DX208635

M0002890SR
1353666



UNIVERSITY OF SURREY LIBRARY

APPLE 1011

ProQuest Number: 10148201

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.



ProQuest 10148201

Published by ProQuest LLC (2017). Copyright of the Dissertation is held by the Author.

All rights reserved.
This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

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

UniS

University of Surrey

Vocoder Model based Variable Rate Narrowband and Wideband Speech Coding below 9 kbps

by **Milos Stefanovic**

Thesis submitted to the University of Surrey
for the degree of
Doctor of Philosophy



Centre for Communication Systems Research
School of Electronic Engineering, IT and Mathematics
University of Surrey, Guildford, United Kingdom

July 1999

Summary

The past two decades have witnessed rapid growth and development within the telecommunications industry. This has been primarily fuelled by the proliferation of digital mobile communication applications and services which have become commonplace and easily within the financial reach of businesses and the general public. Current research trends, involving integration and packetisation of voice, video and data channels into true multimedia communications, promise a similar technological revolution in the next decade. One of the key design issues of the new high quality multimedia services is a requirement for very high data rates. Whilst the available bandwidth in wire based terrestrial network is a relatively cheap and expandable resource, it becomes inherently limited in satellite or cellular radio systems. In order to accommodate ever growing numbers of subscribers whilst maintaining high quality and low operational costs, it is necessary to maximise spectral efficiency and reduce power consumption. This has given rise to the rapid development of signal compression techniques, which in the speech transmission domain are known as speech coding algorithms.

The research carried out for this thesis has mainly focused on the design and development of low bit rate narrowband and wideband speech coding systems which utilise a variable rate approach in order to improve their perceptual quality and reduce their transmission rates. The algorithms subsequently developed are based on the existing vocoding schemes, whose rigid fixed rate structure is a major limitation to achieving higher quality and lower rates. The variable rate schemes utilise the time-varying characteristics of the speech signal which is classified according to the developed segmentation algorithms. Two main schemes were developed, a variable bit rate with an average as low as 1.35 kbps and a variable frame rate with an average of 2.1 kbps, both achieving or even surpassing the subjective quality of the existing vocoding standard at 4.15 kbps. Wideband speech exhibits characteristics which are not embodied within narrowband speech and which contribute to the superior perceived quality. A very high quality wideband vocoder operating at rates (fixed and variable) below 9 kbps is presented in this thesis, whereby particular attention is paid to preserving the information in higher frequencies in order to maximise the attainable quality.

Acknowledgements

I would like to take this opportunity to express my sincere gratitude to Prof. Ahmet Kondoz, my PhD. project supervisor, whose moral and academic support and guidance throughout my research were most helpful and much appreciated. I am also indebted to my colleagues in the Multimedia Communications group, especially Chana Sriratanaban and Stéphane Villette, whose support and friendship have been invaluable.

Finally, I would like to thank my fiancée Nevin Bayoumi and my family in Belgrade for their constant support, love and encouragement throughout the course of this research. Special gratitude yet again goes to my fiancée for her meticulous proof reading skills.

To all the peoples of former Yugoslavia
who have been systematically
deprived of their material and spiritual well-being
over the years.

Table of Contents

1. INTRODUCTION 5

1.1 BACKGROUND 5
1.2 OUTLINE OF THESIS 8
1.3 ORIGINAL CONTRIBUTIONS 11

2. DIGITAL SPEECH CODING 13

2.1 INTRODUCTION 13
2.2 DESIGN CRITERIA 14
2.2.1 Quality and Coding Capacity 16
2.2.2 Delay 17
2.2.3 Complexity 18
2.2.4 Robustness to Transmission Errors 19
2.2.5 Noise Reproduction and Non-Speech Signal Requirements 20
2.2.6 Multi-rate Operation 21
2.3 APPLICATIONS OF SPEECH TRANSMISSION 21
2.3.1 Terrestrial Systems 21
2.3.1.1 PSTN 22
2.3.1.2 ISDN 23
2.3.1.3 Multimedia 24
2.3.1.4 Land Mobile Telephony 24
2.3.2 Satellite Systems 27
2.4 SPEECH CODING CATEGORIES 29
2.4.1 Waveform Coding 29
2.4.2 Source Coding 30
2.4.3 Hybrid Coding 31
2.5 CONCLUDING REMARKS 33

3. FUNDAMENTALS OF SPEECH CODING TECHNIQUES 34

3.1 INTRODUCTION 34
3.2 GENERAL QUANTISATION PRINCIPLES 35
3.2.1 Scalar Quantisation 35
3.2.2 Vector Quantisation 36
3.3 VOCAL TRACT MODEL 37
3.3.1 Speech Production Mechanism 37
3.3.2 Speech Production Model 38
3.4 LINEAR PREDICTIVE CODING (LPC) 39
3.4.1 Estimation of LPC Parameters 39
3.4.1.1 The Auto-correlation Method (AM) 42
3.4.2 Practical Implementation of LPC Analysis 43
3.4.3 Application of LPC Analysis 46
3.4.4 Quantisation of LP Coefficients 48
3.4.4.1 Line Spectral Frequencies (LSFs) 48

3.4.4.1.1 Definition 48
3.4.4.1.2 Derivation 49
3.4.4.1.3 Properties 49
3.4.4.1.4 Quantisation 51
3.5 PERCEPTUAL WEIGHTING 52
3.6 PITCH PREDICTIVE CODING 53
3.6.1 Pitch Determination in the Time Domain 54
3.6.1.1 Waveform Peak Picking 54
3.6.1.2 Auto-correlation 55
3.6.1.3 Average Magnitude Difference Function 58
3.6.2 Pitch Determination in the Frequency Domain 59
3.6.2.1 Harmonic Peak Picking 59
3.6.2.2 Synthetic Spectral Matching 59
3.6.3 Pre- and Post-processing Techniques 61
3.7 INTERPOLATION 63
3.7.1 General Principles 63
3.7.2 Linear Interpolation Function 66
3.8 CONCLUDING REMARKS 68

4. VOCODING PRINCIPLES AND STANDARDS 70

4.1 INTRODUCTION 70
4.2 CHANNEL VOCODER 71
4.3 FORMANT VOCODER 72
4.4 HOMOMORPHIC (CEPSTRAL) VOCODER 74
4.5 LPC VOCODER 75
4.6 MULTI-BAND EXCITATION VOCODER 77
4.6.1 Introduction 77
4.6.2 I-MBE Encoder 78
4.6.2.1 Pitch Estimation 79
4.6.2.2 Voicing Decision 80
4.6.2.3 Spectral Amplitude Estimation 81
4.6.3 I-MBE Decoder 82
4.6.4 Quantisation Scheme for the MBE Vocoder at 4.15 kbps 83
4.6.5 I-MBE Performance Evaluation 84
4.7 VOCODING STANDARDS 84
4.8 CONCLUDING REMARKS 85

5. SPLIT BAND LPC VOCODING 87

5.1 INTRODUCTION 87
5.2 TIME ENVELOPE SB-LPC VOCODER AT 2.4 KBPS 88
5.2.1 Encoder 88
5.2.1.1 Segmented Auto-correlation Pitch Detection Algorithm (SAPDA) 90
5.2.1.2 Split Band Voicing Technique 92
5.2.1.3 Time Envelope based on RMS Energy 94
5.2.1.4 Determining and Quantising LP Coefficients 95
5.2.2 Decoder 95
5.2.2.1 Voiced Excitation Generation 96
5.2.2.2 Unvoiced Excitation Generation 96
5.2.2.3 Synthesis Filter and Post-filtering 97

5.2.2.4 Energy Scaling Mechanism 98
5.2.3 Bit Allocation Scheme *99*
5.2.4 Performance evaluation *99*
5.3 AMPLITUDE EXCITED SB-LPC VOCODER AT 2.4 KBPS 100
5.3.1 Encoder *101*
5.3.1.1 Sinusoidal Model Matching Pitch Detection Algorithm (SMM-PDA) 102
5.3.1.2 Improved Split Band Voicing Technique 104
5.3.1.3 Harmonic Amplitude Determination and Quantisation 105
5.3.1.4 Determining and Quantising LP Coefficients 105
5.3.2 Decoder *106*
5.3.2.1 Perceptual Enhancement 106
5.3.2.2 Voiced Excitation Generation 107
5.3.2.3 Unvoiced Excitation Generation 108
5.3.2.4 LPC Synthesis Filter 109
5.3.3 Bit Allocation Scheme *109*
5.3.4 Performance Evaluation *110*
5.4 CONCLUSIONS 111

6. VARIABLE RATE QUANTISATION IN SPEECH CODING 112

6.1 INTRODUCTION 112
6.2 SPEECH CATEGORISATION 115
6.3 VARIABLE RATE QUANTISATION STRATEGY 117
6.3.1 Coding of Silences *118*
6.3.2 Coding of Unvoiced Segments *123*
6.3.3 Coding of Mixed Speech *126*
6.3.4 Coding of Fully Voiced Segments *127*
6.4 PERFORMANCE OF THE VARIABLE RATE QUANTISATION 128
6.5 CONCLUSIONS 131

7. VARIABLE FRAME LENGTH SPEECH CODING 133

7.1 INTRODUCTION 133
7.2 SPEECH SEGMENTATION METHODS 137
7.2.1 Phonetically Based Segmentation Method *138*
7.2.2 Multiple Pitch Segmentation Method *140*
7.2.3 Sub-frame Based Segmentation Method *145*
7.3 TIME ENVELOPE SB-LPC VOCODER WITH THE VARIABLE FRAME LENGTH STRATEGY 147
7.3.1 Encoder *148*
7.3.1.1 Pitch Detection for the Variable Frame Length System 149
7.3.1.2 Voicing Decision for the Variable Frame Length System 151
7.3.1.3 LP Modelling for the Variable Frame Length System 152
7.3.2 Decoder *154*
7.3.2.1 Adaptive Interpolation Algorithm 155
7.3.3 Performance Discussion *157*
7.4 AMPLITUDE EXCITED SB-LPC VOCODER WITH THE VARIABLE FRAME LENGTH STRATEGY 159
7.4.1 Encoder *159*
7.4.1.1 Pitch Detection for the Variable Frame Length System 161
7.4.1.2 Voicing Decision for the Variable Frame Length System 163
7.4.1.3 LP Modelling for the Variable Frame Length System 163

7.4.1.4 Determination and Quantisation of Harmonic Amplitudes 163
7.4.1.5 Improved 'Search-and-Compare' Segmentation Algorithm 165
7.4.2 Decoder *169*
7.4.2.1 Perceptual Enhancement 170
7.4.2.2 Improved Adaptive Interpolation Algorithm 170
7.4.2.3 Voiced Excitation Generation 172
7.4.2.4 Unvoiced Excitation Generation 173
7.4.2.5 LPC Synthesis Filter 174
7.4.3 Bit Allocation Scheme *174*
7.4.4 Performance Evaluation *175*
7.5 AMPLITUDE EXCITED SB-LPC VOCODING WITH PHONETICALLY BASED SEGMENT. 176
7.5.1 Segmentation Structure *177*
7.5.2 Performance Evaluation *179*
7.6 CONCLUSIONS 181

8. CODING OF WIDEBAND SPEECH 184

8.1 INTRODUCTION 184
8.2 CURRENT WIDEBAND SPEECH TECHNIQUES AND METHODS 186
8.3 AMPLITUDE EXCITED SB-LPC WIDEBAND VOCODER AT 8.5 KBPS 189
8.3.1 Encoder *189*
8.3.1.1 Pre-processing Wideband Speech 190
8.3.1.2 LPC Filtering for Wideband Speech Analysis 192
8.3.1.3 Wideband Pitch Detection Algorithm 192
8.3.1.4 Wideband Pitch Refinement 192
8.3.1.5 Wideband Voiced/Unvoiced Determination 195
8.3.1.6 Wideband Harmonic Amplitude Determination and Quantisation 196
8.3.1.7 Phase Calculation and Quantisation 196
8.3.2 Decoder *198*
8.3.2.1 Perceptual Enhancement 199
8.3.2.2 Improved Adaptive Interpolation Algorithm for Wideband Synthesis 199
8.3.2.3 Wideband Voiced Excitation Generation 200
8.3.2.4 Wideband Unvoiced Excitation Generation 202
8.3.2.5 Energy Smoothing 202
8.3.2.6 LPC Synthesis Filter 204
8.3.3 Variable Frame Rate in Wideband Speech *205*
8.3.4 Bit Allocation Scheme *205*
8.3.5 Performance Evaluation *206*
8.4 CONCLUSIONS 207

9. CONCLUSIONS 210

9.1 PREAMBLE 210
9.2 CONCLUDING OVERVIEW 211
9.3 FUTURE WORK 216

APPENDIX 218

LIST OF PUBLICATIONS 218

LIST OF ABBREVIATIONS 219

REFERENCES 221

Chapter 1.

1. Introduction

1.1 Background

Speech represents the most prevalent and convenient means of communication between human beings. It is physically realised in the form of sound waves which can not travel more than a few hundred meters before becoming inaudible and unrecognisable. For communication over long distances, therefore, the sound waves are transformed into electrical signals and transmitted either as continuous (analogue) waveforms or as discrete (digital) signals. This forms the basis of telephony which, since its invention by Alexander Graham Bell in 1876 and widespread adoption in the late 19th century, has mainly utilised analogue forms of transmission. The analogue transmission is either in the base-band, or the waveform is used to modulate a carrier which is then transmitted via cable or radio.

Despite many great advances in analogue telephony, it is fundamentally limited by its inability to consistently deliver good quality to the user. Analogue transmission has a disadvantage, namely that the signal is prone to noise at every stage of the transmission,

therefore limiting the maximum communication distance and ultimately the quality of the received signal. In a digital system, the analogue speech signal is sampled, quantised and converted into a binary data stream. This then allows the speech signal to be manipulated by various tools reserved for the digital domain. The signal can now be regenerated at regular intervals as it becomes independent of a transmission distance. It offers improved security for the passing of sensitive information over a transmission link with the use of digital encryption techniques. Error control techniques, such as Forward Error Correction (FEC) [1], can also be applied, making it more robust to channel errors. Furthermore, it can be multiplexed together with other digitised signals such as video, graphics, facsimile and computer data over the Integrated Services Digital Networks (ISDN) [2]. It also requires less transmission power for a given signal quality and it offers the possibility of using complex Digital Signal Processing (DSP) [3] techniques for filtering, echo control, generation and detection of signalling tones and so on. In addition, digital hardware is cheap, reliable, repeatable and almost immune to the major problems encountered in analogue systems, such as noise, component tolerance and thermal stability.

The digital transmission of speech does however have a disadvantage as well. The bandwidth it requires is dictated by the Nyquist sampling rate of 8 kHz for a telephone signal, traditionally band limited to a range of 300 Hz - 3.4 kHz. The first generation digital telephony, introduced in 1947, used 8 bits/sample for quantisation, which meant a data rate of 64000 bits per second (bps) or 64 kilobits per second (kbps) (7 bits/sample in the U.S. and Japan to give 56 kbps). The system in use was a Log Pulse Code Modulation (Log-PCM) [4], an improved version of the PCM invented in 1938 [5]. In the 1970's, development of second generation systems was based on adaptive quantisation techniques (e.g. ADPCM [6]) which further reduced the transmission bit rate to 32 kbps whilst maintaining high quality speech. However, speech transmission at these high rates has only been feasible on trunk telephone links with relatively large available bandwidths, especially with the advent and widespread adoption of fibre-optic technology.

The past two decades have witnessed the emergence of new telecommunication systems, such as satellite communications (fixed and mobile), digital mobile radio, high frequency (HF) secure radio communications for civil and military use, private and cellular networks. In order to accommodate ever growing numbers of subscribers whilst maintaining high quality and

low operational costs, these systems have highlighted the need for better bandwidth utilisation and reduced power consumption. These two requirements have given rise to a rapid development of signal compression techniques, which, in the speech transmission domain, are known as speech coding algorithms.

Nowadays many speech coding algorithms operate at bit rates in the 600 bps to 64 kbps range with speech quality ranging from synthetic (i.e. sounding like a robot) to toll (i.e. sounding like the original voice) and their use primarily depends on the compression needs of a particular application. High bit rates (>16 kbps) provide toll quality speech and they are typically used in PSTNs, ISDN and other high quality network applications. Speech coders operating in the medium bit rate range of 4.8 kbps to 16 kbps are capable of producing speech with quality ranging from communication to toll. They represent the linchpin in the current development in land and satellite based mobile communication systems. However, with the increasing uptake of mobile telephony, an ever faster growing number of subscribers and new multimedia products, there is an increasing demand for even lower rate speech coders, still capable of maintaining high speech quality, comparable to that of the medium range coders. The growing research interest into the development of various low bit rate speech coding algorithms has been fuelled by the enormous advances in the Very Large Scale Integration (VLSI) technologies which has resulted in the availability of increasingly powerful (in terms of computational capability) and low power consumption digital signal processors (DSPs) at ever decreasing costs.

The work presented in this thesis focuses on the further design and development of variable rate coding systems at very low bit rates (3 kbps and below) for narrowband speech and medium-to-low bit rates (around 9 kbps and below) for wideband speech. The narrowband variable rate coding applications include store and forward machines (e.g. answer phones), video conferencing, Internet phone and future low bit rate packet-switched multimedia communication systems. The wideband variable rate coding is particularly suitable for applications in audio video teleconferencing, high quality voice-mail services and high quality wideband telephony.

1.2 Outline of Thesis

The research work can be generally divided into four inter-related areas:

(i) improving and adapting the current state of the very low (fixed) bit rate coder which serves as a basis for integrating the variable rate schemes

(ii) design and development of a variable rate quantisation system and its integration with the fixed rate coding algorithm

(iii) design and development of a variable frame length scheme and its integration with the fixed bit rate coding structure

(iv) design and development of a high quality variable frame length scheme operating for wideband speech based on adapted narrowband speech coding principles.

In (i), efforts were focused on formulating modifications to the low bit rate coding algorithm for further speech quality enhancement, assessing the pros and cons of the variable rate system implementation and preparing the ground for its integration. In (ii), the work mainly consisted of building an efficient and easy-to-implement variable rate quantisation model based on the source characteristics of the processed speech. Once the variable rate quantisation had been established and its quality performance confirmed, in (iii), a more complex segmentation scheme, again source driven, for varying the frame length of the processed speech was designed and implemented. In (iv), the principles of the developed narrowband speech coding algorithm were then used and adapted to facilitate a simulation of the high quality variable frame length wideband speech coder.

Chapter 2 gives a brief overview of the main characteristics of speech coding algorithms. This includes various design criteria which must be considered and which are particularly relevant when developing high quality low bit rate coders. Issues such as algorithmic delay, coding rate, complexity, storage, cost and robustness are covered and their features are assessed. These issues also bear cardinal importance when designing and implementing the variable rate algorithms. Speech coders are then broadly classified into three conventional categories, with each category being briefly discussed and exemplified. Applications, standards and current research trends in speech coding are also briefly reviewed.

Chapter 3 discusses the fundamental techniques and methods used in very low bit rate speech coding. In particular it addresses the general principles of spectral modelling using the well-known Linear Prediction (LP) model. The very low bit rate speech coding compression algorithm, described in the following chapters, is based on the LP model. This chapter covers the physical basis of the LP analysis in the human speech production system, LP coefficient determination algorithms and properties of LP models. For all LP-based coding schemes it is essential that the LP filter coefficients are transmitted in a form which is robust to the effects of the quantisation. Therefore, before quantising them, the LP coefficients are transformed into an alternative set of Linear Spectral Frequencies (LSFs), which are principally exploited as they can guarantee the stability of the filter provided that certain criteria are met. In this chapter the role of pitch detection in speech coding is also presented. It begins by discussing the degree to which Pitch Detection Algorithm (PDA) choice impacts on the overall speech coder performance. Various pitch detection techniques are then presented together with a brief discussion of their relative merits and disadvantages with respect to both performance and complexity. The chapter concludes with a discussion of the use and applications of LP parameter interpolation, with respect to the methodology, expected benefits and algorithmic implications.

Chapter 4 provides a brief history of vocoder development together with a description of different types of vocoding algorithms in use. As the work presented in this thesis is based on vocoding structure, particular attention is paid to the Multi-Band Excitation (MBE) [7] vocoder, one of the most influential vocoding schemes to be developed in the past ten years. The most important rules of vocoding parameterisation and quantisation are also discussed.

Chapter 5 describes in detail the characteristics of the fixed bit rate vocoding algorithm which served as the starting point of this research. The improvements made to this MBE-LPC based vocoding algorithm, especially the introduction of the split-band voiced/unvoiced classification method, are discussed and analysed. The improved vocoding algorithm forms the basis for design and development of the variable rate systems. A two track approach is taken. The first version of this algorithm is a Time Envelope Split-Band LPC vocoder operating at a fixed bit rate of 2.4 kbps for which a variable quantisation rate scheme was designed and simulated. This vocoder is also used for introducing and integrating a more complex variable rate scheme, called variable frame length scheme or variable frame rate

scheme. However, before this new scheme is introduced and analysed, the quantisation structure for the time envelope parameters of the vocoder is carefully studied. During research it proved to be of high complexity, bearing a heavy computational load when designed to accommodate the variable frame length strategy. Therefore, the time envelope approach is substituted with the spectral amplitude quantisation method producing a 2.4 kbps Amplitude Excited Split-Band LPC vocoder. This new high quality vocoding structure is shown to be very well suited to the particular needs of the variable frame length scheme.

Chapter 6 introduces the concept of variable bit rate coding based on a speech categorisation method, used to characterise different regions which make up a speech signal. The concept utilises the fact that modelling parameters for different speech segments do not necessarily require the same number of bits to encode. The objective is to produce a very low bit rate coder with an average bit rate lower than the fixed bit rate. However, this should be achieved in such a way that the output quality of the new variable rate system is preferably better or at least matches that of the equivalent fixed rate system. The speech categorisation method is source driven, relying on speech coding techniques, such as pitch prediction and voiced/unvoiced classification for accurate speech segment characterisation. It forms the basis of the variable rate quantisation scheme which was developed for the Time Envelope Split-Band LPC Vocoder operating at a fixed bit rate of 2.4 kbps. The performance of the variable rate strategy is then compared to an equivalent fixed rate structure described in Chapter 5.

Chapter 7 shows and analyses the adaptations made to the vocoding structure of the Split-Band LPC vocoder in order to facilitate integration of the variable frame length scheme. The idea behind this scheme is to make better use of speech continuity by observing the short-term and long-term changes in a speech waveform. By varying the frame length, fast speech transitions can be modelled by shorter frame sizes, whereas slowly evolving steady state speech can be represented by longer frame sizes. The expected overall effect, when compared to the fixed frame length scheme, is the potential reduction in the overall number of both voiced and unvoiced frames and higher quality when shaping the dynamically changing speech transients. This scheme introduces a complex strategy of varying the length of a speech segment according to its characteristics, such as pitch period and voicing decision, as it is also source driven. The most important part of this scheme is the encoder's decision on the speech segment size, which is then transmitted to the decoder with all the other speech

parameters. In order to assess the accuracy of the encoder's decision, a phoneme-based speech recognition module is integrated into the variable frame length scheme as a method of testing the speech segmentation process. The variable frame length scheme is developed for both the Time Envelope Split-Band LPC Vocoder operating at a fixed rate of 2.4 kbps and the Split-Band LPC Vocoder based on spectral amplitude quantisation operating at a fixed rate of 2.6 kbps. The performance of the variable frame length strategy is then compared to an equivalent fixed rate structure which is described in Chapter 5.

Chapter 8 starts with a detailed look at the characteristics of wideband speech. In particular it highlights features which ought to be represented in a coding scheme in order that the quality or fidelity of the synthetic speech is maintained. The Split-Band LPC vocoder based on the spectral amplitude quantisation is then examined for its suitability to code wideband speech. The narrowband algorithmic structure of the vocoder is modified in order to achieve and maintain the quality of the wideband speech. The performance of the wideband Split-Band LPC vocoder based on spectral amplitude quantisation and operating at around 8.5 kbps is then compared to the current state of the art coder in wideband coding standards, the ITU G.722 coder.

Chapter 9 is a summary of the results presented in the preceding chapters. It consolidates the conclusions made in the previous chapters, underlining the most significant work and achievements and their contribution towards the field of modern telecommunication systems. In addition, the chapter discusses possible future work which may further improve both the synthetic speech quality of the variable rate systems and the efficiency of their algorithmic structure.

1.3 Original Contributions

The original achievements contended in this thesis can be summarised as follows:

- A comprehensive review of the linear prediction method, including its physical basis, calculation, quantisation and its suitability in deriving vocoding algorithms and structures.

• A study of the existing very low bit rate MBE-LPC vocoding algorithm and its further improvement by introducing the split-band voiced/unvoiced classification.

• Development of a new variable quantisation scheme and its integration with the Split-Band LPC vocoding algorithm.

• Further improvement to the Split-Band LPC vocoder by way of replacing the time envelope quantisation structure with the spectral amplitude quantisation model.

• Design and development of the complex variable frame length scheme and its integration into both the Time Envelope Split-Band LPC vocoder and the Amplitude Excited Split-Band LPC vocoder.

• Integrating the phoneme-based speech recognition module with the variable frame length structure in order to test and further enhance the accuracy of the in-built segmentation algorithm.

• A comprehensive study of wideband speech coding, as opposed to the narrowband speech coding, and its influence on the perceptual speech quality and fidelity.

• Successful transformation and adaptation of the Split-Band LPC vocoder from the 2.75 kbps narrowband to around 8.5 kbps wideband coding domain.

Some of the above work has been published in various papers, a full list is included in Appendix A.

Chapter 2.

2. Digital Speech Coding

2.1 Introduction

Digital speech coding may be defined as a technique of converting an analogue speech signal into a digital form and then either transmitting it over a digital communication path or storing it on a digital storage medium. Once converted into a digital domain, the speech signal can be manipulated and controlled in various ways by means of sophisticated digital speech coding techniques and methods, such as error protection, encryption, multiplexing, packetisation and mixing with other digitised signals (video, facsimile, data, etc.).

The idea of transmitting digitally coded analogue signals, such as speech, was born in 1938 when Alec H. Reeves invented pulse code modulation (PCM). The first PCM telephone system was built at the Bell Telephone Laboratories around 1947 [8][5], but it was not until more than a decade later that the new system could be fully exploited. The reason for this delay lay with the fact that the necessary supporting hardware, i.e. digital integrated circuit technology, was in its early stages of development and therefore somewhat limited at the

time. However, when a 64 kbps PCM digital speech coding system [9] was accepted as an international standard by the International Telegraph and Telephone Consultative Committee (CCITT) in the early 1970's, it became a driving force behind the development of digital speech coding, its future widespread adoption and low-cost use. Today, almost all of the Public Switched Telephone Networks (PSTN) are based upon PCM, much of it using fibre optic technology which is particularly suited to the transmission of digital data.

Over the years, PCM has clearly demonstrated all the advantages of digital speech coding over its analogue counterpart and predecessor. The availability of various sophisticated tools and techniques for manipulating the digitised signal, such as error correction, encryption and signal compression, coupled with far simpler hardware requirements are just a few of them. However, a disadvantage with PCM coding schemes is the amount of channel capacity they require for transmitting binary data. This problem becomes a major obstacle when communicating over satellite, radio or transoceanic cable links where the available data bandwidth is limited and at a premium cost. Due to the rapid growth in the user demand for these transmission media over the past two decades, a considerable research effort has been directed into producing more efficient methods of bandwidth utilisation in these systems. One of the main methods in solving this problem is a compression of source data, i.e. the speech signal, by designing and developing algorithms capable of compressing the digital speech information by exploiting natural redundancies contained within human speech. Yet again, the early efforts in this research domain were severely restricted by the limited technology available at the time. However, recent breakthroughs in development of the Very Large Scale Integration (VLSI) circuit technology have produced a new set of powerful Digital Signal Processing (DSP) tools resulting in rapid advancements in the domain of speech compression. This, in turn, has opened up the speech compression technology to market acceptance and widespread use.

2.2 Design Criteria

A digital speech communication system, used to transmit and receive digital speech information, generally comprises three components, namely, a speech coder, a channel coder

and a modem. Each component has its subsystem on both the transmitting and receiving ends of the communication system. On the transmitter side of the system, there is a speech encoder, a channel encoder and a modulator. On the receiver side, there is a demodulator, a channel decoder and a speech decoder. The block diagram of this system is depicted in Figure 2.1. A speech encoder compresses the digital input signal for better bandwidth utilisation whilst preserving the speech quality. A channel encoder adds some redundancy in order to protect the bit stream against transmission channel errors. A modulator seeks to maximise the bit rate that can be supported in a given channel without causing an unacceptable level of bit error probability. On the receiver's end of the system, the order and the task of the components are reversed with the purpose of reproducing a faithful replica of the original input signal.

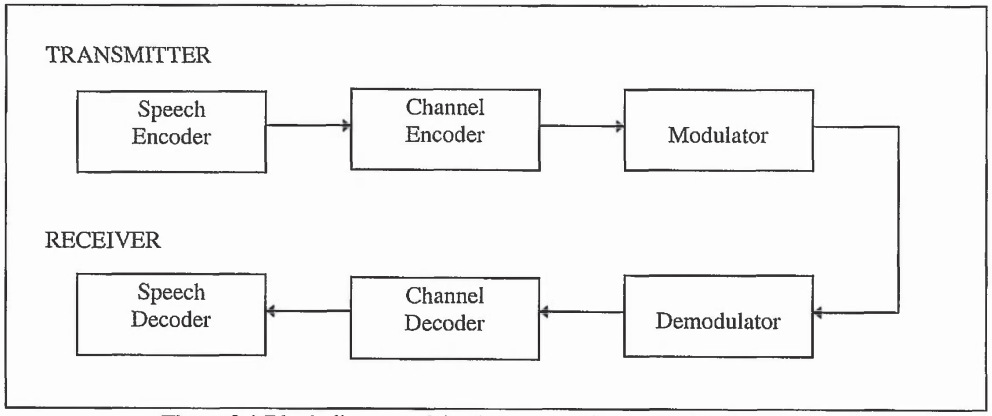


Figure 2.1 Block diagram of the digital speech communication system

The speech coder can be considered to play a vital role in the digital communication system as the final output speech quality is largely determined by its performance. Design and development of the digital speech coder are often driven by its target application and entail numerous factors that attempt to keep in balance often conflicting objectives. Usually the improvement in one aspect of the coder's performance results in the degradation of another. The key factors that need to be carefully assessed, according to their relative importance in the design of the coder are briefly discussed in the following subsections.

2.2.1 Quality and Coding Capacity

Output speech quality and coding capacity are considered to be the two most important design issues when choosing the most suitable speech coding algorithm for any particular target application. However, they already represent two conflicting factors in speech coding, as it is now widely agreed that the higher the transmission or output bit rate for a given coding algorithm, the better the output speech quality. Therefore, one cannot expect to majorly improve the speech quality while reducing the bit rate or, in other words, one cannot significantly reduce the bit rate and still expect to maintain the high speech quality. It has been observed that each individual type of coding algorithm is particularly suited to a certain bit rate range, below which its quality rapidly deteriorates and above which it quickly achieves a point of diminishing returns, i.e. where any further increase in bit rate leads to insignificant improvement in quality.

One of the problems that designers are faced with, when assessing the performance of a given coder, is the choice of a quality measure. In higher rate coders, objective distortion measures, such as Segmental Signal to Noise Ratios (SNR), may represent the correct choice to evaluate the output quality. For example, waveform coders, such as PCM, use the SNR measure to get an indication of quality on a sample by sample basis. However, speech coders with lower bit rates employ a parametric speech modelling of the human vocal tract and, hence, their coded waveform often bears no resemblance to the original. Instead, alternative objective measures are used, such as Dynamic Rhyme Test (DRT) [10] and Diagnostic Acceptability Measure (DAM) [11]. DRT requires listeners, or test subjects, to distinguish between two words differing by only one phoneme or consonant. The DRT score is defined as the percentage of correct responses. DAM provides listeners with an opportunity to indicate directly not only how acceptable a speech sample file is, but also to independently indicate to what extent various perceived qualities are present in the sample without regard to how they may affect their evaluation of acceptability. A listener makes a total of 21 ratings during the course of a speech sample file; ten ratings are concerned with perceptual qualities of the signal, eight ratings are concerned with the perceptual qualities of the background, and three ratings are concerned with perceived intelligibility, pleasantness, and overall acceptability. The DAM ratings are then combined to make an overall result report.

In any speech algorithm, a subjective test, i.e. the way listeners perceive the output quality of a given coder, is the most valuable quality assessment measure. This is particularly pertinent to low bit rates where a subjective test measure, such as Mean Opinion Score (MOS) [12], is often more revealing and useful than the objective measures. A MOS scale is depicted in Table 2.1. The MOS figure for a given algorithm is determined by averaging the scores marked by a large number of test subjects for individual speech material. Subjectively, a toll quality (MOS>4) speech coder produces output speech as intelligible as the original and free of distortions. Communication quality, where 3<MOS<4, means that some distortion is present but not obvious and the output speech has very high intelligibility, whereas synthetic quality, where MOS<3, is associated with good intelligibility but less speaker identifiability.

Table 2.1 MOS scale

MOS	Subjective Opinion	Speech Quality
5	Excellent	Broadcast
4	Good	Toll
3	Fair	Communication
2	Bad	Synthetic
1	Poor	-

The quality of a given coder can also be affected by a sampling rate used to digitise the speech signal. As expected, the higher the sampling rate, the higher the output quality of the coder. The higher sampling rate means that a greater portion of spectral information is included in the signal analysis. This fact has lead to widespread use of so-called wideband coders, whose analysis bandwidth has been extended from a conventional 4 kHz (narrowband) to 7 kHz (wideband).

2.2.2 Delay

The end-to-end delay in communication systems is yet another conflicting issue which faces a speech coding designer. Every type of speech coder incurs a certain delay due to its buffering and processing requirements. The length of the input and output speech buffer largely depends on the prediction algorithm of the coder. In general, lower bit rate coders require

greater prediction mechanisms, which in turn entails longer buffering of the input speech signal. In addition, the time taken for an encoder and a decoder to analyse and synthesise the signal respectively, incurs a certain processing delay. There are also other factors that can further extend the delay in a communication system, such as transmission and channel coding.

The problem with delay occurs when there is an impedance mismatch between the switching equipment and telephone hybrid circuits. When the impedance mismatch is combined with the excessive delay factor, it produces an echo, which in turn can be subjectively disturbing to both a talker and a listener. Studies [13] have shown that an acceptable relative signal level of echo is a function of an incurred delay, i.e. an increase in delay causes a decrease in the tolerable echo energy level. Systems in which the end-to-end delay approaches the maximum recommended by regulatory bodies have to employ sophisticated echo cancellation or echo suppression hardware which substantially increases the overall system cost. This equipment has already become an integral part of mobile and satellite communication systems characterised by excessive propagation delays. However, the transmission delay for land based systems is typically very low [14] and it does not require expensive echo cancellation or echo suppression equipment. The introduction of any speech compression algorithm into a land based system can therefore significantly increase the end-to-end delay due to its buffering and processing requirements. The buffering delay can be minimised by shortening the analysis windows or by employing backward prediction method. However, both these methods require an increase in frame update rate which often leads to higher bit rates. The processing delay can be minimised by employing sophisticated DSP hardware which can lower the complexity of a coder.

2.2.3 Complexity

The complexity of the speech coding algorithm is related to a number of computations carried out during the encoding and decoding processes. In order to be able to perform complex algorithms, a given speech coder requires sophisticated and, therefore, usually expensive DSP hardware. Although recent years have witnessed a substantial progress in DSP development, which has resulted in yet cheaper and faster DSP hardware, they have also seen a rise in

complexity of the new generation of speech coding algorithms. One of the successful approaches aiming at reducing the cost of hardware is a conversion from a floating point processor to an equivalent fixed point processor. However, the price to pay for converting a speech coding algorithm to operate in fixed point arithmetic is the addition of extra design stages which cause further increase in development costs.

The complexity issue becomes conflicting when it is faced with problems of power consumption and size. The power consumption for a DSP device fabricated using CMOS technology is almost linearly related to the operating clock frequency and the number of transistors used. Consequently, more sophisticated algorithms consume more power and can cause an increase in hardware size. Therefore, the choice of a speech coding algorithm for a given application is not only influenced by the required output quality but also by the complexity factor. A successful trade-off between the complexity and the quality is of paramount importance for mass market manufacturers, such as digital mobile handset producers.

2.2.4 Robustness to Transmission Errors

Transmission channels in most speech communication applications frequently suffer from degradation which can greatly influence the quality of transmission by introducing errors into the bit stream. As most modern day speech coders are based on the speech production model rather than the use of waveform coding, the result of an error in the bit stream can have varying effects depending on which transmitted model parameter is affected. In order to combat this problem, one has to rely on the inherent robustness of a given coder and, if necessary, on a certain degree of Forward Error Correction (FEC). For example, an inherently robust coder employed in a PSTN network can probably tolerate errors in its expected Bit Error Rate (BER) range of 10^{-6} to 10^{-5}. However, in mobile and satellite networks, where a transmission link is frequently subjected to severe fading conditions and where an expected BER can lie anywhere in the region of 1% to 5%, it is simply not enough to design inherently robust algorithms, and extensive use of FEC techniques may be vital.

However, the large use of the FEC also introduces a high degree of redundancy in the bit stream resulting in the inefficient use of the transmission bandwidth. For instance, a 13 kbps Full-Rate GSM coder has an overall rate of 22.8 kbps [15], the excess capacity being allocated for channel coding. Any reduction in FEC redundancy will lead to a high residual BER which can only be combated by a greater use of the coder's built-in error control measures, such as error detection and concealment, zero redundancy techniques and lost frame reconstruction techniques. This brings us back to the point of highlighting the importance of the inherent robustness of the coding algorithm. It is a serious issue when designing a speech coder and must be taken into consideration from the very start of a design process.

2.2.5 Noise Reproduction and Non-Speech Signal Requirements

The high output quality of a given speech coder depends not only on its ability to code human speech signals but also to faithfully reproduce the acoustic background noise present in transmission systems. As well as faithfully reproducing the acoustic background noise, which is essential for maintaining the naturalness of the speech during conversation, the coder is often faced with a task of suppressing the background noise capable of severely contaminating the speech itself. Voice Activity Detection (VAD) [16] is a method often employed to combat the latter problem.

As a general rule, the lower the coder bit rate, the more specialised towards modelling human speech the coder becomes, and as a result, it has greater difficulty in reproducing non-human sounds. However, the ability to carry non-speech signals for speech communication channels can be essential. For example, a PSTN network is used to sending signalling tones based on a Dual Tone Multi-Frequency (DTMF) system for the transmission of telephone digits and modem tones for transmitting voice-band data. Severe distortions in coding these tones can significantly degrade the performance of PSTN.

In addition, the speech coder must be able to adjust its dynamic input range in order to avoid any amplification or clamping distortions caused by the variable level of an input signal. The minimum value of a dynamic input range usually recommended lies in the order of 25 dB.

2.2.6 Multi-rate Operation

The capability to seamlessly switch from one operational bit rate to another is one of the strong points of the new generation of speech coding algorithms [17]. This is particularly desirable in radio transmission links which often suffer from shadowing and multi-path fading effects [18]. These effects can significantly reduce the signal strength which in turn results in a high increase in BER. The multi-rate operation can help to combat this problem by reducing the operational bit rate, hence, increasing the number of available bits for the error protection mechanisms. This way the speech quality may degrade gradually depending on the channel conditions, but it is unlikely to fall to an unacceptable level.

Lowering the bit rate of a multi-rate coder can also have other positive implications on channel performance. It can make the channel more efficient by accommodating higher traffic during peak traffic hours or by multiplexing speech and data services over the same channel.

2.3 Applications of Speech Transmission

Digital speech coding plays a vital role in various communication system applications. Some of these applications have been standardised and operational for years, and are now entering into a second or third generation, employing more sophisticated bandwidth reduction methods and speech coding algorithms. Others have only recently taken off with completed standardisation proceedings. There are various ways of classifying the communication systems. The approach adopted in this thesis is to divide them into two broad categories, terrestrial and satellite.

2.3.1 Terrestrial Systems

Terrestrial systems are digital speech transmission systems that generally do not employ satellite channels. The systems which are briefly reviewed in this thesis are Public Switched Telephone Networks (PSTNs), Integrated Services Digital Networks (ISDNs), multimedia, and land mobile telephony.

2.3.1.1 PSTN

Public switched telephone networks are regulated by international bodies, such as the International Telecommunication Union (ITU). The system requirements are therefore very strict, as the main task of the network is to produce very high quality speech, ranging from toll to broadcast. Other important requirements include coding delay less than 10 ms, robustness at random channel errors up to 10^{-3}, and ability to handle signalling tones and voice-band data.

In order to meet such high quality speech transmission requirements, PSTN operators world-wide tend to employ speech coders with very high bit rates. In 1972, a 64 kbps PCM (A- and μ-law versions) coding algorithm [9] was adopted by CCITT, the ITU predecessor, as an international standard for transmitting digital speech over PSTN networks. In order to accommodate a substantial increase in subscriber numbers over the years and hence increase revenue, a 64 kbps PCM was in 1985 replaced by a more efficient 32 kbps ADPCM coding standard [6]. Both of these coders are capable of producing toll quality speech. Being easy to implement and having less than a millisecond coding delay are also the features that have made them so popular among the international telecommunication community. However, the breathtaking pace of the speech coding research coupled with widespread adoption of digital standards has lead in recent years to the recommendation of a third generation of digital speech coding standards for PSTN use. A 16 kbps Low-Delay Code Excitation Linear Prediction (LD-CELP) coder [19], standardised in 1991, has much higher complexity than its predecessors and it is, therefore, the most expensive among the three. However, LD-CELP is also capable of producing toll quality speech and has a very low delay of around 2 ms. Since 1991, while the LD-CELP standard has slowly been adopted, especially by PSTNs with very high subscriber demand, speech coding researchers around the world have already started looking in the area of 8 kbps where they are hoping to design the new global PSTN standard of the future.

The biggest challenge for the modern day PSTN is the digitisation of dense local loops as opposed to the trunk links with high bandwidth and optical fibres. A dense local loop is characterised by a limited bandwidth of the twisted pair wires. The most efficient use of this limited bandwidth can be made by a sub-4.8 kbps speech coder which is still waiting to be

standardised. The other possible solution, to replace the narrowband local loop with broader bandwidth media in order to accommodate higher rate coders, is still prohibitively expensive.

2.3.1.2 ISDN

The first generation of ISDN links appeared in the late 1980's. The widespread research interest in this area has been further spurred by the growing need for transmitting speech integrated with data. The channel capacity of the first generation of ISDN was only 64 kbps whereas the second generation, introduced in the early 1990's, was capable of providing 144 kbps capacity, enough to accommodate two 64 kbps data, voice or video channels and a 16 kbps channel for call control [20].

Although the use of ISDN has seemed to be growing over the past few years, any substantial breakthrough in its further development has been slowed down by two major obstacles. The first one is the absence of a common world-wide ISDN standard, similar to those accepted as ITU standards for PSTN use. The second one is the high cost involved in a nation-wide provision of ISDN network. However, some market analysts believe that once a common standard is agreed among the main producers of ISDN (Europe, USA and Japan), it will quickly release its market potential on a global scale.

The most important speech requirement of the ISDN link is the toll quality of speech and the robustness at random channel errors up to 10^{-4}. In 1984, a 64 kbps Sub-Band ADPCM (SB-ADPCM) coder [21] was adopted by CCITT as a standard for teleconferencing on ISDN. This algorithm fully met the requirements of the ISDN link. In this algorithm, the input speech is band limited to 7 kHz wideband instead of the conventional 3.4 kHz narrowband. The sub-band coding is applied to split the initial bandwidth into two separate sub-bands. Each sub-band is then individually encoded using 32 kbps ADPCM. However, recent research has also proposed a 7 kHz bandwidth 32 kbps speech coder based on transform coding for a new generation ISDN standard [22].

2.3.1.3 Multimedia

The term multimedia generally stands for a computer-based programming platform capable of supporting at least three of the following seven media types: text, audio, data, graphics, still pictures, animation and moving pictures [23]. Human voice, music and sound effects are all grouped under the audio category. The ultimate aim of the multimedia provider is to offer very high quality services transmitted across the ISDN type network. Therefore, the main speech coding requirements of the multimedia services will be similar to those currently supported by the ISDN network. The only exception is the additional use of wideband (0-7 kHz) audio signals. Currently, the multimedia products present in the consumer electronics market are making use of a 64 kbps narrowband PCM (G.711), and 64 kbps wideband SB-ADPCM (G.722). However, further research has opted for lower rate algorithms, such as 16 kbps narrowband ADPCM (G.726) [24], in order to make the products more bandwidth efficient.

2.3.1.4 Land Mobile Telephony

The simplest definition of land mobile communication systems says that they incorporate networks, excluding the use of satellites, over which subscribers can transmit and/or receive messages, at or from an arbitrary location within a coverage area, while either on the move or stationary. Typical users of land mobile systems are haulage and dispatch companies; emergency services like police, fire and ambulance; any other professional who needs to be reached and to reach others at any time, anywhere; and all those people whose lifestyle cannot be contented with services provided by the fixed network [25]. While the development of land mobile systems has progressed at an unprecedented pace over the past few years, numerous new services have appeared on the market, including paging, dispatching and voice mail. However, the most commonly used service has always been land mobile telephony. Land mobile telephony systems could broadly be classified into the following three categories: cellular mobile radio systems, cordless radio mobile systems and personal communication networks (PCNs).

The first generation of the cellular mobile radio systems, employing analogue speech transmission, was gradually introduced two decades ago. However, large and expanding demand led to the development of second generation digitally based systems that could offer secure, interference free, cost effective and feature-rich services. An example of a typical cellular mobile radio system is shown in Figure 2.2, which also includes an interface between the PSTN (or ISDN) network and the cellular mobile radio system.

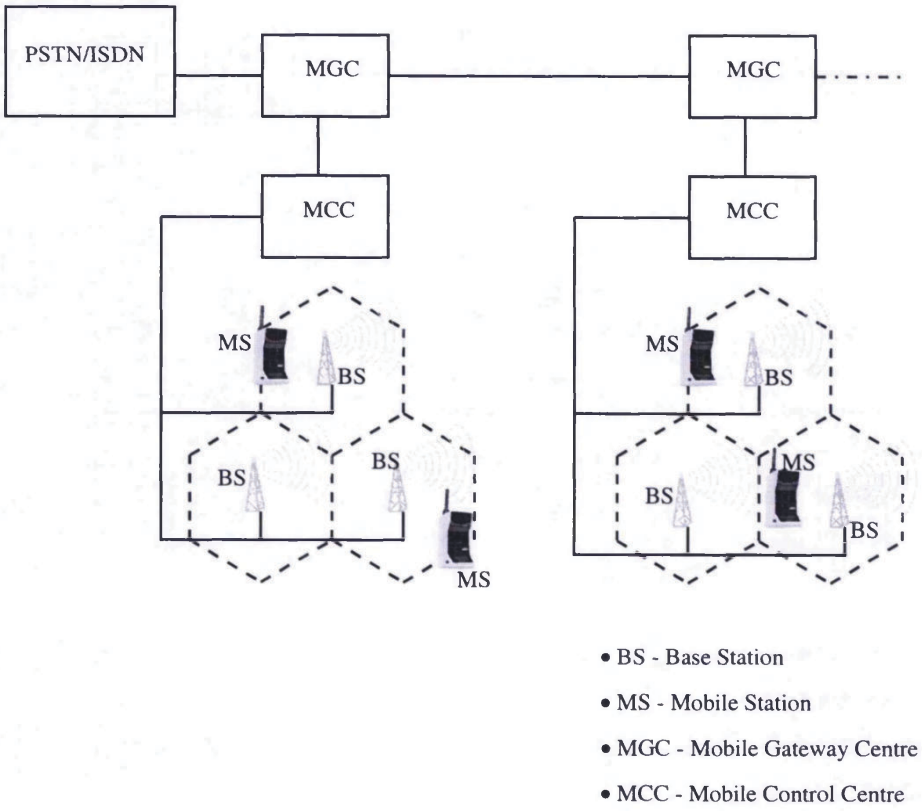


Figure 2.2 Typical cellular land mobile radio system

The speech coding requirements of the cellular mobile radio system are: coding delays less than 70 ms, robustness at random channel errors up to 10^{-2}, communication quality and ability to pass signalling tones [26]. The most serious problem that a cellular mobile transmission channel suffers from is severe bursty errors due to the co-channel interference and multi-path fading effects. Therefore, the coder must display robustness at high rate bursty channel errors (up to 3%) and in mobile environment background noise. The practical requirements also

include the need for low computational complexity which, in turn, decreases the power consumption and implementation costs.

Throughout the last decade several digital speech coding algorithms have been standardised for cellular mobile radio applications, the most prominent being the Pan-European Digital Cellular Mobile Radio system (GSM) introduced in 1988. The Full Rate GSM (FR-GSM) operates at a gross bit rate of 22.8 kbps which accrues from 13 kbps for speech coding based on the Regular Pulse Excited LPC (RPELPC) with long term prediction (RPE-LTP) and 9.8 kbps for channel coding based on a half-rate convolutional coder used to protect the most sensitive bits [27]. FR-GSM produces communication quality speech, has a 25-30 ms delay and is robust to channel errors and mobile background noise. FR-GSM has already been employed in more than 50 countries world-wide. This huge market demand has accelerated the development and standardisation of a Half-Rate GSM (HR-GSM) [28] system operating at a gross bit rate of 11.4 kbps and based on a 5.6 kbps Vector Sum Excitation Linear Prediction (VSELP) coder [29].

Cordless mobile radio systems can be regarded as ordinary telephone receivers with the wire between the handset and the rest of the unit replaced by a radio link. The handset only works with its base station. During the last decade, the majority of the analogue terminals from the first generation have been substituted by second generation digital terminals. The speech coding requirements for digital cordless telephony are more or less the same as those for digital cellular mobile radio systems. High bit rate speech coders, such as 32 kbps ADPCM (G.721) standard, generally meet all these requirements.

Personal Communication Networks (PCN) are based on the desire to have communication terminals for people instead of places or locations. The basic idea behind the first generation of PCNs, currently operated in UK by Orange and Mercury One2One [30], is that they use much smaller cells than those adopted in mobile networks, with lower powered hand-held transmitters, smaller and cheaper handsets, and much higher user capacity. A lot of on-going research has been devoted to the realisation of PCN, however, only two imperfect solutions exist at the moment, based on either cellular radio systems or cordless telephone systems. The ideal PCN of the future will probably operate like a single wireless ISDN-type network able to accommodate other digital services. The 32 kbps ADPCM (G.721) standard is again a

notable proponent for a successful speech compression algorithm in this area. However, various constraints that will be placed on the design and functioning of PCN will probably find any reduction in bit rate highly desirable.

2.3.2 Satellite Systems

Satellite systems are used for a range of applications, such as communications, meteorology and scientific research. The communication satellites, which are of prime concern here, operate as active repeaters in space, usually in the 24-hour geostationary orbit with an altitude of 35786 km. This orbit imposes minimum tracking operations requirements on the earth stations [31]. The use of satellite systems is primarily for long distance communication as they have a wide coverage area and point-to-point and point-to-multi-point connection capability. The three main types of satellite services defined by the ITU are: Fixed Satellite Service (FSS), Mobile Satellite Service (MSS) and Broadcast Satellite Service (BSS). FSS provides services, such as television-relay, telephony and data communication, to fixed earth stations. MSS, on the other hand, provides services, such as maritime, aeronautical or land, to both fixed and mobile earth terminals. BSS covers the domain of Direct Broadcast Satellites (DBS) which mainly provide television and radio programmes broadcasting services to smaller earth stations on domestic premises.

The key problems that speech coders for satellite services are confronted with are related to the space segment signal propagation. Similar to the land mobile communication services, a given speech coder is faced with a limited bandwidth and bursty channel degradation due to fading from multi-path effects and shadowing. Moreover, there is an inherent and substantial delay in the system when compared to the land mobile system use. Therefore, at first, in order to maximise bandwidth efficiency and thus minimise satellite transmission power, the coder must operate at low bit rates. Secondly, it must employ reasonable error control techniques to reduce channel degradation effects. And thirdly, the coder must exhibit limited speech and channel coding delays.

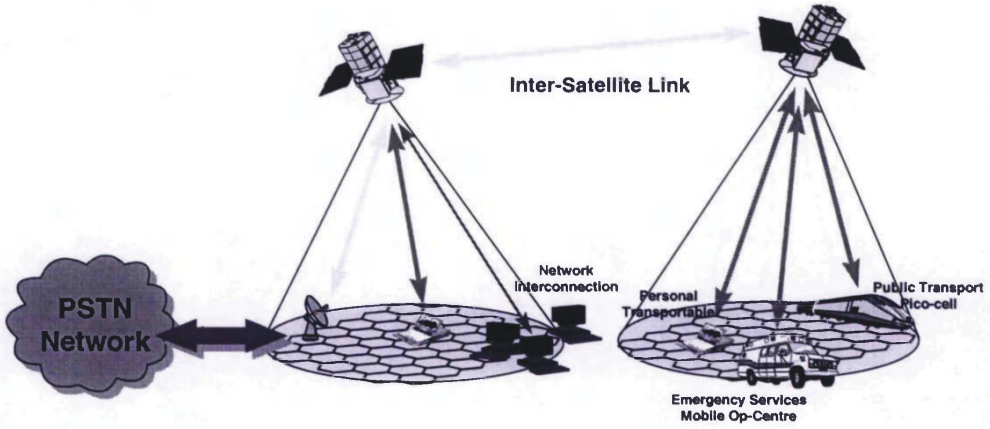


Figure 2.3 Diagram of a typical Satellite Services network

The most prominent example of a coder used for providing MSS, as depicted in Figure 2.3, in the field of satellite-based land and maritime mobile communications is a Multi-Band Excitation (MBE) vocoder [32] operating at a gross rate of 6.4 kbps, selected in 1990 by the International Maritime Satellite (INMARSAT) organisation, producing high quality speech and being very robust to channel errors. However, the main disadvantage of this system has been its complexity. In the field of FSS and BSS services, where it is aimed at the broadcasting quality of communication, satellites from the International Telecommunication Satellite organisation (INTELSAT), for example, provide services at 64 kbps and 32 kbps, although with the advent of 16 kbps LD-CELP, further reduction in the transmission rate is expected.

2.4 Speech Coding Categories

Speech coding schemes can be broadly classified into three main categories which also represent the most widely researched areas in digital speech coding:

- Waveform coding
- Source coding
- Hybrid coding

The general function of these coding schemes is signal analysis, removal of redundancies and efficient coding of the non-redundant parts of speech in a perceptually acceptable manner.

2.4.1 Waveform Coding

Waveform coders, as the name implies, attempt to code the signal waveform of speech, i.e. to quantise the shape of the waveform on essentially a sample by sample basis. Waveform coding is not necessarily speech specific, as it can be used for coding of any input waveform that is bound by certain criteria. The performance of the waveform coders is measured in terms of the signal-to-noise ratio (SNR) since quantisation represents the major source of distortion in the output waveform. The main drawback of this scheme lies in its inefficiency, i.e. a typical waveform coding structure requires a very high bit rate (>16 kbps) to deliver toll quality speech. Due to the fact that human speech lies in the range of 300-3400 Hz, which dictates a sampling rate of 7-8 kHz, and in order to achieve a sufficient dynamic range, it is necessary to use a 13-bit linear quantisation scale which leads to an excessive bit rate of over 100 kbps. However, by using the logarithmic companding techniques, such as A-law and μ-law [33], based on the logarithmic sensitivity of the human auditory system, it was possible to reduce the quantisation rate to 8 bits per sample. This marked the birth of the world's first and most dominant 64 kbps Pulse Code Modulation (PCM) waveform coder with logarithmic companding which has been in widespread use since the 1960's and was internationally standardised by CCITT in 1972. The toll quality of the PCM-coded speech has since served as a benchmark for comparing coding performances of lower bit rate coders. Apart from its

toll quality speech output, the popularity of PCM is mainly due to its simplicity and ease of implementation, low delay and robustness to transmission errors. Further improvements to the 64 kbps PCM coder were achieved by incorporating regular adaptation of the quantisation scheme and predictor parameters which depended upon the characteristics of the coded speech. These changes made a new coder more efficient, whilst still preserving its toll quality speech output. As a result, in 1985 CCITT approved this new waveform coder in the form of the 32 kbps Adaptive Differential PCM (ADPCM) as a new standard for digitally coded speech. However, any further attempts to lower the bit rate using the waveform coding approach have resulted in a rapid deterioration of speech output quality.

2.4.2 Source Coding

Source coding, also referred to as vocoding (a contraction of the words voice and coding), can be regarded as an opposite approach to waveform coding, as it makes no attempt to shape the input speech waveform. Instead, vocoders are very speech specific in their principles and rely on a theoretical model of the human speech production system. A vocoder consists of an analyser and a synthesiser. The analyser extracts from the original speech a set of transmission parameters. These parameters represent the perceptually important components of the theoretical model. They are then digitally encoded and passed onto the receiver. At the decoder end, the parameters of the speech production model are used to define a character of the synthesised output speech. As there is no waveform matching, the accuracy of the speech production model and the reliability of the parameter determination by the encoding algorithm set an upper limit on the attainable output speech quality, as indicated in Figure 2.4. However, the most important advantage of coding the speech production model is that it offers a far greater bit rate reduction than the waveform coding. A typical operating range of the vocoders lies between 600 bps, where speech is merely intelligible, and 4.8 kbps, where quite natural sounding speech quality can be achieved. Recent advances in this area have resulted in vocoders also producing natural sounding speech around 2.4 kbps [34].

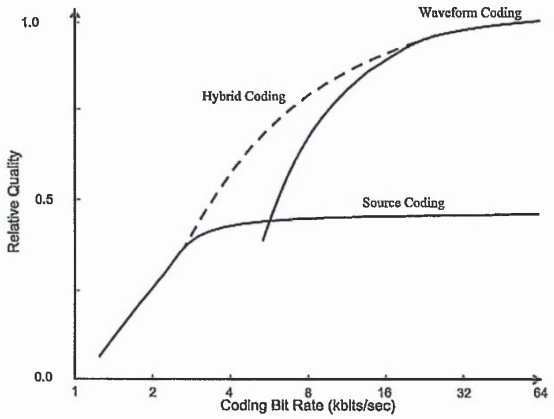


Figure 2.4 Quality versus Bit Rate for different speech coding techniques

As expected, the SNR distortion measures are of no use for vocoders due to the absence of waveform matching. Instead, vocoders rely on subjective quality measures, such as the mean opinion score (MOS), the diagnostic rhyme test (DRT) and the diagnostic acceptability measure (DAM).

One of the most prominent vocoding algorithms that has been introduced is the U.S. Government standard Linear Predictive Coding (LPC-10) [35] coder operating at 2.4 kbps. Similar to other vocoders with synthetic output quality, LPC-10 has also found its application in non-commercial communication systems, e.g. military communications. Another well-known vocoder is the 4.15 kbps Improved Multi-Band Excitation (I-MBE) coder. However, unlike LPC-10, the I-MBE is able to faithfully reproduce the speaker's characteristics, and is more robust to channel errors and background noise. Thanks to these qualities, it is capable of producing communication quality and, even, near-toll quality speech. These characteristics meant that it has a great potential for finding application not just in non-commercial but also in commercial communication systems.

2.4.3 Hybrid Coding

Hybrid coders attempt to exploit the best characteristics of the waveform and source coding schemes. The most important consequence of this merger is that the hybrid coders can produce near-toll quality speech and operate at medium bit rates. This has made them of

particular interest to speech coding research, as they can be used to bridge the gap between toll quality high bit rate coders and communication quality low bit rate coders.

A source coding part of a hybrid coder uses a speech production model to reduce the correlation between neighbouring speech samples. A waveform coding part then takes over the resulting residual or excitation signal which is waveform coded on a sample by sample basis. Early hybrid coding systems operated on an Analysis-and-Synthesis (AaS) basis, because of the lack of processing power available at the time. AaS most commonly involved removing the short-term speech correlation with the inverse of the speech production model. The long-term correlation was then removed from the residual signal using the long-term predictor and the error vector was then coded using waveform matching techniques. The coefficients of the short-term and long-term predictors were then transmitted to the decoder together with the quantised error signal. The most prominent examples of AaS scheme are Adaptive Predictive Coding (APC) [36] algorithm producing high quality speech at 16 kbps and above, and Residual Excited Linear Prediction (RELP) [37] algorithm with a lower bit rate range (9.6 kbps and above) and hence a slightly lower speech quality.

The emergence of a second generation of hybrid coders was mainly due to a large increase in DSP computational power. The new scheme was called an Analysis-by-Synthesis (AbS) technique. AbS coders applied the waveform matching process at the output of the speech production model, a so-called closed loop approach in finding the optimum excitation. Each candidate block of excitation was passed through the speech production model and the block that provided the best match with the original was chosen. There are many variants of the AbS scheme that have been proposed over the years. The most notable example is the Code Excited Linear Predictive (CELP) algorithm [38] which can nowadays be found in many commercial applications, such as Private Mobile Radio (PMR), cellular radio systems and satellite communication systems. AbS coders are capable of producing near toll quality speech in a bit rate range of 4.8 to 9.6 kbps.

2.5 Concluding Remarks

This chapter has briefly reviewed the most crucial factors that need to be addressed when designing a speech coding algorithm. These include speech quality, complexity, bit rate, delay, robustness to channel errors and background noise, as well as non-speech signal requirements and multi-rate operation. The chapter has also highlighted the fact that many of the algorithms presented are very application specific, especially in the lower end of the bit rate range.

Various terrestrial and satellite communication applications of digital speech coders, many of which are internationally recognised standards operating at bit rates in the range from 2.4 to 64 kbps, have been surveyed. The main speech coding requirements and constraints of each application have been identified and briefly analysed. It has emerged that a current trend in speech coding research is to reduce bit rate but preserve high speech quality, in order to utilise the existing bandwidth more efficiently and thus accommodate the rapidly growing user demand. The research work undertaken and presented in this thesis is in line with this trend.

The three main speech coding schemes, namely, high bit rate waveform coders, low bit rate source coders (vocoders) and medium bit rate hybrid coders, have also been introduced. As the work reported in this thesis is primarily focused on the improvement in speech coding in the very low bit rate range, further attention is paid to more detailed understanding and description of the vocoding model in the forthcoming chapters.

Chapter 3.

3. Fundamentals of Speech Coding Techniques

3.1 Introduction

The need for low bit rate, low complexity and high quality digital speech coding algorithms for a range of applications was clearly illustrated in the previous chapter. In order to design and develop such algorithms, it is necessary to understand the human speech production and perception mechanisms. The majority of current low bit rate speech coders use a parametric model of the human speech production system. This model is based on the fact that a typical speech signal consists of components or correlations of a repetitive nature which are exploited to reduce the overall coding bit rate.

This chapter begins with an introduction to some general principles of quantisation. It is followed by a brief description of the human speech production mechanism on which a vocal tract model is based. Various bit rate reduction tools, which have been developed over the years, are then introduced and formulated. The two most efficient and well-known low bit rate techniques are reviewed, that of Linear Predictive Coding, commonly referred to as LPC

analysis, and pitch prediction. In LPC analysis the short-term correlation between speech samples are efficiently modelled and removed. The pitch prediction methods are then used to remove remaining long term correlations in speech. The chapter also portrays other essential speech coding tools, such as post-filtering and interpolation, which are commonly used for the purpose of efficiently coding the speech signal.

3.2 General Quantisation Principles

In order for the analogue speech signal to be digitally processed or coded, it has to be converted into a digital domain. The digital conversion consists of two distinct processes; sampling, which dicretises an analogue signal by measuring the signal value at regular time intervals, and quantisation, whose function is to reduce the infinite range of samples to a finite set of levels using a bit representation. The quantised signal can then be transmitted as a set of bits from an encoder to a decoder, or be stored in a digital format. The main drawback of the quantisation process is that it suffers from quantisation noise or error that stems from the difference between the continuous amplitude signal and the discrete amplitude signal. Nowadays, there is a multitude of various types of signal quantisation methods and techniques, each of them aiming at (a) reducing the incurred signal parameter quantisation error and (b) making the coding structure more efficient by lowering the overall system bit rate. Scalar quantisation and vector quantisation are two main approaches to parameter quantisation in low bit rate speech coding.

3.2.1 Scalar Quantisation

Scalar quantisation represents a separate quantisation of each of a set of discrete amplitude values. It is characterised by low computational complexity, robustness to errors and low storage requirements. The main problem with this type of quantisation is that the quantisation error may increase dramatically if the number of available bits is reduced. This has prompted the design of various types of scalar quantisers. In the *uniform quantiser*, all of the quantiser intervals (steps between the levels of quantisation) are of the same width. However, this assumes even probability density function (PDF) across a finite range which is atypical for the

majority of speech-like signals, where the probability of small amplitudes occurring is much higher than that of large amplitudes. *Non-uniform quantisers* are, therefore, better suited to this type of application, as they can adjust their quantisation levels according to the statistical distribution of the parameter and hence produce higher signal-to-noise ratio (SNR) than the uniform quantiser. In speech coding, two methods of non-uniform scalar quantisation based on variable parameter PDF have generally been adopted, the Cumulative Frequency Distribution (CFD) and the Mean Square Error (MSE) approach [39]. Other types of scalar quantisation, such as adaptive quantisation or differential quantisation, are also used and have been well documented [33].

3.2.2 Vector Quantisation

Vector quantisation, also known as block quantisation or pattern matching quantisation, is the process in which a set of discrete values are quantised jointly as a single vector. The key advantage of the vector quantisation is that it can offer fractional rates in bits per sample. However, the price to pay is a very large demand on the computational capability and storage. Its other advantage is that, according to Shannon's rate distortion theory, it should always provide better performance than scalar quantisation for a given transmission rate [40][41].

In vector quantisation, an N-dimensional vector $\boldsymbol{x}$, whose elements are real valued discrete variables, is mapped onto another real valued N-dimensional vector $\boldsymbol{y}$, which is chosen from a finite set of values Y, called a codebook. The size of the codebook may be considered to be equivalent to the number of levels in scalar quantisers. A process used to design such a codebook is known as training or populating the codebook. In order to compromise between the computational and storage demand and quantiser performance, a number of codebook types have been developed over the years [42]. Some codebooks are static, i.e. pre-computed before use, while others are adaptive, i.e. updated during quantisation. One of the most popular methods for codebook design is an iterative clustering algorithm known as the Linde Buzo Gray (LBG) algorithm [43].

3.3 Vocal Tract Model

3.3.1 Speech Production Mechanism

Figure 3.1 shows the human vocal tract [44] which extends from the vocal cords to the lips, including the nasal cavity. Speech is produced by passing air, expelled from the lungs, through a restriction in the vocal tract.

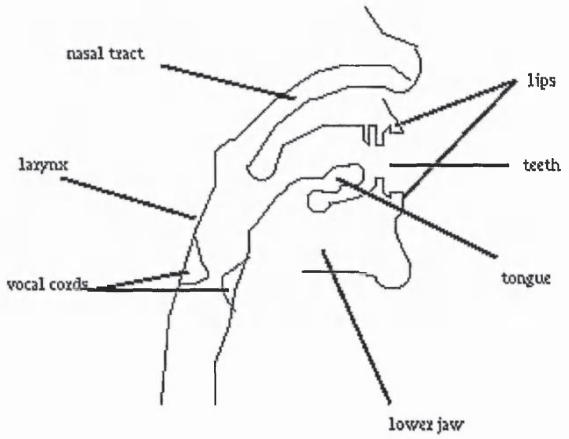


Figure 3.1 Vocal tract for speech production

The waves generated by the vibration of the vocal cords travel along the vocal tract being spectrally shaped by the varying diameter of the larynx and by the position of the tongue, jaw, lips and teeth. The closure or otherwise of the nasal tract also has some effect. The frequency, or pitch, of the vibration is dependent upon the tension of the vocal cords, which is under muscular control. Such sounds, referred to as 'voiced', are usually high in energy and periodic. A second group of sounds, referred to as 'unvoiced', are formed by a turbulent air flow past a constriction in the vocal tract, e.g. tongue, lips or teeth. They contain less energy than the voiced sounds, no periodicity and the waveform appears random though with some limited spectral shaping. Fully unvoiced sounds, better known as whispering, are produced by a constriction in the larynx when the vocal cords are open. Some unvoiced sounds, known as fricatives, may be accompanied by voiced excitation generated by the vocal cords, forming a 'mixed' excitation, a third group of sounds. A fourth category, also known as 'silence', may

be observed when no speech is present. The waveform has very low energy and is mostly random.

3.3.2 Speech Production Model

The speech production model can be described as a source of energy whose output is modulated by the acoustic filtering action of the vocal tract, a highly non-linear filter. However, a more realistic approach can be made by adopting a simplified linear approximation in which an excitation source, consisting of either a periodic train of pulses (which is in itself an approximation) for voiced speech or a random (Gaussian) source for unvoiced sounds, is supplied to a linear filter. Therefore, a speech waveform at the output of the model is considered a convolution of an excitation function and the linear filter response.

In an attempt to achieve a maximum bit reduction, many speech coders employ a parametric approach to speech coding, in which a mathematical model of the speech production mechanism is developed. The underlying idea is that instead of representing the speech signal on a sample-by-sample basis, typical for high bit rate coders (e.g. PCM, ADPCM), artificial or synthetic speech having perceptually similar characteristics to the original speech is generated. Essentially, the synthetic speech is produced using the parameters of the speech production model. Among many speech production models proposed over the years, the Source-Filter model [45] has been widely adopted for the vocoder applications.

The Source-Filter model is depicted in Figure 3.2. The main assumption is that the excitation source is independent of the modulation, i.e. no feedback path exists and no interaction between the vocal tract shape and larynx occurs. The second key assumption is that the vocal tract is represented as a time-varying linear transfer filter, and it is either excited by random noise for unvoiced speech or pitch period pulses for voiced speech. The excitation is then scaled by a suitable factor to produce the synthetic output. The Source-Filter model essentially acts as a lossless acoustic tube [46], consisting of p equal length sections of varying cross-sectional area. Adopting these two assumptions results in large reductions in terms of computational complexity.

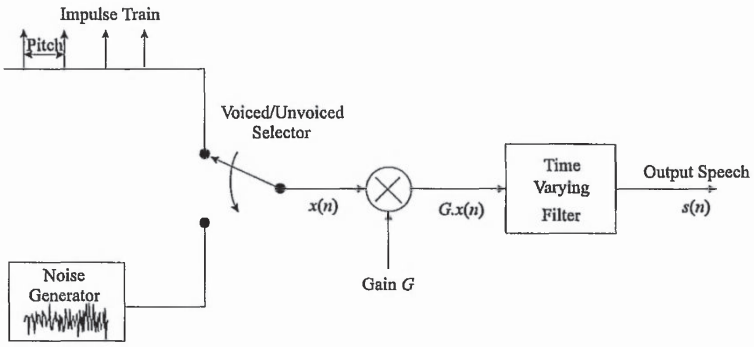


Figure 3.2 Block diagram of the simplified Source-Filter model

3.4 Linear Predictive Coding (LPC)

The LPC analysis [47] is a convolutional parametric analysis method based on the time-varying digital filter model. It is one of the most powerful analysis methods used in speech coding. The basic idea of the LPC analysis is modelling the short-term correlations that exist between successive speech samples in the form of a short order filter. By using the inverse function of the LPC filter, the short-term correlations are removed, which is one of the main applications of the LPC analysis.

3.4.1 Estimation of LPC Parameters

The time-varying digital filter which represents the combined effects of the vocal tract, glottal flow and the radiation of the lips can be given as a pole-zero transfer function

$$H(z) = \frac{S(z)}{X(z)} = \frac{G(1-\sum_{k=1}^{q}\beta_k z^{-k})}{(1-\sum_{j=1}^{p}\alpha_j z^{-j})} \tag{3.1}$$

where $S(z)$ is the speech output, $X(z)$ is the excitation (periodic or random) and G is the gain factor. In the speech signal, the input samples can never be observed directly and so the estimation of the transfer function of the vocal tract is made using only the knowledge of present and past output samples. To achieve this, an all-pole model, also known as the Auto

Regressive model, is used, which sets the order of the denominator high enough, so that $H(z)$ can be approximated, as given by

$$H(z) = \frac{G}{1 - \sum_{j=1}^{p} \alpha_j z^{-j}} = \frac{G}{A(z)} \tag{3.2}$$

where

$$A(z) = 1 - \sum_{j=1}^{p} \alpha_j z^{-j} \tag{3.3}$$

If this is transformed into the sampled time domain, it can be written in the form of a difference equation

$$s(n) = Gx(n) + \sum_{j=1}^{p} \alpha_j s(n-j) \tag{3.4}$$

Equation 3.4 states that the value of the present output sample, $s(n)$, can be precisely obtained by summing the weighted present input, $Gx(n)$, and a weighted sum of the most recent past output samples. If the scaling term is assumed to take a unity value, the next step is to determine the predictor coefficients, α_j for j=1,2...,p where p is the order of the filter. It is the determination of the predictor coefficients that lies at the heart of the LPC analysis.

The difference between the predicted value for the current speech sample and the actual original sample value is the prediction error signal $e(n)$, termed a prediction residual, given by

$$e(n) = s(n) - \sum_{j=1}^{p} \alpha_j s(n-j) \tag{3.5}$$

The set of optimum prediction coefficients may be determined using a least mean square (LMS) error criterion given by

$$\min E = E\left[e^2(n)\right] = E\left[\left(s(n) - \sum_{j=1}^{p} \alpha_j s(n-j)\right)^2\right] \tag{3.6}$$

where E represents the expectation value of the expression. The LMS solution, and hence the α_j coefficients, can now be determined by equating the partial derivatives of the above with respect to α_j to zero,

$$\frac{\partial\left(E\left[e^2(n)\right]\right)}{\partial\alpha_j} = \frac{\partial\left(E\left[s(n)-\sum_{j=1}^{p}\alpha_j s(n-j)\right]^2\right)}{\partial\alpha_j} = 0 \tag{3.7}$$

This results in the solution,

$$E\left[\left(s(n)-\sum_{j=1}^{p}\alpha_j s(n-j)\right)s(n-i)\right] = 0 \qquad 1 \le i \le p \tag{3.8}$$

As prediction error *e(n)* is orthogonal to *s(n-i)* for *i=1,..,p*, Equation 3.8 can be rearranged to give a set of linear equations,

$$\sum_{j=1}^{p}\alpha_j\phi_n(i,j) = \phi_n(i,0), \qquad 1 \le i \le p \tag{3.9}$$

where

$$\phi_n(i,j) = E\left[s(n-i)s(n-j)\right] \tag{3.10}$$

A key assumption in the derivation of Equation 3.9 is that the speech signal is stationary for a short period of time (up to 40 ms). In order to limit the LPC analysis to short-time speech segments, Equation 3.9 can be rewritten as

$$\phi_n(i,j) = \sum_m s_n(m-i)s_n(m-j) \qquad 1 \le i \le p,\ 0 \le j \le p \tag{3.11}$$

where m is the length of the sample sequence. Once the quantities $\phi_n(i,j)$ are computed, the set of LP coefficients, α_j, can be obtained.

There are two basic approaches to solving Equation 3.11: the Auto-correlation Method (AM) and the Covariance Method (CM). Both these methods consist of two steps: computing a matrix of correlation values and solving a set of linear equations. There is also another approach, called the Lattice Method (LM), in which the two steps have in a sense been combined into a recursive algorithm. Although the three methods have similar principles, AM has the advantage of being both computationally somewhat less intensive than the other two and, with respect to the CM, able to guarantee the stability of the LP coefficients. In this chapter, a brief description of only the AM is given due to its importance for the speech

coding work detailed in this thesis. A more thorough explanation of the CM and LM can be found in [48].

3.4.1.1 The Auto-correlation Method (AM)

For an accurate solution of AM, the speech segment $s_n(m)$ is assumed to be zero outside the windowed sequence as given by

$$s_n(m) = s_n(n+m)w(m) \tag{3.12}$$

where $w(m)$ is an N sample duration window function which is also zero outside the interval $0 \le m \le N-1$. Since $s_n(m)$ is non-zero for the interval $0 \le m \le N-1$, the corresponding prediction error $e_n(m)$ over the extended interval $0 \le m \le N-1+p$ will be affected due to the non-zero filter memory of p samples. The consequence is that the error will be large at both the beginning of the segment, as the signal is being predicted from samples which have been predominantly set to zero, and at the end of the segment, where the predictor is trying to predict zero samples from previous samples which are clearly non-zero. A way to solve this problem is to employ the window function, $w(m)$, that tapers the ends of the signal segment $s_n(m)$ towards zero.

Assuming that the future prediction performance is of main interest, the limits in Equation 3.11 can be set at

$$\phi_n(i,j) = \sum_{m=0}^{p+N-1} s_n(m-i)s_n(m-j) \qquad 1 \le i \le p, 0 \le j \le p \tag{3.13}$$

or

$$\phi_n(i,j) = \sum_{m=0}^{N-1-(i-j)} s_n(m).s_n(m+i-j) \qquad 1 \le i \le p, 0 \le j \le p \tag{3.14}$$

Equation 3.14 can then be reduced to the short-term auto-correlation function as given by

$$\phi_n(i,j) = R_n(|i-j|), \qquad 1 \le i \le p, 0 \le j \le p \tag{3.15}$$

where

$$R_n(j) = \sum_{m=0}^{N-1-j} s_n(m)s_n(m+j) \tag{3.16}$$

Substituting the Equation 3.9 into the above Equation yields the following expression

$$\sum_{j=1}^{P} \alpha_j R_n(|i-j|) = R_n(i), \qquad 1 \le i \le P \tag{3.17}$$

which can be expressed in matrix form as

$$\begin{bmatrix} R_n(0) & R_n(1) & \cdots & R_n(P-1) \\ R_n(1) & R_n(0) & \cdots & R_n(P-2) \\ \vdots & \vdots & \vdots & \vdots \\ R_n(P-1) & R_n(P-2) & \cdots & R_n(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R_n(1) \\ R_n(2) \\ \vdots \\ R_n(P) \end{bmatrix} \tag{3.18}$$

The above auto-correlation matrix bears two distinct properties of a Toeplitz matrix: it is symmetrical and it has equal elements along a given diagonal. Solving the Toeplitz matrix requires very efficient recursive methods, which have been devised on the basis of its characteristics. The most widely used is Durbin's algorithm [49] which is detailed below.

$$E_n^{(0)} = R_n^{(0)} \tag{3.19}$$

$$k_i = \frac{\left[R_n(i) - \sum_{j=1}^{i-1} \alpha_j^{i-1} R_n(i-j) \right]}{E_n^{(i-1)}} \qquad 1 \le i \le p \tag{3.20}$$

$$\alpha_i^{(i)} = k_i \tag{3.21}$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \qquad 1 \le j \le i-1 \tag{3.22}$$

$$E_n^{(i)} = (1 - k_i^2) E_n^{(i-1)} \tag{3.23}$$

After recursively solving Equations 3.20 to 3.23 for i=1,2,...,p, the resulting α_j is given by

$$\alpha_j = \alpha_j^{(p)} \quad 1 \le j \le p \tag{3.24}$$

3.4.2 Practical Implementation of LPC Analysis

There are several issues that need to be addressed in the practical implementation of the LPC analysis, such as performance, stability, computational complexity, window function (if used), frame size and filter order. The performance, stability and computational complexity are determined by the chosen estimation method (e.g. AM, CM, etc.). The type and size of the window function is another important factor. There is a number of different window functions that can be used in the AM: Rectangular, Hamming, Hanning, Kaiser, Blackman and Bartlett [50][51][52]. The two most desirable window characteristics are: high frequency resolution

and minimum spectral smearing. The requirement of the former is a narrow main lobe, whereas that of the latter is low side lobes. There is no single window function which is able to meet both requirements and hence the best overall performance. The choice of an adequate window function depends on a given speech coding application. The two windows widely used in this thesis are the Hamming window and the Kaiser window. The Hamming window represents a good compromise between the frequency resolution and spectral smearing, and is defined as

$$w_h(m) = \begin{cases} 0.54 - \left(0.46\cos\left(\dfrac{2\pi m}{N-1}\right)\right) & 0 \le m \le N-1 \\ 0 \end{cases} \tag{3.25}$$

The Kaiser window has a wider main lobe but very low side lobes, and is mathematically described as

$$w_k(m) = \begin{cases} \dfrac{I_0\left(\beta\sqrt{1-\left(\dfrac{2m}{N-1}-1\right)^2}\right)}{I_0(\beta)} & 0 \le m \le N-1 \\ 0 \end{cases} \tag{3.26}$$

where I_0 is zero order Bessel function given by

$$I_0(\beta) = \sum_{k=0}^{\infty} \frac{\beta^{2k}/2}{(k!)^2} \tag{3.27}$$

The main advantage of the Kaiser window is that it incorporates a β 'tuning' factor, which controls the behaviour of the main lobe width and side lobe leakage.

The size of the window function, N, plays an important role in the accurate LPC analysis. If N is small (e.g. 30-40 samples), the short-time energy will fluctuate very rapidly. In addition, a short window size also increases the analysis update rate. If N is too large (e.g. several pitch periods), the short-time spectral energy variations will be averaged over a long time, which in turn will cause the LPC filter to model the rapidly time varying spectral changes rather than the spectral envelope. Also, a long window size may require a large algorithmic delay due to the buffering of the future samples. It can be concluded that there is no satisfactory window size. However, a useful rule of thumb is that a chosen window size must be able to represent the harmonic structure fairly accurately by being able to accommodate more than one or two

pitch periods in each window. As the pitch period can vary from about 16 samples for a high pitch female or a child, up to 150 samples for a very low pitch male, a compromise is made by setting a suitable practical value of the window to N=120-240 samples. Switching between various window sizes depending on the identified pitch period of the speech segment can also enhance the LPC analysis. This option is detailed in Chapter 7. The choice of the frame size obviously corresponds to that of the used window size as the frame is made to fit within the window boundaries.

An important aspect of the process of windowing is the so-called block edge effect. The block edge effect stems from the low weighting given to the near-end window samples. This is particularly relevant in cases where these samples describe a significant short-term speech event (e.g. onset), which will be suppressed by any window weighting function other than the rectangular window. This problem can be overcome by the use of either overlapping windows or parameter interpolation.

The order of the LPC filter largely depends on the spectral size of the processed speech. In *narrowband* speech coding, speech is typically sampled at 8 kHz, rendering a 4 kHz analysis spectrum with up to 4 resonant spectral peaks, usually referred to as *formants*. As each formant requires a two pole conjugate pair for its representation, the minimum number of LP coefficients is eight. However, a 10th order LPC filter is generally employed as it provides a higher accuracy in modelling not just the formant structure but also the general spectral shape. In *wideband* speech, the sampling rate is 16 kHz, rendering an 8 kHz analysis spectrum with up to 7 formants. Such formant structure calls for a minimum of 14 coefficients, although a 16-pole filter is normally in use, again for higher accuracy. The graph in Figure 3.3 justifies the use of the filter order in both the narrowband and the wideband. It shows a plot of a prediction gain as a function of the LPC order for female speech. It can be seen from the plot that the performance of the LPC predictor saturates at the 10th order and 16th order for the narrowband and wideband, respectively. The prediction gain is a useful objective measure of the spectral modelling performance.

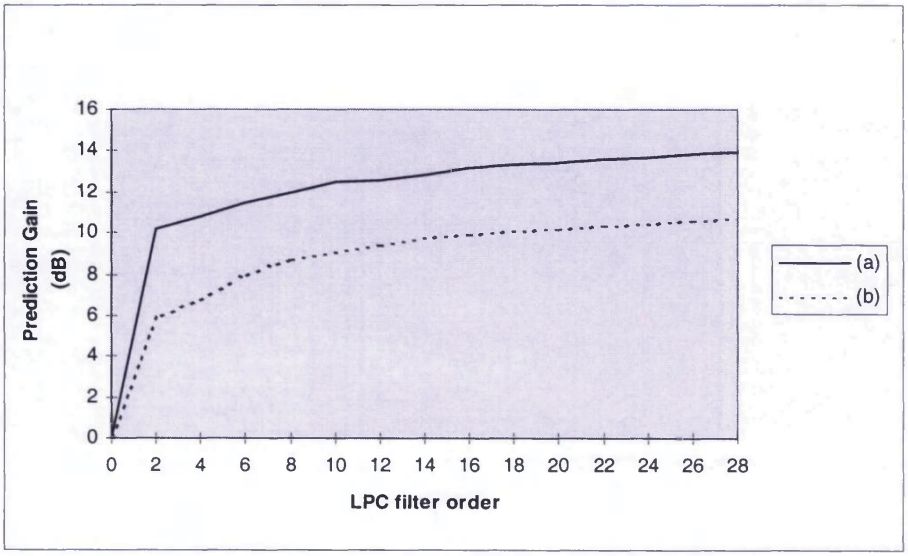


Figure 3.3 LPC prediction gain vs. LPC order for (a) wideband speech and (b) narrowband speech

In order to increase its accuracy, the LPC filter is often preceded by two pre-processing stages, namely, *notch filtering* and *pre-emphasis*. The notch filter removes any residual DC component in speech which can adversely affect the high frequency regeneration in the decoder [53]. The pre-emphasis serves two purposes. It improves the quantisation process as it reduces the signal dynamic range, and it enhances the modelling of the low amplitude high frequency formants. Pre-emphasis on the encoder side is matched with de-emphasis on the decoder side.

3.4.3 Application of LPC Analysis

Inverse LPC filtering is the process of removing the short-term correlation in the speech signal, which leaves the resultant error or excitation signal. This excitation signal is characterised by its long-term correlation in the time domain, which is depicted in the example given in Figure 3.4, and a relatively flat spectrum in the frequency domain, as shown in the corresponding frequency response given in Figure 3.5.

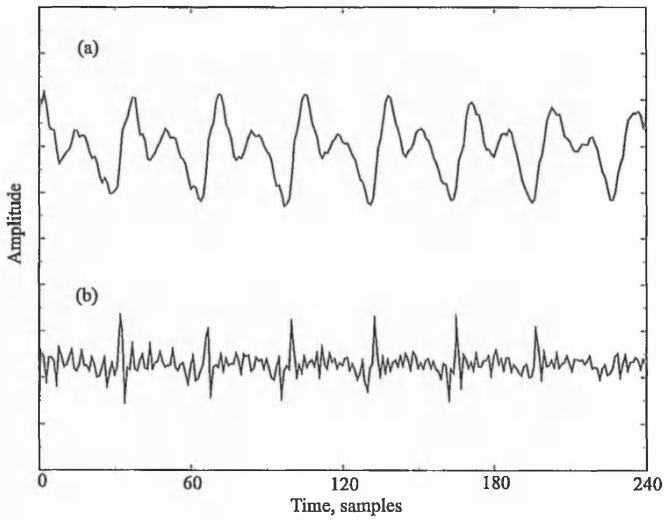


Figure 3.4 Time domain plots of the (a) original speech and (b) LPC residual

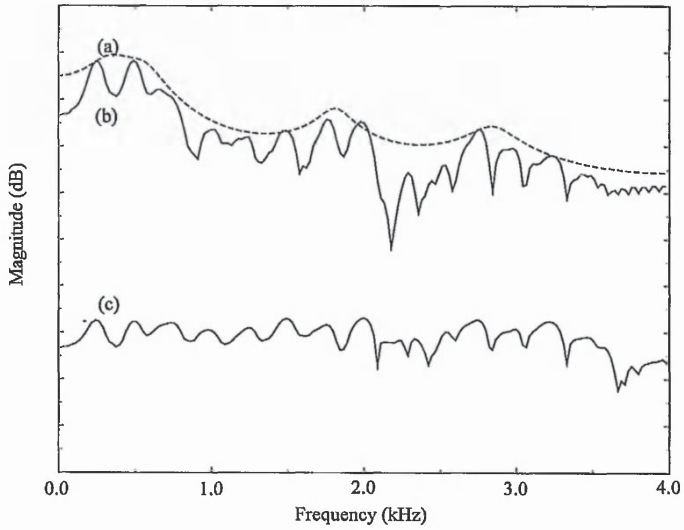


Figure 3.5 Frequency domain plots of the (a) original speech envelope, (b) original speech and (c) LPC residual

From Figure 3.5, it can also be observed that the LPC spectral envelope matches the speech spectrum more accurately in the spectral peaks than in the spectral valleys. This feature is mainly due to the fact that the filter transfer function has enough poles to model the formant peaks but no zeros to model the valleys.

3.4.4 Quantisation of LP Coefficients

The LP coefficients of the *H(z)* transfer function provide an accurate and compact representation of the speech spectral envelope. Therefore, maintaining the accuracy of the coefficients, by keeping all the poles within the boundary of the z-domain unit circle, is essential for a good and reliable coder performance. However, the direct quantisation of the LP coefficients poses a great design problem, as even a small disturbance in any of the coefficients can lead to significant spectral distortions. The practical considerations must also be taken into account, as the LP parameters cover a large dynamic range and, therefore, require a large number of bits for an accurate quantisation. Moreover, the straightforward interpolation of the LP parameters between frames can also result in an unstable filter. In order to overcome all the aforementioned problems, the LP coefficients are mapped into various alternative representations before undergoing any quantisation or interpolation. These include PARCOR coefficients, log area ratios (LARs), arc-sine reflection coefficients (ASRCs), and line spectrum frequencies (LSFs) or pairs (LSPs). A lot of research has been done in determining the relative advantages and disadvantages for using any of these representations [54][55][56]. The representation used in this thesis is that of the LSFs, whose properties, derivation and quantisation are described in the following sections.

3.4.4.1 Line Spectral Frequencies (LSFs)

3.4.4.1.1 Definition

Recent research [57][58] has shown that the LSF parameters can provide not only high coding efficiency with respect to quantising the LP coefficients, but also some other very useful properties that could be successfully exploited in the area of low bit rate speech coding. By definition, for any given stable LP filter of even order p, there exists a set of p LSFs which completely describes the filter.

3.4.4.1.2 Derivation

LSFs are related to the poles of the LPC filter $H(z)$. For the derivation of LSFs, the inverse filter $A(z)$ function from Equation 3.2 is decomposed into

$$A(z) = \frac{P(z) + Q(z)}{2} \tag{3.28}$$

where two formulated polynomials are,

$$P(z) = A(z) + z^{-(p+1)} A(z^{-1}) = z^{-(p+1)} \prod_{j=0}^{p+1} \left(z - a_j\right) \tag{3.29}$$

and

$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1}) = z^{-(p+1)} \prod_{j=0}^{p+1} (z - b_j) \tag{3.30}$$

The roots of $P(z)$ and $Q(z)$ lie on the unit circle in complex conjugate pairs, except for the roots at $z=-1$ in the case of $P(z)$ and $z=1$ in the case of $Q(z)$, with $p/2$ on the upper half and $p/2$ on the lower half. Since the angles of the roots on the lower half are negative of the ones on the upper half, a total of p unknowns must be found, i.e. the arguments of a_j and b_j. The cosine term of the arguments a_j and b_j arises from the determination procedure of the Real Root method [59], one of the numerous methods of solving the roots of the polynomials $P(z)$ and $Q(z)$ that have been reported in the literature over the years [60][61]. The Real Root method converts the polynomials with complex roots into the polynomials with real roots and then applies the Newton-Raphson method in an iterative search for these roots. The inverse transformation from LSFs to LPCs is less computationally intensive than the forward transformation and can be achieved by either a direct expansion method or an LPC synthesis filter method [62].

3.4.4.1.3 Properties

LSFs are defined as the angular positions of the roots of P(z) and Q(z) lying in the range of 0 to π radians. Filter stability is guaranteed providing the natural ordering or monotonicity of LSFs for a given speech frame. In addition, the LSF parameters exhibit strong correlation both within a frame, termed *intra*-frame correlation, and between frames, known as *inter*-frame correlation. These two properties are depicted in Figure 3.6, which contains plots of

typical LSF trajectories obtained for both a 10th order narrowband LPC filter and a 16th order wideband LPC filter.

The ordering property can be used for the detection of any transmission errors, i.e. by checking for crossovers of the LSFs in a given frame. The strong correlation property may be exploited for the purpose of quantising the LSF coefficients which can yield a 30% bit reduction over an equivalent PARCOR quantiser [63].

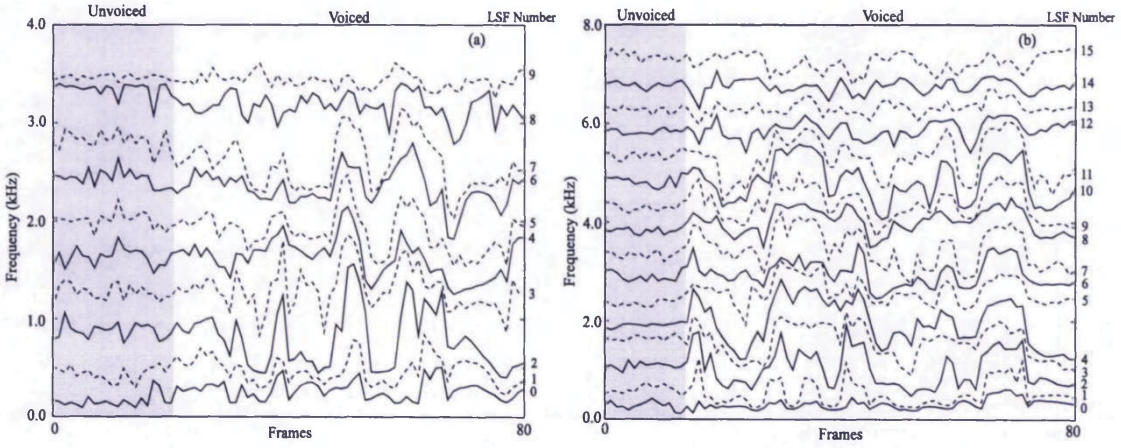


Figure 3.6 Typical LSF trajectories for (a) narrowband (10th order LPC filter) and (b) wideband speech (16th order LPC filter)

The LSFs correlate directly with the LPC magnitude spectrum, as can be seen in Figure 3.7. Closely grouped LSFs indicate the presence of a formant, where the degree of closeness is a measure of the bandwidth or the formant strength. Each LSF coefficient is characterised by a local spectral sensitivity. As a result, a spectral distortion incurred in a particular LSF will not affect the rest of the spectrum. This last property shows that the LSF representation can exhibit a high degree of robustness against transmission channel errors, a feature particularly pertinent to low bit rate applications such as cellular and satellite mobile communications.

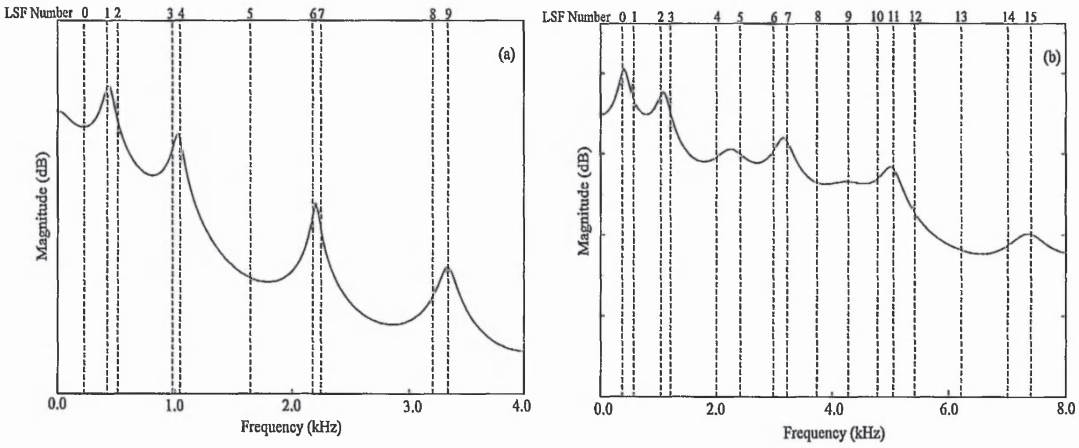


Figure 3.7 LSF parameters co-plotted with the corresponding LP spectrum for (a) narrowband speech and (b) wideband speech

3.4.4.1.4 Quantisation

Designing an LSF quantiser which introduces little or no perceptual distortion and is low in complexity, is one of the most important tasks in low bit rate speech coding. The main reason for it is that it normally requires the largest share of the bit allocation scheme which ultimately determines the overall coding bit rate.

Vector quantisation is the predominant technique used in quantisation of the LSFs at present. A split-vector quantisation approach is often taken, in which the LSF parameters are split into groups and each group is then individually vector quantised. Previous research [58] has shown that the optimum coding performance can be achieved if the dimension and size of LSF vector quantisers is set to be as large as possible within the limits imposed by the requirements of storage and computational complexity.

As already mentioned, LSF parameters exhibit strong correlation both on the inter-frame and on the intra-frame level. Both types of correlation can be successfully exploited in the course of the quantisation process.

The intra-frame correlation property is utilised in a method that involves normalising groups of LSFs to 2 or 3 adjacent single LSFs that have already been quantised. For example, the quantised values of LSF 8 and LSF 4 can be used as normalisation factors. In this case, LSFs 8, 9 and 10 are vector quantised, at first. Then, the quantised value of LSF 8 is used as a division factor for LSFs 4,5,6 and 7, before their vector quantisation. This routine is then repeated for LSFs 1,2 and 3, in which case LSF 4 plays the role of the division factor. The advantages of this method are the removal of some degree of correlation between neighbouring LSFs and a guaranteed filter stability. The latter one is true even when some bits are corrupted during transmission, since the trained normalised vector elements will always lie between 0 and 1, and will also be arranged in ascending order. Other similar methods have also been proposed in the literature [64].

The inter-frame correlation between successive frames of LSF parameters may be exploited using predictive LSF quantiser schemes, where a moving average predictor makes an estimate of the current LSF value based on the previous LSF parameter. The LSF prediction residual is then quantised by, usually, a split-vector quantiser with the LBG algorithm [46].

The LSF quantisation scheme is a combination of the techniques based on the intra-frame and inter-frame correlation properties. Initially, LSFs are normalised to account for intra-frame correlation. The inter-frame correlation is then removed for each normalised LSF and the resulting residuals are quantised using the split-vector quantisation of the respective normalised LSF groups.

The LSF quantisation scheme can also be applied to wideband speech, though a slight adjustment in the normalisation routine is needed to accommodate 16 LSFs.

3.5 Perceptual Weighting

The optimisation criterion used in the LPC analysis is based upon the commonly used mean squared error (MSE) principle, due to its simplicity and adequate performance. However, the MSE metric cannot maintain its waveform matching performance at very low bit rates, where

there is frequently only one bit per sample (or less) available. At these very low bit rates, an error criterion is required which should be more in line with the perceptual characteristics of the human ear. This error criterion is termed a weighting function or a post-filtering function.

The main characteristic of the perceptual weighting is that it can improve the performance of the LSF quantisation method, especially in the spectral valley regions between adjacent formants, which would otherwise be perceived as roughness or noise in the synthetic speech. Numerous weighting schemes have been developed over the years to lessen this effect and increase the overall synthetic speech quality[65][66][67]. However, they also suffer from various side effects. For example, an adaptive post-filtering [68] introduces a spectral tilt which reduces the energy in the higher frequencies, making the speech sound dull. An opposing spectral tilt filter is needed to counter such effect and raise the high frequencies. The other potential side effect is the variable amplification of the formants, which can also affect the synthetic speech quality.

3.6 Pitch Predictive Coding

Pitch period represents long-term correlation or periodicity contained in voiced speech. As such, its contribution is very significant in retaining the natural quality of speech. Moreover, the accurate pitch information is often a fundamental prerequisite of many of the vocoding functions, such as the voicing determination algorithm. The accuracy and reliability of the pitch detection algorithm (PDA) therefore largely influences the performance and output quality of low bit rate vocoding systems.

By definition, pitch period indicates the time between two glottal pulses, which may vary from one pulse to the next. A periodic speech waveform containing several glottal pulses has a comb like spectrum consisting of equally spaced harmonics shaped by the filtering action of the vocal tract. During the analysis, the PDA attempts to extract a single most prominent pitch value from a speech segment or frame. This is a very difficult task due to the fact that the majority of speech signals displays quasi-periodic non-stationary characteristics. They are usually mixed in nature, consisting of a periodic waveform contaminated with noise. Other

intrinsic features that can greatly affect the PDA performance are strong formant interaction, rapidly changing energy, pitch fluctuations, etc. Some extrinsic effects also present problems to the PDA, such as the presence of background noise, multiple speakers and the general need to band-limit the telephone speech, which may attenuate the fundamental frequency. In fixed frame size algorithms, it can prove difficult for the PDA to detect the exact beginning and the end of the pitch period.

The pitch estimation is performed either in the time domain or in the frequency domain, although there are also some hybrid versions that exploit both domains. A brief review of some standard approaches to pitch detection is given in the following sections.

3.6.1 Pitch Determination in the Time Domain

The most obvious characteristic of periodic signals is the distant similarity of the waveform in the time domain. Many existing PDAs are based on this principle of waveform similarity. The best known methods are the waveform peak picking, the auto-correlation and the average magnitude difference function.

3.6.1.1 Waveform Peak Picking

The waveform peak picking (WPP) is based on the observation that periodicity in speech can be represented by a high energy pulse that decays throughout the remainder of the period. The WPP algorithm attempts to determine the positions of local waveform minima and maxima. A sub-set of these candidates for local pulse position selection is then chosen empirically, showing the most likely glottal excitation instants between which the pitch period lies [69]. The key advantage of this method, its very low computational requirements, has been limited since the advent of VLSI technologies that enabled the use of more sophisticated PDAs. Moreover, the WPP can be highly prone to errors in the presence of noise or formant interference. These are the main reasons why it is rarely used today.

3.6.1.2 Auto-correlation

The auto-correlation function is to find similarities between a speech waveform and its time-shifted version, as given in Equation 3.31,

$$R(\tau) = \sum_{n=0}^{N-1} s(n)s(n-\tau) \tag{3.31}$$

where N is the analysis frame length and τ is the shifted distance, called *lag* or *delay*. The pitch is equal to the value of τ, which results in the maximum $R(\tau)$.

An important issue when designing an auto-correlation based PDA is the size of the analysis frame. It is generally accepted that the window must be at least twice the length of the maximum pitch period in order to properly accommodate an entire pitch cycle. However, if the window length is too large, it may become difficult for the PDA to detect both small pitch fluctuations, due to a lower temporal resolution, and large pitch fluctuations, due to the 'smearing' of the auto-correlation peaks.

The most common task that auto-correlation PDAs face is prevention of erroneous selection of a pitch period multiple. This is particularly relevant for cases of non-stationary speech with rapid energy fluctuations where auto-correlation peaks corresponding to pitch multiples are far greater than the actual pitch peak. An evident example of this phenomenon occurs at speech offsets, as depicted in Figure 3.8 [70], where the auto-correlation value increases with the lag size.

The normalised auto-correlation method partially solves this problem by including the normalisation factor, which is dependent on the RMS energy of the shifted frame. This auto-correlation method with shift energy normalisation is given in Equation 3.32 below

$$R_n(\tau) = \frac{\sum_{n=0}^{N-1} s(n)s(n-\tau)}{\sqrt{\sum_{n=0}^{N-1} s(n-\tau)s(n-\tau)}} \tag{3.32}$$

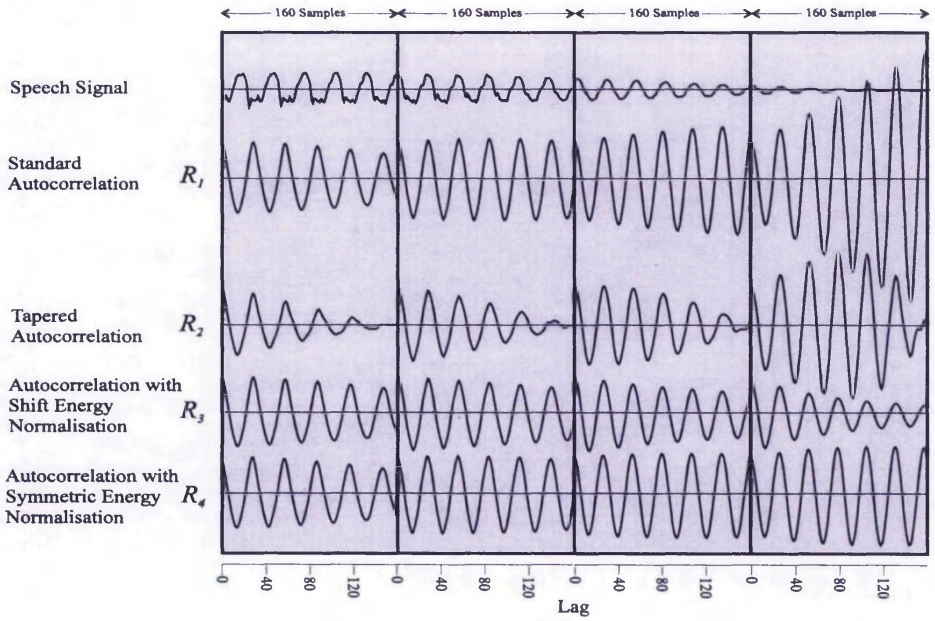


Figure 3.8 Auto-correlation pitch detection at speech offsets

However, the normalisation routine can cause substantial problems at speech onsets, as depicted in Figure 3.9, where the $R_n(\tau)$ gets smaller as the lag increases, hence its reciprocal value becomes very large and noisy.

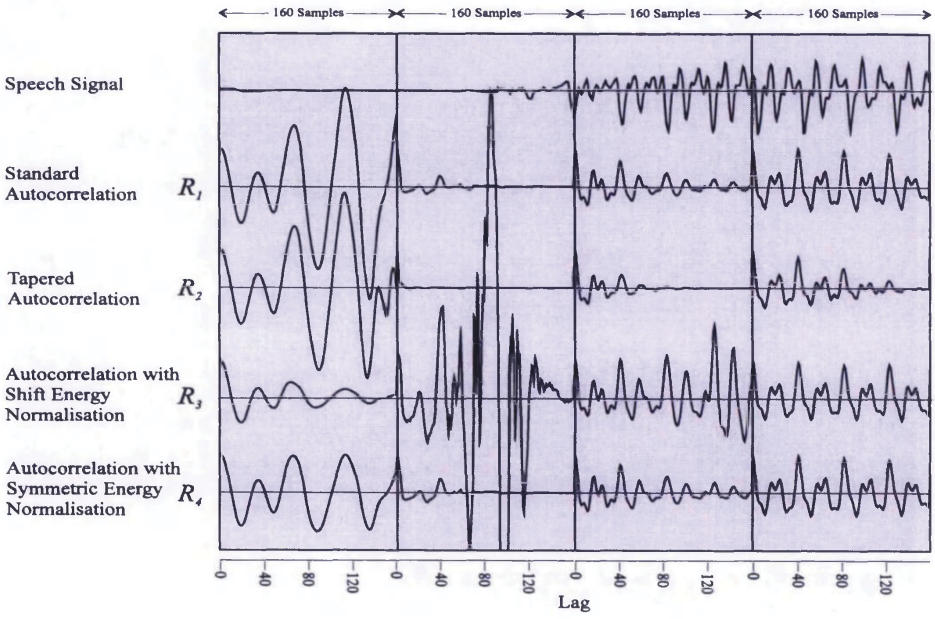


Figure 3.9 Auto-correlation pitch detection at speech onsets

This effect calls for the inclusion of the RMS energy of the frame itself, which will balance off the auto-correlation performance at both ends of the speech signal. Equation 3.33 depicts this modified version of the $R_n(\tau)$, called the auto-correlation with symmetric energy normalisation

$$R_{nn}(\tau) = \frac{\sum_{n=0}^{N-1} s(n)s(n-\tau)}{\sqrt{\frac{1}{2}\left[\sum_{n=0}^{N-1} s(n-\tau)s(n-\tau) + \sum_{n=0}^{N-1} s(n)s(n)\right]}} \tag{3.33}$$

Although it offers very reliable performance at onsets and offsets, the above solution still does not guarantee 100% success in combating the multiple pitch or half pitch errors that often occur during various other parts of non-stationary speech. The generic function of various sub-multiple pitch checking algorithms [70] is to detect the largest auto-correlation peak and any other peak lying at sub-multiple pitch positions. The ratio between the largest peak and the sub-multiple peaks is compared against a set threshold value or a dynamic threshold value. The smallest sub-multiple which produces a ratio greater than the threshold is accepted as the pitch period.

The auto-correlation based PDAs also encounter great difficulties in solving the formant interference problem. It takes place during highly resonant speech, where a strong formant harmonic may tend to dominate the overall spectrum. The solution to this problem can be found in the use of spectrum flattening and pitch tracking, which are described in the later sections of this chapter. The effect of formant interference becomes particularly reinforced when the pitch changes during the analysis window. Using a single auto-correlation in these cases can have a blurring effect which can deteriorate the PDA performance. However, the introduction of the segmented auto-correlation method, originally proposed in [70], can significantly improve the time domain PDA performance. The idea behind the segmented auto-correlation method is that, by segmenting the auto-correlation into shorter subsets and by calculating the median peak separation and the lag within each subset, it is possible to track the pitch changes within the analysis frame and thus improve the reliability of the auto-correlation method. The segmented auto-correlation PDA (SAPDA) is explained in more detail in Chapter 5, where it is used as a part of the Time Envelope Split-Band LPC vocoding algorithm.

3.6.1.3 Average Magnitude Difference Function

The Average Magnitude Difference Function (AMDF) is defined as

$$E(\tau) = \frac{1}{N}\sum_{n=0}^{N-1}|s(n) - s(n-\tau)| \tag{3.34}$$

The AMDF measures the signal disagreement as the pitch period corresponds to function minima, as shown in Figure 3.10 below.

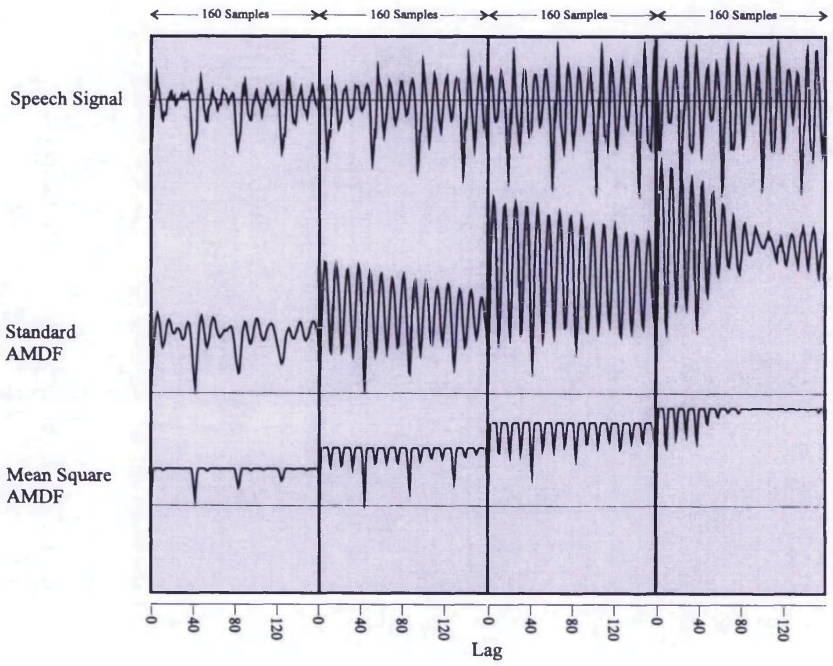


Figure 3.10 Average Magnitude Difference Function

The advantages of the AMDF over the traditional auto-correlation method lie with its lower computational complexity, relatively small dynamic range and narrower valleys for stationary signals [71]. However, the performance of AMDF based PDAs is also greatly affected by the changing speech energy, formant interference and, generally, during non-stationary sections of speech, where minima may not appear so clearly. Moreover, experimental investigation [70] has implied that the auto-correlation method is superior to the AMDF.

3.6.2 Pitch Determination in the Frequency Domain

The main frequency domain feature of a periodic signal is its harmonic structure, with the distance between harmonics being the reciprocal of the time domain pitch period. Therefore, in any frequency domain PDA the speech signal is transformed into its spectral domain by means of a Discrete Fourier Transform (DFT) or, more frequently, a Fast Fourier Transform (FFT) [3]. The transformation can only be performed on the windowed speech, hence the use and choice of a window function (as already discussed in Section 3.4.2) have a great impact on the signal spectral representation. The frequency domain PDAs briefly described in the following sections are: Harmonic Peak Picking and Synthetic Spectral Matching.

3.6.2.1 Harmonic Peak Picking

The Harmonic Peak Picking (HPP) algorithm selects the most likely pitch (fundamental) frequency out of the local peaks detected in the spectrum. This may be achieved either by detecting the largest common factor of the peak frequencies or by detecting the most common inter-peak distance. The HPP has been found [70] to perform reasonably well in the presence of background noise or unvoiced sounds. It has also shown reliable performance when a fundamental harmonic is filtered out, a potential problem when band-passing the telephone speech. Spectral up-sampling may be used to clarify the spectral peaks and enable peak picking with greater accuracy.

3.6.2.2 Synthetic Spectral Matching

The Synthetic Spectral Matching (SSM) method [72] detects all of the harmonic peaks and then measures the fundamental frequency by matching the original spectrum with the comb-like structure of the reconstructed synthetic voiced spectrum generated for each candidate fundamental frequency. The SSM method uses the mean squared error (MSE) spectrum similarity measure to find the best match for the original signal spectrum. The MSE measure is given by

$$E(\omega_0) = \frac{1}{2\pi}\int_{-\pi}^{+\pi}\left|S(\omega)-\hat{S}(\omega,\omega_0)\right|^2 d\omega \qquad (3.35)$$

The value of the fundamental frequency, ω_0 (in radians), must be of a high precision as even a small deviation from its true value can severely degrade the MSE matching process. $S(\omega)$ represents the original windowed speech spectrum, given by

$$S(\omega) = \sum_{k=-\infty}^{\infty} s(k)w(k)e^{-j\omega k} \tag{3.36}$$

$\hat{S}(\omega,\omega_0)$ is the reconstructed, pitch dependent spectrum, which is an approximation to the convolution of an impulse train, spaced at ω_0 intervals, with the spectrum of the windowing function over $N=2M+1$ points, as given by

$$\hat{S}(\omega,\omega_0) = \sum_{k=-M}^{M} A_k(\omega_0)W(\omega - k\omega_0) \tag{3.37}$$

where the transformation of the windowing function, $W(\omega)$, is defined as

$$W(\omega) = \sum_{k=-\infty}^{\infty} w(k)e^{-j\omega k} \tag{3.38}$$

The MSE harmonic amplitudes are given in Equation 3.39, which represents the cross-correlation of the speech spectrum with the harmonic lobe spectrum, divided by the auto-correlation of the harmonic lobe spectrum.

$$A_k(\omega_0) = \frac{\int_{(k-0.5)\omega_0}^{(k+0.5)\omega_0} S(\omega)W^*(\omega - k\omega_0)d\omega}{\int_{(k-0.5)\omega_0}^{(k+0.5)\omega_0} |W(\omega - k\omega_0)|^2 d\omega_0} \tag{3.39}$$

The limits $(k-0.5)\omega_0$ and $(k+0.5)\omega_0$ define the calculation range of a harmonic interval.

The obvious advantage of the SSM method is that it can offer a fundamental frequency search with very fine resolution. However, a full search over all possible pitch frequencies is computationally too demanding. Therefore, this method is usually used as a pitch refinement technique in the second stage of the PDA where it refines the initial estimate of the pitch value. The major problem that the SSM method encounters is the potential incorrect selection of pitch period multiples and sub-multiples. This problem can be solved by imposing restrictions upon the harmonic magnitudes, so that they may not fall far below the speech spectral envelope. This is the basis of the Sinusoidal Model Matching PDA which is

described in Chapter 5, where it is used as a part of the Amplitude Excited Split-Band LPC vocoder.

3.6.3 Pre- and Post-processing Techniques

Pre- and post-processing techniques can significantly improve the accuracy of a PDA. Their function is to supplement the chosen PDA before and after the pitch detection stage respectively.

Spectrum flattening is the most dominant pre-processing stage used as part of many PDAs in low bit rate speech coding. The function of the spectrum flattening method is to remove the formants of the original speech, which can affect the accuracy of the PDA, so that the pitch detection is performed on the residual spectrum. The linear spectrum flattening is achieved using the LPC inverse filter, as described in Section 3.4.3. Figure 3.11 depicts the effect of linear spectrum flattening.

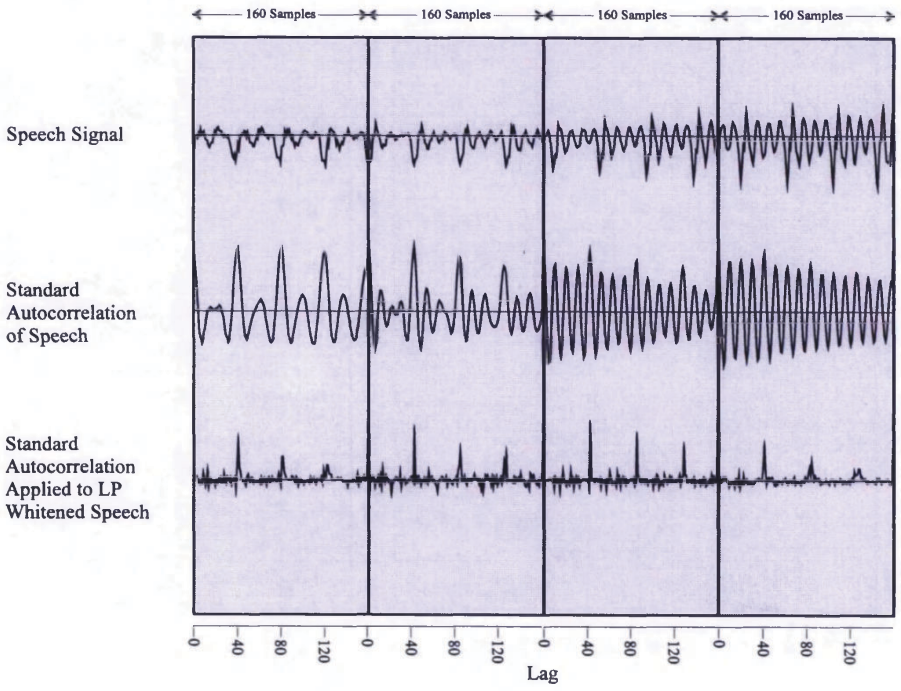


Figure 3.11 The effect of linear spectrum flattening

Although the formants are removed and the peaks are narrow, the auto-correlation output is much noisier than before the pre-processing. This can become a major problem when there is a ramping pitch period that effectively reduces the size of the peaks and worsens the pitch detection. Another mode of failure of this technique is for high-pitch speech, whereby the inverse LPC filtering, which is simply an energy minimisation procedure, can place a zero at the fundamental frequency. A solution to this problem is the use of bandwidth expansion [64] of the LP coefficients prior to filtering, in order to control the level of spectrum flattening.

Non-linear spectrum flattening is usually achieved by centre clipping the speech signal. There are several centre clipping combinations which display varied performance, especially for low-pitch speakers. The overall performance can be improved if the non-linear processing stage is placed before low-pass filtering. The main problem in applying the centre clipping is the choice of the clipping threshold. For more information on this subject see [73].

Pitch tracking is based on the continuity characteristic of pitch. For speech, in general, it was shown [74] that pitch period does not usually fluctuate by more than 25% from one frame to the next. This percentage would decrease even further if silence and unvoiced regions were excluded from the measurements. The pitch tracking can be used either before the PDA, in which case it is a pre-processing technique, or after the PDA, in which case it is a post-processing method also known as pitch smoothing. As a pre-processing technique, the pitch tracking uses the speech information, such as pitch history, voicing history and speech power history, from the neighbouring frames in order to estimate the pitch in the current frame in a global sense of a minimum path error. As a post-processing technique, the pitch smoothing forces the pitch contour to be smooth by utilising the continuity characteristics. The key task of the pitch tracking mechanism, both at the pre- and post-processing stages, is to prevent the occurrence of the single event pitch errors by keeping the pitch fluctuations within a narrow range. Unfortunately, this can occasionally lead to smoothing out of abrupt changes in pitch or extending the pitch sub-multiple error from the beginning of a word. These cases are however very rare and cannot significantly reduce the reliability and accuracy of this technique.

3.7 Interpolation

3.7.1 General Principles

The process of encoding and decoding speech is performed on segments of speech usually in the range of 20-30 ms. The limitation on the update frame size is based on the assumption that speech characteristics remain stationary within a given frame size. However, this assumption does not always hold, as speech continuously evolves from one frame to another. This is particularly pertinent to the spectral envelope, which varies continuously with time. One of the main design issues in low bit rate speech coding is the ability of a given coder to capture and model the majority of speech events and to maintain the continuity of the modelling parameters, especially that of the spectral envelope. One possible option in order to solve this issue would be to increase the update rate of the spectral parameters, although it would inevitably lead to higher bit rates. The other possible way is to introduce a variable frame length scheme, whereby the frame size is made dependent on the length of the actual speech event (see Chapter 7). However, whether it is a fixed rate choice or a variable rate choice, they both still have to overcome the problem of speech continuity beyond the frame size, i.e. they have to allow speech to progress 'naturally' from one frame to another. This process of 'smoothing' out the changes in modelling parameters between frames is called *interpolation* and is used widely in low bit rate speech coding.

Any interpolation algorithm is based on the time delay between the encoding process and decoding process. The decoder has to wait for the encoder to transmit two sets of LPC analysis windows with the corresponding modelling parameters associated with the beginning and the end of a frame, before it proceeds with the synthesis. The synthesis algorithm is based on the intermediate set of parameters that are calculated at any position within a decoded frame using interpolation. The time delay is typically one half of a frame size, although it can be longer for variable frame length schemes.

The main task of interpolation is to reduce the step change in LPC (or LSF) parameters as well as in the other modelling parameters. This reduction is particularly important for voicing transition regions (onsets and offsets), where non-interpolated speech can give rise to

perceptually disturbing, harsh unvoiced sounds. The effect of interpolation at an offset voicing transition can be seen in Figure 3.12. The non-interpolated synthetic speech dramatically changes its frequency content, as it displays much more high frequency energy, especially during the transition. On the other hand, the interpolated synthetic speech gradually changes its frequency content from low to high frequency energy. The difference between the two waveforms, particularly at the transition point, is highlighted in Figure 3.12 using a dashed rectangle.

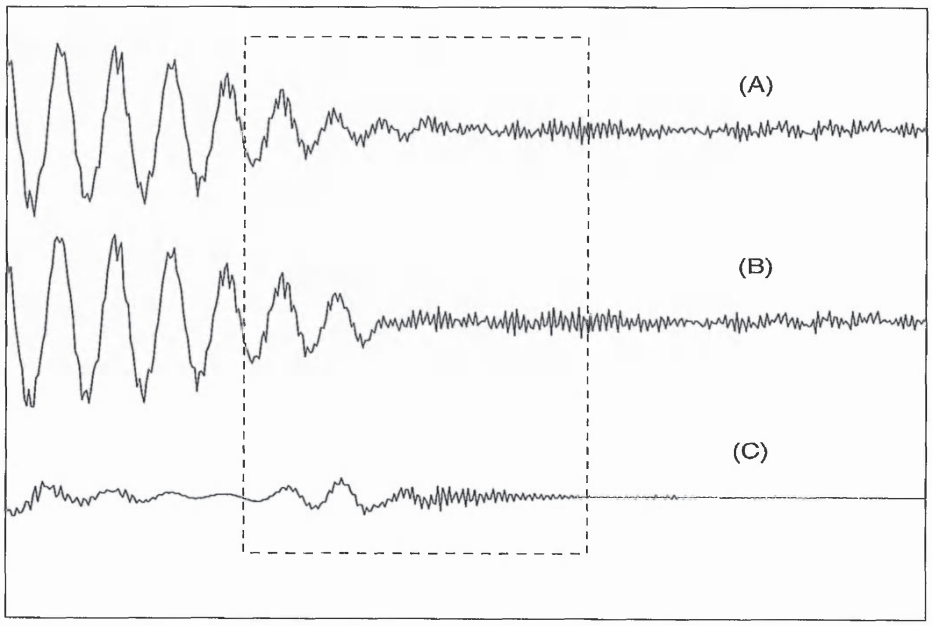


Figure 3.12 The effect of interpolation at the speech offset.
(A) Interpolated waveform (B) Non-interpolated waveform (C) Difference waveform

Parameter interpolation not only improves the quality of speech at voicing transitions, but also that of highly periodic voiced speech, during which, any sudden change in waveform shape produces highly audible 'click' and 'pop' sounds. An example of a non-interpolated versus interpolated waveform is shown in Figure 3.13, where the application of interpolation produces a smoothly varying waveform shape. The areas within the dashed rectangles display strong differences between the interpolated and non-interpolated waveforms, as highlighted by the difference waveform.

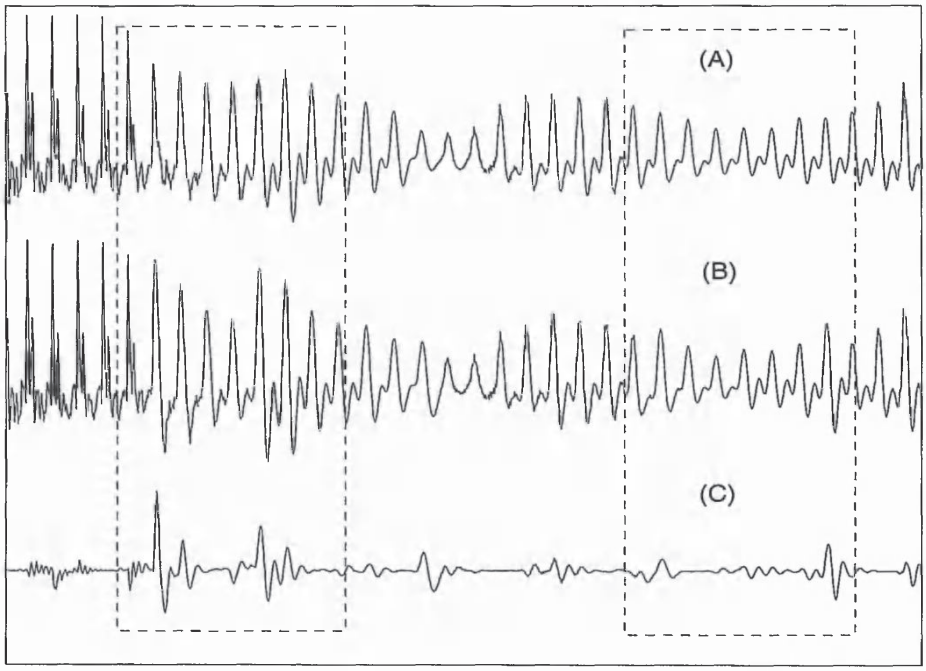


Figure 3.13 The effect of interpolation during steady state voiced speech.
(A) Interpolated waveform (B) Non-interpolated waveform (C) Difference waveform

As already discussed in Section 3.4.4, the quantisation of LP coefficients is performed in one of the alternative domains, due to a range of design problems that quantising LP coefficients can create. Consequently, the interpolation of LP coefficients is also performed in the alternative domain. Whichever interpolation domain is chosen, the following set of design criteria must be met:

- guaranteed filter stability in both starting and terminating parameter sets
- low sensitivity to the parameter fluctuation
- constant sensitivity throughout the parameter range
- linear perceptual change.

The domain of LSF is frequently chosen for interpolation purposes, as it exhibits all the aforementioned design characteristics. In addition, the LSF parameters are already available at the decoder for use in interpolation in many speech coders, including the algorithms presented in this thesis. The stability of the LSF domain is guaranteed by the monotonically increasing LSF parameters. Linear interpolation between two stable sets of LSF parameters will always produce a stable intermediate filter, since the two input sets of LSF differences

are positive and any linear interpolation therefore results in all positive differences, i.e. a stable filter. The sensitivity of the LSF domain is adequately constant throughout parameter change. However, precautions must be taken to keep the adjacent LSFs apart by a distance which is more than 40 Hz [70] to avoid the non-linear sensitivity of the filter.

3.7.2 Linear Interpolation Function

The linear interpolation function must meet the following two conditions:

$$z_j = ax_j + by_j$$
$$a + b = 1 \tag{3.40}$$

where z is the linear combination of the two functions, a and b. If x_j represents a set of LSF parameters at the beginning of a frame and y_j represents a set of LSFs at the end of a frame on the decoder side, then z_j represents the set of intermediate LSFs that characterise the spectral envelope of the interpolated frame. The interpolation may be performed once per frame sample, although, in order to reduce the computational complexity, the interpolated LSFs are usually calculated on a sub-frame basis, where sub-frame size usually lies between 20 and 40 samples.

The main disadvantage of the linear interpolation method is that it does not account for the energy of the original speech signal, i.e. the amount of interpolation with no energy consideration (per frame sample or per sub-frame) remains the same from one frame to another. An example of an energy unaided linear interpolation is depicted in the top waveform of Figure 3.14 for a typical onset transition, where equal bias is given to each parameter set on both sides of the interpolated transition frame. As a result, the onset is dominated by the presence of high frequency energy, producing a very harsh, perceptually noticeable transition. Figure 3.14 shows that when the bias in the LSF interpolation is made energy dependent, the obtained waveform matches the original speech more closely resulting in higher perceptual quality at the transition. Generally, the interpolated LSF coefficients tend towards the spectrum of the frame with the higher energy content. The energy information represents one of the essential vocoding parameters and it must therefore be transmitted from the encoder to the decoder either in time domain or frequency domain.

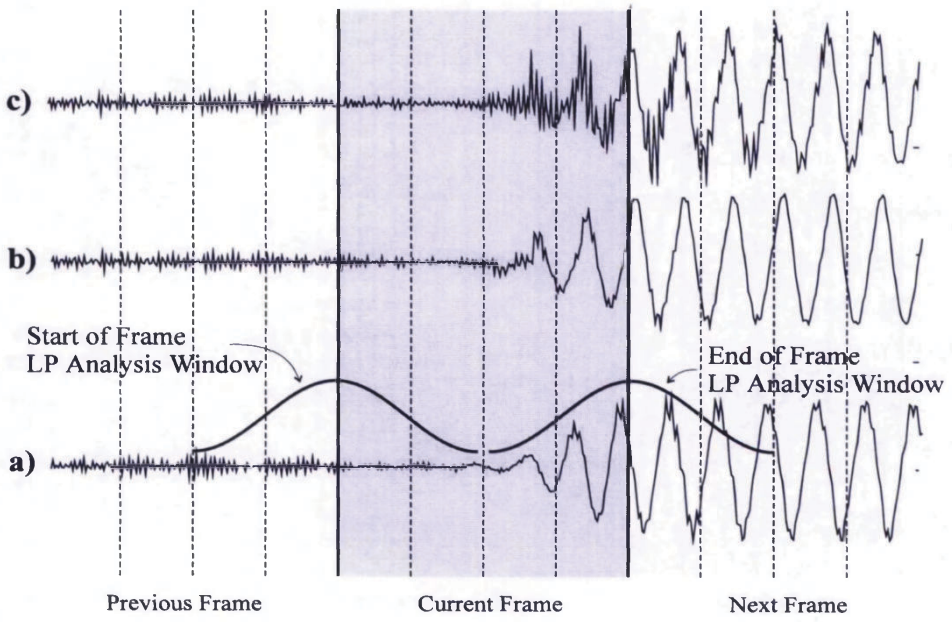


Figure 3.14 Comparison of Linear LSF interpolation (top waveform **c**) and Energy Dependent LSF interpolation (middle waveform **b**) versus Original Speech (bottom waveform **a**)

Figure 3.15 depicts a plot of the energy dependent linear interpolation function for various relative energies. The function is monotonic and in the range 0 to 1.

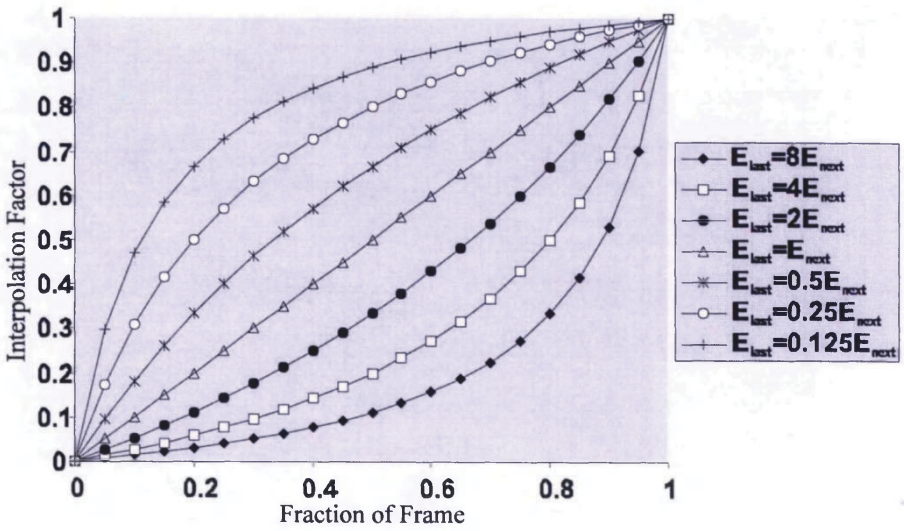


Figure 3.15 Energy Dependent LSF interpolation

The energy interpolation factor is used to modify the initial linear interpolation functions a and b in Equation 3.31. If $a=1-b$, then new energy dependent b_{mod} is given by

$$b_{\text{mod}} = \frac{E_{next}}{E_{next} + E_{last}} \tag{3.41}$$

where the interpolation is performed on either a sample basis or a sub-frame basis.

3.8 Concluding Remarks

The majority of current low bit rate speech coders, such as vocoders also presented in this thesis, use the parametric model of the human speech production mechanism. The aim of this parametric model is to reduce the overall bit rate whilst retaining the high perceptual quality of processed speech. This chapter has presented a range of coding techniques which are essential to achieve that aim.

The most powerful bit rate reduction tool is the *LPC analysis*, which is used to represent the short-term correlation in the speech signal. The LPC analysis is block based in order to facilitate a computation of a realisable DSP algorithm. The main assumption used for this analysis is that the speech signal is considered stationary. This is, however, not true for many cases where inaccuracies in the calculation of the vocal tract response can lead to undesirable spectral distortions. In order to reduce the effect of these spectral distortions and to maintain high perceptual quality of the overall parametric model, other speech coding tools, such as *post-filtering* and *interpolation*, are used.

Special consideration has been given to the notion of parameter *quantisation*, particularly that of quantising the LSF parameters. The LSF parameters are closely related to the speech spectral characteristics, in particular the formants, and therefore represent an alternative way of quantising the LP coefficients. The key advantages of the LSFs are their inherent robustness to channel errors and increased quantisation efficiency.

The long-term correlation in speech is detected by the *pitch prediction* mechanism. The principle techniques used in pitch prediction, both in time and frequency domain, have been

presented together with their respective merits and disadvantages. In the design and development of the vocoding systems presented in this thesis, both the time domain and frequency domain approach to pitch prediction have been explored and utilised.

Chapter 4.

4. Vocoding Principles and Standards

4.1 Introduction

The primary aim in vocoding is to parameterise the speech production model which was described in the previous chapter. The vocoder modelling and parameterisation can take different approaches, which lead to different classes of vocoding. However, there is a common underlying principle, which has remained unchanged since the inception of the vocoding idea in 1928 [75], shared by all vocoder types. This common principle of vocoding is to extract the modelling parameters from the following speech features:

- vocal tract characteristics
- periodicity
- phonetic status (such as voicing information).

This chapter gives a brief overview of various vocoding algorithms with particular attention paid to the Improved Multi-Band Excitation (I-MBE) vocoder, on whose similar structure

most of the speech coding work in this thesis is based. An outline of the current vocoding contributions to the world-wide speech coding standardisation is given at the end of the chapter.

4.2 Channel Vocoder

The first channel vocoder was introduced at the World Fair in New York in 1939 [76]. Its initial design was carried out using analogue circuitry and, generally, very limited hardware technology, which resulted in poor synthetic speech quality. However, it was one of the first speech coding methods that managed to significantly reduce the bit rate. As a result, it was soon put into operation to efficiently transmit encrypted speech during the Second World War. With the advent of modern technology in the last 60 years, especially the introduction of transistors and digital circuits, the output rate of the channel vocoder has further fallen whilst the speech quality has improved.

Channel vocoding is based on the principle of modelling the spectral envelope by a bank of bandpass filters, the number of which can range from 14 to 38, each of about 100-300 Hz bandwidth. In doing so, the vocoder neglects the spectrum phase information, since the human hearing mechanism is considered to be largely insensitive to the short-term spectrum phase change. Figure 4.1 depicts the encoder and decoder sides of the channel vocoder. The excitation, selected by the voicing decision, consists of a noise generator and a pulse generator, whose period is controlled by the determined pitch. The coding rate of the channel vocoder is typically around 2.4 kbps, although it can be reduced to as low as 1.2 kbps [69].

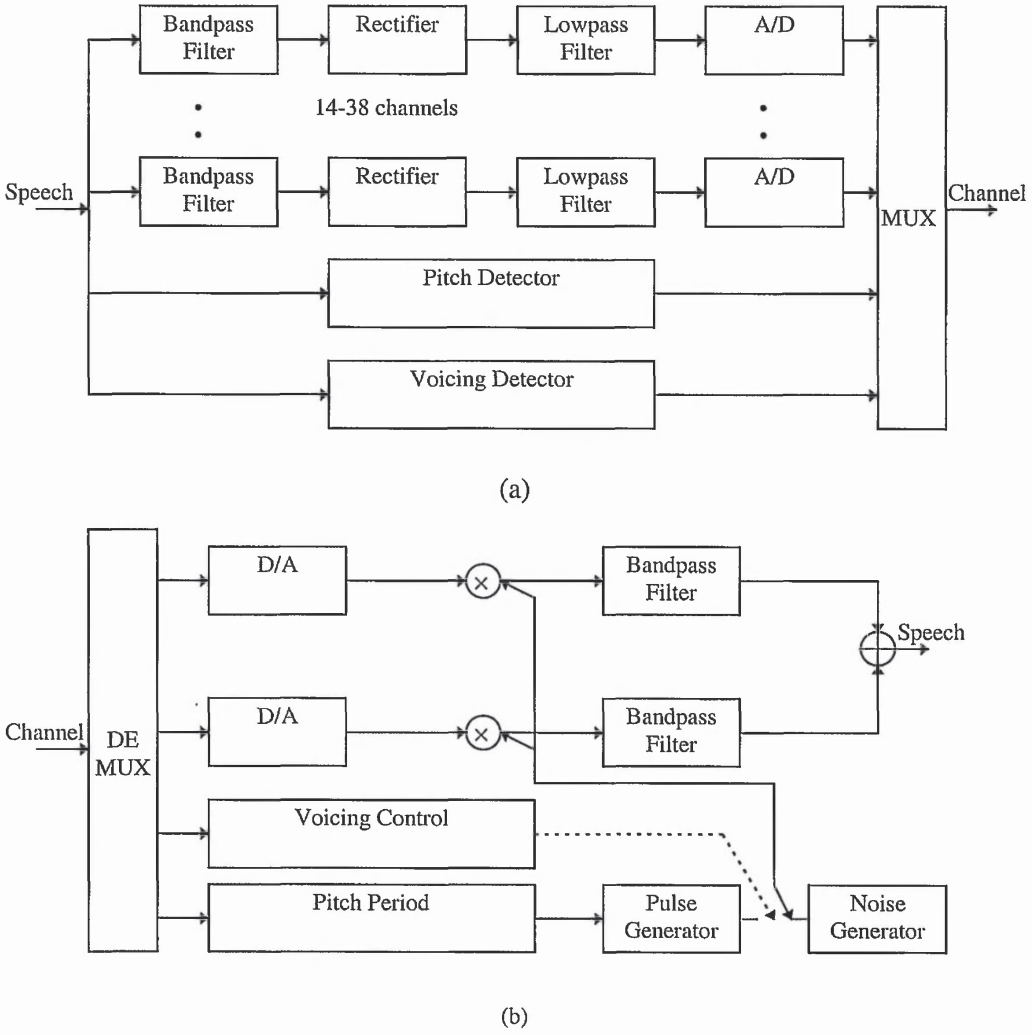


Figure 4.1 Channel Vocoder Structure - (a) Encoder and (b) Decoder

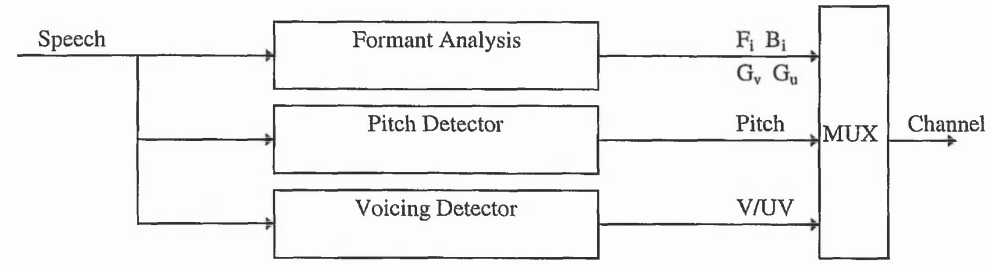
4.3 Formant Vocoder

A formant vocoder [77] represents a variation on the channel vocoder. Unlike the channel vocoder which splits the spectrum into several frequency bands, the formant vocoder identifies and quantises only the spectral formants. As there are usually 3-4 formants present in the spectrum, a significant saving in the number of bits needed for the spectral envelope quantisation is achieved with respect to channel vocoding, resulting in typical operating rates

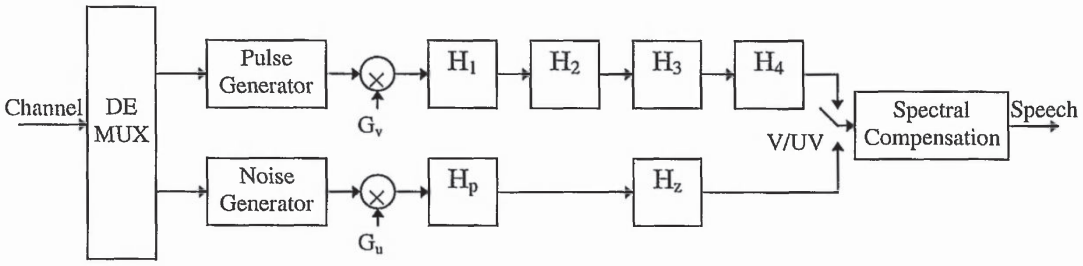
below 1 kbps. Normally, formants are most meaningful for voiced speech, however, their use can also be extended to representing unvoiced speech. For unvoiced speech, the spectral envelope can be modelled using only two filters, each of which contains a complex pole pair and a complex zero pair. Figure 4.2 depicts the encoder and decoder parts of a typical formant vocoder. The shape of each formant is modelled by a digital resonator characterised by the transfer function $H(z)$, given by

$$H_i(z) = \frac{1 - 2e^{-B_i T}\cos(\omega_i T) + e^{-2B_i T}}{1 - 2e^{-B_i T}\cos(\omega_i T)z^{-1} + e^{-2B_i T}z^{-2}} \tag{4.1}$$

The output perceptual quality of the formant vocoder often exceeds that of the channel vocoder due to its formant shaping approach which more closely resembles the original spectrum. However, the key problem in this approach is the inherent difficulty in providing a reliable and accurate method for detecting and tracking the formant frequencies. Both cepstral analysis and LPC analysis have been used [78] to combat this problem and improve the output synthetic quality.



(a)



(b)

Figure 4.2 Formant Vocoder Structure - (a) Encoder and (b) Decoder

4.4 Homomorphic (Cepstral) Vocoder

The function of the cepstral analysis in speech coding is to separate the linearly convolved speech signal into two parts: the excitation $E(\omega)$, and the vocal tract filter response $H(\omega)$. The homomorphic vocoder [79] applies a logarithm operation to this convolution function, as given by

$$\log|S(\omega)| = \log|E(\omega)H(\omega)| = \log|E(\omega)| + \log|H(\omega)| \tag{4.2}$$

where both the excitation $E(\omega)$ and the vocal tract $H(\omega)$ spectra are complex functions.

Taking the logarithm of a complex function inevitably leads to phase ambiguities, which is one of the main problems of the homomorphic approach. A process known as phase unwrapping must be applied to ensure that the phase function progresses smoothly and without discontinuities [79]. Since the human auditory system is largely insensitive to phase, the phase problem can be avoided by using only the real spectral magnitudes for calculation purposes. However, the removal of the phase information also means the removal of much of the temporal information contained in the speech signal thus reducing the accuracy of the whole process.

The logarithmic spectra of the excitation and the vocal tract are inverse Fourier transformed to give two separated *quefrency domain* signals, which form the *cepstrum* of the speech signal $C(t)$, as given by

$$C(t) = C_E(t) + C_H(t) = F^{-1}\{\log|E(\omega)|\} + F^{-1}\{\log|H(\omega)|\} \tag{4.3}$$

In the quefrency domain, the excitation cepstrum and the vocal tract cepstrum are typically separated. The former represents the rapidly-varying spectral details and is placed in the upper half of the speech cepstrum. The latter represents the slow-varying spectral envelope characteristics and is situated in the lower half of the speech cepstrum.

The homomorphic vocoder applies the low-quefrency window in the cepstral domain, thus, filtering out the vocal tract information, which is then encoded using scalar quantisation. The encoder transmits this quefrency information together with the extracted pitch and voicing decision. At the decoder, an inverse process is applied to the quefrency information to obtain

the vocal tract filter, used to synthesise speech by shaping the generated excitation. Figure 4.3 shows the block diagram of this process.

The homomorphic vocoders have shown to perform well at rates as low as 2.4 kbps [79]. However, they are rarely a preferred vocoder choice nowadays due to the severe performance degradation in the background noise conditions; the additive noise further reduces the accuracy of the cepstral deconvolution.

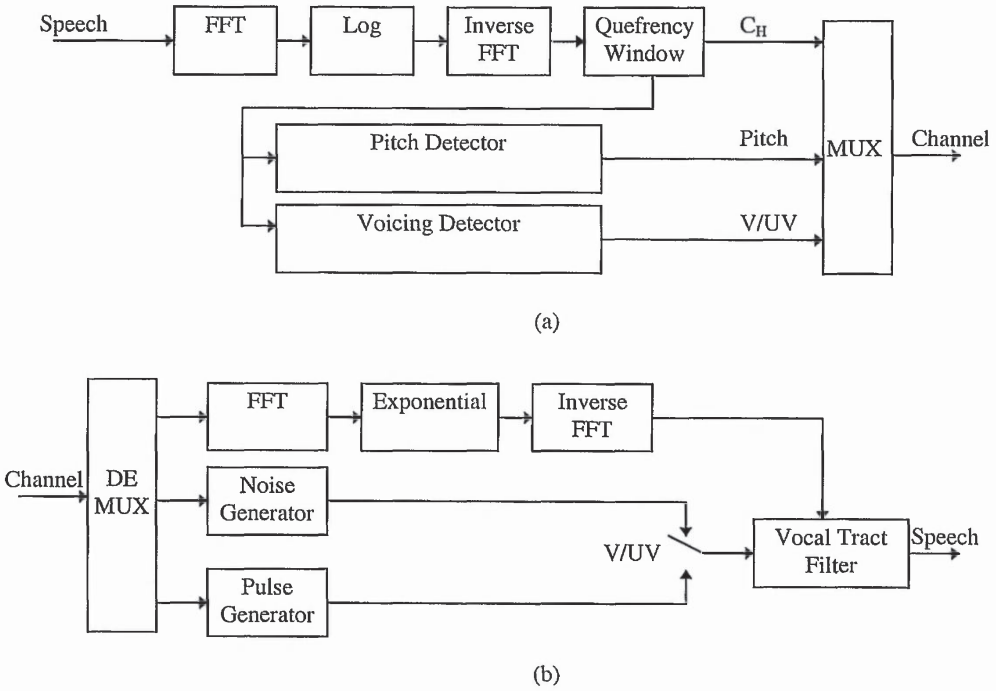


Figure 4.3 Homomorphic Vocoder Structure - (a) Encoder and (b) Decoder

4.5 LPC Vocoder

The operation of the LPC vocoder is based on the LPC analysis, which was described in detail in the previous chapter. The key function of this vocoding system is modelling the speech spectral magnitude information using a linear predictive filter. Generally, an all-pole filter is employed to represent the spectrum as it offers a significant reduction in computing the filter coefficients. Figure 4.4 depicts the block diagram of the LPC vocoder model.

The LPC vocoders have been shown to produce high intelligibility at bit rates of 2.4 kbps and below. The first widely used LPC vocoding algorithm to be standardised was the U.S. Federal Government Standard, Fed-Std 1015, [35] which was based upon 10th order LPC analysis at a bit rate of 2.4 kbps.

The two major drawbacks of the LPC vocoders are the difficulty in coding the nasal sounds, as they are characterised by a null in the spectrum which the all-pole filter can only approximate with the excess of poles, and the poor performance in the presence of additive noise. Produced mechanical or buzzy sounds arise from the inability of a simple pulse train to reproduce all kinds of voiced speech.

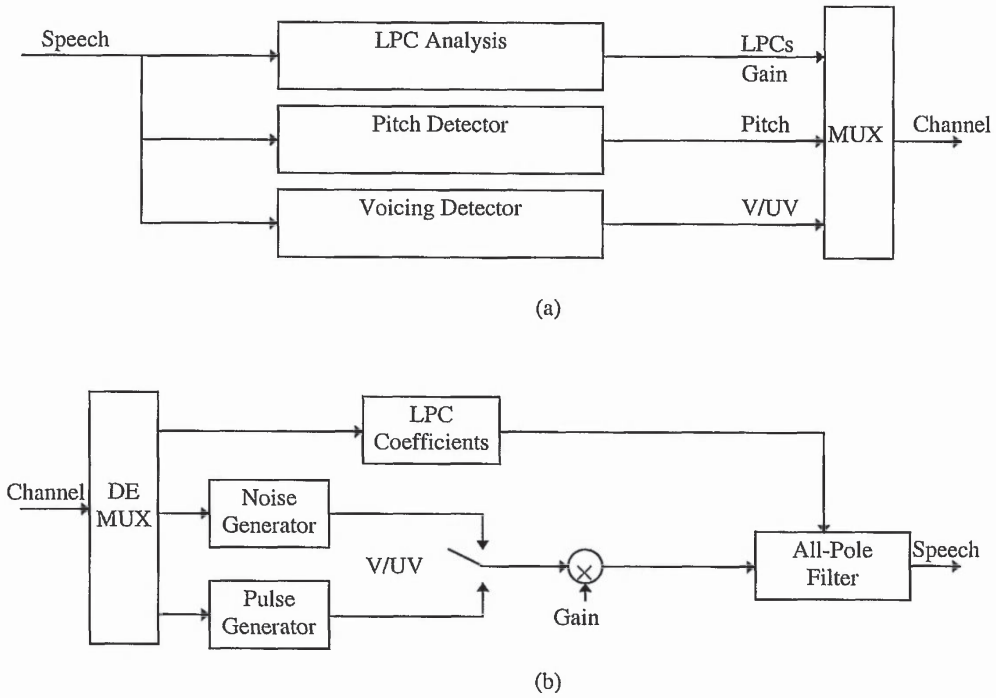


Figure 4.4 LPC Vocoder Structure - (a) Encoder and (b) Decoder

One of the recently standardised LPC-type vocoder is the 2.4 kbps Mixed Excitation Linear Predictive (MELP) vocoder [80]. The MELP vocoder is based on the traditional LPC parametric model, but also includes four additional features: mixed excitation, aperiodic pulses, pulse dispersion and adaptive spectral enhancement. The mixed excitation model produces more natural sounding speech, because it can represent a richer ensemble of

possible speech characteristics. The MELP is particularly robust in difficult background noise environments, such as those encountered in commercial and military communication systems.

4.6 Multi-Band Excitation Vocoder

4.6.1 Introduction

The Multi-Band Excitation (MBE) vocoder, first introduced in 1987 [72], like all other model-based vocoder types, attempts to construct a synthetic speech signal that contains the same perceptual information as an original speech signal. The two primary advantages of the MBE vocoder over other vocoders are the more robust multi-band excitation speech model and the use of more sophisticated algorithms to estimate the speech model parameters.

The major difference between traditional speech models and the MBE speech model is the extraction and representation of the excitation signal. In conventional speech models a single voiced/unvoiced decision (V/UV) is used for each speech segment. In contrast, the MBE speech model divides the excitation spectrum into a number of non-overlapping frequency bands and makes a V/UV decision for each band. This allows the excitation signal to be a mixture of periodic (voiced) energy and noise-like (unvoiced) energy. With this added degree of freedom in modelling the excitation, the MBE can generate higher quality speech and be more robust in the presence of background noise than the conventional speech models.

In the MBE speech model the excitation spectrum is obtained from the pitch period (or the fundamental frequency) and the V/UV decisions. The V/UV determination is performed such that frequency bands are declared voiced where the ratio of periodic energy to noise-like energy is high, otherwise they are declared unvoiced. A periodic spectrum is used in the frequency bands declared voiced, while a random noise spectrum is used in the frequency bands declared unvoiced. The periodic spectrum is generated from a windowed periodic impulse train, whereas the random noise spectrum is generated from a windowed random noise sequence.

The main principle of the MBE vocoder, its multi-band voicing decision, proved to be capable of producing very high quality speech at very low bit rates. This section overviews the principles of the multi-band excitation vocoding technique as adopted by the Improved MBE (I-MBE) vocoder, proposed in 1991 [81], operating at 4.15 kbps. The I-MBE vocoder is similar to the MBE vocoder except that in the I-MBE the phase information of the harmonics was found to be redundant as it could be estimated on the decoder side using a simple phase generator. The I-MBE vocoder can be seen as one of the starting points in this research aimed at producing good to high quality vocoders at rates around 2.4 kbps.

4.6.2 I-MBE Encoder

The I-MBE excitation and spectral envelope parameters are estimated simultaneously, so that the synthesised spectrum is closest in the least squares sense to the original spectrum. A block diagram of the analysis algorithm for the I-MBE encoder is shown in Figure 4.5.

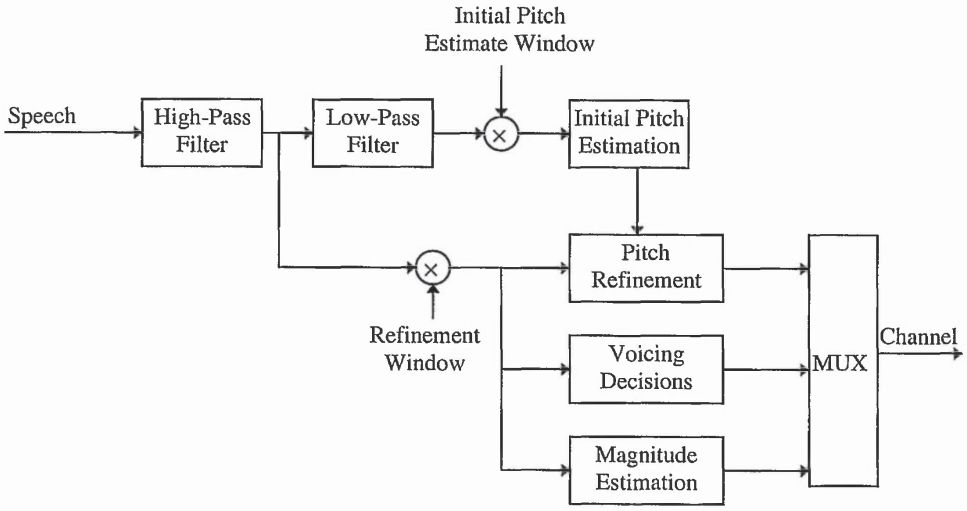


Figure 4.5 The block diagram of the encoder analysis model

The I-MBE speech model parameters estimated for each speech frame are:

- pitch period or fundamental frequency
- voicing decisions
- spectral amplitudes of the spectral envelope.

4.6.2.1 Pitch Estimation

The pitch estimation algorithm attempts to preserve some continuity of the pitch between neighbouring speech frames. When determining the pitch of the current frame, a pitch tracking algorithm deployed considers the pitch from previous and estimated future frames.

The pitch is determined using a two-step procedure. First, an *initial pitch estimate* is obtained. It is restricted to be a member of the set (21, 21.5,...,113.5,114 samples). It is then refined and the final estimate of the fundamental frequency is obtained, which has a ¼ sample accuracy. The *pitch refinement* method is based on the Synthetic Spectral Matching model described in Chapter 3. This two-part procedure is used in part to reduce the computational complexity, and in part to improve the robustness of the pitch detection.

Figure 4.6 shows the initial pitch estimation model. For each speech frame two different pitch estimates are computed. P_b is the backward estimate which maintains pitch continuity with previous speech frames. P_f is a forward estimate which maintains pitch continuity with the future speech frames. The backward pitch estimate is calculated with the look-back pitch tracking algorithm, while the forward pitch estimate is calculated with the look-ahead pitch tracking algorithm [81]. These two estimates are compared with the set of decision rules, and one of them is chosen as the initial pitch estimate.

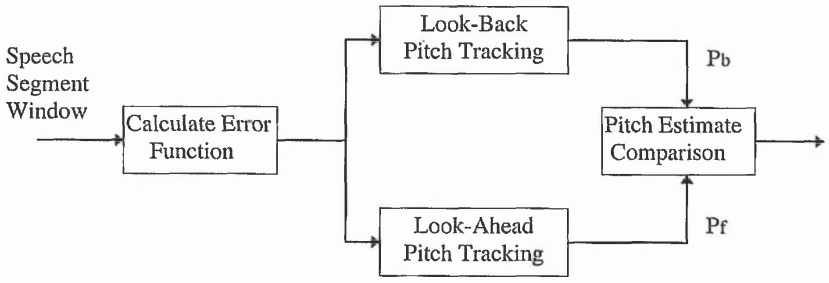


Figure 4.6 Block diagram of the initial pitch estimate

4.6.2.2 Voicing Decision

The voiced/unvoiced decisions are found by dividing the spectrum into K frequency bands and evaluating a voicing measure, D_k, for each band. The number of frequency bands is a function of L, a total number of harmonics within the frame spectrum, where $1 \leq l \leq L$ and l indicates a single harmonic number.

The voicing measure D_k is given as

$$D_k = \frac{\sum_{bandsize(k)} \left| S(\omega) - S_{ref}(\omega, \hat{\omega}_0) \right|^2}{\sum_{bandsize(k)} \left| S(\omega) \right|^2} \tag{4.4}$$

where $S(\omega)$ represents the original spectrum, and $S_{ref}(\omega, \hat{\omega}_0)$ is the reconstructed spectrum for the given refined pitch frequency $\hat{\omega}_0$.

The parameters D_k, for $1 \leq k \leq K$, are compared with the threshold function Θ_k whose value depends on the set of energy parameters, such as the energy of the current segment ε_0, the local average energy ε_{avg}, the local maximum energy ε_{max}, and the local minimum energy ε_{min}. These energy parameters are updated every frame and are then used to calculate the voicing decision threshold function Θ_k for each frequency band. If D_k is less than the threshold function, the frequency band is declared voiced; otherwise, the band is declared unvoiced. A block diagram of this procedure is given in Figure 4.7.

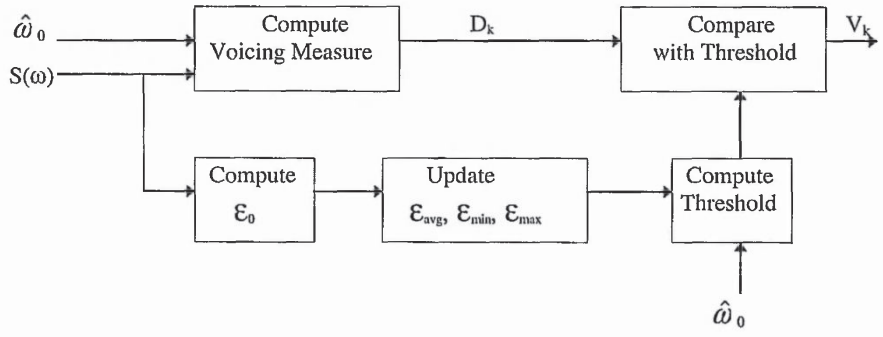


Figure 4.7 Block diagram of the I-MBE voicing routine

With the exception of the highest frequency band, each band, V_k, contains three harmonics of the refined fundamental frequency. The highest frequency band may contain more or less than three harmonics of the refined fundamental frequency. If a particular frequency band is declared voiced, then all of the harmonics within that band are defined as voiced harmonics. Similarly, if a particular frequency band is declared unvoiced, then all of the harmonics within that band are defined as unvoiced.

4.6.2.3 Spectral Amplitude Estimation

The block diagram of the I-MBE spectral amplitude estimation technique is shown in Figure 4.8. For each harmonic of the fundamental frequency a spectral amplitude, M_l, is determined in the following way.

If the frequency band is declared voiced, then spectral amplitude for each harmonic within that band is calculated as

$$M_l = \left|A_l(\hat{\omega}_0)\right| = \left| \frac{\int_{(l-0.5)\hat{\omega}_0}^{(l+0.5)\hat{\omega}_0} S(\omega)W^*(\omega - l\hat{\omega}_0)d\omega}{\int_{(l-0.5)\hat{\omega}_0}^{(l+0.5)\hat{\omega}_0} \left|W(\omega - l\hat{\omega}_0)\right|^2 d\hat{\omega}_0} \right| \tag{4.5}$$

where $S(\omega)$ refers to the windowed original speech spectrum, and $W(\omega)$ refers to the refined window spectrum.

Alternatively, if the frequency band is declared unvoiced, the spectral amplitude for each harmonic within that band is calculated as

$$M_l = \frac{\left[\frac{\sum_{(l-0.5)\hat{\omega}_0}^{(l+0.5)\hat{\omega}_0} \left|S(\omega)\right|^2}{\hat{\omega}_0} \right]^{\frac{1}{2}}}{\sum_{n=-F/2}^{F/2} w(n)} \tag{4.6}$$

where $F+1$ is the size of the refinement window.

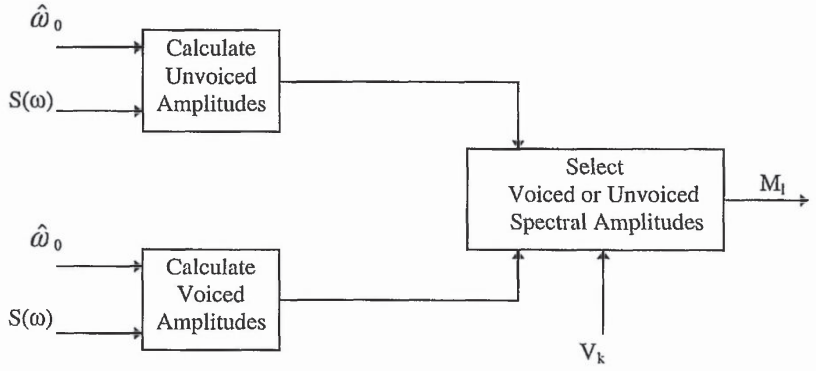


Figure 4.8 Block diagram of the I-MBE Spectral Amplitude Estimation

4.6.3 I-MBE Decoder

The I-MBE speech synthesis algorithm uses the reconstructed model parameters to generate a speech signal which is perceptually similar to the original speech signal. The model parameters consist of the fundamental frequency, the voicing decisions and the spectral amplitudes for each frequency band. For each new set of model parameters, the frame of synthetic speech is interpolated between the previous set of model parameters and the current set of model parameters. The synthetic speech signal is divided into a voiced component and an unvoiced component. These two components are synthesised separately, as shown in Figure 4.9, and then summed to form an output speech signal.

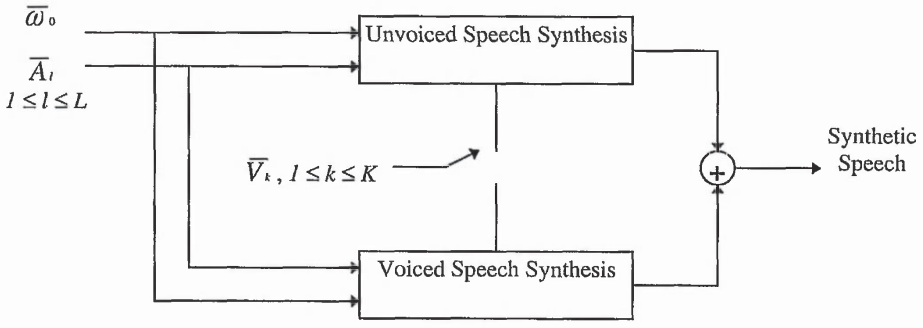


Figure 4.9 Block diagram of the I-MBE decoder

The *unvoiced part* is synthesised from the randomly generated white noise sequence. After being windowed and converted into frequency domain, the synthesised spectrum is divided into a number of voiced/unvoiced frequency bands, $\overline{V}_k$, (according to the fundamental frequency information, $\overline{\omega}_0$, from the encoder). All the voiced bands are then zeroed out, while the unvoiced bands are preserved and each unvoiced harmonic is scaled according to its respective magnitude information passed from the encoder. This spectrum is then converted back into time domain and by using the weighted Overlap-and-Add algorithm on the unvoiced speech of the previous and current segment, the unvoiced synthesis frame is generated.

The *voiced part* is synthesised by summing sinusoids for each voiced spectral amplitude. The voice synthesis algorithm attempts to match the l'th voiced spectral amplitude of the current frame with the l'th voiced spectral amplitude of the previous frame. The unvoiced bands are zeroed out. The algorithm recognises 4 different synthesis cases:

- both previous and current harmonic are declared voiced
- both previous and current harmonic are declared unvoiced
- previous harmonic is voiced and current harmonic is unvoiced
- previous harmonic is unvoiced and current harmonic is voiced.

Once the separate synthesis for both voiced and unvoiced parts take place, the two frames are merged on the sample-by-sample basis.

4.6.4 Quantisation Scheme for the MBE Vocoder at 4.15 kbps

In order for the I-MBE vocoder to operate at 4.15 kbps, 128 bits/frame are needed for its design; 45 bits are reserved for error correction, and 83 bits are divided among the model parameters (as depicted in Table 4.1). The total operating rate (including error correction) amounts to 6.4 kbps, but the effective coding bit rate is 4.15 kbps.

Table 4.1 I-MBE quantisation scheme at 4.15 kbps

Modelling Parameter	Bits/frame
Fundamental Frequency	8
V/UV decisions per band	K
Spectral Amplitudes	75 - K
Total	83

where $K \leq 12$ (K- number of frequency bands).

4.6.5 I-MBE Performance Evaluation

The I-MBE vocoder has proved to be one of the most efficient low bit rate speech coding methods, which is capable of producing high quality speech [81] and performing well in the background noise conditions. Its multi-band voicing technique renders more flexibility and better accuracy in modelling the speech parameters than most of the conventional coders. However, reconstructing the multi-band voicing spectrum still requires a large number of bits, hence, the I-MBE vocoder's quality starts to deteriorate below 4.15 kbps.

The I-MBE vocoder was used throughout the work presented in this thesis as a benchmark for low bit rate speech quality when evaluating the perceptual performance of the newly designed and developed vocoders.

4.7 Vocoding Standards

The extensive research into the area of speech coding since the introduction of the first world-wide speech coding standard, G.711 PCM, has resulted in prolific development of a variety of different coding algorithms and methods. The setting of standards has enabled the equipment manufacturers to co-ordinate their research activities and resolve the equipment compatibility issues with other manufacturers and service providers. World-wide standardisation has initiated the development of a very competitive telecommunication services market which has resulted in the lowering of product prices and a rapid take up of telecommunication products

by the general public. This high market demand has in turn called for more efficient speech coding algorithms to be designed, developed and standardised.

The major speech coding standards developed in the past twenty-seven years, since the adoption of the PCM algorithm, are shown in Table 4.2. From Table 4.2 the importance that vocoders play in present day standardisation is clearly indicated by the various MBE standards and the MELP standard adopted since the introduction of the INMARSAT-M I-MBE vocoder operating at 4.15 kbps and standardised in 1991 for land mobile satellite applications. Moreover, in the race to set new high speech quality standards targeting ever lower bit rates, the vocoder modelling approach evidently represents a formidable contender.

4.8 Concluding Remarks

This chapter has presented a brief overview of the main vocoding techniques, which have been developed in the past 60 years, and their impact on the world-wide speech coding standardisation process. The most prominent vocoder among them, the I-MBE vocoder, has been discussed in more detail, as it represents one of the starting points of the research presented in this thesis.

The key advantages of the I-MBE vocoder over the other vocoders are the more robust multi-band excitation speech model and the use of more sophisticated algorithms to estimate the speech modelling parameters. The I-MBE has shown to be capable of producing very high quality at its designated bit rate of 4.15 kbps. However, its speech quality and coding efficiency start to deteriorate rapidly with the reduction in bit rate largely due to its relatively inefficient reconstruction of the multi-band spectrum. The reasons the I-MBE represents the starting point of the research presented in this thesis lie primarily in the high value of its coding structure and also in the room for potential improvement of its performance at lower bit rates.

The research undertaken during the course of this project has predominantly focused on the development of very low bit rate narrowband vocoding algorithms operating in the region of

2.4 kbps and below. Their performance has been compared to that of the 4.15 kbps I-MBE adopted standard. The wideband vocoding algorithm, operating in the region of 8.5 kbps, has also been developed and its performance evaluated against the G.722 wideband standard.

Table 4.2 Major digital speech coding standards adopted in the last 27 years

Coding Bit Rate (kb/s)	Coding Standard	Application Area	Coding Algorithm	Year of Adopti on
64	ITU-T G.711	PSTN	PCM	1972
64,56 and 48	ITU-T G.722 (Multi-rate)	ISDN (Wideband)	SB-ADPCM	1984
32	ITU-T G.721	PSTN	ADPCM	1984
2.4	US Federal-Standard 1015	Secure Voice (Fixed and Mobile)	LPC-10e	1984
16	INMARSAT Standard-B	Maritime	APC	1985
13	Full-Rate GSM	Pan-European Digital Cellular Mobile Radio and DCS-1800	RPE-LTP	1988
4.8	US Federal-Standard 1016	Secure Voice	CELP	1989
8.9	Skyphone INMARSAT	Aeronautical Mobile Standard	MPE-LPC	1989
7.95	IS-54	North American Digital Cellular Mobile Radio	VSELP	1989
4.8	NASA MSAT-X	Mobile Satellite	VAPC	1990
16	ITU-T G.728	PSTN	LD-CELP	1991
6.7	Full-Rate PDC	Japanese Personal Digital Cellular Mobile Radio	VSELP	1991
4.15	INMARSAT Standard-M	Land Mobile Satellite	IMBE	1991
4.4	APCO	North American Police Officer Mobile Radio	IMBE	1992
4.53	TETRA	Trans-European Trunk Radio	ACELP	1993
8→1 (variable)	IS-95	North American CDMA Digital Cellular Mobile Radio	QCELP	1993
8	ITU-T G.729	PSTN	ACELP	1995
5.7	Half-Rate GSM	Pan-European Digital Cellular Mobile Radio	VSELP	1995
3.45	Half-Rate PDC	Japanese Personal Digital Cellular Mobile Radio	PSI-CELP	1995
4.8	INMARSAT Mini-M	Land Mobile Satellite (Notebook sized terminal)	AMBE	1995
13	Enhanced-FR GSM	Pan-European Digital Cellular Mobile Radio	ACELP	1995
13	PCS-1900	North American Personal Communication Systems	ACELP	1995
2.4	U.S. Federal Standard MELP Vocoder	Commercial and Military Communication Systems	MELP	1997
4.75-12.2	AMR-GSM	Adaptive Multi-Rate (AMR) GSM service	MR-ACELP	1998

Chapter 5.

5. Split Band LPC Vocoding

5.1 Introduction

Various speech coders, operating at bit rates in the region of 2.4 kbps and below, have been available for many years. However, their synthetic output has struggled to offer communication quality of speech, which has largely restricted their potential use to applications where mere intelligibility is of prime concern. The recent explosive growth in demand for mobile personal communication products has revitalised the interest in very low bit rate speech coders largely due to limited available radio bandwidth. The general public, accustomed to toll quality of the PCM telephone speech at the rate of 64 kbps, is however unlikely to accept any telecommunication system that could barely provide communication quality of service. Improving the perceptual quality of very low bit rate speech coders by making it 'near-toll' is therefore one of the key driving forces in telecommunication research, and it is in the light of these recent developments that the research into the Split Band LPC (SB-LPC) vocoding at 2.4 kbps has been undertaken.

This chapter is essentially divided into two sections according to the approaches taken in the SB-LPC vocoding. The first section describes the principles of the Time Envelope SB-LPC vocoder, covering the important issues of parameter extraction and modelling, such as pitch detection, voicing determination strategy and time envelope representation. The subject of the second section is concerned with the fundamentals of Amplitude Excited SB-LPC vocoding. Although based on the techniques already developed for the Time Envelope SB-LPC vocoder, this new vocoding technique with predominantly frequency domain functions represents an improved alternative to the time domain approach. An overview of the two vocoding strategies and objective comparison of their strengths and weaknesses are presented in this chapter.

The fixed rate Time Envelope SB-LPC vocoder provides foundations for simulating the variable quantisation rate scheme capable of lowering the average bit rate whilst maintaining the already achieved speech quality, as described in Chapter 6. Both the time domain and frequency domain vocoders are then used for designing and developing the variable frame length system, as described in Chapter 7, where the frame rate for voiced speech is reduced by up to 22% and that for unvoiced speech by up to 32%, and where the attained speech quality is higher or at least matches that of the fixed rate systems.

5.2 Time Envelope SB-LPC Vocoder at 2.4 kbps

The Time Envelope SB-LPC vocoder differs from the conventional LPC vocoders because it places more emphasis on the time domain information. Accurate modelling and preservation of the time envelope of speech, coupled with the parameter interpolation, are its key features that are capable of reproducing highly intelligible speech, in particular at speech transitions.

5.2.1 Encoder

The block diagram of the encoder section of the Time Envelope SB-LPC vocoder is shown in Figure 5.1. The speech analysis is performed on 20 ms speech frames for the operating bit rate of 2.4 kbps (at 8000 Hz sampling frequency).

The parameters that are used to represent the speech model are:

- pitch period
- split band voicing decision
- shape and gain of the time envelope
- 10th order LPC filter.

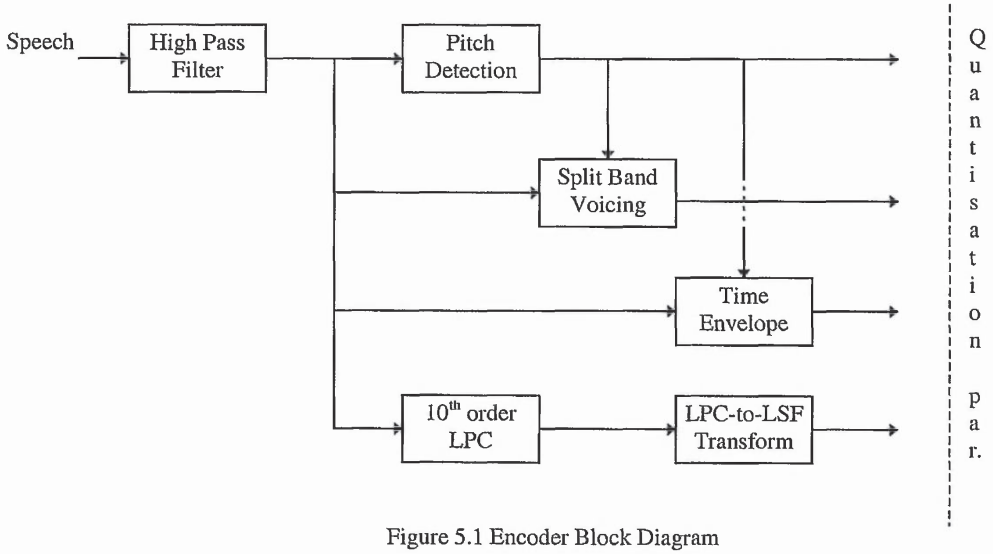


Figure 5.1 Encoder Block Diagram

At first, the original speech is passed through the high pass filter with the cut-off frequency at 50 Hz. The function of this filter is to remove any DC signal which may be superimposed upon the input speech. The filtered speech is then processed by the segmented auto-correlation pitch detection algorithm (SAPDA) using a half frame delay to allow pitch interpolation on the decoder side. The obtained pitch period, with a single sample accuracy, is refined using the synthetic spectral matching (SSM) method, as described in Chapter 3, which offers a ½ sample accuracy. The accurate pitch detection is of paramount importance for the successful encoding process as it largely influences the SB voicing decision and the work of the time envelope detector. The SB voicing decision is made once per frame of speech, whereas the shape and gain of the time envelope are calculated on the sub-frame basis, where one frame consists of eight sub-frames. The 10th order LPC filter coefficients are quantised in the LSF domain using a two stage split vector quantisation method [82]. The pitch period and the time envelope parameters are quantised using the logarithmic scalar quantiser [82].

5.2.1.1 Segmented Auto-correlation Pitch Detection Algorithm (SAPDA)

The SAPDA pitch detector is auto-correlation based, however, the auto-correlation is segmented into smaller sections whose window length and lag are determined from the local maxima and minima of the speech signal. This ensures the use of the minimum possible correlation window lengths which attains maximum temporal resolution. SAPDA has shown [82] to perform particularly well at speech onsets as it allows the pitch analysis window to be extended without increasing the usual smearing effect. The block diagram of Figure 5.2 represents this pitch detector algorithm.

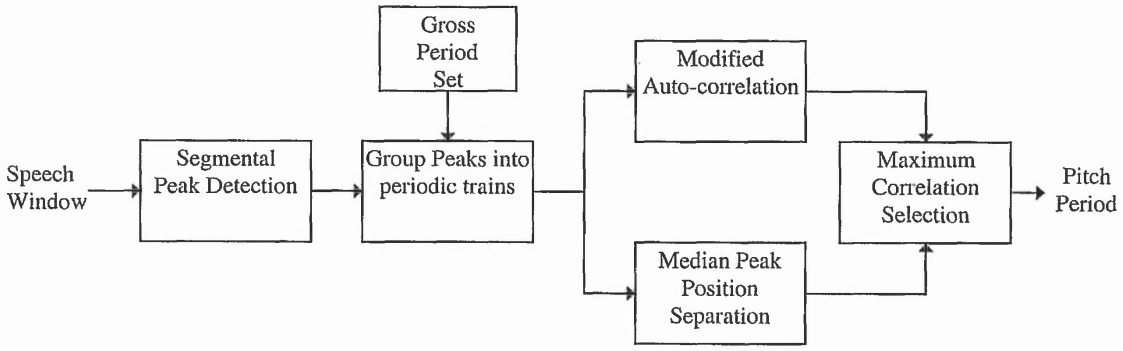


Figure 5.2 SAPDA Block Diagram

Each segmented pitch window is analysed in order to separately determine both positive and negative gross period peaks which are stored in separate lists. The peaks are determined using the following criteria:

- adjacent peaks must have a minimum separation of 5 samples
- each peak must be the largest value within ±5 samples of its position
- the smallest peak must be at least 1/20 of the amplitude of the largest peak.

The selection of gross pitch values forms an exponentially increasing set of numbers. For each selected gross pitch, a subset of peaks is determined assuming that the largest peak is included in the set and that the subset's adjacent peak separation lies within ±25% of the current gross pitch period. The latter assumption maintains the quasi-periodicity within the subset.

The segmented auto-correlation and the median peak separation are calculated for each subset. The median peak separation of a subset, whose segmented auto-correlation is greatest, represents the pitch period of the current frame, subject to pitch period multiple checking. The SAPDA algorithm operates effectively over the pitch period range of 16 to 180 samples. Figure 5.3 illustrates the segmented auto-correlation process, where the shaded rectangles represent the speech segments over which the auto-correlation and median peak separation calculations are performed.

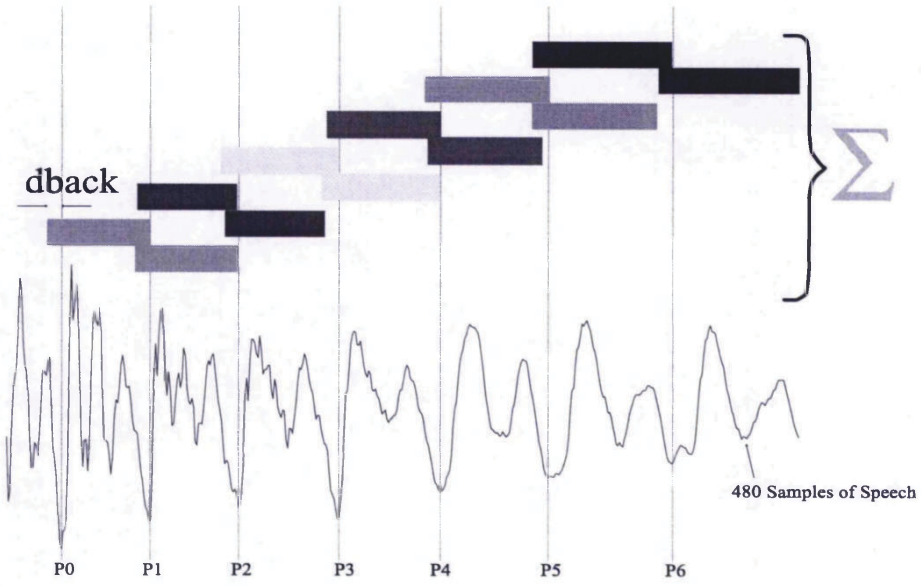


Figure 5.3 Segmented Auto-correlation

The advantages of SAPDA over a single auto-correlation are as follows:

- reduction in the blurring effect caused by the pitch change within a frame
- rejection of the outlier pitch values and better pitch contour matching
- low computational load
- reduction in formant interference.

The SAPDA is however faced with two distinct problems which stem from the time domain nature of its algorithm. Its key disadvantage is its susceptibility to background acoustic noise. Another problem with SAPDA is degraded performance when presented with speech

exhibiting mixed voicing or changing waveform shape which often leads to multiple pitch errors. Despite these problems, SAPDA has proved to be a reliable pitch detector, especially in clean speech conditions.

5.2.1.2 Split Band Voicing Technique

The SB voicing technique is similar to the frequency domain voicing scheme used by the I-MBE vocoder, where a single voicing decision is applied to blocks of three harmonics and the voiced/unvoiced (V/UV) decision is made according to the voicing threshold function, as shown in Chapter 4. Although the multi-band approach produces higher quality than the traditional single voicing per sub-frame approach, it still requires a relatively large number of bits (typically 12 bits as opposed to 3 bits used for the single voicing per sub-frame) for the V/UV quantisation. This becomes an unacceptable requirement if the target operating bit rate of the Time Envelope SB-LPC vocoder is to be around 2.4 kbps. The solution to this problem is to assemble the voicing classification in a more efficient manner, the topic which was investigated during the course of this research.

The biggest disadvantage of the multi-band approach using three harmonics per voicing band is that it imposes restrictions upon the voicing combinations, especially when there are only a few harmonics present in the spectrum. Therefore, an alternative approach is to narrow the voicing band to either a single harmonic or a pair of harmonics. During simulations the individual voicing classification was found to be unsatisfactory when speech spectrum contained only a few harmonics as it proved difficult for the MSE matching process to reliably determine the harmonic voicing decision. The paired selection, proposed earlier in [83], was adopted as a compromise between the I-MBE approach and the individual harmonic voicing classification. However, this paired approach still required a large amount of bits, so another more general approach was needed to restrict the number of allowable voicing combinations.

By observing typical speech spectra, it is readily seen that the strong harmonic structure tends to occupy lower frequency regions, conversely the upper portion of the spectrum tends to be unvoiced. This observation can be exploited by enforcing the following split band rule: all

harmonics up to a certain frequency ('frequency marker') are declared voiced and those above are declared unvoiced. This rule, which represents the essence of the SB voicing technique, determines the voiced band which lies at the highest frequency in the spectrum whose upper end represents the voicing cut-off frequency or frequency marker as depicted in Figure 5.4. In this way, the number of bits for voicing decision can be significantly reduced as only the position of the frequency marker is transmitted.

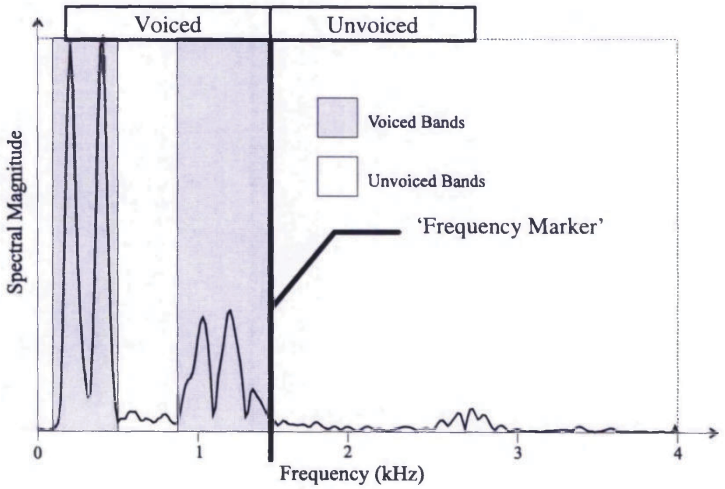


Figure 5.4 Split Band voicing spectrum

Theoretically, this method, when compared to the multi-band approach, may degrade the speech quality, since the variety of V/UV band combinations is reduced. However, in practice, this method produced very similar speech quality and was even cleaner [84]. For example, in Figure 5.4, redeclaring the second frequency band as voiced will not greatly affect the speech quality. The output speech will be slightly more voiced, but this will perceptually be difficult to distinguish. By rule of thumb, giving the speech more voiced character is the perceptually preferred choice, although the speech can sound slightly metallic, than making it noisier, which inevitably gives rise to the perceptually disturbing hoarse sounding speech.

However, simply declaring the highest frequency voiced band as the frequency marker does not provide an optimum solution to the SB approach. Unlike the case illustrated in Figure 5.4, where a conversion of only one unvoiced band into a voiced band does not produce much adverse effect, other cases, where more unvoiced bands are converted into voiced bands,

show a much more disturbing effect of metallic sounding speech. The key problem in the SB approach is therefore the attainable degree of accuracy when determining the frequency marker in the spectrum. Having the voiced harmonics scattered around the spectrum means that just a single inaccurately detected voiced band at very high frequency can dramatically shift the cut-off frequency. A similar case occurs when no voiced harmonics appear in the lower part of the spectrum. Conditions have to be made therefore to remove all isolated voiced bands in the higher frequencies, and to make an entire frame unvoiced in cases where the lower part of the spectrum is entirely unvoiced.

The method adopted for determining the frequency marker was suggested in [83], which proved to perform significantly better than the previously suggested methods [85]. In this method, an 8-level linearly varying voicing decision codebook is used to evaluate the voicing sum of the weighted original speech harmonic amplitudes. Each level marks the point in the spectrum up to which the declared voiced harmonic amplitudes are added to the voicing sum and the declared unvoiced harmonic amplitudes are subtracted from the voicing sum. Above this marker the process is reversed. The quantised single frequency marker becomes the level with the highest calculated voicing sum, i.e. with the highest energy ratio with respect to the original spectrum.

5.2.1.3 Time Envelope based on RMS Energy

Equation 5.1 represents the RMS energy of the time envelope calculated over the length of a pitch period, P. It is important to use the whole pitch period as the window length for calculating the time envelope RMS energy in order to preserve the overall energy balance of the pitch cycle.

$$e(n) = \sqrt{\frac{1}{P}\sum_{j=0}^{P}\left[s\left(n + j - \left\lfloor \frac{P}{2} \right\rfloor\right)\right]^2} \tag{5.1}$$

The time envelope coding is performed using an MSE gain/shape codebook [82]. Firstly, the envelope is decimated to N samples per frame using an averager as the decimation filter. The function $e(n)$ is evaluated at the decimated sampling positions. The optimum size of the decimated sub-frame used is 20 samples (2.5 ms), which represents a balance between the

envelope resolution and the required size of the quantiser tables. A 5-bit shape quantisation table (32 vectors) places an emphasis upon speech transitions. The gain quantisation requires 7 bits per frame.

5.2.1.4 Determining and Quantising LP Coefficients

The 10th order LPC filter is based on Durbin's analysis which is performed on the pre-emphasised speech using a 221-sample Kaiser window (β=6.0). The LP coefficients are transformed into the LSF domain where they are quantised using the 26-bit two-stage split vector quantisation method [82].

5.2.2 Decoder

A block diagram of the decoder of the Time Envelope SB-LPC vocoding scheme is given in Figure 5.5.

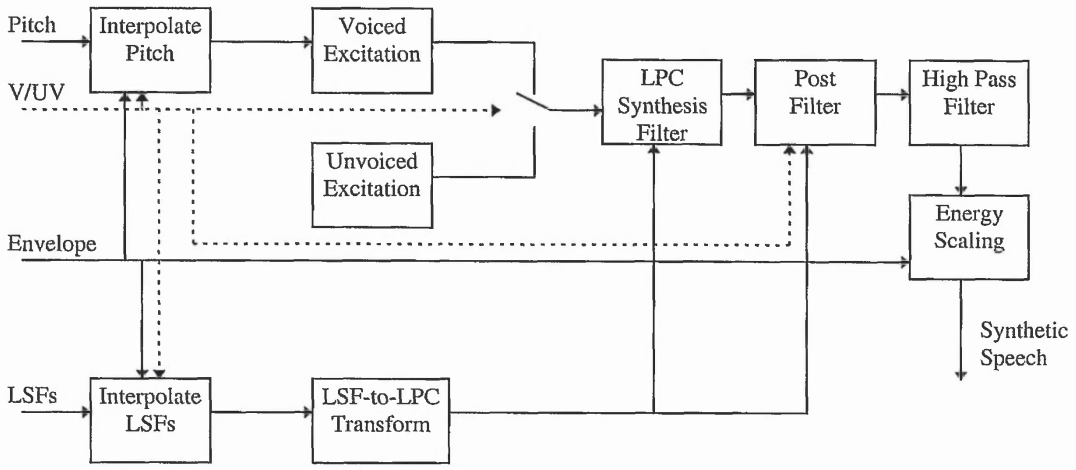


Figure 5.5 Block Diagram of the Decoder

Once the speech parameters have been decoded, either a quasi-periodic excitation or noisy excitation is selected by the voicing information. The quasi-periodic unity energy voiced excitation is generated using an interpolated pitch contour. The unity energy unvoiced

excitation is generated using a high pass noise source. The chosen excitation is fed into the LPC synthesis filter whose parameters have been obtained from the interpolated LSF information. The LSF interpolation factor is determined on the sub-frame basis by the decoded envelope and voicing information. The LPC synthesis filter is followed by the post filter. A high pass filter is also included in order to remove any steady state bias which may affect the energy level matching process.

5.2.2.1 Voiced Excitation Generation

There are several different ways of generating the voiced excitation in the Time Envelope vocoder [82]. Their common requirement is to reproduce the harmonic structure of the original speech at the decoder, whilst maintaining a flat average excitation power spectrum. The method adopted in the Time Envelope SB-LPC vocoder is based on the periodic impulse train generation which may be built up as a summation of sinusoidal harmonics that characterise the spectrum of the original speech, as shown in Equation 5.2 below

$$e(n) = \sum_{i=1}^{N} \cos(\frac{2\pi ni}{T}) \tag{5.2}$$

where T represents the period of the impulse train. The key advantage of this method lies in its ability to control the behaviour of each individual harmonic in terms of its phase and amplitude. Another important advantage is that the pitch period and impulse positions are not restricted to integer values.

In the case of the Time Envelope SB-LPC vocoder, the harmonics summation excitation also has the ability to vary the periodicity throughout the frequency spectrum by incorporating a controlled amount of phase jitter into each harmonic. As a consequence, the level of metallic sounds in the synthetic speech can be reduced, which renders more natural sounding speech output.

5.2.2.2 Unvoiced Excitation Generation

The white Gaussian noise source is generally used for unvoiced excitation, as it can meet the two basic conditions of unvoiced excitation generation - a flat spectrum and absence of any

harmonic structure. In order to subjectively improve the unvoiced excitation, the Gaussian source can be clipped and high pass filtered [82]. The clipping is performed in order to eliminate the occasional occurrence of large noise 'spikes'. The high pass filter is used to remove any steady state bias below 50 Hz.

5.2.2.3 Synthesis Filter and Post-filtering

Figure 5.6 depicts the synthesis filter structure followed by an adaptive post-filter [68], whose coefficients are dependent on the voicing decision. The function of the *perceptual weighting filter* is to reduce the perceived noise level by slightly emphasising the poles of the LPC filter whilst suppressing the noise in the valleys of the LP spectrum. However, sufficient noise reduction can only be achieved with slight muffling in the filtered speech. This phenomenon is due to the fact that the frequency response of the all-pole filter generally has a low-pass spectral tilt for voiced speech. A first-order *spectral tilt filter* is therefore included in order to provide a slightly high-passed spectral tilt and thus aid in further reducing the muffling effect. The tests [82] showed that this adaptive post-filter performs particularly well during voiced speech sounds, when greater perceptual shaping is employed, resulting in a more natural sounding of speech. During unvoiced speech, the combination of the moderately applied weighting filter and the spectral control produces an increased sharpness of these sounds. The main disadvantage of the post-filtering is that it requires an energy balance control at the beginning and at the end of this process. However, the time envelope energy control algorithm, already present in the Time Envelope SB-LPC vocoder, performs this task at no extra computational cost.

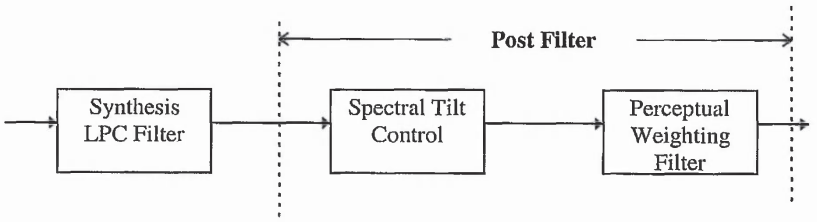


Figure 5.6 Synthesis Filter and Post Filter structure

5.2.2.4 Energy Scaling Mechanism

The energy scaling method used in the Time Envelope SB-LPC vocoder is the direct scaling of LP filtered unit variance excitation, investigated in [82]. In this method, the energy matching procedure is applied directly to the output of the LPC spectral shaping filter. The function of this method is to find the ratio between the filtered time envelope energy $e_f(j)$ and the decoded envelope energy $e_d(j)$, as given in

$$x = \frac{e_d(j)}{e_f(j)} \tag{5.3}$$

where x represents the scaling factor.

The calculation of the filtered time envelope is performed on the sub-frame basis, in the same manner as the encoded time envelope, where the envelope window length is adjusted to match the pitch period. This requires extending the decoded time envelope beyond the current frame end in order to accommodate the scaling of the sub-frames near to the frame end. The extension of the decoded time envelope can be as much as half of the maximum allowed pitch period, which in turn increases the overall coder delay.

The calculation of the scaling factor is performed, in conjunction with the envelope smoothing, on the sub-frame basis. The envelope smoothing is achieved using the first order IIR filter, as given in

$$y(n) = \alpha y(n-1) + (1-\alpha)x(n) \tag{5.4}$$

where $\alpha=0.85$ for general case smoothing and $\alpha=0.65$ for interpolating the detected speech onsets; y(n) is the scaling factor of the current output sample and x(n) is the energy of the current decoded sub-frame envelope.

The key features of the direct scaling of LP filtered unit variance excitation method are *reliability* in combating perceptually disturbing artefacts stemming from the energy variations and impulsive noise, *flexibility* regarding the choice of excitation source used and *ease of implementation.*

5.2.3 Bit Allocation Scheme

Table 5.1 shows the overall bit allocation scheme for the 2.4 kbps Time Envelope SB-LPC vocoder.

Table 5.1 Bit Allocation Scheme

		Bit Allocation					
Frame Length (ms)	Sub-Frame Length (ms)	LSF	Pitch	TE Gain	TE Shape	Voicing	Bit Rate kbits/ sec
20	**5**	**26**	**7**	**7**	**5**	**3**	**2.4**

5.2.4 Performance evaluation

The 2.4 kbps Time Envelope SB-LPC vocoder generally produces high quality natural sounding speech comparable to that offered by the I-MBE at 4.15 kbps as indicated by the informal tests conducted in [82] and in Chapter 7. However, the key disadvantage of this coder is its high degree of dependency upon the reliability of the time domain pitch detector algorithm. Pitch doubling errors are highly noticeable and reduce the overall quality assessment of the coder, even when the pitch errors are very occasional. The solution to this problem is the adoption of the frequency domain pitch search algorithm presented in the following section.

Another key problem with the time domain approach lies in its inability to more closely control the spectral behaviour of speech. Although an introduction of the 'soft' split band voicing decision has shown superior performance over the 'hard' single voicing decision, the Time Envelope vocoder is still totally dependent upon LP modelling to represent the speech spectral energy distribution, hence LP modelling inaccuracies and imperfections show up in the synthetic speech. Numerous speech tests [86] showed that the LP residual waveform remains highly intelligible containing important spectral information. Better modelling of the LP excitation is therefore required to improve the vocoder's quality and can be achieved by incorporating spectral amplitude information using the MBE type approach. The key advantage of the spectral amplitude representation is that it follows closely the split band voicing classification and enables more accurate voiced and unvoiced excitation generation.

The introduction of the amplitude information, as performed in the following section, can also significantly improve the interpolation process on the decoder side.

The fundamental idea behind this research is the design and development of variable rate schemes using the existing improved vocoding techniques. As far as the variable rate quantisation approach is concerned, as indicated in Chapter 6, the Time Envelope vocoder poses no predefined restrictions in its structure. However, adopting the variable frame length approach can be adversely affected by the fixed time envelope quantisation structure. Quantising the variable sized time envelope can therefore be regarded as another big disadvantage of this vocoder.

5.3 Amplitude Excited SB-LPC vocoder at 2.4 kbps

Although the Time Envelope vocoder presented in the previous section produced highly intelligible speech at a fixed bit rate of 2.4 kbps, its basic coding structure was based on the assumptions and simplifications which limited the overall speech quality, and was found to be relatively inflexible towards the implementation of variable rate schemes. The Amplitude Excited SB-LPC vocoder abandons the varying time envelope representation and switches to the spectral amplitude representation. The main advantage of this frequency domain approach, with respect to the attainable speech quality, is a significant improvement of the LP modelling technique. Another key advantage, with respect to the future implementation of the variable frame rate scheme, is that its spectral amplitude structure remains of a fixed size in the variable length coding conditions and, thus, is much easier to quantise.

Generating voiced and unvoiced excitation in the frequency domain using the spectral amplitudes is the main feature of the 4.15 kbps I-MBE vocoder, as described in the previous chapter. The key disadvantage of the I-MBE approach is that it requires a lot of bits in order to accurately encode the spectral amplitudes; these may be far more efficiently encoded when aided by the LP modelling technique. The Amplitude Excited SB-LPC vocoder operating at 2.4 kbps is an MBE-LPC type vocoder. Its main characteristics are the use of both the LP

modelling technique to represent the overall speech spectral envelope, and an MBE type approach to model the LP excitation.

5.3.1 Encoder

The block diagram of the encoder section of the Amplitude Excited SB-LPC vocoder is shown in Figure 5.7.

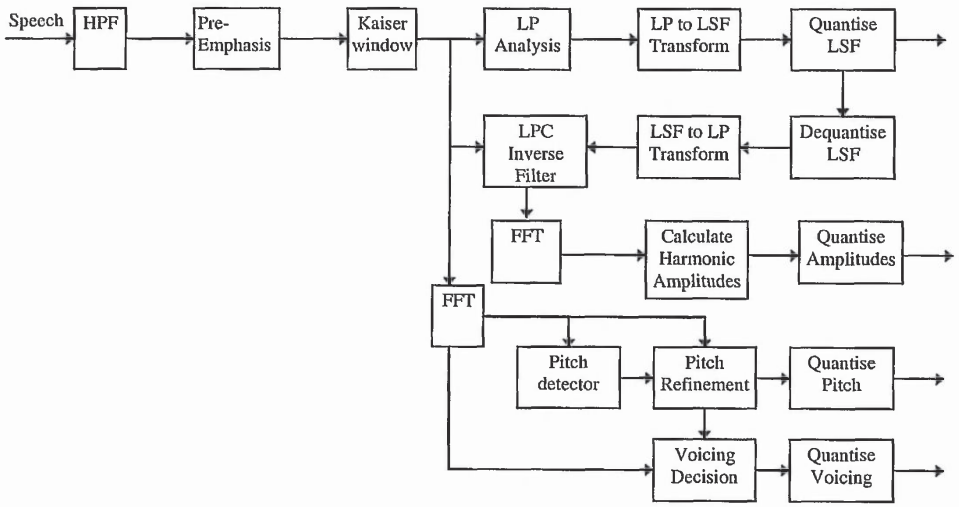


Figure 5.7 Encoder Block Diagram

The speech analysis is performed on 20 ms speech frames for the operating bit rate of 2.4 kbps (at 8000 Hz sampling frequency). The parameters that are used to represent the speech model include:

- pitch period
- split band voicing decision
- harmonic amplitudes
- 10th order LP filter.

At first, the original speech is high pass filtered, pre-emphasised and windowed using a 221-point Kaiser window (β=6.0). The calculation and quantisation of the LPC parameters and SB voicing decision are performed in a similar way to the Time Envelope encoding routine. The novelty is the presence of the LPC inverse filter which is used to find the LP residual

waveform required for determination of the excitation harmonic amplitudes. Another novelty is the frequency domain pitch detection method called a Sinusoidal Model Matching method. The Synthetic Spectral Matching pitch refinement method is also performed in the frequency domain as in the Time Envelope vocoder.

5.3.1.1 Sinusoidal Model Matching Pitch Detection Algorithm (SMM-PDA)

The SMM-PDA method is a frequency based pitch detection algorithm which is performed on the FFT spectrum of the original speech. The preference towards the original speech rather than its LP residual waveform lies with the fact that the LP filtering action can sometimes not only remove the formant interference (its key purpose) but also lower the speech harmonics to the noise floor level, significantly reducing the accuracy of the pitch detection. The SMM-PDA algorithm was originally developed by [87] and then further enhanced by [86]. The present algorithm used in the Amplitude Excited SB-LPC vocoder is an enhanced version of the SMM-PDA.

In the SMM-PDA the speech signal is assumed to be composed of a number of harmonically unrelated sinusoidal components, given by

$$s(n) = \sum_{l=1}^{L} A_l \exp(jn\omega_l + j\theta_l) \tag{5.5}$$

where A_l, ω_l and θ_l represent the spectral amplitudes, frequency and phase respectively. The method then attempts to detect the fundamental frequency, ω_0, by generating harmonically related synthetic sinusoidal signals, given by

$$\hat{s}(n,\omega_0) = \sum_{l=1}^{K(w0)} \hat{A}_l \exp(jnl\omega_0 + j\theta_l) \tag{5.6}$$

where $\hat{A}_l$ represents harmonic amplitudes of the synthetic spectral envelope. The fundamental pitch is found by minimising the MSE between $s(n)$ and $\hat{s}(n)$. The complex searching solution is based on a number of assumptions and perceptual enhancements, of which the most important are:

• both spectra of $s(n)$ and $\hat{s}(n)$ are well resolved and are approximated by a distance function $D(\omega - k\omega_0)$, which usually takes the shape of a modified *sinc* function [86] to account for the

ear's increasing tolerance to pitch errors at high frequencies, at each component frequency scaled by the respective harmonic magnitudes

• each sinusoidal component of $s(n)$ is matched by the usually closest candidate frequency of $\hat{s}(n)$

• minimisation of the error in $\sum_{k=0}^{N}\left[x(k)-y(k,a,b,c,...)\right]^2$ over variables a,b,c,... is equivalent to maximising the function $\sum_{k=0}^{N} y(k,a,b,c,...)\left[x(k)-\frac{1}{2}y(k,a,b,c,...)\right]$.

The fundamental pitch is then determined by maximising $\rho(\omega_0)$ defined in Equation 5.7 below

$$\rho(\omega_0)=\sum_{k=1}^{K(\omega_0)}\hat{A}(k\omega_0)\left\{\max_l\left[A_l D(\omega_l-k\omega_0)\right]-\frac{1}{2}\hat{A}(k\omega_0)\right\} \tag{5.7}$$

where the number of harmonics in $\hat{s}(n)$ is frequency dependent, as defined in Equation 5.8,

$$K(\omega_0)=\left\lfloor\frac{\pi}{\omega_0}\right\rfloor \tag{5.8}$$

It can be shown [86] that the $\rho(\omega_0)$ function ensures that the probability of choosing multiples or sub-multiples of the fundamental frequency is minimal. In order for this process to operate effectively, a simple spectral peak interpolation algorithm is usually sufficient to generate the spectral envelope of the $\hat{s}(n)$, which should ideally pass through the peak spectral points of $s(n)$, thus minimising the mismatch at correctly positioned synthetic harmonics. However, determining the accurate and reliable spectral envelope requires the knowledge of the pitch period, so that minor inter-harmonic peaks may be distinguished from the true peaks. This gives rise to the circular dilemma [87], a problem which was successfully tackled by the improved version of the SMM-PDA algorithm proposed in [86]. The improved algorithm is characterised by:

• the use of simple metrics to reject spectral peaks which are likely to lie in inter-harmonic valleys

• the spectral envelope detection which does not rely upon the pitch period knowledge

• the biased selection towards lower pitch periods (within the first 1.4 kHz)

• the addition of the pitch tracking element to the pitch decision.

The improved SMM-PDA showed very high resilience towards the pitch multiple and sub-multiple errors and was therefore a preferred choice over the time-domain SAPDA method used in the Time Envelope SB-LPC vocoding scheme.

5.3.1.2 Improved Split Band Voicing Technique

The SB voicing technique used for the Amplitude Excited SB-LPC vocoder resembles in its key points the one used in the Time Envelope approach described in the previous section. The experimental additions include:

- the zero-crossing rate
- low-band to high-band energy ratio
- fractional energy ratio.

All these inclusions represent ancillary tools in distinguishing between the voiced and unvoiced speech. As the SB voicing technique is based on accurate pitch detection, these tools play an important role in cases where this pitch detection fails.

The *zero-crossing rate* is defined as the number of times the signal changes sign, divided by the number of samples used in the observation. It is important that any DC bias is removed from the signal before the analysis, in order to correctly identify zero-crossing points. Previous research [83] indicated a certain degree of reliability when the zero-crossing metric is used to differentiate between voiced and unvoiced speech, as it is generally accepted that the unvoiced sounds have much higher zero-crossing rate than the voiced sounds. The best results were obtained for regions of speech where the detected pitch periods are very large (pitch > 75 samples). For shorter pitch sizes, the transition region between voiced and unvoiced classification remained too broad, which incurred reduced reliability of this metric.

The traditional *low-band to high-band energy ratio* can also be used to help the voicing classification. The noise signal generally tends to contain high frequency, therefore a large part of its energy should lie in the upper part of the spectrum. The voiced signal usually displays the opposite character with much of its formant energy situated in the lower half of the spectrum. The main drawback of this metric is defining the cut-off points for low-pass

and high-pass filters used, and adjusting the ratio thresholds. The alternative to this method is the full-band to low-band energy ratio [45] which has been shown to provide a higher level of reliability in voicing classification than the conventional energy ratio approach.

The *fractional energy ratio* measures the ratio between the full-band energy of the current frame of speech and the maximum energy detected in the processed signal. The idea here is to track large fluctuations in energy levels and to potentially bias the overall voicing decision.

5.3.1.3 Harmonic Amplitude Determination and Quantisation

The spectral amplitude is determined for each harmonic of the LP residual signal of the current frame according to the values of the unquantised refined pitch and the quantised voicing information. In the case of the voiced harmonics, the amplitude calculation uses the MSE matching method applied to the Kaiser window in the frequency domain, similar to the pitch refinement search technique. In the case of the unvoiced harmonics, the amplitude calculation determines the RMS spectral energy over the unvoiced harmonic band.

The RMS value of the excitation harmonic amplitudes is quantised using a 6-bit logarithmic scalar quantiser. The harmonic amplitudes are normalised to their RMS values and then quantised. The first eight harmonics are quantised using a 6-bit quantiser, whereas the remainder of harmonics are defined as having unity values [86].

5.3.1.4 Determining and Quantising LP Coefficients

The 10th order LPC filter is based on Durbin's analysis which is performed on pre-emphasised speech using a 221-sample Kaiser window (β=6.0). The LP coefficients are transformed into the LSF domain where they are quantised using the 26-bit two-stage split vector quantisation method.

5.3.2 Decoder

Figure 5.8 presents the decoder schematic of the Amplitude Excited SB-LPC vocoder operating at the fixed rate of 2.4 kbps. The *LSF parameters* are decoded, interpolated and transformed into the LP coefficients. The LP coefficients are then used at the final stage of the decoder in the LPC synthesis filter. The *voicing* information is decoded and used to control the amount of voiced and unvoiced excitation generation. The decoded *pitch* period also plays an important part in generating the separate excitation signals. The *amplitudes* are decoded and then modified in such a way as to enhance the perceptual performance of the vocoder. The voiced excitation is generated using the time domain harmonic summation of the voiced amplitudes. The unvoiced excitation is generated in the frequency domain and then transformed into time domain where it is combined with the voiced excitation. Finally, the combined excitation signal is passed through the LPC filter to produce the synthetic speech.

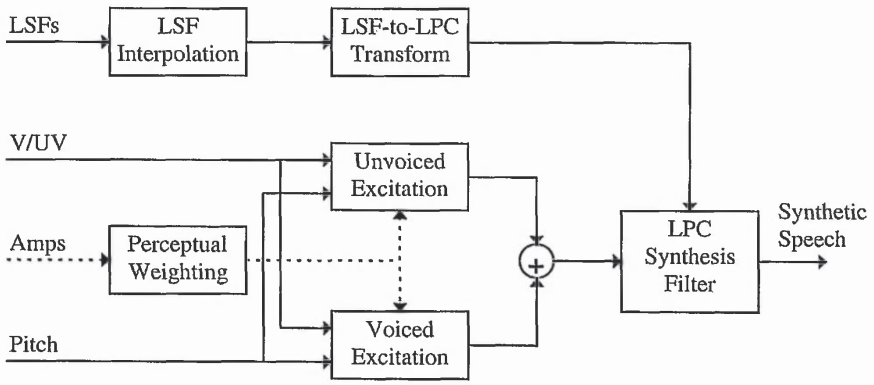


Figure 5.8 Decoder Block Diagram

5.3.2.1 Perceptual Enhancement

The post-filtering technique applied in the decoder part of the Time Envelope SB-LPC vocoder described in the previous section showed a high degree of noise attenuation in the spectral valleys, which contributed to the level of cleanness in the synthetic speech. However, the side effect of such a function is the selective amplification of the formant structure, which may introduce distortion into the synthetic speech.

In this vocoder an alternative perceptual enhancement technique replaces the post filter. The modification of the spectral amplitudes before the LP filtering, as suggested in [86], demonstrates the valley suppressing effect on the LP spectrum and shows no signs of the selective formant amplification. The amount of amplitude modification depends on the difference between the LP spectrum and its equivalent peak interpolated LP spectral shape, as given in

$$Amp_{\text{mod}}(i) = Amp(i)\left[\frac{H(iw_0)}{P(iw_0)}\right]^{\gamma} \tag{5.9}$$

where $H(i\omega_0)$ represents the LP spectrum and $P(i\omega_0)$ is the peak interpolated LP spectrum, both sampled at the harmonic frequencies. The value of the suppression factor γ typically lies in the range of 0.1 to 0.35.

5.3.2.2 Voiced Excitation Generation

Equation 5.10 defines the generation of the voiced excitation. The principle is similar to the Time Envelope SB-LPC vocoder's voiced excitation generation, except that there is no more phase jitter and each harmonic is individually amplitude scaled.

$$e_{voiced}(i) = \sum_{j=1}^{L} Amp_{\text{mod}}(j)\cos[j\phi(i)] \tag{5.10}$$

where $Amp_{mod}(j)$ takes on the modified amplitude value if the jth harmonic is voiced, or $Amp_{mod}(j)=0$ when the jth harmonic is unvoiced.

Equation 5.11 represents the fundamental phase $\phi(i)$ as the sum of the interpolated fundamental frequency function $\omega(i)$, defined in Equation 5.12, and the fundamental phase of the previous sample.

$$\phi(i) = \phi(i-1) + \omega(i) \tag{5.11}$$

$$\omega(i) = \begin{cases} \omega_{last}\left(\frac{i+0.5}{N}\right) + \omega_{next}\left(1-\frac{i+0.5}{N}\right), & |\omega_{last} - \omega_{next}| \le 0.25(\omega_{last} + \omega_{next}) \\ \omega_{last}, & |\omega_{last} - \omega_{next}| > 0.25(\omega_{last} + \omega_{next}) \text{ and } i \le \frac{N}{2} \\ \omega_{next}, & |\omega_{last} - \omega_{next}| > 0.25(\omega_{last} + \omega_{next}) \text{ and } i > \frac{N}{2} \end{cases} \tag{5.12}$$

The interpolation process shown in Equation 5.12 is only performed when the pitch value deviates by less than 25% of its value from one frame to the next. During voiced excitation, the linear harmonic amplitude interpolation from one frame to the next is employed, whereby the amount of interpolation depends on the voicing information of each interpolated frame.

5.3.2.3 Unvoiced Excitation Generation

Figure 5.9 depicts the block diagram of the unvoiced excitation generation function. A Gaussian noise source is used to produce a random noise signal whose spectrum is approximately flat. The noise is then windowed and transformed into the frequency domain using a 512-point FFT.

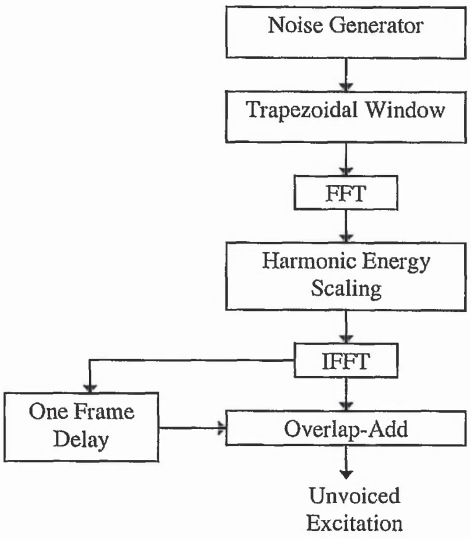


Figure 5.9 Unvoiced Excitation Generator Block Diagram

In the frequency domain, the noise spectral energy components lying within a voiced harmonic band are set to zero. The noise spectral energy in the remaining unvoiced harmonic bands is then averaged and scaled against the corresponding decoded unvoiced harmonic amplitudes. The inverse FFT of the scaled noise spectrum produces the time signal containing the correct spectral energy distribution and no harmonic structure. In order to maintain a smooth transition in the unvoiced excitation from the previous frame to the next, an Overlap-Add method is employed [86], as shown in Figure 5.10.

This method of unvoiced excitation generation proved to give an improved performance when compared to its simpler version employed in the Time Envelope SB-LPC vocoder.

Accurate and efficient parameter interpolation between the adjacent speech frames, both in voiced and unvoiced excitation generation domain, played an important role in designing and developing variable frame length systems, as described in Chapter 7.

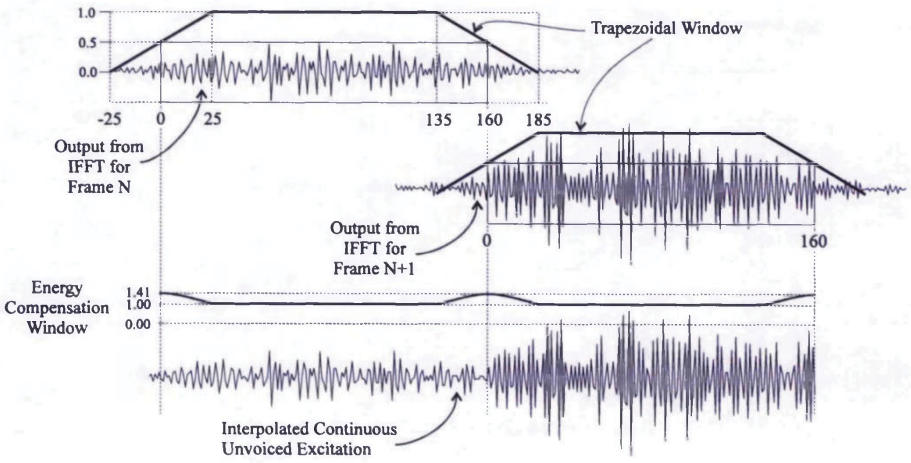


Figure 5.10 Unvoiced Excitation Generator Block Diagram

5.3.2.4 LPC Synthesis Filter

The LP excitation signal is the summation of both the generated voiced and unvoiced excitations. The filtered signal is synthesised on the sub-frame basis, i.e. every 5 ms, using the interpolated LP coefficients. Spectral de-emphasis is performed on the filtered signal in order to produce the synthetic speech output.

5.3.3 Bit Allocation Scheme

Table 5.2 shows the overall bit allocation scheme for the 2.4 kbps Amplitude Excited SB-LPC vocoder.

Table 5.2 Bit Allocation Scheme

		Bit Allocation					
Frame Length (ms)	Sub-Frame Length (ms)	LSF	Pitch	Energy	1st 8 Harmonics	Voicing	Bit Rate kbits/ sec
20	**5**	**26**	**7**	**6**	**6**	**3**	**2.4**

5.3.4 Performance Evaluation

The 2.4 kbps Amplitude Excited SB-LPC vocoder generally produces high quality natural sounding speech. Its perceptual test evaluation, as conducted in [86] and also shown in Chapter 7, indicates an improved subjective performance over the Time Envelope SB-LPC vocoder at 2.4 kbps and the I-MBE vocoder at 4.15 kbps.

The two key problems of the Time Envelope approach have been successfully solved by the Amplitude Excitation approach. Firstly, the frequency domain pitch detection routine has proved to be much more reliable than the time domain PDA, especially in eliminating the pitch multiple and sub-multiple cases. Also, the introduction of the spectral amplitude information has significantly improved the subjective performance of the LP modelling technique used in the Time Envelope approach. Secondly, the problem of quantising the time envelope of the frame whose size is allowed to vary has now been resolved by switching to the frequency domain and the amplitude representation approach. Namely, the amplitude quantisation is performed in the spectral domain, which does not change size regardless of the actual analysis frame length. This constant spectrum size means that the coding of the parameters in the variable frame length technique can be done in a more efficient way. Also, the parameter interpolation at the decoder side between the adjacent frames of various lengths can be done with a greater degree of accuracy by using the amplitude excited approach. Therefore, the Amplitude Excited SB-LPC vocoder represents a preferred choice over the Time Envelope SB-LPC vocoder for designing and simulating variable frame length scheme, a topic discussed in Chapter 7.

5.4 Conclusions

The subject of this chapter has been the design and simulation of the fixed rate speech coding systems, namely, the Time Envelope system and the Amplitude Excited system, based on the SB-LPC vocoding structure and operating at 2.4 kbps. For each system, detailed information about the basic structure and principles of operation of the encoder and decoder have been provided. The advantages and disadvantages of both coders have also been assessed in terms of their technical merits and their output perceptual quality.

The thorough analysis of the fixed rate coding structures has mainly been aimed at assessing their particular suitability to provide the grounds for the development of the variable rate coding systems, which are presented in the following chapters. Both the objective and subjective tests have shown that the Amplitude Excited SB-LPC vocoder is much more acceptable as the basis for designing the variable rate scheme. Not only its subjective quality performance but also its flexible coding structure completely outweighs the Time Envelope approach.

In the course of this research the Time Envelope system was the first one to be designed and was consequently the first one to adopt the variable rate system approach; the variable rate quantisation system, presented in Chapter 6, was developed using solely the Time Envelope vocoder. However, the complexity and bit allocation requirements of the variable frame length approach, described in Chapter 7, called for a more sophisticated and more flexible system approach than the one already provided by the Time Envelope vocoder. As a result, the Amplitude Excited SB-LPC vocoder was used, although both systems were initially designed and tested for the conditions of this novel variable rate approach.

Chapter 6.

6. Variable Rate Quantisation in Speech Coding

6.1 Introduction

One of the main research objectives in speech coding is the preservation of high speech quality whilst minimising bit rate. This is becoming increasingly relevant due to the growing demand for better bandwidth utilisation over speech, video, and, in general, multimedia channels. In speech coding many well-documented algorithms, such as CELP, MBE vocoders or LPC coders [88], are designed to operate at a fixed bit rate, usually in the very low bit rate range of 2.4 - 8 kbps. However, in the speech storage applications, such as voice mail or an answering machine, even this bit rate range is too high when only a restricted memory space is available. Moreover, the possibility of varying the bit rate of a coder is particularly suited for use in the packet-switched networks.

The bit rate of fixed rate speech coders depends on the quality to which the most perceptually difficult regions of speech are coded. These regions occur usually within voiced speech segments, especially during dynamically changing speech transients, such as voiced onsets

and offsets. In these segments, especially the coding of the residual signal consumes most of the transmission bits. However, in unvoiced parts of speech fewer bits are necessary to code the residual as well as the LPC parameters. A further bit rate reduction is possible during silence or background noise segments where only few bits are necessary to code spectral shape and energy of a stationary noise signal.

The variable bit rate systems utilise the time varying characteristics of the speech signal such that fewer bits are used to code highly correlated steady state segments and unvoiced regions, whilst a fixed rate is maintained for finer quantisation of speech transients. In order to realise the advantages of a variable rate coder, which selects the bit rate according to the time varying characteristics of the speech segments, development of accurate and robust criteria to classify these segments is of utmost importance. These criteria should be chosen in such a way that the output quality of variable rate systems is preferably better or at least matches that of the equivalent fixed rate systems.

The concept of Variable Bit Rate Coding (VBRC) has in recent years become a focal point of considerable research activity [89][90][91][92] showing that variable bit rate techniques can augment and enhance many fixed rate systems, producing similar speech quality at lower average bit rates. This activity is due to some new and rapidly expanding voice communication services, such as Voice Store and Forward systems, which can take maximum advantage of the bit rate variations of the VBRC systems. However, the most important future application of the VBR coders is in integrated packet switched communication networks, such as the newly emerging GPRS (General Packet Radio Service) which is expected to offer packet data services within the same physical layer environment as GSM services [93]. The transport technology of choice for the GPRS architecture is more than likely to be based on ATM (Asynchronous Transmission Mode) [94], which utilises a statistical bandwidth allocation strategy to provide a virtually continuous range of bit rates within physical network limits. Before the ATM era, packets of data were transmitted in a fixed physical channel structure where fixed rate speech codings, such as log-PCM and ADPCM, were conventionally being used. This was necessary because variable rate coding imposed complexity and excess delay for bit rate smoothing (bit rate adaptation to the channel) in the packetisation process. However, with the advance of the ATM networks, the physical channel structure becomes extinct, making variable bit rate coding very attractive. The other packet

switched services that could make use of the VBRC system are the Wireless Voice-over-IP (Internet Protocol) and Multimedia-over-IP applications [95][96].

The variable rate coders could broadly be classified into two groups: source driven VBR coders and network driven VBR coders [97]. For the source driven VBR coding, the encoding bit rate varies dynamically in harmony with the time varying characteristics of the input source. For the network driven VBR coders, as the name implies, the encoder output can be adjusted according to the traffic status and general requirements of the communications network.

The source driven VBR coding is implemented by varying the bit rate used for parameters encoding whilst maintaining the fixed duration of input speech segments. This frame based criterion can be based on either the short-term characteristics of the input speech or a coding accuracy measure that is a function of the difference formed between the original and synthesised speech frames or a hybrid combination of these two methods.

The network driven VBR coding can be implemented as either multi-rate coding or embedded coding. In multi-rate coding the network requests, at relatively infrequent intervals, the VBR coder to operate at an appropriate rate. In response, the coder, while employing a given fixed architecture, adjusts its parameters to yield one of a small number of possible output bit rates [98][17]. The embedded coding is a technique which allows simple bit dropping in a given bit stream. In other words this means that embedded in the 'faster' bit-stream (higher output rate) there is a 'slower' signal (lower output rate) that can be decoded without seriously affecting the receiver's synchronisation with the transmitter and its output quality [99][100].

This Chapter discusses the variable quantisation scheme which was designed and developed for the Time Envelope SB-LPC vocoder, described in the previous chapter, operating at a fixed rate of 2.4 kbps. This new scheme utilises the fact that the modelling parameters for different speech segments do not necessarily require the same number of bits to encode, hence it could be classified as the source driven variable bit rate coder. The objective is to produce a very low bit rate coder with an average bit rate lower than the fixed bit rate. However, this should be achieved in such a way that the output quality of the new variable rate system is preferably better or at least matches that of the equivalent fixed rate system. At

first the chapter describes the segmentation and categorisation of the speech signal. It then discusses the differences in coding strategies of the variable and fixed rate quantisation schemes. Finally, it assesses the quality of the performance of the variable rate quantisation depicting its advantages and disadvantages over the fixed rate system.

6.2 Speech Categorisation

It is well known that speech can be considered as a sequence of phonemes where each phoneme is characterised by a set of features, such as voicing, nasality, sustention, graveness, sibilation and compactness [101]. Different type phonemes attach greater importance to some features in order to preserve their phonetic integrity. Since these features are represented by certain combinations of parameters which are to be quantised by the coder, an ideal coding scheme should be able to find the critical parameters which correspond to the most important features and place the coding emphasis on these parameters by assigning more bits to them. Such a variable bit quantisation scheme will then be able to make a fixed rate algorithm more efficient and produce if not better then at least similar output perceptual quality.

In order to achieve such a goal, speech should be segmented into phonetically distinct categories where each category requires its own coding analysis and bit allocation scheme. In general, each phonetically distinct segment can be categorised separately so that a coding structure and bit allocation can be precisely matched to each category. However, a large number of phonetic categories would make an overall coding scheme too complex and would require an excessive amount of side information. The main speech modelling parameters used in this work for categorisation of speech segments were pitch period, voicing decision and time-envelope energy. According to their values the following distinct speech segmentation categories were chosen:

- silence
- unvoiced
- mixed
- fully voiced.

Unvoiced regions are non-stationary and have a noise-like waveform with fairly weak near sample correlation.

Silence or non-speech can be described as a sub-class of the unvoiced category as it generally consists of a low energy background acoustical noise. However, the silence segment assumes great importance in the variable quantisation scheme due to its reduced bit consumption when compared to the unvoiced class and its frequent occurrence during the typical telephone conversation. In this work it is, therefore, regarded as a separate segmentation category.

Mixed sounds, as the name implies, contain a mixture of both periodic and unvoiced speech which usually occurs in the form of dynamically changing transient speech. Rapid transitions from an unvoiced to a voiced sound (onset) or vice versa (offset) also belong to this mixed category.

Fully voiced sounds exhibit steady state quasi periodic features. It is essential to distinguish between the mixed and fully voiced categories so that an efficient coding strategy and bit allocation scheme can be formulated. For example, the mixed segment will require a maximum number of bits for the quantisation of its parameters as it is typically non-stationary and perceptually very important. The fully voiced segment is on the other hand locally stationary with steady state speech and therefore requires less bits for the parameter quantisation.

Figure 6.1 illustrates the four different segmentation categories of speech.

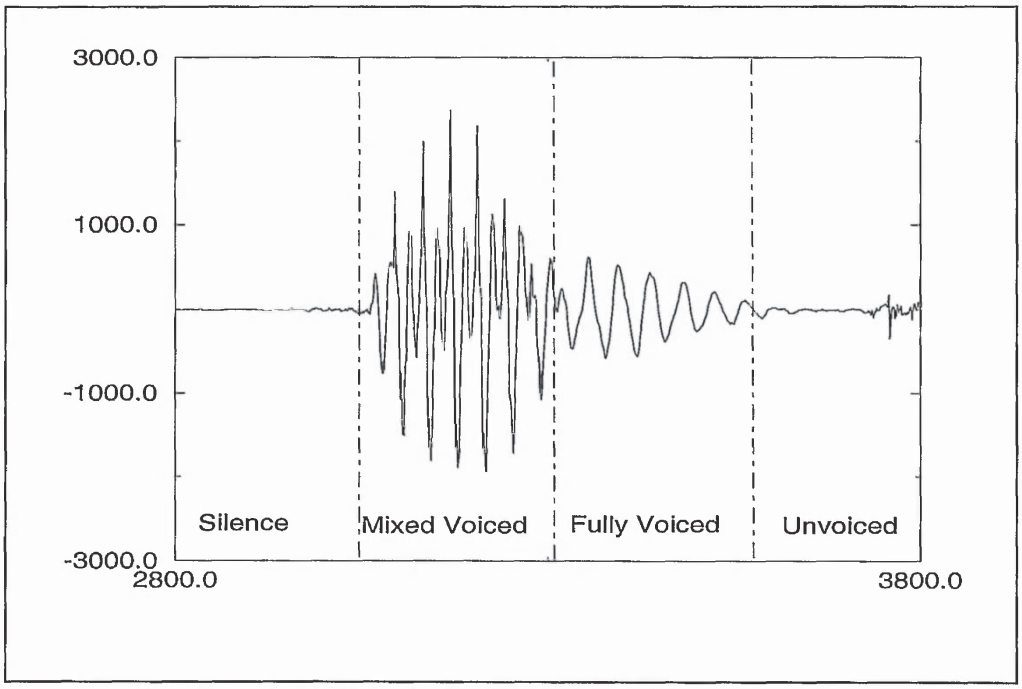


Figure 6.1 Speech segmentation classes

6.3 Variable Rate Quantisation Strategy

The variable quantisation scheme has been developed for the Time Envelope SB-LPC vocoder described in detail in Chapter 5. This vocoding method models each 20 ms speech frame using the following parameters: pitch period, split-band voicing decision, gain and energy of a time envelope and Line Spectral Frequencies (LSFs) for quantising ten LP coefficients. The number of bits required for encoding each of these parameters is depicted in Table 6.1(a). According to this Table, the overall fixed bit rate is 2.4 kbps assuming a sampling frequency of 8 kHz.

Table 6.1(a) Quantisation scheme for the fixed bit rate vocoder

Frame Length (ms)	LSF	Pitch	Energy	Shape	Voicing	Bit Rate kbits/ sec
20	**26**	**7**	**7**	**5**	**3**	**2.4**

The new variable quantisation scheme utilises the fact that the modelling parameters for different speech segments do not necessarily require the same bit allocation structure. Speech

frames, of a constant length of 20 ms, are classified into four categories, namely, silence, unvoiced, mixed and fully voiced. Table 6.1(b) summarises the segment classification and coding structure used for these classes by specifying coder parameters and bit allocation for each of the four categories.

Table 6.1(b) Quantisation scheme for the variable bit rate vocoder

	Silence	Unvoiced	Mixed	Fully Voiced
Header	2	2	2	2
Mode	1	1	--	--
Pitch	--	--	7	7
Voicing	--	--	3	--
LSFs	--	20/26	26	26
Energy	--	7	7	7
Shape	--	5	5	5
Bits/Frame	3	35/41	50	47

As can be seen in Table 6.1(b), the minimum quantisation rate is achieved when only silent frames are transmitted by the encoder. Transmission of a single silent frame requires 3 bits. However, if a series of four consecutive silent frames have been detected by the embedded pre-processor on the encoder side, the mode bit is set and only one silent frame is sent to the decoder. This implies that 3 bits (2 header bits plus 1 mode bit) cover not one but four silent frames, producing an average of 0.75 bits/frame. Therefore, when only silence is present, the average bit rate can drop down to around 0.0375 kbps which is a minimum bit rate of the variable quantisation scheme. One can also observe that a maximum rate of 2.5 kbps is necessary for transmission of the mixed frames. This figure is slightly higher than that of the fixed quantisation rate coder due to the extra two bits overhead which represent header classification bits carrying speech category information for each transmitted frame.

6.3.1 Coding of Silences

A substantial reduction in average bit rate can be attained by exploiting voice activity patterns in conversational speech. In general, one side of a two way telephone conversation is silent roughly 60% of the time, i.e. the voice activity factor is 0.4. This silence or non-speech

generally consists of a low-energy level background acoustic noise which must be coded at low rate or otherwise regenerated as comfort noise on the decoder side to preserve the natural quality of the synthesised signal. The variable quantisation scheme of the Time Envelope SB-LPC vocoder employs a simple voice activity detection (VAD) scheme based on the Discontinuous Transmission (DTX) system approach adopted in the GSM [16].

When no speech is present, the DTX comes into operation by switching the transmitter off. This increases system capacity by reducing co-channel interference and also reduces transmitter power consumption, which is an important consideration for hand-held portable telephones. In a typical conversation, each speaker talks for less than 40% of the time, and the estimation is that the DTX could approximately double the capacity of the radio system. Figure 6.2 shows a block diagram of the DTX system.

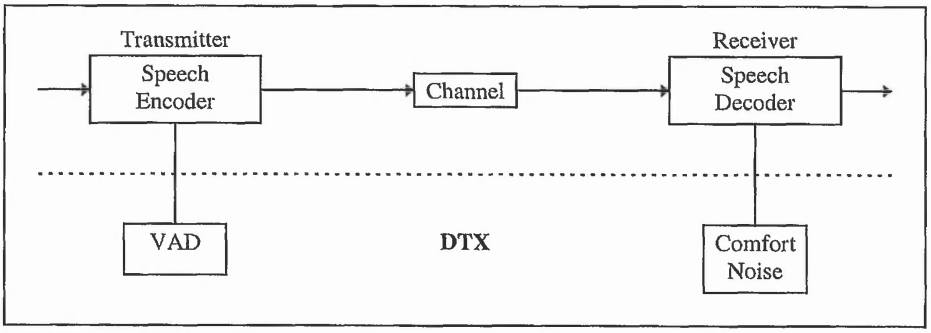


Figure 6.2 DTX block diagram

The computational complexity of the DTX algorithm is very low as it works alongside the speech coding algorithm. Essentially, the DTX consists of the VAD at the encoder and the comfort noise generator at the decoder. The function of the VAD is to distinguish between speech superimposed on the background noise and noise without speech present. The VAD will degrade output speech quality by clipping the speech if it fails to detect every speech event. On the other hand, if the VAD identifies noise as speech too often, the effectiveness of the DTX will be greatly reduced. The DTX is a potential source of signal degradation, so it should only be made operational at times when the system usage rises above a certain threshold. Comfort noise is generated by the receiver when the DTX has switched the transmitter off, in order to eliminate the unpleasant subjective effect of switching to and from silence.

The DTX system for the variable rate quantisation Time Envelope SB-LPC vocoder comprises an encoding pre-processor with a simple VAD on the transmitter side and a comfortable noise generator on the receiver side. The encoding pre-processor, as the name implies, is operational before the actual encoder, i.e. it takes the incoming speech frame first and determines, according to the frame's speech characteristics, whether it is a silent frame or not. The frame is then passed onto the encoder for further speech analysis. If the silence is detected, the pre-processor sets its internal clock and inputs the next frame. Each successive input frame that has been declared as silent increases the internal clock of the pre-processor. When four successive input frames have been detected as silence, the internal clock is reset, a silence mode bit is set and only three bits (two-bit header and one mode bit) are transmitted to the decoder. Therefore, instead of four silence frames, only one frame is transmitted. The period of four silence frames is chosen because it can fit well inside a maximum allowable delay in the GSM-type mobile communications ($\leq$ 90 ms). However, all four frames must be passed through the encoder to maintain the encoding continuity. Each time a non-silent frame is detected, the internal clock of the pre-processor is automatically reset. This also means that, for example, if a non-silent frame (unvoiced or voiced) follows two or three successive silent frames stored in the pre-processor, the internal clock is reset and the stored silent frames are then passed individually through the encoder and onto the decoder. For the successful operation of the encoding pre-processor, a maximum algorithmic delay of 80 ms, representing four successive frames, is necessary.

Figure 6.3 shows an example of the correct silence detection as performed by the encoding pre-processor. It can be observed that a group of four consecutive silent frames is detected by the encoding pre-processor and encoded, but only one silent frame is transmitted, i.e. the encoder transmits two header bits indicating a silent frame and a silent mode bit in the set mode indicating a group of four silent frames. However, the following frame just before the onset is also detected as silence by the encoding pre-processor, but encoded and transmitted individually, i.e. the encoder transmits two header bits indicating a silent frame and a silent mode bit in the reset mode indicating a single silent frame.

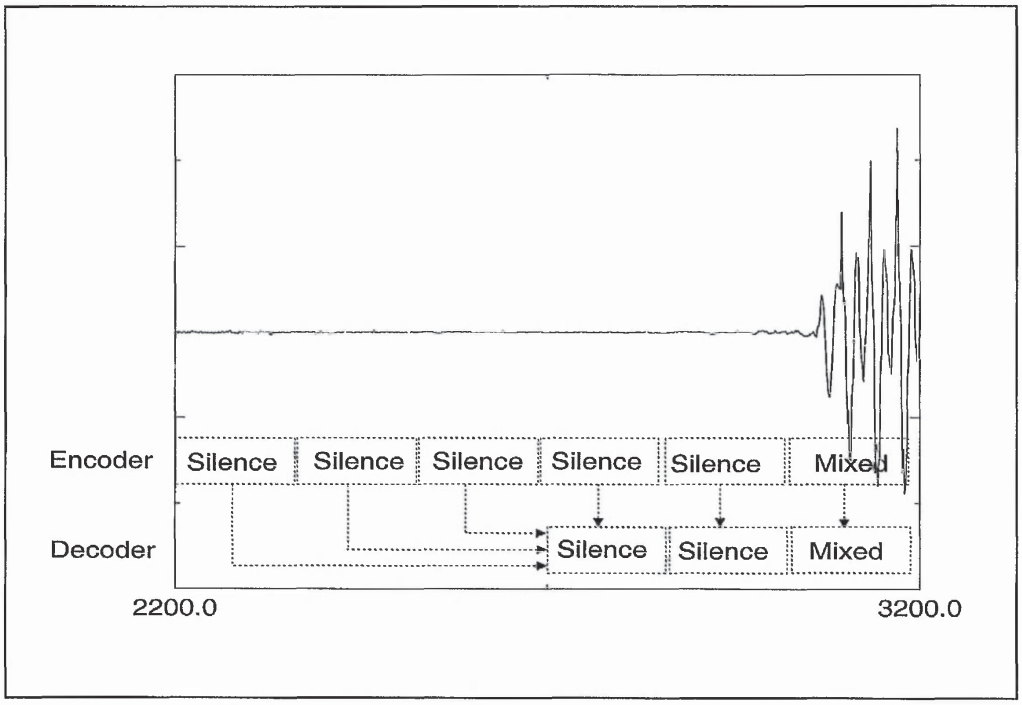


Figure 6.3 Silence frame detection mechanism

As can be seen in Table 6.1(b), a silent frame requires no other information except the two header bits and a mode bit. As it is an unvoiced frame by nature, any information on pitch period, voicing decision and LP shape would be unnecessarily transmitted. The other two parameters are the energy and shape of a time envelope. However, they have also been found largely redundant after the inclusion of an adaptive energy threshold calculated at the encoding pre-processor. The function of the adaptive energy threshold is to distinguish between the silent and the unvoiced frames, whereby a silent frame is detected when its time envelope energy, as given in Equation 6.1,

$$e\_t = \sqrt{\frac{\sum_{n=0}^{N-1} s(n)^2}{N}} \tag{6.1}$$

where N is a constant frame size, lies below the adaptive energy threshold.

In order for the adaptive energy threshold to function properly, the encoding pre-processor must have some prior knowledge of the energy of the input speech. This means that the pre-processor does not come into action at the very beginning of the speech processing. Instead, it

monitors the speech energy levels, especially that of unvoiced speech parts, at the beginning of the input speech, during which it adapts the energy threshold. After this initial adaptation period, the energy threshold function sets its value at around -30 dBr relative average power [102] and starts detecting and extracting the silent frames. As soon as the adaptive energy threshold is set, the encoder transmits a Silence Description (SID) frame [16]. The purpose of this SID frame is to trigger the comfort noise generation at the receiver with similar amplitude and spectrum to the noise at the transmitter. The comfort noise characteristics, including the adaptive energy threshold, can be further monitored and updated at regular intervals, according to the fluctuations in the background noise conditions, by the transmission of SID frames.

During silence the comfort noise generator produces a signal similar to the telephone line activity signal. The values of pitch period and LSFs are randomised for each frame in order to maintain continuity of the speech synthesis. The value of the time envelope energy is kept below the adaptive energy threshold, whereby the shape of the time envelope is randomised.

The threshold between voiced speech and silence is found using only signal power due to the large energy difference. However, the threshold between unvoiced speech and silence requires an extra discrimination stage due to the small energy difference. Therefore, the zero crossing rate, already described in Chapter 5, is also utilised for the finer discrimination between the two categories. The simulations and previous study [103] revealed that the zero crossing margin can be fixed at around 30% of the number of samples. Signals with the time envelope energy below the adaptive energy threshold are judged silent if their zero crossing number exceeds 30% of the number of samples. Otherwise, if the zero crossing number is less than 30%, the signals are declared unvoiced. The tests also showed that a forced conversion of a silent frame into an unvoiced frame can provide higher perceptual quality in the speech transition regions, due to the higher modelling accuracy of the unvoiced frame. However, the cost of this change is the inevitable increase in the bit rate, from 3 bits/frame for silence to 36-41 bits/frame for unvoiced.

The silent frames offer a substantial reduction in the average bit rate. A transmission of a single silent frame requires only 3 bits/frame. Moreover, when a group of four consecutive silent frames have been detected, the mode bit is set and only one silent frame is again sent to

the receiver. This implies that 3 transmitted bits (header and mode) can cover not one but four silent frames, producing an average of 0.75 bit/frame. When only silence is present, the average bit rate can drop down to 37.5 bps which represents the minimum attainable bit rate of the variable quantisation scheme simulated using the Time Envelope SB-LPC vocoder.

6.3.2 Coding of Unvoiced Segments

Unvoiced speech has a noise-like waveform with fairly weak near sample correlation. In addition, it exhibits very little long-term correlation. The key benefit from these two characteristics is the potential for bit rate reduction with respect to the fixed quantisation scheme. The variable quantisation scheme of the Time Envelope SB-LPC vocoder requires either 35 bits/frame, if a 6th order LPC filter is quantised, or 41 bits/frame, if a full 10th order LPC filter is quantised, for this speech category. The achieved saving lies in the range of 7-13 bits/frame when compared to the fixed quantisation scheme. If only unvoiced frames are coded and transmitted, the average bit rate can be kept as low as 1.75 kbps.

Referring to Table 6.1(b), it can be observed that no bits are allocated for the transmission of two parameters, a voicing decision and a pitch period. A split-band voicing decision, which requires 3 bits/frame in the fixed quantisation scheme, defines where the spectrum splits into voiced (lower frequency) and unvoiced (higher frequency) regions. However, being unvoiced, i.e. having no voiced harmonics, the unvoiced frame will always carry the same voicing decision. At the same time, as noise does not have a significant long-term correlation it will not require a pitch parameter. The pitch period of the unvoiced frame can be randomly generated during the speech synthesis on the decoder side.

The other significant saving in a bit allocation scheme is the reduction in the number of LP coefficients used by the LPC filter to represent the spectral envelope of the unvoiced segments. The statistical distribution of the spectral envelopes for unvoiced segments differs substantially from the general distribution of voiced segments. Typically, low frequency formants are most common in voiced speech, but for the majority of the unvoiced segments, especially unvoiced phonemes such as /z/, /s/ or /Θ/, most of the spectral energy is concentrated in the high frequency bands [101]. Therefore, a lower accuracy of shaping the

low frequency formants in the unvoiced spectrum should not greatly affect the overall spectral envelope representation of these segments. Simulations revealed that the 6th order LPC filter was sufficient for satisfactory modelling of the spectrum during unvoiced speech. When converted to an equivalent number of LSFs, it required only 20 bits/frame for quantisation, a saving of 6 bits/frame when compared to the quantisation of the 10 LSF parameters of the fixed quantisation scheme.

However, the simulations also showed that it was necessary to keep all 10 LP coefficients for modelling the unvoiced frames near the speech transition regions, in order to preserve the perceptual quality of the fixed quantisation scheme. The problem of locating the 'near-transition' unvoiced frames was solved with the introduction of an embedded clock buffer within the encoder. The task of this buffer is to aid in detecting the onset region. The clock of the buffer is set when an unvoiced frame has been declared by the encoder. The buffer then clocks in two more consecutive input frames. If all three frames happen to be declared unvoiced, the LP coefficients of the first input frame are obtained using the 6th order LPC filter. However, if any of the following two input frames is found to be voiced, the first input frame will use a 10th order LPC filter. The operation of the embedded clock buffer is depicted in Figure 6.4.

An extra signalling bit, called an LPC mode bit, is also transmitted during unvoiced frames to inform the decoder whether the 10th order LPC filter or the 6th order LPC filter is in operation.

The time envelope of the fixed rate Time Envelope SB-LPC vocoder is quantised using 12 bits/frame. The shape quantisation requires 5 bits/frame and the envelope energy quantisation requires 7 bits/frame. Unlike silence, the unpredictable nature of the noise time envelope forces the unvoiced speech category to use the full 12 bits/frame to maintain an accurate envelope quantisation and preserve the perceptual quality, especially in the near-transition regions.

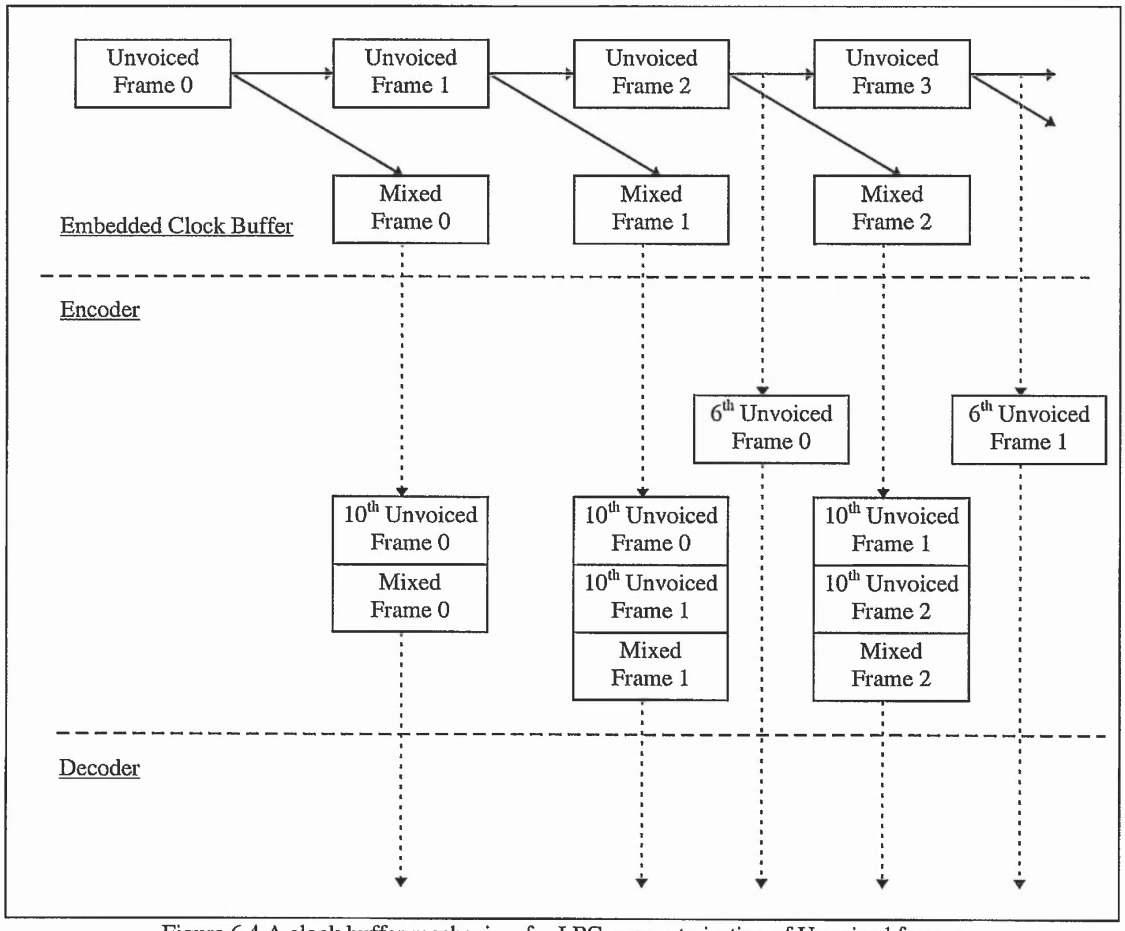


Figure 6.4 A clock buffer mechanism for LPC parameterisation of Unvoiced frames

The transmission of pure unvoiced speech requires only 35 bits/frame. In the mixed telephone speech, a small percentage of 'near-transition' unvoiced frames will be coded with 41 bits/frame, including 26 bits/frame for quantising the full 10th order LPC filter. One can therefore conclude that 35 bits/frame can be accepted as a general quantisation figure for the unvoiced frames, as shown in Table 6.1(b). An algorithmic delay of 60 ms is necessary for correct coding of the unvoiced frames. As it is shorter than the silence detection (80 ms) algorithmic delay, it does not affect the overall algorithmic delay of the encoding system.

6.3.3 Coding of Mixed Speech

A mixed sound, as the name implies, contains a mixture of both periodic and unvoiced speech which usually occurs in the form of dynamically changing transient speech. Rapid transitions from an unvoiced to a voiced sound (onset) or vice versa (offset) also belong to this mixed category. On the other hand, fully voiced sounds exhibit steady state quasi periodic features. It is essential to distinguish between these two voiced categories so that an efficient coding strategy and bit allocation scheme can be formulated. For example, the mixed segment will require a maximum number of bits for the quantisation of its parameters as it is typically non-stationary and perceptually very important. The fully voiced segment is on the other hand locally stationary with steady state speech and therefore requires less bits for the parameter quantisation.

Observing the bit allocation scheme presented in Table 6.1(b), leads to a conclusion that the quantisation of the mixed voiced segments is the most demanding in terms of number of bits used. With the overall number of 50 bits/frame it even exceeds the quantisation requirements of the fixed rate version, due to the extra two bits coming from the header speech classifier. If only mixed frames were to be transmitted, the variable rate quantisation of the Time Envelope SB-LPC vocoder would need to perform at its peak transmission rate of 2.5 kbps.

The greatest difficulty in conventional low bit rate speech coding lies with a detection and quantisation of onsets. The onsets correspond to rapid transitions from unvoiced to voiced sound, in which both energy and spectrum change rapidly. The signal is then only poorly predictable, and the prediction gain is usually small. Interestingly, for such segments the ear's sensitivity to spectral errors is also small, so that greater spectral quantising errors can be tolerated during transitions [104]. On the other hand, onsets carry vital phonetic information for distinguishing consonants, with exact timing being crucial. Therefore, the coder has to place more emphasis on onset coding by employing every available bit for the correct quantisation.

For mixed voiced frames both long-term (pitch period) and short-term (LPC parameters) predictions are needed to extract most of the speech information. At very low bit rates, the quantisation of the LPC parameters becomes especially critical since large spectral distortions

can severely affect the overall perceptual quality of the synthesised speech. It is, therefore, absolutely necessary to reserve a maximum available number of bits (26 bits/frame) for accurately quantising the 10th order LPC filter.

The time envelope of the fixed rate Time Envelope Split-Band LPC vocoder is quantised with 12 bits/frame. The shape quantisation requires 5 bits/frame and the envelope energy quantisation requires 7 bits/frame. Just like with the unvoiced sounds, the unpredictable nature of the voiced time envelope forces this speech category to use the full 12 bits/frame to maintain an accurate envelope quantisation and preserve the perceptual quality.

6.3.4 Coding of Fully Voiced Segments

The fully voiced speech, as the name implies, is the voiced segment of speech that does not contain any noise in its spectrum envelope. It could be regarded as a sub-class of the mixed voiced speech category was it not for the fact that it is considerably simpler to be quantised than the mixed class and can provide a further bit reduction in the variable quantisation scheme.

The main characteristic of the fully voiced speech segment is its steady state voiced nature with a slowly changing spectral envelope. Figure 6.1 illustrates the difference between the mixed voiced and fully voiced segments. The fully voiced frame can provide a bit saving on both the *intra-frame* and *inter-frame* level. This work only concentrated on the intra-frame level. At this level, the total saving when compared to the mixed voiced frame quantisation is 3 bits/frame coming from the redundant voicing decision information. Therefore, if only fully voiced speech is being transmitted, the average bit rate will be 2.35 kbps, just slightly less than the fixed rate due to the presence of a 2-bit speech classification header. On the inter-frame level, further bit savings can be achieved by exploiting the steady-state nature of the consecutive fully voiced speech frames. Taking into account that the spectral envelope of the fully voiced speech changes slowly over time, i.e. there is a high correlation between consecutive spectral envelopes, extra bits can be saved by inter-frame LPC parameter coding. The same idea can be applied to an inter-frame long-term (pitch period) coding, as the fully

voiced speech usually exhibits steady state pitch fluctuations. The inter-frame coding is an option that will need some future research attention.

6.4 Performance of the Variable Rate Quantisation

The performance of the variable quantisation scheme of the Time Envelope SB-LPC vocoder was assessed using several speech files consisting of both male and female speakers where approximately 30-40% was classified as unvoiced or silence. The tests conducted were aimed at achieving two goals. The first goal was to assess the speech quality of the new variable rate scheme. The second goal was to establish the average operating bit range of the variable rate coder.

The 2.4 kbps (fixed rate) Time Envelope SB-LPC vocoder was used as a benchmark for evaluating the perceptual quality of the variable rate quantisation coder. In the following Chapter 7, the MOS score of the fixed rate version was reported to be 3.30. This result was only slightly lower than the MOS score of the standardised 4.15 kbps I-MBE vocoder [81], which scored 3.45 during the same tests.

In order to assess the quality of the variable rate quantisation, the comparative A-B tests were conducted between the variable and fixed rate schemes. The MOS score of the fixed rate version served as an indication of the perceptual quality attained by the variable rate coder. The aim was to show how closely the two bit rate schemes perform, proving that no substantial reduction in speech quality occurs when the fixed rate version is substituted by the more flexible and efficient variable rate scheme.

The subjective comparison tests were made between the synthetic outputs of the variable rate coder and fixed bit rate coder. Tests were performed on 9 different speech files with a group of 20 listeners (both trained and untrained). For each speech file, both the variable bit rate version and the fixed bit rate version were played to the listener, who compared them, categorising one as 'better', 'slightly better', 'same', 'slightly worse' or 'worse' with respect to the other. The final results are shown in Table 6.2 and visually presented in the shape of a

pie-chart in Figure 6.5. The percentage results depict the frequency of choice for each category. It can be seen that the highest percentage came under the category 'same', followed closely by the 'slightly better' and 'slightly worse' categories. The resulting percentage of both extreme categories, 'better' and 'worse', was almost insignificant. These results show that a large proportion of the listeners were indecisive when asked to give even a slight preference to either a fixed or a variable quantisation scheme. This indecisiveness was good proof that the variable bit rate introduced very little or no quality degradation over the fixed bit rate. In almost 30% of all cases it performed even better than the fixed bit rate coder. The aim of this experiment was therefore achieved.

Table 6.2 Variable Bit Rate vs. Fixed Bit Rate A-B test results

	Better	Slightly Better	Same	Slightly Worse	Worse
Variable Rate vs. Fixed Rate	5.6%	24.5%	37.2%	29.4%	3.3%

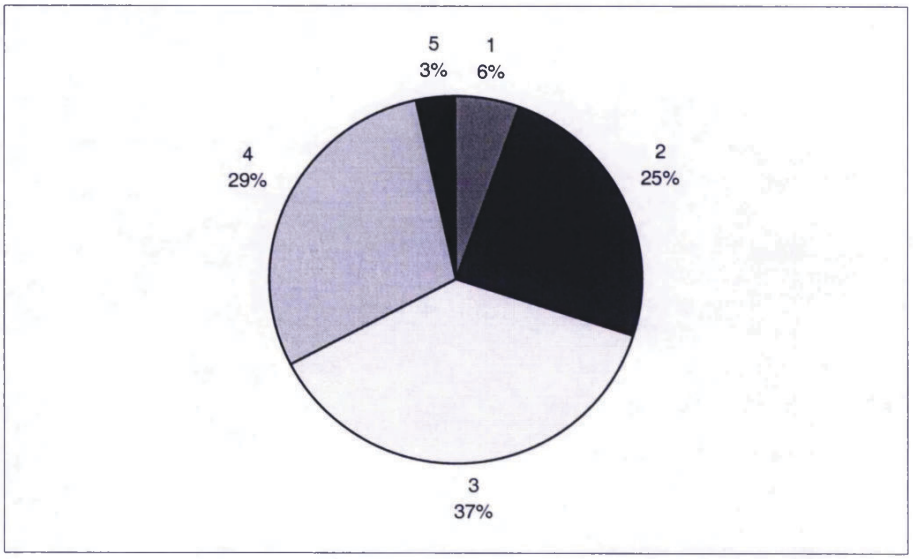


Figure 6.5 A pie-chart representation of the A-B comparison results

Legend:

1. Variable Rate better than Fixed Rate
2. Variable Rate slightly better than Fixed Rate
3. Variable Rate same as Fixed Rate
4. Variable Rate slightly worse than Fixed Rate
5. Variable Rate worse than Fixed Rate

Table 6.1(b) depicts the bit occupancy for each of the four speech categories - silence, unvoiced, mixed and fully voiced. The silent frames require the lowest rate of 3 bits/frame. Moreover, when four successive silent frames are detected, the silence mode bit is set and only one silent frame is transmitted to the decoder, lowering the transmission rate to just 0.75 bits/frame. This corresponds to the lowest bit rate of 0.0375 kbps. The highest rate of 50 bits/frame is needed for encoding mixed voiced frames. This corresponds to the highest bit rate of 2.5 kbps, which is only slightly higher than the fixed bit rate of the Time Envelope SB-LPC vocoder.

In order to find the average bit rate, the following simulation was performed on several speech files. For each speech file, the total number of frames for each speech category was extracted, and the average bit rate was calculated. The example results of this speech classification test are shown in Table 6.3.

Table 6.3 Speech Categorisation for three sample speech files

File	Silence	Unvoiced	Mixed	Fully Voiced	Average Rate
Sample 1	1%	26%	61%	12%	2.23 kbps
Sample 2	8%	20%	59%	13%	2.14 kbps
Sample 3	26%	12%	48%	14%	1.76 kbps
Sample 4	41%	13%	35%	11%	1.35 kbps

The tests showed that the operating average bit rate was in the range of 1.35-2.23 kbps. For the maximum average bit rate of 2.23 kbps, 73% of frames was voiced speech with only 1% of frames declared as silence. On the other hand, the minimum average bit rate of 1.35 kbps was achieved when 46% of frames was found to be voiced with 41% silences. It must not be forgotten though that these tests were performed only on the individual speech files that can be regarded as a one-way communication line. As pointed out in part 6.3.1, when the variable quantisation scheme is employed in a telephone link for a two-way conversation, where one side is silent roughly 60% of the time, one can expect further reductions in the operating average bit rate. Nevertheless, the presented results already show the viability of the new variable rate scheme.

6.5 Conclusions

In this chapter a new variable quantisation scheme has been proposed and integrated with a 2.4 kbps (fixed rate) Time Envelope Split-Band LPC vocoder. The scheme is based on an accurate source-dependent classification of different speech regions. The main speech categories identified are: silence, unvoiced, mixed and fully voiced. Each speech category is assigned a unique bit allocation scheme. By varying the allocation of bits in relation to the identified speech region, a variable coding rate is achieved.

The minimum bit rate is attained when only silence is transmitted. A 3 bit/frame option for a single silent frame can be further reduced to 0.75 bit/frame when four consecutive silent frames are detected and substituted by only one silent frame before the transmission. In this way, the minimum attainable bit rate is 37.5 bps. However, the silence detection mechanism requires an algorithmic delay of up to 80 ms, equivalent of four successive frames, four times longer than the algorithmic delay of the fixed bit rate scheme. This prolonged algorithmic delay is the only disadvantage of the variable rate system, price to be paid for a more efficient coding scheme.

The maximum bit rate is achieved when only mixed segments, containing both voiced and unvoiced speech, are transmitted. This maximum bit rate of 50 bits/frame (2.5 kbps) is slightly higher than the 48 bits/frame (2.4 kbps) utilised for the fixed bit rate due to 2 additional bits/frame representing the speech classification header.

The tests on individual speech files showed that average operational bit rate lies in the range of 1.35-2.23 kbps. A further reduction in the average bit rate can be attained when the variable rate quantisation scheme is employed as part of a telephone link for a two-way conversation, where one side is silent roughly 60% of time. The comparison tests between the variable bit rate and fixed bit rate indicated that any difference in the speech quality between the two versions was hardly noticeable.

The variable quantisation method can significantly reduce the bit rate, especially for silent and unvoiced speech categories. In order to make this method even more efficient, one must focus on a better bit rate utilisation when coding voiced speech. One idea would be to

introduce inter-frame coding of a pitch period and LPC parameters, especially for fully voiced speech, where there is a high correlation between consecutive spectral envelopes.

A more complex approach to make this variable rate scheme not just more efficient but also of higher perceptual quality would include the introduction of a variable frame length method, where a short frame is used to model dynamically active transient regions and a long frame is used for representing steady state speech bursts. The work on the variable frame length method, discussed in the following chapter, was a natural step forward in the exploration and investigation of new variable rate schemes applied to very low bit rate speech coding systems.

Chapter 7.

7. Variable Frame Length Speech Coding

7.1 Introduction

The variable rate quantisation system presented in the previous chapter proved it possible to significantly reduce the average coding bit rate by adopting the proposed speech classification method and varying the allocation of bits according to the identified speech region. The new variable rate coder managed to bring down the fixed bit rate of 2.4 kbps to a minimum average bit rate of 1.35 kbps without compromising the fixed bit rate quality. However, its performance was still limited by the use of a conventional fixed frame length scheme.

The fixed frame length scheme is based on the notion that speech can be assumed stationary when segmented into blocks of around 20 ms duration, the equivalent of 160 samples for an 8 kHz-sampled signal. Each speech block is windowed out and analysed for its parametric information to be extracted and transmitted. The 20 ms period represents a compromised and preferred solution for building an accurate parametric speech model. If the block length is shorter on one hand, the speech short-term energy will change very rapidly, which, in turn, can seriously affect the harmonic structure of the extracted signal; the shorter block length

also requires a more frequent update rate, i.e. a higher transmission bit rate. On the other hand, a longer block length will cause averaging of the short-term energy over a long time and, therefore, will not adequately reflect the changing properties of the speech signal. The block length of 160 samples provides a suitable practical solution capable of representing the harmonic structure fairly accurately and encompassing the long-term correlation (pitch period) range, which can span from about 16 samples for a high pitch female or a child, up to 150 samples for a very low pitch male.

However, the output speech quality of fixed frame length speech coders depends on their ability to code the most perceptually difficult regions of speech. These regions occur usually within voiced speech segments, especially during dynamically changing speech transients, such as voiced onsets and offsets. It is very difficult to accurately code very sharp onsets or offsets by using the fixed frame length, as their characteristics can dramatically change within the length of a frame. In order to account for these fast transitions within a frame, one has to resort to complex interpolation schemes. However, the interpolation can only provide a partial solution to this problem.

A more practical approach to accurately modelling the speech transitions is to make the frame length more flexible. In the case of speech onsets and offsets, there is a clear need to reduce the analysis frame size. By shaping the transition with a sequence of shorter frames, one can expect to be more successful in detecting and modelling its rapid dynamics. Shorter frames rely less on the use of sophisticated interpolation techniques, except for eliminating the block edge effects. The introduction of shorter frames can improve the output speech quality at the speech transition regions, but it also carries the burden of an increase in frame update rate. Figure 7.1 depicts typical voiced onset and offset regions and the two possible approaches to shaping it: using the fixed frame length or using the variable shorter frame sizes.

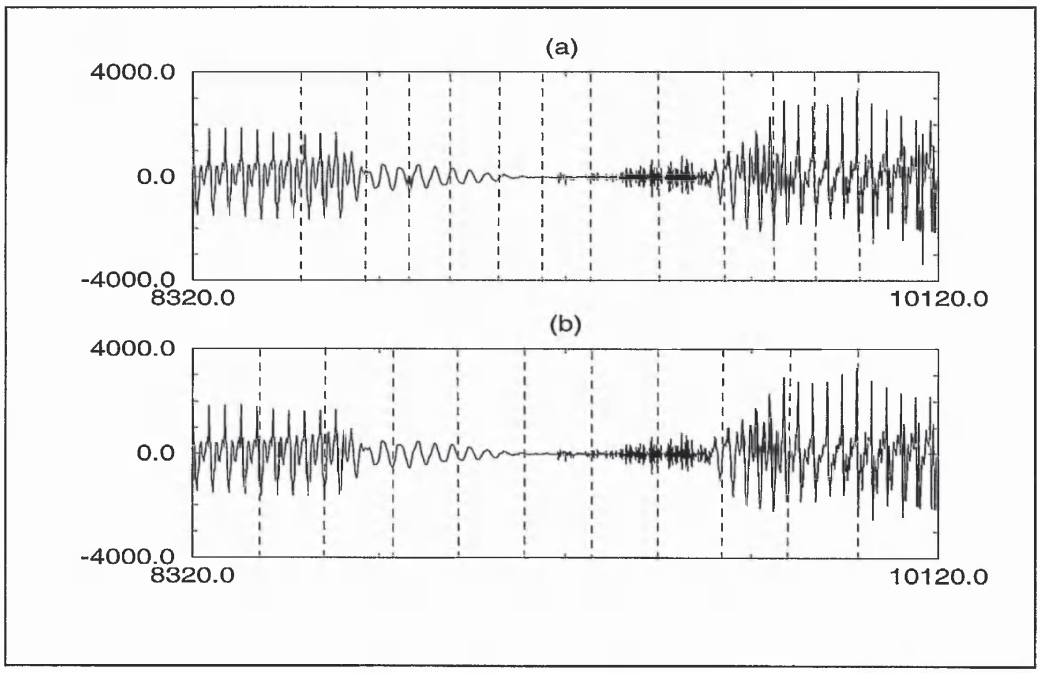


Figure 7.1 Variable frame length (a) vs. Fixed frame length (b)

The successful implementation of the shorter frame sizes inevitably questions the potential use of longer frame sizes. This becomes particularly relevant when modelling speech regions which do not significantly change their characteristics over a long period of time. The efficiency of the variable rate quantisation system lay in its ability to substantially reduce the average bit rate for non-speech regions, such as unvoiced and silence. The variable frame length technique can provide a further cut in frame rate by making use of longer frames when there is no speech present. Moreover, the frame extension method can also be applied in parts of speech which exhibit steady state quasi periodic features. Assuming that the main speech characteristics, such as pitch period, voicing and LP shape, vary at a very slow pace over a period of time, one can efficiently make use of longer frames for shaping these regions. However, there is a problem of potential averaging of the short-term energy due to the longer analysis frames. It can be solved by making use of modelling parameters, such as either time envelope energy or spectral amplitudes, which are able to adequately reflect the changing properties in the longer frame sizes. Although the biggest advantage offered by the use of the longer frame lengths is the reduction in the overall frame rate, the perceptual quality already achieved by the fixed frame length scheme in the steady state regions must not be compromised.

The function of the variable frame length system (VFLS) is based on the time varying characteristics of the speech signal such that longer frame lengths are used to code highly correlated steady state speech segments and unvoiced regions, while shorter frame sizes are employed for shaping the speech transients. The most important part of the VFLS is the accurate speech segmentation algorithm, for which the development of robust source-dependent criteria is of the utmost importance. The speech segmentation criteria are expected to produce a speech coder whose overall output quality is preferably better or at least matches that of the equivalent fixed frame length system. As it is a source-dependent algorithm, the average frame rate is expected to constantly vary; a speech material with predominantly short bursts of speech will require a higher update rate, whereas the one that contains longer speech periods will have its update rate significantly reduced. The variable frame length coding with its frame rate optimisation is highly suitable for use in voice communication services, such as Voice Store and Forward systems and digital answering machines, Internet phone, videoconferencing, or to augment low bit rate integrated digital packet networks.

This chapter discusses the variable frame length scheme which was applied to the Split-Band (SB) LPC vocoder described in Chapter 5. At first it introduces the concept of a speech segmentation and describes its various approaches and strategies that were investigated and developed during the course of this research. The chapter then gives a detailed system analysis of the *Time Envelope* version of the SB-LPC vocoder and assesses its performance under the variable frame length conditions. A similar approach is then adopted for analysis and evaluation of the *Amplitude Excited* SB-LPC vocoder where the Time Envelope concept is replaced by a more efficient and perceptually better Amplitude Excitation model. A third model, termed a *hybrid* model, is also introduced and described. It represents a synergy between the Amplitude Excited vocoder and the phonetically based speech recognition module. The performance of the three models is then compared and discussed, which is followed by a conclusions section, where thoughts on some future work are presented.

7.2 Speech Segmentation Methods

The basis of an accurate and reliable variable frame length scheme is the speech segmentation algorithm. The main characteristic of such an algorithm is its capability to segment the speech into blocks of various sizes, where each block can individually be described in terms of the speech modelling parameters, such as pitch period or voicing content. The segmented length can be extended as long as the main speech characteristics are preserved within that length. This idea introduces a concept of *frame length flexibility*, whereby speech is modelled according to its latent characteristics and not according to some pre-determined fixed block length mechanism.

The two most important segmentation features are the *selection of threshold levels* for the speech modelling parameters and the *selection of frame sizes*. According to the former feature the fluctuations in each modelling parameter must be confounded within a certain set of limits. This definition implies that if the size of the frame is to be extended, the changes in value of every one of the parameters used for analysing and modelling that frame must lie within some pre-defined threshold limits. For example, if the pitch period of a frame is used as an analysis factor, it can only be allowed to vary in its size by a maximum of ±2 samples during the frame extension, as it can be shown that any further increase or decrease in its value can lead to significant perceptual changes. These changes, in turn, can affect the fluctuations of other modelling parameters. The threshold values must therefore be carefully calculated. The latter rule deals with the selection of frame sizes. It concerns not just the minimum and the maximum frame lengths, but also the intermediate segmented step sizes, i.e. by what portion of its size is the frame length allowed to be extended. The frame size rule defines the criteria for incrementing or decrementing the speech segment length within its given minimum and maximum size values.

During the course of this research, three methods with different segmentation criteria were investigated. These segmentation methods can be classified as: phonetically-based segmentation method, multiple pitch segmentation method and sub-frame based segmentation method.

7.2.1 Phonetically Based Segmentation Method

The notion that speech can be considered as a sequence of phonemes was already introduced and analysed in the previous chapter. Phonemes can be defined as linguistically distinct speech sounds characterised by various acoustic properties. A phonetically based segmentation method is based on the theory of acoustic phonetics which postulates that there exist finite, distinctive phonetic units in spoken language which can be broadly characterised by a set of properties that are manifested in the speech signal, or its spectrum, over time. Even though the acoustic properties of phonetic units are highly variable, both with speakers and with neighbouring phonetic units, it is assumed that the rules governing the variability are straightforward and can readily be learned and applied in practical situations [105]. This acoustic-phonetic approach involves segmenting the speech signal into discrete (in time) regions where the acoustic properties of the signal are representative of one (or possibly several) phonetic units. An example of such an approach is shown in Figure 7.2.

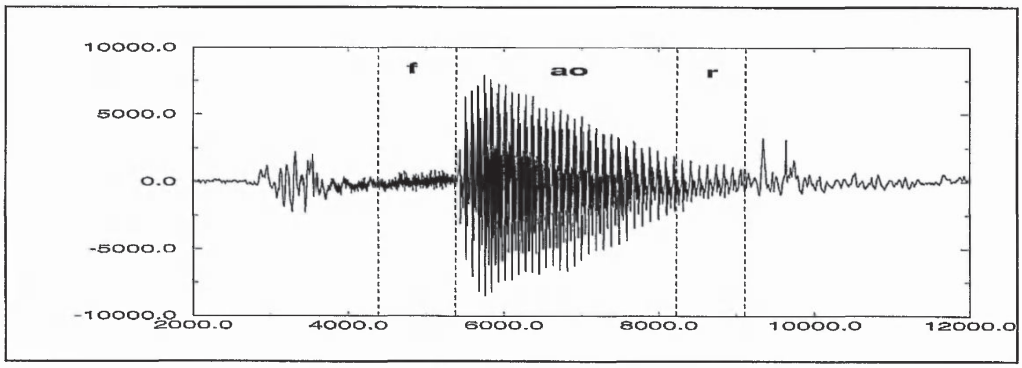


Figure 7.2 The phonetic segmentation of the word 'four' in time domain

The acoustic-phonetic approach in segmenting speech is quite often used and analysed in speech recognition systems [106]. The key aim of this type of speech recognition is the creation of a large vocabulary containing a range of phonemes, which can then be used in speech segmentation. One of the well-known phonetic libraries is the Texas Instruments/Massachusetts Institute of Technology (TIMIT) acoustic-phonetic continuous speech corpus of read speech [107]. TIMIT has been designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems. The research presented in this thesis also made use of TIMIT library contents in conjunction with a pre-processing speech recognition module

[108]. The task of the adopted speech recognition module is to perform phonetically based segmentation of speech that is to be coded by the developed variable frame length low bit rate speech coder. The two-module system is presented in Figure 7.3.

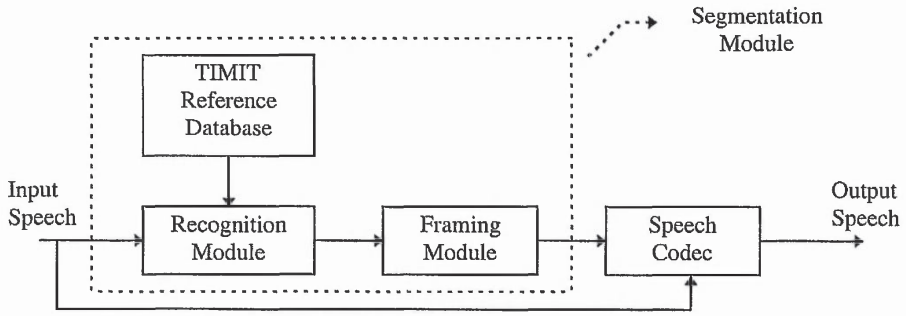


Figure 7.3 The block diagram of the two-module system

The phonetically based segmentation performed in the pre-processing speech recognition module can be split into three steps. The first step is the *speech analysis*, which provides an appropriate spectral representation of the characteristics of the time-varying speech signal. The second step in the segmentation is the *feature-detection*. The idea here is to convert the spectral measurements to a set of features that describe the broad acoustic properties of the different phonetic units. Among the features proposed are nasality, frication, formant locations, etc. [101]. The third step is the *segmentation and labelling,* whereby the system attempts to find stable regions (where the features change very little over the region) and then to label the segmented region according to how well the features within that region match those of individual phonetic units. The result of the acoustic-phonetic approach is the breakdown of the speech into a set of phonemes of various lengths. Before being passed onto the speech coder, each phoneme is further segmented into frames using the *framing module*, so that they fit within the frame size limits of the variable frame length coding method used.

Many problems are, however, associated with both the pre-processing speech recognition module that performs the speech segmentation and the speech coding module that performs the variable frame length analysis. The most prominent ones detected in the acoustic-phonetic approach of the speech recognition section are [105]:

- the method requires extensive knowledge of the acoustic properties of phonemes
- the choice of features is mostly based on intuition and is not optimal in a well-defined and meaningful sense

• phonetic classification is also not optimal

• no well-defined, automatic procedure exists for tuning the method (i.e. adjusting decision thresholds, etc.) on real, labelled speech. In fact, there is not even an ideal way of labelling the training speech in a consistent manner and one agreed upon uniformly by a wide class of linguistic experts.

The main design problem in coding the phonetically-segmented speech is the wide variety of the input phoneme sizes. The length of the phoneme extracted from the TIMIT speech corpus can lie anywhere between 5 ms and 400 ms (and longer) [107]. Although the assumption is made that the speech characteristics within a phoneme do not significantly change, one has to account for both the time-envelope variations, especially for longer frames, and inter-phoneme transition regions.

Due to all the aforementioned problems with designing an accurate and reliable speech recognition module, the phonetically based segmentation approach is seen as a novel approach in variable speech rate coding which remains an interesting idea but one that requires further research and understanding. However, quite a successful attempt was made during the course of this research to explore this new idea and develop a 'hybrid system' containing both the speech recognition and the speech coding module.

7.2.2 Multiple Pitch Segmentation Method

Pitch has been conventionally accepted as a description of periodicity contained in a speech signal. It can be represented as a non-zero time shift which results in a shifted speech segment which has the greatest perceptual similarity with the original segment. In the context of the human speech production mechanism, this representation is the elapsed time between two successive glottal pulses or the average interval of the glottal pulse for a short segment of the speech signal. Pitch frequency is the reciprocal of the pitch period.

A close study of the speech signal showed that the pitch period usually varies slowly and gradually in the voiced regions. It is only during transitions, such as onsets at the beginning and offsets at the end of a voiced speech interval, that the pitch period can change in a more

dramatic fashion. Accurate pitch period detection is a fundamental prerequisite for extracting other speech model parameters, e.g. voicing or LP shape. It can be observed that these other parameters change their values in the same fashion that the pitch period does; a slowly varying pitch is usually accompanied by a slowly varying voicing or LP shape, as depicted in Figure 7.4.

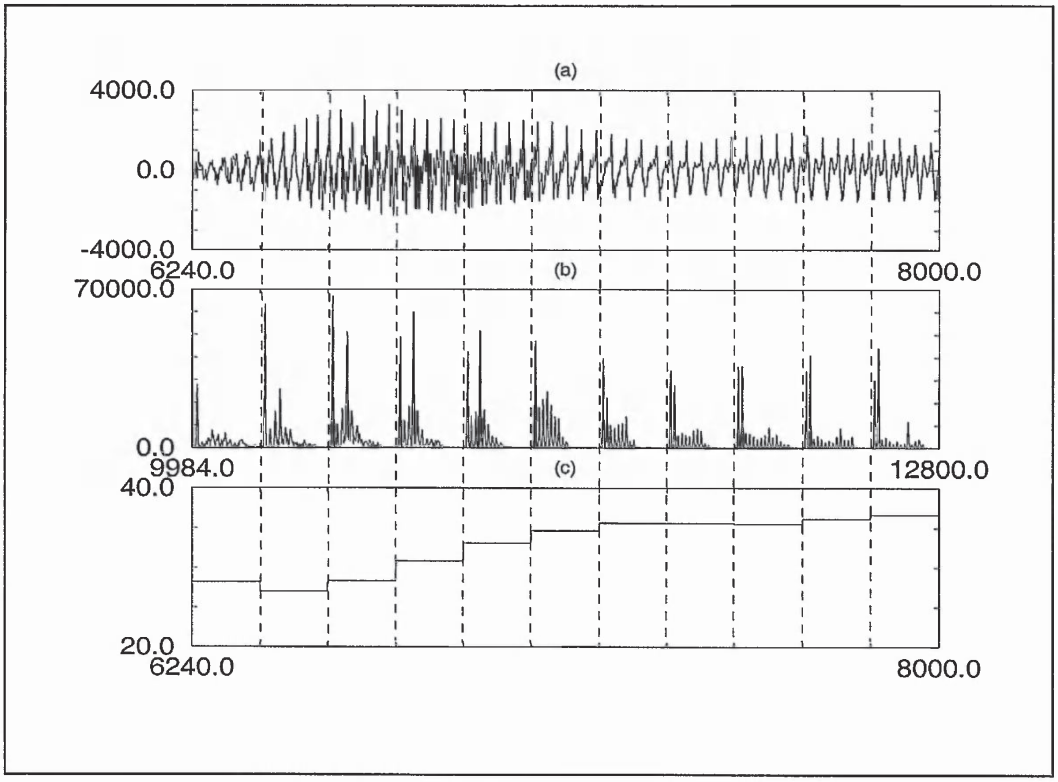


Figure 7.4 An example of a slowly varying speech

(a) time domain representation

(b) frequency domain representation

(c) pitch period fluctuations

Speech segmentation can very much benefit from these slow varying speech characteristics. The multiple pitch segmentation method is a variable frame length speech coding method that performs pitch-related segmentation of speech. This method relies on the accurate extraction and reconstruction of pitch information by the in-built pitch determination algorithm (PDA). The main aim of this method is to determine the size of the speech frame which is analysed at the encoder. The frame size is pitch dependent, whereby the minimum size of the frame equals the pitch period within that frame. The algorithm then attempts to extend the frame

length by adding the same pitch period found in the starting frame. If the analysis of the new frame size, now consisting of two pitch period units, shows no significant change in the pitch period, the algorithm adds another pitch period unit and repeats the analysis. The maximum size of the frame can be determined as either a certain length, such as for example 320 samples, or a certain permissible number of pitch period units. An example of how this method operates is shown in Figure 7.5, where a 315-sample frame is built using nine pitch period units, each unit characterised by a constant pitch period of 35 samples.

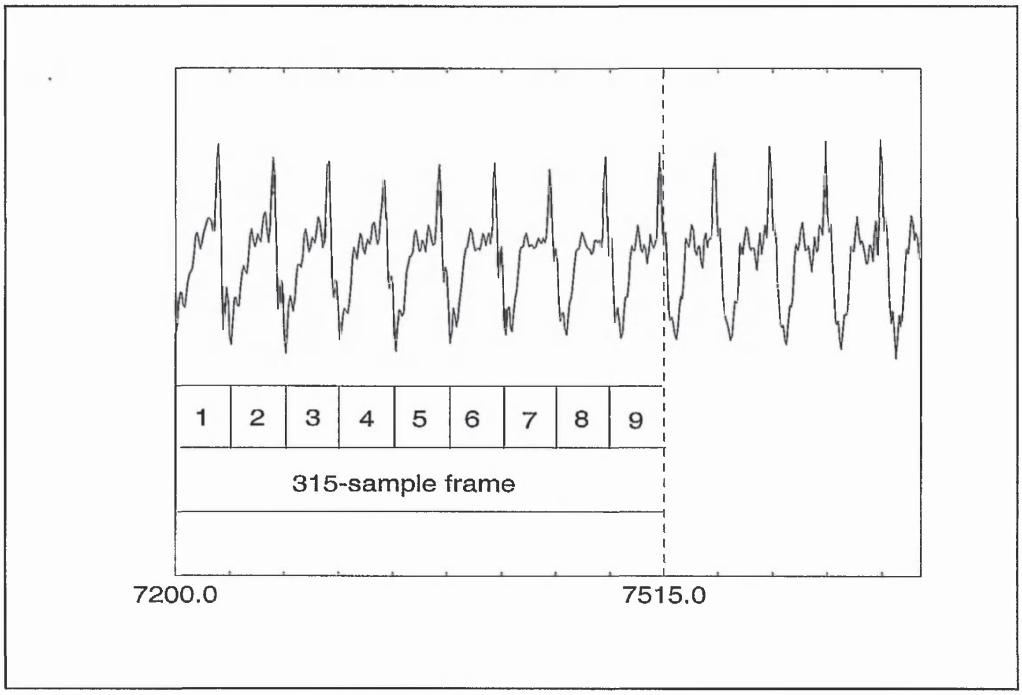


Figure 7.5 An example of frame extending using pitch period units

During the tests the PDA function of the Time Envelope SB-LPC vocoder, described in Chapter 5, was used, where the pitch period is allowed to vary between 20 samples and 116 samples. The starting frame length, which is used to extract the starting pitch period unit, is always 160 samples long, the equivalent of 20 ms at a sampling rate of 8 kHz. The reason behind this is that speech can be assumed stationary during 160-sample frame analysis, an option which is normally exploited in the conventional speech coding. Once the pitch period is detected, the starting frame length takes the size of the pitch period unit. The frame length is then extended by the same pitch period unit and again analysed for its periodicity. If the

pitch period of the new frame size equals the starting pitch period, the frame length is extended again and the calculations are repeated.

In order to make this method more efficient, the pitch period is allowed to vary to a certain extent within the extended frame length. Just how much it is permitted to fluctuate depends on the set threshold sizes. This research showed that the pitch period can fluctuate to a maximum deviation of ±2 samples. As the accurate pitch information influences the entire segment analysis, it can be shown that any deviation exceeding the above figure can have a detrimental effect especially on the frequency domain analysis.

The major drawback of the multiple pitch segmentation method is the wide variety of possible frame lengths. The maximum frame length must be limited to a certain number of pitch period units that will not require too many extra bits of information. In this work, the number of pitch period units was limited to four, as it required only two extra bits for transmission. By transmitting the pitch period and the number of pitch units, the decoder can easily determine the total synthesis frame length. However, this solution is not very efficient, as it ultimately limits the maximum frame size, especially for frame lengths containing short pitch periods. For example, the maximum size of a segment containing a 25-sample pitch period unit will only be 100 samples. The method is also inefficient with respect to very long pitch periods. For example, if a segment characterised by a 70-sample pitch period is extended to 140 samples for further analysis (twice the pitch period unit) and it shows a big change in pitch value, the transmitted frame length will again only be 70 samples long.

Large inefficiency is not the only problem that the multiple pitch segmentation method is faced with. There is also the question of complexity when synthesising the pitch-dependent variable frame lengths on the decoder side. In the fixed frame length solution, the interpolation and quantisation of parameters are based on the equal number of equally sized sub-frames. In the variable frame length coding, the conducted tests identified two general solutions for performing the interpolation and quantisation of the variably sized time envelope, but none can be regarded as efficient or low in complexity. Figure 7.6 illustrates the two approaches to solving the problem of pitch-dependent time envelope interpolation and quantisation.

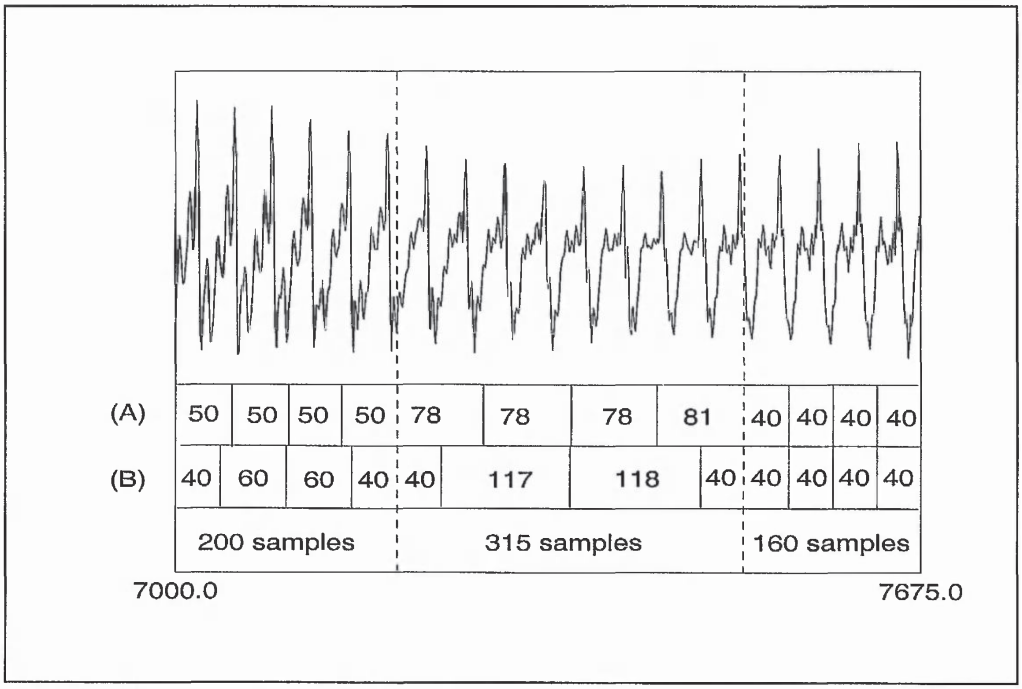


Figure 7.6 Multiple pitch segmentation method
(A) with equal number of sub-frames with (nearly) equal lengths
(B) with equal number of sub-frames, whereby the edge sub-frames are kept constant for interpolation purposes

Method A breaks the pitch-dependent variable frame length into a fixed number of sub-frames, where each sub-frame contains an average or nearly average number of samples. However, this solution makes the interpolation between the adjacent frames very difficult and complex. Method B keeps the fixed number of samples per sub-frame at the edges of the frame and varies the number of samples in the sub-frames positioned in the centre of the transmitted frame. This solution can improve the interpolation between the adjacent frames, but it can also seriously affect the interpolation inside the synthesised frame. In addition, the quantisation of the variably sized time envelope can prove to be too complex.

The multiple pitch segmentation method revealed during the course of this research that the optimal approach to efficiently and accurately segmenting speech must rely on accurate pitch detection and its preservation within the variable frame length. However, largely due to its inefficiency and design complexity when quantising and interpolating the time envelope, the pitch-dependent segmentation mechanism was abandoned and replaced with a sub-frame

based segmentation method that proved to be capable of solving most of the aforementioned problems regarding the efficiency of frame extending, the ease of interpolating and the simplicity of segmenting the variable frame length.

7.2.3 Sub-frame Based Segmentation Method

A major limitation of the fixed frame length coding is the inefficiency of its rigid coding configuration where the waveform is divided into frames with a fixed size extracted at suitable time intervals. The speech analysis and encoding is performed on the extracted frames whose characteristics are assumed to be instantaneous measurements which describe the speech waveform in the vicinity of a particular point in time.

The rigidity of the encoding structure can be 'softened' by the introduction of a sub-frame analysis. Although the speech analysis using the fixed frame length of a standard 20 ms duration can provide accurate information about the underlying pitch period, voicing harmonic structure and LP shape, it is unable to track down fast energy changes within the frame itself. The sub-frame analysis requires splitting the frame into shorter sub-frames of equal size. In the Time Envelope approach, sub-frames are analysed for their energy content; in the Amplitude Excitation approach, the inverse LP filtering on the sub-frame basis provides more accurate spectral amplitude determination of the signal, as shown in Chapter 5.

The key advantage of the sub-frame segmentation method is the *uniformity* of its approach. The idea of the sub-frame analysis can be adopted for any frame size, which itself can be further extended by adding more sub-frames. Instead of the pitch period unit of the variable length on which the multiple pitch segmentation method was based, the sub-frame segmentation method is based on the uniformly sized sub-frame units which substantially reduces the system's complexity on both the encoder and the decoder side and increases its overall efficiency.

The pitch detection algorithm is again the key segmentation criterion, as in the multiple pitch approach, but the difference here lies in the frame extension mechanism. This means that the starting analysis frame is extended by the sub-frame unit instead of the pitch period unit.

Although the extension steps are usually shorter in this case (the sub-frame unit is typically 2.5 ms or 5 ms long, whereas the pitch unit can range anywhere between 2 ms and 14 ms), the sub-frame segmentation can take advantage of the intermediate steps within a pitch cycle. For example, if the pitch in the starting frame is 70 samples or 8.75 ms (when sampling at 8 kHz), the next testing stage for the multiple pitch segmentation method is twice the size of the pitch (two pitch units) which equals 140 samples. If any of the thresholds that control the fluctuations in speech parameters, such as pitch or voicing, is exceeded, the extracted frame size remains at 70 samples. Considering the same example, the sub-frame segmentation begins its analysis with an 80-sample frame comprising two 40-sample sub-frames, as it is the shortest frame length that contains the initial pitch period and is the integer number of sub-frames. If none of the thresholds is exceeded, the method adds another 40-sample unit and repeats the threshold check. In this way, the minimum transmitted frame length is 80 samples or 120 samples long, which is more efficient than the multiple pitch segmentation method. Figure 7.7 illustrates this point by showing the two frame extending approaches with the detected pitch period of 50 samples.

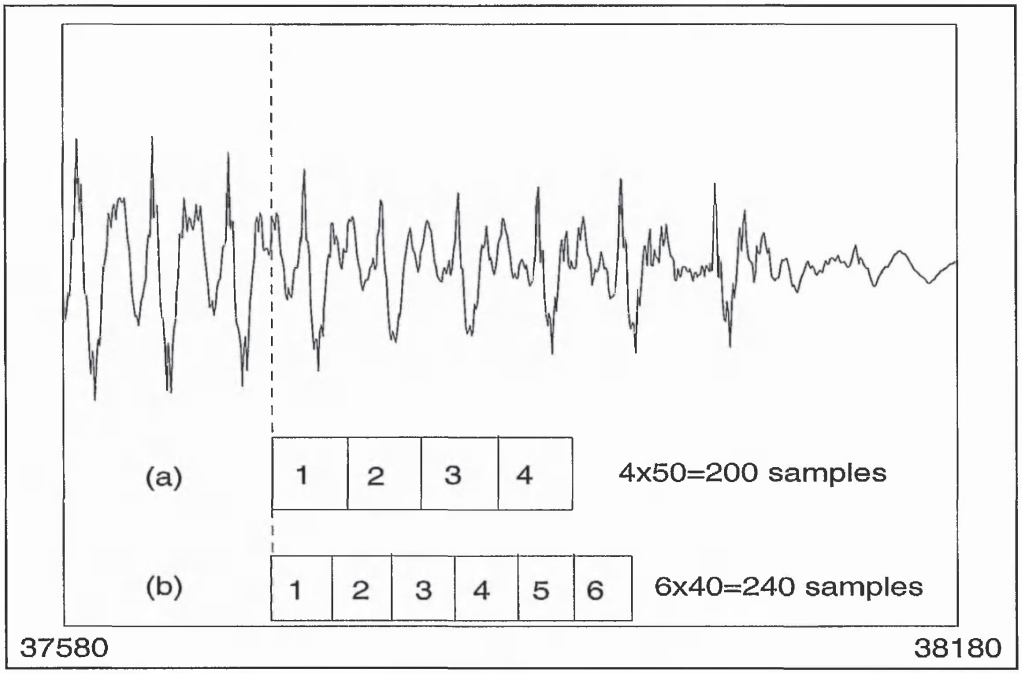


Figure 7.7 (a) Multiple pitch segmentation method vs. (b) Sub-frame segmentation method

In Figure 7.7, the frame size of the multiple pitch segmentation approach contains 4 pitch period units and is 200 samples long. The frame size of the sub-frame segmentation approach contains 6 sub-frame units of 40 samples and is 240 samples long yielding a saving of 20% in frame rate when compared to the pitch segmentation method. The efficiency of the sub-frame segmentation can be further increased for longer pitch sizes.

The lower complexity of the sub-frame segmentation method in relation to the pitch segmentation method is most prominent in the areas of analysis window utilisation, energy content representation and interpolation. A uniformly sized sub-frame uses a uniformly sized analysis window which can significantly reduce the computational load on the encoder side. The energy content of the sub-frame, viewed as either a time envelope or used for calculating spectral magnitudes, can be represented in a clear and simple fashion. However, sufficient care must be taken when encoding the variable number of sub-frame units, as is the case in the Time Envelope approach. The interpolation scheme on the decoder side can also be substantially simplified and made more accurate by adopting the sub-frame segmentation approach. It can now be performed on the sub-frame basis both between the adjacent frames and the adjacent sub-frames providing even better 'smoothing' of the changing speech parameters.

The research reported in this thesis successfully adopted and implemented the sub-frame based segmentation method as the preferred choice among the three speech segmentation approaches. The key advantages of this method are the efficiency of its frame extending mechanism, ease of interpolating and simplicity of segmenting the variable frame length. The sub-frame based segmentation method was simulated on the fixed frame length structure of the SB-LPC vocoder for both the Time Envelope and Amplitude Excitation approach.

7.3 Time Envelope SB-LPC Vocoder with the Variable Frame Length strategy

In this work, the variable frame length structure was first implemented on the Time Envelope SB-LPC vocoder operating at fixed bit rate of 2.4 kbps. Whilst the basic functioning of the

fixed rate vocoder was presented in Chapter 5, here we concentrate on the design issues and development strategy of the variable frame length system implemented upon the existing fixed rate coding structure. The main technical problems encountered are highlighted and discussed. The subjective performance of the new variable rate system is evaluated and compared with that of the fixed rate codec and the I-MBE speech coding standard.

7.3.1 Encoder

Figure 7.8 shows the block diagram of the encoder. Speech analysis is performed on 10 ms (80 samples) to 40 ms (320 samples) speech frames depending on the parameter characteristics of the 8 kHz-sampled speech. The frames are further divided into 2.5 ms (20 samples) sub-frames.

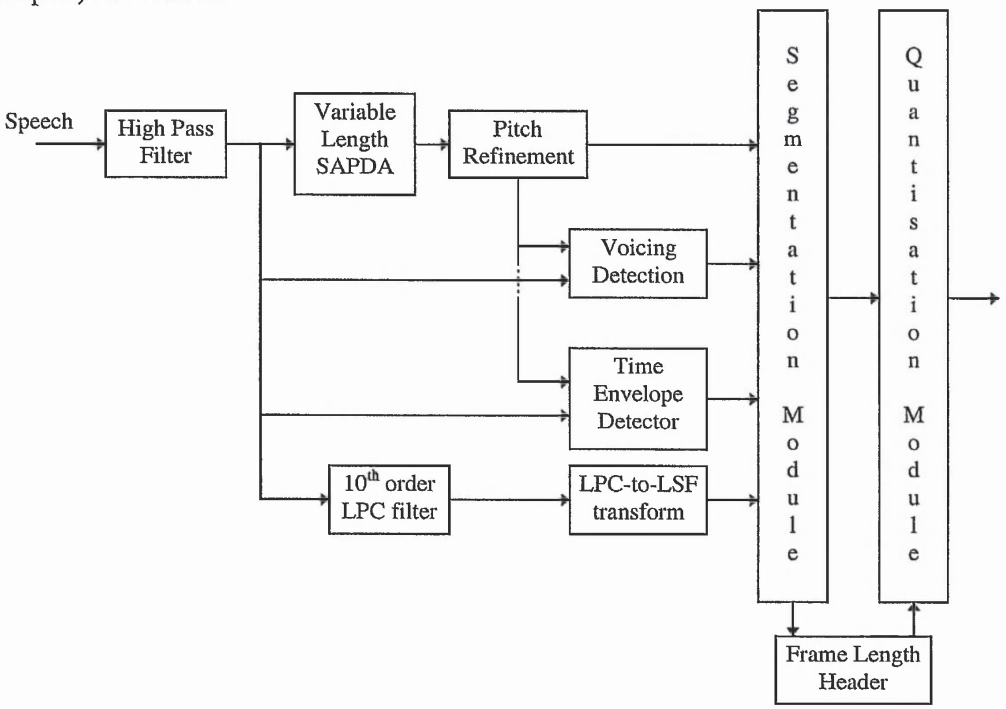


Figure 7.8 Encoder block diagram

The time domain based PDA, described in Chapter 5, is employed to determine the pitch period. However, the encoder does not look half-a-frame in advance as it conveniently did in the fixed frame length structure. Instead, a frame buffer is used on the decoder side, with a maximum length of 40 ms, for storing the current frame, while the previous frame is

synthesised through interpolation. A split-band voicing decision is made once per frame. The 10th order LPC filter is used for spectral modelling. The LP coefficients are quantised in the LSF domain using a two stage split vector quantiser. As with the fixed rate system, the logarithmic scale quantisation is used to encode the pitch information, whereas the time envelope energy is left unquantised due to the inherent complexity of the variable sized vector quantisation. The size of a frame is determined by a sophisticated 'search-and-compare' segmentation algorithm, based on the subjective and objective measurements of the speech parameters.

7.3.1.1 Pitch Detection for the Variable Frame Length System

The SAPDA pitch detection algorithm is similar to that of the time envelope fixed rate scheme which is used to detect the pitch period for the 20 ms (160 samples) frame size. However, the frame sizes are now allowed to vary between 80 samples and 320 samples which poses two major problems for accurate pitch period detection. Firstly, there is the problem of the pitch detection accuracy for the frames which are shorter than the standard 160 samples. The second problem lies on the other side of the scale with the pitch detection of the frames longer than 160 samples.

As the time domain PDA adopted for the fixed rate codec does not rely on pre-determined pitch period windowing, like the I-MBE codec as shown in Chapter 4, but on sub-frame analysis, the pitch detection can be confined within the boundaries of the short frames (80 or 120 samples). However, previous studies of the other codec structures that utilise the 80-sample or shorter frame lengths [109] as well as the conducted tests, all showed that the use of short frames for the pitch detection can be prone to errors, especially when longer pitch periods are dominantly present in speech. Tests also revealed that using a 160-sample length for the pitch detection yields very similar pitch values to that of shorter frames when the pitch steadily evolves. The search for the accurate pitch for the shorter frame sizes can then be performed by the pitch refinement routine in the frequency domain. The shorter frame sizes are very useful in tackling the problem of sharp speech onsets that may pass 'unnoticed' by the encoder if only 160-sample frames are used, as illustrated in Figure 7.9.

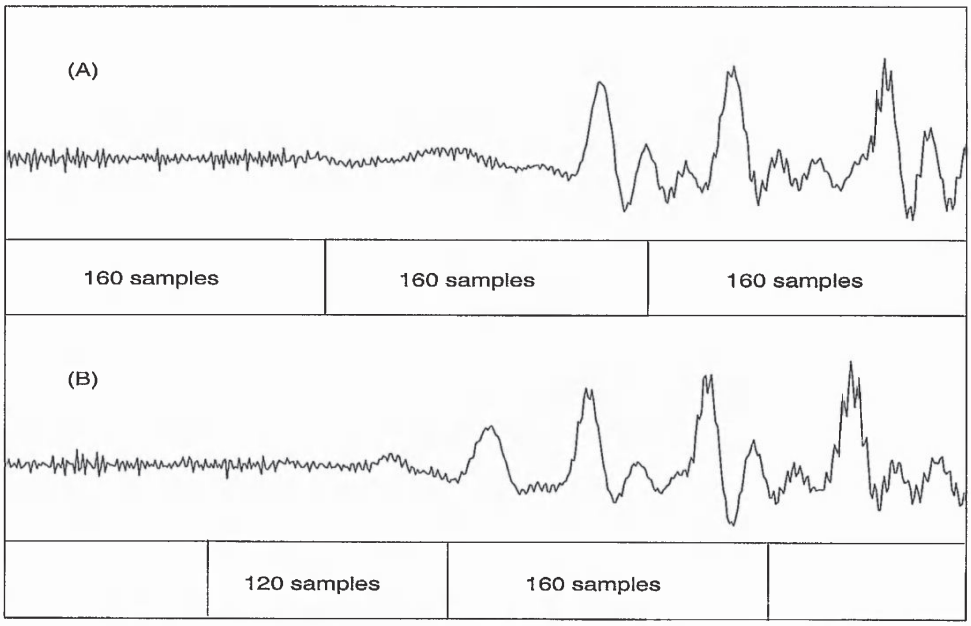


Figure 7.9 Onset detection using (A) fixed frame and (B) variable frame lengths

Figure 7.9 depicts a typical onset situation where a sudden change in pitch values and voicing at the beginning of the onset region is more accurately detected and modelled by the variable frame length method than the standard fixed frame length method.

The longer frame sizes also present a problem for reliable and accurate pitch detection. The study showed two possible approaches to its solution. The first approach uses the entire frame size to detect the pitch period. This option is logically and theoretically most appropriate. However, although the pitch may stay relatively constant, with small fluctuations within a large frame, this may not always be true for voiced spectral content. Tests revealed that even a small change in the pitch value can cause a significant change in the spectral structure and, consequently, a large change in the voicing decision. Therefore, the voicing decision is calculated only within a maximum of 160-sample frame, as in the fixed rate structure. This analysis frame is then conveniently shifted by a 40-sample sub-frame for the frame lengths equal to or longer than 200 samples, as depicted in Figure 7.10.

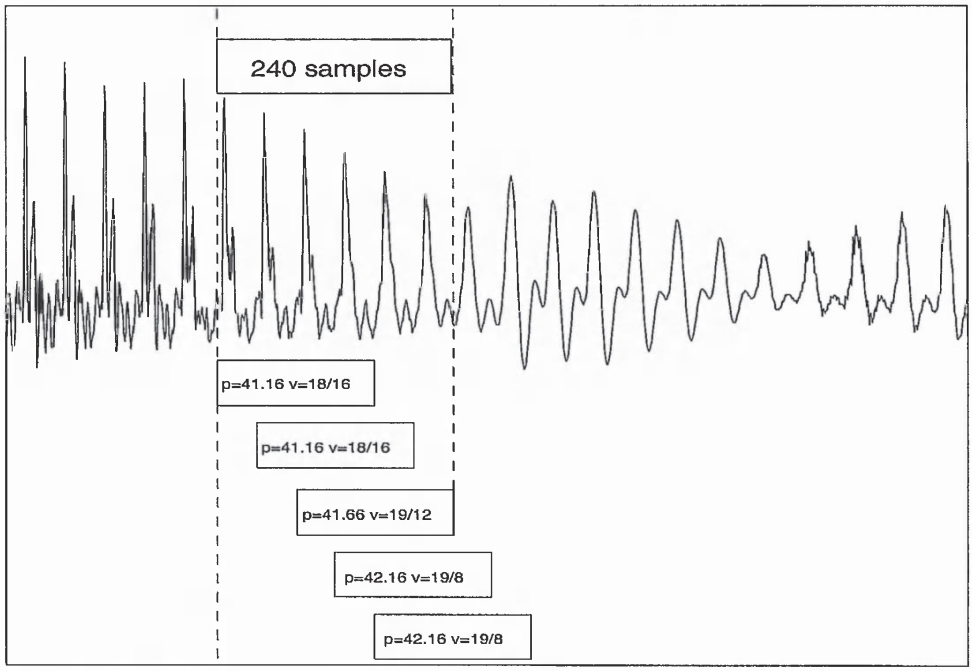


Figure 7.10 Detecting the change in pitch and voicing using the analysis frame shift method

In Figure 7.10 the frame is allowed to extend to 240-sample length as it is considered that both the pitch period and voicing information do not fluctuate by a large margin up to that length. However, any further extension would yield a significant change in the spectral structure of the synthetic speech.

This second approach, the sliding 160-sample analysis window was adopted for both pitch detection and voicing determination for longer frame sizes. The two key advantages of this approach are the improved accuracy and the reduction in the encoder's computational complexity.

7.3.1.2 Voicing Decision for the Variable Frame Length System

The split-band voicing decision, introduced for the fixed rate structure, remains structurally unchanged in the variable frame length scenario. The accurate voicing decision is primarily based on the accurate PDA, and, therefore, the size of the voicing analysis window equals that of the PDA analysis window for both the short and long frames.

7.3.1.3 LP Modelling for the Variable Frame Length System

The tenth order LPC filter is used for spectral modelling. The LP coefficients are quantised in the LSF domain using a two stage split vector quantiser.

7.3.1.4 Envelope Energy Calculations for the Variable Frame Length System

The RMS energy and time envelope shape are calculated on a sub-frame basis. The only difference to the existing fixed rate algorithm is presented in the number of sub-frames within a given frame. This does not create problems in calculating the parameters but it does put a great burden upon the quantisation routine. The time envelope energy is, therefore, left unquantised due to the inherent complexity of the variable sized vector quantisation.

7.3.1.5 'Search-and-Compare' Segmentation Algorithm

The search for the most suitable frame size is source-dependent. The 80-sample or 120-sample starting frame, the size which is dependent on the calculated pitch period, is analysed for its parameters and then extended by 40-sample segments up to a maximum length of 320 samples. At each stage of frame extension, the speech parameters are calculated and compared with those that characterised the starting frame. The novel three-level 'search-and-compare' segmentation algorithm contains a set of objective and subjective thresholds, shown in Table 7.1, used to detect any significant change in each tested speech parameter.

Table 7.1 A three-level decision making algorithm for optimum frame size

Decision level	Parameter	Threshold
1.	Pitch period	≤ ±2 samples
2.	Voiced Harmonics 2-12 Voiced Harmonics 13-18 Voiced Harmonics 19-24 Voiced Harmonics 25-	≤±2 ≤±3 ≤±4 ≤±5
3.	Envelope energy change	≤50%

The pitch period threshold of ±2 samples was derived objectively. The tests showed that if the pitch value changes by more than this threshold, it will automatically trigger the change in the spectral structure, as the total number of harmonics changes. Pitch fluctuations smaller than the above threshold were also found to affect the spectral structure, especially the voicing

decision. Therefore, a variable threshold on the voicing decision variation was developed which further controls the frame extending mechanism.

The thresholds determining the number of voiced harmonics were derived through subjective listening tests. They clearly highlight the importance of the low frequency harmonic structure, as only a minimum fluctuation of ±2 is allowed within the first 12 harmonics. However, the threshold gets further relaxed as the number of harmonics increases. An upper limit of around ±5 is set at the 25th harmonic or above.

The RMS energy envelope is allowed to change its value up to a maximum of 50%. This percentage threshold was set through the objective tests in order to aid the onset indication or to prevent the use of longer frames for parts of speech containing large energy fluctuations. Two different approaches were identified during the course of the simulations. At first, the average energy envelope over the entire current frame was compared with that of the previous frame. However, this strategy failed to detect the sharp energy changes, especially within the longer frames. The adopted second approach is sub-frame based, whereby the RMS energy change is monitored from one sub-frame to the next. In this way, the energy fluctuations can be identified in the course of the frame extending regardless of the overall frame size.

The optimum frame that is encoded is the longest frame whose speech parameters lie within the thresholds at all three decision levels. As a rule of thumb, the speech transitions are usually modelled with shorter frames, whilst the slowly-varying speech segments and the unvoiced regions (including silence) are modelled with longer frames, examples of which are shown in Figure 7.11.

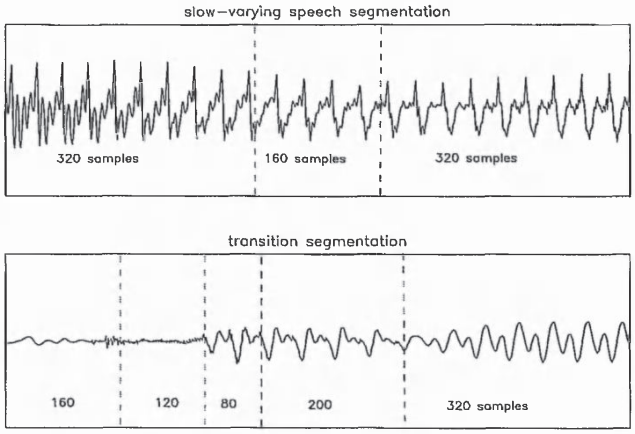


Figure 7.11 Frame segmentation example for slow-varying and fast-varying speech

In order to notify the decoder on the size of the encoded frame, an additional 3-bit header is also incorporated in the bit rate structure yielding a theoretical overall bit rate of 2.55 kbps (if quantisation of the variable length time envelope is assumed to require equal number of bits used for the fixed length time envelope).

7.3.2 Decoder

Figure 7.12 shows the decoder block diagram. Once the speech parameters are transmitted, the pitch information is used to produce a linearly interpolated pitch contour. As in the fixed rate decoder, this contour is used to generate unity energy synthetic voiced excitation based on the sinusoidal summation approach. The LSF interpolation is performed once per sub-frame using the decoded envelope and voicing information to determine the linear interpolation factor. Unity unvoiced energy is also generated and the voicing decision is used to switch either the unvoiced or voiced excitation into input of the LPC synthesis filter which is followed by the perceptual weighting filter. A high pass filter is included to prevent any steady state bias present in the signal from effecting the energy scaling process.

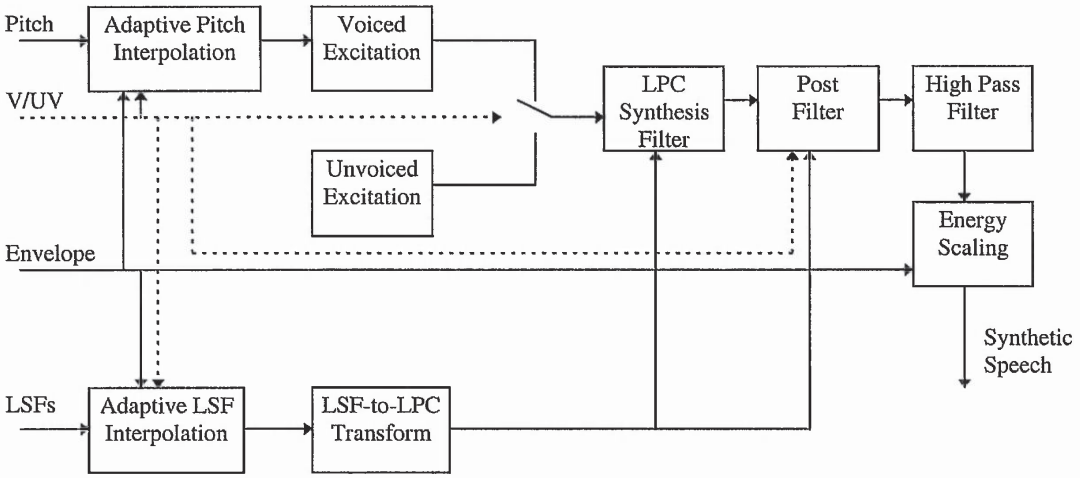


Figure 7.12 Decoder block diagram

The main problem with the variable frame length synthesis is caused by the fact that the decoder can no longer utilise the standard interpolation routine of the fixed frame rate scheme. The half-a-frame delay used by the fixed rate decoder is now replaced by a one-frame delay rendering a maximum delay of 40 ms (320 samples). The one-frame delay offers the possibility of backward and forward interpolation of the decoded frame. It means that the interpolation can be performed at both ends of the decoded frame. This is very important when interpolating between frames of different sizes. However, new rules concerning the amount of pitch and LSF interpolation needed for various frame sizes had to be introduced during the course of this work, which led to the development of an adaptive interpolation algorithm.

7.3.2.1 Adaptive Interpolation Algorithm

The function of the adaptive interpolation algorithm is to define the amount of interpolation needed to smooth out transitions between the adjacent frames of variable sizes. The algorithm is concerned primarily with the interpolation of the pitch period and the LSF parameters that characterise the current decoded frame. As the frames are allowed to vary from 80 samples to 320 samples, the decoder has to adapt the interpolation size according to the lengths of the adjacent frames. The adaptive interpolation algorithm must therefore be able to identify and accommodate a wide range of interpolation sizes regarding both the back frame transition

(interpolation between the end of the previously decoded frame and the start of the current frame) and the front frame transition (interpolation between the end of the current frame and the start of the buffered follow-up frame). An example of the variable frame length interpolation is given in Figure 7.13.

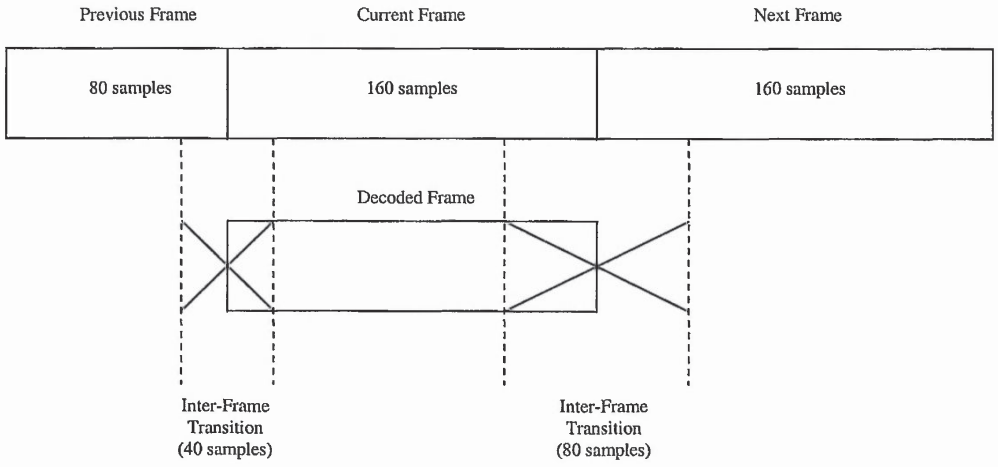


Figure 7.13 Adaptive inter-frame interpolation

The adaptive interpolation algorithm is based on the frame length boundary classifier and operates on the sub-frame basis. The algorithm takes into account the voicing of the adjacent frames and the change in the pitch period. The frame length boundary classifier distinguishes between all the possible cases of the frame length boundary conditions, starting from the case where the interpolation is performed between two 80-sample frames, then between an 80-sample frame and a 120-sample frame, up to the case where two 320-sample frames are adjacent. The interpolation is energy dependent and linear, where the value of the gradient depends on the number of boundary sub-frames that are included in the adaptive interpolation algorithm. The complex structure of the algorithm can be observed in Table 7.2. The variety of inter-frame transition sizes is a result of objective and subjective tests conducted during the research work.

Figure 7.2 Rules of the adaptive interpolation

Frame A (in samples)	Frame B (in samples)	Inter-frame Transition (in samples)
80 or 120	80-320	40
160 or 200	80-120	40
	160-320	80
240 or 280	80-120	40
	160-200	80
	240-320	120
320	80-120	40
	160-200	80
	240-280	120
	320	160

The scaling of the time envelope and the interpolation of the sub-frame based envelope energy segments are performed according to rules similar to those used by the fixed frame length decoder.

7.3.3 Performance Discussion

The performance of the variable frame length structure of the Time Envelope SB-LPC vocoder was assessed using the MOS scale, whereby the subjective quality of the variable frame length scheme was compared with the subjective quality of the fixed frame length scheme. Moreover, the relevant fixed rate speech coding standard (4.15 kbps I-MBE) was included in the tests as a reference point. Sixteen subjects, mostly experienced listeners, were used to conduct the informal MOS listening test that included two male and two female sentences. The results are presented in Table 7.3 below.

Table 7.3 MOS test results for variable frame length Time Envelope SB-LPC vocoder

Type of codec	MOS
4.15 kbps I-MBE vocoder	3.45
2.4 kbps Time Envelope SB-LPC vocoder (fixed rate)	3.30
120-320 sample variable frame length Time Envelope SB-LPC vocoder	3.28
80-320 sample variable frame length Time Envelope SB-LPC vocoder	3.22

The listening test results show that the variable frame length codec versions performed only slightly worse than the equivalent fixed rate codec. The variable frame length algorithm that included the 80-sample frames was given the lowest overall MOS mark. There are at least two major reasons for such performance. Firstly, this result proved the level of difficulty and complexity involved in both the encoding and decoding of the 80-sample frames, especially with respect to the sudden pitch and voicing changes that usually characterise these short frames. Secondly, the presence of such a short frame length causes the adverse block edge effects which characterise the short interpolation length within the frame boundaries. However, other tests revealed that some 80-sample frames were very important in modelling sharp onsets in speech. Therefore, a total removal of this option would not be an ideal solution. On the contrary, a possible solution to this problem would be the selective use of the 80-sample frame, i.e. the inclusion of the 80-sample frame length only for the set of specific, primarily onset, conditions. Defining these conditions represents an area of potential future work on this codec.

Tests were also conducted to show the potential reduction in frame rate of the variable frame length codec as opposed to the 2.4 kbps fixed rate codec. The tests revealed that the 120-320 sample variable frame length algorithm reduced the total number of voiced frames by up to 20% and unvoiced frames by up to 32%. The amount of reduction was largely speaker-dependent and it averaged at 7% for voiced frames only. On the other hand, the 80-320 sample variable frame length structure did not show any significant saving in the overall frame rate, the result which further questions the potential of this scheme unless the use of the 80-sample frames is made selective, as already mentioned.

The variable frame length scheme showed the best performance for steady state speech with gradual parameter change, both in terms of perceptual quality and frame rate reduction. At this stage of work, the 120-320 sample variable frame length algorithm proved to be the optimal choice as it was capable of retaining good speech quality and lowering the fixed frame rate.

The biggest drawback of this Time Envelope approach is more objective than subjective in nature. Although the perceptual quality attained by the variable frame rate Time Envelope SB-LPC vocoder was comparable to that of the fixed frame rate scheme, the key problem

represents the full implementation of the algorithm, namely, the quantisation of the variably sized time envelope. The variable frame length strategy requires a more flexible approach to quantisation, whereby the modelling parameters are not constrained by the fixed frame length structure. Switching to the Amplitude Excitation approach, where the majority of parameters are based in the frequency domain free of any frame length constraints, represented therefore a natural way forward in this research.

7.4 Amplitude Excited SB-LPC Vocoder with the Variable Frame Length Strategy

The variable frame length structure of the Time Envelope SB-LPC vocoder proved to perform well when compared to its fixed rate alternative. However, the need to further improve the perceptual performance coupled with the practical difficulty in quantising the variably sized time envelope caused the research interest to shift at this point from the predominantly time domain approach to the frequency domain approach. The time envelope information was replaced by the spectral amplitude information, which has the advantage of being independent from the size of the frame. This makes the amplitude approach more suitable for variable size quantisation, as the number of bits used for spectral envelope quantisation remains constant for any frame size.

In the following part of the thesis we concentrate on the design aspects of the new amplitude driven scheme, most of which is based on the existing time envelope scheme. Again, the main technical problems encountered are highlighted and discussed. Also, the subjective performance of the new variable rate system is evaluated and compared with that of the fixed rate codec and of the current speech coding standard.

7.4.1 Encoder

Figure 7.14 shows the block diagram of the encoder. Speech analysis is performed on 10 ms (80 samples) to 40 ms (320 samples) speech frames depending on the parameter

characteristics of the 8 kHz-sampled speech. The frames are further divided into 5 ms (40 samples) sub-frames.

Unlike the Time Envelope algorithm that used one frame length delay, the encoder here looks half-a-frame in advance as it conveniently did in the fixed frame length structure. The half-a-frame delay offers a higher degree of overlap between the previous and current encoded frames during the interpolation phase on the decoder side, where a frame buffer of a maximum delay of 20 ms is used. The frequency domain based pitch detector algorithm, SMM-PDA, described in the second part of Chapter 5, is employed here to determine the pitch period. A split-band voicing decision is made once per frame. The 10th order LPC filter is used for spectral modelling. The LP coefficients are quantised in the LSF domain using a two stage split vector quantiser. The quantised LP parameters are then used to find the LP residual, which is required for determining the excitation harmonic amplitudes. Both the speech signal and the LP excitation are transformed into the frequency domain using a 512-point FFT. The size of a frame is determined by the 'search-and-compare' segmentation algorithm, already introduced, with an added spectral comparison stage.

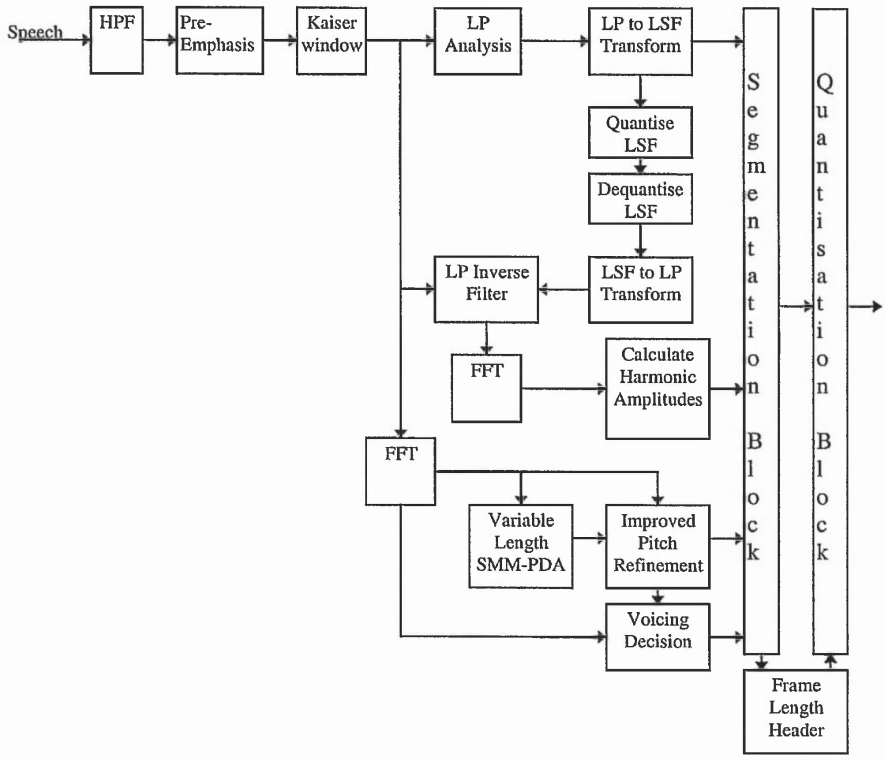


Figure 7.14 Encoder block diagram

7.4.1.1 Pitch Detection for the Variable Frame Length System

The SMM-PDA, used in the fixed rate Amplitude Excited SB-LPC vocoder operating at 2.4 kbps and described in Chapter 5, is also employed here for the variable frame length system. However, the SMM-PDA had to be modified during the design stages as the frame sizes were now allowed to vary between 80 samples and 320 samples. This frame size variability posed two major problems for accurate pitch period detection. Firstly, there was the problem of pitch detection accuracy for the frames shorter than the standard 160-sample frame. Secondly, there was the problem on the other side of the scale with pitch detection of the frames longer than 160 samples.

The key feature of the SMM-PDA algorithm is that it performs the pitch detection entirely in the frequency domain. This frequency domain approach proved to be more accurate and more reliable than the time envelope approach, as shown in Chapter 5. However, the main drawback of the SMM-PDA is that it does not operate on the sub-frame basis but relies on the FFT conversion of the entire frame into the frequency domain. Therefore, adjusting the SMM-PDA for the variable frame length approach was not as straightforward a task as it had been with the time envelope SAPDA algorithm.

The study of the SMM-PDA algorithm offered two potential solutions to its variable frame length modification. The first solution took a very straightforward approach: each frame size was 'windowed out' by its pre-determined window size and converted into the frequency domain using a constant 512-sample FFT function, which could generally give sufficient accuracy and resolution for the proposed frame size range. Although complex, due to the presence of various window sizes, this method was theoretically sound as the pitch detection was performed using the entire length of a given frame. However, tests showed that this method could not be applied successfully due to the complexity of the SMM-PDA algorithm which had to be tuned for every single window size, i.e. each frame length required retuning of the algorithm's factors and parameters. This first approach was therefore abandoned for the second, more flexible solution based on the use of the constant window size throughout the frame length search.

The idea behind the second solution to the problem of adapting the SMM-PDA for the variable frame length approach was to keep the window size constant regardless of a given frame size. The 221-sample Kaiser window proved to work well not only for the 160-sample fixed frame length, but also for all the frame lengths in the region from 80 to 200 samples, as they can all be positioned in the centre of the 221-sample window. However, frames longer than 200 samples can not be placed in the centre of the Kaiser window and hence a portion of the frame is to be inevitably lost for the windowing purposes. The solution was to place the window in the centre of the frame length instead. As the maximum frame size is allowed to be 320 samples, the 221-sample Kaiser window is still able to contain a very large portion of any frame longer than 200 samples. This was confirmed by the tests, which showed high accuracy of the SMM-PDA for longer frame lengths of up to 240 samples and slightly reduced accuracy for the frame lengths of 280 and 320 samples. As the very long frames, such as 280 or 320 samples, are rarely used due to the typical non-stationary character of speech, the SMM-PDA proved to be highly reliable and accurate as the variable frame length pitch detector. It offers ½ sample precision performance which is further enhanced by the modified SSM pitch refinement method with its ¼ sample accuracy.

The key modification made to the improved SSM pitch refinement method is the introduction of the pitch-dependent refinement window size for the frames shorter than 160 samples. The reason for this is the fact that the 221-sample Kaiser window, used for gross pitch extraction, proved to be too long for accurate pitch refinement. The tests showed that this was particularly valid in cases where a short frame contained a short pitch period. Therefore, the pitch refinement window needed to be adjusted according to the gross pitch period detected for the short frame lengths. An adaptive pitch refinement windowing method, previously suggested in [70], was introduced according to the following decision rules given in Table 7.4 below:

Table 7.4 Rules on pitch refinement window size

Rough pitch period	Pitch refinement window
>= 70 samples	221 samples
>= 55 samples	181 samples
>= 40 samples	161 samples
< 40 samples	121 samples

The tests showed that the accuracy of the SSM pitch refinement method for the shorter frames (80 and 120 samples) was greatly improved by the introduction of this pitch-dependent windowing technique.

7.4.1.2 Voicing Decision for the Variable Frame Length System

The split-band voicing decision, introduced for the fixed rate structure, remains structurally unchanged in the variable frame length scenario, including the additions, such as the zero-crossing rate, low-band to high-band energy ratio and fractional energy ratio. The accurate voicing decision is primarily based on the accurate PDA, and, therefore, the size of the voicing analysis window equals that of the PDA analysis window for both the short and long frames.

7.4.1.3 LP Modelling for the Variable Frame Length System

The tenth order LPC filter is used for spectral modelling. The LP coefficients are quantised in the LSF domain using a 24-bit split vector quantiser based on moving average prediction similar to the one used in [110][111][112] .

7.4.1.4 Determination and Quantisation of Harmonic Amplitudes

The determination and quantisation of the spectral content of the LP residual signal is performed in a similar manner to that applied in the fixed rate scheme described in Chapter 5. However, the introduction of the variable frame length approach inevitably required some system modifications to be made during the design stages.

The *first modification* was the introduction of the variably sized window length for the LP residual. The LP residual is now calculated on the 2.5 ms (20 samples) sub-frame basis and its window length depends on the current frame size under analysis.

The objective of the *second modification* was to improve the accuracy of the amplitude determination, especially for longer frame sizes. This was achieved by performing amplitude

determination on a half-frame basis, i.e. the set of amplitudes is calculated twice per frame. This extra spectral information proved to be of particular importance for the frame sizes longer than 160 samples, which was confirmed by the tests. The tests showed that although the pitch and voicing decision usually stay relatively unchanged and stable even for longer frame sizes, the actual spectral content typically varies more significantly. Therefore, calculating amplitudes on a half frame basis essentially doubles the accuracy of the spectral content which is transmitted to the decoder. This extra spectral information proved to be essential for the improved accuracy of the decoder's interpolation function. An example of this new modification to the amplitude determination routine is depicted in Figure 7.15.

After splitting the current frame into two half-frames, each half-frame residual is 'windowed out' according to the pitch period contained within the frame. This pitch-dependent windowing follows the rules earlier established in Table 7.4. In the case of the voiced harmonics, the amplitude calculation uses the MSE matching method applied in the frequency domain, similar to the pitch refinement search technique. In the case of the unvoiced harmonics, the amplitude calculation determines the RMS spectral energy over the unvoiced harmonic band.

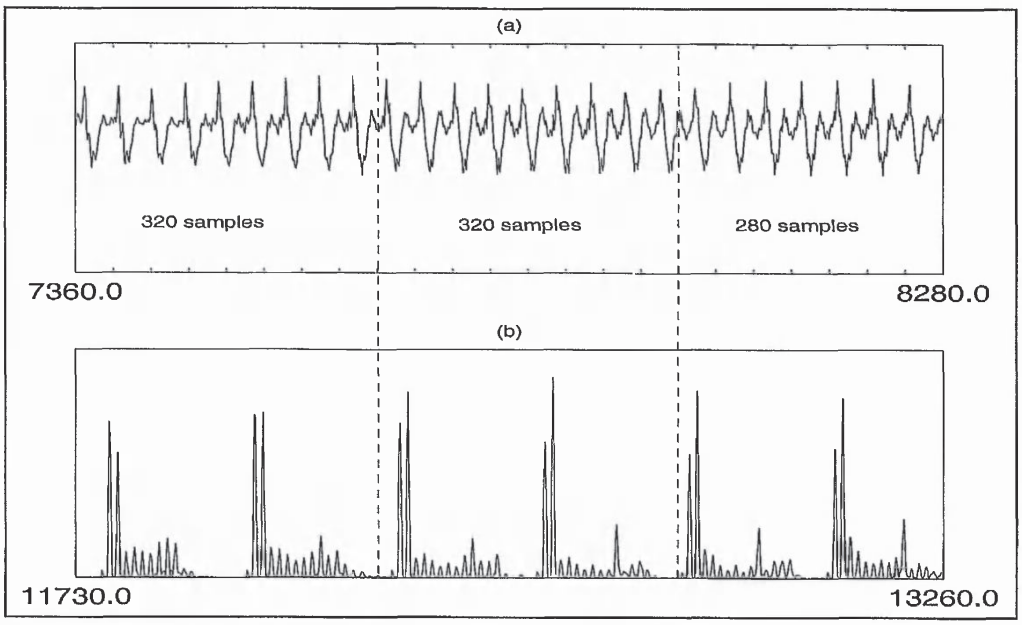


Figure 7.15 Calculating amplitude content twice per frame: (a) time domain (b) frequency domain

Quantisation of the harmonic amplitudes of the variable frame length includes quantising the RMS energy content for each half frame and quantising the actual harmonics for the full frame. Each half frame amplitude set individually requires 6 bits for RMS energy quantisation, where the harmonic amplitudes are normalised to their RMS values and then quantised. Quantisation of the actual harmonics of the full frame is performed using 6 bits.

7.4.1.5 Improved 'Search-and-Compare' Segmentation Algorithm

The 'search-and-compare' segmentation algorithm used in the variable frame length scheme of the Time Envelope SB-LPC vocoder, was primarily based on pitch, voicing and time envelope information. The frame extending was applied when the fluctuations in each of the above parameters were confined within the specified threshold limits. The design of the Amplitude Excited SB-LPC vocoder introduced spectral amplitudes as a coding parameter. The estimation of the spectral amplitudes offered an opportunity for taking a closer look at the spectral behaviour of speech within a given frame. Tests revealed a number of cases where the speech parameters used (pitch, voicing and envelope energy) were within their threshold limits during frame extending, however, the spectra of the extended frames showed a relative change in magnitudes of the individual amplitudes, as illustrated in Figure 7.16.

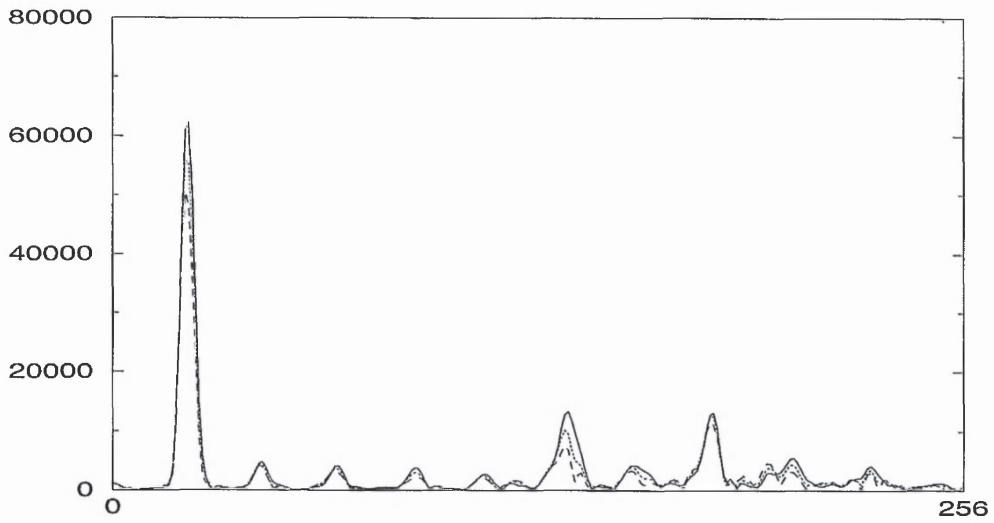


Figure 7.16 An example of the varying spectra for three frames with equal pitch and voicing

The consequence of disregarding these spectral differences is the elimination of some short-term speech events and therefore a decrease in speech sharpness. There was thus a clear need to modify and improve the 'search-and-compare' segmentation algorithm which lead to the inclusion of a fourth decision level where the spectra of the frame extensions are compared against the spectrum of the starting frame. The comparison procedure is performed along the vertical and horizontal axes of the frequency domain.

The *vertical comparison* entails setting the energy thresholds for the fluctuations in harmonic magnitudes. A closer study of the spectral behaviour resulted in the following conclusions:

- the energy comparison can only be performed on the voiced harmonics; the unvoiced harmonics usually carry an insignificant amount of energy
- the voiced harmonic magnitudes must be compared on the individual basis; the tests showed that any comparison among the harmonics on the MBE basis, i.e. as bands of 2 or 3 harmonics, can lead to large errors, especially in the lower frequencies, where individual harmonics usually vary widely in their magnitudes.

The following rules for the vertical spectral comparison were adopted after both objective and subjective tests had been undertaken:

- Rule 1: the behaviour of the first voiced harmonic is the most significant and it should therefore be treated separately from any other common energy threshold rules. The tests showed that the variation in energy level of the first harmonic should not exceed 15% of its original value in the starting frame.

- Rule 2: the voiced harmonics under scrutiny can vary widely in their magnitude levels. Tests indicated that any deviation in larger magnitudes is more significant than that of smaller sized magnitudes. A way of drawing the line between the two cases is to determine the average energy of the voiced part of the spectrum and deal with magnitude variations of those harmonics whose magnitudes are above the average energy level. In order to calculate the average energy level, only the lower part of the spectrum is considered to be energy significant, as given in Equation 7.1 which estimates a divisional factor D_l

$$D_l = \frac{N/2}{b_l - a_l} \tag{7.1}$$

where N represents the spectrum size and $(b_l\text{-}a_l)$ is the width of the harmonic l. The average spectral energy for the lower half of the spectrum is then calculated as given in Equation 7.2,

$$E_{av} = \frac{E_{lowfreq}}{D_l} \tag{7.2}$$

The estimated average spectral energy is then compared against the spectral energy of each individual voiced harmonic. The average energy of the voiced part of the spectrum is a threshold parameter that can be adjusted by a given factor. By reducing this threshold, a greater number of harmonics will become energy significant and will consequently have a greater effect upon the spectral comparison procedure which in turn will increase the frame rate. The graph in Figure 7.17 shows three typical examples of the increase in number of processed frames due to the reduction in the average energy threshold level.

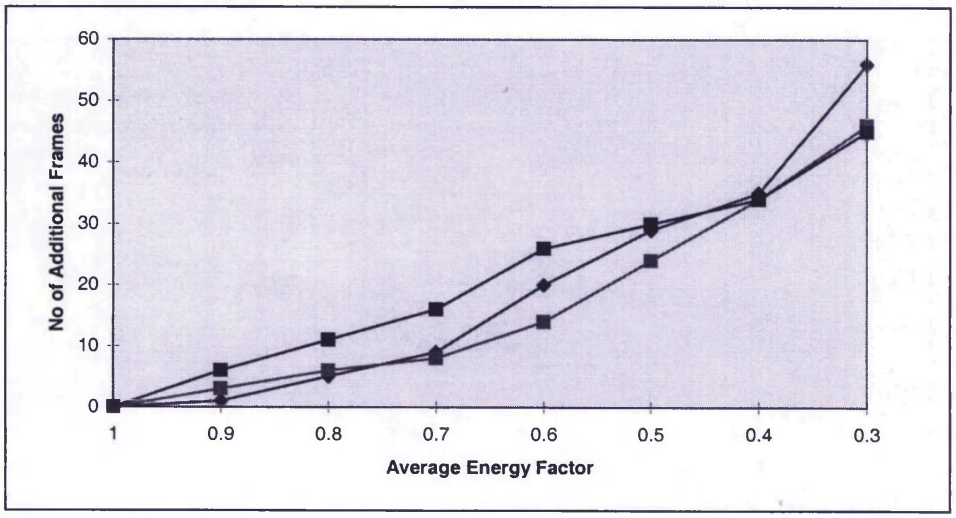


Figure 7.17 Number of additional frames vs. Average energy factor graph

Two relatively sharp points can be detected on the graph; firstly, at 30% reduction in the average energy threshold level and then at 60% reduction in the average energy threshold. Therefore, a choice among the energy factor values of 1.0, 0.7 and 0.4 can be made as a trade-off between the affordable decrease in frame rate and a required perceptual quality level.

• Rule 3: The energy threshold for each voiced harmonic magnitude (excluding the first harmonic), that lies above the average energy threshold, is set at 30% of its corresponding voiced harmonic magnitude in the starting frame. This figure is again a trade-off between the

produced quality and the reduction in the overall frame rate. The tests indicated that if any of the voiced harmonics with significant energy content surpasses this threshold, the potential for speech smoothing is very high. Figure 7.18 depicts three typical examples of how the frame rate decreases with the increase of the Rule 3 threshold. The figure of 30% was chosen as the optimum level which can meet both the subjective (perceptual quality) and objective (low frame rate) constraints of the variable frame length vocoding. The tests showed that any lowering of this figure, to for example 20%, will significantly increase the overall bit rate, whereby the perceptual quality will not be noticeably improved. If the figure is increased to, for example, 40% or 50%, the frame rate will not be much changed, however, the output speech quality will be affected by occasional smoothing patterns.

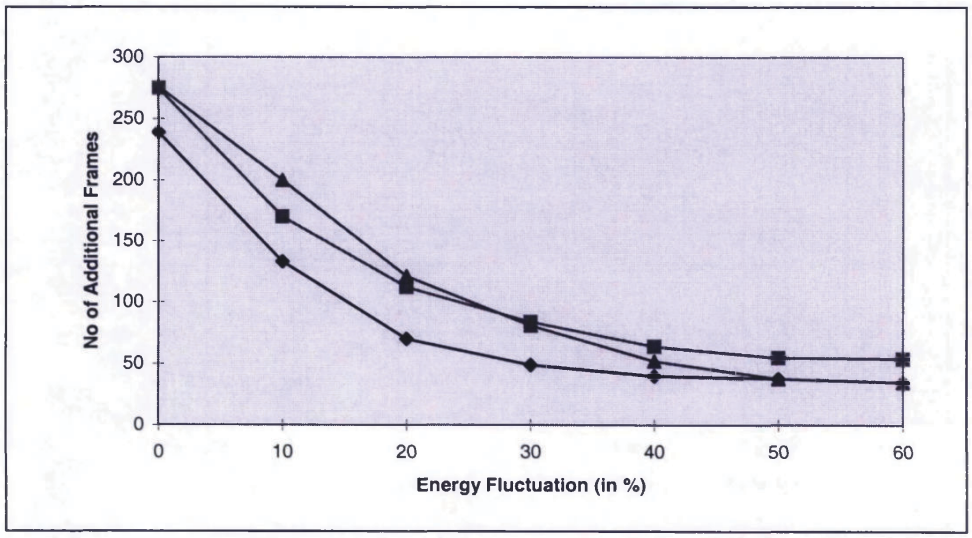


Figure 7.18 An example of the decrease in the bit rate due to the larger energy threshold values

The *horizontal comparison* represents setting the thresholds for the deviations in pitch frequencies and related widths of harmonic bands of the comparing spectra. The tests indicated that a shift in pitch frequency during frame extending procedure can give rise to perceptual disturbances, whereby the larger the shift, the greater the disturbance. A closer scrutiny of the spectral behaviour during frame extension showed that pitch frequency deviation should not be allowed to exceed ±1 sample in time domain, which means that the initial ±2 sample threshold, used in the Time Envelope approach, had to be replaced. The tests confirmed that any horizontal shift on the frequency scale that is kept within this new

threshold does not significantly affect the perceptual quality of the vocoder. However, the price to pay for this newest restriction is the slight increase in the overall frame rate.

The improved four-level 'search-and-compare' decision algorithm is depicted in Table 7.5 below.

Table 7.5 A four-level decision making algorithm on an optimum frame size

Decision level	Parameter	Threshold
1.	Pitch period	≤ ±1 samples
2.	Voiced Harmonics 2-12 Voiced Harmonics 13-18 Voiced Harmonics 19-24 Voiced Harmonics 25-	≤±2 ≤±3 ≤±4 ≤±5
3.	Envelope energy change	≤50%
4.	Spectral comparison: First voiced harmonic Remaining voiced harmonics	 ≤15% of its magnitude ≤30% of energy variation per voiced harmonics and above the average energy threshold line

7.4.2 Decoder

Figure 7.19 represents the decoder schematic of the Amplitude Excited SB-LPC vocoder operating at the variable frame rate.

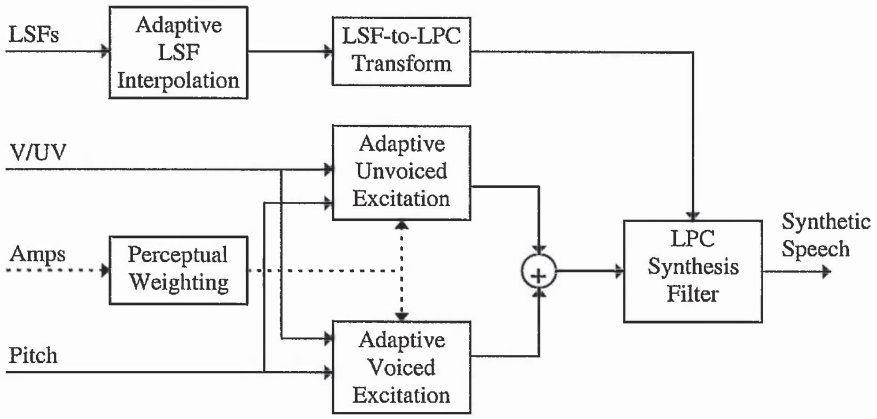


Figure 7.19 Decoder Block Diagram

The *LSF parameters* are decoded, interpolated and transformed into the LP coefficients, which are then used at the final stage of the decoder in the LPC synthesis filter. The decoded *voicing* and *pitch period* information control the amount of voiced and unvoiced excitation generation. The *amplitudes* are decoded and then modified in such a way as to enhance the perceptual performance of the vocoder. The voiced excitation is generated using the time domain harmonic summation of the voiced amplitudes. The unvoiced excitation is generated in the frequency domain and then transformed into time domain where it is combined with the voiced excitation. Finally, the combined excitation signal is passed through the LPC filter to produce the synthetic speech.

7.4.2.1 Perceptual Enhancement

The perceptual enhancement technique is used to modify the two sets of spectral amplitudes before the LPC filtering, the method already used by the fixed frame rate version of this vocoder and described in Chapter 5. The amount of amplitude modification depends on the difference between the LP spectrum and its equivalent peak interpolated LP spectral shape.

7.4.2.2 Improved Adaptive Interpolation Algorithm

The synthesis structure of this novel variable frame length decoding algorithm preserves the *half frame delay* mechanism used in the fixed frame length structure. The reason behind it is the fact that the full frame delay of the Time Envelope SB-LPC vocoder operating at variable frame rates, as described in the previous section of this Chapter, proved to be too complex and yet its variable interpolation method could not accurately match that of the fixed rate system. The adaptive interpolation algorithm applied here is less complex in structure, however, it allows a higher degree of overlap, as it not only controls the change between the neighbouring frames but also between the two half frames within a current decoded frame.

The half frame delay mechanism means that the synthesised frame often has a different length to the two encoded frames whose parameters are being interpolated on the decoder side. An example is shown in Figure 7.20. In this Figure, an overlap between a 120-sample encoded frame and a 200-sample encoded frame produces a 160-sample synthetic frame.

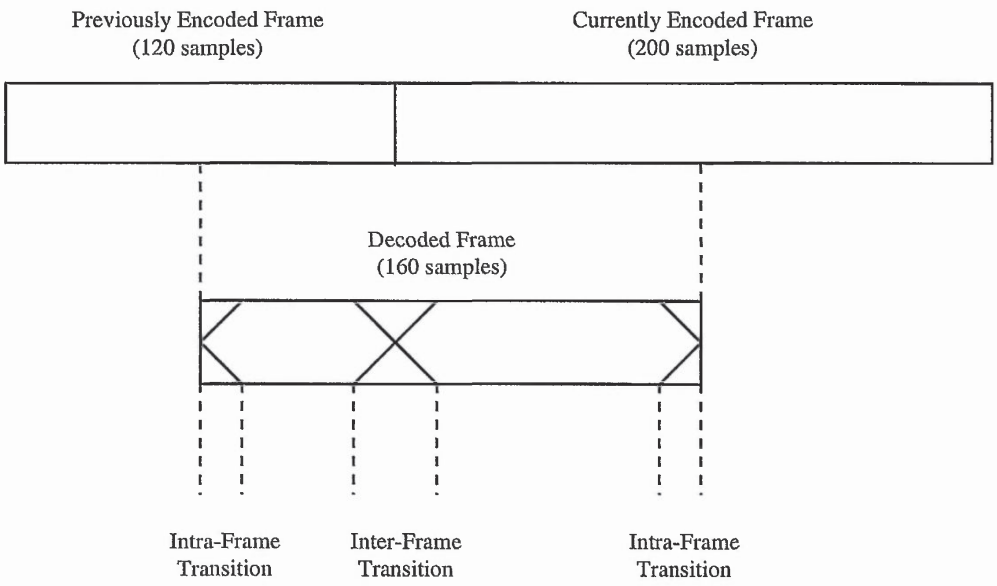


Figure 7.20 Adaptive interpolation algorithm based on half frame delay and both inter-frame and intra-frame overlap

There are two general principles on which this adaptive interpolation algorithm is based. The first principle concerns with the *intra-frame* boundary between the two half frames of a given encoded frame, as each half frame uses a separately transmitted set of spectral amplitudes. In order to smooth out the transition between the two amplitude sets, a short transition region of a constant length of one decoded sub-frame (20 samples) on each side is established. Setting up of this rule proved to be very important when a long frame is encoded whose amplitude energy content changes from one half frame to the other.

The second principle concerns with the *inter-frame* transition from one encoded frame to another. Due to the fact that the lengths of the adjacent encoded frames are allowed to vary, the size of the transition region is also permitted to vary. Although there is quite a number of possible combinations of pairs of adjacent encoded frames, in order to reduce the decoder's complexity, the inter-frame transition region is allowed to take the sizes of 2, 4 or 6 sub-frames, i.e. 40, 80 or 120 samples respectively. Which size the transition will take depends solely on the paired combination of encoded frame lengths, following similar transition rules as depicted earlier in Table 7.2. The paired combination of shorter frames take on the 2 sub-

frame length transition, those of the middle range length take on the 4 sub-frame transition size and only the combinations of longest frame sizes adopt the 6 sub-frame length transition.

The biggest advantage of this adaptive interpolation method lies with the fact that the transition boundaries, for both intra-frame and inter-frame transitions, are kept constant during the synthesis of an output frame. That is, the same transition boundaries are used for the energy dependent linear harmonic interpolation of the voiced excitation, linear interpolation of the unvoiced excitation and during the LP synthesis of the output frame. In this way, unlike the variable frame length adaptive interpolation of the Time Envelope SB-LPC vocoder, this interpolation method provides higher consistency to the overall decoding structure.

7.4.2.3 Voiced Excitation Generation

Equation 7.3 defines the generation of the voiced excitation. The principle is very similar to the voiced excitation generation of the fixed frame rate version, as described in Chapter 5. The novelty lies with the fact that there are now two sets of amplitudes for each half of the encoded frame size.

$$e_{voiced}(i) = \sum_{j=1}^{L} Amp_{\text{mod}}(j)\cos\left[j\phi(i)\right] \tag{7.3}$$

where $Amp_{mod}(j)$ takes on the modified amplitude value if the jth harmonic is voiced, or $Amp_{mod}(j)=0$ when the jth harmonic is unvoiced.

Equation 7.4 represents the fundamental phase $\phi(i)$ as the sum of the interpolated fundamental frequency function $\omega(i)$, defined in Equation 7.5, and the fundamental phase of the previous sample.

$$\phi(i) = \phi(i-1) + \omega(i) \tag{7.4}$$

$$\omega(i) = \begin{cases} \omega_{last}\left(\frac{i+0.5}{N}\right) + \omega_{next}\left(1-\frac{i+0.5}{N}\right), & \left|\omega_{last}-\omega_{next}\right| \le 0.25\left(\omega_{last}+\omega_{next}\right) \\ \omega_{last}, & \left|\omega_{last}-\omega_{next}\right| > 0.25\left(\omega_{last}+\omega_{next}\right) \text{ and } i \le \frac{N}{2} \\ \omega_{next}, & \left|\omega_{last}-\omega_{next}\right| > 0.25\left(\omega_{last}+\omega_{next}\right) \text{ and } i > \frac{N}{2} \end{cases} \tag{7.5}$$

The interpolation process shown in Equation 7.5 is only performed when the pitch value deviates by less than 25% of its value from one frame to the next.

During voiced excitation, the linear harmonic amplitude interpolation adheres to the transition principles and boundaries set by the improved adaptive interpolation mechanism. The amplitude interpolation is performed at both the intra-frame boundaries and inter-frame boundary, as shown in the previous section. The tests showed that the perceptual quality of this vocoder is significantly increased by adopting the two sets of amplitudes per encoded frame rather than the traditional one set of amplitudes per encoded frame, especially in cases when long frame sizes are interpolated and synthesised.

7.4.2.4 Unvoiced Excitation Generation

Figure 7.21 depicts the block diagram of the unvoiced excitation generation function.

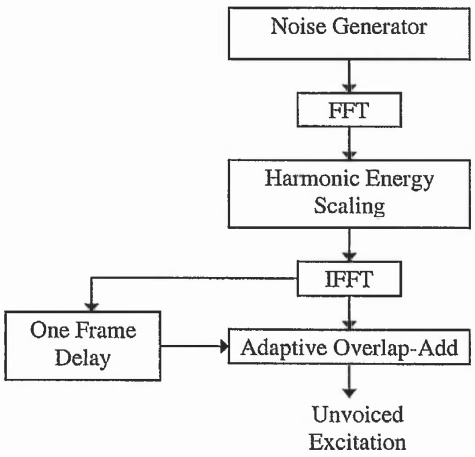


Figure 7.21 Unvoiced Excitation Generator Block Diagram

Although largely similar to the unvoiced excitation generation scheme of the fixed frame rate vocoding described in Chapter 5, there are three major differences that characterise this new unvoiced excitation generation scheme. Firstly, the unvoiced excitation is generated twice per frame as each half frame can now be characterised by a set of harmonic magnitudes. Secondly, the variability of frame sizes has eliminated the trapezoidal windowing of the generated noise prior to the FFT. And thirdly, the Overlap-Add method is now performed

both in the middle and at the ends of the synthesised frame, which adheres to the transition principles and boundaries set by the improved adaptive interpolation mechanism.

The new method of unvoiced excitation generation has shown to produce good perceptual quality. Moreover, its relatively complex structure has also shown to fit well within the overall synthesis procedure of the variable frame length vocoder.

7.4.2.5 LPC Synthesis Filter

The LP excitation signal is the summation of both the generated voiced and unvoiced excitations. The filtered signal is synthesised on the 20-sample sub-frame basis, i.e. every 2.5 ms, using the interpolated LP coefficients. The interpolation of the LP coefficients is performed according to the principles of the adaptive interpolation method for both intra-frame and inter-frame transition boundaries, as described in the previous section. Spectral de-emphasis is performed on the filtered signal before synthetic speech output is produced.

7.4.3 Bit Allocation Scheme

Table 7.6 depicts the overall bit allocation scheme for the 2.75 kbps Amplitude Excited SB-LPC vocoder if a fixed frame length of 20 ms is assumed. This bit rate differs from the starting fixed frame rate structure described in Chapter 5, where the constant bit rate was 2.4 kbps. One can notice the slight increase in the bit rate due to the use of two sets of amplitudes per frame and the frame size header. It can also be observed that the bit allocation level for the LSF quantisation has been reduced from 26 bits to 24 bits.

Table 7.6 Bit Allocation Scheme

	Bit Allocation						
Frame Length (ms)	Frame Size Header	LSF	Pitch	Energy	Shape of 1st 8 Harmonics	Voicing	Fixed Bit Rate (kbps)
20	**3**	**24**	**7**	**12**	**6**	**3**	**2.75**

7.4.4 Performance Evaluation

The performance of the variable frame length structure of the SB-LPC vocoder was assessed using the MOS scale, whereby the subjective quality of the variable frame length scheme was compared with the subjective quality of the fixed frame length scheme. The relevant fixed rate speech coding standard (4.15 kbps I-MBE vocoder) was also included in the tests as a reference point. Twenty-four subjects, mostly experienced listeners, were used to conduct the informal MOS listening test that included two male and two female sentences.

Table 7.7 MOS Test Results

Vocoder Type	Frame Length (samples)	MOS	Variance	Change in average frame rate (voiced frames only)
4.15 kbps I-MBE	160 (fixed)	2.93	0.29	None
2.6 kbps SB-LPC	160 (fixed)	3.54	0.52	None
2.75 kbps SB-LPC	80-320	3.51	0.52	+1%
	120-240	3.63	0.41	+2%
	120-320	3.47	0.50	-2%
	160-320	3.52	0.29	-22%

The test results, as given in Table 7.7, show that the perceptual quality and the change in frame rate depend on the selected frame size range of the variable frame rate codec.

The short range version that permits frame extending only (160-320 samples) produced similar quality of speech as the fixed rate codec. However, its real gain over the fixed rate structure is the average drop in voiced frame rate by 22%. The moderate range version (120-240 samples) performed better than the fixed rate codec with an insignificant increase in frame rate. Both the short and the moderate range versions represent the real success of the variable frame length coding.

The full range versions (80-320 samples and 120-320 samples) performed only slightly worse than the fixed rate codec, although they did not experience any large fluctuations in frame rate. This decrease in speech quality indicates the level of difficulty and complexity involved in jointly coding very short and very long frames. This is especially related to the adverse block edge effects which characterise the variable interpolation sizes within the frame boundaries. However, the tests also confirmed the importance of selectively using highly

accurate shorter frames in modelling sharp speech onsets. Restricting the use of shorter frames (80-sample and 120-sample) to speech onset regions proved beneficial to the variable frame rate scheme both in terms of the subjective quality and the reduction in average frame rate.

The Amplitude Excited SB-LPC vocoder, including its fixed and variable frame rate structures, outperformed the I-MBE standard during the tests. This outcome was mainly due to a large difference in cleanness and smoothness between the two codec structures as perceived by a considerable majority of the test subjects. As it was the same I-MBE standard used for reference purposes during the subjective testing of the variable frame rate structures simulated on the Time Envelope SB-LPC vocoder, it can also be concluded that the Amplitude Excitation version of the SB-LPC vocoder far outperforms the Time Envelope vocoder.

The potential reduction in frame rate is slightly decreased in the case of the Amplitude Excitation vocoder largely because of an extra stage in the 'search-and-compare' segmentation method. However, the use of the spectral comparison procedures has resulted in more accurate segmentation decision. This fact was confirmed by both objective and subjective testing of the produced synthetic speech.

7.5 Amplitude Excited SB-LPC Vocoding with Phonetically Based Segmentation

The idea of segmenting speech according to its phonetic characteristics was already introduced and elaborated in Section 7.2.1, where phonemes were defined as linguistically distinct speech sounds characterised by various acoustic properties. This acoustic-phonetic approach, which is often used in speech recognition systems, involves segmenting the speech signal into discrete (in time) regions where the acoustic properties of the signal are representative of one (or possibly several) phonetic unit(s). The key aim of this type of speech recognition is the creation of a large vocabulary containing a range of phonemes, which can then be used in segmenting the speech. As the speech characteristics are mainly kept constant

within a phoneme, this method should theoretically yield the most accurate segmentation results, which in turn could further improve the performance of the variable frame rate coding. During the course of this research a quite successful attempt was made to integrate the phonetically based segmentation module with the Amplitude Excited SB-LPC vocoder operating at variable frame rates. The block diagram of the two-module system is given in Figure 7.22.

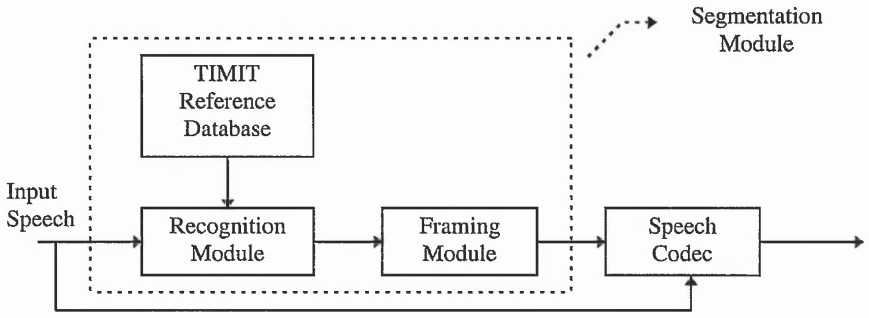


Figure 7.22 The block diagram of the integrated two-module system

7.5.1 Segmentation Structure

The speech segmentation module consists of a TIMIT reference database, recognition module and framing module. TIMIT represents a phonetic library designed to provide speech data for the acquisition of acoustic-phonetic knowledge used for segmenting input speech. The task of the recognition module is to perform phonetically based segmentation of speech, where each phoneme is built as a multiple of 5 ms (40 samples) [108]. An array of speech phonemes is then passed onto the framing module, where an inner segmentation of the extracted phonemes is performed using the frame sizes within the variable frame length range (e.g. 120-240 samples) of the speech coding module. The two key principles of the inner segmentation of phonemes are:

- the use of longer frames to cover the phoneme's length
- the use of shorter frames of constant length for framing the phoneme's edges.

The first principle is based on the assumption that the speech characteristics barely change within a segmented phoneme. Therefore, longer frames can be employed in order to keep the

frame rate low. The second principle is oriented more towards the preservation of speech quality at the phoneme transitions by employing short frame sizes. The fact that they are of constant length significantly reduces the coder's overall complexity. The final output of the segmentation module is a set of segmented frame sizes that are passed with the corresponding input speech onto the speech codec. Figure 7.23 provides an example of segmented speech using the two-module segmentation system.

The main design problem in coding the phonetically segmented speech is the wide variety of the input phoneme sizes. The length of the phoneme extracted from the TIMIT speech corpus can lie anywhere between 5 ms and 400 ms (and longer) [105]. Although the assumption is made that the speech characteristics within a phoneme do not significantly change, one has to account for both the time-envelope variations, especially for longer frames, and inter-phoneme transition regions.

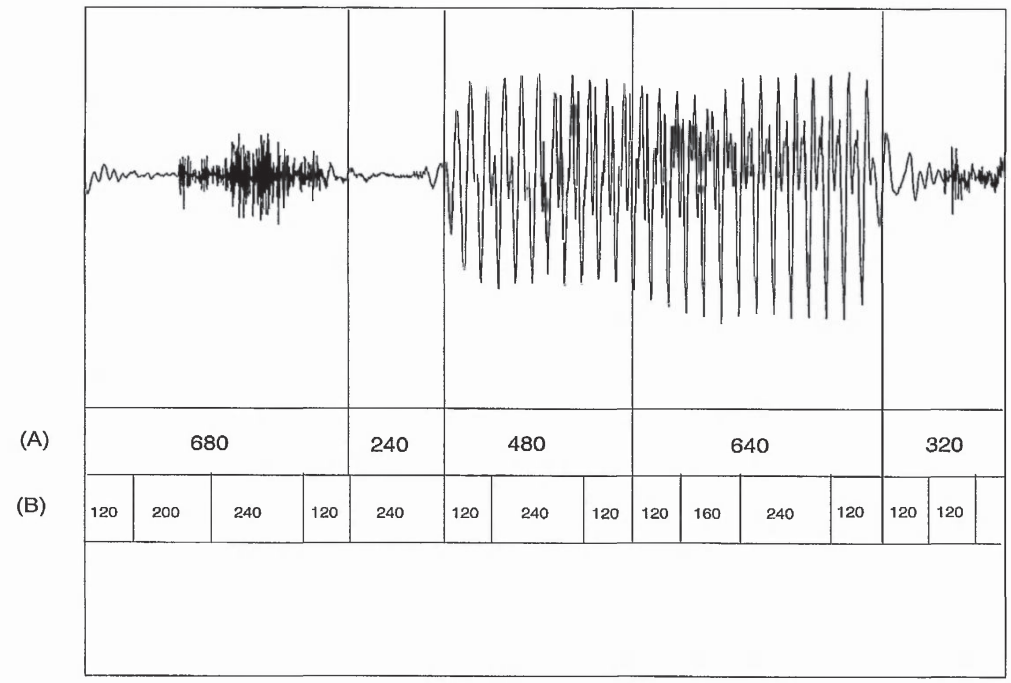


Figure 7.23 'Hybrid' segmentation of speech: (a) Phonetically based speech segmentation
(b) Sub-frame based segmentation of phonemes

For longer phonemes the time envelope variations can be accounted for by using the limited range of longer frame sizes, for example up to 40 ms (320 samples). In this way, both the

reduction in frame rate and the preservation of speech quality can be attained. A greater problem arises when attempting to frame shorter phonemes and inter-phoneme transitions as it can only be solved as a trade-off between the speech quality, the attainable frame rate and the coder complexity. As each phoneme is built as a multiple of 5 ms (40 samples), the number of potential short frame candidates is reduced to three options: 40-sample, 80-sample and 120-sample sizes. The objective and subjective tests showed that the use of 40-sample and 80-sample frames caused a large increase in frame rate and a considerable increase in the coder's complexity. The slight improvement in the achieved speech quality could therefore not justify their use. The solution was thus a 120-sample frame which was used as the minimum frame length for the short phonemes. The tests showed that the use of 120-sample frames for phoneme segments shorter than 120 samples did not produce any significant drop in speech quality. As far as phoneme transitions are concerned, the tests compared the speech quality attained by constantly using 120-sample frames and 160-sample frames. As expected, the use of the 120-sample frames was found to produce slightly higher speech quality at the phoneme edges than the 160-sample frames. However, the option of using the 120-sample frames at the edges caused significant increase in attainable frame rate.

7.5.2 Performance Evaluation

The performance of the phonetically based segmentation system coupled with the variable frame length structure of the Amplitude Excited SB-LPC vocoder was assessed using the MOS scale. Four different phonetic segmentation structures were compared in terms of perceptual quality and the change in frame rate. Table 7.8 lists the four proponents with their respective frame length range and the static boundary frame size, e.g. 120-240 is the range and (160) is the inter-phoneme transition frame size. The change in average frame rate (for voiced frames only) was made with respect to the fixed frame rate vocoding structure with 160-sample frame size. Twenty-four subjects, mostly experienced listeners, were used to conduct the informal MOS listening test that included one male and one female sentence for which the TIMIT library and the recognition module performed the phonetic segmentation. This test was conducted as part of the MOS test for the Amplitude Excited SB-LPC vocoder operating at variable frame rates reported in the previous section. The aim was to show the relative difference in the perceived quality among the four phonetic segmentation structures.

Table 7.8 MOS Test Results

Vocoder Type	Frame Length (samples)	MOS	Variance	Change in average frame rate (voiced frames only)
2.75 kbps	120-240 (160)	3.92	0.60	-13%
SB-LPC	120-240 (120)	4.04	0.39	-3%
with phonetic	160-320 (160)	3.42	0.34	-19%
segmentation	120-320 (160)	3.75	0.46	-15%

The test results, as given in Table 7.8, again show that the perceptual quality and the change in the frame rate much depend on the employed frame size range of the variable frame length codec.

The short range version that permits frame extending only (160-320 samples) using 160 samples at the transitions produced the lowest quality of speech. The main reason for this low performance lies in its inability to accurately model short phonemes, as its minimum length stays at 160 samples at all times. However, its real gain is the average drop in voiced frame rate by 19%. This result is very much in line, both in terms of attainable quality and the reduction in frame rate, with those obtained for the variable frame rate vocoder with no phonetic segmentation and using the same frame length range, as described in the previous section.

Quite a significant improvement in subjective quality, with only a small increase in frame rate, is offered by the 120-320 sample segmentation structure also using 160-sample boundary frames. These results can be accepted to fully justify the use of 120-sample frame sizes, if not at the phonetic boundaries then at least during the framing process within a phoneme.

Both moderate range versions of 120-240 samples managed to push the MOS score even further up the scale. Their respective frame rates depended on the transition frame sizes used. This increase in the MOS score when compared to the short range versions can be seen as a consequence of limiting the long frame sizes to 240 samples. Just like in the previous section, these results indicate the level of difficulty and complexity involved in jointly coding very short and very long frames. This is again related to the adverse block edge effects which characterise the variable interpolation sizes within the frame boundaries. Also, when a comparison is made between the two moderate range versions (120-240 samples with 160-sample transitions and 120-240 samples with 120-sample transitions), higher quality is

attained by the coder employing 120-sample frames at its phonetic transitions. However, the frequent use of 120-sample frames also affects the frame rate performance, as shown in Table 7.8.

Overall, these results indicate that the choice of the segmentation range largely depends on the two conflicting requirements - the output quality and the frame rate. The tests indicated that the segmentation system that emerges as an optimum trade-off between the two requirements is the 120-240 sample range system that employs 160-sample transition frames. It can offer very high quality of speech whilst still retaining a large drop in frame rate.

The phonetically based segmentation approach can be seen as a novel approach in very low bit rate speech communication as it unifies the areas of speech recognition and speech coding. The preliminary results showed a lot of potential in this field both in terms of preserving the high quality of speech and reducing the frame rate. As far as this research is concerned, quite a successful attempt was made to explore this new idea that certainly requires further work and understanding.

7.6 Conclusions

The research presented in this chapter has introduced the idea of variable frame length coding based on the time varying characteristics of the speech signal, where longer frame lengths are used to code highly correlated steady state speech segments and unvoiced speech, whilst shorter frame sizes are employed to shape speech transients. The most important part of the variable frame rate coding is the accurate speech segmentation algorithm. Three different segmentation principles have been investigated, the most accurate and also the most complex being phonetically based. Although some preliminary work has been done using the phonetic segmentation method, the largest part of variable frame rate system design and development presented here has been based on the speech segmentation method that uses robust source-dependent criteria.

The Time Envelope SB-LPC vocoder is the first system that has been developed for variable frame rate coding. Its main strengths lie in its ability to very accurately segment speech according to the adopted source-dependent criteria and adaptively interpolate frames according to their varying sizes. The MOS test showed that the variable frame rate structure does not significantly trail behind the subjective performance of the fixed frame rate algorithm, whilst reducing the average frame rate by 7% for voiced frames. However, the main disadvantages of this vocoder have been attributed to its time domain based pitch detection algorithm, which is prone to pitch multiple and sub-multiple errors, and the practical complexity of quantising its variably sized time envelope. Nevertheless, the Time Envelope vocoder has proved to be very useful in further developing the variable frame rate system, as most of its key functions have been adopted in designing the Amplitude Excited SB-LPC vocoder.

The Amplitude Excited SB-LPC vocoder has succeeded in significantly improving the performance of the Time Envelope version by switching its troubled time domain functions, such as pitch detection algorithm and time envelope calculation, into the frequency domain. The new frequency domain pitch detection algorithm has proved to be much more resilient to the pitch multiple and sub-multiple errors. As for the time envelope, this vocoder uses and quantises spectral amplitudes placed within a finite spectral length independent of its frame size, thus eliminating the need for variable length time envelope quantisation. Other major improvements have been made to the source-dependent segmentation method on the encoder side and the adaptive interpolation method on the decoder side. All this has significantly improved the subjective quality of the vocoder as shown in its MOS results. Not only has this variable frame rate vocoder managed to produce similar or even higher quality of speech than its fixed frame rate version, but it has also outperformed the I-MBE vocoding standard as a result of its increased smoothness and cleanness.

A successful attempt has also been made to integrate the phonetically based speech recognition module with the Amplitude Excited SB-LPC vocoder. The preliminary results showed a lot of potential in this field both in terms of preserving high quality of speech and reducing the frame rate. The further progress in this area will possibly require a joint segmentation effort by both the phonetic based recognition module, which could serve as a first step in accurately segmenting speech into smaller units, and the source-dependent

segmentation algorithm, whose function could be to fine tune the frame division within phonemes and at phonetic transitions.

Chapter 8.

8. Coding of Wideband Speech

8.1 Introduction

In the preceding chapters this research has mainly dealt with the operational traits and potential low bit rate applications of the developed narrowband speech coders. These narrowband speech coders proved to be capable of providing near toll quality speech for a telephone bandwidth of 300 Hz-3400 Hz at bit rates ranging from 2.4 kbps (and lower) to 4.8 kbps. However, there are also many other telecommunication applications where it is not necessary to operate at such low bit rates. On the contrary, an increase in speech quality using a wider bandwidth is regarded as highly desirable. An example includes the emerging Integrated Services Digital Network (ISDN), which requires speech coding algorithms capable of producing high quality speech over a larger bandwidth. Other applications include high quality commentary grade channels for use in audio video teleconferencing, high quality voice-mail services and high quality wideband telephony [113].

The main characteristic of wideband speech is that its bandwidth lies in the region of 50 Hz to 7000 Hz, which is much wider than the standard telephony bandwidth of 300 Hz to 3400 Hz; wideband speech is sampled at 16 kHz for subsequent digital processing, as opposed to the 8 kHz sampling rate of the narrowband speech. The key perceptual advantage of the wideband over the narrowband speech is that its extra low frequency components increase the voice naturalness, whereas the added high frequencies make the speech sound crisper, more intelligible and 'brighter'. It is now widely accepted that the quality produced by the wideband speech equates that of an FM radio announcer [113], with a richness notably greater than telephone bandwidth speech together with very high intelligibility and naturalness.

The wideband speech can be regarded as essentially consisting of two bands: a base band (50-3500 Hz), which contains approximately 80% of the perceptually important speech spectral information [114], and a high band which contributes to the overall perceived intelligibility of speech. The base band spectral components of voiced speech are structured and larger in magnitude than the high band components, which are highly unstructured and relatively small in amplitude for both voiced and unvoiced speech. Figure 8.1 shows the magnitude spectrum for a frame of voiced male speech, where this high spectral dynamic range is clearly evident. The difference in the mean energy levels between low and high bands can be as high as 20-40 dB [115]. This contrast between low and high band spectral information makes wideband speech harder to code than normal narrowband. The output perceptual quality of a wideband speech coding algorithm depends on its ability to produce high quality speech for both the lower and upper spectral content. It must, therefore, accurately code the perceptually important lower frequency components, and yet retain enough of the higher frequency information such that the perceived richness and fidelity of the original speech is preserved.

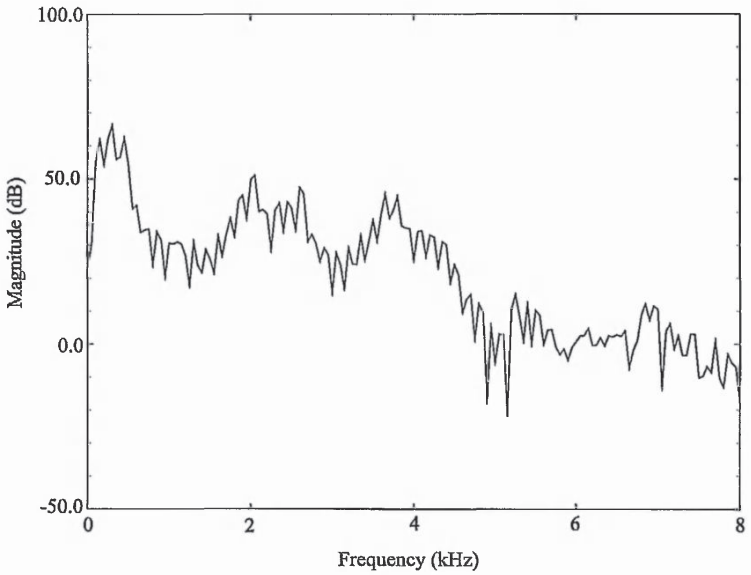


Figure 8.1 Magnitude spectrum for a frame of voiced male speech

This chapter begins with an introduction to key techniques and approaches presently adopted for coding wideband speech. It is followed by the detailed description of the design issues and characteristics of the 8.5 kbps Amplitude Excited SB-LPC vocoder operating in wideband conditions. The performance of the wideband vocoder is then assessed by comparing it with the performance of the current ITU-T wideband standard G.722 [116] operating at three mode rates: 64, 56 and 48 kbps. The concept and application of the variable frame rate wideband coding is also addressed. In the concluding section some potential guidelines for future work are discussed.

8.2 Current Wideband Speech Techniques and Methods

Two techniques that constitute the key approaches to coding wideband speech are:

- the split band scheme
- the full band scheme.

The most notable representative of the *split band approach* is the ITU-T G.722 standard for 7 kHz audio coding. The G.722 coding algorithm splits the speech into two bands using a 24-tap Quadrature Mirror Filter (QMF), and then uses ADPCM to code each band separately.

The coder is designed to operate between three basic rates: 64 kbps, 56 kbps and 48 kbps. The rate of operation is determined by the number of bits used to quantise each lower sub-band sample: 6 bits for 64 kbps, 5 bits for 56 kbps and 4 bits for 48 kbps. Each sample in the higher sub-band is quantised using 2 bits. The use of QMF filters ensures that the aliased energy is cancelled when the two bands are merged. A split band scheme can offer both subjective and objective benefits over a full band scheme [117]. These advantages consist in a more flexible structure and in constraining the quantisation noise to be almost fully confined to the band where it is produced. This allows for unequal bit allocation for each sub-band, i.e. more bits can be allocated to the perceptually important low sub-band and less for the higher frequency components, a fact which the G.722 algorithm exploits. Figure 8.2 depicts the split band approach to the wideband coding.

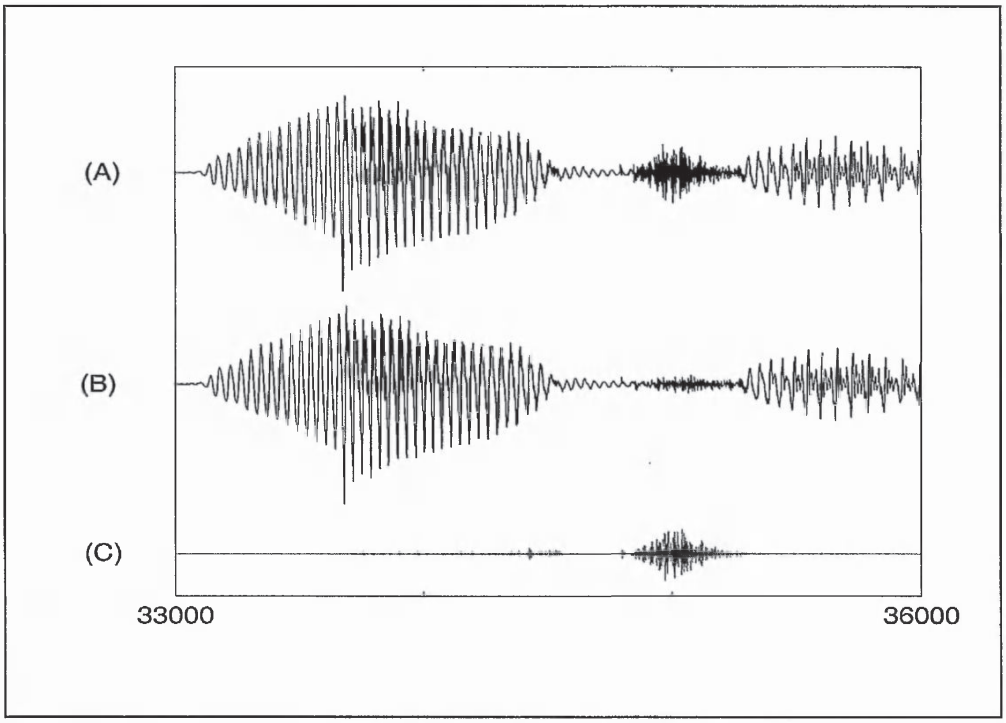


Figure 8.2 Split band approach: (A) full band; (B) lower sub-band; (C) higher sub-band.

Lowering the bit rates for wideband speech coding below those used for the G.722 standard, i.e. below 48 kbps, whilst retaining high perceptual quality, has been a topic of considerable research interest in recent years [118][113][119][120]. Lower bit rates can increase the efficiency of the transmission bandwidth and accommodate more potential subscribers, which is of particular interest to mobile communications. One of the main methods for lowering the

bit rates in wideband coding is combining CELP based algorithms with the sub-band approach adopted by the G.722. The main disadvantage of this method is an increase in coding delay as the G.722 has an inherent delay of 1.5 ms. However, for many applications this is not an important design criterion. Typically these algorithms employ QMF filters as in the G.722 or sharp cut-off filter banks to split the wideband signal into two sub-bands [115][117][121][122]. The sub-bands are then coded separately and reconstructed at the decoder using the complementary filter banks. The lower, perceptually more important sub-band is treated as a normal narrowband signal and encoded using a standard CELP coder with a 10th order LPC filter [117]. The higher sub-band is coded in a different manner, as it does not usually contain any long term correlation. However, there is still a significant amount of short term correlation present in the signal which can be exploited in the encoding process. Various methods have been proposed for coding this information, such as a narrowband CELP coder employing a 4th order LPC filter and a 4-bit Gaussian codebook [117]. Another scheme models the higher sub-band signal with a second order linear predictor and a small non-overlapping excitation codebook with a vector dimension of 200 [122]. The common trait of both these methods is the exploitation of the perceptual imbalance that exists between each band with a majority of the coding 'effort' concentrated on coding the lower sub-band signal.

The biggest disadvantage of the split band methods is the additional computational complexity required to split each frame of speech into two or more sub-bands, which inevitably incurs an extra delay. Representing each sub-band separately also requires a larger portion of bits needed for their individual quantisation. The *full band approach* does not suffer from the computational complexity, but it is, on the other hand, characterised by a less flexible structure and the spread of the quantisation noise throughout the 7 kHz spectrum. However, the key advantage of this approach is the reduced number of bits required to quantise the wideband spectral information, which can potentially lower even further the overall wideband coding bit rate, whilst still preserving the perceptual quality achieved [120].

The vocoding structure presented here takes a full band approach to wideband speech coding. It is based on the notion that a high quality narrowband coder can potentially be converted into a high quality wideband coder operating at relatively low-to-medium bit rates. The Amplitude Excited SB-LPC vocoder operating at 2.75 kbps (narrowband) was used as the

starting coding structure for designing and simulating the wideband vocoder with the output quality expected to be on a similar level with that produced by the ITU-T G.722 standard operating at 48 kbps. In the following section the description of the structure and performance of this newly proposed wideband vocoder is presented.

8.3 Amplitude Excited SB-LPC Wideband Vocoder at 8.5 kbps

The fundamental principles of this wideband algorithm are similar to those of the coding structure of its narrowband equivalent described previously in Chapters 5 and 7. The main adaptations were made above all to accommodate longer frame sizes (320 samples within a 20 ms frame), different windowing options and wider spectral bandwidth (7 kHz) for voicing analysis and harmonic amplitude calculation. Various control factors within the pitch detection, voicing decision and amplitude calculation were also retuned for the wideband conditions. Completely new additions for improving the wideband performance include the phase information module on the encoder side and the energy smoothing algorithm on the decoder side.

8.3.1 Encoder

The block diagram of the encoder section of the Amplitude Excited SB-LPC wideband vocoder is shown in Figure 8.3. The full band speech analysis is performed across the 8 kHz spectrum using 20 ms speech frames for the operating bit rate of 8.5 kbps (at 16000 Hz sampling frequency). The 20 ms speech frame size, like in the narrowband case, seems to offer the best compromise between speech quality and overall transmission rate such that the coder remains within the bounds of medium bit rate. At the same time, the speech stationarity principle is observed.

The parameters that are used to represent the speech model include:

- pitch period
- split band voicing decision
- harmonic amplitudes
- 16th order LPC filter
- phases.

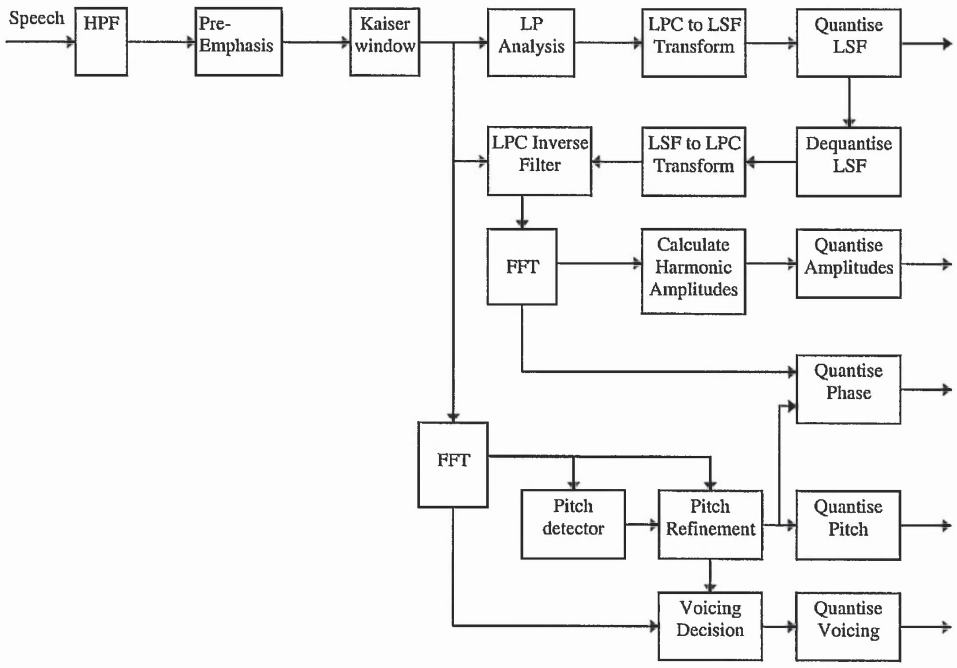


Figure 8.3 Encoder Block Diagram

8.3.1.1 Pre-processing Wideband Speech

At first, the original speech is high pass filtered in order to remove the DC bias. The high pass filter is followed by the pre-emphasis stage. Unlike narrowband speech coding, which can successfully operate even without pre-emphasis, wideband speech coding relies significantly more on this pre-processing routine. The reason for this is the fact that wideband speech contains important information at high frequencies whose accurate representation produces crisper, more intelligible and 'brighter' sounds.

Conventional *fixed pre-emphasis*, in the region from 0.80 to 0.95, as used in the narrowband coding, represents also an acceptable choice here. However, the key drawback of the constant use of large pre-emphasis is the potential reduction in the coding accuracy of the lower frequencies. This phenomenon can contribute to the reduction in the overall output quality. The solution to this problem is the introduction of *adaptive pre-emphasis*. Adaptive pre-emphasis operates as an added LPC filter with a small number of LP coefficients. In this research, four LP coefficients were found adequate to reduce the spectral envelope of a given frame according to its low and high frequency information. The adaptive pre-emphasis filter on the encoder side must be matched by an appropriate de-emphasis filter on the decoder side. In the case of the half frame delay incurred by this vocoding structure, the de-emphasis stage has to undergo an energy dependent interpolation process, as is done with the LPC synthesis filter calculations.

The test results showed a slight increase in perceptual quality in the more significant lower frequencies when the adaptive pre-emphasis was employed. However, the energy dependent interpolation of the de-emphasis filter coefficients caused the mismatch in the higher frequencies. This mismatch resulted in the introduction of the noisier sounds which had a greater overall negative impact on the perceptual quality than the improved performance in the lower frequencies. The tests also showed that the exclusion of the half frame delay removes this disturbing high frequency mismatch. Therefore, the adaptive pre-emphasis should ideally be used in the coding structures, such as CELP, that do not rely on the half frame delay during the speech analysis. The other drawback of the adaptive pre-emphasis is that it requires additional bits for encoding the 4th order LPC filter used. As a result of this experimentation, the adaptive pre-emphasis was abandoned for a more traditional fixed pre-emphasis approach.

The wideband speech analysis requires a change in windowing policy adopted in the narrowband domain. The tests with various window sizes showed that a simple doubling of the Kaiser window size used for the narrowband was the best choice in the wideband domain. Hence, the fixed Kaiser window size chosen was 441 samples with β=6.0.

8.3.1.2 LPC Filtering for Wideband Speech Analysis

The wideband speech spectrum requires a 16th order LPC filter for accurate shaping of the spectral envelope, as explained in Chapter 3. The filter is based on Durbin's analysis which is performed on pre-emphasised speech using a 441-sample Kaiser window (β=6.0). The LP coefficients are transformed into the LSF domain where they are quantised using the 50-bit two-stage split vector quantisation method [113]. This bit allocation choice is driven by the expected high speech quality. The LPC inverse filter is used to find the LP residual waveform required for determination of the excitation harmonic amplitudes and the phase information.

8.3.1.3 Wideband Pitch Detection Algorithm

The frequency based SMM-PDA module, whose accuracy and reliability were well proved in the narrowband domain, is again the chosen pitch detection algorithm. The main reason for this is the simplicity of its narrowband-to-wideband conversion. The main feature of the SMM-PDA is that it performs the pitch detection within the first 1400 Hz of the spectrum regardless of whether it is a case of wideband or narrowband speech. In this way, the accuracy of the pitch detection algorithm is well preserved, although the precision for the wideband domain is now increased to 1 sample (as opposed to ½ sample for narrowband).

8.3.1.4 Wideband Pitch Refinement

The issue of the pitch refinement technique is presented separately from the main pitch detection algorithm due to the significance of its adaptation for the wideband domain and its influence on the voicing determination.

The wideband pitch refinement consists of two modules. The first module is the conventional *synthetic spectral matching (SSM) method*, which was successfully exploited for pitch refinement purposes in the narrowband analysis, as shown in Chapter 3. Adaptation of the SSM method for the wideband conditions included broadening the spectral range in terms of number of harmonics and doubling of the window sizes. The second module is called the *reconstructed spectrum refinement module*. The function of this module is to realign the

peaks of the reconstructed spectrum so that they fit closer to the original spectrum. The improved accuracy of the spectral match is an important prerequisite for an improved voicing determination procedure. An example of how the reconstructed spectrum refinement module operates is illustrated in Figure 8.4.

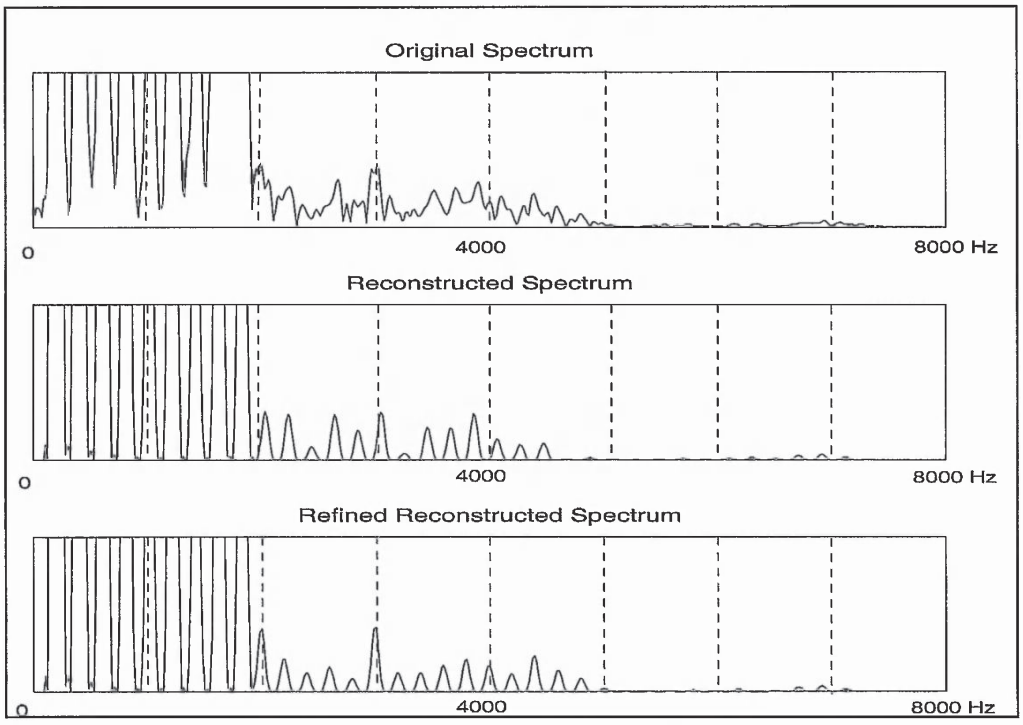


Figure 8.4 Improved spectral matching of the refined reconstructed spectrum

It can be seen in Figure 8.4 that the matching accuracy of the refined reconstructed spectrum with the original spectrum is much higher than that of the reconstructed spectrum and original spectrum. This is especially evident in the 2000-3000 Hz, 3000-4000 Hz and 4000-5000 Hz regions, where the refined reconstructed spectrum matches the highest peaks (sitting at around 2000 Hz, 3000 Hz and 4400 Hz).

The conventional pitch refinement method attempts to find the best match between the original and reconstructed spectra for which it identifies the refined pitch period with ½ sample accuracy for wideband speech. In the multi-band excitation approach it is considered that, for a voiced sub-band, the spectrum lines are centred on the harmonics of the fundamental (pitch) frequency. However, analysis of a whole voiced spectrum and of its

corresponding synthetic spectrum shows that even if the spectrum lines fit well for the lower frequencies, they can shift for the high frequencies causing a considerable mismatch between the original and synthetic spectra. This phenomenon was already observed in the narrowband but was judged to be perceptually less important. However, in wideband speech analysis, this mismatch of the magnitude lines in the higher frequencies must be taken into consideration as it can lead to erroneous amplitude calculation and voicing determination. As the reconstructed spectral harmonic lines do not match the original spectrum lines, the amplitude calculation will incur an error adding to a significant distortion in the high band spectrum. Also, a sub-band could be declared as unvoiced when the error between the original spectrum and the supposed voiced synthetic spectrum is superior to a given threshold. This error can produce very annoying artefacts.

The perceptual quality is even further affected by an increase in the fundamental frequency; the higher the fundamental frequency the greater is the relevant frequency information located elsewhere rather than centred on the harmonics. Therefore, coding of a female voice in the wideband, typically containing half the number of harmonics present in male voice, represents a considerable challenge which has not be adequately tackled by the narrowband multi-band type coders.

Analysis of the spectrum of a whole voiced frame shows that if the harmonic structure appears clearly for the lower band of the spectrum, the spectrum lines shift in the higher frequencies leading to a problem of false evaluation of the amplitudes and a false voiced/unvoiced determination. As the lines still appear for the high frequencies, but seem to be centred on the harmonics of a frequency (F0+Δf), the method implemented here is to divide the spectrum into L equal sub-bands and apply L times the pitch refinement procedure [123][124][125]. During the tests, the spectrum was split to various degrees, but the best results were obtained when it was divided into 8 segments of 1000 Hz width, as shown in Figure 8.4. Dividing the spectrum into a larger number of segments did not considerably improve the achieved output quality, hence an 8-segment split was considered to be the optimum solution. This new procedure lead to audible improvements, especially for high frequencies, and eliminated some very annoying artefacts in both male and female speech. Better matching of the spectral lines not only reduced the frequency distortion but also the magnitude distortion, which is the main cause of the audible degradation. Refinement of the

reconstructed spectrum has no implications on the overall bit rate as it is solely performed on the encoder side.

Another potential way of further improving the perceptual quality, especially in the higher frequencies, is to model the spectral error pattern. It can be shown [126] that the narrowband multi-band structure provides an efficient spectrum model for low fundamental frequency voices (mostly of male speakers) for which all the information is centred on or in close proximity of the harmonics. However, high fundamental frequency voices (mostly female speakers) contain a large part of their information content spread around the harmonics. Thus, the spectral modelling error for female voices is typically much higher than for male voices. One possible way of reducing the spectral distortion is to model this spectral error and transmit it to the decoder. The synthetic signal is now obtained by summing up the MBE synthetic signal and the modelling error signal. It can be shown that the error spectrum is of a quasi harmonic nature, especially in the high frequencies, and can thus be modelled applying the same procedure used to model the harmonic original spectrum. The listening evaluation performed in [127] showed a significant gain in quality, in particular for female voices, which sounded more natural and less metallic. However, this extra spectral information requires a significant increase in the bit rate needed for transmitting the error spectrum, for which reason it could not be used in this vocoding structure.

8.3.1.5 Wideband Voiced/Unvoiced Determination

Adapting the voiced/unvoiced determination procedure from the narrowband to wideband domain largely consisted of careful retuning of the control energy factors and accommodating a larger number of harmonics. Due to the doubling of the spectral range, 5 bits are required for more accurate SB voicing information. The correct selection of the voiced and unvoiced bands still heavily depends on the accuracy of the pitch refinement and the reconstructed spectrum refinement procedures.

8.3.1.6 Wideband Harmonic Amplitude Determination and Quantisation

The two key issues of the harmonic amplitude determination in the wideband domain are the application of the correct window sizes and the optional inclusion of the adaptive pre-emphasis prior to any LP residual filtering procedure. Whereas the simple doubling of the variably sized pitch related window functions provided the best choice during the tests, the optional adaptive pre-emphasis had to be carefully aligned with the LP residual filtering based on the half frame encoding strategy.

The spectral amplitude is determined for each harmonic of the LP residual signal of the current half-frame according to the values of the unquantised refined pitch and the quantised voicing information. The pitch-dependent residual signal windowing follows the rules earlier established in Chapter 7. In the case of the voiced harmonics, the amplitude calculation uses the MSE matching method applied in the frequency domain, similar to the pitch refinement search technique. In the case of the unvoiced harmonics, the amplitude calculation determines the RMS spectral energy over the unvoiced harmonic band.

The RMS value of the excitation harmonic amplitudes is quantised using a 16-bit logarithmic scalar quantiser. The harmonic amplitudes are normalised to their RMS values and then quantised. The RMS quantisation is performed for each half frame yielding a total of 32 bits for energy quantisation. Similar to the narrowband structure, the first eight harmonics are quantised individually. However, the remainder of harmonics are split into four groups and then vector quantised. The total length of the 16-bit shape vector quantiser used is therefore 12 (8 for individual harmonics and 4 for the average value of the remaining harmonic amplitudes). The demand for a large number of bits for harmonic amplitude quantisation is primarily driven by the requirement for high quality of synthesised wideband speech.

8.3.1.7 Phase Calculation and Quantisation

In the typical vocoding structure, the phase information is regarded as redundant and is, therefore, not transmitted by the encoder. The main reason for this is that the human hearing system is largely phase-insensitive to the narrowband speech. Instead, the phase prediction mechanism, usually employed during the voiced synthesis on the decoder side, has proved to

be sufficient when producing communication or even near-toll quality of speech. However, the wideband vocoding targets the toll speech quality whereby the role of the correct phase representation becomes very important, especially in coding the higher frequencies.

In the Amplitude Excited SB-LPC wideband vocoder, calculation of the phase information is performed on the LP residual signal spectrum using the refined pitch period, as depicted in Equation 8.1,

$$Phase(j) = \tan^{-1}\left[\frac{SWres(i).im}{SWres(i).re}\right] \quad 1 \le j \le L \tag{8.1}$$

where L represents the total number of harmonics in the current frame and i represents the point on the residual spectral scale determined as the multiple of the fundamental frequency, as shown in Equation 8.2 below.

$$i = j * \frac{(DFT\_size/2)}{(pitch/2)} \tag{8.2}$$

where both the DFT window and the pitch period are given in the time domain sample representation. Phase calculation must also fully adhere to the conditions of the tan^{-1} function[1].

Adding the phase information significantly increases the perceptual speech quality, especially in the higher frequencies. Therefore, the phase calculation and quantisation module was added to the encoder structure during the design and test simulations.

The key problem with phase representation is its quantisation stage. Conventional vector quantisation of phase requires a large number of bits, predominantly due to the large span of the phase value (0-360 degrees). In order for this vocoder to operate in the low-to-medium bit rate range, a compromise had to be made between the attainable quality and the transmission bit rate. During the course of this research, two main ideas surfaced as the optimum approaches to efficiently quantising the phase information. Both approaches largely

[1] When phase is calculated, care must be taken with respect to the sign of the complex spectral residual value. If the real part of the complex number is negative, π is either added to the positive imaginary part or subtracted from the negative imaginary part.

concentrated on the accurate representation and preservation of the first few phases, which are regarded as carrying the largest amount of phase information.

The first approach was the use of the *pulse dispersion filter* used by the Mixed Excitation Linear Prediction (MELP) coder operating at 2.4 kbps [128]. The original function of the 65th order FIR pulse dispersion filter is to spread the excitation energy within a pitch period in order to reduce some of the harshness in the synthetic speech. In phase information terms, this filter reduces the large differences that usually exist among the phases within a speech frame. The idea here was that, rather than taking into account individual phases, the phase spreading information over a frame could be vector quantised with each vector having a codebook entry. The main drawback of this approach was the complexity of the codebook training process.

The second approach represented a *hybrid phase quantisation* method, whereby the first 8 phases are quantised individually and the rest are predicted using the modified phase prediction mechanism on the decoder side. This hybrid phase model proved to be much simpler to implement and therefore more efficient with almost unnoticeable loss of the overall perceptual quality. The individual phase quantisation required 56 bits, whereby each phase was scalar quantised using 7 bits [129]. The test showed that this was more than adequate for accurate phase representation, making the second approach a more appropriate design choice.

8.3.2 Decoder

Figure 8.5 represents the decoder schematic of the Amplitude Excited SB-LPC wideband vocoder operating at 8.5 kbps. The *LSF parameters* are decoded, interpolated and transformed into the LP coefficients, which are then used for the energy smoothing of the excitation signal and in the LPC synthesis filter. The decoded *voicing* and *pitch period* information control the amount of voiced and unvoiced excitation generation. The *amplitudes* are decoded and then modified in such a way as to enhance the perceptual performance of the vocoder. The voiced excitation is generated using the time domain harmonic summation of the voiced amplitudes. The added *phase* information is partially transmitted from the encoder

and partially predicted on the decoder side. The unvoiced excitation is generated in the frequency domain and then transformed into the time domain where it is combined with the voiced excitation. The combined excitation signal is passed through the novel energy smoothing stage and then through the LPC filter to produce the synthetic speech.

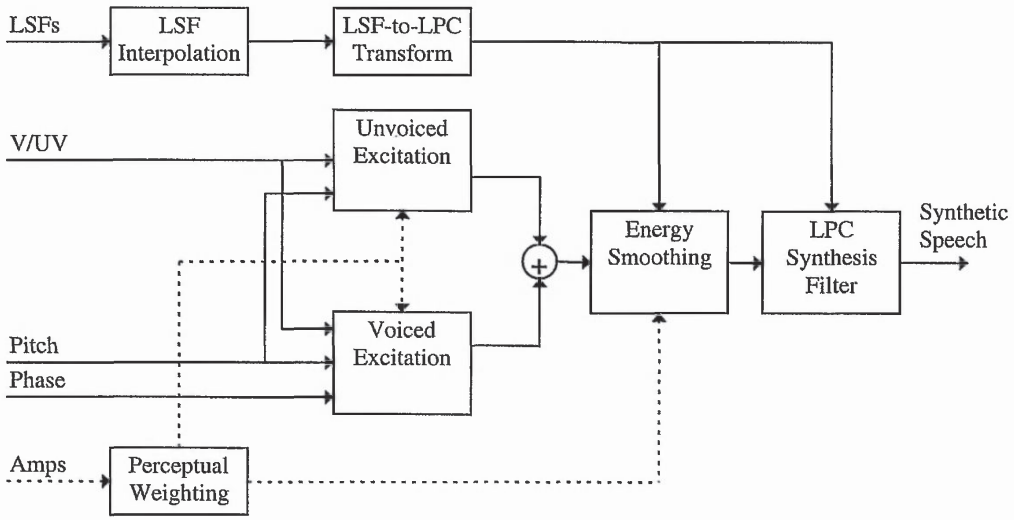


Figure 8.5 Decoder Block Diagram

8.3.2.1 Perceptual Enhancement

The perceptual enhancement technique is used to modify the two sets of spectral amplitudes before LPC filtering, the method already employed in the narrowband version of this vocoder and described in Chapter 5. The amount of amplitude modification depends on the difference between the LP spectrum and its equivalent peak interpolated LP spectral shape. The wideband adaptations included the increased number of LP coefficients and of the available peaks in the interpolated LP spectrum.

8.3.2.2 Improved Adaptive Interpolation Algorithm for Wideband Synthesis

The adaptive interpolation algorithm used in the narrowband domain was adapted in order to accommodate the potential use of variable frame lengths in the wideband domain. The main

characteristic of this algorithm remains the preservation of the *half frame delay* mechanism allowing a higher degree of overlap, as it not only controls the change between the neighbouring frames but also between the two half frames within a current decoded frame.

If compared to the example given for the narrowband domain in the previous chapter, it can be seen that both the intra-frame and inter-frame boundary conditions are also valid in the wideband case. The only real difference is the size of the transition blocks which have been made to meet the wideband frame length criteria.

The biggest advantage of this adaptive interpolation method lies with the fact that the transition boundaries, for both intra-frame and inter-frame transitions, are kept constant during various parts of the synthesis procedure of an output frame. That is, the same transition boundaries are used for the linear harmonic interpolation of the voiced excitation, linear interpolation of the unvoiced excitation, during energy smoothing of the summed excitation and during the LP synthesis of the output frame.

8.3.2.3 Wideband Voiced Excitation Generation

Generation of the wideband voiced excitation is carried out partly using the transmitted phases and partly using the principles of the phase prediction mechanism of the narrowband decoder. The tests showed that the addition of 'true' phases to the voiced excitation generation significantly improves the perceptual speech quality of this wideband vocoder.

Equation 8.3 defines the generation of the voiced excitation, whose principles were described earlier in Chapter 5. In order to meet the variable frame rate conditions and to improve the overall quality, there are again two sets of amplitudes for each half of the encoded frame size.

$$e_{voiced}(i) = \sum_{j=1}^{L} Amp_{\mathrm{mod}}(j)\cos\left[j\phi(i)\right] \tag{8.3}$$

where $Amp_{mod}(j)$ takes on the modified amplitude value if the jth harmonic is voiced, or $Amp_{mod}(j)=0$ when the jth harmonic is unvoiced.

The phase prediction mechanism is similar to the one employed in the narrowband decoder, although modifications had to be incorporated during the design stages with respect to the wideband synthesis frame length conditions. Equation 8.4 represents the fundamental phase $\phi(i)$ as the sum of the interpolated fundamental frequency function $\omega(i)$, defined in Equation 8.5, and the fundamental phase of the previous sample.

$$\phi(i) = \phi(i-1) + \omega(i) \tag{8.4}$$

$$\omega(i) = \begin{cases} \omega_{last}\left(\frac{i+0.5}{N}\right) + \omega_{next}\left(1 - \frac{i+0.5}{N}\right), & |\omega_{last} - \omega_{next}| \le 0.25(\omega_{last} + \omega_{next}) \\ \omega_{last}, & |\omega_{last} - \omega_{next}| > 0.25(\omega_{last} + \omega_{next}) \text{ and } i \le \frac{N}{2} \\ \omega_{next}, & |\omega_{last} - \omega_{next}| > 0.25(\omega_{last} + \omega_{next}) \text{ and } i > \frac{N}{2} \end{cases} \tag{8.5}$$

The interpolation process shown in Equation 8.5 is only performed when the pitch value deviates by less than 25% of its value from one frame to the next.

The transmitted 'true' phases are added on to the phase prediction mechanism. The phase synthesis is mainly based on modifying the predicted phase using the estimated phase difference method. The first part of this method is the estimation of the difference between the transmitted 'true' phases and the predicted phases, as shown in Equation 8.6,

$$ph\_diff_n(j) = ph_n(j) - \phi_n((N-1)j) \tag{8.6}$$

where n is the currently transmitted frame, j is one of the L harmonics ($1 \le j \le L$) and N is the size of the synthesis frame. The obtained phase difference is then compared with the phase difference of the previously decoded synthesis frame producing the total estimated phase difference, as shown in Equation 8.7 below

$$tot\_est\_ph\_diff(j) = ph\_diff_n(j) - ph\_diff_{n-1}(j) \tag{8.7}$$

where n is the current frame, n-1 represents the previously transmitted frame and the total estimated phase difference is kept within the bounds of $-\pi$ and $+\pi$. The synthesis phase is calculated as the sum of the predicted phase, phase offset and phase memory, as shown in Equation 8.8

$$phase(i) = \phi(i) \times j + i \times offset + memory \tag{8.8}$$

where

$$offset = \frac{tot\_est\_ph\_diff(j) / syn\_frame\_length}{2\pi}$$
$$memory = \frac{ph_{n-1}(j)}{2\pi} \quad (8.9)$$

During voiced excitation, the linear harmonic amplitude interpolation adheres to the transition principles and boundaries set by the adaptive interpolation mechanism. The amplitude interpolation is performed at both the intra-frame boundaries and the inter-frame boundary, as shown in the previous section. The tests proved again that the perceptual quality of this vocoder is significantly increased by adopting two sets of amplitudes per encoded frame rather than the traditional one set of amplitudes per encoded frame, especially in cases when long frame sizes are interpolated and synthesised.

8.3.2.4 Wideband Unvoiced Excitation Generation

Unvoiced excitation generation follows the rules established during the narrowband synthesis, as explained in Chapter 7. It is generated twice per frame with no trapezoidal windowing prior to the FFT and with the Overlap-Add method performed both in the middle and at the ends of the synthesised frame. The major adaptation change included the switch from the narrowband frame sizes and conditions to the wideband.

8.3.2.5 Energy Smoothing

The function of the energy smoothing algorithm is to eliminate the energy spikes and energy fluctuations in synthetic speech, which may arise because of the potential spectral mismatch occurring during the interpolation procedure in the LPC synthesis filter. This spectral mismatch is largely caused by the deficiency in the linear interpolation function, which, although energy dependent, does not provide an optimum solution during sudden changes in the speech envelope. The energy smoothing algorithm detects the energy fluctuations in the synthetic speech frame and attempts to smooth them out. The energy smoothing is performed before the LPC synthesis stage using the generated summed excitation, the set of interpolated harmonic magnitudes and LPC coefficients. The energy smoothing algorithm operates on a sub-frame basis.

The *first stage* of this algorithm is the interpolation of the harmonic magnitudes using the adaptive interpolation method, as depicted in Equation 8.10

$$Mag\_\text{int}(j) = ad\_fac \times Mag_{hf} + (1 - ad\_fac) \times Mag_{hf-1} \tag{8.10}$$

where *ad_fac* is the adaptive interpolation factor, and *hf* and *hf-1* represent the current half frame and previous half frame, respectively, whereby the position of the sub-frame determines which half frames are used for the adaptive interpolation. The interpolated amplitudes serve to extrapolate the FFT magnitude spectrum.

The *second stage* of the energy smoothing algorithm is the extrapolation of the LP spectral envelope using the interpolated LP coefficients. A similar procedure has already been applied in the decoder for perceptual enhancement purposes.

In the *third stage* of this algorithm the energy factor is computed for each sub-frame using Equation 8.11

$$energy\_fac = \sum_{i=0}^{FFTsize} Mag_{amps}(i) \times Mag_{lpc}(i) \tag{8.11}$$

where Mag_{amps} are the spectral magnitudes of the extrapolated amplitude spectrum and Mag_{lpc} are the spectral magnitudes of the extrapolated LPC spectrum. The energy factor is determined for each sub-frame of the synthesis frame. It is also calculated for the end points and middle point of the synthesis frame, which serve as control points for energy smoothing. The energy line is extrapolated between the control points. Each sub-frame energy is scaled with respect to the extrapolated energy line.

Figure 8.6 illustrates an example of the energy smoothing mechanism before and after its application. The smoothing effect at the transition from Frame (n-1) to Frame (n) can be clearly seen.

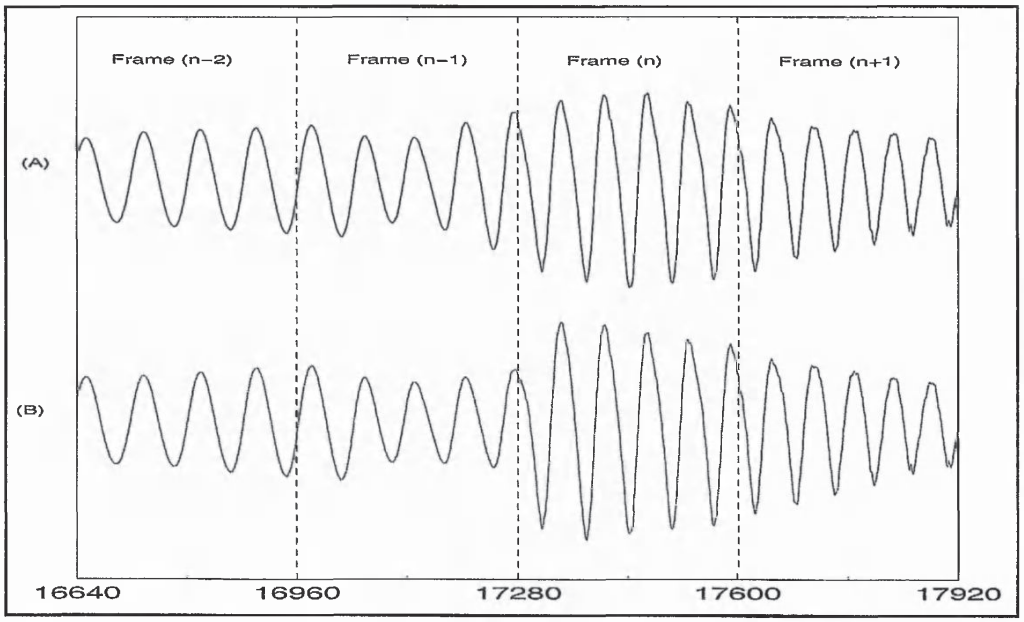


Figure 8.6 An example of the energy smoothing mechanism:
(A) Synthetic speech with the smoothing application
(B) Synthetic speech with no smoothing application

The energy smoothing algorithm significantly increased the cleanness and general smoothness of the synthetic speech during the tests. Moreover, its function proved to be applicable to both the narrowband and wideband domain.

8.3.2.6 LPC Synthesis Filter

The LP excitation signal is the summation of both the generated voiced and unvoiced excitations. The filtered signal is synthesised on a 80-sample sub-frame basis, i.e. every 2.5 ms, using the interpolated LP coefficients. The interpolation of the LP coefficients is performed according to the principles of the adaptive interpolation method for both intra-frame and inter-frame transition boundaries, as described in the previous section. Spectral de-emphasis, fixed or adaptive, is performed on the filtered signal before synthetic speech output is produced.

8.3.3 Variable Frame Rate in Wideband Speech

The variable frame rate concept was also introduced into the wideband speech domain. The rules of the variable frame rate already established for the narrowband speech processing were adapted for the wideband speech conditions. The most notable change was the adaptation of the frame sizes with respect to both the analysis on the encoding side and interpolation and synthesis on the decoding side. The other important factor is the 'search-and-compare' segmentation algorithm, which also required careful retuning.

The main aim of introducing the variable frame rate method into the wideband domain was to test the viability of this concept in the new, wider bandwidth environment with respect to both the reduction in the frame rate and the speech quality preservation of the fixed frame rate.

8.3.4 Bit Allocation Scheme

Table 8.1 depicts the overall bit allocation scheme for the 8.5 kbps Amplitude Excited SB-LPC wideband vocoder when a fixed frame length of 20 ms is used.

Table 8.1 Wideband Bit Allocation Scheme

	Bit Allocation							
Frame Length (ms)	Frame Size Header	Pitch	Voicing	LSF	Ampl Energy	Ampl Shape	Phase	Fixed Bit Rate (kbps)
20	**3**	**8**	**5**	**50**	**32**	**16**	**56**	**8.5**

The number of bits allocated for the frame size header, pitch period and voicing information was optimised for an efficient and high quality performance. However, the amount of bits allocated for quantisation of LSFs, amplitudes and phases were purposely exaggerated in order to lift the overall speech quality of this wideband vocoder to its maximum. Optimum bit allocation was not a design issue here as long as the overall bit rate remained within the low-to-medium bit rate range.

8.3.5 Performance Evaluation

The performance of the variable frame length structure of the SB-LPC wideband vocoder was assessed using the MOS scale, whereby the subjective quality of the variable frame length scheme was compared with the subjective quality of the fixed frame length scheme. The relevant fixed rate wideband speech coding standard (G.722 ITU-T with its three operational mode rates of 64, 56 and 48 kbps) was included in the tests as a reference point. Sixteen subjects, mostly experienced listeners, were used to conduct the informal MOS listening test. The subjects were divided into four listening groups, whereby each group was given a separate set of randomly selected sentences, two male and two female.

Table 8.2 MOS Test Results

Vocoder Type	Frame Length (samples)	MOS	Variance	Change in average frame rate (voiced frames only)
64 kbps G.722	320 (fixed)	4.25	1.12	None
56 kbps G.722	320 (fixed)	3.93	1.94	None
48 kbps G.722	320 (fixed)	3.65	1.52	None
8.5 kbps SB-LPC (wideband)	320 (fixed)	3.54	1.15	None
8.5 kbps SB-LPC (wideband)	160-640 240-480 320-640	3.34 3.43 3.48	0.79 0.98 1.06	+2% +1% -14%

The results, as given in Table 8.2, show that all versions of the Amplitude Excited SB-LPC wideband vocoder performed comparably to the G.722 standard coder operating at 48 kbps during the tests. It can therefore be concluded that the main aim of designing a high quality low-to-medium bit rate wideband vocoder was achieved.

The outcome of this MOS test has also further proved the fact that the G.722 standard coder can provide toll quality speech at the rates of 64 and 56 kbps, but its performance starts to degrade at 48 kbps with an increase in the granular noise. As part of the testing procedure, an attempt was made to assess the G.722 performance at 40 kbps and, even, at 32 kbps. The reason behind this was the need to provide the 8.5 kbps Amplitude Excited vocoder with a comparable bit rate coder for evaluation purposes. However, in the course of these preparatory tests it soon became clear that the 8.5 kbps wideband vocoder, with its high level

of cleanness and smoothness as perceived by a considerable majority of the test subjects, could easily outperform the G.722 standard operating at these rates. Therefore, the G.722 rates below 48 kbps were not considered during the tests.

The performance of the fixed and variable frame rate versions of the Amplitude Excited SB-LPC wideband vocoder has indicated that the perceptual quality and the change in frame rate depend on the selected frame size range of the variable frame rate codec.

The fixed frame rate version provided the best output quality with respect to all the other variable frame rate versions during the tests. However, the difference between the attained quality of the fixed frame rate and the variable frame rates was almost negligible, a fact confirmed by the variance parameter.

The short range version that permits frame extending only (320-640 samples) produced similar quality of speech as the fixed rate codec. However, its real gain over the fixed rate structure is the average drop in voiced frame rate by 14%. The moderate range version (240-480 samples) and the full range version (160-640 samples) performed slightly worse than the fixed frame rate, although they did not experience any large fluctuations in frame rate. This decrease in speech quality indicates the level of difficulty and complexity involved in jointly coding very short and very long frames, the conclusion already reached during the narrowband design of the variable frame rate. This is especially related to the adverse block edge effects which characterise the variable interpolation sizes within the frame boundaries. However, the tests also confirmed the importance of selectively using highly accurate shorter frames in modelling sharp speech onsets. Restricting the use of shorter frames (160-sample and 240-sample) to speech onset regions proved beneficial to the variable frame rate scheme both in terms of the subjective quality and the reduction in average frame rate.

8.4 Conclusions

The research presented in this chapter was aimed at improving the perceptual quality of wideband speech at low-to-medium bit rates. The current wideband speech standard, G.722

ITU-T coder, can offer very high quality speech at rates of 48 kbps and above. However, there is a growing need to reduce this high bit rate whilst maintaining high speech quality, for purposes such as audio video teleconferencing, high quality voice-mail services and high quality wideband telephony. This work was carried out in response to such a need.

Nowadays, most coding schemes, including the G.722 standard, utilise a *split band* approach to code wideband speech. This allows for a more flexible coding structure, however, it also increases coding delay and complexity. In order to overcome these difficulties, the Amplitude Excited SB-LPC wideband vocoder presented in this chapter treats the signal as a *full band* one.

The fundamental principles of this wideband algorithm are similar to those of the coding structure of its narrowband equivalent described previously in Chapters 5 and 7. The main adaptations had to be made above all to accommodate longer frame sizes (320 samples within a 20 ms frame), different windowing options and wider spectral bandwidth (7 kHz) for voicing analysis and harmonic amplitudes calculation. Various control factors within the pitch detection, voicing decision and amplitude calculation were also retuned for the wideband conditions. Completely new additions for improving the wideband performance included the phase information module on the encoder side and the energy smoothing algorithm on the decoder side.

The Amplitude Excited SB-LPC wideband vocoder nominally operates at 8.5 kbps. Variable frame rates were also simulated and tested. The key listening test results showed that this wideband vocoder performs comparably to the G.722 standard operating at its lowest bit rate mode of 48 kbps. The main quality points of this wideband vocoder, as identified by the test subjects, were the level of cleanness and smoothness attained. The variable frame rate modules also performed comparably to the G.722 standard at 48 kbps, although their performance was generally slightly lower than that of the fixed frame rate.

The potential of the Amplitude Excited SB-LPC wideband vocoder has certainly not been fully exploited. The initial design and results presented in this chapter can only indicate the underlying quality of this vocoder. There is however a lot of room for improving the perceptual quality of both the fixed and the variable frame rates, of which the most important

are the optimisation of the 'search-and-compare' segmentation algorithm and further retuning of the in-built factors that control the estimation of the wideband modelling parameters. In addition, the vocoder's bit allocation scheme has been adopted in order to boost its perceptual performance to its maximum. Further investigation into more efficient quantisation methods and routines could very well lead to the development of an even lower bit rate high quality wideband vocoder.

Chapter 9.

9. Conclusions

9.1 Preamble

The subject of this thesis has been the design and development of speech compression systems based on variable rate coding principles. The new algorithms include the variable bit rate and the variable frame rate coding schemes which were developed for narrowband speech at very low bit rates (2.75 kbps and below). A variable frame rate system was also adapted and upgraded to operate for wideband speech at low-to-medium bit rates (at around 8.5 kbps). The work presented in this thesis is divided into three main research areas.

1. The primary aim has been to *improve the perceptual quality* of both narrowband and wideband speech by adopting a variable frame rate approach. The underlying idea behind this is that the variable frame rate approach offers much needed flexibility to the coded frame sizes, which can more accurately capture and model speech events of various duration than the traditional fixed frame length coding systems.

2. Not only the enhancement of speech quality, but also the *reduction in the transmission rate* has topped the agenda of this research effort. Lowering the bit/frame rate whilst preserving the high perceptual quality is the ultimate design goal in low bit rate speech coding. Both the simulated variable bit rate scheme and the developed variable frame rate systems have mainly focused on this issue, showing the potential to significantly reduce the transmission rate by almost ¼ on average.

3. Finally, the introduction of the wider speech bandwidth has always carried a potential for further improving the perceptual quality of speech. The wideband speech coding algorithm was designed and developed with the key aim of attaining very high wideband speech quality whilst operating at low-to-medium bit rates. The variable frame rate scheme was also integrated to operate in the wideband domain.

In the following sections a brief overview of the research achievements reported in this thesis is presented. This is followed by a short discussion on possible directions of future research activity, which may be undertaken to build upon and further enhance the coding schemes detailed in this thesis.

9.2 Concluding Overview

The main body of the thesis can be broadly divided into two parts. The first part comprises a survey of current speech transmission systems and speech coding techniques, whilst the second part presents variable rate vocoding schemes which have been designed and developed during the course of this research.

In Chapter 2 a brief analysis has been made of the factors that influence the design or choice of a speech coding algorithm for a given application. These include speech quality, complexity, bit rate, delay, robustness to channel errors and background noise, as well as non-speech signal requirements and multi-rate operation. Various applications of the main speech coders have been surveyed. The speech coding requirements and constraints of each application have been identified and briefly analysed. It has emerged that a current trend in

speech coding research in the *communication* to *near toll* speech quality range is to reduce bit rates below 4 kbps whilst still preserving high speech quality, in order to utilise the existing bandwidth more efficiently and thus accommodate the rapidly growing user demand. The research work undertaken and presented in this thesis is in line with this trend.

The study of the parametric model of the human speech production system and the associated coding techniques used by a majority of current low bit rate speech coders has been presented in Chapter 3. It has been shown that extremely high coding efficiency can be achieved using *linear prediction* to encode short-term spectral energy information, especially when combined with various *quantisation* mechanisms. The long-term correlation in speech is detected using *pitch prediction*. The principle pitch prediction techniques have been investigated together with their respective merits and disadvantages. Special consideration has also been given to other speech coding tools, such as *post-filtering* and *interpolation*, which have proved capable of reducing the effects of possible spectral distortions and maintaining high perceptual quality. Due to its high coding efficiency and perceptual speech quality, the parametric model and its pertinent coding tools were also adopted and further improved during the design and development of the low bit rate coders presented in this research.

A brief overview of the main vocoding techniques has been provided in Chapter 4. The most prominent vocoder among them, the I-MBE vocoder, has been discussed in more detail, as it served as one of the starting points of the research presented in this thesis. The key advantages of the I-MBE vocoder, capable of producing very high quality at its designated bit rate of 4.15 kbps, have been analysed, such as its robust multi-band excitation speech model and sophisticated algorithms to estimate the speech modelling parameters. However, its speech quality and coding efficiency start to deteriorate rapidly with the reduction in bit rate largely because of its relatively inefficient reconstruction of the multi-band spectrum. Nevertheless, the I-MBE represented one of the main starting points in this research. Adapting and further improving its vocoding structure so that it can operate efficiently and produce high quality even at very low bit rates by combining it with the linear prediction model represented one of the key design goals during the course of this work.

The subject of Chapter 5 has been the design and simulation of the fixed rate linear prediction vocoding systems, namely, the Time Envelope SB-LPC system and the Amplitude Excited

SB-LPC system, operating at 2.4 kbps. The two vocoding schemes employ the newly developed split band voicing approach, which proved more efficient at very low bit rates than the MBE voicing approach. For each system, detailed information about the basic structure and principles of operation of the encoder and decoder have been provided. The advantages and disadvantages of both coders have also been assessed in terms of their technical merits and their output perceptual quality. The thorough analysis of the fixed rate coding structures was mainly aimed at assessing their particular suitability to provide the grounds for the development of the variable rate coding systems. Both the objective and subjective tests showed that the Amplitude Excited SB-LPC vocoder is much more acceptable as the basis for designing the variable frame rate scheme. Not only its subjective quality performance but also its flexible coding structure completely outweighed the Time Envelope approach.

In Chapter 6 a new variable quantisation scheme has been proposed and integrated with a 2.4 kbps Time Envelope SB-LPC vocoder, as it was the first vocoder to be studied and further developed during the course of this research. The scheme is based on the accurate source-dependent classification of different speech regions where each speech category (silence, unvoiced, mixed and fully voiced) is assigned a unique bit allocation scheme. By varying the allocation of bits in relation to the identified speech region, a variable coding rate is achieved.

The minimum bit rate of only 37.5 bps is attained when only silence is transmitted. However, the silence detection mechanism requires an algorithmic delay of up to 80 ms, four times longer than the algorithmic delay of the fixed bit rate scheme. This prolonged algorithmic delay is the only disadvantage of this otherwise highly efficient variable rate coding scheme. The maximum bit rate is achieved when only mixed segments, containing both voiced and unvoiced speech, are transmitted. This maximum bit rate of 2.5 kbps is slightly higher than the 2.4 kbps utilised for the fixed bit rate because of 2 additional bits/frame representing the speech classification header.

The tests on individual speech files showed that the average operational bit rate lies in the range of 1.35-2.23 kbps, although a further reduction in the average bit rate can be attained when the variable rate quantisation scheme is employed as part of a telephone link for a two-way conversation, where one side is silent roughly 60% of the time. The comparison tests

between the variable bit rate and fixed bit rate indicated that any difference in the speech quality was hardly noticeable.

The research presented in Chapter 7 has introduced the idea of variable frame length coding based on the time varying characteristics of the speech signal, where longer frame lengths are used to code highly correlated steady state speech segments and unvoiced speech, whilst shorter frame sizes are employed to shape speech transients. The most important part of the variable frame rate coding is the accurate speech segmentation algorithm. Three different segmentation principles have been investigated, the most accurate and also the most complex being phonetically based.

The variable frame rate coding was at first simulated using the Time Envelope SB-LPC vocoder. Its main strengths lie in its ability to accurately segment speech according to the adopted source-dependent criteria and adaptively interpolate frames according to their varying sizes. The MOS test showed that the variable frame rate structure does not significantly trail behind the subjective performance of the fixed frame rate algorithm, whilst reducing the average frame rate by 7% for voiced frames. However, the main disadvantages of this vocoder have been attributed to its time domain based pitch detection algorithm, which is prone to pitch multiple and sub-multiple errors, and the practical complexity of quantising its variably sized time envelope. Nevertheless, the Time Envelope vocoder proved to be very useful in further developing the variable frame rate system, as most of its key functions were adopted in designing the Amplitude Excited SB-LPC vocoder.

The Amplitude Excited SB-LPC vocoder succeeded in significantly improving the performance of the Time Envelope version by switching its troubled time domain functions, such as pitch detection algorithm and time envelope calculation, into the frequency domain. The new frequency domain pitch detection algorithm proved to be much more resilient to the pitch multiple and sub-multiple errors. As for the time envelope, this vocoder uses and quantises spectral amplitudes placed within a finite spectral length independent of its frame size, thus eliminating the need for variable length time envelope quantisation. The switch to frequency domain also offered closer control of the spectral amplitude behaviour which increased coding flexibility and accuracy, especially with respect to the coder's interpolation mechanism. Other major improvements were made to the source-dependent segmentation

method on the encoder side and the adaptive interpolation method on the decoder side. All this significantly improved the subjective quality of the vocoder as shown in its MOS test results. This variable frame rate vocoder managed not only to produce similar or even higher quality of speech than its fixed frame rate version, but also to outperform the I-MBE vocoding standard during the tests as a result of its increased smoothness and cleanness.

A successful attempt was also carried out to integrate the phonetically based speech recognition module with the Amplitude Excited SB-LPC vocoder. The preliminary results showed a lot of potential in this field both in terms of preserving high quality of speech and reducing the frame rate.

The research presented in Chapter 8 aimed at improving the perceptual quality of wideband speech at low-to-medium bit rates for purposes such as audio video teleconferencing, high quality voice-mail services and high quality wideband telephony. This work was carried out in response to such a need. Nowadays, most coding schemes, including the G.722 standard, utilise a *split band* approach to code wideband speech. This allows for a more flexible coding structure, however, it also increases coding delay and complexity. In order to overcome these difficulties, the Amplitude Excited SB-LPC wideband vocoder presented in this chapter treats the signal as a *full band* one.

The fundamental principles of this wideband algorithm are similar to those of the coding structure of its narrowband equivalent. The main adaptations had to be made above all to accommodate longer frame sizes, different windowing options and wider spectral bandwidth for voicing analysis and harmonic amplitudes calculation. Various control factors within the pitch detection, voicing decision and amplitude calculation were also retuned for the wideband conditions. Completely new additions for improving the wideband performance included the phase information module on the encoder side and the energy smoothing algorithm on the decoder side.

This wideband vocoder nominally operates at 8.5 kbps, although variable frame rates were also simulated and tested. The informal listening test results showed that it performed close to the G.722 standard operating at its lowest bit rate mode of 48 kbps. The main quality points of this wideband vocoder are the level of cleanness and smoothness attained. The variable

frame rate modules also performed comparably to the G.722 standard at 48 kbps, although their performance was generally slightly lower than that of the fixed frame rate.

9.3 Future Work

Although the variable rate speech coding algorithms developed during the course of this research produced high quality speech, a number of design issues that may lead to future expansion of this research direction have also been identified.

• The variable quantisation method has been shown to significantly reduce the bit rate, especially for silent and unvoiced speech categories. However, in order to make this method even more efficient, one must focus on a better bit rate utilisation when coding voiced speech. One idea would be to introduce inter-frame coding of slow varying pitch period and LPC parameters, especially for fully voiced speech, where there is high correlation between consecutive spectral envelopes. More work is also needed on the network system level, e.g. in terms of efficiently packetising the variable bit rate information.

• In the variable frame rate domain any further progress will certainly require a joint segmentation effort by both the phonetic based recognition module, which could serve as a first step in accurately segmenting speech into smaller units, and the source-dependent segmentation algorithm, whose function could be to fine tune the frame division within phonemes and at phonetic transitions. Another useful idea may be to make the selection of variable frame lengths more suited to the segmented phoneme sizes. In this research, the maximum frame size was selected to be 40 ms long for, above all, purely practical reasons. As the actual phoneme length can be in excess of 400 ms, it is certainly plausible to allow the segmented frames to follow the phoneme length selection more closely. Implementation of such an idea should at least further reduce the average frame rate.

• The variable frame rate method described in this thesis was based on the efficient extraction of fixed frame and sub-frame sizes, which was found suitable for integration with the MBE-LPC vocoding model. However, our new thinking has shown that the sub-frame approach can

potentially be replaced by the prototype waveform interpolation (PWI) method, which is based on the accurate pitch period extraction [130]. The newest research trends in very low bit rate speech coding show that PWI coders are capable of producing very high quality [131][132]. A PWI coder will certainly require a multi-pitch speech segmentation approach if it is to be integrated with the variable frame rate scheme similar to the ones implemented in this thesis.

• Current industry trends indicate that in the drive for higher quality the industry is ready to accept longer processing delays [133][134]. In terms of the variable rate speech coding proposed in this thesis, it means that further pre- and post-processing modules and techniques will be welcome to be added to the existing variable rate schemes. The development of these new coding stages may prove significant in achieving more accurate and more efficient segmentation of speech, and thus higher output quality.

• The potential of the Amplitude Excited SB-LPC wideband vocoder has certainly not been fully exploited. Although the initial design and results presented have indicated the underlying quality of this vocoder, there is a lot of room for improving the perceptual quality of both the fixed and the variable frame rates, of which the most important are the optimisation of the segmentation algorithm and retuning of the in-built factors that control the estimation of the wideband modelling parameters. In addition, the vocoder's bit allocation scheme has been adopted in order to boost its perceptual performance to its maximum. Further investigation into more efficient quantisation methods and routines could very well lead to the development of even lower bit rate high quality wideband vocoders.

Appendix

List of Publications

1. Villete S., Stefanovic M, Atkinson I. A., Kondoz A. M., "High Quality Split-Band LPC Vocoder and its Fixed Point Real Time Implementation", *Proc. EUROSPEECH '97,* Vol. 3, pages 1243-1246, Rhodes, Greece, September 1997.

2. Villette S., Stefanovic M., Kondoz A. M., "Split Band LPC Based Adaptive Multi-Rate GSM Candidate", *Proc. Int. Conf. on Acoustics Speech and Signal Processing,* paper no. 1798, Arizona, USA, May 1999.

3. Villette S., Stefanovic M., Kondoz A. M., "A Multi-Rate Speech and Channel Codec: A GSM AMR Half-Rate Candidate", *Proc. EUROSPEECH '99,* Vol. 4, pages 1755-1758, Budapest, Hungary, September 1999.

4. Stefanovic M., Kondoz A. M., "Source-Dependent Variable Rate Speech Coding Below 3 kbps", *Proc. EUROSPEECH '99,* Vol. 3, pages 1487-1490, Budapest, Hungary, September 1999.

5. Stefanovic M., "Variable Rate Speech Coding for Satellite Mobile Communications", Winner of the annual INMARSAT prize, October 1999.

6. Stefanovic M., Sriratanaban C., Kondoz A.M., "Split-Band LPC Vocoding with Phonetically based Segmentation below 3 kbps", *Proc. Int. Conf. on Acoustics Speech and Signal Processing,(submitted for publication),* Istanbul, Turkey, June 2000.

List of Abbreviations

AaS	Analysis and Synthesis
AbS	Analysis by Synthesis
ADPCM	Adaptive Differential Pulse Code Modulation
AM	Auto-correlation Method
AMDF	Average Magnitude Difference Function
AMR	Adaptive Multi-Rate
APC	Adaptive Predictive Coding
ATM	Asynchronous Transmission Mode
BER	Bit Error Rate
CCITT	International Telegraph and Telephone Consultative Committee
CELP	Code Excited Linear Prediction
CFD	Cumulative Frequency Distribution
CM	Covariance Method
DAM	Diagnostic Acceptability Measure
DFT	Discrete Fourier Transform
DRT	Dynamic Rhyme Test
DSP	Digital Speech Processing/Processor
DTMF	Dual Tone Multi-Frequency
DTX	Discontinuous Transmission
FEC	Forward Error Correction
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FR-GSM	Full Rate GSM
GPRS	General Packet Radio Service
GSM	Pan-European Digital Cellular Mobile Radio System
HF	High Frequency
HPP	Harmonic Peak Picking
HR-GSM	Half Rate GSM
ICASSP	International Conference on Acoustics, Speech and Signal Processing
I-MBE	Improved Multi-Band Excitation
INMARSAT	International Maritime Satellite
IP	Internet Protocol
ISDN	Integrated Services Digital Network
ITU	International Telecommunication Union
LBG	Linde Buzo Gray
LD-CELP	Low Delay Code Excited Linear Prediction
LM	Lattice Method
LMS	Least Mean Square

LP	Linear Prediction
LPC	Linear Predictive Coding
LSF	Line Spectral Frequency
LSP	Line Spectral Pair
LTP	Long Term Prediction
MBE	Multi-Band Excitation
MELP	Mixed Excitation Linear Prediction
MOS	Mean Opinion Score
MSE	Mean Square Error
PCM	Pulse Code Modulation
PCN	Personal Communications Network
PDA	Pitch Detection Algorithm
PDF	Probability Density Function
PSTN	Public Switched Telephone Network
QMF	Quadrature Mirror Filter
RELP	Residual Excited Linear Prediction
RMS	Root Mean Square
RPELPC	Regular Pulse Excited Linear Predictive Coding
SAPDA	Segmented Auto-correlation Pitch Detection Algorithm
SB	Split Band
SB-ADPCM	Sub-Band Adaptive Differential Pulse Code Modulation
SID	Silence Description
SMM-PDA	Sinusoidal Model Matching Pitch Detection Algorithm
SNR	Signal to Noise Ratio
SSM	Synthetic Spectral Matching
TE	Time Envelope
TIMIT	Texas Instruments/Massachusetts Institute of Technology
VAD	Voice Activity Detector
VBRC	Variable Bit Rate Coding
VFLS	Variable Frame Length Coding
VLSI	Very Large Scale Integration
VQ	Vector Quantisation
VSELP	Vector Sum Excitation Linear Prediction
V/UV	Voiced/Unvoiced
WPP	Waveform Peak Picking

References

[1] P. Sweeney. "Error Control Coding: An Introduction". Prentice Hall, 1991.

[2] R. Haslem and C. Johnson. "ISDN Service and Applications Support in the UK". British Telecom Eng., 11:213-219, October 1992.

[3] J.G. Proakis and D.G. Manolakis. "Digital Signal Processing - Principles, Algorithms and Applications". Third Edition, Prentice Hall, 1996.

[4] L.A. Meacham and E. Peterson. "An experimental multichannel pulse code modulation system of toll quality". *Bell Sys. Technical Journal*, 27:1-43, 1948.

[5] B.M. Oliver, J.R. Pierce and C.E. Shannon. "The Philosophy of PCM". *Proc. IEEE*, pp.1324-1331, November 1948.

[6] ITU-T. "CCITT Recommendation G.721: 32 kbps Adaptive Differential Pulse Code Modulation (ADPCM)". International Telecommunications Union, 1984.

[7] D.W. Griffin. "Multi-Band Excitation Vocoder". PhD Thesis, M.I.T., USA, 1987.

[8] L.A. Meacham and E. Peterson. "An experimental multichannel pulse code modulation system of toll quality". *Bell Sys. Technical Journal*, 27:1-43, 1948.

[10] ITU-T. "CCITT Recommendation G.711: 64 kbps Pulse Code Modulation (PCM) of Voice Frequencies". International Telecommunications Union, 1972.

[11] W. Voiers. "Diagnostic Evaluation of Speech Intelligibility". Speech Intelligibility and Recognition, Dowden Halden Ross, 1977.

[12] Dynastat Inc. "Subjective Testing and Evaluation of Voiced Communication Systems". Dynastat Inc., USA, 1997.

[13] W.Daumer. "Subjective Evaluation of Several Efficient Speech Coders". *IEEE Transactions on Communications*, pp. 655-662, 1982.

[14] F.F. Mazda. "Telecommunications Engineer's Reference Book". Butterworth-Heinemann Ltd., pp. 30/5-30/7, 1993.

[15] I. Boyd. "Speech Coding for Telecommunications". *IEE Electronics and Communication Engineering Journal*, pp. 273-283, October 1992.

[16] K. Hellwig, P. Vary, D. Massaloux and J. Petit. "Speech Codec for the European Mobile Radio System". *Proc. IEEE Global Telecommunications Conf.*, Vol. 2, pp. 1065-1069, Dallas (USA), November 1989.

[17] D.K. Freeman, G. Cosier, C.B. Southcott and I. Boyd. "The Voice Activity Detector for the Pan-European Digital Cellular Mobile Telephone Service". *Proc. IEEE Int. Conf. On Acoustics, Speech and Signal Processing (ICASSP)*, pp. 369-372, Glasgow, Scotland, May 1989.

[18] S. Villette, M. Stefanovic and A. Kondoz. "Split Band LPC Based Adaptive Multi-Rate GSM Candidate". *Proc, IEEE ICASSP*, paper no. 1798, Phoenix, Arizona (USA), February 1999.

[19] R. Steele. "Mobile Radio Communications". Pentech Press, London, pp. 92-185, 1992.

[22] CCITT: Rec. G.728. "16 kbps LD-CELP". 1992.

[23] R. Haslem and C. Johnson. "ISDN Service and Applications Support in the UK". British Telecom Eng., 11:213-219, October 1992.

[24] CCITT: Rec. G.722. "64 kbps SB-ADPCM", 1988.

[25] S. Quackenbush. "A 7 kHz Bandwidth, 32 kbps Speech Coder for ISDN". *Proc. IEEE ICASSP*, pp. 1-4, 1991.

[26] J. Ayre, J. Callaghan and S. Hoffos. "The MULTIMEDIA Yearbook". Interactive Media Publications, London, Chapter 1, 1995.

[24] CCITT: Rec. G.726. "Digital Circuit Multiplication Equipment", 1988.

[25] M.R. Suddle. "Speech Coding in Private and Broadcast Networks". PhD Thesis, University of Surrey, 1996.

[26] M. Decina and G. Modena. "CCITT Standards on Digital Speech Processing". *IEEE Journal on Selected Area in Communications*, 6(2):227-233, February 1988.

[27] J.E. Natvig et al. "Speech Processing in the Pan-European Digital Mobile Radio System (GSM) - System Overview". *Proc. IEEE ICASSP*, pp. 1060-1064, 1989.

[28] ETSI/TC/SMG: Rec. 06.20. "European Digital Cellular Telecommunications System Half Rate Speech Part 2: Half Rate Speech Transcoding". January 1995.

[29] I.A. Gerson and M.A. Jasiuk. "A 5600 bps VSELP Speech Coder Candidate for HR-GSM". *Proc. of EUROSPEECH Conf*, pp. 253-256, Berlin, Germany, September 1993.

[30] G. Vincent. "Personal Communications: The Dream and the Reality". IEE Review, pp. 299-303, September 1990.

[31] B.G. Evans. "Satellite Communication Systems". 3rd Edition, IEE Telecommunication Series 38, UK, 1999.

[32] DVSI (Digital Voice Systems Inc.). "INMARSAT-M Voice Coding System Description". Draft 1.3, 1991.

[33] N.S. Jayant and P. Noll. "Digital Coding of Waveforms - Principles and Applications to Speech and Video". Pages 115-220, Prentice Hall, 1984.

[34] A. McCree, K. Truong, E. Bryan George, T.P. Barnwell and V. Viswanathan. "A 2.4 kbps MELP Coder Candidate for the New U.S. Federal Standard". *Proc. IEEE ICASSP*, pp. 200-203, Atlanta, Georgia (USA), May 1996.

[35] T. Tremain. "The Government Standard Linear Predictive Coding Algorithm: LPC-10". Speech Technology, pp. 40-49, April 1982.

[36] B. Atal. "Predictive Coding of Speech at Low Bit Rates". *IEEE Trans. On Communications*, Vol. 30, No. 4, pp. 600-614, April 1982.

[37] B. Fette. "High Quality Secure Voice Communication". Speech Technology, pp. 40-48, October 1989.

[38] M. Schroeder and B. Atal. "Code-Excited Linear Prediction (CELP) High Quality Speech Coding at Very Low Bit Rates". *Proc. IEEE ICASSP*, pp. 937-940, Tampa, Florida (USA), March 1985.

[39] S. Loyd. "Least Squares Quantisation in PCM". *IEEE Transactions on Information Theory*, pp. 129-136, March 1982.

[40] C.E. Shannon. "Coding Theorems for a Discrete Source with a Fidelity Criterion". IRE National Convention Record: Part 4, pp. 142-163, 1959.

[41] C.E. Shannon. "A Mathematical Theory of Communication". *Bell Systems Technical Journal*, 27:623-656, October 1968.

[42] A. Gersho and R. Gray. "Vector Quantisation and Signal Compression". Kluwer Academic Publishers, 1992.

[43] Y. Linde, A. Buzo and R. Gray. "An Algorithm for Vector Quantiser Design". *IEEE Transactions on Communications*, Vol. COM-28, No. 1, January 1980.

[44] Edward Chilton. "Factors Affecting the Quality of Linear Predictive Coding of Speech at Low Bit Rates". PhD Thesis, University of Surrey, UK, October 1990.

[45] L. Rabiner and R. Schafer. "Digital Processing of Speech Signals". Signal Processing, Prentice Hall, pp. 396-461, 1978.

[46] A. Black. "Analysis-by-Synthesis Coding of Narrowband and Wideband Speech at Medium Bit Rates". Chapter 3, PhD Thesis, University of Surrey, UK, January 1997.

[47] J. Makhoul. "Linear Prediction: A Tutorial Review". *Proc. of IEEE*, Vol. 63, No. 4, pp. 561-580, April 1975.

[48] L. Rabiner and R. Schafer. "Digital Processing of Speech Signals". Signal Processing, Prentice Hall, pp. 396-461, 1978.

[49] J. Durbin. "The Fitting of Time-Series Models". Rev. Inst. Int. Statistics, Vol. 28, No. 3, pp. 233-243, 1960.

[50] N.C. Geckinli and D. Yavuz. "Some Novel Windows and Concise Tutorial Comparison of Window Families". *IEEE Trans. On Acoustics, Speech and Signal Processing (ASSP)*, 26(6):501-507, December 1978.

[51] C. Rowden. "Speech Processing". McGraw-Hill, UK, 1992.

[52] S. Saito and K. Nakata. "Fundamentals of Speech Signal Processing". Digital Signal Processing, Academic Press, pp. 9-30, 1985.

[53] J.E. Natvig et al. "Speech Processing in the Pan-European Digital Mobile Radio System (GSM) - System Overview". *Proc. IEEE ICASSP*, pp. 1060-1064, 1989.

[54]B. Atal, R. Cox and P. Kroon. "Spectral Quantisation and Interpolation for CELP coders". *Proc. IEEE ICASSP*, pp. 69-72, 1989.

[55] D. O'Shaughnessy. "Speech Communication: Human and Machine". Digital Signal Processing, Addison-Wesley, pp. 336-379, 1987.

[56] T. Parsons. "Voice and Speech Processing". McGraw-Hill Inc., USA, 1987.

[57] K.K. Paliwal and B.S. Atal. "Efficient Vector Quantisation of LPC Parameters at 24 bits/frame Coders". *Proc. IEEE ICASSP*, pp. 661-664, 1991.

[58] I. Atkinson. "Advanced Linear Predictive Speech Compression at 3.0 kbps and below". Chapter 3, PhD Thesis, University of Surrey, UK, 1997.

[59] A.M. Kondoz. "Digital Speech: Coding for Low Bit Rate Communications Systems". John Wiley, Chichester, UK, pp. 87-88, 1994.

[60] A.M. Kondoz. "Digital Speech: Coding for Low Bit Rate Communications Systems". John Wiley, Chichester, UK, pp. 86-95, 1994.

[61] K.Y. Lee. "Analysis by Synthesis Linear Predictive Coding". PhD Thesis, University of Surrey, UK, 1990.

[62] S. Saito and K. Nakata. "Fundamentals of Speech Signal Processing". Academic Press, pp. 141-168, 1985.

[63] J. Gruber and N. Le. "Performance Requirements for Integrated Voice/Data Networks". *IEEE Journal of Selected Areas in Communications*, Vol. 1, pp. 981-1005, 1983.

[64] A.M. Kondoz. "Digital Speech: Coding for Low Bit Rate Communications Systems". John Wiley, Chichester, UK, pp. 79-116, 1994.

[65] R. Laroia, N. Phamdo and N. Farvardin. "Robust and Efficient Quantisation of Speech LSP Parameters using Structured Vector Quantisers". *Proc. IEEE ICASSP*, pp. 641-644, 1991

[66] K.K. Paliwal and B.S. Atal. "Efficient Vector Quantisation of LPC Parameters at 24 bits/frame Coders". *Proc. IEEE ICASSP*, pp. 661-664, 1989.

[67] F.K. Soong and B.-H. Juang. "Optimal Quantisation of LSP Parameters using Delayed Decision". *Proc. IEEE ICASSP*, pp. 185-188, 1990.

[68] J.-H.Chen and A. Gersho. "Real-Time Vector APC Speech Coding at 4800 bps with Adaptive Postfiltering". *Proc. IEEE ICASSP*, pp. 2185-2188, 1987.

[69] B. Gold, P. Blankenship and R. McAulay. "New Applications of Channel Vocoders". *IEEE Trans. ASSP*, Vol. 29, pp. 12-23, 1981

[70] I. Atkinson. "Advanced Linear Predictive Speech Compression at 3.0 kbps and below". Chapter 4, PhD Thesis, University of Surrey, UK, 1997.

[71] A.M. Kondoz. "Digital Speech: Coding for Low Bit Rate Communications Systems". John Wiley, Chichester, UK, pp. 60-65, 1994.

[72] D.W. Griffin and J.S. Lim. "Multi-Band Excitation Vocoder". *IEEE Trans. On ASSP*, 36(8):664-678, August 1988.

[73] A.M. Kondoz. "Digital Speech: Coding for Low Bit Rate Communications Systems". John Wiley, Chichester, UK, pp. 67-78, 1994.

[74] D. O'Shaughnessy. "Speech Communication: Human and Machine". Digital Signal Processing, pp. 232-234, Addison-Wesley, 1987.

[75] L.R. Schroeder. "Vocoders: Analysis and Synthesis of Speech". Proc. IEEE, 54:720-734, May 1966.

[76] H. Dudley. "The Vocoder". Bell Lab. Rec., 17:122-126, 1939.

[77] R.W. Schafer and L.R. Rabiner. "System for Automatic Formant Analysis of Voiced Speech". *Journal of the Acoustic Society of America*, 47:634-648, February 1970.

[78] P.E. Papamichalis. "Practical Approaches to Speech Coding". Prentice-Hall, Englewood Cliffs, N.J., pp. 87-112, 1987.

[79] A. Oppenheim. "Speech Analysis-Synthesis System Based on the Homomorphic Filters". *Journal of the Acoustic Society of America*, 45:458-465, February 1969.

[80] Atlanta Signal Processors, Inc. "MELP Vocoder Algorithm - The New 2400 bps Federal Standard Speech Coder". Revised May 1996.

[81] Digital Voice System Inc. "INMARSAT M System Definition Manual (SDM) Module 1 - INMARSAT M Voice Codec". INMARSAT, August 1991.

[82] I. Atkinson. "Advanced Linear Predictive Speech Compression at 3.0 kbps and below". Chapter 6, PhD Thesis, University of Surrey, UK, 1997.

[83] I. Atkinson. "Advanced Linear Predictive Speech Compression at 3.0 kbps and below". Chapter 5, PhD Thesis, University of Surrey, UK, 1997.

[84] S. Yeldener. "Sinusoidal Model Based Low Bit Rate Speech Coding for Communication Systems". PhD Thesis, University of Surrey, UK, 1993.

[85] M. Stefanovic. "Split-Band LPC Vocoding at Very Low Bit Rates (Performance at fixed rates, variable rates and variable frame length rates)". MPhil-to-PhD Transfer Report, University of Surrey, UK, September 1996.

[86] I. Atkinson. "Advanced Linear Predictive Speech Compression at 3.0 kbps and below". Chapter 7, PhD Thesis, University of Surrey, UK, 1997.

[87] R.J. McAulay and T.F. Quatieri. "Pitch Estimation and Voicing Detection Based on a Sinusoidal Speech Model". *Proc. IEEE ICASSP*, Vol. 2, pp. 249-252, 1990.

[88] A.M. Kondoz. "Digital Speech: Coding for Low Bit Rate Communications Systems". John Wiley, Chichester, UK, 1994.

[89] H. Nakada and K.I. Sato. "Variable Rate Speech Coding for ATM". *IEEE Trans. On Communications*, Vol. 38:3, pp. 277-284, March 1990.

[90] P. Meyer, W. Peters and J. Paulus. "Variable Rate Speech Coding Using Perceptive Thresholds and Adaptive VUS Detection". *Proc. of EUROSPEECH*, pp. 809-812, September 1991.

[91] V. Viswanathan, J. Makhoul et al. "Variable Frame Rate Transmission: A Review of Methodology and Application to Narrowband LPC Speech Coding". *IEEE Trans. On Communications*, Vol. 30, pp. 674-686, April 1982.

[92] A. Gersho and E. Paksoy. "Variable Bit Rate Coding". Signal Processing VII, Theories and Applications, pp. 1169-1173, EUSIPCO 1994.

[93] J. Cai and D.J. Goodman. "General Packet Radio Service in GSM". *IEEE Communications*, pp. 122-131, October 1997.

[94] H. Nagabuchi and N. Kitawaki. "Evaluation of Coded Speech Quality Degraded by Cell Loss in ATM Networks". *Proc. of IEICE(A)*, Vol. J74-A, No. 11, pp. 1615-1622, November 1991.

[95] T.J. Kostas, M.S. Borella et al. "Real-Time Voice Over Packet-Switched Networks". *IEEE Network*, pp. 18-27, Jan/Feb 1998.

[96] J. Wang, P.J. McCann et al. "Wireless Voice-Over-IP and Implications for Third Generation Network Design". *Bell Labs Technical Journal*, pp. 79-97, September 1998.

[97] C. Xydeas and B. Cao. "Source Driven Variable Bit Rate Prototype Interpolation Coding". *Proc. IEEE ICASSP*, pp. 220-223, 1996.

[98] S. Villette, M. Stefanovic and A. Kondoz. "A Multi-Rate Speech and Channel Codec: A GSM AMR Half-Rate Candidate". EUROSPEECH '99 Conference (Accepted Publication).

[99] R. Drogo de Iacovo and D. Sereno. "Embedded CELP Coding for Variable Bit Rate between 6.4 and 9.6 kbps". *Proc. IEEE ICASSP*, pp. 681-683, 1991.

[100] D.J. Goodman. "Embedded DPCM for Variable Bit Rate Transmission". *Proc. of ICC-80*, paper 42.2.

[101] L. Rabiner and B.-H. Juang. "Fundamentals of Speech Recognition". Prentice Hall Signal Processing Series, pp. 11-41, 1993.

[102] H. Nakada and K.I. Sato. "Variable Rate Speech Coding for ATM". *IEEE Trans. On Communications*, Vol. 38:3, pp. 277-284, March 1990.

[103] H. Nakada and K.-I. Sato. "Variable Rate Speech Coding for Asynchronous Transfer Mode". *IEEE Trans. On Communications*, Vol. 38, No. 3, march 1990.

[104] S. Wang and A. Gersho. "Phonetically-Based Vector Excitation Coding of Speech at 3.6 kbps". *Proc. IEEE ICASSP*, pp. 49-52, 1989.

[105] L. Rabiner and B.-H. Juang. "Fundamentals of Speech Recognition". Prentice Hall Signal Processing Series, pp. 42-64, 1993.

[106] L. Rabiner and B.-H. Juang. "Fundamentals of Speech Recognition". Prentice Hall Signal Processing Series, 1993.

[107] J.S. Garofolo, L.F. Lamel et al. "DARPA TIMIT: Acoustic-Phonetic Continuous Speech Corpus". National Institute of Standards and Technology, 4930, February 1993.

[108] C. Sriratanaban. "Speech Recognition Based Phonetic Segmentation for Low Bit Rate Speech Coder". MPhil-to-PhD Report, University of Surrey, UK, November 1996.

[109] Y. Asakawa, P. Rao and H. Sekine. "8 kbps Low-Delay Speech Coding with 4-ms Frame". *IEICE Trans. Fundamentals*, Vol. E78-A, No. 8, August 1995.

[110] Y. Shoham. "Coding the Line Spectral Frequencies by Jointly Optimised MA Prediction and Vector Quantisation". *IEEE Workshop on Speech Coding*, Porvoo, Finland, June 1999.

[111] ITU-T Rec. G.729. "Coding of Speech at 8 kbps using Conjugate Structure Algebraic Code Excited Linear Prediction (CS-ACELP)".

[112] M. Stefanovic and A. Kondoz. "Source-Dependent Variable Rate Speech Coding Below 3 kbps". EUROSPEECH '99 Conference, (accepted for publication), September 1999.

[113] A. Black. "Analysis-by-Synthesis Coding of Narrowband and Wideband Speech at Medium Bit Rates". Chapter 7, PhD Thesis, University of Surrey, UK, January 1997.

[114] D. O'Shaughnessey. "Speech Communication Human and Machine", Addison-Wesley, pp. 164-203, 1987.

[115] V. Abea-Sernández and D. Docampo-Amoedo. "A Multipulse-Deconvolution Codec for Wideband Speech". *Proc. EUROSPEECH*, pp. 49-52, Madrid, September 1995.

[116] CCITT. "7 kHz Audio Coding within 64 kbps". Rec. G.722, Vol. Fascicle III.4 of Blue Book, pp. 269-341, Melbourne, 1988.

[117] R. Drogo de Jacovo, R. Montagna et al. "Some Experiments of 7 kHz Audio Coding at 16 kbps". *Proc. IEEE ICASSP*, pp. 192-195, 1989.

[118] E. Ordentlich. "Low-Delay Code-Excited Linear Predictive Coding of Wideband Speech at 32 kbps". *Proc. IEEE ICASSP*, pp. 9-12, 1991.

[119] A. Amodio and G. Feng. "Wideband Speech Coding Based on the MBE Structure". *Proc. EUROSPEECH*, pp. 1499-1502, Rhodes, 1997.

[120] B. Bessette, R. Salami et al. "A Wideband Speech and Audio Codec at 16/24/32 kbps using Hybrid ACELP/TCX Techniques". *IEEE Workshop on Speech Coding*, Porvoo, Finland, June 1999.

[121] J.W. Paulus and J. Schnitzler. "16 kbps Wideband Speech Coding Based on Unequal Subbands". *Proc. IEEE ICASSP*, pp. 255-258, Atlanta 1996.

[122] C. McElroy, B.Murray and A. Fagen. "Wideband Speech Coding in 7.2 kbps". *Proc. IEEE ICASSP*, pp. 620-623, Minneapolis 1993.

[123] A. Amodio and G. Feng. "Wideband Speech Coding Based on the MBE Structure". *Proc. EUROSPEECH*, pp. 1499-1502, Rhodes, 1997.

[124] C. Garcia-Mateo, J.L. Alba-Castro and E.R. Banga. "Speech Coding Using Bi-harmonic spectral modelling". Signal Processing VII, pp. 391-394, 1994.

[125] M. Stefanovic and A. Kondoz. "Adaptive Passband Estimation for Television Audio". CCSR Report, University of Surrey, UK, DICTION D3.3, June 1999.

[126] A. Amodio and G. Feng. "Wideband Speech Coding Based on the MBE Structure". *Proc. EUROSPEECH*, pp. 1499-1502, Rhodes, 1997.

[127] A. Amodio and G. Feng. "Wideband Speech Coding Based on the MBE Structure". *Proc. EUROSPEECH*, pp. 1499-1502, Rhodes, 1997.

[128] Federal Information Processing Standards Publication. "Specifications for the Analogue to Digital Conversion of Voice by 2400 bps Mixed Excitation Linear Prediction". Draft, January 1998.

[129] R. Lefebvre, R. Salami et al. "High Quality Coding of Wideband Audio Signals using Transform Coded Excitation (TCX)". Proc. IEEE ICASSP, pp. 193-196, Adelaide, South Australia, 1994.

[130] W.B. Kleijn. "Continuous Representation in Linear Predictive Coding". *Proc. IEEE ICASSP*, Vol. 1, pp. 201-204, 1991.

[131] W.T.K. Wong, R.M. Mack et al. "Low Rate Speech Coding for Telecommunications". *BT Technology Journal*, Vol. 14, No. 1, pp. 28-43, January 1996.

[132] W.B. Kleijn and J. Haagen. "A Speech Coder based on Decomposition of characteristic waveforms". *Proc. IEEE ICASSP*, pp. 508-511, 1995.

[133] H. Tasaki and S. Takahashi. "Post Noise Smoother to Improve Low Bit Rate Speech Coding Performance". *IEEE Workshop on Speech Coding*, Porvoo, Finland, June 1999.

[134] R. Martin and R. Cox. "New Speech Enhancement Techniques for Low Bit Rate Speech Coding". *IEEE Workshop on Speech Coding*, Porvoo, Finland, June 1999.

UNIVERSITY OF SURREY LIBRARY