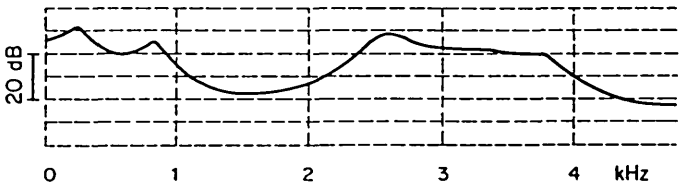
[7.20] MAKHOUL, J., "Spectral Linear Prediction: Properties and Applications", *IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP-23, No.* 3, *June* 1975, *pp* 283-296.



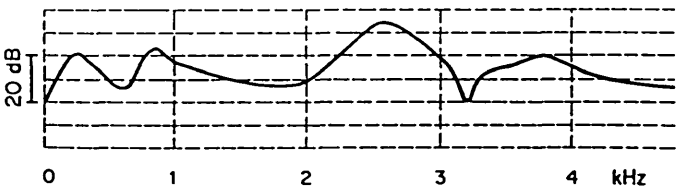
(a) Spectrum of the Nasalized Consonant /m/



(b) Spectral Envelope obtained by Cepstral Smoothing



(c) Spectrum Corresponding to a 12-pole Linear Predictive Analysis of (b)



(d) Spectrum Corresponding to a 12-pole 10-zero Model with Zero Estimation by Inverse LP Modelling

Figure 7.1 Pole-Zero Analysis of Nasalized Speech Based on Inverse LP Modelling (after Oppenheim [7.8])

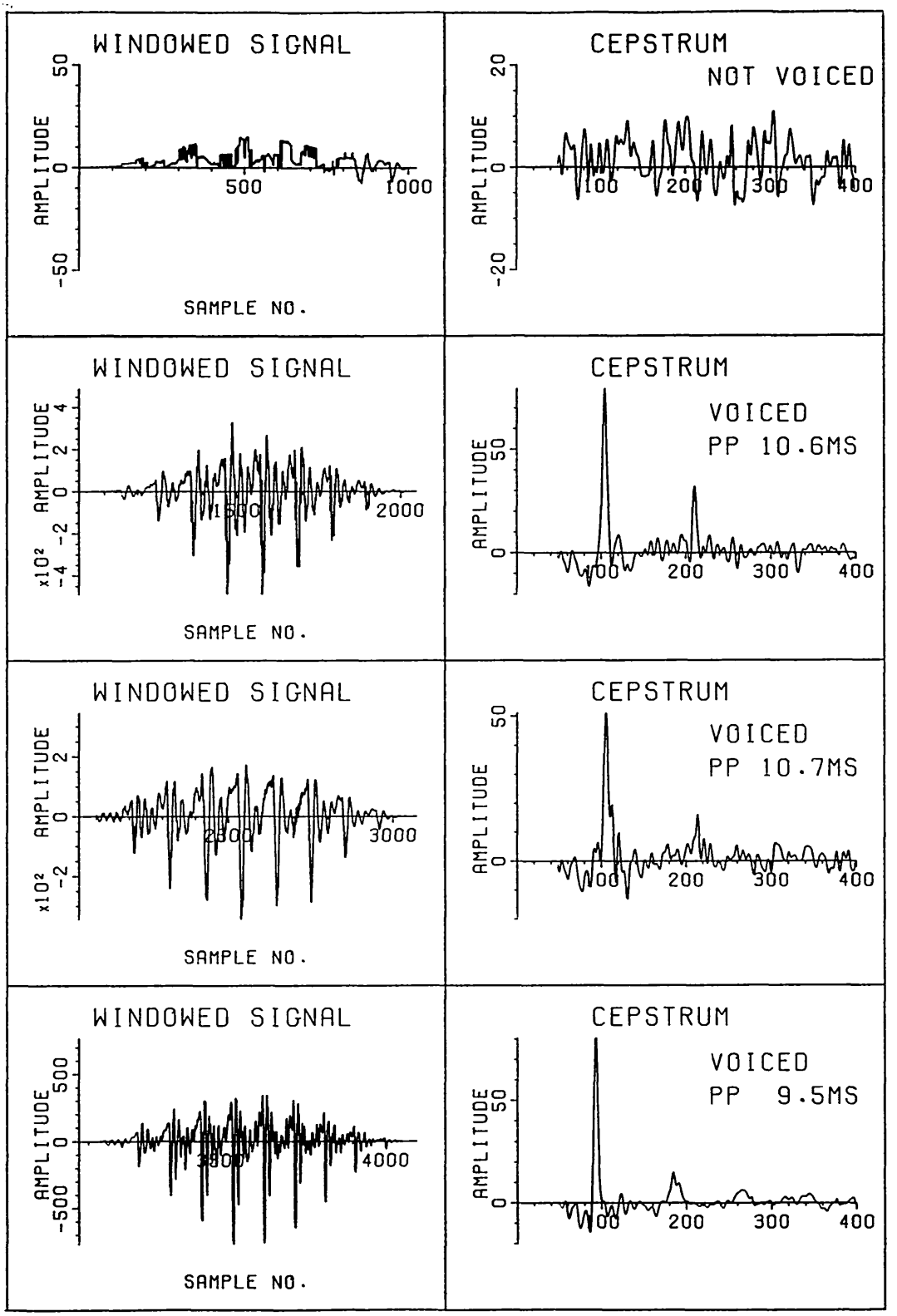


Figure 7.2 Illustration of Voiced/Silence Discrimination and Pitch Estimation using Homomorphic Processing

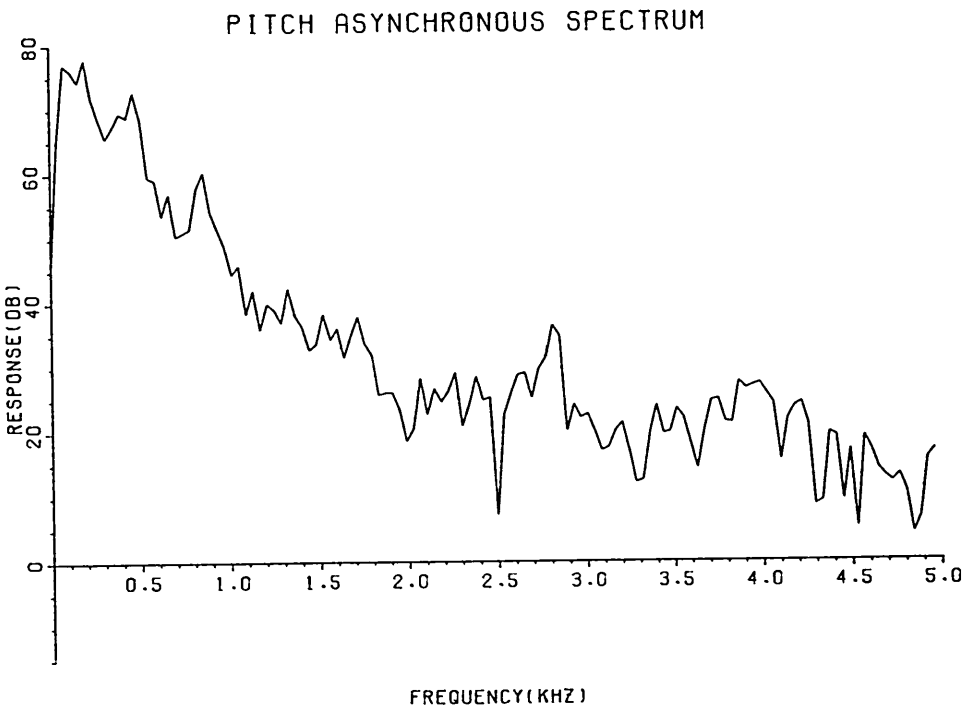


Figure 7.3 Pitch Asynchronous Spectrum of a Nasalized Speech Sound

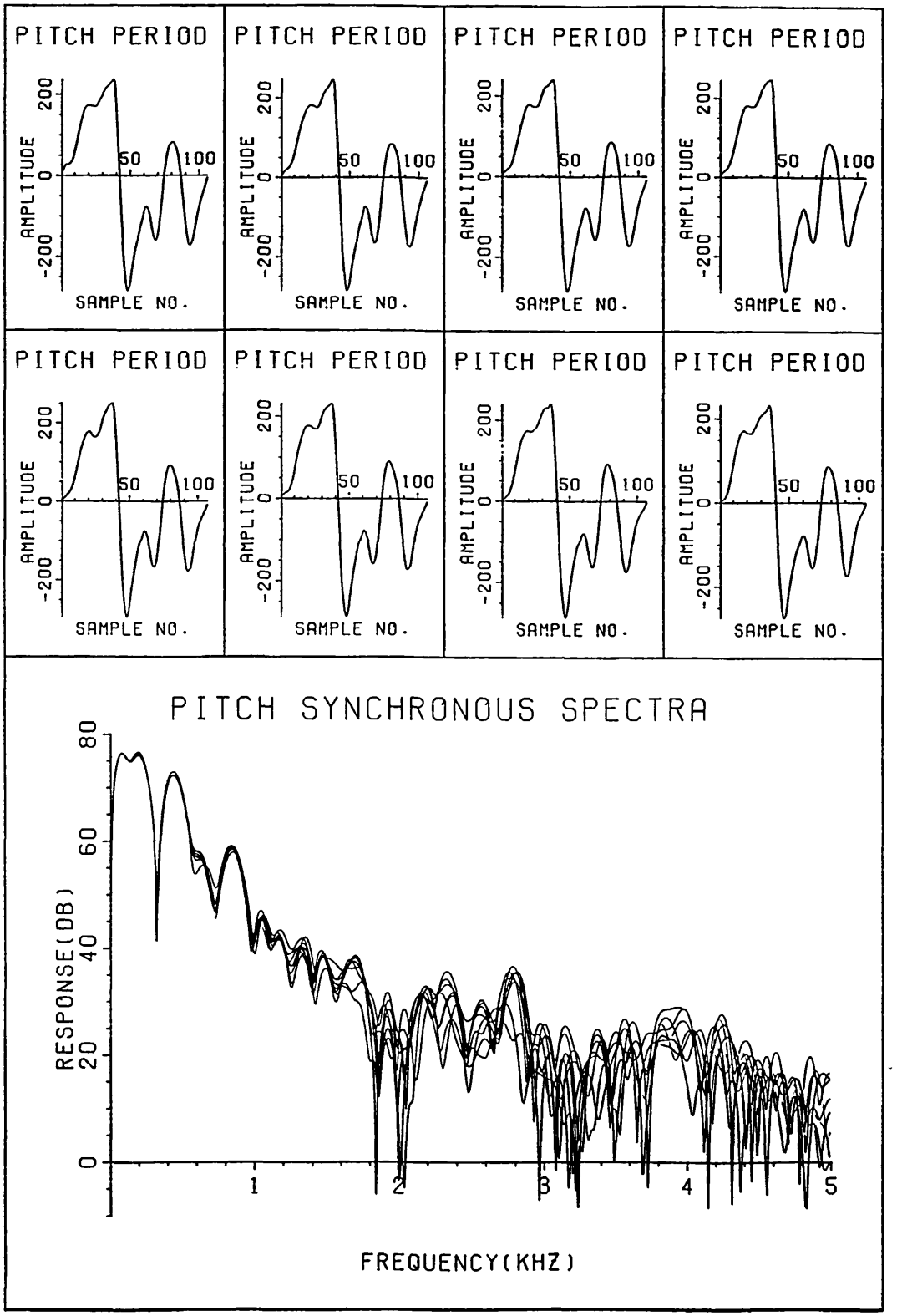


Figure 7.4 Isolated Pitch Periods and Corresponding Pitch Synchronous Spectra for 8 Consecutive Cycles of a Nasalized Utterance

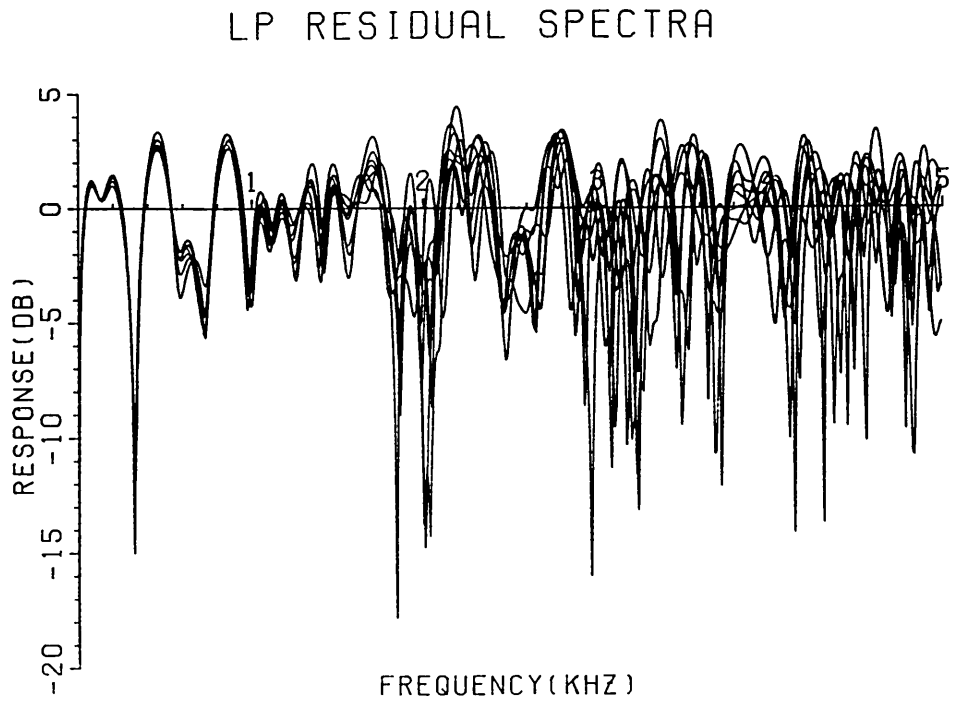


Figure 7.5 Linear Predictive Residual Spectra based on Pitch Synchronous Analysis of 8 Pitch Periods from a Nasalized Utterance

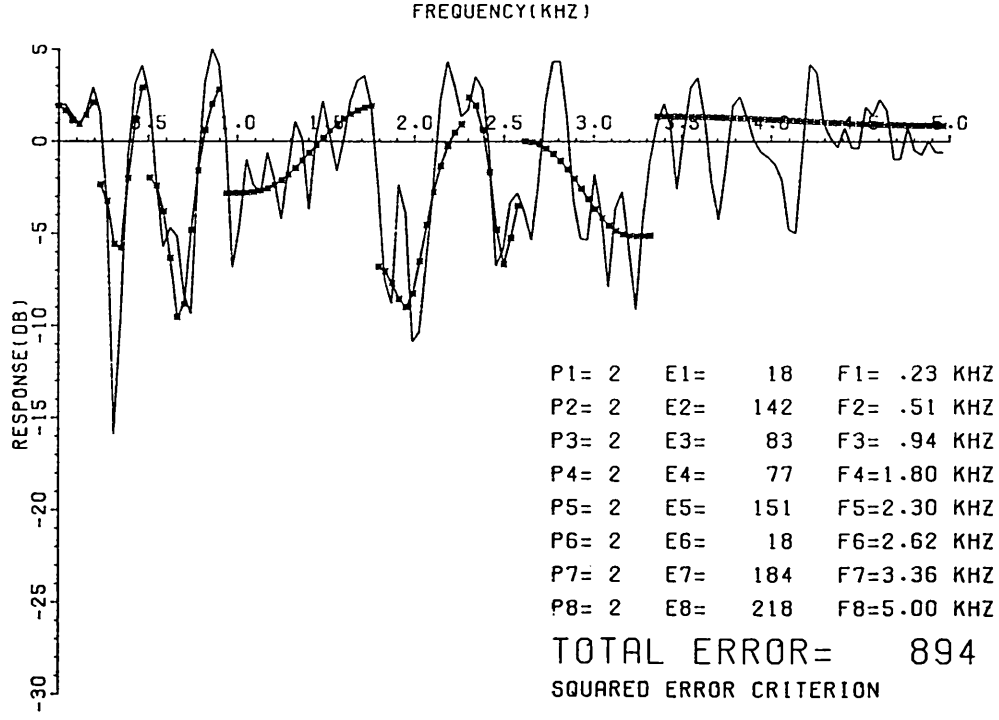


Figure 7.6 Spectral Segmentation into Antiresonance Regions for a Pitch Synchronous LP Residual Spectrum

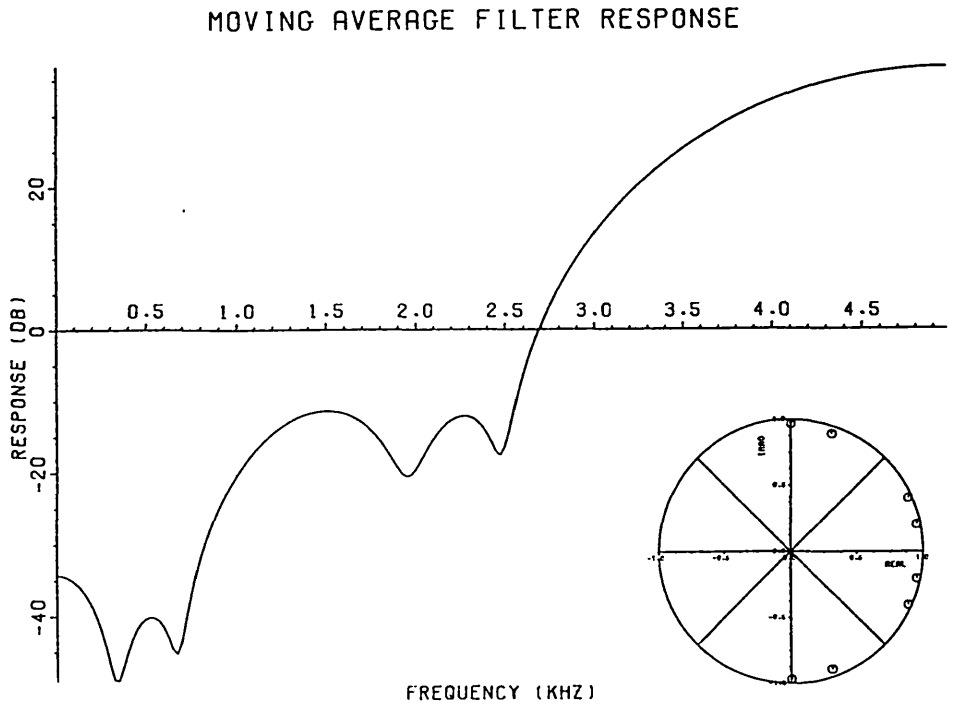


Figure 7.7 Frequency Response of Moving Average Filter obtained from Procedure for Spectral Segmentation into Antiresonances

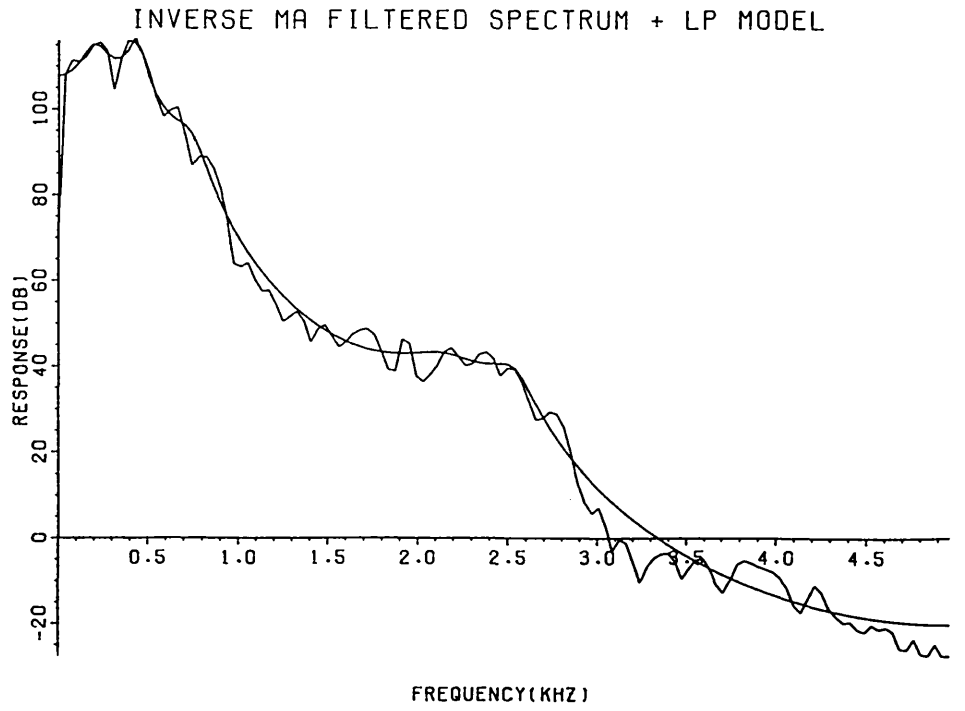


Figure 7.8 Spectra for Inverse MA Filtered Speech Signal and Corresponding 10th Order Linear Predictive Model

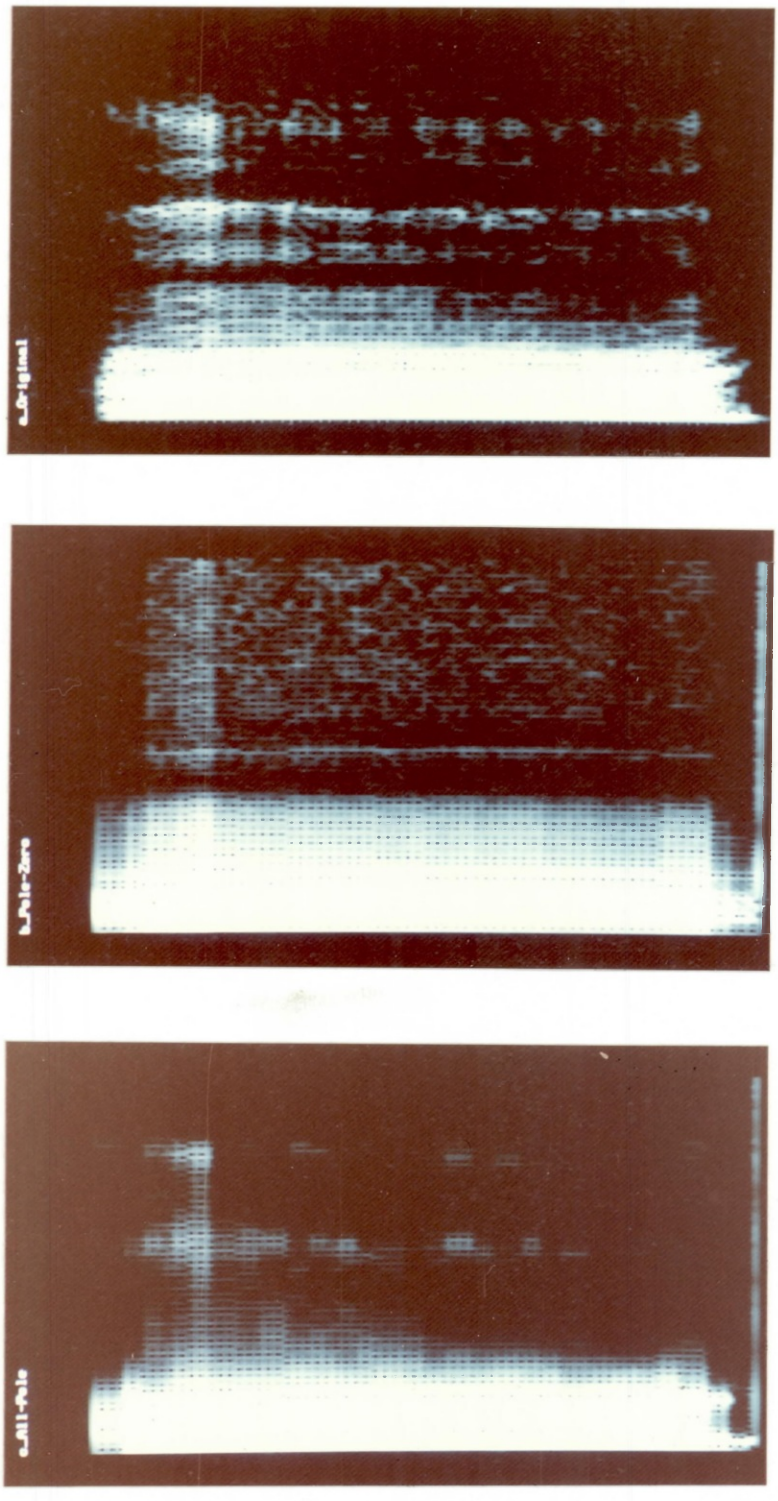


Figure 7.9 Narrowband Spectrograms for
(a) Nasalized Speech Sound
(b) Speech Synthesised from Combined Pole-Zero Model based on Spectral Segmentation into Antiresonances
(c) Speech Synthesised from Standard All-Pole LP Analysis

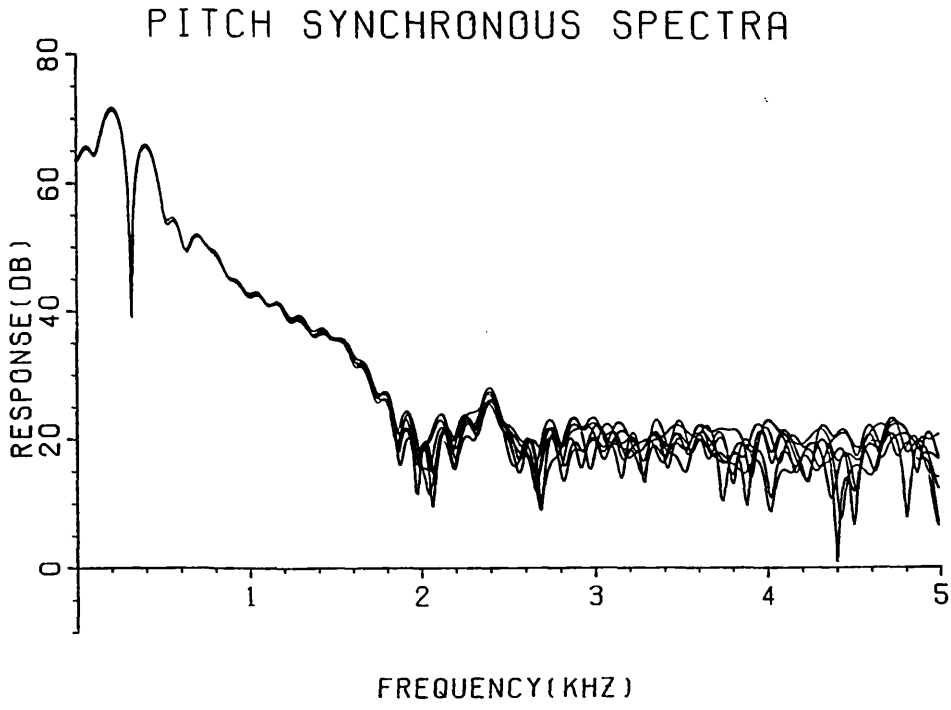


Figure 7.10 Pitch Synchronous Spectra for 8 Cycles of Synthetic Speech Derived from a Combined Pole-Zero Model of Nasalized Speech

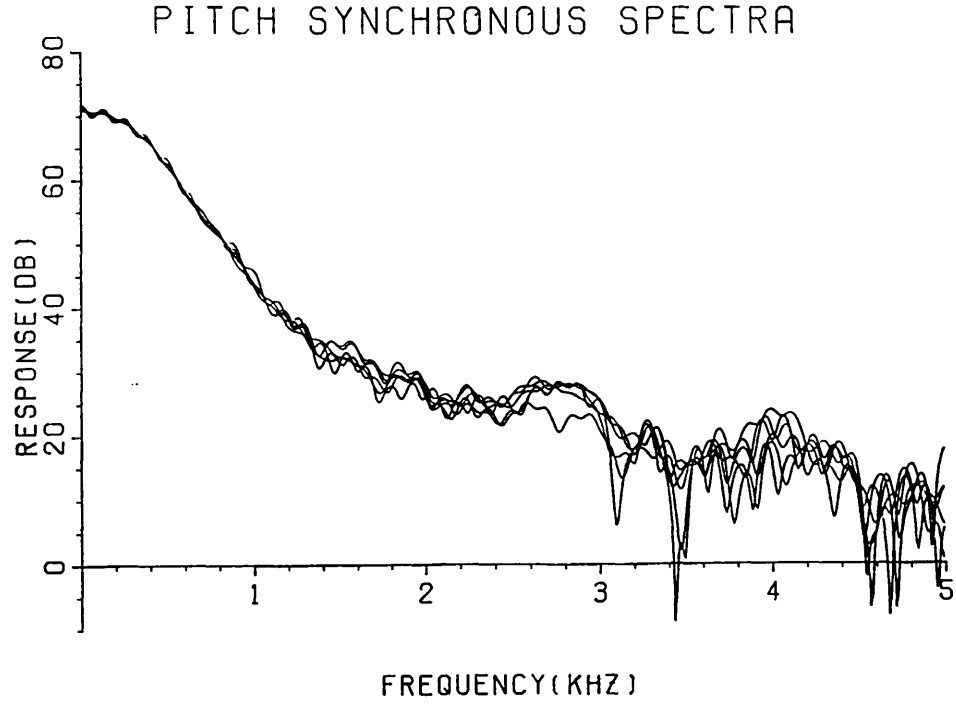


Figure 7.11 Pitch Synchronous Spectra for 8 Cycles of Synthetic Speech Derived from a Conventional LP Analysis of Nasalized Speech

CHAPTER 8

CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH

CHAPTER 8. CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH

8.1 DISCUSSION

Although numerous models have been postulated for the description of the speech process [8.2], no single characterisation has been developed which can account for all the observed complexities of human speech. However, the application of Linear Predictive (LP) techniques to the modelling of speech behaviour is one of the most powerful and widely used methods of digital speech analysis. This widespread utility has arisen primarily since the LP approach is relatively straightforward, computationally inexpensive, its properties are reasonably well understood and, most importantly, it provides an extremely effective model of the speech signal for many vocalisations [8.1].

These characteristics contribute largely to the predominance of LP methods in most of the fundamental areas of speech processing such as low bit rate transmission or storage of speech signals, speech synthesis, speaker/speech recognition and the accurate estimation of the basic speech parameters such as fundamental frequency, formants etc. [8.2]. Another important aspect of LP analysis is its applicability in the area of speech spectral modelling, which is further enhanced by the fact that a frequency domain interpretation is also available for LP analysis such that the linear predictive model may be derived directly from a signal spectrum.

This concept can also be extended to the modelling of a particular portion of the discrete signal spectrum as detailed in section 4.2 for the Selective LP procedure. Using this basic technique, it is possible to obtain accurate spectral estimates as described in Chapter 4 based on an optimal segmentation of the speech frequency response into several regions, each of which is characterised by a Selective LP model covering a particular frequency range and associated with a certain predictor order. This procedure yields the sets of break frequencies

separating the individual segments and the allocation of predictor orders for a given overall model order required to minimise the discrepancy between the signal and Selective LP spectra on the basis of a suitable error criterion.

The approach employed in this thesis to perform this optimal partitioning is that of Dynamic Programming (DP) and the fundamentals of DP are also outlined in Chapter 4 as a methodology for the solution of a class of multi-variate optimisation problems. There are several reasons for the suitability of the DP technique to the segmentation algorithm developed in Chapter 4. The first consideration is that the introduction of certain forms of constraints can be handled effectively by dynamic programming formulations whereas these same restrictions may pose severe difficulties for procedures based on classical methods [8.3].

The two types of constraint which are particularly relevant to the work described in this thesis are the imposition of integrality which is required for the modelling of discrete spectra and the incorporation of limits on the domains of variation for the state and decision variables — as it is necessary to restrict the minimum and maximum possible values at each stage for the region lengths, endpoints and predictor orders as discussed in section 4.4.

The ability of DP to obtain global extrema rather than local optima is also a property which is not generally associated with other computational methods [8.4]. Another advantageous characteristic of dynamic programming is concerned with its "embedded" nature whereby the solution of a particular problem will also yield solutions for all the other subproblems which are embedded in the larger problem. In the case of speech spectral segmentation, this is of consequence since the criterion developed in Chapter 5 for the number of regions to employ in the procedure for spectral segmentation into formants requires not only the final optimal segmentation into R regions, but also the optimal partitions for $1,\ldots,R-1$ segments.

Another important consideration in the development of the optimal breakpoint selection/predictor order allocation technique outlined in Chapter 4 is the incorporation of a spectral error measure which is able to quantify the mismatch between the signal and model spectra within the given Selective LP region. Various error criteria were considered and each of these was found to result in considerably improved spectral estimates when compared to standard LP modelling. This investigation also illustrates the ease with which a change in the required form of the error criterion may be implemented in the segmentation algorithm using the DP formulation. A similar alteration could, in contrast, require a complete reappraisal of the optimisation methodology for classical methods.

The optimal segmentation technique of Chapter 4 was modified in Chapter 5 to enable the partitioning of a non-nasalized voiced speech spectrum into second order Selective LP regions which contain single formant resonances. This objective required the incorporation of certain constraints which were specifically aimed at isolating formant-type regions and, in some cases, actually resulted in a considerable computational saving over the algorithms developed in Chapter 4. On the basis of the reduction of the overall error between real and model spectra with an increasing number of regions in this form of segmentation, an effective criterion was also developed for the selection of an appropriate number of segments (and thereby formants) for the characterisation of the speech spectral density.

This result is of significance since a value for the number of segments employed in the model also yields an overall predictor order for the Selective LP models whereas the choice of a predictor order for a conventional linear predictive system always represents a complex and problematic decision. This consideration is even more critical in the case of procedures aimed at speech formant estimation since an inappropriate value will either lead to an inadequate number of poles to characterise the speech formants or the selection of extraneous peaks in the speech spectrum. In contrast to the criterion employed in the procedure for spectral segmentation into formants, the model order selection criteria utilised in standard LP analysis

were generally found to be associated with serious theoretical and implementational drawbacks which impair their effectiveness for the specific purpose of formant modelling.

A procedure was also introduced in Chapter 5 for the estimation of the positions of those formants characterised by each of the second order Selective LP systems. As the predictor order is only 2 for these models, the calculation of the pole parameters from the model coefficients in the Selective LP domain is relatively straightforward and a simple inverse mapping can then be employed to convert these values to formant centre frequency and bandwidth estimates in the actual frequency domain. The effectiveness of this technique was then compared for an artificial speech signal to those methods which are traditionally employed in the estimation of formant positions from LP models. The segmentation approach was found to produce significant improvements in the accuracy of formant estimates in comparison to those calculated from standard LP analysis.

As a means of assessment for those formant parameters derived from real speech data using the formant extraction algorithms outlined in Chapter 5, these estimates were incorporated into suitable formant synthesiser configurations in Chapter 6. For the representation of the non-nasalized voiced utterances studied in the spectral segmentation method, it was suggested that the most appropriate form of model network is a cascade formant synthesiser and a suitable system of this type was outlined.

A network of this form was employed in the production of cascade model spectra and these results indicated that the procedure for spectral segmentation into formants produced closer approximations to the original speech spectral densities than the model spectra based on formants calculated using standard LP methods. It was also found that there are several disadvantages associated with these conventional LP techniques which can lead to severe aberrations in the cascade model outputs. These problems include the overestimation of formant bandwidths, an incorrect spectral emphasis for the synthesised frequency response and, in some cases, a total absence of some speech formants from the model spectra.

The extension of the formant extraction and synthesis methods to complete speech utterances was achieved through the use of pitch synchronous processing. The reasons for adopting this approach were also discussed and include factors such as the difficulties associated with a calculation of precise formant estimates — especially formant bandwidths — from pitch asynchronous speech data [8.5]. In Chapter 6, the technique used to isolate the individual pitch periods constituting the speech waveform was based on the electroglottographic (EGG) signal which was recorded as a second synchronised channel of information alongside the speech waveform for a non-nasalized voiced utterance.

The pitch synchronous spectra corresponding to these distinct cycles of the speech signal were then used as the basis of the formant extraction algorithms. Some slight modifications were also incorporated into the segmentation procedure to take account of the actual nature of speech formants. The synthetic speech output available from the cascade model configurations displayed similar characteristics to the results for isolated speech frames with the procedure for spectral segmentation into formants again producing a more accurate representation of the original vocalisation than that provided by standard LP methods.

Although these results indicate that the segmentation approach has certain advantages for the objective of formant estimation, it is also associated with a greater computational expense than conventional LP techniques. One of the most important applications for the work described in Chapters 5 and 6 may, therefore, lie in the area of "*reference formant analysis*". Thus the procedure for spectral segmentation into formants could be regarded as a reference against which the performance of other formant analysers can be compared. Alternatively, the formant estimates derived from the segmentation methods could provide high quality data for comparison with formant frequencies produced by rule-generated synthesis from text [8.6].

It may also be possible to increase the computational speed for the calculation of optimal segmentations by implementing a hardware system realisation of the methods

described in Chapters 4,5,6. This is since the dynamic programming formulation is such that, even though the computational load may be high, the individual steps required to compute the segmentation are themselves relatively simple and straightforward to implement.

The analysis of nasalized speech in Chapter 7 was aimed at the development of an effective pole-zero model for the representation of those resonances and antiresonances which generally characterise this form of utterance. Traditional methods of pole-zero analysis can either be classified as complex iterative schemes or non-iterative techniques which are often speculative or are associated with problems such as inaccuracies in the calculation of antiresonance bandwidths. All of these methods also require a reasonable a priori estimate of the number of poles and zeros to be derived during the analysis procedure [8.7] and this situation is obviously a disadvantage for speech signals since it is often difficult to assess beforehand the number of resonances and antiresonances likely to be exhibited by a particular speech spectrum.

Pitch synchronous processing was again employed in the work described in Chapter 7 since this form of analysis tends to reduce the effect of two major causes of spectral zeros in pitch asynchronous spectra. These are the pitch harmonics caused by the excitation source [8.5] and the zeros resulting from the application of an analysis window to the speech data frame [8.2]. The required antiresonances caused by the vocal tract/nasal tract combination will thus be enhanced in the pitch synchronous spectra.

A fundamental requirement for the pole-zero analysis method investigated in Chapter 7 is the development of an algorithm for the representation of a single spectral zero by a second order Selective moving average (MA) model. This procedure was introduced through a suitable transformation of the equations describing a standard MA model and then deriving that model which is approprate for the representation of spectral antiresonances. This method, therefore, constitutes a general procedure for the calculation of complex conjugate zero positions from a particular portion of the signal spectrum containing a single spectral zero.

In order to obtain a suitable frequency response for the application of this analysis scheme, the pitch synchronous LP residual spectrum was evaluated since the required antiresonances were found to be clearly in evidence for this form of response. A procedure was then developed for the segmentation of this LP residual spectrum into second order Selective MA regions containing isolated antiresonances and a similar process to that described in Chapter 5 was employed in the evaluation of the zero parameters. A technique was also outlined for the estimation of an appropriate number of antiresonances in the signal spectrum based on a calculation of antiresonance bandwidths in the segmented model.

The zero estimates obtained from the spectral segmentation can then be incorporated into an inverse filter which removes the effect of these antiresonances from the nasalized speech spectrum and thereby yields a characterisation which should consist entirely of spectral poles. Both the linear predictive coefficients resulting from the analysis of this all-pole spectrum and the zero positions together represent a combined pole-zero model. The final analysis stage involved the deployment of this model in a synthesiser configuration to produce an output which was found to represent a better characterisation of the original speech signal with the nature of the spectral antiresonances being well preserved in relation to the synthetic speech derived from a direct LP model of the nasalized vocalisation.

8.2 SUGGESTIONS FOR FURTHER RESEARCH

The research presented in this thesis is constituted of several spectral modelling techniques which are employed in a variety of applications such as speech spectral estimation, formant extraction and synthesis and the pole-zero modelling of nasalized speech. The results obtained in the course of this work demonstrate the viability of these techniques, but there are several areas in which it may be beneficial to either extend the present results or improve the performance of the existing algorithms. One such area is the analysis of a greater variety of speech data ranging from a larger vocabulary of speech sounds to female speech or an

assessment of the various segmentation procedures in the presence of extraneous noise.

For the formant extraction method discussed in Chapter 5, one disadvantage is that the procedure for spectral segmentation into formants is based on a single characteristic of the acoustic waveform — i.e. the speech spectrum derived from a Fourier Transform of the original signal. Although this representation is adequate for many speech sounds, a situation can sometimes occur when two formant frequencies are so close together that the individual peaks can not be resolved [8.8]. Under these circumstances, it may be advantageous to employ an "*off-axis spectrum*" [8.1] — which is a z-transform evaluated on a circle inside the unit circle and, therefore, closer to the poles of the spectral model. This spectrum can then be employed in exactly the same manner as the conventional frequency response in order to evaluate the formant estimates.

It is important to note that the formant analysis/synthesis results presented in Chapter 6 for complete speech utterances did not involve the incorporation of any temporal constraints for the smoothing of the formant tracks. This is in contrast to standard LP analysis for which the estimated formant parameters are often merely regarded as 'raw data' requiring further processing [8.9]. These results are then either processed by eye or through the deployment of decision-making algorithms [8.10] which may employ frame-to-frame comparisons or constraints on formant bandwidths or centre frequencies.

From the narrowband spectrograms of Chapter 6, it is clear that the procedure for spectral segmentation into formants generally produces reasonable formant tracks. However, there are occasional discontinuities which probably arise as a result of an inappropriate choice for the number of regions in the segmented models which were chosen on the basis of the criterion developed in Chapter 5. In these cases, the introduction of some simple temporal constraints for these numbers of segments may produce a smoother set of formant tracks and thereby an improvement in the quality of the synthetic speech signal.

The discussion of section 6.2 revealed that the parallel formant synthesiser configuration does not appear to offer any significant advantages over the cascade equivalent, while resulting in a considerably increased system complexity, for the modelling of non-nasalized vowel or vowel-like sounds. However, if the methods discussed in Chapters 5 and 6 are applied to other vocalisations such as the nasals, fricatives and plosives, the parallel network may be preferred since required spectral features such as zeros can be introduced without resorting to additional filter paths [8.6] — as is necessary for cascade models. It may even be possible to employ the 2nd order Selective LP model gains in the calculation of the appropriate resonator amplitudes required for the parallel configuration.

The excitation signals employed in the synthesiser networks of Chapters 6 and 7 were simply impulse train sequences spaced at the pitch period of the speech signal. Although this form of excitation will produce an acceptable quality of output speech, Holmes [8.11] has suggested that the synthetic output of a formant synthesiser can be greatly improved through the use of input waveforms which more accurately reflect the true nature of the speech excitation source. It is noted that the most natural speech is obtained using a synthetic glottal pulse derived by inverse filtering the waveform corresponding to a typical natural vowel from the same speaker. However, improvements in quality are also achieved from idealized glottal pulse shapes such as a combination of cosine segments [8.12] and, therefore, it may prove fruitful to consider the use of these forms of excitation for the synthesis configurations considered in this thesis.

8.3 REFERENCES

[8.1] MARKEL, J.D., GRAY, A.H. Jr., "Linear Prediction of Speech", *Springer-Verlag, Berlin-Heidelberg-New York*, 1976.

[8.2] RABINER, L.R., SCHAFER, R.W., "Digital Processing of Speech Signals", *Prentice-Hall, Englewood Cliffs, New Jersey*, 1978.

[8.3] COOPER, L., COOPER, M.W., "Introduction to Dynamic Programming", *Pergamon Press, Oxford-New York*, 1981.

[8.4] DREYFUS, S.E., "Dynamic Programming and the Calculus of Variations", *Academic Press*, 1965.

[8.5] PINSON, E.N., "Pitch-Synchronous Time-Domain Estimation of Formant Frequencies and Bandwidths", *J. Acoust. Soc. Amer., Vol* 35, *August* 1963, *pp* 1264-1273.

[8.6] RYE, J.M., HOLMES, J.N., "A Versatile Software Parallel-Formant Speech Synthesiser", *JSRU Research Report No.* 1016, *November* 1982, *Joint Speech Research Unit, Cheltenham, Glos., UK.*

[8.7] MAKHOUL, J., "Linear Prediction: A Tutorial Review", *Proc. IEEE, Vol.* 63, *April* 1975, *pp* 561-580.

[8.8] WITTEN, I.H., "Principles of Computer Speech", *Academic Press, London*, 1982.

[8.9] FALLSIDE, F., WOODS, W.A., "Computer Speech Processing", *Prentice-Hall, Englewood Cliffs, New Jersey*, 1985.

[8.10] MARKEL, J.D., "Automatic Formant and Fundamental Frequency from a Digital Inverse Formulation", *Intl. Conf. Speech Commun. Processing, paper* 89, 1972, *pp* 81-84.

[8.11] HOLMES, J.N., "The Influence of Glottal Waveform on the Naturalness of Speech from a Parallel Formant Synthesiser", *IEEE Trans. Audio and Electroacoustics, Vol. AU-21, No.* 3, *June* 1973, *pp* 298-305.

[8.12] ROSENBERG, A.E., "Effect of Glottal Pulse Shape on the Quality of Natural

Vowels", *J. Acoust. Soc. Amer.*, *Vol* 49, *Feb.* 1971, *pp* 583-590.

APPENDICES

APPENDIX A. THE LEVINSON-DURBIN ALGORITHM

This algorithm proceeds in a recursive manner in that it computes the coefficients $\{\alpha_j^{(i)} : j=1,\ldots,i\}$ of an i-th order LP model from the coefficients $\{\alpha_j^{(i-1)} : j=1,\ldots,i\text{-}1\}$ of the corresponding $(i\text{-}1)$th order model. This can be seen as follows. From section 3.3.1, the equations to be solved in order to obtain the LP model parameters for an $(i\text{-}1)$th order predictor are given by:

$$R(0) \; - \; \sum_{k=1}^{i-1} \alpha_k^{(i-1)} \, R(k) \; = \; E^{(i-1)} \; = \; \left(G^2\right)^{(i-1)} \tag{A.1}$$

$$\& \qquad \sum_{k=1}^{i-1} \alpha_k^{(i-1)} \, R(l\text{-}k) \; = \; R(l) \, , \qquad 1 \leq l \leq i\text{-}1 \tag{A.2}$$

which can also be written as

$$-R(l) \; + \; \sum_{k=1}^{i-1} \alpha_k^{(i-1)} \, R(l\text{-}k) \; = \; 0 \, , \qquad 1 \leq l \leq i\text{-}1 \tag{A.3}$$

These two sets of equations may also be characterised by the following matrix representations:

$$\begin{bmatrix} R(0) & -R(1) & -R(2) & \cdots & -R(i\text{-}1) \end{bmatrix} \begin{bmatrix} 1 \\ \alpha_1^{(i-1)} \\ \alpha_2^{(i-1)} \\ \vdots \\ \alpha_{i-1}^{(i-1)} \end{bmatrix} \; = \; E^{(i-1)} \tag{A.4}$$

$$
\& \quad \left[\begin{array}{c|cccc} -R(1) & R(0) & R(1) & \cdots & R(i\text{-}2) \\ -R(2) & R(1) & R(0) & \cdots & R(i\text{-}3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -R(i\text{-}1) & R(i\text{-}2) & R(i\text{-}1) & \cdots & R(0) \end{array}\right] \left[\begin{array}{c} 1 \\ \alpha_1^{(i-1)} \\ \alpha_2^{(i-1)} \\ \vdots \\ \alpha_{i-1}^{(i-1)} \end{array}\right] = \left[\begin{array}{c} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{array}\right] \tag{A.5}
$$

These two sets of matrices can be combined as follows:

$$
\left[\begin{array}{c|cccc} R(0) & -R(1) & -R(2) & \cdots & -R(i\text{-}1) \\ \hline -R(1) & R(0) & R(1) & \cdots & R(i\text{-}2) \\ -R(2) & R(1) & R(0) & \cdots & R(i\text{-}3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -R(i\text{-}1) & R(i\text{-}2) & R(i\text{-}3) & \cdots & R(0) \end{array}\right] \left[\begin{array}{c} 1 \\ \alpha_1^{(i-1)} \\ \alpha_2^{(i-1)} \\ \vdots \\ \alpha_{i-1}^{(i-1)} \end{array}\right] = \left[\begin{array}{c} E^{(i-1)} \\ 0 \\ 0 \\ \vdots \\ 0 \end{array}\right] \tag{A.6}
$$

The $(i \times i)$ matrix in the LHS of equation (A.6) can be expanded by a row and column in such a manner as to produce the corresponding matrix for an i-th order model, i.e.

$$
\left[\begin{array}{c|cccc|c} R(0) & -R(1) & -R(2) & \cdots & -R(i\text{-}1) & -R(i) \\ \hline -R(1) & R(0) & R(1) & \cdots & R(i\text{-}2) & R(i\text{-}1) \\ -R(2) & R(1) & R(0) & \cdots & R(i\text{-}3) & R(i\text{-}2) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -\mathrm{R(i\text{-}1)} & \mathrm{R(i\text{-}2)} & \mathrm{R(i\text{-}3)} & \cdots & \mathrm{R(0)} & \mathrm{R(1)} \\ \hline -\mathrm{R(i)} & \mathrm{R(i\text{-}1)} & \mathrm{R(i\text{-}2)} & \cdots & \mathrm{R(1)} & \mathrm{R(0)} \end{array}\right] \left[\begin{array}{c} 1 \\ \alpha_1^{(i-1)} \\ \alpha_2^{(i-1)} \\ \vdots \\ \alpha_{i-1}^{(i-1)} \\ 0 \end{array}\right] = \left[\begin{array}{c} E^{(i-1)} \\ 0 \\ 0 \\ \vdots \\ 0 \\ \beta \end{array}\right] \tag{A.7}
$$

where β will be assigned the following value such that equation (A.7) is valid,

$$\beta = -R(i) + \sum_{j=1}^{i-1} \alpha_j^{(i-1)} R(i\text{-}j) \tag{A.8}$$

Due to the nature of the expression in equation (A.7), this result can also be rewritten in the following alternative[1] form,

$$\left[\begin{array}{c|cccccc} R(0) & -R(1) & -R(2) & \cdots & -R(i\text{-}1) & -R(i) \\ \hline -R(1) & R(0) & R(1) & \cdots & R(i\text{-}2) & R(i\text{-}1) \\ -R(2) & R(1) & R(0) & \cdots & R(i\text{-}3) & R(i\text{-}2) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -\mathrm{R(i\text{-}1)} & \mathrm{R(i\text{-}2)} & \mathrm{R(i\text{-}3)} & \cdots & \mathrm{R(0)} & \mathrm{R(1)} \\ -\mathrm{R(i)} & \mathrm{R(i\text{-}1)} & \mathrm{R(i\text{-}2)} & \cdots & \mathrm{R(1)} & \mathrm{R(0)} \end{array}\right] \left[\begin{array}{c} 0 \\ -\alpha_{i-1}^{(i-1)} \\ -\alpha_{i-2}^{(i-1)} \\ \vdots \\ -\alpha_1^{(i-1)} \\ 1 \end{array}\right] = \left[\begin{array}{c} \beta \\ 0 \\ 0 \\ \vdots \\ 0 \\ E^{(i-1)} \end{array}\right] \tag{A.9}$$

In spite of this transformation, it should be noted that the square matrices in both equations (A.9) and (A.7) remain identical. Consider multiplying equation (A.9) by a constant k_i and adding the result to equation (A.7) such that,

---

[1] *This rearrangement is achieved by swapping all (i+1) rows in equation (A.7) such that row m becomes i+2-m and vice versa. All the α coefficients are then reversed in sign and the square matrix is finally adjusted to take account of this reversal.*

$$\left[\begin{array}{c|cccccc} R(0) & -R(1) & -R(2) & \cdots & -R(i\text{-}1) & -R(i) \\ \hline -R(1) & R(0) & R(1) & \cdots & R(i\text{-}2) & R(i\text{-}1) \\ -R(2) & R(1) & R(0) & \cdots & R(i\text{-}3) & R(i\text{-}2) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -R(i\text{-}1) & R(i\text{-}2) & R(i\text{-}3) & \cdots & R(0) & R(1) \\ -R(i) & R(i\text{-}1) & R(i\text{-}2) & \cdots & R(1) & R(0) \end{array}\right]\left[\begin{array}{c} 1 \\ \alpha_1^{(i-1)} - k_i\,\alpha_{i-1}^{(i-1)} \\ \alpha_2^{(i-1)} - k_i\,\alpha_{i-2}^{(i-1)} \\ \vdots \\ \alpha_{i-1}^{(i-1)} - k_i\,\alpha_1^{(i-1)} \\ k_i \end{array}\right] = \left[\begin{array}{c} E^{(i-1)} + k_i.\beta \\ 0 \\ 0 \\ \vdots \\ 0 \\ \beta + k_i.E^{(i-1)} \end{array}\right] \tag{A.10}$$

Now it is easy to see that by choosing an appropriate value for k_i, equation (A.10) will achieve a form for an i-th order predictor identical to that in equation (A.6) for an (i-1)th order model. In fact, by choosing

$$k_i = -\frac{\beta}{E^{(i-1)}} \tag{A.11}$$

equation (A.10) becomes of the correct form since the last element in the RHS is then zero as required. Therefore, the results described by equations (A.10) & (A.11) represent the normal equations for an i-th order predictor. Hence, by comparing this result to equation (A.6), the coefficients $\{\alpha_j^{(i)} : j=1,\ldots,i\}$, $E^{(i)}$ of this i-th order model can finally be computed by updating the coefficients $\{\alpha_j^{(i-1)} : j=1,\ldots,i\text{-}1\}$, $E^{(i-1)}$ as follows:

$$k_i = -\frac{\beta}{E^{(i-1)}} = \frac{\left(R(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)}\, R(i\text{-}j)\right)}{E^{(i-1)}} \tag{A.12a}$$

$$\alpha_i^{(i)} = k_i \tag{A.12b}$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i.\alpha_{i-j}^{(i-1)}, \qquad j = 1,\ldots,i-1 \tag{A.12c}$$

$$E^{(i)} = E^{(i-1)} + k_i \cdot \beta = E^{(i-1)} - k_i \cdot k_i E^{(i-1)}$$

$$= \left(1 - k_i^2\right) E^{(i-1)} \qquad \text{(A.12d)}$$

An initial condition arises from considering equation (A.1) with $i-1=0$ such that

$$E^{(0)} = R(0) \qquad \text{(A.13)}$$

Equations (A.12a)-(A.12d) are, therefore, solved recursively for $i=1,\ldots,p$ and the final result is given by $\{\alpha_j^{(p)} : j=1,\ldots,p\}$ & $E^{(p)}=G^2$.

APPENDIX B. EFFECT OF THE SYSTEM TRANSFER FUNCTION ON THE SPECTRUM OF A SECOND ORDER LINEAR PREDICTIVE MODEL

The second order Selective LP models used in the modelling of speech formants in Chapter 5 were divided into two possible classes. The first type of model is characterised by a system transfer function which consists of a pair of real poles while the second form is made up of two complex conjugate poles. An important requirement for these models is that they produce model spectra which are similar in form to individual speech formants, i.e. they exhibit a resonance peak which matches the formant peak. This appendix examines the conditions under which these two types of model fulfil this requirement of producing a maximum in the model frequency response.

From Chapter 5, the transfer function $H(z)$ of a 2nd order LP model is given by

$$H(z) = \frac{G}{1 - \alpha_1 z^{-1} - \alpha_2 z^{-2}} \tag{B.1}$$

and thus the corresponding magnitude response is such that

$$\begin{aligned} \left|H\left(e^{j\omega}\right)\right| &= \left|\frac{G}{1 - \alpha_1 e^{-j\omega} - \alpha_2 e^{-2j\omega}}\right| \\ &= \frac{G}{\sqrt{(1 - \alpha_1 \cos\omega - \alpha_2 \cos 2\omega)^2 + (\alpha_1 \sin\omega + \alpha_2 \sin 2\omega)^2}} \end{aligned} \tag{B.2}$$

As $f(x) = \frac{1}{\sqrt{x}}$, $x > 0$, is a strictly monotonic decreasing function, a maximum in the $\left|H\left(e^{j\omega}\right)\right|$ response will correspond to a minimum of the function $T(\omega)$ and vice versa, where $T(\omega)$ is defined as:

$$\begin{aligned} T(\omega) &= (1 - \alpha_1 \cos\omega - \alpha_2 \cos 2\omega)^2 + (\alpha_1 \sin\omega + \alpha_2 \sin 2\omega)^2 \\ &= 1 + {\alpha_1}^2 + {\alpha_2}^2 - 2\alpha_1 \cos\omega + 2\alpha_1\alpha_2 \cos\omega - 2\alpha_2 \cos 2\omega \end{aligned} \tag{B.3}$$

Thus, $$\frac{dT}{d\omega} = 2\alpha_1 \sin\omega\,(1-\alpha_2) + 4\alpha_2 \sin 2\omega \tag{B.4}$$

An extremum of the function $T(\omega)$ will, therefore, occur at the frequency $(\hat{\omega})$ such that

$$2\alpha_1 \sin\hat{\omega}\,(1-\alpha_2) + 8\alpha_2 \sin\hat{\omega}\cos\hat{\omega} = 0 \tag{B.5}$$

$$\Longrightarrow \qquad \cos\hat{\omega} = \frac{\alpha_1(\alpha_2-1)}{4\alpha_2} \tag{B.6}$$

This result implies that the following condition must hold for an extremum to exist for $T(\omega)$,

$$\left|\frac{\alpha_1(\alpha_2-1)}{4\alpha_2}\right| < 1 \tag{B.7}$$

Also, from equation (B.4), at the frequency $\hat{\omega}$,

$$\left.\frac{d^2T}{d\omega^2}\right|_{\omega=\hat{\omega}} = 2\alpha_1 \cos\hat{\omega}\,(1-\alpha_2) + 8\alpha_2 \cos 2\hat{\omega}$$

$$= 2\alpha_1 \cos\hat{\omega}\,(1-\alpha_2) + 8\alpha_2\left(2\cos^2\hat{\omega} - 1\right) \tag{B.8}$$

Substituting for $\cos\hat{\omega}$ from equation (B.6),

$$\left.\frac{d^2T}{d\omega^2}\right|_{\omega=\hat{\omega}} = -\frac{\alpha_1{}^2(\alpha_2-1)^2}{2\alpha_2} + 8\alpha_2\left[2\left(\frac{\alpha_1(\alpha_2-1)}{4\alpha_2}\right)^2 - 1\right]$$

$$= 8\alpha_2\left[-\left(\frac{\alpha_1(\alpha_2-1)}{4\alpha_2}\right)^2\right] + 8\alpha_2\left[2\left(\frac{\alpha_1(\alpha_2-1)}{4\alpha_2}\right)^2 - 1\right]$$

$$= 8\alpha_2\left[\left(\frac{\alpha_1(\alpha_2-1)}{4\alpha_2}\right)^2 - 1\right]$$

$$\left.\frac{d^2T}{d\omega^2}\right|_{\omega=\hat{\omega}} = 8\alpha_2\left(cos^2\hat{\omega} - 1\right) \tag{B.9}$$

Now, for $\left|H\left(e^{j\omega}\right)\right|$ to exhibit a maximum at the frequency $\omega=\hat{\omega}$, it is required that $T(\hat{\omega})$ should be a minimum. Therefore,

$$\left.\frac{d^2T}{d\omega^2}\right|_{\omega=\hat{\omega}} > 0 \tag{B.10}$$

From equation (B.9), this situation will only occur when

$$\alpha_2 < 0 \tag{B.11}$$

Hence, equation (B.7) represents the condition for there to be an extremum for the function $\left|H\left(e^{j\omega}\right)\right|$ and equation (B.11) indicates the requirement for this extremum to be a maximum. Now, from equation (B.7),

$$\frac{{\alpha_1}^2\,(\alpha_2-1)^2}{16{\alpha_2}^2} < 1 \tag{B.12}$$

Also, from section 5.2, the 2nd order system function will consist of real poles if

$${\alpha_1}^2 + 4\alpha_2 > 0 \tag{B.13}$$

Thus, from equation (B.12), for a real-pole system,

$$\frac{-4\,\alpha_2\,(\alpha_2-1)^2}{16{\alpha_2}^2} < 1 \tag{B.14}$$

$$\Longrightarrow \qquad \frac{(\alpha_2-1)^2}{-4\,\alpha_2} < 1 \tag{B.15}$$

As $\alpha_2 < 0$,

$$(\alpha_2 - 1)^2 < -4\alpha_2$$

or $$(\alpha_2 + 1)^2 < 0 \qquad \text{(B.16)}$$

Since it is not possible for this result to occur for any value of α_2, this expression provides a *reductio ad absurdum* proof that it is impossible for there to a peak in the magnitude response of a 2nd order LP model if the poles corresponding to the model are both real.

There is no such restriction for the condition when the poles are complex conjugate (i.e. $\alpha_1^2 + 4\alpha_2 < 0$) and there may or may not be a peak in the spectrum. In practice, a maximum will not occur only when both the complex radii are small (i.e. high formant bandwidth) or when the corresponding angles are very close to $\omega = 0$ or $\omega = \pi$. For instance, when the two poles are given by ($z = 0.7\, e^{j\pi/20}$, $z^* = 0.7\, e^{-j\pi/20}$), then

$$\frac{\alpha_1(\alpha_2 - 1)}{4\alpha_2} = 1.05 \qquad \text{(B.17)}$$

and this result then violates the condition expressed by equation (B.7). However, under these circumstances, the predictor parameters can still be related (as shown in Chapter 5) to the formant bandwidth and centre frequency as long as the poles are complex conjugate. In practice the above situation rarely occurs for true speech formants and thus a 2nd order Selective LP spectrum modelling an actual speech formant will usually exhibit a clear resonance peak.