Multimode Speech Coding below 6 kbps

Nilantha N. Katugampala

Submitted for the Degree of
Doctor of Philosophy
from the
University of Surrey

University of Surrey
Centre for Communication Systems Research
School of Electronic Engineering, Information Technology and Mathematics
University of Surrey
Guildford, Surrey GU2 7XH, U.K.

October 2001

© Nilantha N. Katugampala 2001
Summary

The past two decades have witnessed a rapid expansion of the telecommunications industry. This growth has been primarily fuelled by the proliferation of the digital communication systems and services which have become easily available through wired and wireless networks. 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. The available bandwidth in wire based terrestrial network is a relatively cheap and expandable resource. However in satellite and cellular radio systems the bandwidth is inherently limited and an expensive resource. In order to accommodate ever growing numbers of subscribers whilst maintaining high quality and low operational costs, it is essential to maximise the spectral efficiency. The research presented in this thesis has focused on the development of new source compression algorithms, tailored for human speech in order to improve the spectral efficiency of digital transmission systems.

Recently there is an increasing interest on speech coding algorithms which combine various existing technologies in order to improve the speech quality whilst maintaining the low transmission rate of the existing coding techniques. The aim of the research presented in this thesis was to develop a complete hybrid coding algorithm which combines harmonic and waveform approximating coding techniques. In order to integrate the two coding paradigms novel phase synchronisation and classification techniques were developed. The perceptual quality of the speech synthesised using the unquantised hybrid model achieves nearly transparent quality. The hybrid model was used to develop variable bit rate coders, which are particularly advantageous for voice storage, Code Division Multiple Access (CDMA) wireless networks, packet switched networks, and statistical multiplexing of speech for multi channel communications.

Key words: Hybrid coding, Harmonic excitation, ACELP

Email: n.katugampala@eim.surrey.ac.uk
WWW: http://www.eim.surrey.ac.uk/
Acknowledgements

I would like to take this opportunity to express my sincere gratitude to Prof. Ahmet Kondoz, my PhD project supervisor, whose support and guidance throughout my research was most helpful and much appreciated. I am also indebted to my colleagues in the Multimedia Communications Research Group, especially Stéphane Villette, Yong Duk Cho, and Stewart Worrall whose support and friendship have been invaluable. I would also like to thank the CCSR support staff, especially Lakshmi Chennell for her help.

Finally I would like to thank my parents for their constant support and encouragement throughout the course of this research.
## Contents

1 Introduction ................................................................. 1
   1.1 Background ................................................. 1
   1.2 Outline of thesis ........................................... 2
   1.3 Original achievements ...................................... 4

2 A review of speech coding ........................................... 6
   2.1 Introduction ................................................ 6
   2.2 Speech coding techniques ............................... 7
      2.2.1 Parametric coders .................................... 7
      2.2.2 Waveform approximating coders .............. 9
      2.2.3 Hybrid coding of speech ....................... 10
   2.3 Design criteria ............................................ 11
      2.3.1 Bit rate ............................................... 11
      2.3.2 Quality ................................................ 11
      2.3.3 Delay .................................................. 12
      2.3.4 Complexity ........................................... 13
      2.3.5 Channel error sensitivity ...................... 13
      2.3.6 Diverse input signal requirements ........... 14
   2.4 Applications of speech coding ......................... 14
      2.4.1 Terrestrial voice communication systems ...... 14
      2.4.2 Satellite communication systems .............. 15
   2.5 Standardisation ............................................. 16
   2.6 Concluding remarks ....................................... 18
## 3 Fundamental speech coding techniques

3.1 Introduction ................................................................. 20

3.2 Quantisation ................................................................. 21
   3.2.1 Scalar quantisation ..................................................... 21
   3.2.2 Vector quantisation ..................................................... 22

3.3 Vocal tract model ............................................................ 22
   3.3.1 Source filter model ..................................................... 24
   3.3.2 Linear Predictive Coding (LPC) ..................................... 24

3.4 Quantisation and interpolation of LPC coefficients ................. 28
   3.4.1 Line Spectral Frequencies (LSFs) ................................. 29

3.5 Pitch determination .......................................................... 35
   3.5.1 Time domain approaches ............................................. 35
   3.5.2 Frequency domain approaches ..................................... 37

3.6 Concluding remarks ......................................................... 41

## 4 Harmonic coders

4.1 Introduction ................................................................. 42

4.2 Sinusoidal analysis and synthesis ........................................ 43

4.3 Parameter estimation ....................................................... 44
   4.3.1 Voicing determination ................................................. 45
   4.3.2 Harmonic amplitude estimation ..................................... 48

4.4 Common harmonic coders .................................................. 50
   4.4.1 Sinusoidal Transform Coding (STC) .............................. 50
   4.4.2 Improved Multi Band Excitation (IMBE) ......................... 52
   4.4.3 Split Band Linear Predictive Coding (SB-LPC) ................ 54

4.5 Concluding remarks ......................................................... 56

## 5 Analysis by synthesis coding of speech

5.1 Introduction ................................................................. 57

5.2 Long Term Prediction (LTP) .............................................. 60

5.3 Innovation structures ....................................................... 64

5.4 Perceptual weighting ....................................................... 67

5.5 Algebraic Code Excited Linear Prediction (ACELP) ............... 69
   5.5.1 Conjugate Structure ACELP (CS-ACELP) ITU G.729 coder . 71

5.6 Concluding remarks ......................................................... 74
6 Integration of harmonic and analysis by synthesis coders

6.1 Introduction ................................................................. 75

6.2 Challenges in designing a hybrid coder ................................... 78
  6.2.1 Reliable speech classification ........................................ 78
  6.2.2 Phase synchronisation ............................................... 79

6.3 The existing hybrid coders ................................................ 79
  6.3.1 Limitations of the existing hybrid coders ..................... 82

6.4 Synchronised Waveform matched Phase Model (SWPM) .......... 84
  6.4.1 Extracting the Pitch Pulse Location (PPL) ..................... 85
  6.4.2 Estimating the Pitch Pulse Shape (PPS) ......................... 92
  6.4.3 Synthesis using the generalised cubic phase interpolation ... 98

6.5 Hybrid encoder ............................................................. 100
  6.5.1 Synchronised harmonic excitation ............................... 102
  6.5.2 Pros and cons of SWPM ........................................... 103
  6.5.3 Offset target modification ........................................ 108
  6.5.4 Onset harmonic memory initialisation ......................... 112
  6.5.5 White noise excitation ............................................. 114

6.6 Speech classification ...................................................... 116
  6.6.1 Open loop initial classification ................................... 117
  6.6.2 Closed loop transition detection ................................. 120
  6.6.3 Plosive detection ................................................... 124

6.7 Hybrid decoder ............................................................ 126

6.8 Performance evaluation .................................................. 127

6.9 Concluding remarks ..................................................... 130

7 Variable bit rate hybrid coding ........................................ 132

7.1 Introduction ............................................................... 132

7.2 Unvoiced gain quantisation ............................................ 133

7.3 Harmonic parameter quantisation ..................................... 134
  7.3.1 Harmonic amplitude quantisation ............................... 139
  7.3.2 Onset harmonic parameter quantisation ....................... 143

7.4 Quantising the transitions ............................................. 145
Chapter 1

Introduction

1.1 Background

Speech is the simplest and the most convenient means of communication between human beings. As a result, the telephone has become the primary means of communication around the globe. Since its invention in 1876 the telephone has been subjected to many improvements, and one of the most significant advances made during the last century is the digitisation of the telecommunication services. The digital transmission of speech offers many advantages over the analogue transmission, which include perfect signal reconstruction, forward error correction, encryption, and multiplexing. Moreover the digital hardware is cheap, reliable, and can replicate the performance despite the existence of component tolerances, electrical noise, and thermal stability issues.

The major drawback of the digital transmission of speech is that the required bandwidth is greater than that required by the original analogue signal. The data rate of the original PCM is 128 kbps, with 16 bits per sample and 8 kHz sampling rate. Companded PCM model the sensitivity of the human ear to quantisation noise using a logarithmic scale in the quantiser design and halved the original data rate down to 64 kbps. ADPCM exploit the redundancies in the speech signal and utilise adaptive quantisation techniques to further reduce the data rate to 32 kbps. While these high bit rates are acceptable for trunk telephone lines, and especially for fibre optic channels
which have large bandwidths, the satellite and terrestrial radio channels demand more compression due to the limited radio bandwidth.

The increasing interest in mobile communications coupled with the rapid advances in the VLSI technology since the mid 1980's lead to extensive research into the area of low to medium bit rate speech compression. As a result, many speech coding algorithms have been developed which offer synthetic to toll quality speech, operating around the bit rates 2~16 kbps. An interesting feature of those coders is that the majority of the coders operating below 4 kbps are based on parametric modelling of speech, while the majority of the coders operating above 4 kbps are based on analysis by synthesis coding techniques.

The work presented in this thesis focuses on integrating the parametric coding and analysis by synthesis coding in order to circumvent the limitations of each technique and utilise them to maximise the synthesised speech quality. A novel hybrid coding algorithm is presented, which offers the flexibility of choosing the best coding method for a given speech segment, optimising their performance for the intended speech characteristics, and producing various constant or variable bit rates by utilising different configurations for each coding method. The target applications include, transparent quality for PSTN at 8 kbps, toll quality for mobile telephony at 4 kbps, and variable bit rate applications such as answer phones, CDMA wireless networks, and ATM networks.

1.2 Outline of thesis

The rest of the thesis consists of three main parts: an introduction to digital speech coding, popular speech coding techniques, and integration of harmonic and analysis by synthesis coders.

An introduction to digital speech coding

Chapter 2: A review of speech coding

A brief overview of digital speech coding is provided. The basic speech coding paradigms: parametric and waveform approximating coders are introduced. The issues related to
1.2. Outline of thesis

the design of a speech coder such as bit rate, quality, delay, and complexity are discussed, followed by a description of applications of speech coding. Finally a summary of the existing speech coding standards is provided.

Chapter 3: Fundamental speech coding techniques

The common speech coding techniques, such as vector quantisation and modelling of the spectral envelope using Linear Predictive Coding (LPC) are presented. The conversion of the LPC coefficients to Line Spectral Frequencies (LSF) for efficient quantisation and interpolation and the commonly used pitch determination algorithms are also described.

Popular speech coding techniques

Chapter 4: Harmonic coders

The basic sinusoidal analysis and synthesis model is introduced, followed by a description of adapting the basic model for low bit rate applications. Three common harmonic coders are described in some detail: Sinusoidal Transform Coding (STC), Improved Multi Band Excitation (IMBE), and Split Band Linear Predictive Coding (SB-LPC).

Chapter 5: Analysis by synthesis coding of speech

The basic principles behind the Analysis by Synthesis (AbS) coding concept are introduced. Excitation signal generation by summing the Long Term Prediction (LTP) contribution and the innovation contribution is explained. The use of perceptual weighting, in order to shape the quantisation noise spectrum is also described. Finally an example AbS scheme, Algebraic Code Excited Linear Prediction (ACELP) and the ITU 8 kbps G.729 standard, which is based on ACELP are introduced.

Integration of harmonic and analysis by synthesis coders

Chapter 6: Integration of harmonic and analysis by synthesis coders

This chapter describes a novel hybrid coding algorithm in detail. The chapter begins with a description of the advantageous and disadvantageous of the harmonic and ACELP coding techniques, and the potential of a combined hybrid coding algorithm
1.3. Original achievements

to overcome the limitations of each other. A description of the existing hybrid coders is provided identifying their limitations. The hybrid coding algorithm designed during the course of this research, employs a new phase model for the harmonic excitation Synchronised Waveform matched Phase Model (SWPM), which minimises the phase discontinuities when switched between the harmonic and ACELP modes. A two stage speech classification algorithm is also designed, which employs a novel analysis by synthesis transition detection technique. The chapter concludes by presenting the performance of the speech classification algorithm and the results of the subjective listening tests.

Chapter 7: Variable bit rate hybrid coding

This chapter presents the quantisation of the model parameters and the development of variable rate hybrid coders based on the hybrid coding algorithm explained in chapter 6. The advantageous and applications of variable rate coders are also described. The performance of the variable rate hybrid coders are then compared with the standard coders, ITU G.723.1 and ITU G.729, with the standard coders operating at least at the maximum bit rate of the variable rate coders.

Chapter 8: Acoustic noise and channel error performance

This chapter investigates the robustness of the hybrid coding algorithm against acoustic noise and errors in the mode bits due to random channel errors. The measures taken to improve the robustness and suggestions for further improvements are also presented.

1.3 Original achievements

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

- A comprehensive review of harmonic and analysis by synthesis coding techniques. This included identifying their pros and cons and understanding the potential benefits of a combined hybrid coding scheme.

- A comprehensive review of the existing hybrid coders along with their strengths and shortcomings.
1.3. Original achievements

- Design of a novel hybrid coding algorithm based on a new phase model for the LP based harmonic excitation signals, which synchronises the synthesised speech signal with the original signal.

- Design of a novel two stage speech classification algorithm based on a new analysis by synthesis transition detection technique.

- Development of variable rate hybrid coders based on the designed hybrid coding model.

- Investigating and improving the robustness of the hybrid coding algorithm against the background noise and mode bit errors.

Some of these achievements have been published in various international and European conference proceedings. These are listed in Appendix A.
Chapter 2

A review of speech coding

2.1 Introduction

The invention of Pulse Code Modulation (PCM) in 1938 by Alec H. Reeves was the beginning of digital speech communications. Unlike the analogue systems, PCM systems allow perfect signal reconstruction at the repeaters of the communication systems, which compensate for the attenuation, provided that the channel noise level is insufficient to corrupt the transmitted bit stream. In the early 1960's as the digital system components became widely available, PCM was implemented in private and public switched telephone networks. Today, all most 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. The additional advantages of PCM over the analogue transmission include, the availability of sophisticated hardware for error correction, encryption, multiplexing, switching, and compression.

The main disadvantage of PCM is that the required transmission bandwidth is greater than that required by the original analogue signal. This is not desirable when using expensive and bandwidth restricted channels such as satellite and cellular mobile radio systems. This resulted in extensive research into the area of speech coding during the last two decades. The recent advances of the VLSI technology has allowed to implement highly complex speech compression algorithms in real time applications, further fuelling the need of sophisticated speech coding schemes.
2.2 Speech coding techniques

Traditionally, speech coders have been separated into two classes: waveform approximating coders and parametric coders. Kleijn [1] defines them as follows:

**Waveform approximating coders:** Coders which produce a reconstructed signal which converges towards the original signal with decreasing quantisation error.

**Parametric coders:** Coders which produce a reconstructed signal which does not converge to the original signal with decreasing quantisation error.

Typical performance curves for waveform approximating coders and parametric coders are shown in Figure 2.1. Some authors [2], [3] classify the time domain analysis by synthesis coders as a third technique, and use the term hybrid coders. Analysis by synthesis coders use linear predictive and pitch predictive coding to model the short term and long term correlation present in the human speech respectively. The prediction residual is encoded using waveform coding. Hence the term hybrid coding is used for analysis by synthesis coders. However in this thesis the analysis by synthesis coders are classified as an improved waveform coding technique, since the synthesised speech waveform converges towards the original signal, as the quantisation error of the prediction residual is decreased. The term hybrid coding is used in a different context as described below and in detail in chapter 6.

2.2.1 Parametric coders

Parametric coders model the speech signal using a set of model parameters. The extracted parameters at the encoder are quantised and transmitted to the decoder. The decoder synthesises speech according to the specified model. The speech production model does not account for the quantisation noise nor it tries to preserve the waveform similarity between the synthesised speech and the original speech signal. The model parameter estimation is an open loop process and has no feedback from the quantisation or the speech synthesis. These coders only preserve the features included in the speech production model, e.g. spectral envelope, pitch and energy contour. The speech quality of parametric coders do not converge towards the transparent quality of the
2.2. Speech coding techniques

![Quality Vs. Bit rate for different speech coding techniques](image)

Figure 2.1: Quality Vs. Bit rate for different speech coding techniques

original speech with better quantisation of model parameters and tend to saturate at a lower quality, due to the limitations of the speech production model, see Figure 2.1. Furthermore, they do not preserve the waveform similarity and the measurement of SNR is meaningless, often the SNR is negative, when expressed in dB. The SNR has no correlation with the synthesised speech quality and the quality should be assessed subjectively.

**Linear prediction based vocoders**

The Linear Prediction (LP) based vocoders are modelled after the human speech production mechanism [4]. The vocal tract is modelled by a linear prediction filter. The glottal pulses and turbulent air flow at the glottis are modelled by periodic pulses and Gaussian noise respectively, which form the excitation signal of the linear prediction filter. The vocal tract model is described in detail in chapter 3. The LP filter coefficients, signal power, binary voicing decision i.e. periodic pulses or noise excitation, and the pitch period of the voiced segments are estimated and transmitted to the decoder. The main weakness of the LP based vocoders is the binary voicing decision of the excitation, which fails to model the mixed type signals with both periodic and noisy components. By employing frequency domain voicing decision techniques the performance of the LP
2.2. Speech coding techniques

based vocoders can be improved [5].

**Harmonic coders**

Harmonic or sinusoidal coding represents the speech signal as a sum of sinusoidal components. The model parameters i.e. the amplitudes, frequencies and phases are estimated at regular intervals from the speech spectrum. The frequency tracks are extracted from the peaks of the speech spectra, and the amplitudes and frequencies are interpolated in the synthesis process [6]. The general sinusoidal model does not restrict the frequency tracks to be harmonics of the fundamental frequency. Increasing the parameter extraction rate converges the synthesised speech waveform towards the original, if the parameters are unquantised. However at low bit rates the phases are not transmitted and modelled at the decoder, and the frequency tracks are confined to be harmonics. Therefore the waveform similarity is not preserved. See chapter 4 for more details on harmonic coding.

**2.2.2 Waveform approximating coders**

Waveform coders minimise the error between the synthesised and the original speech waveforms. The early waveform coders such as companded Pulse Code Modulation (PCM) [7] and Adaptive Differential Pulse Code Modulation (ADPCM) [8] transmit a quantised value for each speech sample. However ADPCM employs an adaptive pole zero predictor and quantises the error signal, with an adaptive quantiser step size. ADPCM predictor coefficients and the quantiser step size are backward adaptive and updated at the sampling rate.

**Analysis by synthesis coders**

The recent waveform approximating coders based on time domain analysis by synthesis such as Code Excited Linear Prediction (CELP) [9], explicitly make use of the vocal tract model and the long term prediction to remove the correlations present in the speech signal. CELP coders buffer the speech signal and perform block based analysis
and transmit the prediction filter coefficients along with an index for the prediction residual vector. They also employ perceptual weighting so that the quantisation noise spectrum is masked below the signal level. See chapter 5 for more details on analysis by synthesis coding.

2.2.3 Hybrid coding of speech

Most of the existing speech coders apply the same coding technique, regardless of the widely varying character of the speech signal. Examples include, Adaptive Differential Pulse Code Modulation (ADPCM) [8], Code Excited Linear Prediction (CELP) [9], [10], and Improved Multi Band Excitation (IMBE) [11], [12]. When the bit rate is reduced, the perceived quality of these coders tend to degrade for some speech segments while remaining adequate for the others. In order to circumvent this problem hybrid coders, which combine different coding techniques to encode different types of speech segments were introduced [13], [14], [15].

A hybrid coder can switch between a set of predefined coding modes. Hence they are also referred to as multimode coders. A hybrid coder is an adaptive coder, which can change the coding technique or mode according to the source, selecting the best suited mode for the local character of the speech signal. Network or channel dependent mode decision is also possible [16]. Those coders can adapt to the network load or the channel error performance, by varying the modes and the bit rate, and changing the relative bit allocation of the source and channel coding [17].

In source dependant mode decision, the speech classification can be based on fixed length frames or variable length frames. The number of bits allocated for frames of different modes can be the same or different. The overall bit rate of a hybrid coder can be fixed or variable. In fact variable rate coding can be seen as an extension of hybrid coding.
2.3 Design criteria

The design and coding capacity of a particular coding algorithm is often determined by the target application. However, many of the issues involved in the design process have conflicting requirements. Therefore designing a speech coder involves various trade offs between those conflicting requirements.

2.3.1 Bit rate

Reducing the transmission bit rate is the primary motivation of speech coding. Depending on the application fixed rate or Variable Bit Rate (VBR) coding can be used. VBR coders need to define criteria, which determine the bit rate for a particular segment of speech. A VBR coder allocates the minimum number of bits required to maintain sufficient speech quality for a given segment of speech. However, they are more complex in design compared to the fixed rate coders. Most of the existing speech coding standards are based on fixed rate algorithms. Variable bit rate coders are particularly advantageous for voice storage, Code Division Multiple Access (CDMA) [18] wireless networks, packet switched networks, and statistical multiplexing of speech for multi channel communications. It is likely that in the future a large fraction of speech coders will be based on variable bit rate algorithms.

2.3.2 Quality

In general lowering the bit rate results in a reduction in the quality of coded speech. Quality measurements based on SNR can be used to evaluate the coders, which preserve the waveform similarity, usually the coders operating at bit rates above 8 kbps. Low bit rate parametric coders do not preserve the waveform similarity and the SNR based quality measures become meaningless. For the parametric coders perception based subjective measures are more reliable. The Mean Opinion Score (MOS) [19] scale shown in Table 2.1 is a widely used subjective quality measure.
2.3. Design criteria

Table 2.1: Mean Opinion Score (MOS) scale

<table>
<thead>
<tr>
<th>Grade</th>
<th>Subjective Opinion</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Imperceptible</td>
<td>Transparent</td>
</tr>
<tr>
<td>4</td>
<td>Perceptible, but not annoying</td>
<td>Toll</td>
</tr>
<tr>
<td>3</td>
<td>Slightly annoying</td>
<td>Communication</td>
</tr>
<tr>
<td>2</td>
<td>Annoying</td>
<td>Synthetic</td>
</tr>
<tr>
<td>1</td>
<td>Very annoying</td>
<td>-</td>
</tr>
</tbody>
</table>

2.3.3 Delay

The delay in a speech communications system consists of three components, algorithmic delay, processing delay, and any extra delay added by the communications network. The speech encoders buffer the input samples to obtain the analysis frames. The buffering process introduces a delay equivalent to the frame length. A look ahead is required for operations such as parameter interpolation and pitch tracking. The sum of the look ahead and the buffering delay is called algorithmic delay. The processing delay may be reduced by using a low complexity algorithm or by using a faster processor. The delay introduced by the communications network include the signal propagation delay and the multiplexing delay.

The end to end delay is an important factor for transmission applications and real time interactive applications, such as telephony and audio conferencing. Large communication delays of 400 ms or greater make full duplex conversation impossible. In addition, delays of 50 ms or more increase the problem of disturbing echoes generated at the hybrid interface between two wire and four wire lines in the terrestrial network. This introduces the need for echo cancellation [20], either within the network or in the terminal equipment. This has an impact on the communication system cost and complexity.
2.3. Design criteria

2.3.4 Complexity

The computational complexity and memory requirements of a coding algorithm determine the cost and the power consumption of the Digital Signal Processor (DSP). Lower bit rates and higher quality can be achieved by increasing the complexity of an algorithm. However, in real time applications the processing delay of a frame should be less than the frame length, in order to process the next frame. Furthermore, minimising the power consumption and cost is essential for the hand held mobile equipments. DSPs become faster, cheaper and low power due to the advances in the VLSI technology, making more sophisticated coding algorithms realisable in real time.

Speed is commonly measured as the number of Millions of Instructions Per Second (MIPS) necessary for the real time implementation of the speech coding algorithm. 16 bit fixed point DSPs are commonly used for low cost implementations. Thus, complexity is often specified in terms of fixed point MIPS and the number of 16 bit words of RAM needed for an implementation.

2.3.5 Channel error sensitivity

In many applications, the bit stream received at the far end of the channel is corrupted by channel errors. This may cause annoying artifacts in the reconstructed speech. Therefore the robustness against the channel errors is an important factor. The types of channel errors that speech coders are supposed to handle are usually divided into two classes: random errors and burst errors. The latter class of errors is typical for a mobile telephone environment. These two classes require different strategies to reduce their impact on the reconstructed speech.

To counter random channel errors, the coder should provide reasonable output for a frame, even if a small proportion of the received information within that frame is incorrect. Robustness against such channel errors can be increased by means of index assignment algorithms [21], [22], through proper quantiser design, and by adding redundancy into the transmitted information [23], [24], [25]. It is best to integrate channel coding in the design of the quantisers of a speech coder. However, in many existing communication systems, the speech coder is separate from the channel coder.
In the case of burst errors, error detection schemes are used to classify each frame of received bits as usable or unusable. If a received frame is detected as unusable, the decoder enters into a special mode, which often means that the signal power is gradually reduced, and the synthesised signal is made to converge slowly to a white noise signal.

2.3.6 Diverse input signal requirements

Low bit rate speech coding algorithms are tailored for compression of human speech. However, the ability to carry non speech signals for speech communication channels may be essential. For example, Public Switched Telephone Network (PSTN) is used to send signalling tones based on Dual Tone Multi Frequency (DTMF) system for the transmission of telephone digits and modem tones for transmitting voice band data.

Faithful reproduction of background noise is also essential to maintain the naturalness of the speech during conversation. The speech analysis and parameter estimation process should be robust in the presence of acoustic background noise.

2.4 Applications of speech coding

Transmission of voice is the major application of speech coding. The voice transmission systems can be divided into two broad categories: terrestrial and satellite. Voice storage applications also employ speech compression schemes. Many of the new applications of speech coding, such as packet switched cellular telephony and answering machines, do not require a fixed bit rate. As a result, significant effort has been dedicated to development of variable bit rate speech coders in the recent years. See chapter 7 for more details on variable bit rate speech coding.

2.4.1 Terrestrial voice communication systems

The terrestrial voice communication systems include Public Switched Telephone Networks (PSTN), Integrated Services Digital Networks (ISDN), and cellular mobile radio systems. The first generation PSTN used 64 kbps companded PCM. With the rise of the
subscriber numbers the second generation PSTN employed more bandwidth efficient 32 kbps ADPCM. PSTNs with very high subscriber demands employ the toll quality Low Delay Code Excited Linear Prediction (LD-CELP) at 16 kbps. In the recent years the speech coding research for PSTN has been focused on designing a 8 kbps coder. The relatively low Bit Error Rate (BER of $10^{-6}$ to $10^{-5}$) of PSTN makes it attractive to employ a hybrid coding algorithm for this purpose, since the most sensitive mode bits can be received with a very high accuracy. Hybrid coding has the potential to achieve transparent speech quality at 8 kbps. See chapter 6 for more details on hybrid coding.

ISDN integrates transmission of speech and data. The wide band speech coder, 64 kbps Sub Band ADPCM (SB-ADPCM) [26] is used for ISDN applications. Cellular telephony can be seen as the most important application of speech coding. The limited radio bandwidth and the increasing number of mobile phone subscribers make speech coding essential for mobile communication applications. The most prominent, Pan-European Digital Cellular Mobile Radio System employs 13 kbps GSM Full Rate (GSM FR) [27], 12.2 kbps GSM Enhanced Full Rate (GSM EFR) [28], and 5.6 kbps GSM Half Rate (GSM HR) [29] speech coders. The bursty channel errors, which can be as much as 3% is the most serious problem in mobile communications. Forward Error Correction (FEC) techniques are essential to counter bursty channel errors. The gross bit rate of the GSM FR and the GSM EFR coders is 22.8 kbps and the gross bit rate of the GSM HR coder is 11.4 kbps, with channel coding.

2.4.2 Satellite communication systems

The use of satellite systems is primarily for long distance communications, due to the wide coverage area and point to point and point to multi point connection capability. There are three main types of satellite services are defined by the International Telecommunications Union (ITU): 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 provides services such as maritime, aeronautical or land to both fixed and mobile earth terminals. BSS broadcasts television and radio programmes to smaller earth stations in the
domestic premises. The key problem of the satellite communications is the inherent delay due to signal propagation. Satellite services are also faced with the bursty channel errors due to fading from multi path effects and shadowing, common in the terrestrial mobile channels.

The Inmarsat-M Improved Multi Band Excitation (IMBE) [12] coder used for mobile telephony via satellite is an example of a speech coder used in MSS. The 4.15 kbps IMBE coder is channel coded for a gross channel bit rate of 6.4 kbps. FSS and BSS require broadcast quality audio with data rates of above 32 kbps.

2.5 Standardisation

As a result of the research activities in speech coding since 1960's, especially during the last two decades, a variety of speech coding algorithms have been developed. In the early days, many commercial equipment providers have developed their own in house speech coding algorithms for use in their particular products used in private network applications. This isolated development process resulted in incompatibility between the products from different manufactures. With the introduction of speech coding techniques in publicly available telecommunication services, it was necessary to standardise speech coding algorithms in order that equipment manufacturers could coordinate their research activities and resolve equipment compatibility issues. This has enabled commercial competition between manufactures, lowering product prices and resulting in a rapid increase of mobile product usage by the general public, further fuelling the demand for efficient speech coding algorithms.

The standardisation procedure which identifies and defines the speech coding requirements for the next generation telecommunication products has become the main driving force in speech coding research. Table 2.2 shows some of the well known telephone band speech coding standards. The International Telecommunications Union (ITU) 4 kbps standardisation is yet to be finalised, and the figures given are the requirements set by the ITU. Figure 2.2 illustrates the performance of those standards in terms of the quality against the bit rate [30], [31].
2.5. Standardisation

Figure 2.2: Performance of telephone band speech coding standards

Linear PCM at 128 kbps offers transparent speech quality. The ITU standard G.711 companded PCM has two versions, μ law and A law PCM [7]. The ITU standards, G.726, G.728, and G.729 [32] are based on Variable Bit Rate Adaptive Differential PCM (VBR-ADPCM) [8], Low Delay Code Excited Linear Prediction (LD-CELP) [33], and Conjugate Structure Algebraic CELP (CS-ACELP) [34] respectively. The ITU standard G.723.1 has two rates 5.3 kbps uses Algebraic CELP (ACELP) [35] and 6.3 kbps uses Multi Pulse-Maximum Likelihood Quantisation CELP (MP-MLQ CELP) [36].

European Telecommunications Standards Institute (ETSI) standards GSM Full Rate (GSM FR), GSM Enhanced Full Rate (GSM EFR) [28], and GSM half rate coders are based on Regular Pulse Excitation with Long Term Prediction (RPE-LTP) [27], [37], ACELP, and Vector Sum Excited Linear Prediction (VSELP) [29], [38] respectively.

Telecommunications Industry Association (TIA) standards IS54 and IS96, are based on VSELP and Qualcomm CELP (Q-CELP) [39] respectively. Japanese Digital Cellular
(JDC) coder, which uses VSELP, and JDC half rate coder, which uses Pitch Synchronous Innovation CELP (PSI-CELP) [40] were standardised by the Research and development Center for Radio systems (RCR). Inmarsat-M [11] coder, which is based on Improved Multi Band Excitation (IMBE) [12] was standardised by the International Maritime Satellite Corporation (Inmarsat), and implemented by Digital Voice Systems Incorporated (DVSI).

The secure voice standards FS1015, FS1016 [41], and new FS 2.4 are based on LPC-10 vocoder [42], CELP, and Mixed Excitation Linear Prediction (MELP) [43] respectively. The vast majority of the speech coding standards emerged during the last two decades are based on the variants of CELP, which includes all the terrestrial digital mobile communication systems. It is worth noting that near toll quality has been achieved at 6.3 kpbs (G.723.1) and the next target set by the ITU is toll quality at 4 kbps. Figure 2.2 also shows that the subsequent standards have achieved toll quality at lower bit rates, however their quality is below the transparent quality of linear PCM, even at considerably high bit rates. The research undertaken during the course of this project has focused on development of hybrid coding algorithms, which produce toll to transparent speech quality, i.e. operating in the region above good quality in the MOS scale, at the rates 4 to 8 kbps.

2.6 Concluding remarks

The existing speech coders can be divided into three groups: parametric coders, waveform approximating coders, and hybrid coders. The work presented in this thesis is focused on hybrid coding of speech.

The design process of a speech coder involves several trade offs between some conflicting requirements. These requirements include the target bit rate, quality, delay, complexity, channel error sensitivity, and sending non speech signals.

The final sections of the chapter present a review of the excising telephone band speech coding standards and the applications, and potential future demands of speech coding.
Table 2.2: Comparison of telephone band speech coding standards

<table>
<thead>
<tr>
<th>Standard</th>
<th>Year</th>
<th>Algorithm</th>
<th>Bit rate*</th>
<th>MOS**</th>
<th>Delay***</th>
</tr>
</thead>
<tbody>
<tr>
<td>G.711</td>
<td>1972</td>
<td>Companded PCM</td>
<td>64</td>
<td>4.3</td>
<td>0.125</td>
</tr>
<tr>
<td>G.726</td>
<td>1991</td>
<td>VBR-ADPCM</td>
<td>16/24/32/40 toll</td>
<td>4.3</td>
<td>0.125</td>
</tr>
<tr>
<td>G.728</td>
<td>1994</td>
<td>LD-CELP</td>
<td>16</td>
<td>4</td>
<td>0.625</td>
</tr>
<tr>
<td>G.729</td>
<td>1995</td>
<td>CS-ACELP</td>
<td>8</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>G.723.1</td>
<td>1995</td>
<td>A/MP-MLQ CELP</td>
<td>5.3/6.3</td>
<td>toll</td>
<td>37.5</td>
</tr>
<tr>
<td>ITU 4</td>
<td>-</td>
<td>-</td>
<td>4</td>
<td>toll</td>
<td>25</td>
</tr>
<tr>
<td>GSM FR</td>
<td>1989</td>
<td>RPE-LTP</td>
<td>13</td>
<td>3.7</td>
<td>20</td>
</tr>
<tr>
<td>GSM EFR</td>
<td>1995</td>
<td>ACELP</td>
<td>12.2</td>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td>GSM/2</td>
<td>1994</td>
<td>VSELP</td>
<td>5.6</td>
<td>3.5</td>
<td>24.375</td>
</tr>
<tr>
<td>IS54</td>
<td>1989</td>
<td>VSELP</td>
<td>7.95</td>
<td>3.6</td>
<td>20</td>
</tr>
<tr>
<td>IS96</td>
<td>1993</td>
<td>Q-CELP</td>
<td>0.8/2/4/8.5</td>
<td>3.5</td>
<td>20</td>
</tr>
<tr>
<td>JDC</td>
<td>1990</td>
<td>VSELP</td>
<td>6.7</td>
<td>commun.</td>
<td>20</td>
</tr>
<tr>
<td>JDC/2</td>
<td>1993</td>
<td>PSI-CELP</td>
<td>3.45</td>
<td>commun.</td>
<td>40</td>
</tr>
<tr>
<td>Inmarsat-M</td>
<td>1990</td>
<td>IMBE</td>
<td>4.15</td>
<td>3.4</td>
<td>78.75</td>
</tr>
<tr>
<td>FS1015</td>
<td>1984</td>
<td>LPC-10</td>
<td>2.4</td>
<td>synthetic</td>
<td>112.5</td>
</tr>
<tr>
<td>FS1016</td>
<td>1991</td>
<td>CELP</td>
<td>4.8</td>
<td>3</td>
<td>37.5</td>
</tr>
<tr>
<td>New FS 2.4</td>
<td>1997</td>
<td>MELP</td>
<td>2.4</td>
<td>3</td>
<td>45.5</td>
</tr>
</tbody>
</table>

* Bit rate is given in kbps.

** The MOS figures given above are obtained from different formal subjective tests using different test material. The given MOS figures are therefore useful as a guide, but should not be taken as a definitive indication of codec performance.

*** Delay is the total algorithmic delay, i.e. the frame length and look ahead, and given in milliseconds.
Chapter 3

Fundamental speech coding techniques

3.1 Introduction

The initial speech compression schemes such as PCM and ADPCM quantise the speech signal sample by sample basis. However ADPCM exploits the short term correlation present in the speech waveform, and quantises an error signal, with an adaptive quantiser step size. These techniques were not adequate to further reduce the coding bit rate without significantly degrading the output speech quality. Therefore the concepts such as vector quantisation and model based parametric coding have been introduced.

Most of the low bit rate speech coders are based on the source filter model of speech production, which is a mathematical approximation of the human speech production mechanism. The time varying filter of the source filter model represents the spectral envelope or the short term correlation of the speech signal. The coefficients of the filter are estimated for each frame of buffered speech samples and transmitted. The information relating to the prediction residual of the filter are quantised and transmitted separately. The use of parametric modelling and efficient quantisation techniques have enabled synthesising highly intelligible speech at bit rates as low as 2.4 kbps.
3.2 Quantisation

Quantisation approximates a sampled signal with an infinite precision, i.e. a sampled analogue signal to a set of finite discrete values, so that the signal can be represented in a digital format. Quantisation can also be applied to high precision digital signals in order to reduce their precision. Inevitably quantisation introduces noise which cannot be removed, unlike the sampling of analogue signals above the Nyquist rate. However, by exploiting the properties of the human auditory system, much of the quantisation noise can be masked [44]. PCM quantises the speech signals and transmits a value for each speech sample. The parametric coders estimate the speech parameters for each block of data with a high precision. Direct transmission of those high precision parameters will result in a prohibitive bit rate, which will not be suitable for low bit rate applications. Consequently, the model parameters should be quantised prior to transmission, in order to accommodate in the available bandwidth of the transmission channel. There are two fundamental quantisation techniques: scalar quantisation and vector quantisation.

3.2.1 Scalar quantisation

Scalar quantisation maps each value to the nearest quantiser level from a finite set of levels. The number of bits $B$, required to represent $l$ levels is given by,

$$B = \log_2(l)$$

(3.1)

Usually $l$ is chosen to be a power of 2, so that $B$ is an integer. For a uniform quantiser the quantisation step size is given by $D/l$, where $D$ is the dynamic range of the input signal. Uniform quantisation assumes that the distribution of the input values is even across their dynamic range. However, in many speech coding applications this is not the case and uniform quantisers are rarely used due to their poor performance. The overall quantiser performance can be improved by designing the quantiser levels to match the statistical properties of the signal. For example, more quantisation levels
3.3. Vocal tract model

can be placed in the regions where parameter values are more densely populated. There are two methods to design such non uniform scalar quantisers: Cumulative Frequency Distribution (CFD) and Mean Square Error (MSE) approach [45]. Scalar quantisation is characterised by low computational complexity, robustness to channel errors, and low memory requirements.

3.2.2 Vector quantisation

Vector quantisation maps a sequence of values to a vector chosen from a codebook and represents them using an index. The encoder transmits the indices corresponding to the quantised data sequences, and the decoder extracts the quantised vectors from a replica of the codebook stored at the decoder. Usually the codebook size is chosen to be a power of 2, so that the number of bits required to represent the indices is an integer. According to Shannon's rate distortion theory, in principle, vector quantisation should always give better performance than scalar quantisation for a given transmission rate [46]. However vector quantisation is computationally complex and need memory to store the codebooks. In order to compromise between the computational and storage requirements, and quantiser performance, a number of codebook types have been developed [47]. One of the most popular methods for codebook design is the iterative clustering algorithm known as the Linde Buzo Gray (LBG) algorithm [48].

3.3 Vocal tract model

Figure 3.1 illustrates the human speech production mechanism, excluding the lungs. The speech production mechanism can be best explained by considering it as the result of two interconnected functions: excitation and modulation. The lungs and the vocal cords produce the excitation. The vocal tract modulates the excitation by changing its shape, i.e. the frequency response. The lungs provide the airflow and pressure source for speech, which controls the loudness. Voiced speech is generated by vibrations of the vocal cords, which break the air stream from the lungs into quasi-periodic pulses of air. The fundamental frequency of voiced speech is determined by the tension of the
vocal cords, which is under muscular control. The air waves generated by the vibration of the vocal cords travel along the vocal tract being spectrally shaped by the moving articulators, such as tongue, jaws, lips, and teeth.

Unvoiced sounds are generated as a result of turbulent air flow due the constrictions of the vocal tract. For unvoiced sounds vocal cords are relaxed and do not vibrate. Whispering is produced by a constriction in the larynx. Fricatives are produced by constrictions in the tongue, lips or teeth and may be accompanied by voiced excitation generated by the vocal cords, resulting in mixed excitation. Plosives are generated by completely closing a part of the vocal tract, and then by releasing the accumulated pressure. Nasal sounds are produced by passing the air flow through the nasal cavity, which gives rise to spectral nulls. A detailed description of the human speech production mechanism has been provided by O'Shaughnessy [4].

Figure 3.1: Human speech production mechanism
3.3 Vocal tract model

3.3.1 Source filter model

Speech coding algorithms model the human speech production mechanism mathematically. Source filter model is the most widely used mathematical model of the human speech production mechanism, see Figure 3.2. The vocal tract is modelled as a time varying linear filter, and the excitation is assumed to be either voiced or unvoiced. The voiced excitation is generated as a sequence of impulses separated by the pitch period, and the unvoiced excitation is generated by a random noise generator.

The source filter model is based on a number of assumptions and simplifications. The main assumption is that the source of speech excitation is independent of the modulation, i.e. there is no interaction between the vocal tract shape and larynx. The second major assumption is that the vocal tract has a linear transfer function. Although these assumptions are not strictly true, the resulting computational savings have led to the widespread adoption of the source filter model.

3.3.2 Linear Predictive Coding (LPC)

Linear Predictive Coding (LPC) is used to derive the coefficients of the time varying filter of the source filter model by analysing the speech signal. There are three basic methods for LPC analysis: Autocorrelation Method (AM), Covariance Method (CM), and Lattice Method (LM) [49]. AM is used during the course of this project due to its simpler implementation and comparable performance.
3.3. Vocal tract model

Autocorrelation Method (AM)

The time varying filter which represents the combined effects of the vocal tract, glottal flow, and the radiation of the lips can be given by the following pole-zero filter.

\[
H(z) = \frac{S(z)}{X(z)} = \frac{G \left(1 - \sum_{j=1}^{p} \beta_j z^{-j}\right)}{1 - \sum_{j=1}^{p} \alpha_j z^{-j}}
\]  

(3.2)

In equation 3.2 both poles and zeros exist in the transfer function. However, if the order of the denominator is high enough, \(H(z)\) can be approximated by an all-pole filter as given by equation 3.3.

\[
H(z) = \frac{G}{1 - \sum_{j=1}^{p} \alpha_j z^{-j}}
\]  

(3.3)

The z domain transfer function given in equation 3.3 can be written in the form of a difference equation in time domain,

\[
s(n) = Gx(n) + \sum_{j=1}^{p} \alpha_j s(n - j)
\]  

(3.4)

Equation 3.4 states that the present output, \(s(n)\) can be obtained by summing the weighted present input, \(Gx(n)\) and a weighted sum of the most recent past \(p\) output samples. The scaling term is usually taken as being equal to unity, so the next step is to determine the coefficients of the predictor, i.e. \(\alpha_j\) for \(j = 1, 2, ..., p\) where \(p\) is the order of the filter.

If the speech signal \(s(n)\) is filtered by the inverse of the predictor filter the output \(e(n)\) is called an error or residual signal,

\[
e(n) = s(n) - \sum_{j=1}^{p} \alpha_j s(n - j)
\]  

(3.5)
The main objective is to find the set of predictor coefficients $\alpha_j$ which maximises the prediction gain, i.e. which minimises the residual energy, $E$.

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

The values of $\alpha_j$ that minimise $E$ are found by setting $\frac{\partial E}{\partial \alpha_j} = 0$ for $j = 1, 2, \ldots, p$. This yields $p$ linear equations,

$$\sum_{n=-\infty}^{\infty} s(n-i) s(n) = \sum_{j=1}^{p} \alpha_j \sum_{n=-\infty}^{\infty} s(n-i) s(n-j) \quad \text{for} \quad i = 1, 2, \ldots, p \hspace{1cm} (3.7)$$

in $p$ unknowns $\alpha_j$. The speech signal is windowed to limit the extent of speech under analysis, and the infinite summations are replaced with finite summations, as follows:

$$\sum_{n=-\infty}^{\infty} s(n-i) s(n) = \sum_{j=1}^{p} \alpha_j \sum_{n=0}^{N-1} s(n-i) s(n-j) \quad \text{for} \quad i = 1, 2, \ldots, p \hspace{1cm} (3.8)$$

Where

$$\phi(i) = \sum_{n=0}^{N-1} s_w(n) s_w(n-i) \quad \text{for} \quad i = 1, 2, \ldots, p \hspace{1cm} (3.9)$$

Where

$$s_w(n) = s(n) w(n) \quad \text{for} \quad n = 0, 1, 2, \ldots, N - 1 \hspace{1cm} (3.10)$$

Where $w(n)$ is a Hamming window function given by,

$$w(n) = 0.54 - 0.46 \cos \left( \frac{2\pi n}{N-1} \right) \quad \text{for} \quad n = 0, 1, 2, \ldots, N - 1 \hspace{1cm} (3.11)$$

and $w(n) = 0$ for all the other $n$. 
The $p$ linear equations 3.8 can be expressed in matrix form, considering that $\phi(i-j) = \phi(j-i)$, as follows:

$$
\begin{pmatrix}
\phi(0) & \phi(1) & \phi(2) & \cdots & \phi(p-1) \\
\phi(1) & \phi(0) & \phi(1) & \cdots & \phi(p-2) \\
\phi(2) & \phi(1) & \phi(0) & \cdots & \phi(p-3) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\phi(p-1) & \phi(p-2) & \phi(p-3) & \cdots & \phi(0)
\end{pmatrix}
\begin{pmatrix}
\alpha_1 \\
\alpha_2 \\
\alpha_3 \\
\vdots \\
\alpha_p
\end{pmatrix}
= 
\begin{pmatrix}
\phi(1) \\
\phi(2) \\
\phi(3) \\
\vdots \\
\phi(p)
\end{pmatrix}
$$

(3.12)

In addition to being symmetric, the matrix in equation 3.12 is Toeplitz, i.e. all the elements lying on the same diagonal are equal, and the equation may be solved using an efficient recursive procedure, known as the Levinson-Durbin algorithm [49].

**Levinson-Durbin algorithm**

The algorithm’s initial conditions are given by equations 3.13.

$$E^{(0)} = \phi(0) \quad \text{and} \quad a_i^{(0)} = 0$$

(3.13)

Equations 3.14, 3.15, 3.16, and 3.17 are then evaluated sequentially in each iterative step, for $j = 1, 2, \ldots, p$.

$$k_j = \frac{\phi(j) - \sum_{i=1}^{j-1} a_i^{j-1} \phi(j-i)}{E^{(j-1)}}$$

(3.14)

$$a_j^{(j)} = k_j$$

(3.15)

$$a_i^{(j)} = a_i^{(j-1)} - k_j a_{i-j}^{(j-1)} \quad \text{for} \quad i = 1, 2, \ldots, j - 1$$

(3.16)

$$E^{(j)} = \left(1 - [k_j]^2\right) E^{(j-1)}$$

(3.17)
Finally the LP coefficients are given by,

$$\alpha_i = \alpha_i^{p}$$

(3.18)

A major assumption in the formulation of equation 3.12 is that the speech signal is stationary. For short segments of windowed speech of up to about 30 ms this assumption may be realistic, however when the speech signal is non stationary the LPC coefficients describe the smoothed average of the windowed segment. The inverse filtered residual signal may deviate from the expected ideal case of periodic impulses, even for the voiced speech, due to the simplifications and assumptions made in the derivation of LPC coefficients [50].

### 3.4 Quantisation and interpolation of LPC coefficients

Figure 3.3 depicts z domain spectra of two LPC filters with the corresponding speech spectra. The speech spectra are shifted by -80 dB for clarity. LPC coefficients provide an accurate and compact representation of the speech spectral envelope. However the spectral envelope defined by the LPC coefficients is highly sensitive to small changes in the coefficient values, which may be introduced by the quantisation process. The stability of the LPC filter is also a major concern during the quantisation process, and
3.4. Quantisation and interpolation of LPC coefficients

Checking the stability of the filter directly from the coefficients is a complex process. Moreover, in order to smooth out the block edge effects, LPC analysis is performed using overlapped windows and the speech spectral envelope is interpolated between the analysis points. Interpolating the LPC coefficients directly can cause large changes in the spectral envelope, and often results in unstable intermediate filters. Therefore LPC coefficients should be transformed into an alternative representation, which has the following features [51].

1. Stability of the corresponding LPC filter can be readily checked
2. Low sensitivity to parameter change
3. Constant sensitivity throughout parameter range
4. Stability of the interpolated intermediate filters is guaranteed if both starting and terminating parameter sets are stable
5. Linear interpolation of parameters should provide linear perceptual change

Line Spectral Frequencies (LSFs) [52] satisfy the above conditions and many speech coders quantise and interpolate the LPC coefficients in the LSF domain.

3.4.1 Line Spectral Frequencies (LSFs)

For any LP filter, which is stable and of even order \( p \), there exists a set of \( p \) Line Spectral Pair (LSP) coefficients which completely describes the filter. Line Spectral Frequencies (LSFs) measured in Hz are related to LSPs by the relation,

\[
\text{LSF}_i = \frac{f_s}{2\pi} \cos^{-1} (\text{LSP}_i)
\]  

(3.19)

Where \( f_s \) is the sampling frequency. The LP filter transfer function can be written as,

\[
H(z) = \frac{1}{1 + \sum_{j=1}^{p} \alpha_j z^{-j}} = \frac{1}{A(z)}
\]  

(3.20)

For an even value of \( p \) and a stable filter, \( A(z) \) can be decomposed as,
3.4. Quantisation and interpolation of LPC coefficients

\[ A(z) = \frac{P(z) + Q(z)}{2} \] (3.21)

Where

\[ P(z) = A(z) + z^{-(p+1)}A(z^{-1}) = (1 + z^{-1}) \prod_{i=1,3,\ldots,p-1} (1 - 2\cos(\omega_i)z^{-1} + z^{-2}) \] (3.22)

and

\[ Q(z) = A(z) - z^{-(p+1)}A(z^{-1}) = (1 - z^{-1}) \prod_{i=2,4,\ldots,p} (1 - 2\cos(\omega_i)z^{-1} + z^{-2}) \] (3.23)

The roots of \( P(z) \) and \( Q(z) \) lie on the unit circle and occur in conjugate pairs, with the exception of the roots at \( z^{-1} = -1 \) in the case of \( P(z) \) and \( z^{-1} = 1 \) for \( Q(z) \). Since \( \arg(z) = -\arg(z^*) \), there are \( p \) unknowns to be found, i.e. the arguments, \( \omega_i \) of the roots. The LSP coefficients are given by,

\[ LSP_i = \cos(\omega_i) \] (3.24)

This angular information is sufficient to fully represent the root positions since they lie on the unit circle. Various methods exist for solving the polynomials \( P(z) \) and \( Q(z) \) \([2]\), which include complex root method, real root method, ratio filter method, and Chebyshev series method. The LSF to LPC transformation is much simpler, and achieved by multiplying out the products in equations 3.22 and 3.23, and then substituting the results into equation 3.21.

Properties of LSFs

The LSF representation has a number of properties which can be exploited in order to achieve efficient quantisation of the parameters, such as intra and inter frame correlations \([25]\). The stability of the LPC filter is guaranteed provided that the LSFs are
3.4. Quantisation and interpolation of LPC coefficients

monotonically increasing and are bound between the limits 0 Hz and \( f_s/2 \), where \( f_s \) is the sampling frequency, i.e. \( 0 < \omega_1 < \omega_2 < \ldots < \omega_p < \pi \). This ordering property accounts for the intra and inter frame correlations observed in the LSF tracks. LSFs have a direct correlation with the magnitude spectrum of the LPC filter as shown in Figure 3.4. Closely grouped LSFs indicate the presence of a formant, and the formant bandwidth depends on the closeness of the LSFs, closer the LSFs narrower the formant bandwidth. The LSF sensitivity is fairly uniform and linear interpolation between two stable sets of LSFs will always produce a stable intermediate filter. The spectral sensitivity of each LSF is localised in the region of the LPC spectrum it represents, therefore any errors incurred in a particular LSF will not affect the whole spectrum.

**LSF quantisation**

The early LSF quantisers used scalar quantisation [53], [54] due to the high computational demand of Vector Quantisation (VQ). As a trade off between the high performance and the complexity of VQ, split VQ was designed [55]. Another efficient way of reducing the complexity of VQ is multistage VQ [56].

LSF parameters may be quantised independent of the previous LSF vector, i.e. memory less quantisation. However parameters extracted from speech such as LSFs typically

![Figure 3.4: LPC filter frequency response and LSFs](image_url)
3.4. Quantisation and interpolation of LPC coefficients

(a) Moving average quantiser

(LSFs) → $\Sigma$ → $r$ → 3 Stage VQ → $\hat{r}$

$\alpha \hat{r}_{-1}$

(b) Moving average dequantiser

$\hat{r}$ → $\Sigma$ → Quantised LSFs

$\alpha \hat{z}_{-1}$

Figure 3.5: Moving Average (MA) prediction

show a significant interframe correlation, which can be exploited to increase the coding efficiency using memory based quantisation schemes [57]. However the performance of memory based quantisation schemes may degrade under noisy channel conditions. This imposes another compromise on the VQ design, i.e. the amount of prediction and the performance for noisy channels.

The LSF quantiser used in the course of this research is a Multi Stage Vector Quantiser (MSVQ), with a first order Moving Average (MA) prediction [58]. The MSVQ has three stages, the first and the second stages use 8 bit vector codebooks, and the third stage uses a 7 bit vector codebook, giving a total of 23 bits per frame.

Figure 3.5 illustrates the first order MA prediction process. The three stage vector quantiser quantise the prediction residual, $r$ as shown in equation 3.25.

$$\hat{r} = Q[r] = Q[V_{LSF} - \alpha \hat{r}_{-1}]$$  \hspace{1cm} (3.25)

Where $\alpha = 0.5$ and $Q[\cdot]$ represents the three stage vector quantisation. The quantised LSF vector, $\hat{V}_{LSF}$ is given by,

$$\hat{V}_{LSF} = \hat{r} + \alpha \hat{r}_{-1}$$  \hspace{1cm} (3.26)

An important feature of this MA scheme is that it compromises the prediction gain for channel errors. Note that the decoder uses only two consecutive residual vectors to reconstruct the quantised LSFs, i.e. the present quantised residual vector and the
previous vector. Therefore when a random channel error corrupts one vector it affects only two reconstructed LSF vectors. However this scheme does not completely remove the interframe correlation present, since it subtracts the scaled previous vector of the quantised residual rather than a prediction of the present LSF vector.

Figure 3.6 depicts a block diagram of the three stage vector quantiser. The quantisation error vector of each stage is quantised again in the next stage, and the dequantiser add all the three vectors to produce the final quantised output vector as shown in Figure 3.7. The quantisation error may be minimised by a full search of all the possible combinations, which gives the minimum spectral distortion possible for the given codebook structure. The resulting complexity is $2^8 + 2^8.2^8 + 2^8.2^8.2^7$ vector searches, which is more than the complexity of a 23 bit single codebook search. However the codebook structure shown in Figure 3.6 requires less memory than a 23 bit single codebook. The least possible complexity, $2^8 + 2^8 + 2^7$ may be achieved by a sequential search, at the expense of an increase in the spectral distortion. As a compromise an M-L tree search procedure [59] is used, with $L = 3$ and $M = 16$. Starting with the first codebook, the M codevectors which achieve the lowest distortion are selected and M difference vectors, i.e. M targets for the next stage, are computed. The second codebook is searched M times, i.e. once for each target vector, and the M paths that achieve the lowest overall distortion are selected. The procedure continues for all L stages. After the $L^{th}$ stage, the indices corresponding to the path that give the lowest overall distortion are selected.
3.4. Quantisation and interpolation of LPC coefficients

LSF interpolation

The LPC analysis is performed on buffered blocks of speech samples, typically every 10 to 30 ms. The estimated LPCs are quantised and transmitted in the LSF domain. This process is equivalent to sampling the speech spectral envelope at the frame rate. The spectral envelope shows rapid variations across the speech transitions, and two consecutive frames may have significantly different spectral envelope estimates. However the vocal tract frequency response evolves as a continuous function of time, and changing the spectral envelope for each frame in the speech synthesis may cause audible distortions, due to the sudden changes in the spectral envelope.

In order to produce a smooth synthetic spectrum the spectral envelope is linearly interpolated in the LSF domain. In principle a new set of LPCs may be calculated at the sampling rate. However such a high interpolation rate is computationally demanding and equally good perceptual quality can be achieved by interpolating the LSFs for each sub frame of 5 ms [51]. The decoder has two sets of LSFs for each synthesis frame, corresponding to the spectral envelope estimates of the analysis frames centred at the synthesis frame boundaries, using a half a frame look ahead. All the speech coding algorithms developed during the course of this research estimate the spectral envelope for every 20 ms frame and interpolate every 5 ms, and the LSF vector of the \( i \)th sub frame \( V_i \) is given by,

\[
V_i = \frac{1}{4} [(3.5 - i) V_j + (0.5 + i) V_{j+1}] \quad \text{for} \quad i = 0, 1, 2, 3 \tag{3.27}
\]

Where \( V_j \) is the LSF vector of the \( j \)th analysis frame.
3.5 Pitch determination

Pitch determination may be performed directly on the time domain speech signal or in the frequency domain using the Fourier transform of the speech.

3.5.1 Time domain approaches

Time domain pitch determination algorithms are based on the waveform similarity of the speech signal and the shifted version of itself by a pitch period.

Autocorrelation method

The autocorrelation function with a lag of $\tau$ is defined as,

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

$R(\tau)$ has peaks when $\tau$ is zero or multiples of the pitch period, see Figure 3.8. The pitch period may be extracted by selecting the minimum $\tau$ larger than the minimum possible pitch, corresponding to a peak of $R(\tau)$ larger than a certain threshold. The
3.5. **Pitch determination**

threshold may be defined as the maximum value of \( R(\tau) \) within the pitch search range, multiplied by a factor slightly less than unity, typically 0.9. However the large energy variations at the offsets may cause \( R(\tau) \) to be much larger at the lags corresponding to the pitch multiples compared to the \( R(\tau) \) at the correct pitch lag. This problem can be eliminated by using autocorrelation with symmetric energy normalisation, \( R' \) given by,

\[
R'(\tau) = \frac{\sum_{n=0}^{N-1} s(n) s(n-\tau)}{\frac{1}{2} \left[ \sum_{n=0}^{N-1} s^2(n) + \sum_{n=0}^{N-1} s^2(n-\tau) \right]} \tag{3.29}
\]

When only one term in the denominator of equation 3.29 is used, i.e. either \( s(n) \) or \( s(n-\tau) \), \( R' \) becomes noisy either at the offsets or onsets, due to the small normalisation energy.
3.5. Pitch determination

Average Magnitude Difference Function (AMDF)

Average magnitude difference function [60] is defined by equation 3.30.

\[ A(\tau) = \frac{1}{N} \sum_{n=0}^{N-1} |s(n) - s(n - \tau)| \] (3.30)

\( A(\tau) \) has minima when \( \tau \) is zero or multiples of the pitch period, see Figure 3.9. Therefore additional computations are required to eliminate the multiples of the pitch period. Formant interaction is another difficulty in pitch estimation, which causes minima to appear in AMDF at the multiples of the formant period, see Figure 3.9 (b). Formant interaction may mislead a pitch multiple removal algorithm.

The main advantage of AMDF was its computational simplicity. The subtraction and magnitude operations were much faster than the multiply-add operations in the early processors which did not have hardware multipliers. However the introduction of the pipelined multiply-add instruction with a hardware multiplier has removed this advantage.

3.5.2 Frequency domain approaches

Frequency domain pitch determination algorithms extract the fundamental frequency from the harmonics of the speech spectrum.

Sinusoidal speech model matching

Sinusoidal speech model matching proposed by McAulay [61] assumes the speech signal, \( s(n) \) is composed of a number of harmonically unrelated sinusoidal components, as follows:

\[ s(n) = \sum_{l=1}^{L} A_l \exp (j n \omega_l + j \phi_l) \] (3.31)

The amplitudes \( A_l \), frequencies \( \omega_l \), and phases \( \phi_l \) can be estimated from the points corresponding to the local maxima of the magnitude spectrum. Then a synthetic signal \( \hat{s}(n, \omega_0) \), is generated, composed of entirely harmonically related sinusoids, as follows:
3.5. Pitch determination

\[
\hat{s}(n, \omega_0) = \sum_{l=1}^{K(\omega_0)} \bar{A}_l \exp(j n l \omega_0 + j \theta_l)
\]  

(3.32)

Where the harmonic amplitudes, \( \bar{A}_l \) are obtained from the spectral envelope, and \( \theta_l \) are the harmonic phases. The candidate frequency \( \omega_0 \), which minimises the mean square error \( \varepsilon(\omega_0) \), between \( s(n) \) and \( \hat{s}(n, \omega_0) \) is chosen as the fundamental frequency.

\[
\varepsilon(\omega_0) = \frac{1}{N} \sum_{n=0}^{N-1} [s(n) - \hat{s}(n, \omega_0)]^2
\]  

(3.33)

Direct evaluation of equation 3.33 for all the possible \( \omega_0 \) and choosing the \( \omega_0 \) corresponding to the minimum \( \varepsilon(\omega_0) \) is an extremely computationally intensive process. However McAulay simplified the search procedure based on a number of assumptions and perceptual enhancements, which can be summarised as follows [62]:

1. The spectrum of \( s(n) \) is well resolved and can be approximated by sinc functions at each component frequency, each scaled by \( A_l \), where \( \text{sinc}(x) = \sin(x)/x \).
2. The spectrum of \( \hat{s}(n, \omega_0) \) is well resolved and can be approximated by sinc functions at harmonics of the candidate frequency \( \omega_0 \), scaled by the spectral envelope at the harmonic frequencies.
3. The matching of one sinusoidal component from \( s(n) \) to another from \( \hat{s}(n, \omega_0) \) is given by the product of their harmonic amplitudes, i.e. \( A_l \) and \( \bar{A}_l \) weighted with a distance function \( D(\omega - k\omega_0) \), see equation 3.35.
4. Each sinusoidal component of \( s(n) \) is represented by only one of its counterparts from \( \hat{s}(n) \), the one which has the greatest matching, usually the closest.
5. The value of \( D(\omega - k\omega_0) \) may be pre computed and stored in a look up table.
6. Minimising the error \( \sum_{k=0}^{N} [x(k) - y(k, a, b, c, ...)]^2 \) over the variables \( a, b, c, ... \) is equivalent to maximising the function \( \sum_{k=0}^{N} y(k, a, b, c, ...) [x(k) - 0.5y(k, a, b, c, ...)] \).

After the simplifications, the fundamental frequency can be determined by maximising \( \rho(\omega_0) \) with respect to \( \omega_0 \), where \( \rho(\omega_0) \) is given by,

\[
\rho(\omega_0) = \sum_{l=1}^{K(\omega_0)} \bar{A}_l \left\{ \max_i \left[ A_i D(\omega_l - k \omega_0) \right] - \frac{1}{2} \bar{A}_l \right\}
\]  

(3.34)
3.5. Pitch determination

![Pitch determination diagram](image)

Figure 3.10: (a) Original, shifted by +40 dB and (b) synthetic speech spectra

Where $K (\omega_0) = [\pi/\omega_0]$, 

$$D (\omega_l - k\omega_0) = \frac{\sin \left( \frac{2\pi \omega_l - k\omega_0}{\omega_0} \right)}{2\pi \omega_l - k\omega_0 \omega_0} \quad \text{for} \quad |\omega_l - k\omega_0| \leq \frac{\omega_0}{2} \quad (3.35)$$

and $D (\omega_l - k\omega_0) = 0$ otherwise. An improved version of the sinusoidal speech model matching algorithm, which has a higher resilience against pitch multiple and sub multiple errors [5], [58] is used in the hybrid coding algorithm described in chapter 6.

Synthetic spectral matching

Synthetic spectral matching proposed by Griffin [63] is based on the similarity of the speech spectrum to a synthetic voiced spectrum generated for each candidate fundamental frequency, $\omega_0$. An error criterion is defined as, 

$$\varepsilon (\omega_0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[ S_w (\omega) - \hat{S}_w (\omega, \omega_0) \right]^2 d\omega \quad (3.36)$$

The fundamental frequency is given by $\omega_0$, which minimises $\varepsilon (\omega_0)$. $S_w (\omega)$ is the fourier transform of the windowed speech signal, and $\hat{S}_w (\omega, \omega_0)$ denotes the synthetic spectrum.
In order to reduce the computational complexity, the error minimisation process is divided into non overlapping frequency bands centred at the harmonic frequencies of \( \omega_0 \), and when synthesising \( \hat{S}_w(\omega, \omega_0) \) for a particular harmonic, the spectral leakage from the other harmonics is assumed to be negligible. The simplified error criterion for the \( k^{th} \) harmonic is given by,

\[
\varepsilon_k = \frac{1}{2\pi} \int_{(k-0.5)\omega_0}^{(k+0.5)\omega_0} \left| S_w(\omega) - A_k(\omega_0) W(\omega - k\omega_0) \right|^2 d\omega \quad \text{for } k = 1, 2, ..., K(\omega_0) \quad (3.37)
\]

Where \( K(\omega_0) = [\pi/\omega_0] \) and \( A_k(\omega_0) \) is given by,

\[
A_k(\omega_0) = \frac{\int_{(k-0.5)\omega_0}^{(k+0.5)\omega_0} S_w(\omega) W^*(\omega - k\omega_0) d\omega}{\int_{(k-0.5)\omega_0}^{(k+0.5)\omega_0} W^2(\omega - k\omega_0) d\omega} \quad (3.38)
\]

Where \( W(\omega) \) is the Fourier transform of the window function that used to window the original speech signal. The fundamental frequency is given by the candidate \( \omega_0 \), which minimises \( \varepsilon'(\omega_0) \), defined as,

\[
\varepsilon'(\omega_0) = \sum_{k=1}^{K(\omega_0)} \varepsilon_k \quad (3.39)
\]

Figure 3.10 depicts an original speech spectrum and the synthesised spectrum for the selected \( \omega_0 \). A small deviation in \( \omega_0 \) results in large deviations at the higher harmonics and large error values, \( \varepsilon'(\omega_0) \). Therefore the fundamental frequency can be estimated with high precision. However since a full search at fractional resolutions is computationally intensive, synthetic spectral matching is used to refine the initial pitch estimate obtained from sinusoidal speech model matching in the hybrid coding algorithm described in chapter 6.
3.6 Concluding remarks

The scalar quantisation or the vector quantisation can be used to quantise the speech signals and the speech model parameters for digital transmission. The source filter model is the most widely used speech production model in low bit rate speech coding. The filter coefficients of the time varying filter of the source filter model are derived using Linear Predictive Coding (LPC). The coefficients of the linear prediction filter are converted into Line Spectral Frequencies (LSFs) for interpolation and quantisation. The stability of the LPC filter can be readily checked in the LSF domain and linear interpolation of the LSFs provides a linear perceptual change. Multi Stage Vector Quantiser (MSVQ) with a first order Moving Average (MA) prediction is an efficient LSF quantisation scheme, which is used to quantise the LSFs during the course of this research.

One of the most important parameters in low and medium bit rate speech coding is the pitch period of the speech waveform. Pitch may be extracted in the time domain by considering the waveform similarity of the speech waveform and the shifted version of itself by a pitch period or in the frequency domain by considering the harmonic structure of the speech spectrum. An accurate pitch estimate is essential for the operation of low bit rate harmonic coders. Harmonic coding is described in chapter 4. The analysis by synthesis coders, described in chapter 5 employ a time domain pitch prediction filter, which may use an initial pitch estimate, however the accuracy of the pitch estimate is not crucial for their functionality.
Chapter 4

Harmonic coders

4.1 Introduction

McAulay introduced a general sinusoidal speech analysis and synthesis technique [6] and developed the Sinusoidal Transform Coder (STC) [64] to demonstrate the applicability of the technique in low bit rate speech coding. Sinusoidal coding does not restrict the component sinusoids of the synthesised speech to be harmonics of the fundamental frequency. The frequency tracks of the sinusoidal representation may vary independent of each other between their birth and death. However in harmonic coding the higher frequency components are multiples of the fundamental frequency [65]. Therefore harmonic coding can be seen as a sub set of general sinusoidal coding. At low bit rates STC also restricts the frequency tracks to be harmonics of the fundamental frequency, and deduce the harmonic phases at the decoder, simply because the available bits are not sufficient to encode the large number of parameters of the general sinusoidal representation, see section 4.4.1.

STC was introduced as an alternative to the source filter model, and analysis and synthesis was directly applied to the speech signal. The binary voicing decision of the source filter model is one of its major limitations. STC employs a more general mixed voicing scheme by separating the speech spectrum into voiced and unvoiced components, using a voicing transition frequency above which the spectrum is declared
4.2. Sinusoidal analysis and synthesis

Figure 4.1: General sinusoidal analysis and synthesis

unvoiced. However the most recent harmonic coders operate in the LPC residual domain, e.g. Split Band LPC (SB-LPC) [5]. SB-LPC replaces the binary excitation of the source filter model with a more general mixed voiced harmonic excitation, and filter the excitation signal using an LPC filter. The LPC residual has a simpler phase spectrum than the speech, and the phases can be approximated better by the harmonic synthesis, using the integrals of the component frequencies, without transmitting the phase spectrum. Moreover LPC models the large variation in the speech magnitude spectrum, and simplifies the harmonic amplitude quantisation.

4.2 Sinusoidal analysis and synthesis

Figure 4.1 depicts block diagrams of the sinusoidal analysis and synthesis process introduced by McAulay. The speech spectrum is estimated by windowing the input speech signal using a Hamming window and computing the Discrete Fourier Transform (DFT). The frequencies, amplitudes, and phases corresponding to the peaks of the magnitude
4.3. Parameter estimation

Figure 4.2: Sinusoidal synthesis with matched frequency tracks

spectrum become the model parameters of the sinusoidal representation. Employing a pitch adaptive window length of two and a half times the average pitch improves the accuracy of peak estimation. The synthesiser generates the sine waves corresponding to the estimated frequencies and phases and modulates them using the amplitudes. Then all the sinusoids are summed up to produce the synthesised speech. The block edge effects are smoothed out by applying overlap and add using a triangular window. Overlap and add is effectively a simple interpolation technique, and in the sinusoidal synthesis, it requires parameter update rates of at least 12.5 ms for good quality speech synthesis. At lower frame rates the spectral peaks are matched between the analysis frames to form frequency tracks. The amplitudes of the frequency tracks are linearly interpolated, and the instantaneous phases are generated using a cubic polynomial [6], see Figure 4.2.

4.3 Parameter estimation

The low bit rate sinusoidal coders estimate the amplitudes at the harmonics of the fundamental frequency. The harmonic phases are not transmitted. Instead the phases are deduced from the spectral envelope on the assumption that it is the gain response of a minimum phase transfer function, and added to the integrals of the component frequencies. STC implements the harmonic phases explicitly. LPC based coders implement the phases implicitly through the time domain LPC synthesis filter. Improved Multi Band Excitation (IMBE) coder does not use any kind of phase information and the phases are evolved as the integrals of the component harmonic frequencies. Restricting the component frequencies to the harmonics and modelling the phases at the decoder suits well for stationary voiced segments of speech. However in general the speech signal
is not stationary voiced, and consists of mixed voiced and unvoiced segments. When those segments are synthesised with the phase models described above, the synthesised speech sounds too buzzy. In order to remove this buzzyness the concept of frequency domain voicing was introduced to the low bit rate harmonic coders [66]. Frequency domain voicing allows the synthesis of mixed voiced signals, by separating the speech spectrum into frequency bands marked as either voiced or unvoiced.

The frequency domain voicing decisions are usually made for each harmonic of the speech spectrum. However they may be quantised using different techniques, see section 4.4. Therefore an accurate pitch estimate is a prerequisite of harmonic amplitude and voicing determination. The frequency domain voicing determination techniques based on spectral matching need a high precision pitch estimate for their proper functionality. A small error in the pitch will cause large deviations at the high frequency harmonics, and subsequently declaring them as unvoiced. Furthermore, female voices with short pitch periods require a fractional pitch estimate for good quality speech synthesis, since the percentage error of a fractional pitch error is higher for short pitch periods. In order to reduce the complexity of a high precision pitch estimation, an initial pitch estimate is further refined by performing a limited search around the initial estimate. Pitch estimation and refinement techniques are described in detail in section 3.5.

4.3.1 Voicing determination

Multi band approach

Harmonic voicing is estimated by computing the normalised mean squared error of a synthetic voiced spectrum, \( \hat{S}_w(\omega, \omega_0) \) with respect to the speech spectrum, \( S_w(\omega) \), and comparing it against a threshold function for each harmonic band [63]. The normalised mean squared error, \( D_k \) of the \( k^{th} \) harmonic band is given by:

\[
D_k = \frac{\int_{(k-0.5)\omega_0}^{(k+0.5)\omega_0} \left[ S_w(\omega) - \hat{S}_w(\omega, \omega_0) \right]^2 d\omega}{\int_{(k-0.5)\omega_0}^{(k+0.5)\omega_0} S_w^2(\omega) d\omega}
\]

for \( k = 1, 2, ..., K \)
4.3. Parameter estimation

Figure 4.3: Normalised spectral error, (a) original spectrum $S_w(\omega)$, (b) synthetic spectrum $\hat{S}_w(\omega, \omega_0)$, and (c) $D_k$

Where $K = [\pi/\omega_0]$ and $\omega_0$ is the normalised fundamental frequency obtained from synthetic spectral matching, refer to section 3.5.2. Using the equations 3.36 and 3.37 $D_k$ may be simplified to,

$$D_k = \frac{\varepsilon_k}{\frac{1}{2\pi} \int_{(k-0.5)\omega_0}^{(k+0.5)\omega_0} S^2_w(\omega) d\omega}$$

Figure 4.3 illustrates $D_k$ values of two speech spectra with the corresponding synthetic spectra. If $D_k$ is below the threshold function, i.e. a small error and a good spectral match, the $k^{th}$ band is declared voiced. The initial Multi Band Excitation (MBE) coder used a constant threshold for all the bands. However the most recent versions use several heuristic rules to obtain the best performance [2], e.g. the threshold function is decreased, with the frequency, if the same band of the previous frame was unvoiced, if the high frequency energy exceeds the low frequency energy, and if the speech energy approaches the energy of the background noise.

Sinusoidal model approach

McAulay proposed a different voicing determination technique for his Sinusoidal Transform Coder (STC) [64]. The speech spectrum is divided into two bands, determined
4.3. Parameter estimation

by a voicing transition frequency above which the spectrum is declared unvoiced. This
method estimates the similarity between the harmonically synthesised signal, \( \hat{s}(n, \omega_0) \)
and the original speech signal \( s(n) \). The pitch, \( \omega_0 \) is estimated using the sinusoidal
speech model matching described in section 3.5.2 The Signal to Noise Ratio (SNR), \( \delta \)
between \( s(n) \) and \( \hat{s}(n, \omega_0) \) is given by,

\[
\delta = \frac{\sum_{n=0}^{N-1} s^2(n)}{\sum_{n=0}^{N-1} [s(n) - \hat{s}(n, \omega_0)]^2}
\] (4.3)

Where \( \hat{s}(n, \omega_0) \) is given in equation 3.32 and \( N \) is the analysis frame length. Based on
the assumptions given in section 3.5.2 McAulay simplified the equation 4.3 for reduced
computational complexity, and the simplified \( \delta \) is given by,

\[
\delta = \frac{\sum_{i=1}^{L} A_i^2}{\sum_{i=1}^{L} A_i^2 - 2N \rho(\omega_0)}
\] (4.4)

Where \( A_i \) and \( \rho(\omega_0) \) are given in the equations 3.31 and 3.34 respectively. The voicing
level, \( L_v(\delta) \), i.e. the ratio of the voiced bandwidth to the speech bandwidth \( 0 \leq L_v(\delta) \leq 1 \), is defined as,

\[
L_v(\delta) = \begin{cases} 
1 & \delta > 13dB \\
\frac{1}{6} (\delta - 4) & 4dB \leq \delta \leq 13dB \\
0 & \delta < 4dB
\end{cases}
\] (4.5)

The advantage of estimating the voicing for independent bands is it essentially removes
the spectral tilt, i.e. all the components are equally weighted. When the voicing is
based on a single metric, i.e. \( \delta \), the large amplitudes contribute more to the overall
decision and if these have been corrupted by background noise, resulting a large voicing
error [64]. Therefore the voicing estimates based on independent bands are more robust
against background noise.
4.3. Parameter estimation

4.3.2 Harmonic amplitude estimation

Magnitude spectrum peak picking

Harmonic amplitudes may be estimated by simple peak picking of the magnitude spectrum and searching for the largest peak in each harmonic band. The peak amplitude value, $S_w(m_k)$, should be normalised by a factor depending on the window function used, as follows:

$$A_k = \frac{|S_w(m_k)|}{\kappa} \quad \text{for} \quad -\frac{\omega_0}{2} < \frac{2\pi}{N} m_k - k\omega_0 < \frac{\omega_0}{2} \quad \text{and} \quad k = 1, 2, ..., K \quad (4.6)$$

Where $\omega_0$ is the normalised fundamental frequency, $K = [\pi/\omega_0]$, $\kappa = \sum_{n=0}^{N-1} w(n)$, $w(n)$ is the window function, $N$ is the length of the window, and $S_w(m)$, the windowed speech spectrum, is given by,

$$S_w(m) = \sum_{n=0}^{N-1} s(n)w(n)e^{-j\frac{2\pi}{N}mn} \quad \text{for} \quad m = 0, 1, 2, ..., N \quad (4.7)$$

Spectral similarity

Harmonic amplitudes may be estimated by computing the normalised cross correlation between the harmonic lobes of the speech spectrum and the main lobe of the window spectrum, see equation 4.8. This method is based on the fact that the spectrum of the windowed speech is equivalent to the convolution between the speech spectrum and the window spectrum, and assumes that the speech signal is stationary during the windowed segment and the spectral leakage due to the side lobes of the window spectrum is negligible.

$$A_k = \frac{\sum_{m=a_k}^{b_k-1} S_w(m)W^*(2\pi m/N - k\omega_0)}{\sum_{m=a_k}^{b_k-1} W^2(2\pi m/N - k\omega_0)} \quad \text{for} \quad k = 1, 2, ..., K \quad (4.8)$$
4.3. Parameter estimation

Where

\[ a_k = \max \left( \left[ \frac{N}{2\pi} (k - 1/2) \omega_0 \right], 0 \right) \]  \hspace{1cm} (4.9)

\[ b_k = \min \{ a_k, N/2 \}, \] and \( W(\omega) \) is the spectrum of the window function, given by,

\[ W(\omega) = \sum_{n=0}^{N-1} w(n) e^{-j\omega n} \]  \hspace{1cm} (4.10)

In practice \( W(\omega) \) is computed with a high resolution FFT e.g. \( 2^{14} \) samples, by zero padding the window function and stored in a look up table. The high resolution FFT is required, because in general the spectral samples, \( m \) of \( S_w(m) \) do not coincide with the harmonic locations, \( k\omega_0 \) of the fundamental frequency. Hence \( W(\omega) \) is shifted to the harmonic frequency and down sampled to coincide with the corresponding spectral samples of \( S_w(m) \), as shown in equation 4.8. \( W(\omega) \) is pre computed and stored in order to reduce the computational complexity.

The spectral cross correlation based amplitude estimation gives the optimum gain of the harmonic lobes with respect to the main lobe of the window spectrum, hence a more accurate estimate than the simple peak picking. However the cross correlation based method has a higher complexity and requires a high precision pitch estimate. The unvoiced amplitudes are calculated as the RMS spectral energy over the unvoiced spectral bandwidth, given by,

\[ A_{k,\text{uv}} = \frac{1}{r} \sqrt{\frac{\sum_{m=a_k}^{b_k} S_w^2(m)}{b_k - a_k}} \]  \hspace{1cm} (4.11)

The harmonic amplitude estimation techniques described may be applied to either the speech spectrum or the LPC residual spectrum.
4.4 Common harmonic coders

This section describes three examples of low bit rate harmonic coders: Sinusoidal Transform Coding (STC) [64], Improved Multi Band Excitation (IMBE) [11], and Split Band Linear Predictive Coding (SB-LPC) [5]. STC and IMBE apply sinusoidal analysis and synthesis techniques on the speech signal and SB-LPC applies them on the LPC residual signal. All the three examples restrict the synthesis sinusoidal components to be harmonics of the fundamental frequency.

4.4.1 Sinusoidal Transform Coding (STC)

STC operating at 4.8 kbps divides the speech spectrum into two voicing bands using the sinusoidal model approach described in section 4.3.1. The lower part of the spectrum which is declared as voiced is synthesised as follows:

$$ s_v(n) = \sum_{l=1}^{L_v} A(l\omega_0^k) \exp\left( jli\phi_0(n) + j\phi_s\left(l\omega_0^k\right) \right) \quad \text{for} \quad -N/2 \leq n \leq N/2 \quad (4.12) $$

Where

$$ \phi_0(n) = n\omega_0^k + \phi_0^k \quad (4.13) $$

Where

$$ \phi_0^k = \phi_0^{k-1} + \left(\omega_0^{k-1} + \omega_0^k\right)N'/2 \quad (4.14) $$

Where N + 1 is the frame length, \( \omega_0^k \) is the normalised fundamental frequency of the \( k^{th} \) frame, \( N' \) is the duration between the analysis points, \( A(\omega) \) is the spectral envelope obtained by interpolating the selected peaks of the magnitude spectrum, \( \phi_s(\omega) \) is the phase spectrum derived from the spectral envelope on the assumption that it is the gain response of a minimum phase transfer function, and \( L_v \) is the harmonic just below the voicing transition frequency. Pitch is estimated using the sinusoidal speech model matching pitch detection algorithm described in section 3.5.2.
The upper part of the spectrum which is declared as unvoiced is synthesised as follows:

\[
\hat{s}_{uv}(n) = \sum_{l=L_v+1}^{K(\omega^u_0)} \hat{A}(l\omega^u_0) \exp \left( jl\phi_0(n) + j\phi_s(l\omega^u_0) + jU[-\pi, \pi] \right) \tag{4.15}
\]

Where \( K(\omega^u_0) = [\pi/\omega^u_0] \) and \( U[-\pi, \pi] \) denotes a uniformly distributed random variable in the range \(-\pi\) and \(\pi\). When a frame is fully unvoiced the pitch estimate is meaningless, and pitch frequencies larger than 150 Hz may degrade the perceptual quality of unvoiced speech. In order to synthesise the noise like unvoiced speech with adequate quality the number of sinusoids with random phases should be sufficiently large. Therefore the pitch frequency is set to 100 Hz for unvoiced speech. The synthesised speech of the \(k^{th}\) frame is given by,

\[
\hat{s}(n) = \hat{s}_v(n) + \hat{s}_{uv}(n) \tag{4.16}
\]

Overlap and add method is used with a triangular window to produce the final speech output. Therefore the frame length is equal to twice the duration between the analysis points, i.e. \(N = 2N'\). The frequency response of the spectral envelope is given by,

\[
H(\omega) = \hat{A}(\omega) \exp(j\phi_s(\omega)) \tag{4.17}
\]

Which is approximated by an all pole model,

\[
H(\omega) \cong \frac{g}{1 - \sum_{i=1}^{p} a_i z^{-i}} \quad \text{for} \quad |z| = 1 \tag{4.18}
\]

Where \(g\) is the gain and \(a_i\) are the predictor coefficients. The conventional time domain all pole LPC analysis is performed on the speech signal, and the filter order should be limited to half the smallest pitch period, just to model the formant spectral envelope. LPC filters with large number of taps tend to resolve the harmonic structure. However
in the case of STC all pole modelling is applied to the estimated spectral envelope. Hence the filter order is not restricted, and can be increased depending only on the desired accuracy of the spectral envelope. The 4.8 kbps STC uses a 14th order all pole model and quantises the predictor coefficients in the LSF domain. In addition to the LSFs STC transmits gain, pitch, and voicing.

4.4.2 Improved Multi Band Excitation (IMBE)

IMBE operating at 4.15 kbps divides the speech spectrum into several voiced and unvoiced frequency bands, using the multi band approach described in section 4.3.1. However IMBE makes the voicing decisions for groups of 3 harmonics, and a single bit is allocated for each group. The total number of voicing bits $B_v$ is limited to a maximum of 12 bits and the harmonics beyond the coverage of voicing are declared unvoiced. Pitch is estimated using the synthetic spectral matching described in section 3.5.2, and transmitted using 8 bits. The frame length is 20 ms giving 83 bits per frame at 4.15 kbps, and the remaining bits, i.e. $83 - 8 - B_v$ are allocated for spectral amplitudes. The voiced amplitudes are estimated using equation 4.8 and the unvoiced amplitudes are estimated using equation 4.11. The voiced bands are synthesised as follows:

$$\hat{s}_v(n) = \sum_{k=\text{voiced}} A_k \cos(k\phi_0(n)) \quad \text{for } n = 0, 1, 2, ..., N - 1 \quad (4.19)$$

Where $N$ is the frame length and the fundamental phase evolution, $\phi_0(n)$ is defined by the following equations.

$$\phi_0(n) = \phi_0(n - 1) + \omega_0(n) \quad (4.20)$$

$$\omega_0(n) = \frac{1}{N} (N - n)\bar{\omega}_0^{l-1} + n\omega_0^l \quad (4.21)$$

Where $\phi_0(-1)$ is $\phi_0(N - 1)$ of the previous frame and $\omega_0^l$ is the normalised fundamental frequency estimated at the end of the $l^{th}$ frame. The amplitudes of the voiced harmonics
are linearly interpolated between the analysis points. If the corresponding harmonic of one analysis point does not exist or declared unvoiced then its amplitude is set to zero, and the harmonic frequency is set to a constant, i.e. a multiple of the pitch frequency. However if the pitch estimate is not steady, for all the harmonics neither the pitch nor the amplitudes are interpolated, instead overlap and add method is used.

The unvoiced component is synthesised using filtered white Gaussian noise. White noise is generated in the time domain and transformed into frequency domain, the bands corresponding to the voiced components are set to zero and the unvoiced bands are scaled according to the unvoiced gain factors. The inverse Fourier transform of the modified spectrum gives the unvoiced component. The unvoiced component, $\hat{s}_{uv}(n)$ is produced using overlap and add method. The synthesised speech $\hat{s}(n)$ is given by,

$$\hat{s}(n) = \hat{s}_v(n) + \hat{s}_{uv}(n)$$  \hspace{1cm} (4.22)

An interesting feature of the IMBE coder is its simple phase model. The fundamental phase is computed as the integral of the linearly interpolated pitch frequency, and the multiples of the fundamental phase are used as the harmonic phases. The effect of this phase model is illustrated in Figure 4.4, the coherent phase model used in IMBE concentrates the speech energy at the phase locations corresponding to the multiples
4.4. Common harmonic coders

4.4.3 Split Band Linear Predictive Coding (SB-LPC)

SB-LPC coder operating a 4 kbps employs time domain LPC filtering and uses a multi band type excitation signal. However the excitation signal of SB-LPC consists of only two bands, separated by a frequency marker, below which the spectrum is declared voiced. The estimation of the frequency marker of SB-LPC is different from the technique used in STC. SB-LPC estimates a voicing decision for each harmonic band using the multi band approach described in section 4.3.1. The estimated voicing decisions are used to determine the voicing frequency marker, which has eight possible equally spaced locations in the spectrum, the first being fully unvoiced and the last being fully voiced. One method to decide the frequency marker is placing it at the end of the last voiced harmonic of the spectrum, i.e. all the voiced harmonics are included in the voiced band of the spectrum. This method eliminates declaring the voiced harmonics as unvoiced which may be perceptually disturbing. A better solution for determining the frequency marker, based on a soft decision process is described in [67].

Pitch is estimated using the sinusoidal speech model matching and refined using the synthetic spectral matching, which are described in section 3.5.2. The harmonic amplitudes are estimated using the equations 4.8 and 4.11 for voiced and unvoiced harmonics.
Table 4.1: Bit allocation of 4 kbps SB-LPC coder for a 20 ms frame

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSFs</td>
<td>23</td>
</tr>
<tr>
<td>Pitch</td>
<td>5+7</td>
</tr>
<tr>
<td>Parity bit</td>
<td>1</td>
</tr>
<tr>
<td>Voicing</td>
<td>3+3</td>
</tr>
<tr>
<td>Gain</td>
<td>5+5</td>
</tr>
<tr>
<td>Harmonic amplitudes</td>
<td>14+14</td>
</tr>
</tbody>
</table>

Figure 4.6: Harmonic speech synthesis, (a) original speech, and (b) SB-LPC respectively, however the LPC residual is used instead of the speech signal. LPC parameters are quantised and interpolated in the LSF domain. The shape of the harmonic amplitudes is vector quantised and the gain is scalar quantised separately. A block diagram of the SB-LPC decoder is shown in Figure 4.5, and the bit allocation is shown in Table 4.1.

Figure 4.6 illustrates the same waveforms shown in Figure 4.4 synthesised using SB-LPC coder. The time domain LPC filter adds its phase response to the coherent excitation signal of SB-LPC and disperses the energy of the excitation pulses. However the waveform shape of the synthesised speech is different from the original speech.
4.5 Concluding remarks

The fundamental sinusoidal speech analysis and synthesis techniques and their application to the low bit rate speech coding have been presented in this chapter. The basic sinusoidal model has been modified to reduce the number of parameters in order to adapt it for low bit rates. At low bit rates the frequencies of the sinusoids are restricted to be harmonics of the pitch frequency, and the harmonic phases are modelled at the decoder. The concept of frequency domain voicing is introduced to achieve a compromise between the hoarseness and buzzyness of harmonically synthesised speech.

Three examples of low bit rate harmonic coders have been presented: Sinusoidal Transform Coding (STC), Improved Multi Band Excitation (IMBE), and Split Band Linear Predictive Coding (SB-LPC). It has been reported that SB-LPC quantised at 2.7 kbps produces better quality than 4.15 kbps IMBE [67], and STC quantised at 2.4 kbps produces similar quality to FS1016 4.8 kbps CELP [64]. One of the main limitations of the low bit rate harmonic coders is that their inability to produce adequate quality at the speech transitions. The reasons for this limitation and the proposed solutions are described in detail in chapter 6.
Chapter 5

Analysis by synthesis coding of speech

5.1 Introduction

Analysis by Synthesis (AbS) techniques may be applied in time domain as well as in frequency domain. This chapter describes the principles of analysis by synthesis coding based on time domain Linear Predictive Coding (LPC), a widely used technique in medium bit rate speech coding during the last two decades, see section 2.5. Combining the AbS techniques and LPC was first introduced by Atal in his Multi Pulse Excited LPC (MPE-LPC) algorithm [68]. The performance of the original MPE-LPC was improved by introducing a second prediction filter, which models the long term correlation present in the speech [69]. Figure 5.1 depicts the general structure of AbS coding schemes with the prediction filters. In AbS coders the decoder is a part of the encoder, as shown by the long dashed rectangle.

The LPC filter is also referred to as Short Term Predictor (STP), since it models the short term correlation present in the speech. The LPC filter coefficients are estimated from the input speech for each frame and interpolated for each sub frame using the methods described in chapter 3. The excitation parameters which minimise the weighted mean square error for each sub frame are selected for transmission. Typically
5.1. Introduction

Figure 5.1: Analysis by synthesis coding with STP and LTP

the frame size is around 10 to 30 ms and there are two or four sub frames in a frame. The excitation signal is given by the sum of the Long Term Prediction (LTP) filter output and the innovation sequence. The perceptual weighting filter is a linear operation and may be moved across the summation operation to the other two branches. In order to reduce the complexity the weighting filter and the LPC synthesis filter may be merged, provided that the Finite Impulse Response (FIR) part of the weighting filter is an inverse LPC filter so that they cancel each other. In practice the excitation parameters are estimated sequentially, which significantly reduces the complexity of the error minimisation process.

Figure 5.2 depicts the sequential parameter estimation algorithm used by the AbS coders. The LPC synthesis filter is an Infinite Impulse Response (IIR) filter, and generates non zero output samples, \( m_{LPC}(n) \) even for a zero input sequence depending on the memory of the IIR filter. This memory response will be added to the speech synthesised by the excitation signal to produce the output speech. Therefore in order to estimate the excitation signal a memory free reference signal, \( t_{LTP}(n) \) is generated. The LTP parameters are optimised on the target signal, \( t_{LTP}(n) \) so that the energy of \( t_{LTP}(n) - m_{LTP}(n) \) is a minimum. In practise a correlation term between \( t_{LTP}(n) \)
and $m_{LTP}(n)$ is maximised and the error signal is computed only with the optimised parameters to generate the innovation target, $t_f(n)$. The innovation sequence, $e_f(n)$ which minimises the energy of $t_f(n) - s_f(n)$ is selected for transmission.

Figure 5.3 illustrates the waveforms corresponding to the analysis by synthesis parameter estimation algorithm depicted in Figure 5.2. The dashed lines indicate the sub frame boundaries. Figure 5.3 (b) illustrates the speech waveform, $\hat{s}(n)$ synthesised at the decoder, see Figure 5.2. In fact $\hat{s}(n)$ is the sum of the waveforms $m_{LPC}$, $m_{LTP}$, and $s_f$ before perceptual weighting. The memory response, $m_{LPC}(n)$ of the LPC filter rises at the beginning of the sub frames and decays during the sub frames, see Figure
5.2. Long Term Prediction (LTP)

5.3 (d). The energy of the LPC memory response increases with the energy and the short term correlation present in the speech. The innovation sequence, \( e_I(n) \) has two pulses per sub frame. The synthesised speech waveform converges towards the original signal, with more accurate quantisation of the excitation signal, i.e. if \( e_I(n) \) can produce \( s_I(n) \) so that \( s_I(n) = t_I(n) \), then \( \hat{s}(n) = s(n) \). The following sections describe the principles behind the long term prediction, innovation structures, and Perceptual weighting in detail.

5.2 Long Term Prediction (LTP)

LPC analysis models the short term correlation or the spectral envelope of speech, see Figure 3.3. The fine harmonic structure of the speech is not modelled by the LPC
5.2. Long Term Prediction (LTP)

Figure 5.4 illustrates this phenomenon in the time domain. The LPC prediction residual is generated by inverse filtering the speech signal, and a significant correlation still remains in the residual signal, especially in the voiced regions. The pulses of the LPC residual are separated by the pitch period, recall the source filter model from section 3.3.1, and see Figure 3.2. This long term correlation can be efficiently modelled by a filter with a long buffer [69], which is called the long term predictor or the pitch predictor, and the buffer is referred to as an adaptive codebook, see Figure 5.5. During the voiced speech, the delay, $D$ is usually equal to the pitch period. However $D$ may be a multiple of the pitch period, hence the term long term predictor is preferred. Figure 5.3 (g) illustrates the LTP filter output, which looks similar to the LPC residual signal. Figure 5.5 depicts the structure of a recursive LTP filter, in which when $D$ is shorter
5.2. Long Term Prediction (LTP)

than the sub frame length recursion occurs within sub frames [70]. This gives rise to non	
linear equations in the error minimisation process, which are computationally expensive
to solve. Therefore the LTP filter is usually implemented as a First In First Out (FIFO)
buffer, which is updated only after each sub frame by shifting a block of samples equal
to the sub frame length, and inserting the sum of the innovation sequence and the LTP
output of the sub frame at the end. The LTP filter output for any sub frame depends
only on the memory of the filter, and during the optimisation process the innovation
sequence is set to a sequence of zeros as depicted in Figure 5.2. For delays shorter than
the sub frame length the last $D$ samples of the buffer are periodically repeated. After
the above simplifications $\hat{r}(n)$ can be defined as follows:

if $D \geq L$

$$\hat{r}(n) = e_{LTP}(n - D) \quad \text{for } 0 \leq n < L$$  

(5.1)

if $D < L$

$$\hat{r}(n) = e_{LTP}(n - D) \quad \text{for } 0 \leq n < D$$

$$\hat{r}(n) = \hat{r}(n - kD) \quad \text{for } kD \leq n < L'$$  

(5.2)

Where $L' = \min [(k + 1)D, L], k = 1, 2, ..., \lfloor L/D \rfloor$, and $L$ is the sub frame length.
5.2. Long Term Prediction (LTP)

Deriving the optimum delay, $D$ and gain, $\beta$

The optimum delay, $D$ and gain, $\beta$ are derived by minimising the energy, $E_D$ of $t_{LTP}(n) - m_{LTP}(n)$ over the sub frame length, $L$.

Where

$$E_D = \sum_{n=0}^{L-1} [t_{LTP}(n) - m_{LTP}(n)]^2$$

(5.3)

Where

$$m_{LTP}(n) = h_W(n) * \beta \hat{r}(n) = \beta z(D)$$

(5.4)

Where $h_W(n)$ is the impulse response of the weighted LPC filter, evaluated for $L$ samples, since the result of the convolution is required only for the samples within the
sub frame. For a particular $D$, the gain, $\beta_D$ which minimises $E_D$ is found by setting $\partial E_D/\partial \beta = 0$, and given by,

$$\beta_D = \frac{\sum_{n=0}^{L-1} t_{LTP}(n)z(D)}{\sum_{n=0}^{L-1} z^2(D)}$$  \hspace{1cm} (5.5)

Substituting $\beta_D$ in equation 5.3 gives,

$$E_D = \sum_{n=0}^{L-1} t_{LTP}^2(n) - \left[ \sum_{n=0}^{L-1} t_{LTP}(n)z(D) \right]^2 \sum_{n=0}^{L-1} z^2(D)$$  \hspace{1cm} (5.6)

For a given sub frame the LTP target, $t_{LTP}(n)$ is a constant. Therefore the optimum $D$ is found by maximising the second term in equation 5.6 with respect to $D$, and the optimum $\beta$ is obtained by substituting the optimum $D$ in equation 5.5.

In order to reduce the high computational complexity of the delay search, it may be restricted around an initial pitch estimate, see section 3.5. The performance of the LTP filter may be improved by using a multiple tap filter [71] or fractional delays [72], especially for stationary voiced segments. For example a three tap LTP filter has three coefficients, $\beta_j$ corresponding to the delays $D + j$, for $j = \{-1, 0, 1\}$.

5.3 Innovation structures

The innovation sequence serves two main purposes [2]:

1. Providing the start up information of the LTP memory, this includes any sudden changes in the speech waveform not adequately tackled by the LTP.

2. Providing the filling in information that the LTP has omitted, especially during the unvoiced segments.

The role of the innovation sequence at the onsets is depicted in Figure 5.3. At the onsets the content of the LTP memory is not adequate to provide the excitation signal, and the innovation pulses synthesise the pitch pulses. Consequently the LTP memory builds up with the new pitch information as the the buffer is updated with the new
excitation signal, and the LTP filter becomes more efficient. The difference in most of
the variants of AbS LPC based coders described in sections 2.4 and 2.5 comes in the
structure of the innovation sequence. However the derivation of the optimum innovation
sequence, \( e_I(n) \) remains the same [73], even though the parameter estimation process
can be optimised for reduced complexity by exploiting the specific features of a given
innovation structure. Let \( e_I(n) = g_I x_k(n) \), the optimum gain, \( g_I \) and the innovation
vector, \( x_k(n) \) are derived by minimising the energy, \( E_k \) of \( t_I(n) - s_I(n) \) over the sub
frame length, \( L \). Where

\[
E_k = \sum_{n=0}^{L-1} [t_I(n) - s_I(n)]^2
\]

(5.7)

Where

\[
s_I(n) = h_w(n) * g_I x_k(n) = g_I \delta_k(n)
\]

(5.8)

Where \( h_w(n) \) is the impulse response of the weighted LPC filter, evaluated for \( L \)
samples, since the result of the convolution is required only for the samples within the
sub frame. For a particular \( x_k(n) \), the gain, \( g_I \) which minimises \( E_k \) is found by setting
\( \partial E_k / \partial g_I = 0 \), and given by,

\[
g_I = \frac{\sum_{n=0}^{L-1} t_I(n) \delta_k(n)}{\sum_{n=0}^{L-1} \delta_k^2(n)}
\]

(5.9)

Substituting \( g_I \) in equation 5.7 gives,

\[
E_k = \sum_{n=0}^{L-1} t_I^2(n) - \frac{\left[ \sum_{n=0}^{L-1} t_I(n) \delta_k(n) \right]^2}{\sum_{n=0}^{L-1} \delta_k^2(n)}
\]

(5.10)

For a given sub frame the innovation target, \( t_I(n) \) is a constant. Therefore optimum
\( x_k(n) \) is found by maximising the second term in equation 5.10 with respect to \( k \),
and the optimum \( g_I \) is obtained by substituting the optimum sequence in equation
5.9. The derivation of the optimum parameters of the single tap LTP filter and the
innovation sequence are similar except the fact that LTP uses an adaptive codebook
with the variable \( D \) and the innovation sequence is derived from a fixed codebook with
the variable \( k \).
5.3. Innovation structures

The coding bit rate and the complexity may be reduced by limiting the possible $x_k(n)$, at the expense of a reduction in the speech quality. The high computational complexity is one of the main disadvantages of AbS LPC schemes. In order to evaluate the second term in equation 5.10 with respect to $k$, each $x_k(n)$ should be convolved with $h_W(n)$ to obtain $s_k(n)$ and the cross correlation and the auto correlation terms should be evaluated. Hence designing the fixed codebook structures [35] and efficient search procedures [74] for reducing the complexity is an important issue.

Multi pulse and regular pulse innovation

Multi Pulse Excited LPC (MPE-LPC) [68] innovation structure is defined by,

$$e_I(n) = \sum_{j=0}^{M-1} a_j \delta(n - n_j) \quad (5.11)$$

Where $a_j$ are the pulse amplitudes, $n_j$ are the pulse positions, $M$ is the number of pulses per sub frame, and $\delta(0)$ represents a unit impulse placed at the beginning of the sub frame. The positions of the pulses are not constrained and can be placed at any sample location of the sub frame. The complexity of a joint optimisation of all the possible $a_j$ and $n_j$ is prohibitive. Therefore the pulses are estimated sequentially, by subtracting the contribution of each pulse to obtain the target, of the next pulse. The amplitudes may be reoptimised once the pulse positions are known [69].

Regular Pulse Excited LPC (RPE-LPC) [37] innovation structure is defined by,

$$e_I(n) = \sum_{j=0}^{M-1} a_j \delta(n - jL - i) \quad (5.12)$$

Where $a_j$ are the pulse amplitudes and $M$ is the number of pulses. The pulse locations are given by $jL + i$, where $L$ is the spacing between the pulses and $i$ is estimated for each sub frame, which defines the location of the first pulse, in the range, $0 \leq i < L - 1$ The pulse amplitudes, $a_j$ are jointly optimised for each grid, $i$. The grid and the amplitude vector which minimise the weighted square error are selected for transmission. Algebraic Code Excited Linear Prediction restricts the magnitude of all the pulses to be equal in addition to restricting their locations, see section 5.5.
Code Excited Linear Prediction (CELP)

Code Excited Linear Prediction (CELP) [10] selects the innovation vector, $x_k(n)$ from a codebook stored in memory, similar to vector quantisation. Usually the codewords are populated with Gaussian noise. This may be justified by considering the fact that when all the correlations are removed from the speech signal, the prediction residual signal has a distribution similar to Gaussian. However the synthetic speech quality becomes poor at the transitions, especially at the onsets, since the prediction filters can not track the features of the non stationary waveforms, and the random innovation vectors are inadequate to model the prediction residual.

The computational complexity of CELP is higher than the sparse pulse innovation structures. CELP needs to compute the convolution in equation 5.8 for all the samples in each vector. However for the sparse pulse sequences the convolution is given by the sum of the scaled, delayed impulse responses, with each impulse response corresponding to a pulse. The complexity of CELP can be reduced by using overlapped codebooks [75] or Vector Sum Excited Linear Prediction (VSELP) [29].

5.4 Perceptual weighting

The mean square error minimisation process produces a flat quantisation noise spectrum. However at low bit rates the quantisation noise energy may exceed the speech energy at the spectral valley regions. The ear is frequency selective and sensitive to those high noise spectral components, making them perceptually disturbing. It is well known that a louder signal can reduce the audibility of a weaker signal, which is called masking [44]. The analysis by synthesis coders exploit the masking property of the ear in the frequency domain. The audibility of the quantisation noise can be reduced by spectrally shaping the noise spectrum so that the noise energy is always below the speech energy. The noise energy is redistributed by filtering the error signal prior to mean square error minimisation, see Figure 5.1. The filtering operation places most of the quantisation noise in the spectral regions where the speech signal has formant peaks, and places less noise in the spectral valley regions.
5.4. Perceptual weighting

The perceptual weighting filter used by the analysis by synthesis coders is based on the LPC spectral envelope [76]. The transfer function of the weighting filter \( W(z) \) is given by,

\[
W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)} \quad \text{for} \quad 0 < \gamma_2 < \gamma_1 \leq 1
\]  

(5.13)

Where

\[
A(z/\gamma) = 1 - \sum_{j=1}^{P} \alpha_j \gamma^j z^{-j}
\]  

(5.14)

Where \( \alpha_j \) are coefficients of the LPC filter. The weighting factors, \( \gamma_1 \) and \( \gamma_2 \) should be tuned for each coder by means of subjective listening tests. A commonly used set of values is \( \gamma_1 = 1.0 \) and \( \gamma_2 = 0.8 \). Figure 5.6 depicts the LPC spectral envelope and the magnitude spectra of the weighting filter, with \( \gamma_1 = 1.0 \) and different \( \gamma_2 \) values. The \( \gamma_2 \) value used is shown against each curve. At the formant peaks the weighting filter de-emphasises the quantisation noise presented to the error minimisation process, however the mean square error minimisation process produces a flat noise spectrum.

Figure 5.6: Magnitude spectra of the weighting filter for a given LPC spectral envelope
5.5. **Algebraic Code Excited Linear Prediction (ACELP)**

In order to compensate for this the noise spectrum of the synthesised speech, i.e. the spectrum of \( s(n) - \hat{s}(n) \) in Figure 5.1 has an inverse relationship to the spectral shape of the weighting filter. The weighting filter is moved across the summation operation as illustrated in Figure 5.2, and usually left as a cascaded filter with the LPC synthesis filter. This offers the flexibility of using unquantised LPC parameters and tuning of \( \gamma_t \) to a value other than unity. The use of unquantised LPC coefficients in the weighing filter gives better results, and it should be pointed out that the weighting filter is only a part of the encoder, see Figure 5.1. However the computational complexity is still reduced since the impulse response of the cascaded filter, \( h_W(n) \) used in the error minimisation process can be computed once and stored for all the convolution operations, see equations 5.4 and 5.8.

### 5.5 Algebraic Code Excited Linear Prediction (ACELP)

Algebraic Code Excited Linear Prediction (ACELP) reduces the complexity of the optimum innovation sequence search by utilising a technique called backward filtering [77] and exploiting the structure of the innovation codebook [34]. Recall that the optimum innovation sequence is found by maximising the second term of equation 5.10. Excluding the square, the numerator, \( C_k^2 \) of the second term of equation 5.10 is given by,

\[
C_k = \sum_{n=0}^{L-1} t_I(n) \hat{s}_k(n) \tag{5.15}
\]

Where

\[
\hat{s}_k(n) = \sum_{m=0}^{L-1} h_W(n - m) x_k(m) \tag{5.16}
\]

i.e. the convolution of the candidate innovation sequence, \( x_k(n) \), and the impulse response of the cascaded LPC filter and the weighting filter, \( h_W(n) \). Substituting \( \hat{s}_k(n) \) in equation 5.16 into equation 5.15 gives,

\[
C_k = \sum_{n=0}^{L-1} t_I(n) \sum_{m=0}^{L-1} h_W(n - m) x_k(m) \tag{5.17}
\]
5.5. **Algebraic Code Excited Linear Prediction (ACELP)**

Which may be rearranged as,

\[
C_k = \sum_{m=0}^{L-1} x_k(m) \sum_{n=0}^{L-1} t_I(n) h_W(n-m)
\]  

\[
C_k = \sum_{m=0}^{L-1} x_k(m)c(m)
\]  

(5.18)

(5.19)

Where

\[
c(m) = \sum_{n=0}^{L-1} t_I(n) h_W(n-m)
\]  

(5.20)

c(m) is similar to a convolution operation, however neither of the terms are reversed. Therefore \(c(m)\) is a cross correlation operation between the target signal, \(t_I(n)\) and the weighted impulse response, \(h_W(n)\) delayed by \(m\) samples. It is not difficult to see that \(c(m)\) is independent of \(x_k(n)\), \(t_I(n)\) and \(h_W(n)\) are constant signals for a given sub frame. Therefore \(c(m)\) can be computed once and use to evaluate \(C_k\) for all the iterations. This process is called backward filtering, and reduces the computational load by a factor approximately equal to the LPC order [77].

ACELP restricts the pulses of \(x_k(n)\) to either 1 or \(-1\), which translates to equal magnitudes in the innovation sequence, \(e_I(n)\) after multiplying by the gain, \(g_I\). Therefore equation 5.19 reduces to,

\[
C_k = \sum_{p=0}^{P-1} a_p c(m_p)
\]  

(5.21)

Where \(a_p\) is the sign of the pulse at the location \(m_p\) and \(P\) is the number of pulses. The search complexity is further reduced by setting the sign of a pulse at a certain position, \(m_p\) a priori to the sign of \(c(m_p)\). This choice of signs for a given combination of pulses maximises \(C_k\). The denominator, \(\varepsilon_k\) of equation 5.10 is given by,

\[
\varepsilon_k = \sum_{n=0}^{L-1} \varepsilon_k^2(n) = \sum_{n=0}^{L-1} \left( \sum_{m=0}^{L-1} h_W(n-m) x_k(m) \right)^2
\]  

(5.22)

\[
\varepsilon_k = \sum_{n=0}^{L-1} \sum_{p=0}^{P-1} a_p h_W(n-m_p) \left( \sum_{p=0}^{P-1} a_p h_W(n-m_p) \right)^2
\]  

(5.23)
5.5. Algebraic Code Excited Linear Prediction (ACELP)

\[ \varepsilon_k = \sum_{n=0}^{L-1} \left[ a_0 h_W(n - m_0) + a_1 h_W(n - m_1) + \ldots + a_{P-1} h_W(n - m_{P-1}) \right]^2 \]  \hspace{1cm} (5.24)

\[ \varepsilon_k = \sum_{n=0}^{L-1} \left[ \sum_{p=0}^{P-2} \sum_{q=p+1}^{P-1} a_p a_q h_W(n - m_p) h_W(n - m_q) \right] \] \hspace{1cm} (5.25)

\[ \varepsilon_k = \sum_{n=0}^{L-1} \sum_{p=0}^{P-2} \sum_{q=p+1}^{P-1} a_p a_q \sum_{n=0}^{L-1} h_W(n - m_p) h_W(n - m_q) \] \hspace{1cm} (5.26)

\[ \varepsilon_k = \sum_{p=0}^{P-1} \Phi_{p,p} + 2 \sum_{p=0}^{P-2} \sum_{q=p+1}^{P-1} a_p a_q \Phi_{p,q} \] \hspace{1cm} (5.27)

Where

\[ \Phi_{p,q} = \sum_{n=0}^{L-1} h_W(n - m_p) h_W(n - m_q) \] \hspace{1cm} (5.28)

Equation 5.27 is composed of sums of the cross correlation terms of \( h_W(n) \) with delays corresponding to the pulse locations. Consequently the matrix of correlations of \( h_W(n) \), \( \Phi_{i,j} \) may be precomputed and stored for all the possible delays, \( 0 \leq m_i, m_j < L - 1 \) considering that, \( \Phi_{i,j} = \Phi_{j,i} \).

Selecting \( P \) pulses from any location of a sub frame of length \( L \) gives \( L C_P \) different \( x_k(n) \) sequences, and the number of searches may be reduced by restricting the possible pulse locations. RPE-LPC restricts all the pulses of a sub frame to be in the same grid, however ACELP selects a pulse from each grid, allowing pulses to be placed at the adjacent locations. Table 5.1 shows the pulse position combinations used in 8 kbps G.729 CS-ACELP coder. Another method to reduce the number of searches is to restrict the locations to those which have the highest \( c(m) \) values, see equation 5.20. For example when 1 pulse should be chosen out of 8 locations, the 4 locations which have the highest \( c(m) \) may be included in the search loop.

5.5.1 Conjugate Structure ACELP (CS-ACELP) ITU G.729 coder

The toll quality Conjugate Structure Algebraic Code Excited Linear Prediction (CS-ACELP) [34] ITU G.729 coder uses a frame length of 10 ms, and a look ahead of 5
ms for linear prediction analysis. LP analysis is performed once per frame using the autocorrelation method with a 30 ms asymmetric window. Each frame has two sub frames, the second sub frame uses the estimated LP parameters and the first sub frame uses a parameter set obtained by linear interpolation of the LSFs of the adjacent sub frames.

A single tap Long Term Predictor (LTP) with a fractional resolution of 1/3 sample accuracy, and the fixed codebook structure with four pulses per sub frame described in section 5.5 is used, see Table 5.1. Since G.729 transmits a sign bit for each pulse, \( c(m_i) \) in equation 5.21 and \( \Phi_{i,j} \) in equation 5.27 are precomputed with the corresponding sign, as follows:

\[
C_k = \sum_{p=0}^{3} c^\prime(m_p) \tag{5.29}
\]

Where

\[
c^\prime(m_i) = |c(m_i)| \tag{5.30}
\]

And

\[
\varepsilon_k/2 = \Phi_{0,0} + \Phi_{1,1}^\prime + \Phi_{2,2}^\prime + \Phi_{3,3}^\prime + \Phi_{0,1}^\prime + \Phi_{0,2}^\prime + \Phi_{0,3}^\prime + \Phi_{1,2}^\prime + \Phi_{1,3}^\prime + \Phi_{2,3}^\prime \tag{5.31}
\]

\[
\Phi_{i,j} = sign[c(m_i)] sign[c(m_j)] \Phi_{i,j} \quad \text{and} \quad \Phi_{i,i}^\prime = 0.5\Phi_{i,i} \tag{5.32}
\]

This method is not suitable when the signs of all the pulses are not transmitted explicitly. For example some schemes transmit only the signs of a sub set of the pulses and the rest are determined from the transmitted ones, in order to reduce the bit rate.
5.5. *Algebraic Code Excited Linear Prediction (ACELP)*

Table 5.2: Bit allocation of G.729 CS-ACELP for a 10 ms frame

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSFs</td>
<td>18</td>
</tr>
<tr>
<td>LTP delay index</td>
<td>8+5</td>
</tr>
<tr>
<td>Parity bit</td>
<td>1</td>
</tr>
<tr>
<td>Pulse locations</td>
<td>13+13</td>
</tr>
<tr>
<td>Pulse signs</td>
<td>4+4</td>
</tr>
<tr>
<td>Codebook gains stage 1</td>
<td>3+3</td>
</tr>
<tr>
<td>Codebook gains stage 2</td>
<td>4+4</td>
</tr>
</tbody>
</table>

The fixed codebook gain is predicted using a fourth order MA predictor based on the log energy of the previous fixed codebook contributions. The adaptive codebook gain, \( \beta \) and a correction factor, \( \gamma \) of the fixed codebook gain, \( g_I \) are vector quantised using a two stage conjugate structured codebook [78], by minimising the mean squared error, \( E \) given by,

\[
E = \sum_{n=0}^{L-1} [t_{LTP} - \beta z(D) - g_I \hat{s}_k(n)]^2
\]  

(5.33)

Where \( g_I = \gamma g_{I_p} \) and \( g_{I_p} \) is the predicted fixed codebook gain. \( t_{LTP} \), \( z(D) \), and \( \hat{s}_k(n) \) are given in equations 5.3, 5.4, and 5.8 respectively, see also Figure 5.2.

Where

The bit allocation of G.729 coder is shown in Table 5.2. The LTP delay index is encoded with 8 bits in the first sub frame and differentially encoded with 5 bits in the second sub frame. In order to improve the robustness against random bit errors a parity bit is computed on the six most significant bits of the LTP delay index of the first sub frame. If a parity error is detected, the LTP delay is set to the integer part of the delay of the second sub frame of the previous frame.

An error concealment procedure is applied in the decoder to reduce the degradation in the reconstructed speech when frame erasures are detected in the bit stream. The error concealment technique replaces the missing LPC synthesis filter parameters and the excitation signal, based on previously received information. The excitation signal
is replaced with one of similar characteristics, while gradually decaying its energy.

5.6 Concluding remarks

The time domain Analysis by Synthesis (AbS) coders employ Linear Predictive Coding (LPC) and Long Term Prediction (LTP) to model the short term and long term correlations present in the speech. The prediction residual is quantised using the waveform matching techniques. The most popular Code Excited Linear Prediction (CELP) vector quantises the prediction residual. However the quantisation is performed closed loop using the synthesis filters and comparing the synthesised speech with the input speech signal. Perceptual weighting is used in the error minimisation process in order to mask the quantisation noise below the signal level. In order to reduce the computational complexity the LPC, LTP, and innovation parameters are extracted sequentially.

AbS coders were widely used in the medium bit rate speech coding during the last two decades, and many of them became standards. Most of the AbS coding schemes differ in the way they produce the innovation sequence. Multi Pulse Excited LPC (MPE-LPC) and Regular Pulse Excited LPC (RPE-LPC) use innovation structures based on sparse pulses. Algebraic Code Excited Linear Prediction (ACELP) reduces the search complexity by using innovation structures with sparse pulses of either 1 or −1. The ITU standard G.729 is based on a toll quality 8 kbps Conjugate Structure ACELP (CS-ACELP) coder. The AbS coders are inherently robust against different types of input signals such as music and background noise, due to their waveform matching capability.
Chapter 6

Integration of harmonic and analysis by synthesis coders

6.1 Introduction

Harmonic coders extract the speech parameters in the frequency domain. Synthesised speech is generated as a sum of sinusoids. They produce highly intelligible speech down to about 2.4 kbps [64]. By using the unquantised phases and amplitudes, and by frequent updating of the parameters, i.e. at least every 10 ms, they can even achieve transparent quality [6]. However this requires a prohibitive bit rate, unsuitable for low bit rate applications. The original multi band excitation (MBE) coder operating at 8 kbps transmits the phase information [63]. Harmonic coders operating at 4 kbps and below do not transmit phase information. The spectral magnitudes are transmitted, typically every 20 ms and interpolated in the synthesis. This simplified version used for low bit rate applications suits well for stationary voiced segments. However at the speech transitions such as onsets, where the speech waveform changes rapidly, the simplifications do not hold and degrade the perceptual quality.

Figure 6.1 demonstrates two examples of harmonically synthesised speech, (a) shows a stationary voiced segment and (b) shows a transitory speech segment. In both cases (i) represents the original speech, i.e. 128 kbps linear pulse code modulation, and
6.1. Introduction

(ii) represents the synthesised speech. The synthesised speech is generated using the Split Band Linear Predictive Coding (SB-LPC) harmonic coder operating at 4 kbps [5] described in section 4.4.3. The synthesised waveforms are shifted in the figures in order to compensate for the delay due to look ahead and the linear phase deviation due to loss of phase information in the synthesis. The SB-LPC decoder predicts the evolution of harmonic phases using the linearly interpolated fundamental frequency, i.e. a quadratic phase evolution function. The low bit rate harmonic coders can not preserve waveform similarity as illustrated in the figures, since the phase information is not transmitted. However at the stationary voiced segments phase information has little importance in terms of the perceptual quality of the synthesised speech. Stationary voiced speech has a strong, slowly evolving harmonic content. Therefore the low bit rate harmonic synthesis model of extracting the speech parameters in the frequency domain at regular intervals and interpolating them in the synthesis suits well for stationary voiced segments. However at the transitions, where the speech waveform evolves rapidly this low bit rate harmonic model fails. As depicted in Figure 6.1 (b), the highly non stationary character of the transition has been smeared by the low bit rate harmonic model. This effect reduces the intelligibility of the synthesised speech.

On the other hand analysis by synthesis coders like ACELP [34], [35] encode the tar-
6.1. Introduction

Figure 6.2 shows a stationary voiced segment and (b) shows a transitory segment synthesised using ACELP at 4 kbps. In Figure 6.2 (a) the consecutive pitch cycles have different shapes, which destroys the slowly evolving periodicity of voiced speech, as opposed to Figure 6.1 (a). Therefore despite the fact that waveform similarity is less in Figure 6.1 (a), harmonically synthesised voiced speech is perceptually superior to waveform coded speech at low bit rates. Figure 6.2 (b) shows that ACELP can synthesise the highly non stationary speech transitions better than the harmonic coders, as opposed to Figure 6.1 (b). ACELP may introduce granular noise at the transitions.
as well. However at the transitions the speech waveform changes rapidly, masking the granular noise of ACELP, which is not perceptible down to about 4 kbps.

The above observations suggest a hybrid coding approach, which selects the optimum coding algorithm for a given segment of speech. For example coding of stationary voiced segments using harmonic coding and transitions using ACELP. Unvoiced and silence segments can be encoded with CELP [10] or white noise excitation, see section 6.5.5.

Harmonic coders suffer from other potential problems such as voicing and pitch errors that may occur at the transitions. The pitch estimates at the transitions, especially at the onsets may be unreliable due to the rapidly changing speech waveform. Furthermore, pitch tracking algorithms do not have the history at the onsets, and should be turned off. Inaccurate pitch estimates also account for inaccurate voicing decisions, in addition to the spectral mismatches due to the non stationary speech waveform at the transitions. These voicing decision errors declare the voiced bands as unvoiced and increase the hoarseness of synthetic speech. Encoding the transitions using ACELP eliminates those potential problems of harmonic coding.

6.2 Challenges in designing a hybrid coder

The main challenges in designing a hybrid coder are reliable speech classification and phase synchronisation when switching between the coding modes. Furthermore, most of the speech coding techniques make use of a look ahead and parameter interpolation. Interpolation requires the parameters of the previous frame, when switched from a different mode those parameters may not be directly available. Predictive quantisation schemes also require the previous memory. The techniques developed to eliminate those memory initialisation problems are discussed in sections 6.5.4 and 7.3.2.

6.2.1 Reliable speech classification

A Voice Activity Detector (VAD) can be used to identify speech and silence segments [79], while classification of speech into voiced and unvoiced segments can be seen as
the simplest speech classification technique. However there are coders in the literature, which use up to six phonetic classes [80]. The design of such a phonetic classification algorithm can be complicated and computationally complex, and a simple classification with two or three modes is sufficient to exploit the relative merits of waveform coding and harmonic coding.

The accuracy of the speech classification is critical for the performance of a hybrid coder. For example, using the noise excitation for a stationary voiced segment can severely degrade the speech quality, by converting the high voiced energy of the original speech into noise in the synthesised speech. While use of the harmonic excitation for unvoiced segments gives a tonal artifact, ACELP can maintain acceptable quality for all the types of speech. In section 6.6 a robust, low complexity two stage classification algorithm, which eliminates the above difficulties is described.

### 6.2.2 Phase synchronisation

Harmonic coders operating at 4 kbps and below do not transmit phase information, in order to allocate the available bits for accurate quantisation of the more important spectral magnitude information. They exploit the fact that the human ear is partially phase deaf. The waveform shape of the synthesised speech can be very different from the original speech, often yielding negative SNRs. However the analysis by synthesis coders preserve the waveform similarity. Therefore direct switching between those two modes without any precautions will severely degrade the speech quality due to the phase discontinuities.

### 6.3 The existing hybrid coders

The hybrid coding concept has been introduced in the LPC vocoder [42], which classifies speech frames into voiced or unvoiced, and synthesises the excitation using periodic pulses or white noise respectively. Analysis by synthesis CELP coders with Dynamic Bit Allocation (DBA), which adaptively distribute the bits among coder parameters in a given frame while maintaining a constant bit rate, by classifying each frame into a
certain mode are also reported [81]. However the research carried out for this thesis is particularly focused on the hybrid coders, which combine analysis by synthesis coding and harmonic coding. The advantages and disadvantages of harmonic coding and CELP, and the potential benefits of combining the two methods has been discussed by Trancoso et al. [82]. Improving the speech quality of the LPC vocoder by using a form of multi pulse excitation [83], as a third excitation model at the transitions has also been reported [84].

1. Prototype Waveform Interpolation (PWI) coder

Kleijn introduced Prototype Waveform Interpolation (PWI) in order to improve the quality of voiced speech [13]. The PWI technique extracts prototype pitch cycle waveforms from the voiced speech at regular intervals of 20-30 ms. Speech is reconstructed by interpolating the pitch cycles between the update points. The PWI technique can be applied either directly on the speech signal or the LPC residual. Since the PWI technique does not suite to encode the unvoiced speech segments, unvoiced speech is synthesised using CELP. Even though the motivation behind using two coding techniques is different in the PWI coder, i.e. waveform coding is not used for transitions, it combines harmonic coding and analysis by synthesis coding. The speech classification of the PWI coder is relatively easier, since it only needs to classify speech into either voiced or unvoiced.

At the onset of a voiced section, the past estimated prototype waveform is not present at the decoder for the interpolation process. Kleijn suggests three methods to solve this problem:

1. Extracting the prototype waveform from the reconstructed CELP waveform of the previous frame
2. Set to a single pulse waveform (filtered through LPC) with its amplitude determined from the transmitted information
3. Use a replica of the prototype transmitted at the end of the synthesis frame

The starting phase of the pitch cycles at the onsets can be determined at the decoder from the CELP encoded signal. At the offsets the linear phase deviation between
the harmonically synthesised and original speech is measured and the original speech buffer is displaced, such that the analysis by synthesis coder begins exactly where the harmonic coder has ended.

2. Combined harmonic and waveform coding of speech at low bit rates

This coder proposed by Shlomot et al. consists of three modes: harmonic, transition, and unvoiced [14], [85]. All the modes are based on the source filter model. The harmonic mode consists of two components: the lower part of the spectrum or the harmonic bandwidth, which is synthesised as a sum of coherent sinusoids, and the upper part of the spectrum, which is synthesised using sinusoids of random phases. The transitions are synthesised using a pulse excitation, similar to ACELP and the unvoiced segments are synthesised using white noise excitation.

The speech classification is performed by a neural network, which takes into account the speech parameters of the previous, current, and future frames, and the previous mode decision. The classification parameters include speech energy, spectral tilt, zero crossing rate, residual peakiness, residual harmonic matching SNRs, and pitch deviation measures. At the onsets when switching from the waveform coding mode, the harmonic excitation is synchronised by shifting and maximising the cross correlation with the waveform coded excitation. At the offsets the waveform coding target is shifted to maximise the cross correlation with the harmonically synthesised speech, similar to the PWI coder.

3. A 4 kbps hybrid MELP/CELP coder with alignment phase encoding and zero phase equalisation

This coder proposed by Stachurski et al. consists of three modes: strongly voiced, weakly voiced, and unvoiced [15], [86]. The weakly voiced mode includes transitions and plosives, it is used when neither strong voiced nor unvoiced speech segments are clearly identified. In the strongly voiced mode Mixed Excitation Linear Prediction (MELP) [87], [88] coder is used. Weakly voiced and unvoiced modes are synthesised using CELP. In unvoiced frames the LPC excitation is generated from a fixed stochastic
6.3. The existing hybrid coders

codebook. In weakly voiced frames the LPC excitation consists of the sum of a long term prediction filter output and fixed innovation sequence containing a limited number of pulses similar to ACELP.

The speech classification is based on estimated voicing strength and pitch. The signal continuity at the mode transitions is preserved by transmitting “alignment phase” for MELP [87] encoded frames, and by using “zero phase equalisation” for transitional frames. The alignment phase preserves the time synchrony between the original and synthesised speech. Alignment phase is estimated as the linear phase required in the MELP encoded excitation generation to maximise the cross correlation between the MELP excitation and the corresponding LPC residual. Zero phase equalisation modifies the CELP target signal, in order to reduce the phase discontinuities, by removing the phase component, which is not coded in MELP. Zero phase equalisation is implemented in the LPC residual domain, with a Finite Impulse Response (FIR) filter similar to [89]. The FIR filter coefficients are derived from the smoothed pitch pulse waveforms of the LPC residual signal. For unvoiced frames the filter coefficients are set to an impulse so that the filtering has no effect. The analysis by synthesis target is generated by filtering the zero phase equalised residual signal through the LPC synthesis filter.

6.3.1 Limitations of the existing hybrid coders

Both 1 and 2 use similar methods to ensure signal continuity. At the onsets the initial phases of the harmonic excitation are extracted from the previous excitation vector of the waveform coding mode. This can be difficult at some of the rapidly varying onsets, especially if the bit rate of the waveform coder is low. Moreover, inaccuracies in the onset synchronisation will propagate through the harmonic excitation and make the offset synchronisation more difficult. At the offsets the linear phase deviation between the harmonically synthesised and original speech is measured and the original speech buffer is displaced, such that the analysis by synthesis coder begins exactly where the harmonic coder has ended. This method needs resetting of the accumulated displacement during unvoiced or silent segments, and may fail to meet the specifications of a system with strict delay requirements. Another problem arises when a transition occurs
6.3. The existing hybrid coders

Figure 6.3: A transition within voiced speech

within a voiced speech segment as shown in Figure 6.3, where there are no unvoiced or silent segments after the transition to reset the accumulated displacement. Even though the accumulated displacement can be minimised by inserting or eliminating exactly complete pitch cycles, the remainder will propagate into the next harmonic section. Furthermore, a displacement of a fraction of a sample can introduce audible high frequency distortion, especially in segments with short pitch periods, consequently the displacements should be performed with a high resolution.

The MELP/CELP coder preserves signal continuity by transmitting “alignment phase” for MELP encoded frames, and use of “zero phase equalisation” for transitional frames. Zero phase equalisation may reduce the benefits of the use of analysis by synthesis coding by modifying the phase spectrum, and it has been reported that the phase spectrum is perceptually important [90], [91], [92]. Furthermore, zero phase equalisation relies on accurate pitch pulse position detection at the transitions, which can be difficult.

Harmonic excitation can be synchronised with the LPC residual by transmitting the phases, which eliminates the above difficulties. However this requires a prohibitive capacity making it unsuitable for low bit rate applications. As a compromise in this chap-

amplitude

0 200 400 600
samples

s(n)

r(n)
6.4 Synchronised Waveform matched Phase Model (SWPM)

After a new phase model for the harmonic excitation, Synchronised Waveform matched Phase Model (SWPM) is presented. SWPM facilitates the integration of harmonic and analysis by synthesis coders, by synchronising the harmonic excitation with the LPC residual. SWPM requires only two parameters, and does not alter the perceptual quality of the harmonically synthesised speech. It also allows the ACELP mode to target the speech waveform without modifying the perceptually important phase components or the frame boundaries.

6.4 Synchronised Waveform matched Phase Model (SWPM)

SWPM maintains the time synchrony between the original and the harmonically synthesised speech by transmitting the pitch pulse location (PPL) closest to each synthesis frame boundary. SWPM also preserves sufficient waveform similarity, such that the switching between the coding modes is transparent, by transmitting a phase value, which indicates the pitch pulse shape (PPS) of the corresponding pitch pulse. PPL and PPS are estimated for every frame of 20 ms. SWPM needs to detect only the pitch pulses in the stationary voiced segments, which is somewhat easier than detecting the pitch pulses in the transitions as in [15].

SWPM has the disadvantage of transmitting two extra parameters, PPL and PPS, but the bottleneck of the bit allocation of hybrid coders is usually in the waveform coding mode. Furthermore, in stationary voiced segments the location of the pitch pulses can be predicted with high accuracy, and only an error needs to be transmitted. The same argument applies to the shape of the pitch pulses.

In the harmonic synthesis, cubic phase interpolation [6] is applied between the pitch pulse locations, setting the phases of all the harmonics equal to PPS. This makes the waveform similarity between the original and the synthesised speech highest in the vicinity of the selected pitch pulse locations. However this does not cause difficulties, since switching is restricted to frame boundaries and the pitch pulse locations closest to the frame boundaries are selected. Furthermore, SWPM can synchronise the synthesised excitation and the LPC residual with fractional sample resolutions, even without up sampling either of the waveforms.
6.4. Synchronised Waveform matched Phase Model (SWPM)

6.4.1 Extracting the Pitch Pulse Location (PPL)

The Telecommunications Industry Association (TIA) Enhanced Variable Rate Coder (EVRC) [93], which employs Relaxed CELP (RCELP) [94] uses a simple method based on the energy of the LPC residual to detect the pitch pulses. EVRC determines the pitch pulse locations by searching for a maximum in a 5 sample sliding energy window within a region larger than the pitch period, and then find the rest of the pitch pulses by searching recursively at a separation of one pitch period. It is possible to improve the performance of the residual energy based pitch pulse location detection by using Hilbert Envelope of Windowed LP Residual (HEWLPR) [50], [95]. A robust pitch pulse detection algorithm based on the group delay of the phase spectrum has also been reported [96], however this method has a very high computational complexity.

SWPM requires a pitch pulse detection algorithm, which can detect the pulses at stationary voiced segments with a high accuracy and has a low computational complexity. However the ability to detect the pitch pulses at the onsets and offsets is beneficial, since this will increase the flexibility of transition detection, see section 6.6.2. Therefore an improved pitch pulse detection algorithm based on the algorithm used in EVRC is developed for SWPM.

Figure 6.4 depicts a block diagram of the pitch pulse location detection algorithm. Initially all the possible pitch pulse locations are determined by considering the localised energy of the LPC residual and an adaptive threshold function, \( t(n) \). The localised energy \( e(n) \), of the LPC residual \( r(n) \), is given by equation 6.1.

\[
e(n) = \frac{1}{5} \sum_{j=-2}^{2} |r(n+j)| \quad \text{for } 2 \leq n < N - 2
\]  

Where \( N \), the length of the residual buffer is 240.

The adaptive threshold function, \( t(n) \) is updated for each half pitch period, by taking 0.7 of the maximum of \( e(n) \) in the pitch period symmetrically centred around the half pitch period chosen to calculate \( t(n) \), and \( t(n) \) is given by equation 6.2.
6.4. Synchronised Waveform matched Phase Model (SWPM)

\[ r(n) \rightarrow \text{Localised Energy} \rightarrow e(n) \rightarrow \text{Adaptive Threshold} \rightarrow \text{Pitch} \]

\[ e(n) \rightarrow e(n) > t(n) \rightarrow n_p \rightarrow \text{Probability Function} \rightarrow \text{History Bias} \rightarrow \Lambda_p \rightarrow n_{p*} \]

Figure 6.4: Block diagram of the pitch pulse detection algorithm

\[ t(n) = 0.7 \max[e(n_k - \tau_1/4 + n_\tau/2)] \]

for \( 0 \leq n_\tau < \tau_1 \) and \( 0 \leq n_{\tau/2} < \tau_{1/2} \)  \hspace{1cm} (6.2)

Where \( \tau_1 = \lfloor \tau + \frac{1}{2} \rfloor, \tau_{1/2} = \lfloor \frac{\tau}{2} + \frac{1}{2} \rfloor, \tau_{1/4} = \lfloor \frac{\tau}{4} + \frac{1}{2} \rfloor, n_k = k\tau_{1/2} \) for \( 1 \leq k < \left\lfloor \frac{2N}{\tau} \right\rfloor \), and \( \tau \) is the pitch period.

The exceptions corresponding to the analysis frame boundaries are given in the equations 6.3, 6.4, and 6.5.

\[ e(0) = e(1) = e(N-2) = e(N-1) = 0 \]

(6.3)

\[ t(m) = 0.7 \max[e(\tau_{1/2})] \quad \text{for} \quad 0 \leq m < \tau_{1/4} \]

(6.4)

\[ n_{\lfloor 2N/\tau \rfloor} = N - \tau_1 \]

(6.5)
6.4. Synchronised Waveform matched Phase Model (SWPM)

The sample locations, for which $e(n) > t(n)$, are considered as the regions which may contain pitch pulses. If $e(n) > t(n)$, for more than eight consecutive samples, those regions are ignored, since in those regions the residual energy is smeared, which is not a feature of pitch pulses. The centre of the each remaining region is taken as a possible pitch pulse location, $n_p$. If any of the two candidate locations are closer than eight samples from each other only the one which has the higher $e(n_p)$ is taken.

Applying an adaptive threshold to estimate the pitch pulse locations from the localised energy $e(n)$ is advantageous, especially for the segments where the energy of the LPC residual varies rapidly, giving rise to spurious pulses. Figure 6.5 demonstrates this for two occasions, (a) a male offset and (b) a female onset. The male speech segment has a pitch period of about 80 samples and the two high energy irregular pulses which do not belong to the pitch contour are clearly visible. The female speech segment has a pitch of about 45 samples, which also consists of two high energy irregular pulses. The energy function $e(n)$ and the threshold function $t(n)$ are also depicted in Figure 6.5, both $e(n)$ and $t(n)$ are shifted upwards for clarity. The figures also show that $e(n)$ at the irregular pulses may be higher than $e(n)$ at the proper pitch pulses. Therefore selecting the highest $e(n)$ to detect a pitch pulse location as in [93] may lead to errors. However, since $e(n) > t(n)$, for some of the irregular pulses as well as for proper pitch pulse

Figure 6.5: Irregular pulses at the onsets and offsets
locations, further refinements are required. Moreover, the regions where \( e(n) > t(n) \),
gives only a crude estimation of the pitch pulse location. The algorithm relies on the
accuracy of the pitch estimate for the computation of \( t(n) \) and in the refinement process
described below. However SWPM needs only the pitch pulses in the stationary voiced
segments, for which the pitch estimate is reliable.

For each selected location \( n_p \), probability of being a pitch pulse is estimated, using the
pitch and the energy of the neighbouring locations. First, a total energy metric, \( E_{p_0} \)
for the candidate pulse at \( n_{p_0} \) is computed recursively as follows.

\[
E_{p_0} = \sum l e(n_l)
\]  

Where \( l = p_0 \) and any \( q \) which satisfies the condition given in equation 6.7.

\[ |n_l \pm \tau - n_q| < 0.15\tau \]  

(6.7)

For each term, \(+\tau \) and \(-\tau \), if more than one \( q \) satisfy the condition 6.7 only the one
which minimises \( |n_l \pm \tau - n_q| \) is chosen. Then further locations \( n_q \) which satisfy the
condition 6.7 are searched recursively with any \( n_q \) which already satisfied the condi­
tion 6.7 taken as \( n_l \) in the next iteration. Therefore \( E_{p_0} \) can be defined as the sum of
\( e(n_p) \) of the pitch contour corresponding to the location \( n_{p_0} \). This process eliminates
the high energy irregular pulses, since they do not form a proper pitch contour and the
condition 6.7 detects them as isolated pulses. The probability of the candidate location,
\( n_{p_0} \) containing a pitch pulse \( \Lambda_{p_0} \) is given by equation 6.8.

\[
\Lambda_{p_0} = \frac{E_{p_0}}{\max [E_p]}
\]  

(6.8)

If pitch pulse locations were detected in the previous frame and any of the current
candidate pitch pulse locations form a pitch contour which is a continuation of the
previous pitch contour, a history bias term is added. Adding the history bias term
enhances the performance at the offsets, especially at the resonating tails. Furthermore, the history bias helps to maintain the continuity of the pitch contour between the frames, at the segments, where the pitch pulses become less significant, see Figure 6.6. A discontinuity in the pitch contour adds a reverberant character in voiced speech segments. The biased term \( \Lambda'_i \) for any location \( n_i \) which satisfies the conditions 6.10 or 6.11 is given by equation 6.9.

\[
\Lambda'_i = \Lambda_i + 0.2 \tag{6.9}
\]

The initial value for \( l \) is given by the condition 6.10, with \( \epsilon \) being the minimum possible integer value, which satisfies 6.10. If more than one \( l \) satisfies 6.10 with the same minimum \( \epsilon \) the one which maximises \( e(n_i) \) is taken.

\[
|n_{ist} + \epsilon \tau - n_i| < 0.1\tau \tag{6.10}
\]

Where \( n_{ist} \) is the pitch pulse location selected in the last analysis frame. Then any location \( n_{q} \) which satisfies the condition 6.11 is searched, and further \( n_i \) are found recursively with any \( n_{q} \) which already satisfied the condition 6.11 taken as \( n_i \) in the next iteration. If more than one \( n_{q} \) satisfy the condition 6.11 the one which minimises \( |n_i + \tau + n_{q}| \) is chosen.

\[
|n_i + \tau + n_{q}| < 0.15\tau \tag{6.11}
\]

The final probability of the candidate location, \( n_{p0} \) containing a pitch pulse \( \Lambda_{p0} \) is recalculated as shown in equation 6.12.

\[
\Lambda_{p0} = \frac{\Lambda'_i}{\max \left[ \Lambda'_p \right]} \tag{6.12}
\]

Set of positions, \( n_{pw} \) which have probabilities, \( \Lambda_p > 0.8 \), are selected as the pitch pulse locations, and further refined in order to select the pitch pulse closest to the synthesis
6.4. Synchronised Waveform matched Phase Model (SWPM)

Figure 6.6: Some difficult instances of pitch pulse extraction

frame boundary, as described in section 6.4.2. Figure 6.6 shows some difficult instances of pitch pulse detection along with the estimated probabilities, $A_p$ and the threshold value, (a) a female onset with irregular pulses, (b) a female voiced segment with a high fundamental frequency, (c) and (d) nasal sounds, which have a resonance character. In the case of (c) and (d) the resonating speech waveforms are also shown.

The problem illustrated in Figure 6.6 (b) can be explained both in the time domain and frequency domain. In speech segments with a short pitch period the short term prediction tends to remove some of the pitch correlation as well, leaving the LP residual...
6.4. Synchronised Waveform matched Phase Model (SWPM)

without any clearly distinguishable peaks. Shorter pitch periods in the time domain corresponds to smaller number of harmonics in the frequency domain. Hence the inter harmonic spacing becomes wider, and the formants of the short term predictor tends to coincide with some of the harmonics as depicted in Figure 6.7. The speech spectrum in Figure 6.7 is lowered by 80 dB in order to emphasise the coinciding points of the spectra. The excessive removal of some of the harmonic components by the LPC filter disperse the energy of the residual pitch pulses. It has been reported that large errors in the linear prediction coefficients occur in the analysis of sounds with high pitch frequencies [97].

In the case of nasal sounds, the speech waveform has a very high low frequency content, in fact it can be seen in Figure 6.6 (c) that the fundamental frequency component is dominant. In such cases the LPC filter simply places a pole at the fundamental frequency. A pole in the LPC synthesis filter translates to a zero in the inverse filter, giving rise to a fairly random looking LP residual signal. The figures demonstrate that the estimated probabilities, $\Lambda_p$ exceed the threshold value only at the required pitch pulse locations, despite those difficulties.
6.4. Synchronised Waveform matched Phase Model (SWPM)

6.4.2 Estimating the Pitch Pulse Shape (PPS)

Figure 6.8 depicts a complete pitch cycle of the LPC residual, which includes a selected pitch pulse and the positive half of the wrapped phase spectrum obtained from its DFT. The integer pitch pulse position, is taken as the time origin of the DFT, and the phase spectrum indicates that most of the harmonic phases are close to an average value. This average phase value varies with the shape of the pitch pulse, hence it is called pitch pulse shape (PPS). In the absence of a strong pitch pulse the phase spectrum becomes random, and varies between $-\pi$ and $\pi$.

Figure 6.9 depicts a block diagram of the pitch pulse shape estimation algorithm. The designed pitch pulse shape estimation algorithm employs an analysis by synthesis technique in the time domain to estimate PPS. A prototype pulse, $P(n_s)$ is synthesised as follows:

$$p(n_s) = \sum_{k=1}^{K} a_k \cos (k\omega n_s + \alpha_k) \quad \text{for} \quad -4 \leq n_s \leq 4$$  \hspace{1cm} (6.13)

Where $\omega = 2\pi/\tau$, $\tau$ is the pitch period, $K$ is the number of harmonics, $a_k$ are the
6.4. Synchronised Waveform matched Phase Model (SWPM)

harmonic amplitudes, and the candidate pitch pulse shapes, $\alpha_q$ are given by,

$$\alpha_q = \frac{2\pi q}{8} \quad \text{for} \quad 0 \leq q < 8 \quad (6.14)$$

Figure 6.10 depicts the synthesised pulses, $p(n_s)$ for two different candidate pitch pulse shapes, i.e. $\alpha_q$. A simpler solution to avoid estimating the spectral amplitudes, $a_k$ for the equation 6.13 is to assume a flat spectrum. However the use of spectral amplitudes, $a_k$ gives the relative weight for each harmonic, which is beneficial in estimating the pitch pulse shape. For example, a harmonic component which is relatively small in the LP residual signal and given an equal weight in the prototype pulse, $p(n_s)$ may lead to inaccurate estimates in the subsequent analysis by synthesis refinement process. Considering in the frequency domain, those relatively small amplitudes may be affected from the spectral leakage of the larger amplitudes, giving large errors in the phase spectrum. However since computing the spectral amplitudes for each pitch pulse is a very intensive process, as a compromise, the same spectral amplitudes are used for the whole analysis frame, which are also transmitted to the decoder as the harmonic amplitudes of the LP residual.

Then the normalised cross correlation, $R_j$ and SNR, $E_j$ are estimated between the synthesised prototype pitch pulse $p(n_s)$ and each of the detected LP residual pitch pulse, at the locations $n_p$, where $n_p \in n_{pw}$. $R_j$ and $E_j$ are estimated for each candidate pitch pulse shape, $\alpha_q$ using equations 6.15 and 6.16. The term, $j$ is introduced in $R_j$ and
6.4. Synchronised Waveform matched Phase Model (SWPM)

$E_j$, in order to shift the relative positioning of the LP residual pulse and the synthesised pulse. This compensates for the approximate pitch pulse locations, $n_p$ estimated by the algorithm described in section 6.4.1, by allowing to shift around the initial estimates, with a resolution of one sample.

$$R_j = \frac{\sum_{n_s=-4}^{4} r(n_p + n_s + j) p(n_s)}{\sqrt{\sum_{n_s=-4}^{4} p^2(n_s) \sum_{n_s=-4}^{4} r^2(n_p + n_s + j)}} \quad \text{for} \quad -3 \leq j \leq 3 \quad (6.15)$$

$$E_j = \frac{\sum_{n_s=-4}^{4} r^2(n_p + n_s + j)}{\sum_{n_s=-4}^{4} [p(n_s) - r(n_p + n_s + j)]^2} \quad \text{for} \quad -3 \leq j \leq 3 \quad (6.16)$$

Where $r(n)$ is the LPC residual. All the combinations of $n_p$, $\alpha_q$, and $j$ for which $E_j \leq 1.0$, are excluded from any further processing. $E_j \leq 1.0$, corresponds to an SNR of less than or equal to 0 dB. Then probability of the candidate shape, $\alpha_{q_0}$ being the pitch pulse shape is estimated using equation 6.17.

Figure 6.10: Synthesised pulses, $p(n_s)$

(a) $\alpha_q = 0$

(b) $\alpha_q = \pi/2$
6.4. Synchronised Waveform matched Phase Model (SWPM)

\[ \Lambda_{q_0} = \frac{N_{q_0}}{\max \{N_q\}} \]  \hspace{1cm} (6.17)

Where \( N_q \) is the total number of residual pulses for a given \( q \), for which \( R_j > 0.5 \). If more than one \( j \) satisfy the condition \( R_j > 0.5 \), for a particular set of \( q \) and \( n_p \), \( N_q \) is incremented only once. Set of pitch pulse shape values, \( \alpha_{q_0} \), which have probabilities, \( \Lambda_q > 0.7 \) are chosen for further refinements. If \( \max \{N_q\} \) is zero, then all the \( \Lambda_q \) are set to zero, i.e. no pitch pulses are detected. Figure 6.11 shows the LP residual of an analysis frame and the estimated probability density function of \( \alpha_q \) in the range \(-\pi \leq \alpha_q < \pi \). The pitch pulses of the LP residual in Figure 6.11 (a) have similar shapes to the shape of the synthesised pulse shown in Figure 6.10 (a). Consequently the pdf is maximum around \( \alpha_q = 0 \), the pitch pulse shape used to synthesis the pulse shown in Figure 6.10 (a).

If history bias is used in pitch pulse location detection, see section 6.4.1, then the probability term, \( \Lambda_q \) is not estimated. Instead the pitch pulse shape search is limited to three candidates, \( \alpha_L \) around the pitch pulse shape of the previous frame. During the voiced segments, the pitch pulse shape is fairly stationary and restricting the search range around the previous value does not reduce the performance. Restricting the

Figure 6.11: An analysis frame and the probability density function of \( \alpha_q \)
6.4. Synchronised Waveform matched Phase Model (SWPM)

Search range has advantages such as reduced computational complexity and efficient differential quantisation of the pitch pulse shape. Furthermore, restricting the search range avoids large variations in the pitch pulse shape. Large variations in the pitch pulse shape introduce a reverberant character in the synthesised speech.

\[ \alpha_L = \frac{2\pi (q_L + \delta)}{8} \quad \text{for} \quad -1 \leq \delta \leq 1 \quad (6.18) \]

Where

\[ q_L = \left[ \frac{\alpha_{qst}}{2\pi} + \frac{1}{2} \right] \quad \text{for} \quad 0 \leq \alpha_{qst} < 2\pi \quad (6.19) \]

Then \( R_j \) and \( E_j \) are estimated as before, substituting \( \alpha_q \) with \( \alpha_L \), and all the combinations of \( n_p, \alpha_L, \) and \( j \) for which \( E_j \leq 1.0 \) or \( R_j \leq 0.5 \) are excluded from any further refinements. If none of the combinations of \( n_p, \alpha_L, \) and \( j \) are left, the search is extended to all the \( \alpha_q \) and \( \Lambda_q \) is estimated as before, otherwise the remaining \( \alpha_L \) are chosen for further refinements, i.e. the remaining \( \alpha_L \) form the set, \( \alpha_{qst} \). The pitch pulse closest to the centre of the analysis frame, i.e. closest to the synthesis frame boundary for which \( R_j > \xi \) is selected as the final pitch pulse. The threshold value, \( \xi \) is given by,

\[ \xi = 0.7 \max [R_j] \quad \text{for} \quad \alpha_q \in \alpha_{qst} \quad (6.20) \]

If more than one set of \( j \) and \( \alpha_q \) satisfy the condition \( R_j > \xi \) for the same pitch pulse closest to the synthesis frame boundary, the set of values which maximises \( R_j \) is chosen. The pitch pulse shape and the integer pitch pulse location are given by the chosen, \( \alpha_q \) and \( n_p + j \) respectively. Figure 6.11 (a) also shows the centre of the analysis frame and the selected pitch pulse. It is also possible to select the pitch pulse closest to centre of the analysis frame from the set \( n_{p,w} \), and estimate the shape of the selected pulse. However estimating the pdf of \( \alpha_q \) for the whole analysis frame and including it in the selection process improves the reliability of the estimates, by selecting the most probable \( \alpha_q \).
Then the integer pitch pulse location is refined to a 0.125 sample accuracy, and the initial pitch pulse shape is refined to a \(2\pi/64\) accuracy. In the refinement process a synthetic pulse \(p_u(n_u)\) is generated in an eight times up-sampled domain, i.e. at 64 kHz. Let the selected integer pitch pulse location \(n_0\) and the initial pitch pulse shape \(\alpha_0\).

\[
p_u(n_u) = \sum_{k=1}^{K} a_k \cos(k\omega_u n_u + \alpha_i) \quad \text{for} \quad -40 \leq n_u < 40 \tag{6.21}
\]

Where \(\omega_u = 2\pi/8\), and \(\alpha_i\) is given by,

\[
\alpha_i = \alpha_0 + 2\pi i/64 \quad \text{for} \quad -4 \leq i \leq 4 \tag{6.22}
\]

Then equation 6.23 is used to compute the normalised cross correlation \(R_{i,j}\) for all \(i\) and \(j\), and the indices corresponding to the maximum \(R_{i,j}\) are used to evaluate the refined PPS and PPL, as shown in equations 6.22 and 6.25 respectively.

\[
R_{i,j} = \frac{\sum_{n_r=-4}^{4} r(n_0 + n_r) p_j(n_r)}{\sqrt{\sum_{n_r=-4}^{4} p_j^2(n_r) \sum_{n_r=-4}^{4} r^2(n_0 + n_r)}} \tag{6.23}
\]

Where \(p_j(n_r)\), is the shifted and down sampled version of \(p_u(n_u)\) given by,

\[
p_j(n_r) = p_u(8n_r + j) \quad \text{for} \quad -8 \leq j < 8 \quad \text{and} \quad -4 \leq n_r \leq 4 \tag{6.24}
\]

The final PPL, \(t_0\) refined to a 0.125 sample resolution is given by,

\[
t_0 = n_0 - j/8 \tag{6.25}
\]
6.4. Synchronised Waveform matched Phase Model (SWPM)

Fractional PPL is important for the segments with short pitch periods and when the pitch pulse is close to or at the synthesis frame boundary. When the pitch period is short, a small variation in the pitch pulse location can induce a large percentage pitch error. The pitch pulses closest to the synthesis frame boundaries are chosen in SWPM in order to maximise the waveform similarity at the frame boundaries, since the mode changes are limited to synthesis frame boundaries. However if the selected pitch pulse is within few samples from the frame boundary, especially at the frame boundary, the pulse must be synthesised smoothly across the boundary, in order to avoid audible artifacts. In such cases high resolution PPL and PPS are essential to maintain the phase continuity across the frame boundaries.

It is also possible to compute the cross correlation between $p_u(n_u)$ and eight times upsampled residual signal, in order to evaluate the best indices $i$ and $j$. However this requires more computations and an equally good result is obtained by shifting $p_u(n_u)$ in the up sampled domain and then computing the cross correlation in down sampled domain as shown in equations 6.23 and 6.24.

At the offsets, if no pitch pulses are detected, PPL is predicted from the PPL of the previous frame using the pitch, and PPS is set equal to the PPS of the previous frame. This does not introduce any deteriorating artifacts, since the encoder checks the suitability of the harmonic excitation in the mode selection process, see section 6.6.2. The prediction of PPL and PPS is particularly useful at the offsets with a resonant tail, where the pitch pulse detection is difficult. However at such segments with a high LPC gain, the pitch pulse locations are further refined, see section 6.5.2.

6.4.3 Synthesis using the generalised cubic phase interpolation

In the synthesis phases are interpolated cubically, i.e. quadratic interpolation of the frequencies [6]. In [6] phases are interpolated for the frequencies and phases available at the frame boundaries. But in the case of SWPM the frequencies are available at the frame boundaries and the phases at the pitch pulse locations. Therefore a generalised cubic phase interpolation formula is used, to incorporate PPL and PPS.

The phase $\theta_k(n)$ of the $k^{th}$ harmonic of the $i+1^{th}$ synthesis frame is given by equation
6.4. Synchronised Waveform matched Phase Model (SWPM)

\[ \theta_k(n) = \theta_{ki} + k\omega_i n + \alpha_k n^2 + \beta_k n^3 \quad \text{for} \quad 0 \leq n < N \]  \hspace{1cm} (6.26)

Where \( N \) is the number of samples per frame, \( \theta_{ki} \) and \( \omega_i \) are the phase of the \( k^{th} \) harmonic and the fundamental frequency respectively, at the end of the synthesis frame \( i \), and \( \alpha_k \) and \( \beta_k \) are given by,

\[
\begin{pmatrix}
    t_0^2 & t_0^3 \\
    2N & 3N^2
\end{pmatrix}
\begin{pmatrix}
    \alpha_k \\
    \beta_k
\end{pmatrix} =
\begin{pmatrix}
    \theta_{t0} - \theta_{ki} - k\omega_i t_0 + 2\pi M_k \\
    k\omega_{i+1} - k\omega_i
\end{pmatrix}
\]  \hspace{1cm} (6.27)

Where \( t_0 \) is fractional pitch pulse location (PPL), \( \theta_{t0} \) is PPS estimated at \( t_0 \), and \( M_k \) represents the phase unwrapping and chosen according to the “maximally smooth” criterion used by McAulay [6]. McAulay chose \( M_k \) such that, \( f(M_k) \) is a minimum, equation 6.28.

\[ f(M_k) = \int_{0}^{T} \left( \dot{\theta}_k(t,M_k) \right)^2 dt \]  \hspace{1cm} (6.28)

Where \( \theta_k(t,M_k) \) represents the continuous analogue form of \( \theta_k(n) \), and \( \dot{\theta}_k(t,M_k) \) is the second derivative of \( \theta_k(t,M_k) \) with respect to \( t \). Although \( M_k \) is integer valued, since \( f(M_k) \) is quadratic in \( M_k \), the problem is most easily solved by minimising \( f(x_k) \) with respect to the continuous variable \( x_k \) and then choosing \( M_k \) to be the integer closest to \( x_k \). For the generalised case of SWPM, \( f(x_k) \) is minimised with respect to \( x_k \) and the \( x_{k_{\min}} \) corresponding to the minimum is given by,

\[ x_{k_{\min}} = \frac{1}{2\pi} \left( \theta_{ki} - \theta_{t0} + k\omega_i t_0 + \frac{k(\omega_{i+1} - \omega_i) t_0^2}{2N} \right) \]  \hspace{1cm} (6.29)

\( M_{k_{\min}} = \lfloor x_{k_{\min}} + 0.5 \rfloor \), substituted in equation 6.27 for \( M_k \) to solve \( \alpha_k \) and \( \beta_k \) and in turn the unwrapped cubic phase interpolation function \( \theta_k(n) \).
6.5. Hybrid encoder

A simplified block diagram of the hybrid encoder is presented in Figure 6.12. The hybrid coder designed has a fixed frame length of 20 ms, corresponding to 160 samples at the sampling rate of 8 kHz. For each frame, the mode, which gives the optimum performance is selected. There are three possible modes: scaled white noise coloured by LPC for unvoiced, ACELP for transitions, or harmonic excitation for stationary voiced segments. Any waveform coding technique can be used instead of ACELP. In fact the designed hybrid model does not restrict the choice of coding technique for speech transitions, it merely makes the mode decision and defines the target waveform. The principles of ACELP coding technique are described in detail in chapter 5. The suitability of ACELP to encode the transitions is described in section 7.4. In the white
noise excited mode, the gain estimated from the LPC residual energy is transmitted for every 20 ms. The operation of the white noise excitation is described in detail in section 6.5.5, and synthesising the harmonic excitation is explained in detail in section 6.5.1.

LPC parameters are common for all the modes, are estimated every 20 ms, and interpolated in LSF domain for every 5 ms in the synthesis process. In order to interpolate the LSFs the LPC analysis windows are centred at the synthesis frame boundaries with a look ahead of 12.5 ms and a window length of 25 ms. The LPC analysis, converting to LSFs, LSF interpolation, LSF quantisation, and LSF to LPC transformation are described in chapter 3.

A two stage speech classification algorithm is designed. An initial classification is made based on the tracked energy, low band to high band energy ratio, and zero crossing rate, which decides to use either the noise excitation or one of the other modes. The secondary classification based on analysis by synthesis decides to use either the harmonic excitation or ACELP. Segments of plosives with high energy spikes are synthesised using ACELP, use of ACELP for such frames improves speech quality, see section 6.6.3. If the noise excitation mode is selected there is no need to estimate the excitation parameters of the other modes. However if the noise excitation is not selected the harmonic parameters are always estimated and the harmonic excitation is generated at the encoder, for the analysis by synthesis transition detection. The speech classification is described in detail in section 6.6.

For simplicity, details of LPC and adaptive codebook memory update are excluded in the block diagram. The encoder maintains an LPC synthesis filter synchronised with the decoder, and uses the final memory locations for ACELP and analysis by synthesis transition detection in the next frame. Adaptive codebook memory is always updated with the previous LPC excitation vector regardless of the mode. In order to maintain the LPC and the adaptive codebook memories the LPC excitation is generated at the encoder, regardless of the mode.
6.5. Synchronised harmonic excitation

In the harmonic mode pitch and harmonic amplitudes are estimated for every 20 ms frame. Pitch is estimated from pre-processed speech [98], while the harmonic amplitudes are estimated from the LPC residual. The estimation windows are placed at the end of the synthesis frames, and a look ahead of 15 ms is used to facilitate the harmonic parameter interpolation. The pitch estimation algorithm used is based on sinusoidal speech model matching proposed by McAulay [61], and improved by Atkinson [5] and Villette [58], [99]. The initial pitch is refined to 0.2 sample accuracy using synthetic spectral matching proposed by Griffin [63]. For more details on pitch estimation and refinement see section 3.5.2. The harmonic amplitudes are estimated by simple peak picking of the magnitude spectrum of the LPC residual, see section 4.3.2.

The harmonic excitation $e_h(n)$ is generated at the encoder for the analysis by synthesis transition detection and to maintain the LPC and adaptive codebook memories, which is shown in equation 6.30.

$$e_h(n) = \sum_{k=1}^{K} a_k(n) \cos(\theta_k(n)) \quad \text{for } 0 \leq n < N$$ (6.30)

Where $K$ is the number of harmonics. Since two analysis frames are interpolated to produce a synthesis frame, $K$ is taken as the higher number of harmonics out of the two analysis frames and the missing amplitudes of the other analysis frame are set to zero. $N$ is the number of samples in a synthesis frame and $\theta_k(n)$ is given in equation 6.26 for continuing harmonic tracks, i.e. each harmonic of an analysis frame is matched with the corresponding harmonic of the next frame. For terminating harmonics, i.e. when the number of harmonics in the next frame is smaller, $\theta_k(n)$ is given by equation 6.31.

$$\theta_k(n) = \theta_{k_i} + 2\pi kn/\tau_i$$ (6.31)

Where $\theta_{k_i}$ is the phase of the harmonic $k$ at the end of synthesis frame $i$ and $\tau_i$ is the pitch at the end of synthesis frame $i$. 
6.5. Hybrid encoder

For emerging harmonics \( \theta_k(n) \) is given by equation 6.32.

\[
\theta_k(n) = \theta_{t_0} + 2\pi k(n - t_0)/\tau_{i+1}
\]

(6.32)

Where \( t_0 \) is the PPL at the end of synthesis frame \( i + 1 \), \( \theta_{t_0} \) is the corresponding PPS, and \( \tau_{i+1} \) is the pitch at the end of synthesis frame \( i + 1 \).

Continuing harmonic amplitudes are linearly interpolated as shown in equation 6.33.

\[
a_k(n) = a_{k_i} + \frac{(a_{k_{i+1}} - a_{k_i})n}{N}
\]

for \( 0 \leq n < N \)

(6.33)

Where \( a_{k_i} \) is the amplitude estimate of the \( k^{th} \) harmonic at the end of the synthesis frame \( i \). For terminating harmonic amplitudes a trapezoidal window, unity for 55 samples and linearly decaying for 50 samples is applied, from the beginning of the synthesis frame, as shown in equation 6.34.

\[
a_k(n) = a_{k_i}
\]

for \( 0 \leq n < 55 \),

\[
a_k(n) = a_{k_i} \frac{105 - n}{50}
\]

for \( 55 \leq n < 105 \)

(6.34)

For emerging harmonic amplitudes a trapezoidal window, linearly rising for 50 samples and unity for 55 samples is applied, starting from the 56th sample of the synthesis frame, as shown in equation 6.35.

\[
a_k(n) = a_{k_{i+1}} \frac{n - 55}{50}
\]

for \( 55 \leq n < 105 \),

\[
a_k(n) = a_{k_{i+1}}
\]

for \( 105 \leq n < 160 \)

(6.35)

6.5.2 Pros and cons of SWPM

Figure 6.13 depicts some examples of the waveforms synthesised using the harmonic excitation technique described in section 6.5.1. In each example (i) represents the LPC residual or the original speech signal and (ii) represents the LPC excitation or the
synthesised speech signal. Figure 6.13 (a) shows the LPC residual and the harmonic excitation of a segment which has strong pitch pulses, and Figure 6.13 (b) shows the corresponding speech waveforms. It can be seen that the synthesised speech waveform is very similar to the original. Figure 6.13 (c) shows the LPC residual and the harmonic excitation of a segment which has weak or dispersed pitch pulses, and Figure 6.13 (d) shows the corresponding speech waveforms. The synthesised speech is time synchronised with the original, however the waveform shapes are slightly different, especially in between the major pitch pulses. The waveform similarity is highest at the major excitation pulse locations and decreases along the pitch cycles. This due to the fact that SWPM models only the major pitch pulses, and when the LPC residual energy is dispersed, the excitation can not model the minor pulses present in the residual signal. Furthermore, the dispersed energy of the LPC residual, becomes concentrated around the major pitch pulses in the excitation signal. The synthesised speech also exhibits larger variations in the amplitude around the pitch pulse locations, compared with the original speech.

In order to understand the effects on subjective quality due to the above observations, an informal listening test was conducted by switching between the harmonically synthesised speech and the original speech waveforms at desired synthesis frame boundaries. The informal listening tests show that occasional audible artifacts still remain at the mode transitions, when switching from the harmonic mode to the waveform coding mode, however there are no audible switching artifacts at the onsets. It was found that this is due to two reasons: difficulties in reliable pitch pulse detection and limitations in representing the harmonic phases using the pitch pulse shape at some segments. At some highly resonance segments the LPC residual looks like random noise and it is not possible even to define the pitch pulses. The predicted pitch pulse location, assuming a continuing pitch contour may be incorrect at resonant tails. At such segments the pitch pulse locations are determined by applying analysis by synthesis techniques in the speech domain, such that the synthesised speech signal is synchronised with the original, as described below in subsection: Synchronisation at resonant tails.

At the segments as illustrated using Figure 6.13 (c), it is possible to detect dominant pitch pulses. However the LPC residual energy is dispersed throughout the pitch peri-
ods, making the pitch pulses less significant, as described in section 6.4.1. This effect reduces the coherence of the LPC residual harmonic phases at the pulse locations and the DFT phase spectrum estimated at the pulse locations become random looking. Female vowels with short pitch periods show this type of characteristics. A dispersed phase spectrum reduces the effectiveness of the pitch pulse shape, since the concept of pitch pulse shape is based on the assumption that a pitch pulse is the result of the super imposition of coherent phases, which have the same value at the pitch pulse location. This effect is illustrated in Figure 6.14. The pitch pulse synthesised models the major pulse in the LPC residual pitch period, and concentrates the energy at the
Figure 6.14: An illustration of PPS at a dispersed pitch period, (a) A complete pitch cycle of the LPC residual, (b) Pitch pulse synthesised using PPS to represent the pulse in (a), (c) Positive half of the phase spectrum obtained from the DFT of the pitch cycle in (a), (d) Estimated PPS

pulse location. This is due to the single phase value used in synthesising the pulse, as opposed to the more random looking phase spectrum of the original pitch cycle. This phenomenon introduces phase discontinuities, which accounts for the audible switching artifacts. However the click and pop sounds present at the mode transitions in the speech synthesised with SWPM are less annoying compared to a conventional zero phase excitation, even if the pitch pulse locations are synchronised. This is because SWPM has the additional flexibility of choosing the most suitable phase value (PPS) for pitch pulses, such that the phase discontinuities are minimised. Figure 6.15 illustrates the effect of PPS on the LPC excitation and the synthesised speech signals. For comparison original signals and the signals synthesised using the SB-LPC coder [5] which assumes a zero phase excitation are included in the figure.

The absence of audible switching artifacts at the onsets is an interesting issue. There are two basic reasons for these differences in switching artifacts noticed at the onsets and the offsets: the nature of the excitation signal and the LPC memory. At the onsets, even though the pitch pulses may be irregular due to the unsettled pitch of the vocal cords, they are quite strong and the residual energy is concentrated around them. Resonating segments or dispersed pulses do not occur at the onsets. Therefore the only difficulty
6.5. Hybrid encoder

Figure 6.15: An illustration of speech synthesised using PPS

at the onsets is identifying the correct pulses, and as long as the pulse identification process is successful SWPM can maintain the continuity of the harmonic phases at the onsets. The pitch pulse detection algorithm designed is capable of accurate detection of the pitch pulses at the onsets as described in section 6.4.1. Furthermore, at the onsets waveform coding preserves the waveform similarity, this also ensures the correct LPC memory, since LPC memory contain the past synthesised speech samples. Therefore the mode transition at the onsets is relatively easier, and SWPM guaranties a smooth mode transition at the onsets. However at the offsets the presence of weak pitch pulses is a common feature. Moreover the infinite impulse response LPC filter carries on the phase changes caused by the past excitation signal, especially when the LPC filter gain is high at the offsets. Therefore the audible switching artifacts remain at some of the offset mode transitions. The techniques developed to eliminate those artifacts are described in section 6.5.3.

Synchronisation at resonant tails

At the resonant tails the LPC residual looks like random noise, and the pitch pulses are not clearly identifiable. In those cases analysis by synthesis techniques can be applied directly on the speech signal to synchronise the synthesised speech. This process is
applied only for the frames, which follow a harmonic frame and have been classified as transitions.

Synthesised speech is generated by shifting the pitch pulse location (PPL) at the end of the synthesis frame, ±τ/2 around the synthesis frame boundary with a resolution of one sample, where τ is the pitch period. The location which gives the best cross correlation between the synthesised speech and the original speech is selected as the refined PPL. The pitch pulse shape is not included in the search and set equal to the pitch pulse shape of the previous frame. The excitation and the synthesised speech corresponding to the refined PPL are applied to the closed loop transition detection algorithm, and used only if the transition detection algorithm classifies the corresponding frame as harmonic, otherwise waveform coding is used.

6.5.3 Offset target modification

SWPM minimises the phase discontinuities at the mode transitions, as described in section 6.5.2. However at some mode transitions such as the offsets after female vowels, which have dispersed pulses audible phase discontinuities still remain. These discontinuities may be eliminated by transmitting more phase information. This section describes a more economical solution to remove those remaining phase discontinuities at the offsets, which does not need to transmit additional information. The proposed method modifies some of the harmonic phases of the first frame of the waveform coding target, which follows the harmonic mode. The remaining phase discontinuities can be corrected within the first waveform coding frame, since SWPM keeps the phase discontinuities at a minimum and the pitch periods synchronised.

As a first approach the harmonic excitation is extended into the next frame, and the synthesised speech is linearly interpolated with the original speech at the beginning of the frame in order to produce the waveform coding target. Listening tests were carried out with different interpolation lengths. The waveform coding target was not quantised in order to isolate the distortions due to switching. The tests were extended in order to understand the audibility of the phase discontinuities with the frequency of the harmonics, by manually shifting one phase at a time and synthesising the rest of
the harmonics using the original phases. Phase shifts of $\pi/2$ and $\pi$ are used. Listening tests show that despite of the interpolation length the phase discontinuities below 1 kHz are audible, and an interpolation length as small as 10 samples is sufficient to mask the others. Furthermore, male speech segments with long pitch periods around 80 samples and above, do not cause audible switching artifacts. Male speech segments with long pitch periods have well resolved short term and long term correlations, and produce clear and sharp pitch pulses, which can be easily modeled by SWPM. The phase discontinuities below 1 kHz are more pronounced, because they carry most of the speech energy due to the high amplification of the first formant of the LPC filter. Therefore only the harmonics below 1 kHz of the segments with pitch periods shorter than 80 samples are considered in the offset target modification process.

The harmonic excitation is extended beyond the mode transition frame boundary, and the synthesised speech is generated in order to estimate the harmonic phases of the synthesised speech at the mode transition frame boundary. The phase of the $k^{th}$ harmonic of the excitation is computed as follows:

$$\theta_{k_{i+1}}(n) = \theta_{k_i} + 2\pi kn/\tau_i \quad \text{for} \quad 0 \leq n < N$$ (6.36)

Where $\theta_{k_i}$ is the phase of the $k^{th}$ harmonic and $\tau_i$ is the pitch, at the end of the synthesis frame $i$. The excitation signal is given by,

$$e_{k_{i+1}}(n) = \sum_{k=1}^{K} a_{k_i} \cos \left( \theta_{k_{i+1}}(n) \right)$$ (6.37)

Where $K$ is the number of harmonics and $a_{k_i}$ is the amplitude of the $k^{th}$ harmonic estimated at the end of the synthesis frame $i$. Then the excitation signal is filtered through the LPC synthesis filter to produce the synthesised speech signal, with the coefficients estimated at the end of the synthesis frame $i$. The LPC memories after synthesising the $i^{th}$ frame are used as the initial memories. Then the speech synthesised for the $i^{th}$ frame and $i+1^{th}$ frame are concatenated and windowed with a Kaiser window.
of 200 samples and $\beta$ of 6.0 centred at the frame boundary. The harmonic phases, $\varphi_k$, are estimated using a 512 point FFT.

The original speech signal is windowed at three points: at the end of the synthesis frame $i$, at the centre of the synthesis frame $i + 1$, and at the end of the synthesis frame $i + 1$, using the same window function as before. The corresponding harmonic amplitudes, $A_k$, $A_{k+1/2}$, $A_{k+1}$ and the phases $\phi_k$, $\phi_{k+1/2}$, $\phi_{k+1}$ are estimated using 512 point FFTs. Then the signal component $s_l(n)$, which consists of the harmonics below 1 kHz is synthesised as follows:

$$s_l(n) = \sum_{k=1}^{L} A_k(n) \cos(\Theta_k(n)) \quad \text{for } 0 \leq n < N$$  \quad (6.38)

Where $L$ is the number of harmonics below 1 kHz at the end of the $i^{th}$ synthesis frame, $A_k(n)$ is obtained by linear interpolation between $A_k$, $A_{k+1/2}$, and $A_{k+1}$, and $\Theta_k(n)$ is obtained by cubic phase interpolation \cite{6} between $\phi_k$, $\phi_{k+1/2}$, and $\phi_{k+1}$.

Then the signal $s_m(n)$, which has modified phases is synthesised as follows:

$$s_m(n) = \sum_{k=1}^{L} A_k(n) \cos(\Theta_k(n)) \quad \text{for } 0 \leq n < N$$  \quad (6.39)

Where $\Theta_k(n)$ is obtained by cubic phase interpolation between $\varphi_k$, and $\varphi_{k+1}$. Thus the modified signal, $S_m(n)$ has the phases of the harmonically synthesised speech at the beginning of the frame and the phases of the original speech at the end of the frame. In other words $\Theta_k(n)$, i.e. the rate of change of each harmonic phase is modified such that the phase discontinuities are eliminated, by keeping $\Theta_k(n)$ equal to the harmonic frequencies at the frame boundaries. There is a possibility that such phase modification operations induce a reverberant character in the synthesised signals. However, large phase mismatches close to $\pi$ are rare, because SWPM minimises the phase discontinuities. Furthermore, the modifications are applied only for the speech segments, which have pitch periods shorter than 80 samples, thus a phase mismatch is smoothed out in few pitch cycles. The listening tests confirm that the synthesised speech does not posses
Figure 6.16: An illustration of offset target modification, (a) $s(n)$, (b) $s_t(n)$, (c) $s_i(n)$, and (d) $s_m(n)$

a reverberant character. Limiting the phase modification process for the segments with pitch periods shorter than 80 samples also improves the accuracy of the spectral estimations, which use a window length of 200 samples. The modified waveform coding target of the $i + 1^{th}$ synthesis frame is given by,

$$s_t(n) = s(n) - s_i(n) + s_m(n)$$  \hspace{1cm} (6.40)

Figure 6.16 illustrates the waveforms of equation 6.40. It can be seen that the phases of the low frequency components of the original speech waveform, $s(n)$ are modified in order to obtain $s_t(n)$. The waveforms (c) and (d) depict $s_t(n)$ and $s_m(n)$ respectively, the low frequency components, which have been modified. The phase relationships between the high frequency components account more for the perceptual quality of speech [92], and the high frequency phase components are unchanged in the process.

Some speech signals show rapid variations in the harmonic structure at the offsets, which may reduce the efficiency of the phase modification process. In order to limit
those effects the spectral amplitude and phase estimation process is not strictly confined to the harmonics of the fundamental frequency. Instead the amplitude and phase corresponding to the spectral peak closest to each harmonic frequency are estimated. The frequency of the selected spectral peak is taken as the frequency of the estimated amplitude and phase. When finding the spectral peaks closest to the harmonic frequencies, the harmonic frequencies are determined by the fundamental frequency at the end of the $i^{th}$ synthesis frame, since the pitch estimates at the transition frame are less reliable. In fact the purpose of the offset target modification process is to find the frequency components corresponding to the harmonics of the harmonically synthesised frame, i.e. the $i^{th}$ frame and change the phase evolution of those components such that the discontinuities are eliminated. Moreover, the same set of spectral peak frequencies and amplitudes are used when synthesising the terms $s_l(n)$ and $s_m(n)$, hence there is no need to restrict the synthesis process to the pitch harmonics.

Another important issue at the offsets is the energy contour of the synthesised speech. The harmonic coder does not directly control the energy of the synthesised speech, since it transmits the residual energy. However the waveform coders directly control the energy of the synthesised speech, by estimating the excitation gain using the synthesised speech waveform. This may cause discontinuities at the offset mode transition frame boundaries, especially when the LPC filter gain is high. The final target for the waveform coder is produced by linear interpolation between the extended harmonically synthesised speech and the modified target, $s_t(n)$ at the beginning of the frame for 10 samples. The linear interpolation ensures the discontinuities due to variations of the energy contour are eliminated as well as the phase discontinuities, which are not accounted for in the phase modification process, as described above.

6.5.4 Onset harmonic memory initialisation

The harmonic phase evolution described in section 6.4.3 and the harmonic excitation described in section 6.5.1 interpolate the harmonic parameters in the synthesis process, and assume that the model parameters are available at the synthesis frame boundaries. However at the onset mode transitions, when switching from the waveform coding
mode the harmonic model parameters are not directly available. The initial phases $\theta_k$, and the fundamental frequency $\omega_i$ in the phase evolution formula 6.26, and the initial harmonic amplitudes $a_k$ in the equation 6.33 are not available at the onsets. Therefore those unavailable harmonic model parameters should be estimated at the decoder from the available information. The signal reconstructed by the waveform coder prior to the frame boundary and the harmonic parameters estimated at the end of the synthesis frame boundary are available at the decoder. The use of waveform coded signal in estimating the harmonic parameters at the onsets may be unreliable due to two reasons: the speech signal shows large variations at the onsets and at low bit rates the ACELP excitation at the onsets reduces to few dominant pulses, lowering the reliability of spectral estimates. Therefore the use of waveform coded signal in estimating the harmonic parameters is minimised. The waveform coded signal is used only in initialising the amplitude quantisation memories, see section 7.3.

Since preserving the waveform similarity at the frame boundaries is important, pitch is recomputed such that the previous pitch pulse location can be computed at the decoder. Therefore the transmitted pitch represents the average over the synthesis frame. The other harmonic model parameters transmitted are unchanged, which are estimated at the end of the synthesis frame boundary. Let the pitch, $\tau_{i+1}$ and pitch pulse location, $t_{0_{i+1}}$ at the end of the $i+1^{th}$ synthesis frame, and pitch pulse location at the end of the $i^{th}$ synthesis frame $t_{0_{i}}$. The number of pitch cycles $n_c$ between $t_{0_{i}}$ and $t_{0_{i+1}}$ is given by,

$$n_c = \left\lfloor \frac{t_{0_{i+1}} - t_{0_{i}}}{\tau_{i+1}} + \frac{1}{2} \right\rfloor \quad (6.41)$$

The recomputed pitch, $\tau_{r}$ is given by,

$$\tau_{r} = \frac{t_{0_{i+1}} - t_{0_{i}}}{n_c} \quad (6.42)$$

Then $\tau_{r}$ and $t_{0_{i+1}}$ are transmitted, and $t_{0_{i}}$ is computed at the decoder, as follows,

$$t_{0_{i}} = t_{0_{i+1}} - \tau_{r} \left\lfloor \frac{t_{0_{i+1}} - t'}{\tau_{r}} + \frac{1}{2} \right\rfloor \quad (6.43)$$
Where \( t' \) is the starting frame boundary, and \( t_0 \) is the pitch pulse location closest to \( t' \). The pitch pulse shape, \( \theta_0 \), at the end of the \( i^{th} \) synthesis frame is set equal to the pitch pulse shape, \( \theta_{0+i} \), at the end of the \( i+1^{th} \) synthesis frame. The initial phases \( \theta_k \) in the equation 6.26 are estimated as follows,

\[
\theta_k = \theta_0 - \frac{2\pi k t_0}{\tau_r}
\]  

(6.44)

Both fundamental frequency terms \( \omega_i \) and \( \omega_{i+1} \) in the equation 6.27 are computed using \( \tau_r \), i.e. \( \omega_i = \omega_{i+1} = 2\pi/\tau_r \). The harmonic amplitudes \( a_{ki} \) in the equation 6.33 are set equal to \( a_{k+i} \). Therefore effectively the phase evolution of the first harmonic frame of a stationary voiced segment becomes linear, and the harmonic amplitudes are kept constant, i.e. not interpolated.

6.5.5 White noise excitation

The unvoiced speech has a very complicated waveform structure. ACELP can be used to synthesise unvoiced speech, and it will essentially match the waveform shape. However, a large number of excitation pulses are required to synthesise the noise like unvoiced speech. Reducing the number of pulses introduces sparse excitation artifacts in noise like segments [100]. The synthesised speech also shows the sparse nature, and the pulse locations are clearly identifiable even in the LPC synthesised speech. In fact during unvoiced speech the short term correlation is negligible and the LPC filter gain becomes insignificant.

The sinusoidal excitation can also be used to synthesise unvoiced segments, despite the fact that there is no harmonic structure. Speech synthesised by sampling the magnitude spectrum at 100 Hz and a uniformly distributed random phases for unvoiced segments can achieve very good quality [101]. This method suits the sinusoidal coders using frequency domain voicing without an explicit time domain mode decision, since it facilitates the use of the same general analysis and synthesis structure for both voiced and unvoiced speech. However the designed hybrid model classifies the unvoiced and
silence segments as a separate mode, and a simpler unvoiced excitation generation model, which does not require any frequency domain transforms is used.

It has been shown that scaled white noise coloured by LPC can produce unvoiced speech with quality equivalent to μ-law logarithmic PCM [8], [102]. Implying that the complicated waveform structure of unvoiced speech has no perceptual importance. Therefore in terms of the perceptual quality, the phase information transmitted by ACELP is redundant, and higher synthesis quality can be achieved at lower bit rates using scaled white noise excitation. Figure 6.17 shows a block diagram of the unvoiced gain estimation process. Figure 6.18 shows a block diagram of the unvoiced synthesis process. The band pass filters used are identical and have cutoff frequencies of 140 Hz and 3800 Hz. The transfer function of the fourth order infinite impulse response (IIR) band pass filters is given by equation 6.45.

\[
H_{bp}(z) = \frac{0.8278 - 1.6556z^{-2} + 0.8278z^{-4}}{1 - 0.0662z^{-1} - 1.6239z^{-2} + 0.0451z^{-3} + 0.6855z^{-4}}
\]

(6.45)

The unvoiced gain, \( g_{uv} \) is given by equation 6.46.

\[
g_{uv} = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} r_{bp}^2(n)}
\]

(6.46)
6.6 Speech classification

Where $r_{bp}^2(n)$ is the band pass filtered LPC residual signal and N is the length of the residual vector, which is 160 samples with a look ahead of 80 samples to facilitate overlap and add synthesis at the decoder.

White noise, $u(n)$ is generated by a random number generator, with a Gaussian distribution. Gaussian noise source was found to be subjectively superior to a simple uniform noise source. The scaled white noise excitation, $u_s(n)$ is given by equation 6.47.

$$u_s(n) = \frac{g_{uv}}{\sqrt{\sum_{n=0}^{K} u_{bp}(n)/K}}$$

Where $u_{bp}(n)$ is the band pass filtered white noise and K is the length of the noise vector, which is 240 samples. For overlap and add a trapezoidal window is used with an overlap of 80 samples. For each synthesis frame the filtered noise buffer, $u_{bp}$ is shifted by 80 samples and new 160 samples are appended, this eliminates the need for energy compensation functions to remove the windowing effects [67]. In fact the overlapped segments are correlated, and the trapezoidal windows do not distort the rms energy.

No attempt is made to preserve the phase continuity when switching to or from the noise excitation. When switching from a different mode the unvoiced gain, $g_{uv}$ of the previous frame is set equal to the current value. The validity of these assumptions are tested through listening tests and the results confirm that these assumptions are reasonable and do not introduce any audible artifacts. The average bit rate can be further reduced by the introduction of voice activity detection (VAD) and comfort noise generation at the decoder for silence segments [79], [103].

6.6 Speech classification

The speech classification or the mode selection techniques can be divided into three categories [104].

**Open loop mode selection:** Each frame is classified based on the observations of
parameters extracted from the input speech frame without assessing how the selected mode will perform for the frame concerned.

Closed loop mode selection: Each frame is synthesised using all the modes, and the mode, which gives the best performance is selected.

Hybrid mode selection: The mode selection procedure combines both open loop and closed loop approaches. Typically, a subset of modes is first selected by an open loop procedure, which followed by further refinements using closed loop techniques.

The closed loop mode selection has two major difficulties: high complexity and finding an objective measure which reflects the subjective quality of synthesised speech [105]. The existing closed loop mode selection coders are based on CELP, and select the best configuration such that the weighted MSE is minimised [106], [107]. Open loop mode selection is based on the techniques such as: voice activity detection, voicing decision, spectral envelope variation, speech energy, and phonetic classification [80]. See [4] for a detailed description on acoustic phonetics.

During the course of this project, a hybrid mode selection technique is designed, with an open loop initial classification and a closed loop secondary classification. The speech classification process is shown in Figure 6.12: Block diagram of the hybrid encoder. The open loop initial classification decides to use either the noise excitation or one of the other modes. The secondary classification synthesises the harmonic excitation and makes a closed loop decision to use either the harmonic excitation or ACELP. A special feature of the designed classifier is application of closed loop mode selection for harmonic coding. The Synchronised Waveform matched Phase Model (SWPM) preserves the waveform similarity of the harmonically synthesised speech, making it possible to apply closed loop techniques in harmonic coding.

6.6.1 Open loop initial classification

The initial classification extracts the unvoiced and silence segments of speech, which are synthesised using the white noise excitation. It is based on tracked energy, low band to high band energy ratio, and zero crossing rate of the speech signal. The
six voicing metrics are logically combined to enhance the reliability, since a single metric alone is not sufficient to make a decision with a high confidence. The metric combinations and thresholds are determined empirically, by plotting the metrics with the corresponding speech waveforms. A statistical approach is not suitable for deciding the thresholds, because the design of the classification algorithm should consider that a miss classification of a voiced segment as unvoiced will severely degrade the speech quality, but a miss classification of an unvoiced segment as voiced can be tolerated. A miss classified unvoiced segment will be synthesised using ACELP, however a miss classified voiced segment will be synthesised using the noise excitation.

The tracked energy of speech, $t_e$ is estimated as follows:

$$t_e = \frac{0.00025e_h + e}{0.01e_h + e}$$  \hspace{1cm} (6.48)

where $e$ is the mean squared speech energy, given by,

$$e = \frac{\sum_{n=0}^{N-1} s^2(n)}{N}$$  \hspace{1cm} (6.49)

$N$, the length of the analysis frames is 160, $e_h$ is an autoregressive energy term given by,

$$e_h = 0.9e_h + 0.1e \text{ if } 8e > e_h$$  \hspace{1cm} (6.50)

the condition $8e > e_h$ ensures that $e_h$ is updated only when the speech energy is sufficiently high and $e_h$ should be initialised approximately to the mean squared energy of voiced speech. Figure 6.19 (a) illustrates the tracked energy over a segment of speech.

The low band to high band energy ratio, $\gamma_\omega$ is estimated as follows:
6.6. Speech classification

\[ \gamma_\omega = \frac{\int_0^{1/4} S^2 \left( \frac{\omega}{\omega_s} \right) d \left( \frac{\omega}{\omega_s} \right)}{\int_{1/4}^{1/2} S^2 \left( \frac{\omega}{\omega_s} \right) d \left( \frac{\omega}{\omega_s} \right)} \]  

(6.51)

where \( \omega_s \) is the sampling frequency and \( S(\omega) \) is the speech spectrum. The speech spectrum is estimated using a 512 point FFT, windowing 240 speech samples with a Kaiser window of \( \beta \) set to 6.0. Figure 6.19 (b) illustrates the low band to high band energy ratio over a segment of speech, the speech signal is shifted down for clarity.

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. Figure 6.20 (a) illustrates the zero crossing rate over a segment of speech, the speech signal is shifted down for clarity. Figure 6.20 (b) depicts the voicing decision made by the initial classification. Figure 6.21 depicts the three metrics used and the final voicing decision over the same speech segment.

Even though the plosives have significant amount of energy at high frequencies and a high zero crossing rate, synthesising high energy spikes of the plosives using ACELP improves speech quality. Therefore a special routine is designed to detect the plosives, which are classified as unvoiced by the initial classification, and switch them to ACELP.
6.6. Speech classification

Figure 6.20: Zero crossing rate and initial classification

Figure 6.21: Voicing metrics of the initial classification

mode, see section 6.6.3. An example is depicted at the beginning of the speech segment in Figure 6.20 (b).

6.6.2 Closed loop transition detection

The analysis by synthesis transition detection is performed on the speech segments, which are declared voiced by the open loop initial classification. A block diagram of the analysis by synthesis classification process is shown in Figure 6.22. Analysis by synthesis classification module synthesises speech using SWPM, see section 6.5.1, and
checks the suitability of the harmonic model for a given frame. The normalised cross correlation and squared error are computed in both the speech domain and the residual domain for each of the selected pitch cycles of a synthesis frame. The pitch cycles are selected such that they cover the complete synthesis frame. The mode decision between harmonic and ACELP modes is based on the estimated cross correlation and squared error values.

The squared error, $E_i$ of the $i^{th}$ pitch cycle is given by,

$$E_i = \frac{\sum_{j=0}^{T-1} [s(iT + j) - \hat{s}(iT + j)]^2}{\sum_{j=0}^{T-1} s^2(iT + j)}$$

for $0 < i < I$ (6.52)

The normalised cross correlation, $R_i$ of the $i^{th}$ pitch cycle is given by,

$$R_i = \frac{\sum_{j=0}^{T-1} s(iT + j) \hat{s}(iT + j)}{\sqrt{\sum_{j=0}^{T-1} s^2(iT + j) \sum_{j=0}^{T-1} \hat{s}^2(iT + j)}}$$

for $0 < i < I$ (6.53)

Where $T = [\tau + 0.5]$, $\tau$ is the pitch period, $I = [N/\tau + 1]$, and $N$ is the synthesis frame length i.e. 160 samples.

In order to estimate the normalised cross correlation, $R_i$, and squared error, $E_i$, in the LPC residual domain the computations 6.52 and 6.53 are repeated with $s(n)$ and $\hat{s}(n)$ replaced with $r(n)$ and $\hat{r}(n)$ respectively. Figure 6.23 depicts $E_i$, $R_i$, original speech
6.6. **Speech classification**

$s(n)$, and synthesised speech $\hat{s}(n)$. $E_i$ and $R_i$ are aligned with the corresponding pitch cycles of the speech waveforms, and the speech waveforms are shifted down for clarity. Examples of the residual domain signals, LPC residual $r(n)$, LPC excitation $\hat{r}(n)$, $E_i^e$, and $R_i^e$ are also shown in the figure.

For stationary voiced speech, squared error, $E_i$ is usually much lower than unity, and the normalised cross correlation, $R_i$ is close to unity. However at the transitions the harmonic model fails, which results in larger errors and lower correlation values. The cross correlation and squared error values are estimated on the pitch cycle basis in order to determine the suitability of the harmonic excitation for each pitch cycle. Estimating the parameters over the complete synthesis frames may average out a large error, caused by a sudden transition. In Figure 6.23 (a) the speech waveform has a minor transition. The estimated parameters also indicate the presence of the transition. Such minor transitions are synthesised using the harmonic excitation, and the mode is not changed to waveform coding. Changing the mode for those small variations lead to excessive switching, which may degrade the speech quality, due to the quantisation noise of the waveform coding, when the bit rate of the waveform coder is low. Moreover, the harmonic excitation is capable of producing good quality speech despite those small variations in the waveform. In addition to maintaining the harmonic mode across those minor transitions, in order to limit excessive switching, the harmonic mode is not initialised after ACELP when the speech energy is rapidly decreasing. Rapidly decreasing speech energy indicates an offset, and at some offsets the coding mode may fluctuate between ACELP and harmonic, if extra restrictions are not imposed. At such offsets, the accumulated error in the LPC memories through the harmonic mode is corrected by switching to the ACELP mode, which in turn cause to switch back to the harmonic mode. The additional measures taken to eliminate those fluctuations are described below in subsection: Switching to the harmonic mode.

The estimated normalised cross correlation and squared error values are logically combined to increase the reliability of the analysis by synthesis transition detection. The combinations and thresholds are determined empirically by plotting the parameters with the corresponding speech waveforms. This heuristic approach is superior to a statistical approach, because it allows to include the most important transitions, while
6.6. Speech classification

Figure 6.23: Squared error $E_t$, $E_{i,*}$, and cross correlation $R_t$, $R_{i,*}$ values

the less important ones can be given a lower priority. The analysis by synthesis transition detection compares the harmonically synthesised speech with the original speech, verifies the accuracy of the harmonic model parameters, and decides to use ACELP when the harmonic model fails. This transition detection strategy has the additional advantage of adapting to noisy background conditions, see section 8.2.
Switching to the harmonic mode

In order to avoid mode fluctuations at the offsets, extra restrictions are imposed when switching to the harmonic mode after waveform coding. The rms energy of the speech and the LPC residual are computed for each frame, and a hysteresis loop is added using a control flag. The flag is set to zero when the speech or the LPC residual rms energy is less than 75% of the corresponding rms energy values of the previous frame. The flag is set to one when the speech or the LPC residual rms energy is more than 125% of the corresponding rms energy values of the previous frame. The flag is set to zero if the pitch is greater than 100 samples, regardless of the energy, and it is not changed in any other cases. When switching to harmonic mode after waveform coding, the control flag should be one, in addition to the mode decision of closed loop transition detection. The flag is checked only at a mode transition, once the harmonic mode is initialised the flag is ignored. This process avoids excessive switching at the offsets.

The pitch is used to change the control flag due to different reasons. ACELP produces better quality for male speech with long pitch periods than the harmonic coders even at stationary voiced segments. When the pitch period is long ACELP needs only fewer pulses in the time domain to track the changes in the speech waveform, while the harmonic coders have to encode a large number of harmonics in the frequency domain. Furthermore, it is a well known fact that speech coding schemes which preserve the phase accurately work better for male speech, while the harmonic coders which encode only the amplitude spectrum result in better quality for female speech [91].

6.6.3 Plosive detection

The unvoiced synthesis process described in section 6.5.5 updates the unvoiced gain every 20 ms. While this is sufficient for fricatives, it reduces the quality of the highly non stationary unvoiced components such as plosives. The listening tests show that synthesising plosives using ACELP preserves the sharpness of the synthesised speech and improves the perceptual quality. Therefore a special routine is designed to detect the plosives, which are classified as unvoiced by the initial classification, and synthesise them using ACELP.
Plosives are characterised by isolated pulse like signals with a sharp rise in energy, and this feature is used to distinguish them from the fricatives. The rms energy, $e_j$ of the speech signal is computed for every 10 samples as follows:

$$e_j = \sqrt{\frac{\sum_{n=0}^{9} s^2 (10j + n)}{10}} \quad \text{for } 0 \leq j < 15$$  \hspace{1cm} (6.54)

A plosive detection metric $p_j$ is defined as,

$$p_j = \frac{e_j}{8e_{j-1}}$$  \hspace{1cm} (6.55)

Where $e_{-1}$ is the final energy term of the previous frame. A frame is classified as containing a plosive if $p_j > 1$ at least for one $j$. This algorithm may signal a plosive even when the overall energy level is very low, for example at silence segments, if it detects a large fluctuation in energy. Those low energy segments are ignored by using the tracked energy term, $t_e$ used in the open loop initial classification, see equation 6.48.

It should be noted that the scope of this plosive detection algorithm is reduced to unvoiced segments, since the segments, which include the unvoiced plosives are already identified by the initial classification. The plosive detection algorithm may erroneously identify the highly non stationary onsets and the speech signal near the glottal excitation of the long pitch period segments as plosives, if applied to voiced speech.

Figure 6.24 (a) illustrates the plosive detection metric $p_j$ and an example of a plosive. Figure 6.24 (b) illustrates the detected plosive synthesised using 4 kbps SB-LPC, and 3.7 kbps ACELP (without LTP), see section 7.4. ACELP is used only for the frame, which has the plosive, the rest of the segment is synthesised using white noise excitation. SB-LPC synthesises the speech segment using the noise excitation, which can not adequately represent the plosive.
6.7 Hybrid decoder

A simplified block diagram of the hybrid decoder is presented in Figure 6.25. The decoder extracts the excitation parameters according to the mode and uses the appropriate excitation generation. The synthesised excitation is then fed into the LPC synthesis filter, which produces the final synthetic speech output. The LPC parameters are common for all the modes, and linearly interpolated in the LSF domain with an update interval of 5 ms. The excitation vector is also fed into the ACELP excitation and harmonic excitation generators. The ACELP excitation updates the long term prediction (LTP) buffer with the previous LPC excitation. The harmonic excitation uses the previous excitation at the onsets to initialise the interpolation and prediction parameters.

Figure 6.26 depicts an original speech sample, synthesised speech at the decoder, and the mode used for each synthesis frame. The frame boundaries are also shown in the figure.
6.8 Performance evaluation

The previous sections presented the detailed design and the complete framework of a novel hybrid coding algorithm. The major tasks were developing a reliable classification technique and preserving the phase continuity when switching between the coding modes. The classification algorithm is tested using 64 seconds of modified IRS filtered speech, by comparing the mode decision against manually classified waveforms. Eight English sentence pairs uttered by four male and four female speakers, taken from the NTT (Nippon Telegraph and Telephone) [108] speech database are used as the test material. The silence segments are excluded from the analysis, and synthesised using white noise excitation. The initial classification detects all the voiced frames. Therefore the worst possible classification error, i.e. classifying a voiced frame as unvoiced is eliminated. More than 90% of the unvoiced frames are also detected and the rest of the unvoiced frames are classified as voiced. This bias towards voiced is preferred to miss classifying voiced frames as unvoiced. Since the unvoiced frames classified as voiced will be classified as ACELP by the secondary classification, while a miss classified voiced frame will be synthesised using white noise excitation. The plosive detection algorithm detects all the plosives in the unvoiced frames and does not miss classify other unvoiced frames as plosives.

The transition frames are manually marked by observing the waveforms, in order to test
6.8. Performance evaluation

Figure 6.26: Synthesised speech and classification, A: ACELP, H: Harmonic, and N: Noise excitation

the closed loop transition detection algorithm. Speech frames which have irregular pitch periods and show large variations in the energy are identified as transitions. The closed loop transition detection classifies the frames already classified as voiced by the initial classification into transitory and harmonic. Consequently, all the frames classified as voiced by the initial classification are included in the test, and the unvoiced frames, which are classified as voiced are marked as transitions, since they are expected to be synthesised using ACELP. When testing the transition detection algorithm the use of waveform coding for pitch periods longer than 100 samples is not activated, see section 6.6.2. The transition detection algorithm detects more than 90% of the transition frames and the rest of the transitions are classified as harmonic frames. It also detects more than 90% of the harmonic frames and the rest of the stationary voiced frames are classified as transitions.

Miss classifications may restrict maximising the speech quality because of not choosing
6.8. **Performance evaluation**

the best coding algorithm. However miss classifications of the secondary classification do not degrade the speech quality, due to its closed loop nature. A miss classification of a stationary voiced segment as a transition indicates a harmonic parameter estimation error, and such frames are synthesised using ACELP, perhaps a better solution than synthesising with the erroneous harmonic parameters. A miss classification of a transition as stationary voiced indicates that the harmonic mode is capable of synthesising the particular transitory frame. This may be possible at some transitions, particularly offsets, which usually have a steady pitch contour and a smooth energy variation, where the harmonic interpolation model can fit in.

The phase continuity is tested by listening to the synthesised speech, without introducing quantisation. The tests verify the validity of the hybrid model and there are no perceptible discontinuities. The speech synthesised also indicate the upper bound of the quality achievable by the designed hybrid model. An informal listening test was conducted using 128 kbps linear Pulse Code Modulation (PCM), i.e. the the best narrow band speech quality and 8 kbps ITU G.729, a toll quality speech coder as the reference coders. The speech material used for the test consists of 8 sentences, 4 from male and 4 from female talkers, filtered by modified IRS filter, and a pair of headphones was used to conduct the test. Twelve listeners were asked to indicate their preferences for the randomised pairs of synthesised speech. Both the experienced and inexperienced listeners were participated in the test. The subjective test results are shown in Tables 6.1 and 6.2. The unquantised hybrid model performs better than G.729 and worse than 128 kbps linear PCM. Therefore the quality of the unquantised hybrid model is higher than toll quality and lower than the transparent quality. Some listeners have marked that the unquantised hybrid model is slightly preferred to the 128 kbps linear PCM. However in general the speech encoded and decoded with unquantised model parameters does not sound better than the original speech material, and in this case the synthesised speech quality is very close to the original. The perceived speech quality shows only a slight degradation, even after quantising the harmonic mode parameters at 4 kbps and white noise excitation at 1.5 kbps, with unquantised transitions (128 kbps linear PCM). The hybrid coder achieves toll quality when transitions are quantised with 6 kbps ACELP. See the next chapter for details of quantising the parameters.
6.9 Concluding remarks

In this chapter the principle techniques behind a novel hybrid coding algorithm, which integrates harmonic coding and waveform coding have been presented. The integration of the coding techniques allows to overcome the limitations of each technique under different speech characteristics. The designed hybrid coder is based on a new phase model for the harmonic excitation, Synchronised Waveform matched Phase Model (SWPM). SWPM synchronises and preserves the waveform similarity of the harmonic excitation with the corresponding LPC residual, and facilitates switching between the coding modes. The overall hybrid coder operates in three modes: harmonic excitation for stationary voiced segments, waveform coding for transitions, and white noise excitation for unvoiced segments.

A robust, low complexity two stage classification algorithm was also designed. The first stage of classification decides to use either the noise excitation or one of the other modes. The closed loop secondary classification synthesises speech using SWPM, and switches to waveform coding when the performance of the harmonic excitation is not satisfactory. The classification algorithm has a very good overall accuracy, and employs the coding modes such that occasional miss classifications do not significantly degrade
the speech quality.

Subjective listening test results confirm that the designed hybrid model eliminates the limitations of the existing single model based coders. The listening tests also verify the accuracy of the classification algorithm. In addition to selecting the best coding algorithm, hybrid coding offers the flexibility of tailoring the coding algorithms for the intended types of speech characteristics. This fact is considered when quantising the parameters as described in the next chapter. The coders developed using the designed hybrid model operating at different bit rates are also described in the next chapter.
Chapter 7

Variable bit rate hybrid coding

7.1 Introduction

The speech model parameters should be quantised prior to the transmission in order to accommodate in the bandwidth of the transmission channel. At the decoder end, the dequantisers extract the model parameters from the received bit stream. Inevitably the parameter quantisation degrades the quality of synthesised speech, by adding quantisation noise. The quantisation noise may be reduced by increasing the precision at the expense of an increase in the transmission bit rate.

The hybrid speech coding model introduced in chapter 6 can be adopted for various applications with different quality requirements by quantising the model parameters at different bit rates. For applications, which support variable bit rates the model parameters of different modes may be quantised at different bit rates, allocating the minimum number of bits required for each mode to maintain adequate quality. Variable bit rate coders are particularly advantageous for voice storage, Code Division Multiple Access (CDMA) [18] wireless networks, packet switched networks, and statistical multiplexing of speech for multi channel communications. In the future a large fraction of speech coders will be based on variable bit rate algorithms, with the emergence of the third generation mobile communications network, Universal Mobile Telecommunication System (UMTS) [109], which supports variable bit rate transmissions.
7.2. Unvoiced gain quantisation

The LPC parameters are common for all the modes, and quantised using a fixed number of bits per frame. This is advantageous under noisy channel conditions, since the LPC parameters can be decoded correctly even when the mode bits are in error. The LPC parameters are quantised in the LSF domain using a Multi Stage Vector Quantiser (MSVQ), with a first order Moving Average (MA) prediction [58], see section 3.4. The following sections describe the quantisation of the excitation parameters of the three modes: unvoiced, stationary voiced, and transitions, which use white noise excitation, harmonic excitation with Synchronised Waveform matched Phase Model (SWPM), and Algebraic Code Excited Linear Prediction (ACELP) respectively.

7.2 Unvoiced gain quantisation

The designed hybrid coding algorithm synthesises unvoiced speech using scaled white Gaussian noise as the LPC excitation, see section 6.5.5. Therefore in addition to the LPC parameters only a gain term is required to synthesise unvoiced speech. In order to synthesise the unvoiced plosives with adequate quality the gain term should be updated at least every 5 ms. However the listening tests show that synthesising plosives using ACELP gives better perceptual quality. Therefore the plosives are synthesised using ACELP. The energy of the fricatives does not show rapid fluctuations and updating at the frame rate, i.e. at 20 ms is adequate to synthesise high quality unvoiced fricatives.

The unvoiced gain $g_{uv}$ is quantised using a logarithmic scalar quantiser. The quantised unvoiced gain $g_{uv_i}$ is given by,

$$g_{uv_i} = k \left( \frac{g_{\text{max}} + k}{k} \right)^{\frac{1}{N-1}} - k \quad \text{for } i = 0, 1, 2, \ldots, N - 1$$

(7.1)

Where $N$ is the number of quantiser levels, $g_{\text{max}}$ defines the upper limit of $g_{uv_1}$, and $k$ is a constant which controls the gradient of the exponential function. All the $g_{uv}$ values larger than $g_{\text{max}}$ are clipped at $g_{\text{max}}$, and $g_{\text{max}}$ is chosen as 904 by observing the unvoiced gain values, $g_{uv}$ of 128 seconds of male and female speech. The constant $k$ is set as 16, and 32 quantiser levels were sufficient to produce high quality unvoiced
7.3 Harmonic parameter quantisation

speech, i.e. $N$ is 32. Hence 5 bits are required to transmit the quantised unvoiced gain, $g_{uv}$. Figure 7.1 depicts a plot of the unvoiced gain quantiser levels, with the selected parameter values.

7.3 Harmonic parameter quantisation

The stationary voiced speech segments are synthesised using the synchronised harmonic excitation model described in section 6.5.1. The model parameters of the harmonic excitation with Synchronised Waveform matched Phase Model (SWPM) are pitch period, Pitch Pulse Location (PPL), Pitch Pulse Shape (PPS), harmonic amplitudes, and gain. The pitch period, PPL, PPS, and gain are scalar quantised and the harmonic amplitudes are vector quantised.

The analysis by synthesis transition detection algorithm synthesises the harmonic excitation using SWPM at the encoder to evaluate the suitability of the harmonic mode, see section 6.6.2. Therefore at the encoder quantised or unquantised harmonic parameters may be used for the transition detection. Generally the analysis by synthesis algorithms include the quantisation in the error minimisation loop, so that the quantisation noise...
is also accounted in the parameter estimation process. However in this case the solution is not straightforward, since the decision is between two modes, rather than the best set of parameters of a unimodal coder. One solution to this problem is to perform a full closed loop mode decision with quantised parameters, i.e. synthesising the speech frames with all the modes and selecting the best mode. A weighting factor may be required in the mode selection process, since the harmonic excitation with SWPM may give superior perceptual quality even with a slightly lower SNR compared to ACELP. However such a solution is computationally demanding, since ACELP excitation should be computed for all the frames, excluding the silence and unvoiced frames. Furthermore, defining a suitable weighting factor which reflects the perceptual quality is a difficult task.

A more practical solution is to assume the inclusion of the harmonic parameter quantisation in the mode decision loop depending on the ACELP bit rate. The inclusion of the harmonic quantisation in the closed loop mode decision increases the number of ACELP mode frames. Usually by extending the transitions detected with unquantised harmonic parameters. However occasionally new frames may be switched to ACELP in between the harmonic frames, which severely degrades the perceptual quality, when the bit rate of the ACELP mode is below 8 kbps, due to the sudden discontinuities introduced in the voiced harmonics. In general ACELP operating at 8 kbps or higher is capable of synthesising perceptually superior speech compared to harmonic coding, even at the stationary voiced segments. Therefore the harmonic quantisation can be included in the closed loop mode decision. However when the bit rate of the ACELP mode is low, the quantisation noise becomes audible, hence trying to eliminate the quantisation noise of the harmonic mode by switching to ACELP mode does not improve the perceptual quality. Therefore in the experiments described in this chapter and in the following chapter harmonic parameter quantisation is not included in the transition detection loop.

The sensitivity of the analysis by synthesis transition detection is different for each parameter. The sensitivity is high for the pitch period and PPL. Changes in these parameters dramatically reduce the cross correlation of the original and the synthesised speech, due the resulting time shifts. The spectral amplitudes and the LPC parameters
7.3. Harmonic parameter quantisation

are least sensitive. In fact using either the quantised or unquantised LPC parameters produced the same classification decisions for the speech material used to test the speech classification. The LPC memory locations of the transition detection algorithm are initialised for each frame with the memory locations of the LPC synthesis filter, before weighting, maintained at the encoder, see section 6.5. This avoids drifting the LPC synthesis filter of the transition detection algorithm from the synthesised speech.

Pitch quantisation

The pitch period, $\tau$ is quantised using a non linear scalar quantiser, reflecting the high sensitivity of the human ear to the pitch deviations at the shorter pitch pitch periods. A logarithmic scale is used for the pitch values from 16 to 60 samples and a linear scale is used for the pitch values from 60 to 160 samples. The quantised pitch $\tau_i$ is given by,

$$\tau_i = \tau_{\min} \left( \frac{\tau_0}{\tau_{\min}} \right)^{\frac{i}{N_0 - 1}} \text{ for } i = 0, 1, 2, ..., N_0 - 1$$  (7.2)
7.3. Harmonic parameter quantisation

\[ \tau_i = \tau_0 + \frac{\tau_{\text{max}} - \tau_0}{N - N_0} (i - N_0 + 1) \quad \text{for} \quad i = N_0, N_0 + 1, \ldots, N - 1 \]  \hspace{1cm} (7.3)

Where \( \tau_{\text{min}} \) is 16, \( \tau_{\text{max}} \) is 160, \( \tau_0 \) is 60, \( N_0 \) is 156, and \( N \) is 256. Therefore 8 bits are required to transmit the quantised pitch period. Figure 7.2 depicts a plot of the pitch quantiser levels.

Pitch Pulse Location (PPL) quantisation

The Pitch Pulse Location (PPL) is the location of the pitch pulse closest to the centre of the analysis frame, see section 6.4. PPL may be defined as the distance to the pitch pulse concerned from the centre of the analysis frame, measured in samples. Assuming the maximum possible pitch is 160 samples, PPL varies between -80 and 80. However the pitch pulse location may be normalised with respect to the pitch so that the PPL varies between -0.5 and 0.5. Normalisation of the PPL with respect to the pitch ensures the efficient use of quantiser dynamic range regardless of the pitch, and has the effect of more accurate quantisation of the PPL when the pitch period is shorter.
7.3. Harmonic parameter quantisation

The accuracy of the PPL is more important when it is close to the centre of the analysis frames or the synthesis frame boundaries, i.e. PPL values close to zero. This is due to the fact that the mode changes between ACELP and harmonic excitation may take place at the synthesis frame boundaries. Preserving the continuity of the high energy pitch pulses occurring at or close to the switching frame boundaries is essential to eliminate audible switching artifacts. Therefore the normalised PPL is quantised using a logarithmic scale, quantising the PPL values close to zero more accurately. The quantised PPL, $t_i$ is given by,

$$t_i = k \left( \frac{0.5 + k}{k} \right)^{i+N/2-1} - k \quad \text{for} \quad i = N/2 - 1, N/2, \ldots, N - 1$$

(7.4)

$$t_i = t_{N-2-i} \quad \text{for} \quad i = 0, 1, \ldots, N/2 - 2$$

(7.5)

Where $N$ is the number of quantiser levels and $k$ is a constant, which controls the gradient of the exponential function. The constant $k$ is set as 0.125, and 128 quantiser levels were sufficient to eliminate audible switching artifacts, i.e. $N$ is 128. Hence 7 bits are required to transmit the quantised normalised PPL. PPL is normalised using the quantised pitch so that the decoder can denormalise the received PPL value accurately. Figure 7.3 depicts a plot of the normalised PPL quantiser levels, with selected parameters.

Pitch Pulse Shape (PPS) quantisation

The pitch pulse shape depends on the polarity of the input speech, see Figure 7.4. Inverting the input speech also inverts the residual signal. Figure 7.5 depicts the PDF of 400 PPS values taken from voiced residual signals of male and female speech with positive pulses. The PDF of negative pulses are centred around $\pi$ instead of zero.

The PDF suggests that the overall quantisation noise may be reduced by using a non-linear quantiser, which quantises the PPS values close to zero more accurately, provided
7.3. Harmonic parameter quantisation

Figure 7.4: Effect of changing the polarity of input speech

that the polarity of the input speech signal is set such that the residual pulses are positive. However such a quantisation scheme results in larger quantisation errors in the PPS values close to $\pi$. Large variations in the PPS introduces a reverberant character in the synthesised speech, regardless of the PPS value. Therefore in terms of the perceptual quality all the PPS values are equally important, and a linear quantiser is employed to quantise the PPS using 16 values. The quantised PPS, $\theta_i$ is given by,

$$\theta_i = \frac{2\pi}{N} i - \pi \quad \text{for} \quad i = 0, 1, \ldots, N - 1 \quad \text{and} \quad -\pi \leq \theta_i < \pi$$

(7.6)

Where $N$, the number of quantiser levels is 16 and 4 bits are required to quantise PPS.

7.3.1 Harmonic amplitude quantisation

Harmonic amplitudes of the LPC residual are quantised using Switched Predictive Mel scale based Vector Quantisation (SP-MVQ) [110]. Figure 7.6 depicts a block diagram of SP-MVQ. SP-MVQ converts the variable dimension spectral amplitude vectors into fixed dimension vectors by warping the frequency axis using a logarithmic scale. The warping process emphasises the low frequencies, taking into account the perceptual preferences of the human auditory system. The fixed dimension spectral vector, $\hat{z}$ is
7.3. Harmonic parameter quantisation

decomposed into a predicted vector, $\hat{z}_p$ and a prediction residual vector, $\hat{z}_r$ as follows:

$$\hat{z} = \hat{z}_p + \hat{z}_r$$  \hspace{1cm} (7.7)

Where the predicted vector, $\hat{z}_p$ is obtained using a first order auto regressive method, given by,

$$\hat{z}_p = \Phi (\hat{z}_{-1} - \hat{z}_m) + \hat{z}_m$$  \hspace{1cm} (7.8)

Where $\hat{z}_{-1}$ is the most recently quantised $\hat{z}$, $\hat{z}_m$ is the mean vector, and $\Phi$ denotes a diagonal matrix of prediction coefficients. The prediction residual, $\hat{z}_r$ is quantised using a typical vector quantiser such as Multi Stage VQ (MSVQ) [56], see section 3.4.1. The quantisation becomes Memoryless Mel scale based Vector Quantisation (ML-MVQ) if all the prediction coefficients are zero, and auto regressive Predictive MVQ (P-MVQ) otherwise. The predictive scheme is effective in stationary regions, and may increase spectral distortion at the transitions. Therefore a switching scheme is introduced to switch between P-MVQ and ML-MVQ. The decision between P-MVQ and ML-MVQ is made using analysis by synthesis techniques and based on a weighted spectral distortion.
7.3. Harmonic parameter quantisation

Therefore the quantisation scheme is called Switched Predictive Mel scale based Vector Quantisation (SP-MVQ). Moreover the switching scheme restricts the error propagation under noisy channel conditions.

SP-MVQ quantises spectral amplitudes every 10 ms using 14 bits. The harmonic analysis/synthesis scheme designed in chapter 6 estimates the harmonic parameters every 20 ms. However there are sufficient bits for the allocation of 28 bits per 20 ms frame for spectral amplitudes at 4 kbps, see Table 7.3. Therefore the harmonic analysis/synthesis scheme is modified to update the spectral amplitudes every 10 ms. However the pitch is transmitted only every 20 ms, and linearly interpolated to compute the number of harmonics corresponding to the centre of the synthesis frame or the first sub frame, at the decoder. The spectral amplitude quantisation uses the quantised (second sub frame) or quantised and interpolated (first sub frame) pitch to compute the number of harmonics, in order to ensure the correct dequantisation of the spectral amplitude vectors. In the spectral amplitude quantisation of the first sub frame, if the actual number of harmonics is greater than the computed number of harmonics by interpolation,
the higher harmonics are ignored. If the actual number of harmonics is less than the computed number of harmonics by interpolation, the amplitude vector is zero padded. Usually the pitch values of the stationary voiced segments are fairly unchanged and linear interpolation of the number of harmonics gives a good approximation.

**Harmonic gain quantisation**

The spectral amplitude vectors are normalised before the quantisation, in order to improve the dynamic range. The shape component of the vectors are quantised using Switched Predictive Mel scale based Vector Quantisation (SP-MVQ) as described above, and the gain component is scalar quantised.

The normalised amplitude, $a_{kn}$ of the $k^{th}$ harmonic is given by,

$$a_{kn} = \frac{a_k}{g}$$  (7.9)

Where $a_k$ is the spectral amplitude estimated for the $k^{th}$ harmonic and $g$ is the normalisation factor, given by,

$$g = \sqrt{\frac{K}{\sum_{k=1}^{K} a_k^2}}$$  (7.10)

Where $K$ is the total number of harmonics. Normalisation factor of the second sub frame, $g_2$ is quantised using a logarithmic scale, given by,

$$g_{2i} = k \left( \frac{g_{\text{max}} - g_{\text{min}} + k}{k} \right)^{\frac{1}{N}} - k + g_{\text{min}} \quad \text{for} \quad i = 0, 1, \ldots, N - 1$$  (7.11)

Where $k$ is 8, a constant which controls the gradient of the exponential function, $N$, the number of quantiser levels is 32, i.e. 5 bits are required to quantise the gain of the second sub frame, and $g_{\text{max}}$ and $g_{\text{min}}$ are the maximum and minimum possible quantised normalisation factors respectively. The gain values beyond $g_{\text{max}}$ and $g_{\text{min}}$ are clipped by the quantiser, and $g_{\text{max}}$ and $g_{\text{min}}$ are set as 392.5 and 0.5 respectively, determined by observing the normalisation factors of voiced segments of 128 seconds of male and female speech. The term $g_{\text{min}}$ is introduced in equation 7.11, because only the stationary voiced segments are synthesised using the harmonic excitation and the minimum gain is non zero.
Normalisation factor of the first sub frame, $g_1$, is differentially quantised with respect to the mean of the adjacent two quantised $g_2$ values, as follows:

$$
\delta = g_1 - \frac{g_2 + g_{2-1}}{2}
$$

(7.12)

Where $g_{2-1}$ is the gain of the second sub frame of the previous frame, i.e. the previous $g_2$, and $\delta$ is quantised using 3 bits as follows:

$$
\delta \in \{-20.0, -10.0, -5.0, 0.0, 5.0, 10.0, 20.0, 30.0\}
$$

(7.13)

Finally the spectral amplitude vectors are denormalised by multiplying with the quantised normalisation factors.

### 7.3.2 Onset harmonic parameter quantisation

The harmonic synthesis process interpolates the parameters between the synthesis frame boundaries. However at the onsets, when switching from the waveform coding mode, the harmonic parameters of the initial synthesis frame boundary are not directly available. The techniques developed to overcome this problem are described...
in section 6.5.4. In section 6.5.4 it was suggested to use the spectral amplitudes corresponding to the end of the synthesis frame boundary for the whole synthesis frame. However since the spectral amplitude update rate is increased to twice in a 20 ms frame, see section 7.3.1, the spectral amplitudes of the first sub frame are used during the first half of the synthesis frame and interpolated as usual for the second half. The pitch, PPL, and PPS are estimated as described in section 6.5.4, and quantised as described in the preceding sections.

The quantisation of the spectral amplitudes introduced two additional static variables, which should be initialised at the onsets. The two static variables are the previously quantised spectral vector, \( \hat{z}_{-1} \) and the gain of the second sub frame of the previous frame, \( g_{2,-1} \), see section 7.3.1. These parameters are initialised using the synthesised ACELP excitation signal of the previous frame. The spectral amplitudes of the ACELP excitation signal are estimated by windowing it using an asymmetric window function given by,

\[
w(n) = 0.54 - 0.46 \cos \left( \frac{\pi}{n_1} \right) \text{ for } 0 \leq n < n_1 \tag{7.14}
\]

\[
w(n) = 0.08 + 0.92 \cos \left( \frac{\pi}{2n_2 - 1} \right) \text{ for } n_1 \leq n < n_1 + n_2 \tag{7.15}
\]

Where \( n_1 \) is 140 and \( n_2 \) is 20, and the window function is depicted in Figure 7.7. The asymmetric window function emphasises the excitation signal close to the switching frame boundary.

Spectral amplitude vector of the windowed ACELP excitation signal is obtained by peak picking of the magnitude spectrum, see section 4.3.2, using the received pitch value for the harmonic frame. The rms normalisation factor of the estimated spectral vector is used as \( g_{2,-1} \) of the harmonic frame. The amplitude quantisation memory, \( \hat{z}_{-1} \) is initialised by quantising the normalised shape vector, while forcing SP-MVQ to use memoryless quantisation, see section 7.3.1.
7.4 Quantising the transitions

The transitions are quantised using Algebraic Code Excited Linear Prediction (ACELP), see section 5.5. The pulse innovation of ACELP is capable of synthesising the highly non stationary transitions. The Long Term Prediction (LTP) is not very efficient at the onsets, since the LTP memory buffer has no information regarding the onsets. However LTP is employed, due to two reasons: LTP reduces the sparse excitation artifacts [100] and synthesises a significant amount of the excitation at the offsets. Moreover at the resonance offsets, where the gain of the excitation signal is small, the LTP gain acts as an adaptive gain term and compensates for an inadequate gain quantisation dynamic range of the innovation pulses. Use of multi tap and fractional delay LTP filters [72] are useful only for stationary voiced segments, consequently only integer delays and single tap filters are used to encode transitions.

The LTP gain is close to unity during the stationary voiced segments. However at the transitions LTP gain shows large variations, due to the large variations in the speech energy. Therefore the LTP gain is quantised using a larger dynamic range. A drawback in allowing gain values larger than unity is the LTP filter may become unstable under erroneous channel conditions, see section 8.3.1. The high energy pulses of plosives are synthesised using only the innovation sequence of ACELP. However the plosives are not classified as a separate mode, instead when a plosive is detected the LTP gain is forced to be zero.

7.5 Variable Bit Rate (VBR) coding

Informal listening tests were performed by quantising the stationary voiced segments at 4 kbps and unvoiced segments at 1.5 kbps, see Table 7.3, with unquantised transitions. The synthesised speech quality shows only a slight degradation when compared with using the unquantised model parameters. The speech quality with unquantised model parameters is nearly transparent, see section 6.8. The approach used to quantise the transitions is reducing the bit rate of the transition quantisation to maintain a toll quality output. The informal listening tests show that quantising the transitions at 6
7.5. Variable Bit Rate (VBR) coding

Table 7.1: LTP Gain quantiser table

<table>
<thead>
<tr>
<th>Index</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTP Gain</td>
<td>0.00</td>
<td>0.15</td>
<td>0.30</td>
<td>0.40</td>
<td>0.50</td>
<td>0.60</td>
<td>0.70</td>
<td>0.80</td>
</tr>
<tr>
<td>Index</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>LTP Gain</td>
<td>0.90</td>
<td>1.05</td>
<td>1.20</td>
<td>2.00</td>
<td>3.50</td>
<td>5.50</td>
<td>8.00</td>
<td>10.00</td>
</tr>
</tbody>
</table>

kbps is sufficient to achieve toll quality. Three versions of the coder are designed with quantising the transitions at 4 kbps, 8 kbps, and 8 kbps. The results of the informal listening tests are shown in section 7.6.

Quantising the transitions at 4 kbps

The 4 kbps version uses 10 ms sub frames. For each sub frame LTP delay, LTP gain, and locations, signs, and the gain of two innovation pulses are transmitted. The innovation gain terms of the two sub frames are normalised with respect to the quantised rms energy of the speech signal and the normalisation factor is transmitted for each 20 ms frame. The normalisation reduces the dynamic range required to quantise the innovation sequence gain. Table 7.3 shows the bit allocation of the 4 kbps ACELP parameters. The LTP delay range is 20 to 147, and only the integer delays are allowed, needing 7 bits for the index. The LTP gain is quantised using 4 bits, and the quantiser values are shown in Table 7.1.

The two innovation pulses cover only the first 64 locations of each 80 sample sub frame. Each pulse is chosen from 32 possible locations, either even or odd, and 5 bits are required to transmit the location. The sign of each pulse is transmitted using 1 bit. The pulse gain and the common normalisation factor of the frame are quantised using 3 bits each, see Table 7.2.

Quantising the transitions at 6 kbps

The 6 kbps version uses 5 ms sub frames. For each sub frame LTP delay, LTP gain, and locations, signs, and the gain of two innovation pulses are transmitted. The pulse gain
7.5. Variable Bit Rate (VBR) coding

Table 7.2: Innovation pulse gain quantiser table

<table>
<thead>
<tr>
<th>Index</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pulse Gain</td>
<td>0.0</td>
<td>0.3</td>
<td>0.7</td>
<td>1.1</td>
<td>1.6</td>
<td>2.1</td>
<td>2.7</td>
<td>3.5</td>
</tr>
<tr>
<td>RMS Gain</td>
<td>10</td>
<td>40</td>
<td>90</td>
<td>176</td>
<td>325</td>
<td>584</td>
<td>1030</td>
<td>1800</td>
</tr>
</tbody>
</table>

terms of the four sub frames are normalised with respect to the quantised rms energy of the speech signal and the normalisation factor is transmitted for each 20 ms frame. Table 7.3 shows the bit allocation of the 6 kbps ACELP parameters. The LTP delay and gain are quantised similar to 4 kbps version, using 7 bits and 4 bits respectively.

The two innovation pulses cover only the first 32 locations of each 40 sample sub frame. Each pulse is chosen from 16 possible locations, either even or odd, and 4 bits are required to transmit the location. The signs of the two pulses are forced to be opposite in the error minimisation process, hence only the sign of the first pulse is transmitted using 1 bit. The pulse gain and the common normalisation factor of the frame are quantised using 3 bits each, see Table 7.2.

Quantising the transitions at 8 kbps

The 8 kbps version uses 5 ms sub frames. For each sub frame LTP delay, LTP gain, and locations, signs, and the gain of four innovation pulses are transmitted. The pulse gain terms of the four sub frames are normalised with respect to the quantised rms energy of the speech signal and the normalisation factor is transmitted for each 20 ms frame. Table 7.6 shows the bit allocation of the 8 kbps ACELP parameters. The LTP delay and gain are quantised similar to 4 kbps version, using 7 bits and 4 bits respectively.

The locations and the signs of the four pulses are shown in Table 7.4. The pulse gain of each sub frame is quantised using 4 bits, as shown in Table 7.5. The common normalisation factor, i.e. the rms energy of the speech signal is logarithmically quantised using 7 bits, and the quantised value, $g_{rms_i}$, is given by,

$$ g_{rms_i} = k \left( \frac{g_{max} - g_{min} + k}{k} \right)^{\frac{1}{N-1}} - k + g_{min} \quad \text{for} \quad i = 0, 1, \ldots, N - 1 $$

(7.16)
### Table 7.3: Bit allocation for a 20 ms frame

<table>
<thead>
<tr>
<th>Parameters</th>
<th>White Noise</th>
<th>Harmonic</th>
<th>ACELP 4k</th>
<th>ACELP 6k</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPC</td>
<td>23</td>
<td>23</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td>Pitch</td>
<td>-</td>
<td>8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PPL</td>
<td>-</td>
<td>7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PPS</td>
<td>-</td>
<td>4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Amplitudes</td>
<td>-</td>
<td>14+14</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Gain</td>
<td>5</td>
<td>3+5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>LTP Delay</td>
<td>-</td>
<td>-</td>
<td>7+7</td>
<td>7+7+7+7</td>
</tr>
<tr>
<td>LTP Gain</td>
<td>-</td>
<td>-</td>
<td>4+4</td>
<td>4+4+4+4</td>
</tr>
<tr>
<td>Pulse Locations</td>
<td>-</td>
<td>-</td>
<td>10+10</td>
<td>8+8+8+8</td>
</tr>
<tr>
<td>Pulse Signs</td>
<td>-</td>
<td>-</td>
<td>2+2</td>
<td>1+1+1+1</td>
</tr>
<tr>
<td>Pulse Gain</td>
<td>-</td>
<td>-</td>
<td>3+3</td>
<td>3+3+3+3</td>
</tr>
<tr>
<td>Mode</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td>80</td>
<td>80</td>
<td>120</td>
</tr>
</tbody>
</table>

### Table 7.4: Structure of the 17 bit algebraic codebook

<table>
<thead>
<tr>
<th>Pulse</th>
<th>Amplitude</th>
<th>Position</th>
<th>Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>± 1</td>
<td>0, 5, 10, 15, 20, 25, 30, 35</td>
<td>1+3</td>
</tr>
<tr>
<td>1</td>
<td>± 1</td>
<td>1, 6, 11, 16, 21, 26, 31, 36</td>
<td>1+3</td>
</tr>
<tr>
<td>2</td>
<td>± 1</td>
<td>2, 7, 12, 17, 22, 27, 32, 37</td>
<td>1+3</td>
</tr>
<tr>
<td>3</td>
<td>± 1</td>
<td>3, 8, 13, 18, 23, 28, 33, 38, 4, 9, 14, 19, 24, 29, 34, 39</td>
<td>1+4</td>
</tr>
</tbody>
</table>

### Table 7.5: Innovation pulse gain quantiser table for 8 kbps ACELP

<table>
<thead>
<tr>
<th>Index</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pulse Gain</td>
<td>0.0</td>
<td>0.15</td>
<td>0.3</td>
<td>0.45</td>
<td>0.6</td>
<td>0.8</td>
<td>1.0</td>
<td>1.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Index</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pulse Gain</td>
<td>1.5</td>
<td>1.8</td>
<td>2.1</td>
<td>2.4</td>
<td>2.8</td>
<td>3.2</td>
<td>3.7</td>
<td>4.3</td>
</tr>
</tbody>
</table>
7.5. Variable Bit Rate (VBR) coding

Table 7.6: Bit allocation of 8 kbps ACELP for a 20 ms frame

<table>
<thead>
<tr>
<th>Parameters</th>
<th>ACELP 8k</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPC</td>
<td>23</td>
</tr>
<tr>
<td>Gain</td>
<td>7</td>
</tr>
<tr>
<td>LTP Delay</td>
<td>7+7+7+7</td>
</tr>
<tr>
<td>LTP Gain</td>
<td>4+4+4+4</td>
</tr>
<tr>
<td>Pulse Locations</td>
<td>13+13+13+13</td>
</tr>
<tr>
<td>Pulse signs</td>
<td>4+4+4+4</td>
</tr>
<tr>
<td>Pulse Gain</td>
<td>4+4+4+4</td>
</tr>
<tr>
<td>Mode</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>160</td>
</tr>
</tbody>
</table>

Where k is 80, a constant which controls the gradient of the exponential function, N, the number of quantiser levels is 128, and $g_{\max}$ and $g_{\min}$ are 2720.5 and 0.5 respectively.

7.5.1 Tailoring for storage

Another important application of variable rate coding is voice storage. The voice storage systems do not suffer from the channel errors common in transmission systems. Therefore the model parameters of the Synchronised Waveform matched Phase Model (SWPM), such as pitch, Pitch Pulse Shape (PPS), and Pitch Pulse Location (PPL) can be quantised differentially. The pitch is fairly stationary during the voiced segments, and PPL can be predicted using the previous PPL and the pitch. PPS of the voiced segments also shows significant intra frame correlations. Furthermore there are no delay restrictions in the storage systems. Therefore the speech classification can be extended to include variable frame lengths. For example longer frames can be used for stationary voiced segments.


7.6 Subjective test results

7.5.2 Performance of the harmonic and waveform coders with the bit rate

Quantisation noise of waveform coding becomes audible below 4 kbps even at the transitions, and the speech quality of parametric coders are more tolerable at those low bit rates. This limits the use of hybrid coding for applications which can offer at least 4 kbps for any frame. One solution for this problem is buffering the output bit stream and transmitting at a constant average bit rate. However this increases the overall delay of the target communication system.

The effect of performance variation of the two fundamental coding paradigms on hybrid coding can be viewed from another angle. When a set of constant bit rate hybrid coders are required the speech classification may be designed as a function of the bit rate. This means designing the speech classifier such that at lower bit rates it is biased towards harmonic coding and at higher bit rates it is biased towards waveform coding. This process involves adjusting the classification thresholds for a given bit rate, so that switching is a function of the given bit rate. For example, when the waveform coding mode is operating at 8 kbps or higher, most of the quantisation noise of the waveform coding is not audible and excessive switching will not create problems. Consequently minor transitions as shown in Figure 6.23 (a) can be switched to waveform coding. The quality improvements of such a classification scheme should be assessed through listening tests, hence designing such a classifier as a function of the coding bit rate is an extremely labour intensive process and has not been investigated during the course of this research.

7.6 Subjective test results

Three informal listening tests were conducted to compare the speech quality of the hybrid coder, with transitions quantised at 4 kbps, 6 kbps, and 8 kbps. The synthesised speech was compared against 5.3 kbps ITU G.723.1, 6.3 kbps ITU G.723.1, and 8 kbps ITU G.729 coders respectively. In all the tests stationary voiced segments are quantised at 4 kbps and silence and unvoiced segments are quantised at 1.5 kbps. The speech
7.6. Subjective test results

Table 7.7: 4 kbps ACELP hybrid vs. 5.3 kbps G.723.1

<table>
<thead>
<tr>
<th></th>
<th>Better</th>
<th>Slightly better</th>
<th>Same</th>
<th>Slightly worse</th>
<th>Worse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male(%)</td>
<td>6.2</td>
<td>34.4</td>
<td>28.2</td>
<td>31.2</td>
<td>0.0</td>
</tr>
<tr>
<td>Female%</td>
<td>9.4</td>
<td>31.2</td>
<td>37.5</td>
<td>18.8</td>
<td>3.1</td>
</tr>
<tr>
<td>Average%</td>
<td>7.8</td>
<td>32.8</td>
<td>32.8</td>
<td>25.0</td>
<td>1.6</td>
</tr>
</tbody>
</table>

Table 7.8: 6 kbps ACELP hybrid vs. 6.3 kbps G.723.1

<table>
<thead>
<tr>
<th></th>
<th>Better</th>
<th>Slightly better</th>
<th>Same</th>
<th>Slightly worse</th>
<th>Worse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male(%)</td>
<td>0.0</td>
<td>31.3</td>
<td>43.7</td>
<td>18.8</td>
<td>6.2</td>
</tr>
<tr>
<td>Female%</td>
<td>6.3</td>
<td>28.1</td>
<td>37.5</td>
<td>21.9</td>
<td>6.2</td>
</tr>
<tr>
<td>Average%</td>
<td>3.2</td>
<td>29.7</td>
<td>40.6</td>
<td>20.3</td>
<td>6.2</td>
</tr>
</tbody>
</table>

Material used for each test consists of 8 sentences, 4 from male and 4 from female talkers, filtered by modified IRS filter, and a pair of headphones was used to conduct the test. Twelve listeners were asked to indicate their preferences for the randomised pairs of synthesised speech. Both the experienced and inexperienced listeners were participated in the test. The subjective test results are shown in Tables 7.7, 7.8, and 7.9.

For the speech material used in the subjective tests, after discarding the silence frames, about 64% of the frames used harmonic excitation, 22% used ACELP, and 14% used white noise excitation. Giving average bit rates of 3.65 kbps, 4.1 kbps, and 4.53 kbps for 4 kbps, 6 kbps, and 8 kbps ACELP mode hybrid coders respectively. The 4 kbps ACELP version performs better than G.723.1 at 5.3 kbps. The 6 kbps ACELP version achieves similar quality to G.723.1 at 6.3 kbps. The quality of the 8 kbps ACELP version is worse than G.729 at 8 kbps. However the average bit rate of the 8 kbps ACELP version is 4.53 and more than 50% of the speech synthesised using the 8 kbps ACELP version was adjudged to be the same quality as G.729.
7.7 Concluding remarks

In this chapter the quantisation of the model parameters of the hybrid coding algorithm developed in chapter 6 has been presented. The transitions are quantised using Algebraic Code Excited Linear Prediction (ACELP).

The white noise excitation is quantised at 1.5 kbps and the harmonic excitation with Synchronised Waveform matched Phase Model (SWPM) is quantised at 4 kbps. Three versions of the coder are developed with transitions quantising at 4 kbps, 6 kbps, and 8 kbps. Subjective listening tests show that the 4 kbps ACELP version performs better than G.723.1 at 5.3 kbps. The 6 kbps ACELP version achieves similar quality to G.723.1 at 6.3 kbps. The quality of the 8 kbps ACELP version is worse than G.729 at 8 kbps. However the average bit rate of the 8 kbps ACELP version is 4.53 and more than 50% of the speech synthesised using the 8 kbps ACELP version was adjudged to be the same quality as G.729.

<table>
<thead>
<tr>
<th></th>
<th>Better</th>
<th>Slightly better</th>
<th>Same</th>
<th>Slightly worse</th>
<th>Worse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male(%)</td>
<td>0.0</td>
<td>9.6</td>
<td>65.4</td>
<td>23.1</td>
<td>1.9</td>
</tr>
<tr>
<td>Female(%)</td>
<td>1.9</td>
<td>11.5</td>
<td>55.8</td>
<td>30.8</td>
<td>0.0</td>
</tr>
<tr>
<td>Average(%)</td>
<td>1.0</td>
<td>10.6</td>
<td>60.5</td>
<td>26.9</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Chapter 8

Acoustic noise and channel error performance

8.1 Introduction

The robustness to background noise and channel errors is an important factor for any practical speech coding algorithm. The speech coders designed for mobile and military communication applications are frequently encountered with acoustic noise and channel errors. The background noise may be suppressed before the encoding process using a noise pre processor [111]. However this involves additional complexity and delay, which may not be desirable for mobile communication applications. Therefore the speech coding algorithms are expected to produce intelligible synthetic speech even at the presence of background noise. Generally analysis by synthesis coders perform better under noisy background conditions, than the parametric coders. This inherent robustness of analysis by synthesis coders is due to their waveform matching process. The error minimisation process attempts to synthesise the input waveform regardless of its contents. The model parameters estimated by the parametric coders may not be accurate when the input speech signal is corrupted with noise. Inaccurate model parameters may severely degrade the synthetic speech of parametric coders.

The channel errors are usually divided into two classes: random errors and burst errors. A speech coding algorithm should provide a reasonable output even if a small proportion
of the received bit stream is incorrect due to random bit errors. Robustness against random channel errors can be increased by means of index assignment algorithms [21], [22], through proper quantiser design, and by adding redundancy into the transmitted information [23], [24], [25]. Unequal error protection techniques may be applied to provide a higher degree of protection to the most sensitive bits. For example in CELP coders spectral envelope parameters are the most sensitive to errors, followed by the fixed codebook gain, the adaptive codebook index, the adaptive codebook gain, the sign of the fixed codebook gain, and the fixed codebook index [112]. In the case of sinusoidal coders the gain is the most sensitive to errors, followed by the voicing, the pitch, the spectral envelope parameters, and the spectral amplitudes [113].

In the case of burst errors, error detection schemes are used to classify each frame of received bits as usable or unusable. Another similar problem encountered in the packet voice communication systems is lost packets due to transmission impairments and excessive delays. In order to reduce the annoying artefacts due to lost frames, concealment techniques based on waveform substitution can be used [114]. The burst errors may also be converted to occur in a more random fashion using interleaving techniques.

The difficulties specific to a hybrid coding algorithm are the robustness of the classification algorithm under acoustic noise and the robustness to mode bit errors due to random channel errors. The hybrid coding algorithm described in the preceding chapters is tested to evaluate the performance of the classification algorithm under noisy conditions and the mode bit errors. The performance under background noise can be improved by switching to waveform coding, by exploiting the inherent robustness of waveform coders.

8.2 Performance under acoustic noise

The classification algorithm is tested using 64 seconds of male and female speech corrupted with either babel or vehicular noise. The SNR of the corrupted speech is 10 dB. The speech classification algorithm has two stages, see section 6.6. The initial classification decides to use the white noise excitation for unvoiced and silence frames,
8.2. Performance under acoustic noise

Figure 8.1: Classification of female speech corrupted by babel noise (10 dB SNR), A: ACELP, H: Harmonic, and N: Noise excitation

and leaves the remaining segments, i.e. transitions and stationary voiced segments to be classified by the secondary classification. The secondary classification decides to use either ACELP for transitions or harmonic excitation for stationary voiced segments. The transitions are detected using an analysis by synthesis technique based on Synchronised Waveform matched Phase Model (SWPM). The secondary classification synthesises speech using the harmonic excitation with SWPM and estimates the cross correlation with the original speech. The frames with high cross correlation values are declared as harmonic. The low cross correlation values indicate that the harmonic model is not suitable. The ACELP mode is primarily included to encode the transitions, when the harmonic model fails.

Figure 8.1 depicts the classification of female speech corrupted with babel noise, with an SNR of 10 dB. The initial classification declares only the strong unvoiced segments as unvoiced and all the other frames are left to encode using either ACELP or harmonic excitation, compare Figures 8.1 (b) and 8.2 (b). The weak unvoiced segments which have lower energy than the noise level are not detected as unvoiced. When corrupted with babel or vehicular noise the silence and the low energy unvoiced segments do not have the properties of unvoiced speech. It can be seen in the figures the energy of the noise component is comparable with speech and it has a significant low frequency
8.2. Performance under acoustic noise

Figure 8.2: Classification of clean speech corresponding to Figure 8.1

component, see Figure 8.5. This is expected since essentially babel noise is attenuated and superimposed speech components. Figure 8.3 shows the classification of male speech corrupted with vehicular noise, with an SNR of 10 dB and Figure 8.4 shows the corresponding clean speech segments.

The secondary classification performs very similar to under the clean speech conditions, see section 6.8. Except the occasional classification of frames as ACELP, which were originally classified as harmonic under the clean speech conditions, compare Figures 8.1 (a) and 8.2 (a). This is due to the inability of the harmonic model to adequately synthesise the corrupted signal and the model parameter estimation errors. Therefore in general at the presence of acoustic noise the speech classification algorithm declares more frames as ACELP. These include the silence frames of the original clean speech, unvoiced segments with lower energy than the noise level, and the stationary voiced frames with parameter estimation and harmonic modelling difficulties.

The white noise excitation or the harmonic excitation is not suitable for synthesising the background noise. The spectra of babel and vehicular noise are not white, even after discarding the spectral envelope, see Figure 8.5. Synthesising them using the white noise excitation will degrade the perceptual quality by introducing an unnaturally noisy background. Therefore in fact the classification algorithm detects the most suitable mode, i.e. ACELP to synthesise background noise. However the drawback is a high average bit rate, which may be reduced by using a robust Voice Activity Detection
8.2. Performance under acoustic noise

The correct classification of the stationary voiced segments as harmonic mode under noisy background conditions confirms the robustness of Synchronised Waveform matched Phase Model (SWPM), since the analysis by synthesis classification algorithm synthesises speech using SWPM. Therefore it can be concluded that the Pitch Pulse Location (PPL) and the Pitch Pulse Shape (PPS) detection algorithms described in section 6.4 perform well under noisy background conditions.

An informal listening test was conducted to compare the speech quality of the hybrid coder, under noisy background conditions, with white noise, harmonic excitation, and ACELP quantised at 1.5 kbps, 4 kbps, and 6 kbps respectively, see chapter 7. The synthesised speech was compared against the same noisy speech files synthesised using the 6.3 kbps ITU G.723.1 coder. The speech material used for each test consists of 8 sentences, 4 from male and 4 from female talkers, 4 corrupted with vehicular noise and 4 corrupted with babel noise (10 dB SNR), and a pair of headphones was used to conduct the test. Twelve listeners were asked to indicate their preferences for the randomised pairs of synthesised speech. Both the experienced and inexperienced listeners were participated in the test. The subjective test results are shown in Table 8.1.
8.2. Performance under acoustic noise

The subjective listening test shows a clear preference to the 6.3 kbps ITU G.723.1 coder. It was found that this is due to the metallic character of the stationary voiced speech synthesised by the harmonic excitation. The stationary voiced speech synthesised by the harmonic excitation is cleaner, however there is a pronounced metallic character. The test confirms that the listeners prefer more natural sounding, noisy speech rather than metallic speech.

8.2.1 Improving the performance under acoustic noise

The metallic character is not so pronounced in noisy speech synthesised using Split Band LPC (SB-LPC) harmonic coder [5]. The SB-LPC coder divides the speech spectrum into two bands using a voicing frequency marker, the upper band is declared unvoiced, and synthesised using a filtered noise excitation, see section 4.4.3. For clean stationary voiced speech, most of the spectrum is declared voiced. However in the case of stationary voiced segments of noisy speech, more frequency bands are declared un-
8.2. Performance under acoustic noise

The harmonic excitation model described in section 6.5.1 was designed to synthesise stationary voiced segments and the complete spectrum is synthesised using harmonically related sinusoids. Under noisy background conditions, there are strong spectral components which are not related to the fundamental frequency of the speech. These noise components change the harmonic amplitudes and perceived as metallic sounds in harmonically synthesised speech. This is illustrated in Figure 8.6. Therefore introducing a voicing frequency marker for the harmonic excitation, similar to SB-LPC improves the speech quality of the hybrid coder, especially in noisy background conditions. The hybrid coding algorithm described in chapter 6 has three modes, and 2 bits are allocated to transmit the mode, see section 7.5. Therefore an additional mode may be added to further improve the speech quality. The quality of speech corrupted by acoustic noise can be improved by using the additional mode as another harmonic mode with a constant voicing frequency marker, e.g. 80% of the spectrum is voiced. Figure 8.7 depicts the spectrum of speech corrupted with babel noise (10 dB SNR) and the spectrum of the synthesised speech, with 80% of the spectrum declared voiced and the remaining high frequency components synthesised using filtered and scaled Gaussian noise.
8.2. Performance under acoustic noise

![Figure 8.6: Speech corrupted with babel noise (10 dB SNR)](image)

(a) Spectrum of corrupted voiced speech (b) Harmonically synthesised speech

**Table 8.2: Hybrid vs. 6.3 kbps G.723.1 for noisy speech**

<table>
<thead>
<tr>
<th></th>
<th>Better</th>
<th>Slightly better</th>
<th>Same</th>
<th>Slightly worse</th>
<th>Worse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male(%)</td>
<td>0.0</td>
<td>12.5</td>
<td>40.0</td>
<td>40.0</td>
<td>7.5</td>
</tr>
<tr>
<td>Female(%)</td>
<td>5.0</td>
<td>22.5</td>
<td>35.0</td>
<td>30.0</td>
<td>7.5</td>
</tr>
<tr>
<td>Average(%)</td>
<td>2.5</td>
<td>17.5</td>
<td>37.5</td>
<td>35.0</td>
<td>7.5</td>
</tr>
</tbody>
</table>

An informal listening test was conducted to compare the speech quality of the hybrid coder with harmonic voicing, under noisy background conditions, with white noise, harmonic excitation with voicing, and ACELP quantised at 1.5 kbps, 4 kbps, and 6 kbps respectively, see chapter 7. The synthesised speech was compared against the same noisy speech files synthesised using the 6.3 kbps ITU G.723.1 coder. The speech material used for each test consists of 8 sentences, 4 from male and 4 from female talkers, 4 corrupted with vehicular noise and 4 corrupted with babel noise (10 dB SNR), and a pair of headphones was used to conduct the test. Twelve listeners were asked to indicate their preferences for the randomised pairs of synthesised speech. Both the experienced and inexperienced listeners were participated in the test. The subjective test results are shown in Table 8.2. Comparing with the results shown in Table 8.1, the introduction of the harmonic voicing significantly improves the performance under background noise.
8.3 Performance under channel errors

The inherent robustness of the hybrid coder to the mode bit errors is tested by simulating all the possible mode errors. The hybrid coder has three modes, hence there are six possible mode errors, i.e. each mode may be erroneously decoded with the other two modes. The bit stream of the hybrid coder is shown in Tables 8.3 and 8.4. For each parameter the Most Significant Bit (MSB) is transmitted first. When erroneously decoding a lower rate mode as a higher rate mode, e.g. decoding a white noise excitation frame as harmonic the remaining bits are set to 1. Simulations show that setting the remaining bits to 1 has the worst effect, since the higher indices are mapped to the higher energy levels in the gain quantisers. Using the LTP gain quantiser shown in Table 7.1 results in blasts when the white noise or harmonic frames are erroneously decoded as ACELP. Therefore the maximum LTP gain is limited to 1.2.

All the modes quantise the LSFs using 23 bits, consequently they are transmitted using the same bits. Therefore the LSFs are independent of the mode and the mode bit errors can only affect the excitation parameters. This is particularly attractive for the LSF interpolation and quantisation with first order moving average prediction, see section 3.4.1. The most significant bits of the gain parameters are also transmitted using the same bits. However the gain of each mode is estimated using different criteria.

(a) Spectrum of corrupted voiced speech  (b) Synthesised speech, 80 % voiced

Figure 8.7: Speech corrupted with babel noise (10 dB SNR)
8.3. *Performance under channel errors*

Table 8.3: Transmission bit stream of the hybrid coder

<table>
<thead>
<tr>
<th>Parameters</th>
<th>White Noise</th>
<th>Harmonic</th>
<th>ACELP 6k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mode</td>
<td>1-2</td>
<td>1-2</td>
<td>1-2</td>
</tr>
<tr>
<td>LSF</td>
<td>3-25</td>
<td>3-25</td>
<td>3-25</td>
</tr>
<tr>
<td>Gain (2nd sub frame)</td>
<td>26-30</td>
<td>26-30</td>
<td>26-28</td>
</tr>
<tr>
<td>Gain 1st sub frame</td>
<td>-</td>
<td>31-33</td>
<td>-</td>
</tr>
<tr>
<td>Pitch</td>
<td>-</td>
<td>34-41</td>
<td>-</td>
</tr>
<tr>
<td>PPL</td>
<td>-</td>
<td>42-48</td>
<td>-</td>
</tr>
<tr>
<td>PPS</td>
<td>-</td>
<td>49-52</td>
<td>-</td>
</tr>
<tr>
<td>Amplitudes 1st sub frame</td>
<td>-</td>
<td>53-66</td>
<td>-</td>
</tr>
<tr>
<td>Amplitudes 2nd sub frame</td>
<td>-</td>
<td>67-80</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 8.4: Bit stream of 6 kbps ACELP sub frames

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Sub frame 1</th>
<th>Sub frame 2</th>
<th>Sub frame 3</th>
<th>Sub frame 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTP Delay</td>
<td>29-35</td>
<td>52-58</td>
<td>75-81</td>
<td>98-104</td>
</tr>
<tr>
<td>LTP Gain</td>
<td>36-39</td>
<td>59-62</td>
<td>82-85</td>
<td>105-108</td>
</tr>
<tr>
<td>Pulse Sign</td>
<td>40</td>
<td>63</td>
<td>86</td>
<td>109</td>
</tr>
<tr>
<td>Pulse track 1</td>
<td>41-44</td>
<td>64-67</td>
<td>87-90</td>
<td>110-113</td>
</tr>
<tr>
<td>Pulse track 2</td>
<td>45-48</td>
<td>68-71</td>
<td>91-94</td>
<td>114-117</td>
</tr>
<tr>
<td>Innovation Gain</td>
<td>49-51</td>
<td>72-74</td>
<td>95-97</td>
<td>118-120</td>
</tr>
</tbody>
</table>

Hence the gain quantisers of each mode have different dynamic ranges. Therefore the mode errors affect the dequantisation of the gain. The gain may be quantised independent of the mode by always estimating using the same criterion, e.g. the energy of the input speech signal. However this has the drawback of inability to use the optimum dynamic range for different speech characteristics, e.g. usually the energy content of voiced speech is higher than unvoiced speech.
8.3. Performance under channel errors

Figure 8.8 illustrates erroneous decoding of white noise excitation frames as harmonic and ACELP. The figures illustrate that the errors are contained within the frames which have mode errors. This is because the decoder does not interpolate the unvoiced gain at switching. The present gain is used to synthesise the entire frame when switched from a different mode. However if the next frame after decoding a noise excitation frame as ACELP is also ACELP, the LTP memory propagates the errors, similar to the error propagation of CELP coders [112]. The hybrid coding algorithm has the advantage of limiting the error propagation due to the LTP memory, by switching to a different mode, which also refreshes the LTP memory, see section 6.5.

Figure 8.9 illustrates erroneous decoding of harmonic excitation frames as unvoiced and ACELP. The figures illustrate that the errors are contained within the frames which have mode errors. This is because the decoder reinitialises the harmonic excitation memories when switched from a different mode, and the use of the previous excitation vector is minimised, see section 6.5.4. However if the next frame after decoding a
8.3. Performance under channel errors

(a) Decoding a harmonic frame as unvoiced  (b) Decoding a harmonic frame as ACELP

Figure 8.9: Erroneous decoding of harmonic excitation frames, (i) Original speech, (ii) Synthesised speech, A: ACELP, H: Harmonic, and N: Noise excitation

harmonic excitation frame as unvoiced is also unvoiced, the unvoiced overlap and add process spreads the incorrect gain into the next frame.

Erroneously decoding the ACELP frames

Figure 8.10 illustrates erroneous decoding of ACELP frames as unvoiced and harmonic. In Figure 8.10 (a) the error is contained with the frame which has the mode error. For the next frame the harmonic mode reinitialises the excitation memories, see section 6.5.4. However in Figure 8.10 (b) the next frame after decoding an ACELP frame as harmonic is also harmonic. Hence the error propagates into the next frame, due to the harmonic interpolation process.

The LPC filter may propagate the errors, when the filter response is highly resonance. However the bandwidth expansion of the LPC coefficients ensures that the LPC impulse response dies away quickly. Therefore all the mode errors are localised and the output does not become unstable at the presence of mode errors. This is mainly due to the independent memory initialisation procedures of the coding algorithm when switching between the modes. The white noise excitation mode always sets the previous gain equal to the present one when switched from a different mode. The harmonic excitation
8.3. Performance under channel errors

(a) Decoding an ACELP frame as unvoiced (b) Decoding an ACELP frame as harmonic

Figure 8.10: Erroneous decoding of ACELP frames, (i) Original speech, (ii) Synthesised speech, A: ACELP, H: Harmonic, and N: Noise excitation

mostly depends on the received harmonic parameters when switched from a different mode, only the amplitude quantiser memories are initialised using the previous excitation vector, see section 7.3.2. The LTP buffer is refreshed regardless of the mode with the latest excitation vector, see section 6.5

8.3.1 Improving the performance under channel errors

During the experiments described in the preceding sections the robustness to mode bit errors was improved by limiting the LTP gain to 1.2 and using the same set of bits to transmit the LSFs of all the modes. The encoder and the decoder can not synchronise the random number generators at the presence of mode bit errors. This affects the performance of the LTP when switched from the white noise excitation. However the exact content of the white noise excitation has no significance, and can be represented by any noise excitation vector. Therefore the performance of the LTP was also improved by always reinitialising the LTP buffer to a fixed stored noise excitation vector when switched to ACELP from the white noise excitation.

The robustness to mode bit errors can be further improved by using error detection and correction techniques. If a mode error is only detected and not corrected, the conceal-
ment techniques based on waveform substitution can be used to reduce the resulting annoying artifacts [114]. The decoded parameters and the synthesised waveform may also be used to detect mode errors. As can be seen in the Figures 8.8, 8.9, and 8.10, in general mode errors result in sudden changes in the waveform shape and the signal level, which are unusual for speech signals. Many parameters such as pitch are not quantised using differential techniques, checking their continuity is a useful measure against channel errors. Moreover certain mode patterns are common than the others, e.g. for many speech utterances ACELP to harmonic and back to ACELP occur, while the silence segments before and after are synthesised with the white noise excitation. The transition from white noise to harmonic mode is extremely rare, since generally the onsets request ACELP. Consequently may be abandoned from the scheme, in order to assist detecting the mode errors by limiting the possible switching combinations.

8.4 Concluding remarks

The robustness of the hybrid coding algorithm designed in the preceding chapters has been tested under acoustic noise and channel error conditions. The difficulties specific to hybrid coders are the speech classification under background noise and the mode bit errors due to random channel errors. The classification algorithm is capable of selecting the best mode under noisy background conditions. There is a significant bias towards ACELP at the presence of noise compared to clean speech conditions. This is due to the inability of the white noise excitation or the harmonic excitation to encode the corrupted signals. The noisy speech synthesised using the harmonic mode sounds metallic, which can be reduced by introducing a voicing frequency marker.

The robustness of the hybrid coder to mode errors has been tested by simulating all the possible mode errors. The coder is capable of isolating the mode errors and return to normal decoding almost immediately. This is mainly due to the independent memory reinitialisation of the modes when switched from a different mode.
Chapter 9

Conclusions

9.1 Preamble

The subject of this thesis has been the design, development and testing of a novel hybrid speech coding algorithm. The designed hybrid coder combines two fundamental coding techniques to overcome the limitations of each other. Harmonic coders perform well for stationary voiced speech. However the harmonic model is not suitable to encode speech transitions. The speech quality of Algebraic Code Excited Linear Prediction (ACELP) coders based on Analysis by Synthesis (AbS) techniques degrades as the bit rate is reduced. The stationary voiced speech synthesised using ACELP suffers a particular type of granular noise due to the inadequate representation of the LPC excitation. Therefore a combined hybrid coding algorithm has the potential to overcome the limitations of harmonic and ACELP coders at low bit rates. The work presented can be broadly divided into three main areas.

1. A comprehensive study of the harmonic and ACELP coding techniques, and identifying their limitations and merits. Studying the principle techniques behind the existing hybrid coders and identifying their strengths and shortcomings.

2. Designing a hybrid coding algorithm which overcomes the shortcomings of the existing hybrid coders. This was achieved by designing a new phase model, Synchronised Waveform matched Phase Model (SWPM), for the harmonic excitation, in order to
9.2 Concluding overview

reduce the phase discontinuities when switched between ACELP and harmonic excitation. The designed hybrid coder has three modes: scaled white noise coloured by LPC for unvoiced, ACELP for transitions, and harmonic excitation for stationary voiced segments. A reliable two stage speech classification algorithm was also designed. The first stage identifies the silence and unvoiced segments. The analysis by synthesis secondary classification detects the transitions from the voiced speech.

3. Quantising the model parameters at different bit rates, and evaluating the performance with the standard coders at comparable bit rates. The robustness of the coder was also tested against noisy background conditions and mode errors due to random channel errors.

9.2 Concluding overview

Chapter 2 presented a review of digital speech coding. The main speech coding paradigms, the design criteria, the applications of speech coding, and the existing standards were discussed. The present and future demands of speech coding have also been identified.

Chapter 3 described the fundamental speech coding techniques, common to many coding techniques. Modelling the speech spectral envelope using Linear Predictive Coding (LPC), and efficient quantisation and interpolation of the LPC coefficients by converting them to Line Spectral Frequencies (LSF) were discussed. The pitch determination algorithms based on the time domain and the frequency domain techniques were also introduced.

Chapter 4 introduced the fundamental concepts of harmonic coding of speech. The basic sinusoidal analysis and synthesis model, and the parameter estimation techniques were discussed. In order to adapt the basic sinusoidal analysis and synthesis model for low bit rates the component frequencies are restricted to harmonics of the fundamental frequency and the concept of frequency domain voicing is introduced. Three examples of low bit rate harmonic coders were also described: Sinusoidal Transform Coding (STC), Improved Multi Band Excitation (IMBE), and Split Band Linear Predictive
Coding (SB-LPC).

The fundamental concepts of Analysis by Synthesis (AbS) coding technique were presented in chapter 5. AbS schemes model the long term correlation present in the voiced speech using a Long Term Predictor (LTP). The innovation structures provide the startup excitation vectors at the onsets and filling in information for stationary voiced and unvoiced speech that the LTP can not model. Perceptual weighting enhances the speech quality by redistributing the energy of the quantisation noise spectrum, such that the quantisation noise is masked below the energy of the speech spectrum. Finally an example of an AbS scheme, Algebraic Code Excited Linear Prediction (ACELP) and the ITU G.729 standard, which is based on ACELP was presented in some detail.

The detailed design of a novel hybrid coding model was presented in chapter 6. The advantageous and challenges in designing a hybrid coder, and the limitations of the existing hybrid coders were described. The phase synchronisation when switched between the coding modes and designing a reliable speech classification algorithm are the main challenges in designing a hybrid coder, which combines harmonic coding and ACELP. The speech synthesised using the harmonic excitation is synchronised with the original speech using a new phase model, Synchronised Waveform matched Phase Model (SWPM). SWPM maintains time synchrony between the original and the synthesised speech by transmitting the Pitch Pulse Location (PPL) closest to each synthesis frame boundary. SWPM also preserves sufficient waveform similarity such that the switching between the coding modes is transparent, by transmitting a phase value, which represents the Pitch Pulse Shape (PPS) of the corresponding pitch pulse. The speech classification algorithm has two stages. The first stage identifies the silence and unvoiced segments, and the second stage detects the transitions from the remaining segments. The transition detection algorithm synthesises the harmonic excitation using SWPM and determines the suitability of the harmonic mode for each voiced frame. The frames for which the harmonic excitation is not suitable are declared as transitions. The stationary voiced frames are synthesised using the harmonic excitation, and the transitions are synthesised using ACELP. The silence and unvoiced frames are synthesised using scaled white noise coloured by LPC. It was found that synthesising unvoiced plosives using ACELP improves the speech quality. Consequently a special routine was
9.3 Future work

The research work carried out and presented in this thesis was primarily focused on demonstrating the viability of a novel hybrid coding algorithm in producing high quality speech at low bit rates. This section suggests the possible future research directions which may further improve the synthetic speech quality of the hybrid coder.

1. The spectral amplitude quantiser used for the harmonic excitation was originally designed to detect plosives from the unvoiced segments and synthesised using ACELP. Subjective listening tests suggested that the unquantised hybrid model achieves nearly transparent speech quality.

Chapter 7 presented the development of variable bit rate coders based on the hybrid coding model described in chapter 6. The white noise excitation and the harmonic excitation with SWPM are quantised at 1.5 kbps and 4 kbps respectively. Three versions of the hybrid coder were developed with transitions quantised at 4 kbps, 6 kbps, and 8 kbps using ACELP. The quality of the 4 kbps ACELP hybrid coder is subjectively superior to ITU G.723.1 at 5.3 kbps. The 6 kbps ACELP version achieves similar quality to G.723.1 at 6.3 kbps. The quality of the 8 kbps ACELP version is worse than G.729 at 8 kbps. However the average bit rate of the 8 kbps ACELP version is 4.53 and more than 50 % of the speech synthesised using the 8 kbps ACELP version was adjudged to be the same quality as G.729.

Chapter 8 presented the results of the tests conducted to evaluate the robustness of the hybrid coder under acoustic noise and mode errors due to random channel errors. The speech classification algorithm is capable of selecting the best mode under noisy background conditions. The noisy speech synthesised using the harmonic mode sounds metallic, which was reduced by introducing a voicing frequency marker. The robustness to mode bit errors was improved by transmitting the LSFs using the same bits for all the modes and limiting the dynamic range of the LTP gain quantiser. The hybrid coding algorithm is highly resilient against the mode errors and localises the errors within the frames which have the mode errors.
9.3. Future work

designed for 4 kbps SB-LPC and updates the amplitude vectors every 10 ms. This update rate seems unnecessary for the harmonic excitation of the hybrid coder, since only the stationary voiced segments are synthesised using the harmonic excitation. Therefore it would be possible to design a spectral amplitude quantiser with an update rate of 20 ms, which achieves the same quality at a lower overall bit rate. The extra bits may be used to quantise the frequency domain mixed voicing decisions. Mixed voicing decisions are useful for some offsets which gradually change from voiced to unvoiced. Alternatively the extra bits may be used to quantise perceptually important phase components [92], [115]. Savings in the bit allocation can be achieved by differentially quantising the harmonic phases with respect to PPS.

2. Optimising the coding modes for the intended speech characteristics has the potential to improve the speech quality and reducing the average bit rate. For example, stationary voiced segments may be quantised using longer variable length frames, maintaining the quality while reducing the average bit rate. However this will increase the delay, and suitable only for some applications such as voice storage. The ACELP innovation structure may be optimised to encode transitions [116]. The use of LTP for the onset frames should be more critically evaluated at different bit rates. The sparse excitation artefacts of the innovation pulses may be removed by using different pulse shapes [100], in addition to using LTP.

3. A $10^{th}$ order LPC filter is used for all the modes. However the same quality may be achieved by using a lower order filter for the unvoiced segments. Using LPC filters of different orders will introduce additional complications in the LSF interpolation and quantisation stages. The LPC parameters at the onsets may be extracted by using the first complete pitch cycle, by repeating it to obtain the required analysis length. Such an LPC analysis technique has the potential to provide a higher prediction gain at the transitions. Interpolation of the LSFs according to the energy of the extraction points, i.e. giving more weight to the LSFs extracted from the high energy segments and pitch cycle based interpolation techniques would also improve the perceived speech quality [117].

4. The hybrid coding algorithm described in chapter 6 has three modes, and 2 bits
are allocated to transmit the mode, see section 7.5. Therefore an additional mode may be added to further improve the speech quality or reduce the average bit rate. The potential candidates are VAD, plosives, and harmonic voicing. A VAD may be used to identify the silence frames and only the mode needs to be transmitted, since the decoder inserts comfort noise for the silence frames [79], [103]. Plosives may be synthesised using a pre stored template of the LPC residual, which reduces the number of bits required to transmit the excitation signal. The template waveform is gain scaled and filtered through the LPC synthesis filter, which produces perceptually transparent quality [118]. This scheme may also be incorporated by searching for the plosives at the decoder using the properties of the LPC and excitation parameters, such as the first reflection coefficient and the CELP innovation pulses. The harmonic mode was intended for stationary voiced speech and declares the complete spectrum as voiced. This degrades the perceptual quality of some mixed voiced segments, especially voiced speech corrupted with background noise, sounds metallic. The remaining mode may be used as an additional harmonic mode with a constant voicing frequency marker, e.g. 80 % of the spectrum is voiced.
Appendix A

List of publications


Appendix B

List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AaS</td>
<td>Analysis and Synthesis</td>
</tr>
<tr>
<td>AbS</td>
<td>Analysis by Synthesis</td>
</tr>
<tr>
<td>ACELP</td>
<td>Algebraic Code Excited Linear Prediction</td>
</tr>
<tr>
<td>ADPCM</td>
<td>Adaptive Differential Pulse Code Modulation</td>
</tr>
<tr>
<td>AM</td>
<td>Auto-correlation Method</td>
</tr>
<tr>
<td>AMDF</td>
<td>Average Magnitude Difference Function</td>
</tr>
<tr>
<td>AMR</td>
<td>Adaptive Multi-Rate</td>
</tr>
<tr>
<td>APC</td>
<td>Adaptive Predictive Coding</td>
</tr>
<tr>
<td>ATM</td>
<td>Asynchronous Transmission Mode</td>
</tr>
<tr>
<td>BER</td>
<td>Bit Error Rate</td>
</tr>
<tr>
<td>CCITT</td>
<td>International Telegraph and Telephone Consultative Committee</td>
</tr>
<tr>
<td>CDMA</td>
<td>Code Division Multiple Access</td>
</tr>
<tr>
<td>CELP</td>
<td>Code Excited Linear Prediction</td>
</tr>
<tr>
<td>CFD</td>
<td>Cumulative Frequency Distribution</td>
</tr>
<tr>
<td>CM</td>
<td>Covariance Method</td>
</tr>
<tr>
<td>DAM</td>
<td>Diagnostic Acceptability Measure</td>
</tr>
<tr>
<td>DFT</td>
<td>Discrete Fourier Transform</td>
</tr>
<tr>
<td>DRT</td>
<td>Dynamic Rhyme Test</td>
</tr>
<tr>
<td>DSP</td>
<td>Digital Signal Processing</td>
</tr>
<tr>
<td>DTMF</td>
<td>Dual Tone Multi-Frequency</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>DTX</td>
<td>Discontinuous Transmission</td>
</tr>
<tr>
<td>FEC</td>
<td>Forward Error Correction</td>
</tr>
<tr>
<td>FFT</td>
<td>Fast Fourier Transform</td>
</tr>
<tr>
<td>FIR</td>
<td>Finite Impulse Response</td>
</tr>
<tr>
<td>FR-GSM</td>
<td>Full Rate GSM</td>
</tr>
<tr>
<td>GPRS</td>
<td>General Packet Radio Service</td>
</tr>
<tr>
<td>GSM</td>
<td>Global System for Mobile Communications</td>
</tr>
<tr>
<td>HF</td>
<td>High Frequency</td>
</tr>
<tr>
<td>HPP</td>
<td>Harmonic Peak Picking</td>
</tr>
<tr>
<td>HR-GSM</td>
<td>Half Rate GSM</td>
</tr>
<tr>
<td>ICASSP</td>
<td>International Conference on Acoustics, Speech and Signal Processing</td>
</tr>
<tr>
<td>I-MBE</td>
<td>Improved Multi-Band Excitation</td>
</tr>
<tr>
<td>INMARSAT</td>
<td>International Maritime Satellite</td>
</tr>
<tr>
<td>IP</td>
<td>Internet Protocol</td>
</tr>
<tr>
<td>ISDN</td>
<td>Integrated Services Digital Network</td>
</tr>
<tr>
<td>ITU</td>
<td>International Telecommunication Union</td>
</tr>
<tr>
<td>LBG</td>
<td>Linde Buzo Gray</td>
</tr>
<tr>
<td>LD-CELP</td>
<td>Low Delay Code Excited Linear Prediction</td>
</tr>
<tr>
<td>LM</td>
<td>Lattice Method</td>
</tr>
<tr>
<td>LMS</td>
<td>Least Mean Square</td>
</tr>
<tr>
<td>LP</td>
<td>Linear Prediction</td>
</tr>
<tr>
<td>LPC</td>
<td>Linear Predictive Coding</td>
</tr>
<tr>
<td>LSF</td>
<td>Line Spectral Frequency</td>
</tr>
<tr>
<td>LSP</td>
<td>Line Spectral Pair</td>
</tr>
<tr>
<td>LTP</td>
<td>Long Term Prediction</td>
</tr>
<tr>
<td>MBE</td>
<td>Multi-Band Excitation</td>
</tr>
<tr>
<td>MELP</td>
<td>Mixed Excitation Linear Prediction</td>
</tr>
<tr>
<td>MOS</td>
<td>Mean Opinion Score</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Square Error</td>
</tr>
<tr>
<td>PCM</td>
<td>Pulse Code Modulation</td>
</tr>
<tr>
<td>PCN</td>
<td>Personal Communications Network</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>--------------------------------------</td>
</tr>
<tr>
<td>PDA</td>
<td>Pitch Detection Algorithm</td>
</tr>
<tr>
<td>PDF</td>
<td>Probability Density Function</td>
</tr>
<tr>
<td>PPL</td>
<td>Pitch Pulse Location</td>
</tr>
<tr>
<td>PPS</td>
<td>Pitch Pulse Shape</td>
</tr>
<tr>
<td>PSTN</td>
<td>Public Switched Telephone Network</td>
</tr>
<tr>
<td>QMF</td>
<td>Quadrature Mirror Filter</td>
</tr>
<tr>
<td>RELP</td>
<td>Residual Excited Linear Prediction</td>
</tr>
<tr>
<td>RMS</td>
<td>Root Mean Square</td>
</tr>
<tr>
<td>RPELPC</td>
<td>Regular Pulse Excited Linear Predictive Coding</td>
</tr>
<tr>
<td>SAPDA</td>
<td>Segmented Auto-correlation Pitch Detection Algorithm</td>
</tr>
<tr>
<td>SB</td>
<td>Split Band</td>
</tr>
<tr>
<td>SB-ADPCM</td>
<td>Sub-Band Adaptive Differential Pulse Code Modulation</td>
</tr>
<tr>
<td>SID</td>
<td>Silence Description</td>
</tr>
<tr>
<td>SMM-PDA</td>
<td>Sinusoidal Model Matching Pitch Detection Algorithm</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal to Noise Ratio</td>
</tr>
<tr>
<td>SSM</td>
<td>Synthetic Spectral Matching</td>
</tr>
<tr>
<td>SWPM</td>
<td>Synchronised Waveform matched Phase Model</td>
</tr>
<tr>
<td>TE</td>
<td>Time Envelope</td>
</tr>
<tr>
<td>TIMIT</td>
<td>Texas Instruments/Massachusetts Institute of Technology</td>
</tr>
<tr>
<td>UMTS</td>
<td>Universal Mobile Telecommunication System</td>
</tr>
<tr>
<td>VAD</td>
<td>Voice Activity Detector</td>
</tr>
<tr>
<td>VBRC</td>
<td>Variable Bit Rate Coding</td>
</tr>
<tr>
<td>VFLS</td>
<td>Variable Frame Length Coding</td>
</tr>
<tr>
<td>VLSI</td>
<td>Very Large Scale Integration</td>
</tr>
<tr>
<td>VQ</td>
<td>Vector Quantisation</td>
</tr>
<tr>
<td>VSELP</td>
<td>Vector Sum Excitation Linear Prediction</td>
</tr>
<tr>
<td>V/UV</td>
<td>Voiced/Unvoiced</td>
</tr>
<tr>
<td>WPP</td>
<td>Waveform Peak Picking</td>
</tr>
</tbody>
</table>
Bibliography


[92] Doh-Suk Kim, “Perceptual phase redundancy in speech,” in Proc. Int. Conf. on

[93] TIA/EIA, Enhanced variable rate codec, speech service option 3 for wideband

 cellular applications,” in Proc. Int. Conf. on Acoust., Speech, Signal Processing,


[96] P. Satyanarayana Murthy and B. Yegnanarayana, “Robustness of group delay
based method for extraction of significant instants of excitation from speech sig­
1999.

[97] B. S. Atal and M. R. Schroeder, “Recent advances in predictive coding-
applications to voiced speech synthesis,” in Speech Commun. Seminar, Stock­
holm, 1974.

vocoder with robust parameter determination,” in Proc. Int. Conf. on Acoust.,

[99] M. Stefanovic, Y. D. Cho, S. Villette, and A. M. Kondoz, “A 2.4/1.2 kb/s speech
coder with noise pre-processor,” in Proc. European Signal Processing Conference
(EUSIPCO), Tampere, Finland, Sep. 2000.


