1998

Waveform interpolation speech coding

Jun Ni
University of Wollongong

WAVEFORM INTERPOLATION SPEECH CODING

A thesis submitted in fulfilment of the requirements
for the award of the degree of

Honours Master of Engineering

from

The University of Wollongong

by

Jun Ni
M.S., Academia Sinica, China, 1995
B.S., Nanjing University, China, 1992

School of Electrical, Computer & Telecommunications Engineering
The University of Wollongong
April, 1998

Dedicated to my grandmother

Acknowledgments

The research work reported in this thesis was carried out at the School of Electrical, Computer and Telecommunications Engineering at the University of Wollongong, Australia, under the supervision of Dr. Ian S. Burnett and Professor Joe F. Chicharo. I would like to acknowledge the help of many individuals.

First of all, I would like to express my special thanks to my supervisor Dr. Ian S. Burnett for his valuable guidance and support throughout this research. His advice, encouragement, and drive, was a great factor in the completion of this thesis. Professor Joe F. Chicharo also receives special mention for his academic supervision.

I wish to acknowledge and thank the Motorola Research Center, Australia for providing financial assistance and their staff, especially, Dr. Mark M. Thomson for provision of support for speech DATA, MOS tests and result analysis.

I very much appreciate the assistance and friendship of the people in the School of Electrical, Computer & Telecommunications Engineering at the University of Wollongong, particularly Maree Fryer, Peter Costigan, Philip Ogunbona, Zheng Li, David Atkinson, Li Xue, Jack Parry, Nicky Chong, Matthew Miller and many others.

Finally, I am deeply grateful to my parents who live in China for their sacrifices that have enabled me to pursue higher education in Australia.

Abstract

This thesis deals with waveform interpolation speech coding. Speech coding in the last decade has been dominated by the CELP paradigm. CELP algorithms offer high-quality speech compression at bit rates from 4 to 16 kb/s. Recent research efforts have been oriented to a new generation of speech coding algorithms operating at bit rates of 2.4kb/s and below. CELP and its derivative architectures appear to be inadequate to meet the increasing quality objective. This is due to the small bit budget to adequately represent the original signal. A major source of distortion in CELP is an inaccurate degree of periodicity of the speech signal. The Waveform interpolation (WI) algorithm is intended to preserve natural periodicity by representing speech as an evolving set of pitch cycle waveforms (known as the prototype waveform or Characteristic Waveform). The waveform interpolation (WI) paradigm was found to provide state-of-the-art performance at 2.4kb/s.

Research on WI coding has been focused on quality improvement, complexity reduction and channel error robustness. The key to quality improvement is the efficient decomposition and quantization of the LP residual of the speech signal. New techniques, including an analysis-by-synthesis technique, and SEW and REW quantization techniques are presented in this thesis. WI coders provide good compression quality but suffer from high complexity, compared with other low bit-rate speech coders. A low-complexity algorithm is proposed. The waveform

interpolation architecture is particularly convenient for operating at different bit rates. The performance of WI coders with rates between 2.4kb/s and 3.6kb/s is examined.

CONTENTS

List of Symbols

$A(k)$	*DFT coefficients of the impulse response of LP filter*
$A(z)$	*Linear prediction (LP) filter*
a_n	*nth prediction coefficient*
$\alpha_k(t),\ \beta_k(t)$	*the K time-varying Fourier series coefficients of the Characteristic Waveform*
$\phi(t)$	*phase of the extracted characteristic waveform (prototype)*
$G(t_i)$	*gain of the prototype extracted at time interval* t_i
H_p	*adaptive postfilter*
H_t	*tilt compensation filter*
$h(m)$	*the impulse response of the FIR low-pass filter used in SEW/REW decomposition*
$LSF[k]$	*Kth LSF parameter*
$op(t)$	*reconstructed speech*
$P_{n+1}(z)$	*LP difference filter*
$p(t)$	*pitch value at time t*
$Q_{n+1}(z)$	*LP sum filter*
REW	*rapidly evolving waveform*
$R(d)$	*normalized correlation function*

$r(t)$	*speech signal in LP residual domain*
SEW	*slowly evolving waveform*
$s(t)$	*input speech signal*
$T_0(x), T_1(x),\ldots, T_4(x)$	*the first five shifted Chebyshev polynomials*
$\hat{U}(t_m, \phi)$	*aligned prototype at time interval* t_m
$U'(t_i, k)$	*the speech domain prototype described by DFT coefficients*
$u(t, \phi)$	*two-dimension Characteristic Waveform surface*
$u_{SEW}(t, \phi)$	*two-dimension SEW surface*
$u_{REW}(t, \phi)$	*two-dimension REW surface*
$\hat{v}(t_i, t)$	*unaligned extracted prototype at time interval* t_i
$W(z)$	*perceptual weighting filter*

CHAPTER 1

INTRODUCTION

1.1 Introduction

Speech coding is the field concerned with compression and decompression of the digital information necessary to represent a speech signal. Digital speech brings flexibility for encryption, but is also associated with a high data rate. The objective of speech coding is to represent speech with a minimum bit rate while maintaining its perceptual quality. Speech coders compress signals by exploiting the natural redundancies in speech and the properties of human hearing. Most compression techniques used in speech coding are known as lossy compression, where the reproduced speech is not identical to the original. The signal, however, sounds like the original because of masking properties of the human ear that render a level of certain types of noise inaudible.

Speech coders are used to transmit and store speech for various applications. Examples of transmission applications include wireless cellular, satellite communications, Internet phone, audio and video conference, and secure voice systems. In particular, wireless cellular and satellite communications have been enjoying a tremendous worldwide growth. Storage applications include digital telephone answering machines, voice-mail, Text-To-Speech (TTS) systems. In most of these applications, speech coding is based on telephone bandwidth speech, limited to about 3.2 KHz (200Hz to 3.4KHz). In this thesis, speech is bandlimited to 4 KHz and sampled at 8 KHz [46].

The past decade has witnessed substantial progress in speech coding. Central to this progress has been the development of new speech coders capable of producing high quality speech at low bit rates. These coders exploit models of speech production and auditory perception, and offer a quality that significantly exceeds prior compression techniques. A number of speech coders have already been adopted in regional and international telephone standards [14], [20].

The research in this thesis is concerned with waveform interpolation (WI) speech coding. The Waveform Interpolation (WI) coding paradigm was found to provide state-of-the-art performance at bit rates below 4kb/s [32], [33], [34]. The coder performs very well in terms of perceptual quality and robustness against channel errors and background noise.

The remainder of this introductory chapter is organized as follows: Section 1.2 describes the attributes used to evaluate speech coders. Section 1.3 presents the advances in speech coding. A brief introduction to Waveform Interpolation is given in Section 1.4. Section 1.5 discusses the approach of this thesis. Finally, Section 1.6 presents a brief summary of the contributions.

1.2 Evaluation of Speech Coders

The performance of speech coding algorithms is measured on the basis of five attributes - bit rate, the quality of reproduced (coder) speech, the complexity of the

algorithm, the delay introduced by the coder, and the robustness of the algorithm to channel errors and background noise. In general, high quality speech at low rates is achieved using high complexity algorithms with high delay. Speech coders must, thus, balance speech quality, complexity, delay and robustness [14], [52].

1.2.1 Bit Rate

Bit rate reflects the degree of compression that the coding algorithm achieves. Telephone bandwidth speech is sampled at 8 KHz, and quantized with an 8-bit logarithmic quantizer, making the bit rate of the original speech 64 kbits/s [14]. The degree of compression is than measured by how much the bit rate is lowered from 64 kbits/s. Usually, the term *medium rate* is used for coders working in the range of 8 ~ 16 kbits/s, *low rate* for coders working in the range of 2.4kbits/s ~ 8 kbits/s, and *very low rate* for coders operating below 2.4kbits/s. International standards exist for coders operating at 40, 32, 24 and 16 kbits/s. Cellular standards cover the range from 13 to 3.45 kbits/s. Secure voice coders operate at 4.8, 2.4 and 0.8 kbits/s [14].

1.2.2 Quality

Quality is an important attribute. In digital communication, speech quality is generally classified into four categories: *broadcast*, *network* or *toll*, *communication*, and *synthetic*. *Broadcast* wideband (typically 7 KHz) speech refers to high-quality "commentary" speech. *Network* or *toll* quality refers to quality comparable to the original telephone bandwidth speech. *Communication* speech refers to some-what

degraded speech which is, nevertheless, natural, highly intelligible, and adequate for telecommunication. *Synthetic* speech is usually intelligible but can be unnatural and associated with some distortion. Currently, broadcast speech can be achieved at rates above 64 kbits/s, toll quality can be achieved at medium rate, communication quality at low rate, and synthetic quality at very low rate [52].

Judging the quality of coded speech is an important but also very difficult task. Common objective measures, such as the signal-to-noise ratio (SNR) and the segmental SNR (SEGSNR), are often sensitive to gain variations and delays. They can not account for the perceptual properties of human hearing. Therefore, subjective measures are adopted. Subjective measure procedures such as the Diagnostic Rhyme Test (DRT), the Diagnostic Acceptability Measure (DAM), the Mean Opinion Score (MOS) and the Degradation Mean Opinion Score (DMOS) are based on listener ratings. The Diagnostic Rhyme Test (DRT) is used to measure intelligibility. The Diagnostic Acceptability Measure (DAM), the Mean Opinion Score (MOS) and the Degradation Mean Opinion Score (DMOS) are used to measure quality [14], [52].

The MOS test is widely used to evaluate coded speech quality. The MOS usually involves 50 to 60 listeners who are instructed to rate speech according to a five level quality scale. A MOS of 5 implies *excellent* quality, a MOS of 4 implies *good* quality, a MOS of 3 implies *fair* and 2 implies *poor* [52].

1.2.3 Complexity

Complexity is another essential issue. In general, high-quality speech coding at low rates requires high-complexity algorithms. Complexity affects the implementation of speech coders.

Complexity typically has three components [14]:

- The number of instructions executed per second, which is generally measured in MIPS (millions of instruction per second). Generally, a higher speed DSPU costs more and consumes more power.
- The memory requirement in terms of RAM (random access memory). RAM is used to store the variables used in the coding algorithm.
- The memory requirement in terms of ROM (read only memory). ROM is needed to store the instructions, constant values and codebooks used in the coding algorithm.

1.2.4 Delay

Delay introduced by the coder will be objectionable to communication users, and may require the expensive use of echo cancellers. It is strongly recommended that the delay be no greater than 300ms [14]. However, in voice storage applications, delay is not so important. A delay of one second would be unnoticeable in the latter application.

1.2.5 Robustness

Robustness is the ability of a speech coder to preserve the perceptually important information against channel errors. In some situations, the coder must perform well when speech is corrupted by background noise, including narrow band noise (such as DTMF, modem signal, etc) and wide band noise (such as office noise, machine noise, etc). A robust speech coder should also perform well with a variety of languages and accents [52].

The foregoing description of the five attributes - bit rate, quality, complexity, delay, and robustness, indicates that there are many tradeoffs in setting the requirements of a speech coder for a particular application. For example, digital cellular systems transmit speech over radio channels, where channel interference and fading can cause significant random errors in the bit stream. It is thus essential to transmit the bit stream with error protection. As the percentage of channel capacity used for error protection increases, the number of bits available to the speech coder decreases, resulting in lower quality. A tradeoff thus exists between channel robustness and the speech quality.

1.3 Advances in Speech Coding

Speech coding research started over fifty years ago, and early coding implementations were vocoders based on analog speech representations (rather than the current digital

methods). With progress in VLSI technologies and DSP theory, speech coding has, however, advanced rapidly. Driven by the need for telephone bandwidth and secure transmission in cellular and military communications, research efforts during the 1980's and 1990's have focused upon developing low-rate speech coders. Most of these coders incorporate mechanisms to: represent the spectral properties of speech, provide for speech waveform matching, and optimize the speech quality for the human ear. In particular, Atal and Schroeder [1][2][3] proposed a linear prediction algorithm with stochastic vector excitation called Code Excited Linear Prediction (CELP). CELP is capable of producing medium to low rate speech adequate for communication applications.

1.3.1 Waveform Coders and Vocoders

Speech coding algorithms can be divided into two main categories, waveform coders and vocoders. Waveform coders focus upon representing the speech waveform, approximating the original waveform without necessarily exploiting the underlying speech model. In contrast, vocoders do not reproduce an approximation to the original speech. Instead, parameters that characterize individual speech segments are specified and transmitted to the decoder, which then reconstructs a new and different waveform that will have a similar sound. Vocoders thus rely on speech models. Waveform coders are generally more robust than vocoders because they work well with a wider class of signals including audio signals. However, they also operate at higher bit rates than vocoders.

Code Excited Linear Prediction (CELP) [3] belongs to the class of waveform coders. Other methods in commercial use today include Adaptive Delta Modulation (ADM), Adaptive Differential Pulse Code Modulation (ADPCM), Multipulse Linear Predictive Coding (MP-LPC) [4], [51], and Regular Pulse Excitation (RPE) [37]. A standard that uses a 13kbit/s regular pulse excitation algorithm has been deployed by the "Group Speciale Mobile" (GSM) in Europe, Australia and many other areas of the world.

The most important vocoder historically is the Linear Predictive Coding (LPC) vocoder. It is used extensively in secure voice telephony (FS1015) and is the starting point for some current speech coders. Sinusoidal coding is another vocoder that has emerged in the past decade. Sinusoidal Transform Coding (STC) [40], [41] and Multiband Excitation (MBE) coding [23] are examples of sinusoidal coding. A 6.4 kbit/s Improved Multiband Excitation (IMBE) coder has been adopted for the International Maritime Satellite (INMARSAT-M) system and the Australian Satellite (AUSSAT) system [25].

1.3.2 Existing Speech Coding Standards

Progress in speech coding, enabled recent adoptions of low-rate algorithms for mobile telephone and secure military communications. International standards exist for coders operating at 64, 32, and 16kb/s. Regional cellular standards range from 13 to 3.45kb/s. Secure voice coders operate at 4.8 and 2.4kb/s. These standards indicate the

performance of current speech coders. Some of these standards are listed as follows [14], [20].

CCITT G.711 standard is a Pulse-Code Modulation (PCM) coder at 64kb/s. Speech is sampled at 8 KHz, and its amplitude is quantized with an 8-bit logarithmic scalar quantizer. North America uses u-law PCM, and other countries use A-law PCM. G.711 is generally considered as noncompressed and is often used as a reference for comparison [14].

CCITT G.721 standard operates at 32kb/s. G.721 uses Adaptive Differential Pulse-Code Modulation (ADPCM) techniques, which exploit the signal correlation [14].

Low Delay Code Excited Linear Prediction (LD-CELP) is used for ITU-T Recommendation G.728 [9], [10]. LD-CELP is a Code Excited Linear Prediction (CELP) coder using backward adaptive prediction to reduce delay.

IS-54 (Interim Standard 54) was created as the standard for the U.S. cellular system. A kind of CELP coder, Vector Sum Excited Linear Prediction (VSELP) is adopted [21].

FS1016 - U.S. Federal Standard 1016 is a 4.8kb/s CELP coder for secure voice system applications [17].

FS1015 - U.S. Federal Standard 1015 is a 2.4kb/s LPC speech vocoder used in secure voice systems [14].

FS1017 - U.S. Federal Standard 1017 is a Mixed excitation LPC vocoder (MELP) [42], [43] that provides close quality to the FS1016 while operating at half of the bit rate of the FS1016 coder (2.4kb/s).

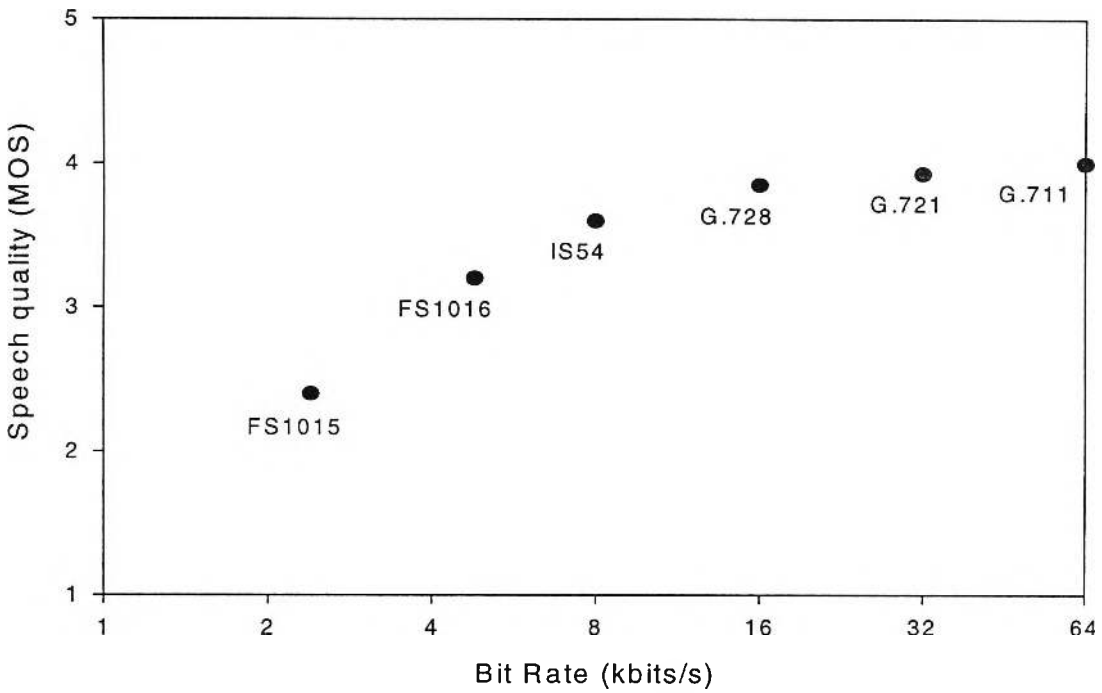


Figure 1.1: Speech quality achieved by coding standards at different bit rates [14].

Figure1.1 illustrates the performance of these coders. It is found that speech coders, such as CELP coding offer good quality for rates in the range of 4 to 16kb/s. The current goal in speech coding is to achieve toll or communications quality below 4kb/s.

1.4 Introduction to Waveform Interpolation Coding

CELP is, perhaps, the most successful speech coder of the past decade. However, speech quality obtained by CELP coding is found to degrade rapidly below 4kb/s. This is because of the sparsity of bits (less than 0.5b for one sample of speech) which makes it impossible to accurately represent the speech waveform. Recently, several new algorithms have emerged in competition with CELP at 4kb/s and below. One promising approach is Waveform Interpolation (WI) coding.

The Waveform Interpolation coding algorithm was proposed by Kleijn in 1991 [29]. In Waveform Interpolation coders, the input speech is represented by a sequence of pitch-cycle waveforms - Characteristic Waveforms(CW). The coded speech is reconstructed by interpolation of the Characteristic Waveforms. Originally, WI was applied to voiced speech only, but in the later work, the algorithm was extended to both voiced and unvoiced speech by decomposition of the Characteristic Waveform [32]. The CWs are decomposed into a slowly evolving waveform (SEW), which represents the voiced component of the speech, and a rapidly evolving waveform (REW), which represents the unvoiced component of the speech. These two waveforms are quantized separately according to their perceptual properties.

The Waveform Interpolation algorithm efficiently exploits the evolutionary nature of speech signal and human perception property. The reproduced speech achieves high perceptual quality even at very low bit rates. Waveform Interpolation coding generally works at 2.4kb/s, but recently, WI coders operating from 1.2kb/s to 4kb/s have been

reported [5], [6], [50]. Recent research is also concentrated in reducing the complexity of WI coders [34], [50].

1.5 Approach of This Thesis

This thesis deals with Waveform Interpolation (WI) speech coding. The primary objective is to develop a Waveform Interpolation coder and improve the implementation of coder. A baseline 2.4kb/s WI coder is developed first. The main procedures in WI coding, signal decomposition, quantization and reconstruction are investigated. Several new techniques are proposed and tested. A series of WI class coders working at different bit rates, and a WI coder with low level of complexity are also developed.

1.5.1 Pitch Detection of WI

In a Waveform Interpolation coder, it is very important that the pitch track is sufficiently accurate. Wrong pitch values may introduce clicks, clunks and other distortion in the reproduced speech. An improved pitch calculation mechanism is thus introduced. The pitch value is determined by a composite correlation function. Possible pitch doubles and multiples are judged by setting a threshold.

1.5.2 Spectral Decomposition in WI

Effective representation of the SEW and REW is the key to coder performance. At low rates, the phase spectrum of the SEW and REW is removed. Only the magnitude information is transmitted. The SEW and REW magnitude are quantized using different VDVQ (Variable Dimension Vector Quantization) algorithms. The REW magnitude is quantized using Chebyshev polynomials. The low frequency part of the SEW magnitude spectrum is represented by eight bins, the high frequency part of the SEW is derived from the REW.

Analysis-by-Synthesis (A-by-S) mechanisms have found favour in the low bit rate speech coders. However, Waveform Interpolation coders depend on open-loop quantization and do not utilise A-by-S techniques. A closed-loop technique for quantization is proposed in this thesis, which incorporates A-by-S mechanisms. The results indicate a better perceptual performance than open-loop schemes.

1.5.3 Scalability of WI

The Waveform Interpolation structure also provides a feasibility to work at different bit rates. The output speech of the WI coder is generated by interpolating the speech prototypes being transmitted. By increasing/decreasing the update rate and/or the codebook size of the prototype parameters, the bit rates of WI coders can be changed. Therefore, the WI coder can work at different bit rates with no or little change in the coder structure. The performance of WI coders working at bit rates above 2.4kb/s is

examined in this thesis. Informal listening tests show successive improvement in speech quality.

1.5.4 WI Complexity

Waveform Interpolation coders provide good-quality speech at low bit rates. However, the coder has a very high level of computational complexity. The high complexity is mainly introduced by the accurate SEW/REW decomposition procedure, including the DFT operation, time alignment and the SEW/REW filtering. At low bit rates, the bits allocated for the SEW and REW is very small. There is no need to generate a high resolution SEW and REW surface. Therefore, simplified SEW/REW decomposition and quantization mechanisms are adopted. The highly complex operations, such as time alignment and filtering are not required. At 2.4kb/s, the quality of the coded speech is similar to the high-complexity version.

1.6 List of Contributions

- A 2.4kb/s WI coder is presented as the baseline coder for future research and development (Chapter 2).

- The main coding operations in the baseline WI coder are investigated. Some new techniques are introduced to the improves the performance of the baseline coder (Chapter 3).

- An improved pitch calculation algorithm is proposed. The reliability of the pitch track is increased even when the pitch period is changing rapidly. The algorithm can also detect pitch doubles and multiples (Chapter 3).

- SEW and REW Quantization Mechanisms are presented. Only the magnitude of the SEW and REW are transmitted. The SEW magnitude is represented by eight bins and the REW magnitude is represented by polynomials (Chapter 3).

- Analysis-by-Synthesis techniques are incorporated in Waveform Interpolation coding architectures. The perceptual performance of the coder is improved, compared with the standard WI coder (Chapter 4).

- Waveform Interpolation coders working at bit rates above 2.4kb/s are presented. The perceptual quality of coded speech can be substantially improved by increasing the bit rate of the WI coder from 2.4kb/s to 3.6kb/s (Chapter 5).

- A low complexity Waveform Interpolation algorithm is proposed. The computational load can be dramatically reduced while the speech quality is maintained (Chapter 5).

CHAPTER 2

Review of Waveform Interpolation Speech Coding

2.1 Introduction

This chapter presents the detail of the Waveform Interpolation (WI) algorithm. The WI coder describes the speech as an evolving sequence of pitch cycle waveforms (Waveform Interpolation) and decomposes the Characteristic Waveforms into a voiced component (SEW) and an unvoiced component (REW). It also utilises some techniques which are used in other speech coders, such as LP analysis and LSF quantization. Further, almost all its parameters are interpolated, resulting in a smooth reconstruction quality. A 2.4kb/s WI coder is introduced as a baseline coder for future research.

Code-Excited Linear Prediction (CELP) coding is a popular speech coding algorithm. The key feature of CELP coding is the use of analysis-by-synthesis (A-by-S) techniques. The CELP algorithm and the A-by-S technique is also described in this Chapter.

This Chapter is organized as follows. In Section 2.2, a survey of the WI coding algorithm is given. Section 2.3 presents the waveform interpolation (WI) algorithm. A brief overview of the CELP algorithm is given in Section 2.4. Finally, Section 2.5 concludes this chapter.

2.2 Survey of the WI algorithm

2.2.1 Prototype Waveform Interpolation Coding

The Prototype-Waveform Interpolation (PWI) [31] coding algorithm is the first-generation WI coder which was designed to code voiced speech at bit rates below 4kb/s. Speech coders that work on a frame-by-frame basis, such as the CELP algorithm, provide good speech quality at bit rates above 4.8kb/s. However, when the bit rate is reduced, the quality of speech generated by CELP based methods degrades rapidly. In particular, for voiced speech, the correct degree of periodicity is no long properly preserved. In contrast, the PWI coding algorithm provides perceptually good speech quality at bit rates below 4kb/s [20], [31].

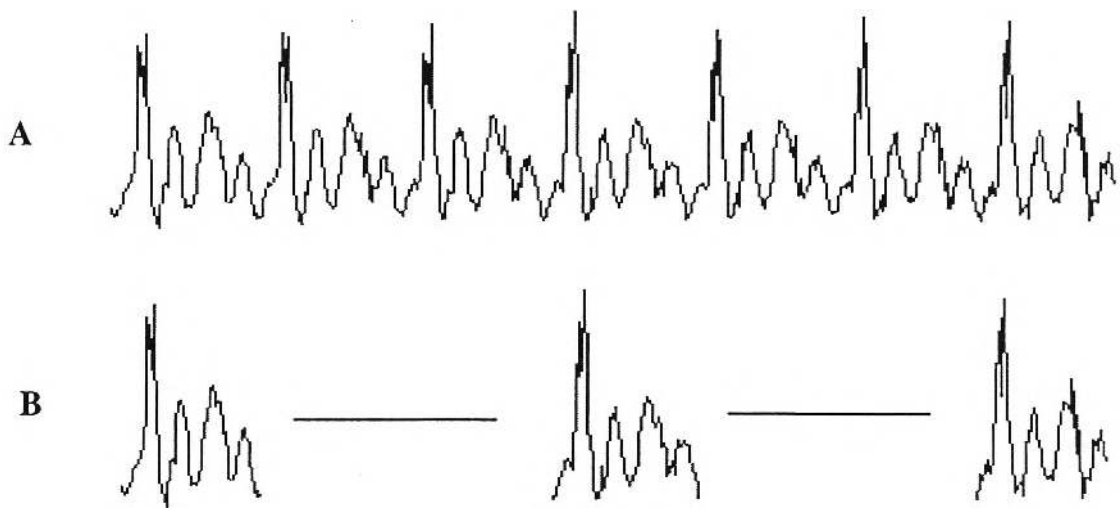


Figure 2.1: An example of one frame of voiced speech (A), showing that voiced speech can be represented by evolving pitch length prototype waveforms (B).

In the PWI coding method, voiced speech is interpreted as a concatenation of evolving pitch-length prototype waveforms. Therefore, voiced speech can be reconstructed by interpolation from a sequence of prototype waveforms with an update rate of one waveform per 20~30 ms interval (see Figure 2.1). Thus, the proper level of periodicity of the voiced speech signal is preserved.

Although voiced speech signals usually evolve slowly during regular intervals of 20-30 ms, there are cases where the waveforms have significant dynamics, such as speech with high levels of aspirations. The pitch-cycle waveforms will not evolve smoothly, especially at frequencies beyond 1500 Hz [31]. Directly using PWI and ignoring the dynamics of the waveform will cause distortion (tonal artifacts) and make reconstructed speech unnatural. Keeping the waveform dynamics suggests the preservation of the signal change ratio (SCR) of the waveform [31]. SCR is defined as a measure of the similarity of waveforms. A long-term SCR (LTSCR) is defined as the SCR between the adjacent transmitted prototype waveforms. By adjusting the LTSCR, the periodicity of the reconstructed speech can be constrained to match the original speech. A short-term SCR (STSCR) is defined as the SCR between the adjacent interpolation waveforms. The dynamics of speech are preserved by replacing an appropriate fraction of the waveforms by noise, according to the STSCR value.

The waveform dynamics of the voiced speech can be preserved by transmitting speech with LTSCR and STSCR adjustments, so that the distortion in the reconstructed speech will be greatly reduced. The complete coder combines WI with CELP coding for unvoiced speech segments [5] [31].

Transmitting prototype waveforms with sufficient information about waveform dynamics requires relatively high bit rates - between 3.0-4.0 kb/s. As PWI is only used for coding voiced speech, and CELP or other speech coding is needed for unvoiced segments, an accurate voiced/unvoiced division is required.

2.2.2 Multiple Prototype Waveform Coding

Recently, a new type of Waveform Interpolation, Multiple Prototype Waveform (MPW) coding was suggested for representing waveforms at low bit rates with the waveform dynamics preserved [6], [32], [33] ,[34]. Multiple Prototype Waveform coding can also describe the unvoiced speech, making the voiced/unvoiced speech division unnecessary.

Prototype-Waveform Interpolation has a low update rate of prototype waveforms, resulting in a high level of periodicity. This makes the algorithm only applicable to voiced speech. An increase in update rate allows a higher evolution bandwidth for prototypes, accommodating both voiced speech, which has a high periodicity, and unvoiced speech, which is less periodic. However, increasing the update rate is, necessarily, associated with an increase in the bit rate if new decomposition mechanisms are not employed.

In multi-prototype waveform (MPW) coding, first a one-dimensional speech signal is transformed to a two-dimensional Characteristic Waveform (CW). Then the

Characteristic Waveform is decomposed into two components, rapidly evolving waveform (REW) and slowly evolving waveform (SEW). The REW and SEW are quantized differently according to perception theory. Because of its low evolution bandwidth, the update rate of SEW can be very low, similar to the update rate of prototype waveforms in a PWI coder. The REW, which has a high evolution bandwidth, is sampled at a high rate, but the quantization accuracy required for REW is low. Thus, Multi-Prototype Waveform (MPW) coding operates at a high update rate, allowing the coding of both voiced and unvoiced speech as well as background noise, with a low bit rate being maintained. The WI coders presented in this thesis all belong to the MPW class of coders. The next section introduces a baseline 2.4kb/s WI coder.

2.3 Waveform Interpolation Algorithm

Using Characteristic Waveforms to describe speech and the subsequent decomposition of the Characteristic Waveforms are key features of WI coding. They are new techniques which are not seen in previous speech coders. This section first gives the definition of Characteristic Waveforms and their decomposition. Then, a WI coding algorithm working at 2.4kb/s is presented. The techniques used in the WI coding, such as LP analysis and quantization, pitch detection and gain quantization are also described.

2.3.1 Waveform Interpolation Principles

2.3.1.1 Characteristic Waveform

Definition of Characteristic Waveforms

In Waveform Interpolation coding, the speech signal is represented by a series of evolving Characteristic Waveforms. Voiced speech is effectively a concatenation of slowly evolving pitch cycle waveforms, and if the pitch cycle waveform and its phase function are always available, then there will be no distortion in the reconstructed speech. Therefore, the one dimensional speech signal $s(t)$ can be represented as a two dimensional signal, $u(t,\phi)$, with the pitch cycle waveform displayed along the phase ϕ axis. While this is natural for voiced speech, it can also be made valid for unvoiced speech. For this reason, the waveform displayed along the ϕ axis will be referred to as a Characteristic Waveform (CW). Aligning the Characteristic Waveform along the time t axis results in a description of the evolution of this waveform (and its sample values), resulting in the two dimensional surface $u(t,\phi)$ [33].

It is convenient to interpret the Characteristic Waveform as being derived from periodic speech. For voiced speech, the period of the speech is pitch period p, while for unvoiced speech, the period of the speech is an arbitrary value.

$u(t,\phi)$ is then a periodic function with a period of 2π along the ϕ axis. For speech with a fixed pitch period, ϕ can obtained by: $\phi(t) = 2\ \pi\ t\ /\ p$. For a time-varying pitch period, the phase is:

$$\phi(t) = \phi(t_0) + \int_{t_0}^{t} \frac{2\pi}{p(t)} dt \qquad \ldots\ldots \qquad (2.1)$$

Then, the one-dimensional speech signal $s(t)$ can be specified by the two dimensional surface $u(t,\phi(t))$:

$$s(t) = u(t,\phi(t)) \qquad \ldots\ldots \qquad (2.2)$$

such that $s(t)$ is a particular trajectory in the t, ϕ plane.

As the CW surface $u(t,\phi(t))$ is obtained from the one-dimensional speech signal $s(t)$ by continuously sampling along a trajectory $(t,\phi(t))$, this method for defining $u(t,\phi(t))$ is called the continuous sampling method.

Discrete CW Surface

In practice, the continuous sampling procedure presented above is too complex for implementation. Instead, the discrete sampling method is used. A discrete CW surface $u(t_i,\phi(t))$ is obtained by sampling at fixed intervals t_i on the time axis. The CW surface $u(t,\phi(t))$ can be reconstructed (approximately or perfectly dependent on sampling rate) by continuous interpolation of the discrete surface $u(t_i,\phi(t))$ [33]. Figure 2.2 shows an example of the discrete CW surface.

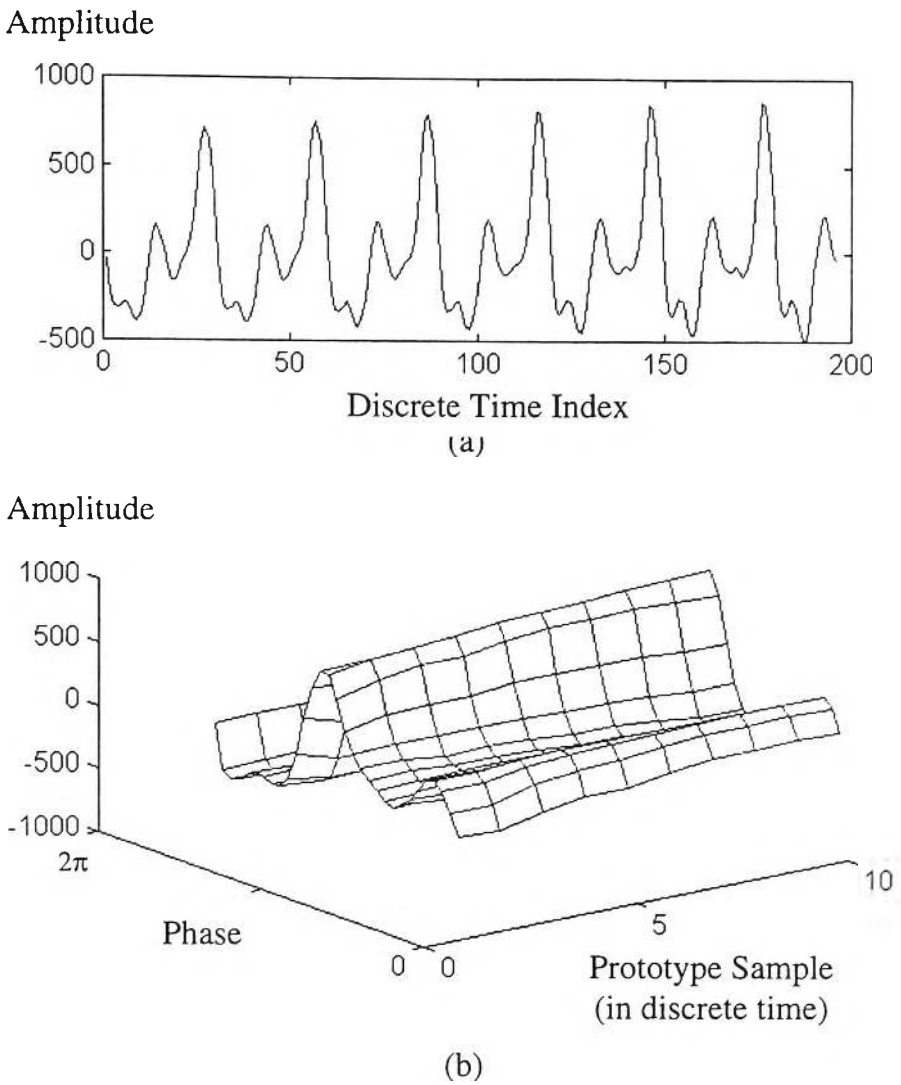


Figure 2.2: (a) One-dimensional speech signal (sampling rate is 8000Hz); (b) Two-dimensional discrete CW surface sampled at 400Hz.

The Fourier-series Description

The Characteristic Waveform can be described in the time domain or in the frequency domain. The Fourier-series description is particularly convenient as it provides flexibility of access to various frequency bands [33]. In this thesis, the Characteristic Waveform $u(t_i, \phi(t_i))$ is represented by a Fourier series,

$$u(t_i,\phi) = \sum_{k=1}^{K} \alpha_k(t_i)\cos(k\phi) + \beta_k(t_i)\sin(k\phi) \qquad \ldots\ldots \qquad (2.3)$$

where $\alpha_k(t_i)$ and $\beta_k(t_i)$ are the K time-varying Fourier series coefficients. In implementation, these coefficients are found by a DFT. The number of harmonics, K, is determined by the pitch of the Characteristic Waveform surface at the point t_i [6].

2.3.1.2 Decomposition of the Characteristic Waveform

Accurate transmission of the CW surface requires a high update rate, particularly for unvoiced sounds. The sampling rate of the CW surface should, in principle, be at least once per pitch period. Table 2.1 shows the MOS for different update rates achieved by Kleijn [33]. However, for the perceptually accurate transmission of the CW surface, only perceptually important information is needed.

CW Sampling Rate (Hz)	50	100	200	400
Mean Opinion Score	2.3	2.8	3.6	4.0

Table 2.1: MOS as a function of the CW sampling rate

It has been found recently that, the perception of voiced speech and unvoiced speech differs greatly [33]. Firstly, for unvoiced speech, only the magnitude spectrum and power contour is important. In contrast, for voiced speech, the phase of voiced speech is important for perception. Furthermore, the magnitude spectrum for voiced speech requires a more precise description than for unvoiced speech. Secondly, for voiced

speech (which is quasi-periodic), the Characteristic Waveform evolves slowly, while for unvoiced speech (which is nonperiodic), the CW evolves rapidly [32], [33].

This suggests a decomposition of the CW into voiced and unvoiced components, which have different quantization requirements. The voiced component of the CW is designated as a slowly-evolving waveform (SEW), and the unvoiced component of the CW is designated as a rapidly-evolving waveform (REW). These two components sum to the entire Characteristic Waveform, such that:

$$u(t_i, \phi(t_i)) = u_{SEW}(t_i, \phi) + u_{REW}(t_i, \phi) \quad \ldots\ldots \quad (2.4)$$

The SEW can be sampled at a low rate, while the REW requires a high sampling rate. Only the magnitude spectrum of the REW is transmitted, and the quantization accuracy required for this magnitude is low.

The SEW/REW decomposition is accomplished with a simple filtering operation. Let *h(m)* represents the impulse response of a low-pass filter, the SEW is then

$$u_{SEW}(t_i, \phi) = \sum h(m) u(t_{i-m}, \phi) \quad \ldots\ldots \quad (2.5)$$

The REW can then be obtained from combining eq. (2.3) and eq. (2.4).

$$u_{REW}(t_i, \phi) = u(t_i, \phi(t_i)) - u_{SEW}(t_i, \phi) \quad \ldots\ldots \quad (2.6)$$

2.3.2 WI Encoder

The Characteristic Waveform surface extraction and the subsequent SEW/REW decomposition introduced above are the common features in the WI coding

paradigms. There are a variety of WI coding schemes, developed by many researchers, which differ in the methods of CW extraction and the representation.

In some early WI coders , the CW extraction is performed in the speech domain [28], [31], but it was found that the residual domain extraction will reduce the discontinuity in the prototype [33]. Residual domain extraction is thus used in a majority of WI coders [6], [32], [33].

The prototype (CW) representations also differ across WI class coders. The prototype can be represented in the time domain [5], as well as the frequency domain (DFT) [6], [33]. Although time domain representations are computationally less complex, the advantage of the DFT representation is that it can efficiently separate the magnitude and phase spectrum of the prototype [6]. This makes it possible to quantize the magnitude and phase spectrum of the prototype separately according to the perceptual properties. The frequency domain (DFT) representation also makes incorporation of the masking properties of the human perception system in the prototype quantization more convenient.

A 2.4kb/s Waveform Interpolation (WI) coder is presented in this section (encoder) and the following section (decoder). Base on above discussion, the entire WI coding procedure operates on the linear-prediction (LP) residual of the input speech, and the extracted discrete CW is described in Fourier series. The basic coding structure is from a 2.84kb/s WI coder developed by Burnett [6]. This coder is designed to operate with a telephone bandwidth (200Hz~3400Hz) sampled at 8000Hz. The coder operates

on speech frames of 25ms corresponding to 200 samples. The speech signal is analyzed to extract the parameters of the WI coder for every 25ms frame. Figure 2.3 provides a block diagram of the encoder.

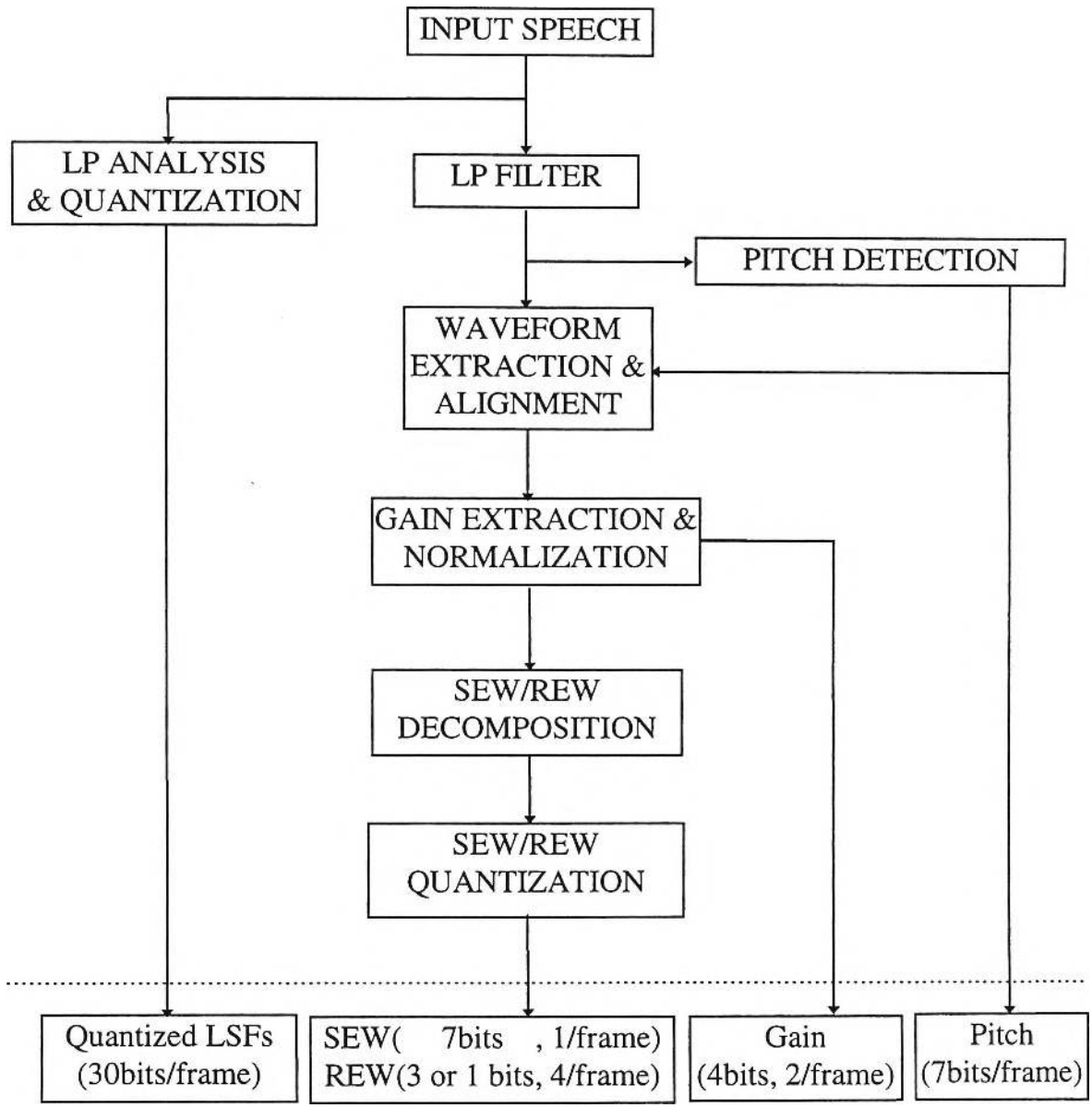


Figure 2.3: Diagram of WI encoder

The speech signal is first converted to the residual domain via a linear-predictive (LP) analysis filter. The LP parameters are calculated once per frame and quantized as LSF

vectors using a split-VQ algorithm (the LSF parameters are linearly interpolated). The pitch period is extracted from this residual signal once per frame. The pitch value is interpolated, and ten (interpolated) pitch length prototypes are extracted from the residual on the time axis and converted to the transform domain by performing a DFT calculation. After alignment, the prototypes form a two-dimensional discrete Characteristic Waveform surface (corresponding to $u(t,\phi)$ downsampled to a rate of 400Hz) in the DFT domain. For convenience and gain quantization purposes, the gain of each Characteristic Waveform is extracted and the CW surface normalized. By filtering this surface along the time axis, the surface is decomposed into two underlying components, the rapidly-evolving waveform (REW) and the slowly-evolving waveform (SEW). The parameters, gain, SEW, REW are down sampled such that the update rates of gain, SEW and REW are 80Hz (twice per frame), 40Hz (once per frame) and 160Hz (four times per frame) respectively. After quantization, the information for all parameters is transmitted.

Parameter	Codebook Size	Update rate per frame	Total per frame
LPC	30	1	30
Pitch	7	1	7
SEW	7	1	7
REW	1 or 3	4	8
Gain	4	2	8
Total			60

Table 2.2: Bit allocation for the 2.4kb/s WI coder

Table 2.2 shows the bit allocation. Details of the encoding procedures are described as follows.

2.3.2.1 LP Analysis and Quantization

LP Analysis

Linear prediction (LP) techniques are widely used in modeling the speech signal in many low bit rate speech coders, including CELP, MBE, and WI. This model assumes an excitation and the vocal tract modelled as an all-pole filter. The excitation signal (LP residual signal) has a white spectrum. The filter coefficients are obtained using one of numerous algorithms [45], [54]. In this thesis the autocorrelation technique attributed to Schur is utilised [45], [54].

In this thesis, a 10th-order linear predictive coding (LPC) filter is used. The LP residual signal *r(t)* is obtained from the speech signal *s(t)* by linear predictive (LP) filtering:

$$r(t) = s(t) + \sum_{n=1}^{10} a_n s(t-n) \qquad \ldots\ldots \qquad (2.7)$$

Figure 2.4 shows a segment of speech signal and LP residual signal sampled at 8000Hz.

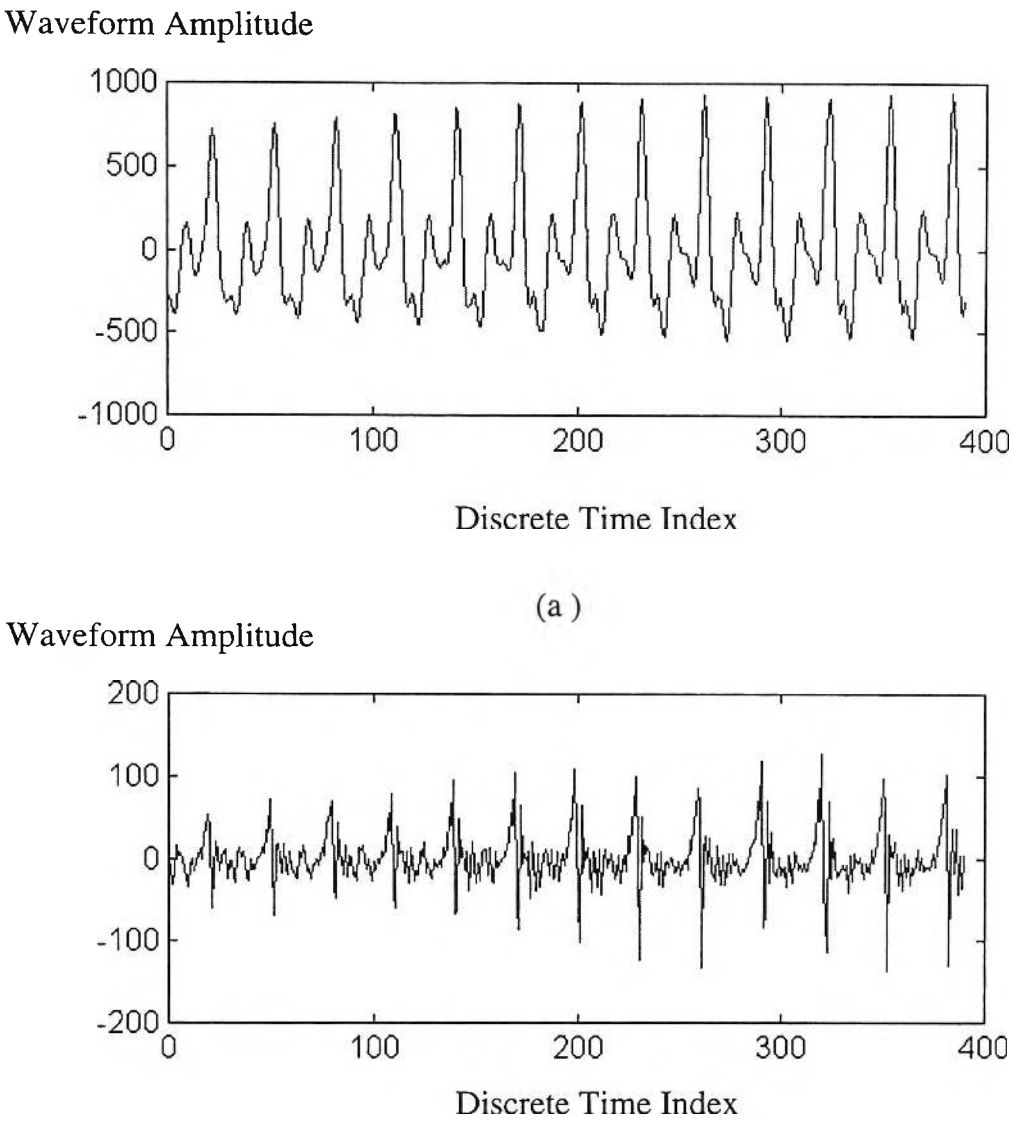


Figure 2.4: (a) Original speech; (b) LP residual of speech.

LSF Calculation

Transmission of the LPC coefficients consumes a large part of the total bit rate, especially at low bit rates. An efficient method of coding the LPC coefficients is the quantization of Line Spectral Frequencies (LSFs), also known as Line Spectral Pairs (LSP) [26], [44].

Prediction and reflection coefficients are frequently used as LPC parameters, however, the implementation of Line Spectral Frequencies (LSFs) provides a more efficient encoding than the prediction and reflection coefficients [26]. LSFs have some intrinsic properties which make it possible to employ significant bit-saving measures. In LSF quantization, one line spectrum only associates with the spectrum near that frequency. Thus, LSFs can be quantized in accordance with properties of auditory perception (i.e., coarse representation of the higher frequency components of the speech spectral envelope). This property also makes it possible to interpolate LSFs in speech coding(leading to smooth evolution of the speech spectrum), which is not possible for LPC prediction and reflection coefficients.

The definition of Line Spectral Frequencies (LSFs) results from the decomposition of the LP analysis filter into even and odd functions [26] [44]. The nth-order LP analysis filter is defined as:

$$A_n(z) = 1 - a_1 z^{-1} - a_2 z^{-2} - \cdots - a_n z^{-n} \qquad \ldots\ldots \qquad (2.8)$$

where a_n is the nth prediction coefficient. By taking a difference and sum between $A_n(z)$ and its conjugate function, the LP analysis filter is decomposed into a difference filter and a sum filter:

$$P_{n+1}(z) = A_n(z) - z^{-(n+1)} A_n(z^{-1}) \qquad \ldots\ldots \qquad (2.9)$$

$$Q_{n+1}(z) = A_n(z) + z^{-(n+1)} A_n(z^{-1}) \qquad \ldots\ldots \qquad (2.10)$$

$P_{n+1}(z)$ is the difference filter, and $Q_{n+1}(z)$ is the sum filter. The LP analysis filter can be reconstructed from these two filters:

$$A_n(z) = \frac{1}{2}[P_{n+1}(z) + Q_{n+1}(z)] \quad \ldots\ldots \quad (2.11)$$

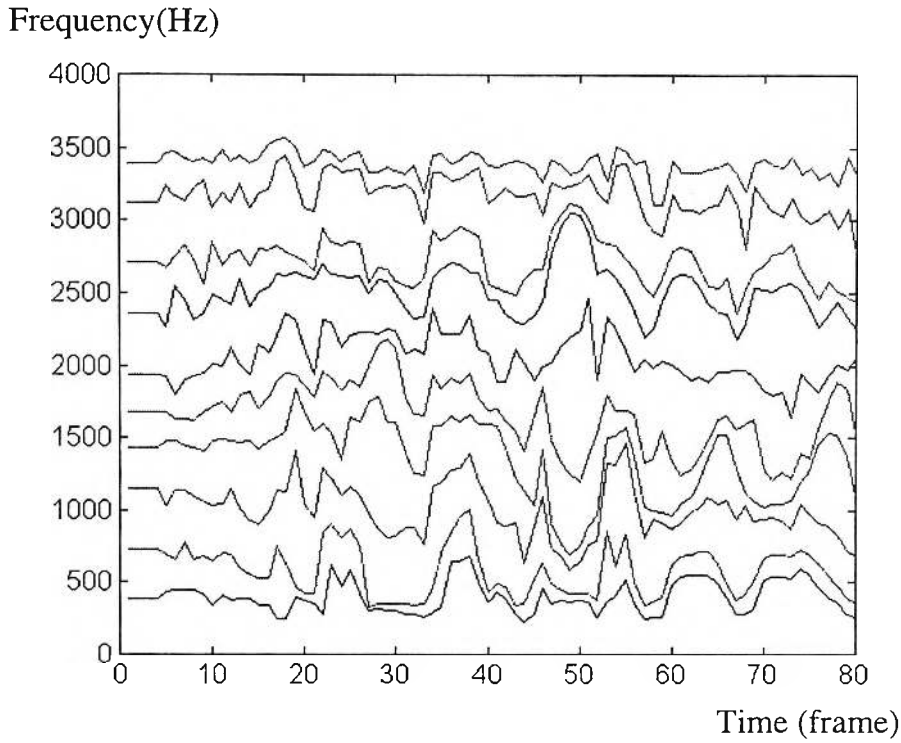


Figure 2.5: An example of LSF trajectories (the frame length is 25ms)

The roots of the difference and sum filter are the lower and upper line-spectra of the LSF. Figure 2.5 shows an example of LSF trajectories. Thus, the difference and sum filter can be described as:

$$P_{n+1}(z) = (1 - z^{-1}) \prod_{k=1}^{n/2} [1 - 2 \times \cos(2\pi f_k / f_s) z^{-1} + z^{-2}] \quad \ldots\ldots \quad (2.12)$$

$$Q_{n+1}(z) = (1 + z^{-1}) \prod_{k=1}^{n/2} [1 - 2 \times \cos(2\pi f_k' / f_s) z^{-1} + z^{-2}] \quad \ldots\ldots \quad (2.13)$$

where f_k and f_k' are the lower and upper line-spectra of the Kth LSF, and f_s is the sampling rate of speech.

The roots of both the difference and sum filters are located on the unit circle of the z-plane, and the roots of the difference and sum filter are interweaved with each other so that the LSFs are in ascending order.

LSF Interpolation

In the WI coder described here, the LPC coefficients are calculated once per frame and converted to LSFs. Every frame is divided into five segments. In each segment, the LSFs are interpolated between the previous, current and future frame. The residual signal in each segment is obtained by using the interpolated LSFs. The LSF interpolation operation in the LP filtering procedure makes the residual signal smoother [33].

LSF Quantization

An error in one line-spectrum only distorts the spectrum of the LPC filter near that line-spectrum, and will not spread over the whole spectrum. Thus, LSFs can be quantized economically by exploitation of human auditory perception. As the low part of the frequency spectrum is perceptually more significant than the high part, the low LSFs are quantized more accurately than the high LSFs. LSF coefficients are represented by three 10-bit vectors from a split-VQ codebook mechanism [44]. Three 10-bit codebooks are assigned for the first three LSFs, the second three LSFs and the last four LSFs respectively. The LSFs are quantized by using mean-squared error (MSE) criteria.

2.3.2.2 Waveform Extraction and Alignment

By successive extraction and alignment of the pitch-cycle prototypes (Characteristic Waveforms), the one dimensional speech signal is transformed into a two dimensional discrete CW surface. This requires an accurate pitch value and a relatively simple prototype extraction process.

Pitch Interpolation & Waveform Extraction

The pitch estimation and waveform extraction procedure operates on the linear-prediction (LP) residual domain. The pitch period is calculated once per frame and ten pitch-length prototypes are extracted from each frame along the time axis. The pitch value of the prototype is obtained by interpolation of the pitch periods between the previous, present and future frames.

The location of the extracted waveform is adjusted by an offset so that the signal energy near the boundaries is minimized. This will prevent significant discontinuities while interpolating between different prototypes [33]. The other advantage of this adjustment is that it can reduce distortion in the prototype if the pitch estimation is wrong. Figure 2.6 gives an example. If a prototype is extracted such that it has high energy boundaries, pitch errors will affect the prototype severely (the pulse in residual is often duplicated or misplaced). For prototypes bounded with low energy samples, pitch errors result in only minor distortion (see Figure 2.6).

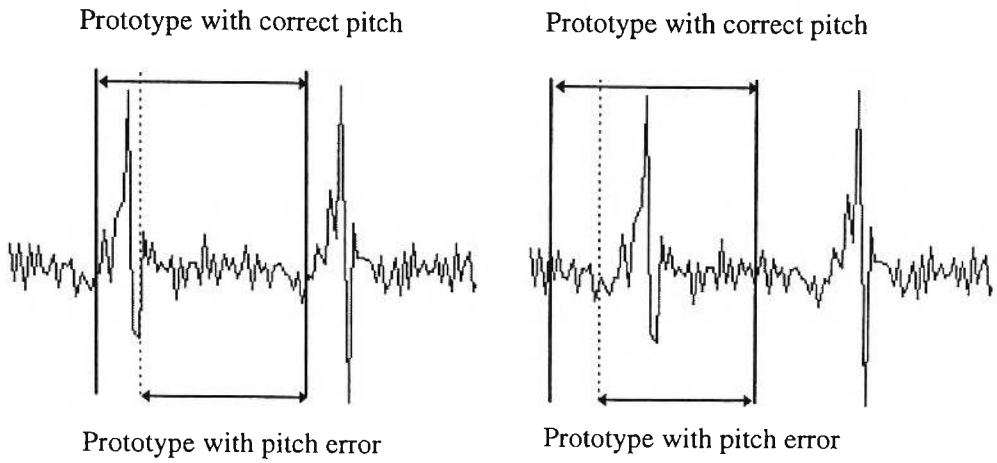


Figure 2.6: (a) Prototype started from high energy part of signal. (b) Prototype started from low energy part.

Since the LP residual signal has a clear pitch pulse and a low-energy portion between pulses, it is convenient to perform this procedure in the residual domain. The extracted prototype at time t_i is then

$$\hat{v}(t_i,t) = r(t_i - \frac{p(t_i)}{2} + t + \Delta), \qquad 0 \le t \le p(t_i) \qquad \ldots\ldots \qquad (2.14)$$

where $r(t)$ is the LP residual signal and Δ is the offset. Δ can be up to 5ms in length. $p(t_i)$ is the discrete pitch length.

$$p(t_i) = \frac{P}{T} \qquad \ldots\ldots \qquad (2.15)$$

where P is the pitch period, T is the sampling interval.

After the time domain prototype is extracted, it is converted to the transform domain by DFT:

$$\hat{V}(t_i,\phi)=\sum_k \alpha_k(t_i)\cos(k\phi)+\beta_k(t_i)\sin(k\phi)=DFT\bullet\{\hat{v}(t_i,t)\}$$

$$0\le k\le p(t_i) \quad \ldots\ldots \qquad (2.16)$$

Alignment

Following the prototype extraction the next step is alignment of the prototype or Characteristic Waveform along the t axis. The phase of the prototype should be adjusted so the smoothness of the Characteristic Waveform surface will be maximized in the t direction. The alignment procedure can be accomplished by alignment in the ϕ axis of the present extracted prototype with the previous prototype. The phase shift is then [33]:

$$\hat{U}(t_{m+1},\phi)=\hat{V}(t_{m+1},\phi+\phi_u) \qquad \ldots\ldots \qquad (2.17)$$

$$\phi_u=\max_{\phi_e}(\sum_{k=0}^{K-1}(\hat{V}(t_{m+1},\phi+\phi_e)\hat{U}(t_m,\phi)^*).re$$

$$K=\max\{p(t_m),\ p(t_{m+1})\}$$

where $p(t_m)$ and $p(t_{m+1})$ are the pitch value of the two prototypes, $\hat{U}(t_{m+1},\phi)$ is the aligned prototype at time t_{m+1}, $\hat{U}(t_m,\phi)$ is the prototype at the previous time interval t_m and ϕ_u is the phase shift. If the two prototypes being aligned have different pitch lengths, the shorter one is padded with zeros at the end to the length of the longer one. After the alignment procedure, the prototype which has been padded with zeros is truncated to its original length.

Pitch Doubling

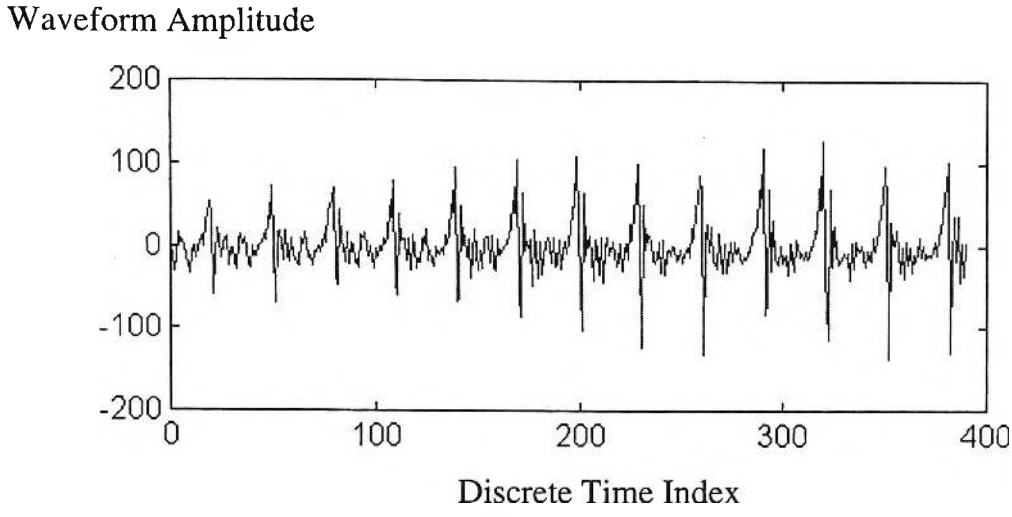
If the pitch doubles between the two prototypes to be aligned, the length of the prototype which contains the single pitch cycle waveform is doubled before alignment. The detail of the procedure is described below. When pitch doubling happens, the prototype will contain two pitch cycle waveforms in the time domain. Equivalently in the DFT domain, the even coefficients of the prototype are zeros or very small values, and the odd coefficients correspondent to the DFT coefficients of the one pitch cycle prototype. Thus the prototype which contains the single pitch cycle waveform can be converted to a prototype containing two pitch cycle waveforms (i.e. a pitch doubled prototype) by:

$$\begin{aligned} \hat{U}_{double}(t_m, 2k) &= \hat{U}(t_m, k) \\ \hat{U}_{double}(t_m, 2k+1) &= (0,0) \end{aligned} \qquad \ldots\ldots \qquad (2.18)$$

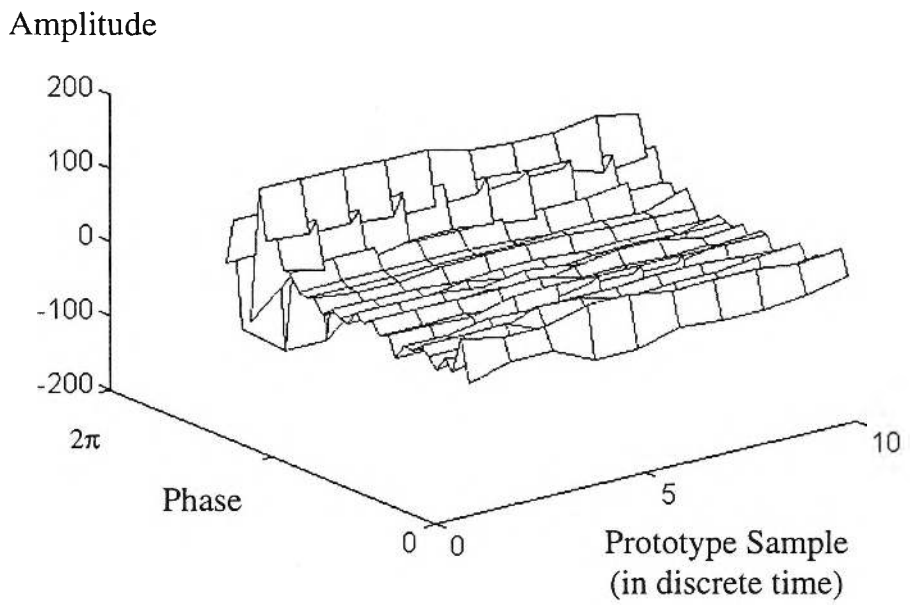
where $\hat{U}_{double}(t_m, k)$ is the prototype with doubled pitch. $\hat{U}_{double}(t_m, k)$ can be converted back to the one pitch cycle prototype by:

$$\hat{U}(t_m, k) = \hat{U}_{double}(t_m, 2k) \qquad \ldots\ldots \qquad (2.19)$$

Figure 2.7 shows the one dimensional speech residual signal and the two dimensional CW surface (in time domain). The residual signal is the same as that shown in Figure 2.5.



(a)



(b)

Figure 2.7: (a) LP residual of speech; (b) Two dimensional Characteristic Waveform in residual domain.

2.3.2.3 Gain Extraction and Quantization

After alignment of the residual domain prototypes, the gain of each prototype is extracted. In practice, the prototype is converted to the speech domain through a LP synthesis filter, and the gain of the prototype is computed in the speech domain. This makes the signal gain independent of the gain of the LP synthesis filter, which means the speech power contour will be reserved even when the LSF transmission or residual parameters are in error. Equations (2.20) and (2.21) are used to extract the gain of the prototype at a given time interval t_i :

$$U'(t_i,k) = \frac{U(t_i,k)}{A(k)} \qquad \ldots\ldots \qquad (2.20)$$

$$G(t_i) = \frac{1}{K}\sum_{k=0}^{K-1}\left|U'(t_i,k)\right| \qquad \ldots\ldots \qquad (2.21)$$

where $A(k)$ is the LP filter, $U'(t_i,k)$ is the speech domain prototype, and $G(t_i)$ is the extracted gain.

The signal gain is then converted to the logarithmic domain and low-pass filtered. The filter used here is a 21-tap FIR filter with a cut-off frequency of 40Hz. The gain is down-sampled to 80Hz (two gain per frame). It is quantized with a differential quantizer using a 4-bit scalar codebook. In the decoder, the gain is decoded, and then upsampled to 400Hz (the sampling rate of the prototype) by interpolation. As some changes in log speech gain can be fast, both linear and step-wise interpolation are used. For small changes in signal gain, the gain is linearly interpolated between

successive intervals. For large changes in signal gain, step-wise interpolation is used according to the following decision process [33]:

$|d(\lg G(t_i))| > 0.3$ step-wise interpolation

$|d(\lg G(t_i))| < 0.3$ linear interpolation

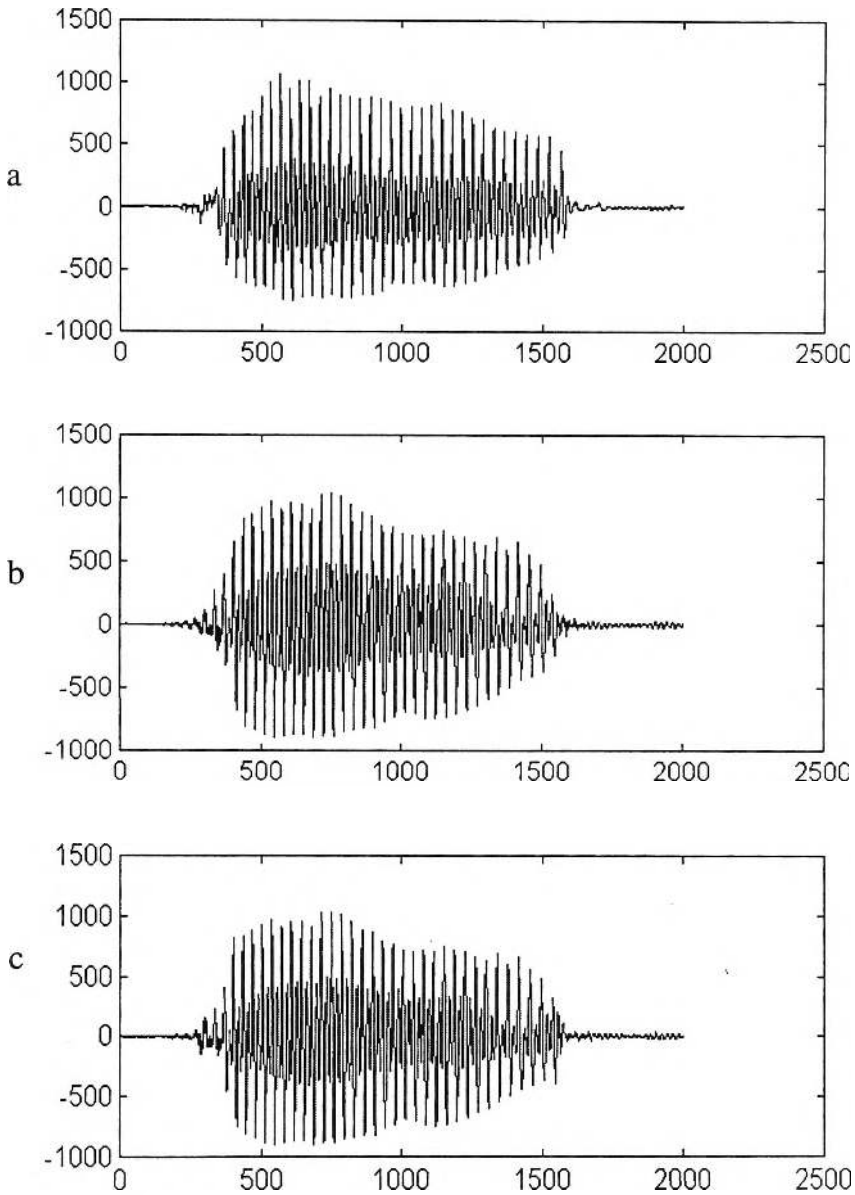


Figure 2.8: (a) Original speech waveform; (b) Coded speech using only linear interpolation gain quantization; (c) Coded speech using both linear and step-wise interpolation gain quantization.

Figure 2.8 gives an example of the gain quantization. At the start and the end of the original speech, the signal power changes brutally. (see in Figure 2.8(a)). Using only linear gain interpolation, the speech signal always changes slowly and fails to catch fast changes in the signal power (see Figure 2.8(b)). By using both the linear interpolation (for small gain changes) and the step-wise interpolation (for large gain changes), fast change of the signal power can be seen duplicated in the output speech. (see Figure 2.8(c)).

2.3.2.4 SEW/REW Decomposition and Quantization

Once the discrete CW surface $\hat{U}(t_i, \phi)$ (sampled at 400Hz) is obtained, it is decomposed into a slowly evolving waveform (SEW) and a rapidly evolving waveform (REW). The SEW can be obtained as the weighted average spectrum of the prototypes within the analysis frame. The REW is the difference between the incoming prototype and the SEW [7].

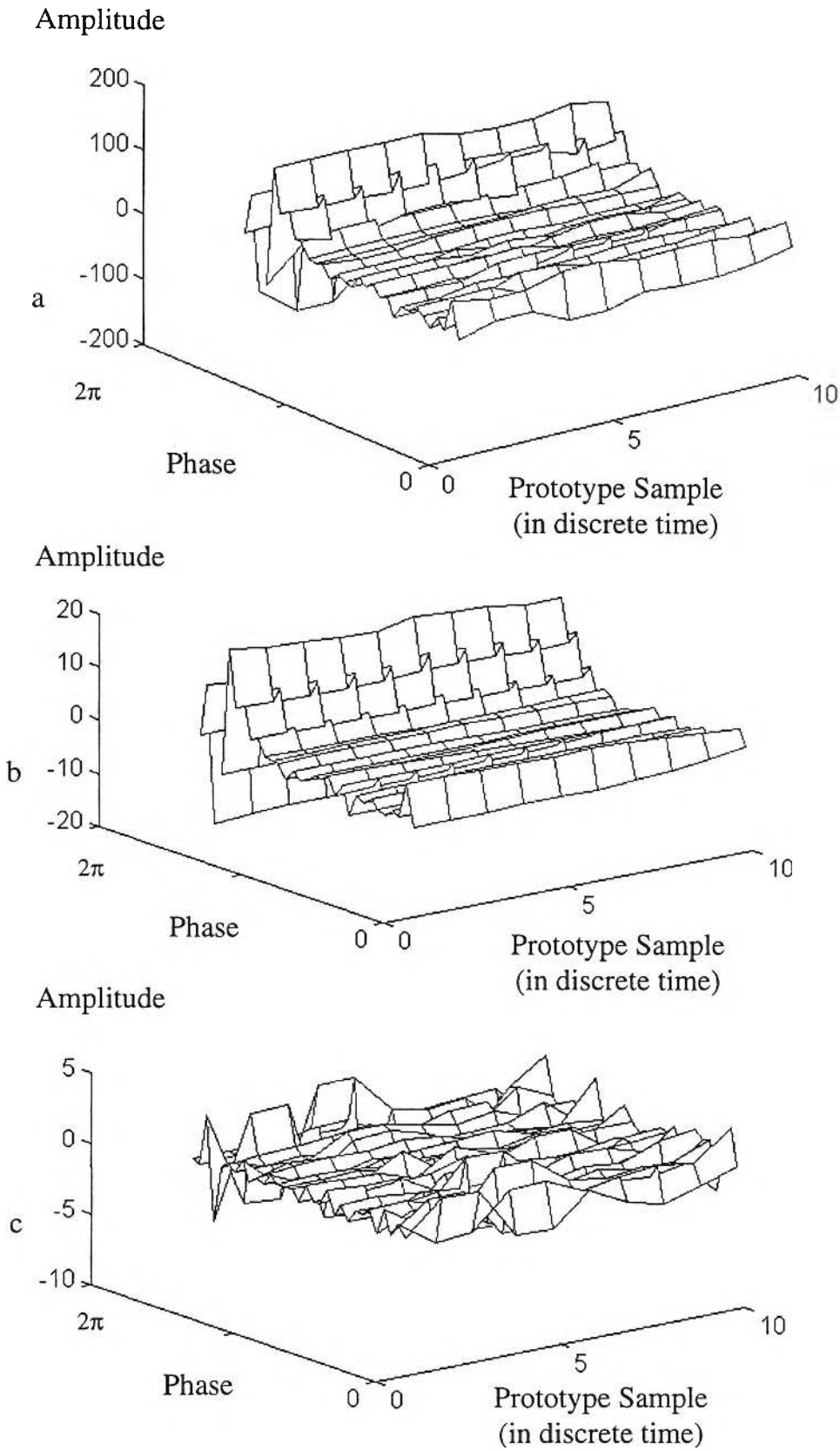


Figure 2.9: (a) Characteristic Waveform surface (b) Slowly-evolving waveform (SEW) (c) Rapidly evolving waveform (REW)

Figure 2.9 gives one example of the SEW/REW decomposition. The CW surface in Figure 2.7 is decomposed into the SEW and REW surfaces.

The SEW and REW surfaces are gain-normalized and then down-sampled. The transmission rate of the SEW is one SEW per frame (40Hz), and the REW are transmitted four times per frame (160Hz), twice as a REW vector index (3bit) and twice as a binary decision between the previous and next transmitted REW. Since the SEW phase spectrum is perceptually significant, in the baseline coder the whole SEW spectrum is quantized as a complex vector. For the REW, the phase and magnitude spectrum are separated. Only the magnitude spectrum of the REW is quantized.

2.3.3 WI Decoder

The decoder diagram is shown in Figure 2.10. The first step is decoding the SEW and the REW. The prototype waveform (Characteristic Waveform) is obtained by adding the SEW and REW together. After, the prototype is converted from the residual domain to the speech domain by a linear-predictive (LP) synthesis filter and post filter. The speech domain waveform is gain-scaled and time-aligned. Then the Characteristic Waveform is converted into output speech through continuous interpolation in the DFT domain.

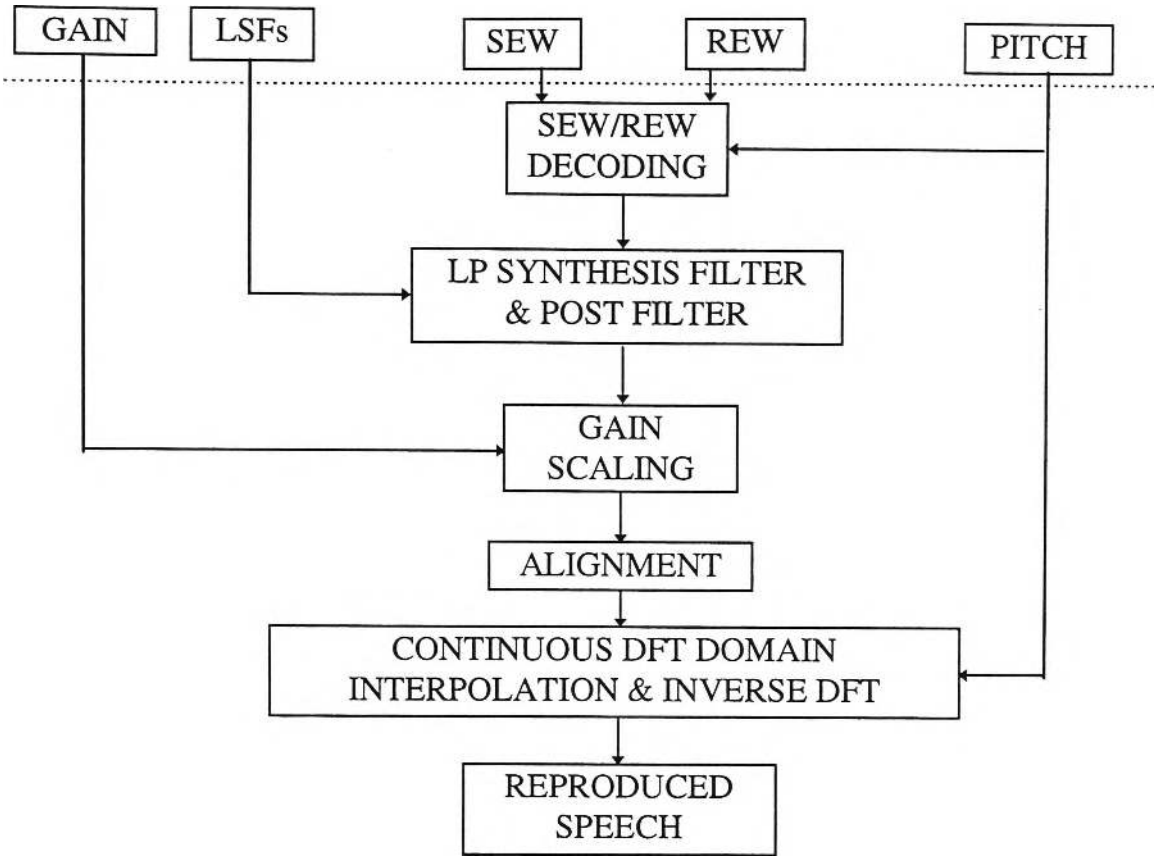


Figure 2.10: Diagram of WI decoder

2.3.3.1 SEW/REW Decoding

In each frame, ten SEWs and REWs are obtained by decoding the transmitted SEW and REW codebook indices. The SEW surface is reconstructed by interpolation of the SEW of the previous, current and future frames. For the REW, the magnitude spectrum is derived from the REW codebook. The REW phase is approximated by a uniformly distributed Gaussian spectrum [7].

Figure 2.11 shows one frame of the decoded SEW and REW surfaces (sampled at 400Hz), correspondent to the original SEW and REW in Figure 2.8.

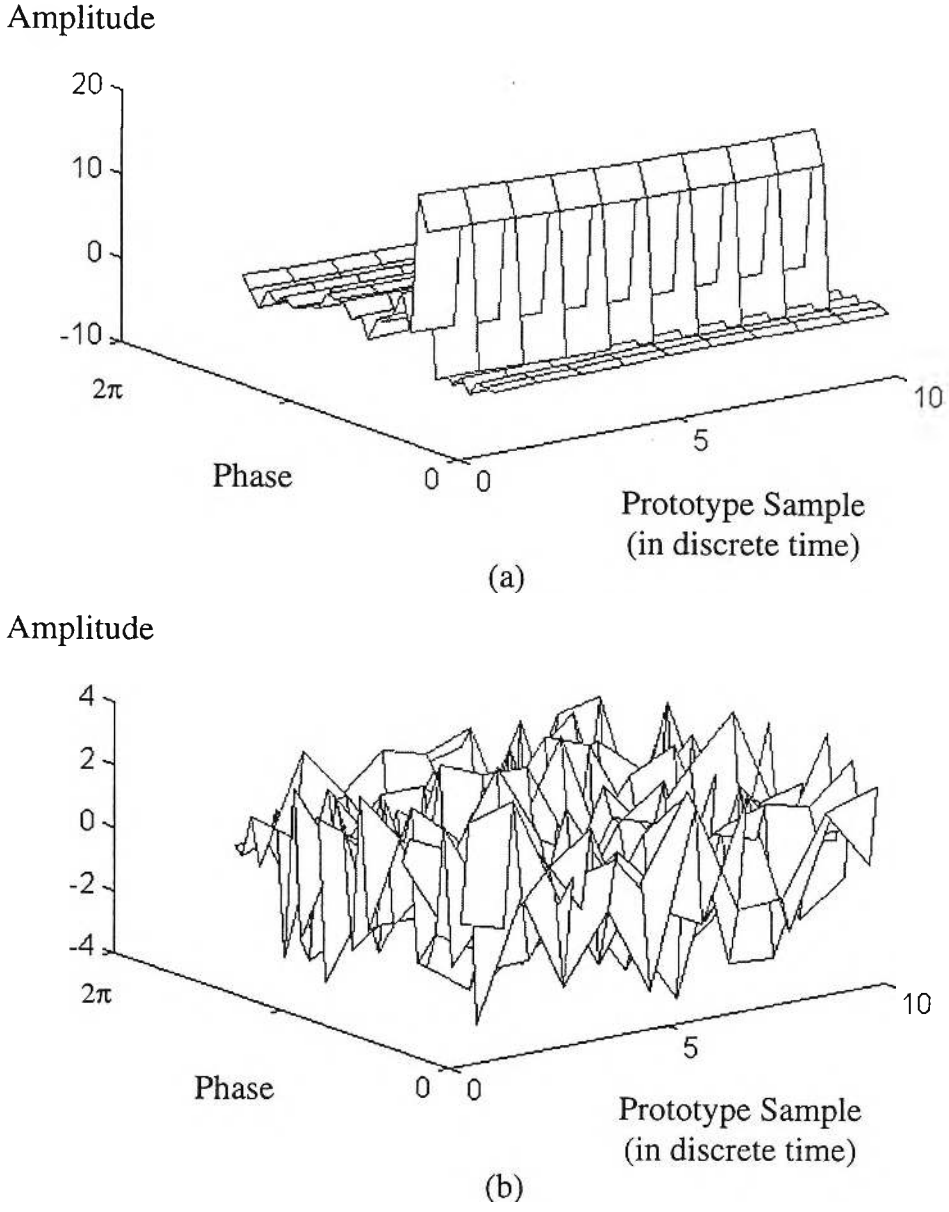


Figure 2.11: (a) Decoded SEW (b) Decoded REW

2.3.3.2 Synthesis Filter

LP Synthesis Filter

The residual domain prototype is obtained by adding the decoded SEW and REW together. The residual domain prototype is then converted into the speech domain

through an LP synthesis filter. The relation between the residual domain and speech domain prototype is described in eq.(2.22):

$$U'(t_i,\phi(t_i)) = U(t_i,\phi(t_i)) - \sum_{n=1}^{N} a_n U'(t_{i-n},\phi(t_{i-n})) \qquad \ldots\ldots \qquad (2.22)$$

$$A = \{a_1, a_2, \cdots, a_N\}$$

where a_n is the coefficient of Nth LP filter A, $U(t_i,\phi(t_i))$ and $U'(t_i,\phi(t_i))$ are the residual domain and speech domain prototypes respectively. The inverse relation is (The prototype are periodic on phase axis.):

$$\begin{aligned} U(t_i,\phi(t_i)) &= U'(t_i,\phi(t_i)) + \sum_{n=1}^{N} a_n U'(t_{i-n},\phi(t_{i-n})) \\ &= U'(t_i,\phi(t_i)) \times A \end{aligned} \qquad \ldots\ldots \qquad (2.23)$$

It is convenient to perform this convolution in the transform domain. From eq.(2.23), we obtain:

$$U(t_i,k) = U'(t_i,k) \times A(k) \quad \text{or}$$

$$U'(t_i,k) = \frac{U(t_i,k)}{A(k)} \qquad \ldots\ldots \qquad (2.24)$$

$$A(k) = DFT \bullet (A)$$

where A(*k*) are the DFT coefficients of the LP filter *A*. In contrast to the time domain LP synthesis filtering, the DFT domain convolution does not add delay to the coder.

Post-filter

Low bit rate coders usually introduce some roughness to the reconstructed speech. A postfilter operation at the decoder's output can enhance the speech quality. The post-filtering procedure exploits the human ear's masking properties to trade off speech distortion vs quantization noise [11]. In speech perception, the formants of speech are perceptually more important than spectral valley regions. Therefore, the postfilter attenuates the components in spectral valleys. The post-filtering procedure reduces perceived noise and only introduces minor distortion in the output speech.

The post-filtering procedure contains an adaptive postfilter H_p and a tilt compensation filter H_t. The adaptive postfilter should follow the formants and valleys of the input speech. As the frequency response of the LP synthesis filter is close to the spectral envelope of speech, the postfilter is derived from the LP filter $A(z)$, by scaling down the poles by a factor of α ($0<\alpha<1$). This filter $A(z/\alpha)$ has lower formant peaks than that of $A(z)$. To reduce the spectral tilt of the all-pole filter $A(z/\alpha)$, an all-zero filter is added [11]. In a similar manner to the LP synthesis procedure, the post-filtering procedure is performed in the DFT domain. The adaptive postfilter is given by:

$$H_p = \frac{A(k/\beta)}{A(k/\alpha)} \qquad \ldots\ldots \qquad (2.25)$$

To achieve the best performance, the values of α and β are selected to be 0.8 and 0.5 respectively [11]. Figure 2.12(b) shows the response of the adaptive postfilter.

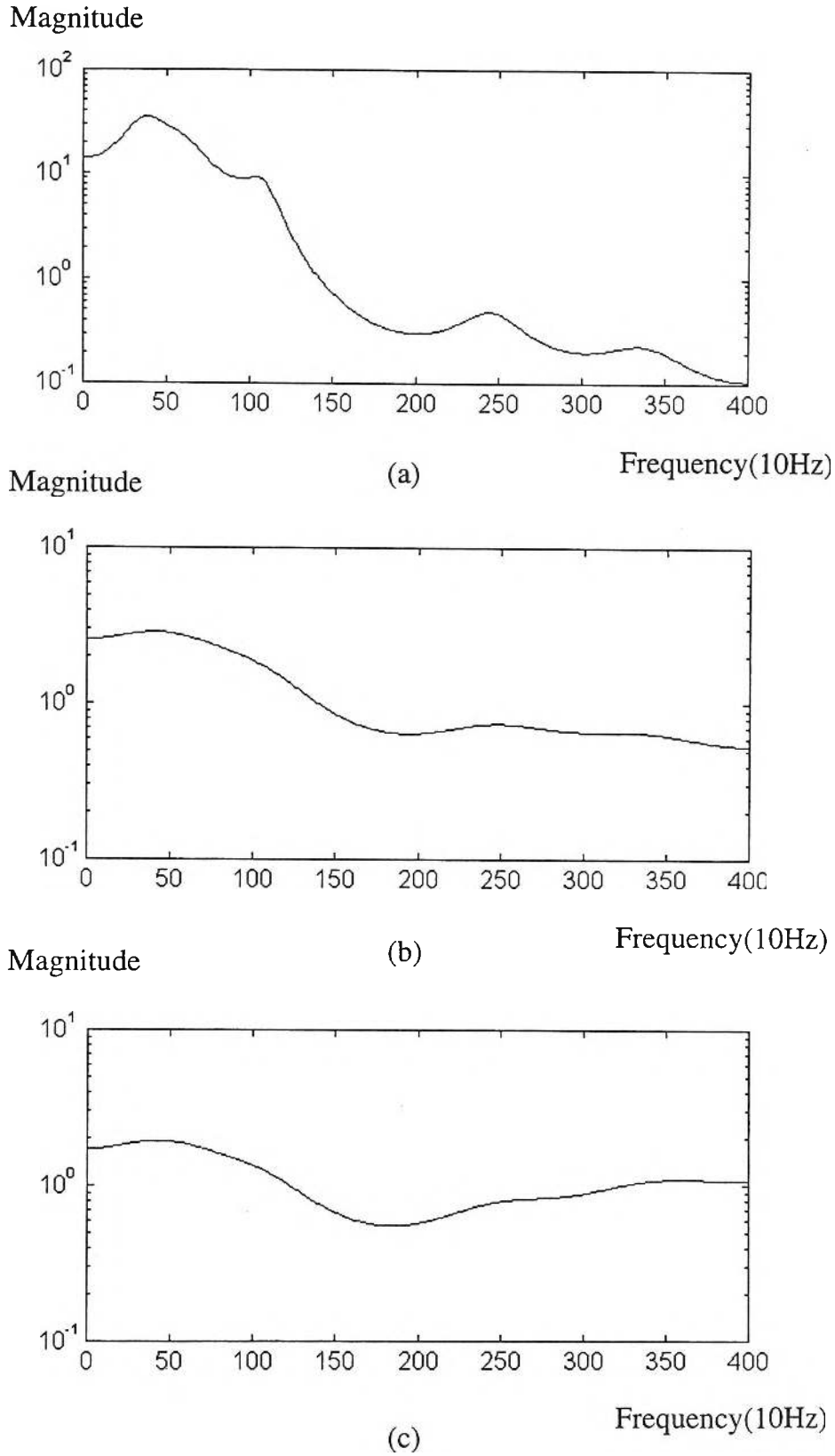


Figure 2.12: (a) Frequency response of the LPC filter; (b) Frequency response of the adaptive post filter; (c) Frequency response of the postfilter (with tilt compensation).

The adaptive postfilter introduces a muffling effect. A first order high-pass filter is used to compensate the tilt effect [11]:

$$H_t = 1 - 0.5 \cdot z^{-1} \qquad \ldots\ldots \qquad (2.26)$$

The overall frequency response of the postfilter is shown in Figure 2.12(c). Note that the frequency response has flat formant peaks, and the spectral tilt is greatly reduced.

Figure 2.13 shows the reconstructed discrete CW (in speech domain) surfaces from the decoded SEW and REW in Figure 2.11.

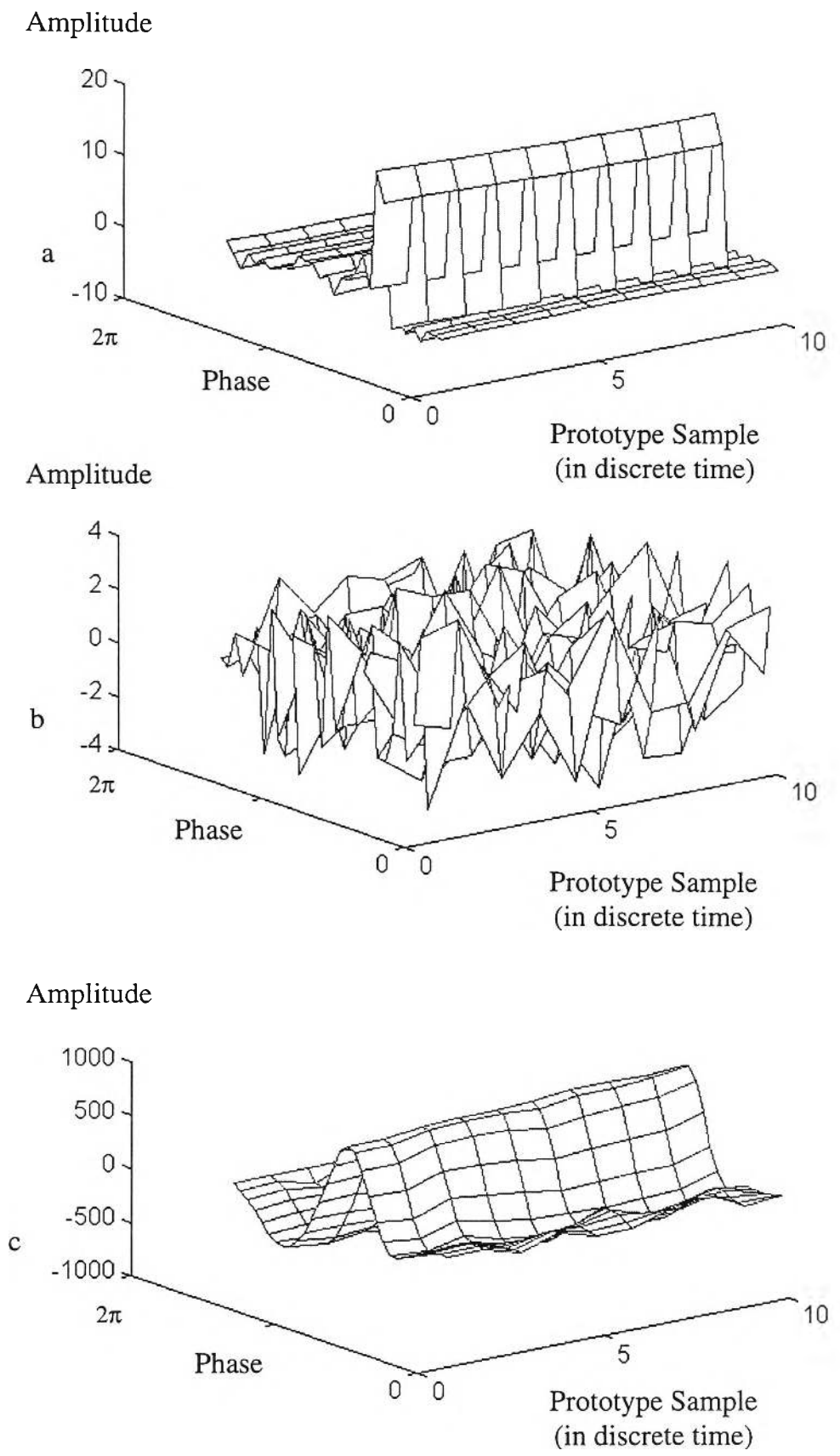


Figure 2.13: (a) Decoded SEW (b) Decoded REW (c) Characteristic Waveform in speech domain (post-filtered)

2.3.3.3 Speech Reconstruction

Gain-Scaling

After the normalized residual domain prototypes are converted into the speech domain, they are gain-scaled in that domain.

$$U'_g(t_i,k) = \frac{U'(t_i,k)}{\sum_{k=0}^{K-1}\left|U'(t_i,k)\right| \Big/ K} \times G(t_i) \qquad \ldots\ldots \qquad (2.27)$$

where $U'_g(t_i,k)$ is the speech domain gain-scaled prototype.

Continuous Interpolation

Finally, after time-alignment, the gain-scaled prototypes are converted into output speech by continuous interpolation. The DFT coefficients of the prototypes are interpolated at every output point, and the reproduced speech is obtained by an effective inverse DFT calculation.

The reconstructed speech *op(t)* at an output point *t* which is between the prototype update interval t_{i-1} and t_i is given by:

$$\begin{aligned} op(t) &= \sum_{k=1}^{K(t)} \alpha_k(t)\cos(k\phi) + \beta_k(t)\sin(k\phi) \\ &= \sum_{k=1}^{K(t)} \alpha_k(t)\cos(k\frac{t-t_{i-1}}{K(t)}) + \beta_k(t)\sin(k\frac{t-t_{i-1}}{K(t)}) \end{aligned} \qquad \ldots\ldots \qquad (2.28)$$

where $\alpha_k(t)$ and $\beta_k(t)$ are the DFT coefficients at time *t*, and *K(t)* is the pitch value (prototype length) at time *t*.

$\alpha_k(t)$, $\beta_k(t)$ and *K(t)* are obtained by continuous interpolation of the parameters of the prototypes transmitted at t_{i-1} and t_i.

$$\alpha_k(t) = \frac{t - t_{i-1}}{t_i - t_{i-1}} \alpha_k(t_i) + \frac{t_i - t}{t_i - t_{i-1}} \alpha_k(t_{i-1}) \quad \ldots\ldots \quad (2.29)$$

$$\beta_k(t) = \frac{t - t_{i-1}}{t_i - t_{i-1}} \beta_k(t_i) + \frac{t_i - t}{t_i - t_{i-1}} \beta_k(t_{i-1})$$

$$K(t) = \frac{t - t_{i-1}}{t_i - t_{i-1}} K(t_i) + \frac{t_i - t}{t_i - t_{i-1}} K(t_{i-1})$$

The DFT coefficients and pitch period of the prototype at time interval t_i are $\alpha_k(t_i)$, $\beta_k(t_i)$ and $K(t_i)$ respectively.

Figure 2.14 shows the output of the WI decoder. Compared with the input speech (Figure 2.5), the speech is closely reconstructed, excepting phase difference between the input and output signal, in contrast with waveform coders such as CELP, the decoded speech is not synchronous with the original speech. These phase differences are caused by the lack of prototype phase information retained in the extraction process.

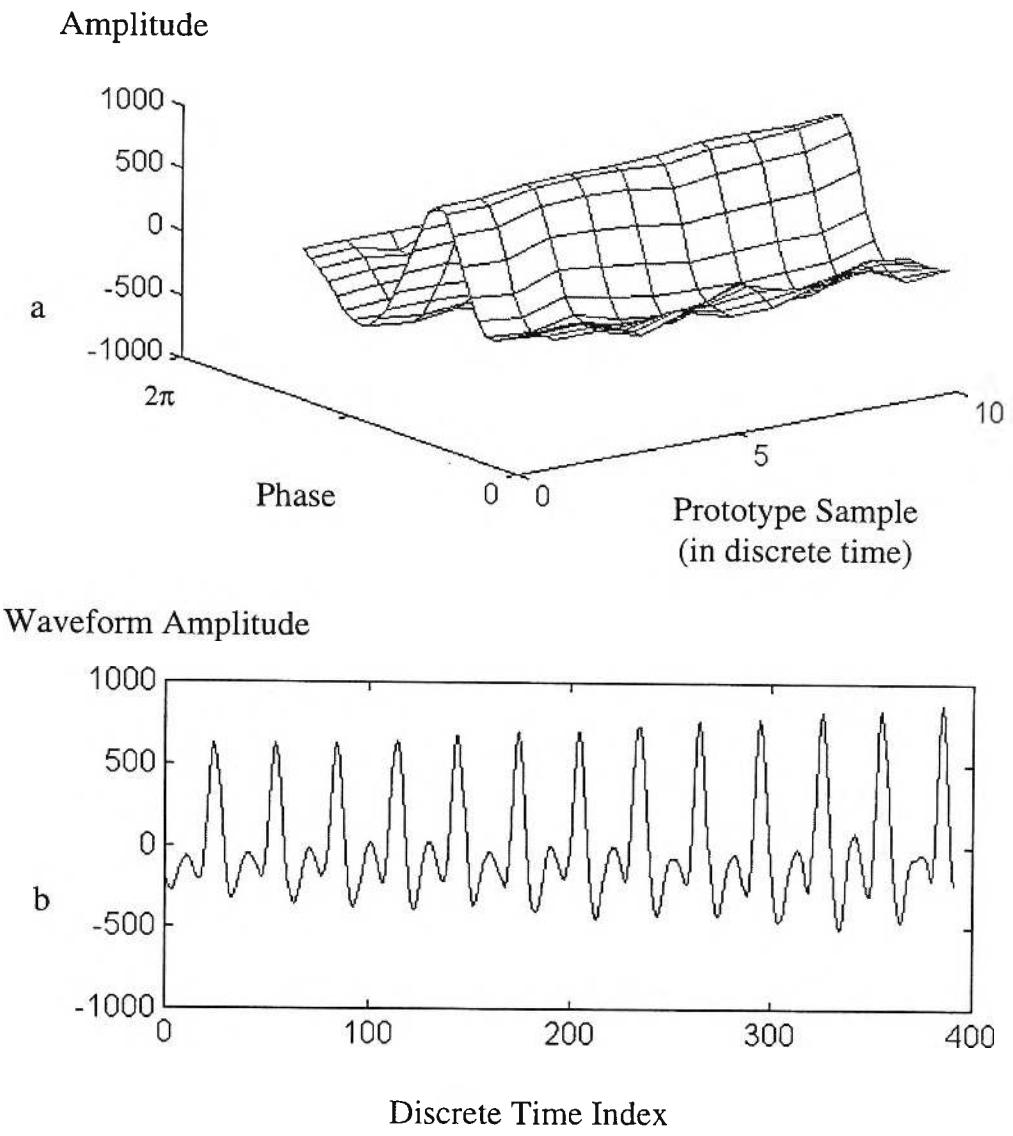


Figure 2.14: (a) Characteristic Waveform; (b) Reconstructed speech.

2.4 CELP Algorithm

2.4.1 Outline of the CELP Coder

Code excited linear prediction (CELP) was proposed in the mid-1980s. A CELP coder [1], [2], [3], [17], [18] consists of a slowly time-varying linear prediction (LP) filter and an excitation signal. The linear prediction filter is periodically updated and is

determined by analysis of the current segment of speech. The CELP algorithm uses vector quantization (VQ) to determine the excitation signal. A set of excitation vectors (Gaussian sequences) is stored in a codebook. The excitation signal is determined by analysis-by-synthesis techniques. The encoder determines the excitation signal by feeding candidate excitations into an LP synthesis filter and selecting the one that minimizes the perceptually weighted error between the original and reproduced speech.

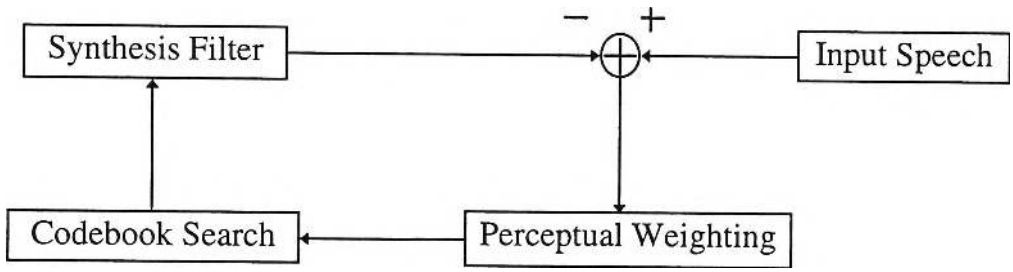


Figure 2.15: Encoding principle of CELP algorithm

2.4.2 Analysis-by-Synthesis Technique in CELP

One of the key features of CELP coding is the use of analysis-by-synthesis techniques [3], [20], which exploit the masking property of the human ear to reduce perceived noise. In a direct VQ scheme, the output quantization noise has equal energy at all the frequencies of the original speech, but frequency masking theory has shown that high levels of noise are undetectable by the human ear in the formant regions where speech signal has high energy. Therefore, the error between the original and reproduced speech is passed through a perceptual weighting filter which emphasizes the error in frequency bands where input speech has valleys and de-emphasizes the error in bands

where input speech has peaks. The perceptual weighting filter is generally an auto-regressive (AR) filter derived from the LP synthesis filter by scaling down the magnitude of the poles [1]. The effect of perceptual weighting is to reduce quantization noise in the spectral valleys and increase it near peaks. Thus, the quantization noise is pushed below threshold at all frequencies.

2.5 Conclusions

A review of Waveform Interpolation coding has been presented in this chapter. A survey revealed that the WI coder was initially designed for coding voice at bit rates below 4kb/s (PWI). By using the waveform decomposition technique, the coder was extended to both voiced and unvoiced speech (MPW). A baseline 2.4kb/s WI coder is presented. The major constituent Waveform Interpolation coding procedures, such as LP analysis, waveform extraction and quantization, are described.

One popular speech coding algorithm, the CELP, was also introduced. The basic feature of the CELP algorithm, the analysis-by-synthesis (A-by-S) encoding procedure is described. This will be incorporated into WI coding in a later Chapter.

CHAPTER 3

IMPROVING THE PERFORMANCE OF THE BASELINE CODER

3.1 Introduction

This chapter introduces an improved WI coder working at 2.4kb/s. The basic coding architecture is the same as the baseline coder described in Chapter 2, but the coding procedures in that baseline coder are reinvestigated and improved. The LP analysis operation, LP filter, gain quantization, waveform continuously interpolation and speech reconstruction are found to work well in the baseline coder, and remain unchanged. However, techniques are developed to improve the pitch detection, LSF quantization, SEW/REW decomposition and SEW/REW quantization mechanisms, especially the SEW/REW quantization, which is the main source of the coder distortion. Results show that the quality of the coded speech is improved.

This chapter is organized as follows. Section 3.2 describes an improved LSF quantization method. Section 3.3 presents a new pitch detection algorithm. Section 3.4 presents the SEW/REW decomposition. The SEW/REW quantization mechanisms are discussed in Section 3.5 and Section 3.6. Coder performance is included in Section 3.7. Section 3.8 concludes the chapter.

3.2 LSF Quantization

In the baseline coder, the LSFs are quantized by using mean-squared error (MSE) criteria. Several researchers have found that a weighted MSE criteria, which quantizes the LSF according to their spectral sensitivities, can improve the perceptual performance [26], [38], [44], [45]. The coefficients of the weighting filter are proportional to the values of LPC power spectrum of the given set of LSFs. Thus, the LSFs near spectrum peaks, which are more sensitive spectrally, are better quantized than those near spectrum valleys. The weighted MSE is defined by [44], [45]:

$$E_k = \frac{1}{K}\sum_{k=0}^{K-1} W_k (LSF[k] - \hat{LSF}[k])^2 \qquad \ldots\ldots \qquad (3.1)$$

where E_k is the weighted MSE, *LSF[k]* is the Kth original LSF parameter, $\hat{LSF}[k]$ is its quantization value, and the W_k is the weighting function given by:

$$W_k = [Q(LSF[k])]^r \qquad \ldots\ldots \qquad (3.2)$$

where $Q(.)$ is the LP sum filter (see section 2.3.2.1). The LSF quantization distortion is determined by minimizing E_k.

LSFs cluster near the frequencies of spectrum peaks, and are spaced sparsely near the frequencies of spectrum valleys. Based on this property, an inverse harmonic mean (IHM) weighting function is introduced and used for LSF quantization in this thesis [38], [45]. For a given LSF set, its spectral error sensitivities can be readily estimated from the distances between the adjacent LSFs. The IHM weighting function is then defined as:

$$W_k = \frac{1}{LSF[k+1] - LSF[k]} + \frac{1}{LSF[k] - LSF[k-1]} \quad \cdots\cdots \quad (3.3)$$

(k=1,2⋯8)

$$W_0 = \frac{1}{LSF[1] - LSF[0]} + \frac{1}{LSF[0]}$$

(k=0)

$$W_9 = \frac{1}{4000 - LSF[9]} + \frac{1}{LSF[9] - LSF[8]}$$

(k=9)

The IHM weighting function has a very small computational load and performs close to or sometimes slightly better than the spectral sensitive weighting (eq.(3.2)) [38]. Table 3.1 gives an example of the performance of a LSF quantizer using three different criteria, e.g., no-weighting, IHM weighting and spectral sensitivity weighting [38].

Rate (bits/frame)	**No-weighting (dB)**	**IHM (dB)**	**Spectral Sensitivity (dB)**
20	1.64	1.58	1.57
24	1.31	1.27	1.27
28	1.06	1.03	1.04
32	0.85	0.83	0.84

Table 3.1: Spectral Distortion (SD) of three different LSF quantization schemes.

3.3 Pitch Detection

The Waveform Interpolation coder requires reliable pitch detection. Errors in pitch estimation will cause discontinuities and distortions in the reconstructed Characteristic Waveform surface. For most pitch estimation methods, the reliability can be increased by increasing the analysis window length. However, for a speech signal where the pitch value changes rapidly, an increase in window size may result in an increase in the estimation error.

3.3.1 Pitch Estimation

Several pitch detection algorithms have been proposed, including the autocorrelation method and glottal closure instant method [12], [25], [30], [47]. A modified pitch estimation method based on the autocorrelation method has been found to provide the best performance [30], [47].

This method increases the estimation reliability even when the pitch period is changing. The pitch period is determined by a composite correlation function. First, the estimation window is subdivided into three segments: past, current and future. For each of these segments, the normalized correlation function is computed.

$$R(d) = \sum S(n)S(n-d) \Big/ \sum S(n)S(n) \qquad \ldots\ldots \qquad (3.4)$$

A composite function $R_{composite}$ is then computed as follows [47]:

$$R_{composite}(d) = R_{current}(d) + \max_{i=-f(d)}^{i=f(d)} \{w(i)R_{past}(d+i)\} + \max_{i=-f(d)}^{i=f(d)} \{w(i)R_{future}(d+i)\} \quad \ldots\ldots \quad (3.5)$$

where d is the candidate pitch value, $w(i)$ is the window, and $f(d)$ is the window size.

The windows size $f(d)$ determines the variation in pitch period allowed between segments. For reasons of convenience, A rectangular window is used here. As pitch period usually changes less than 10% between adjacent segments [47], the window size $f(d)$ is chosen to be equal to $d/10$. As the pitch period changes over time, the composition function is the sum of the correlation function of the current segment and the respective maximum correlation values of the past and future segments.

This method only requires a minor computational increase when compared with the ordinary autocorrelation method [47], but provides a more reliable pitch period estimation.

Figure 3.1 gives an example of the pitch estimation. When pitch changes rapidly (the area where the arrows point to), the standard autocorrelation method makes estimation errors, while the proposed method still provides a correct pitch trajectory.

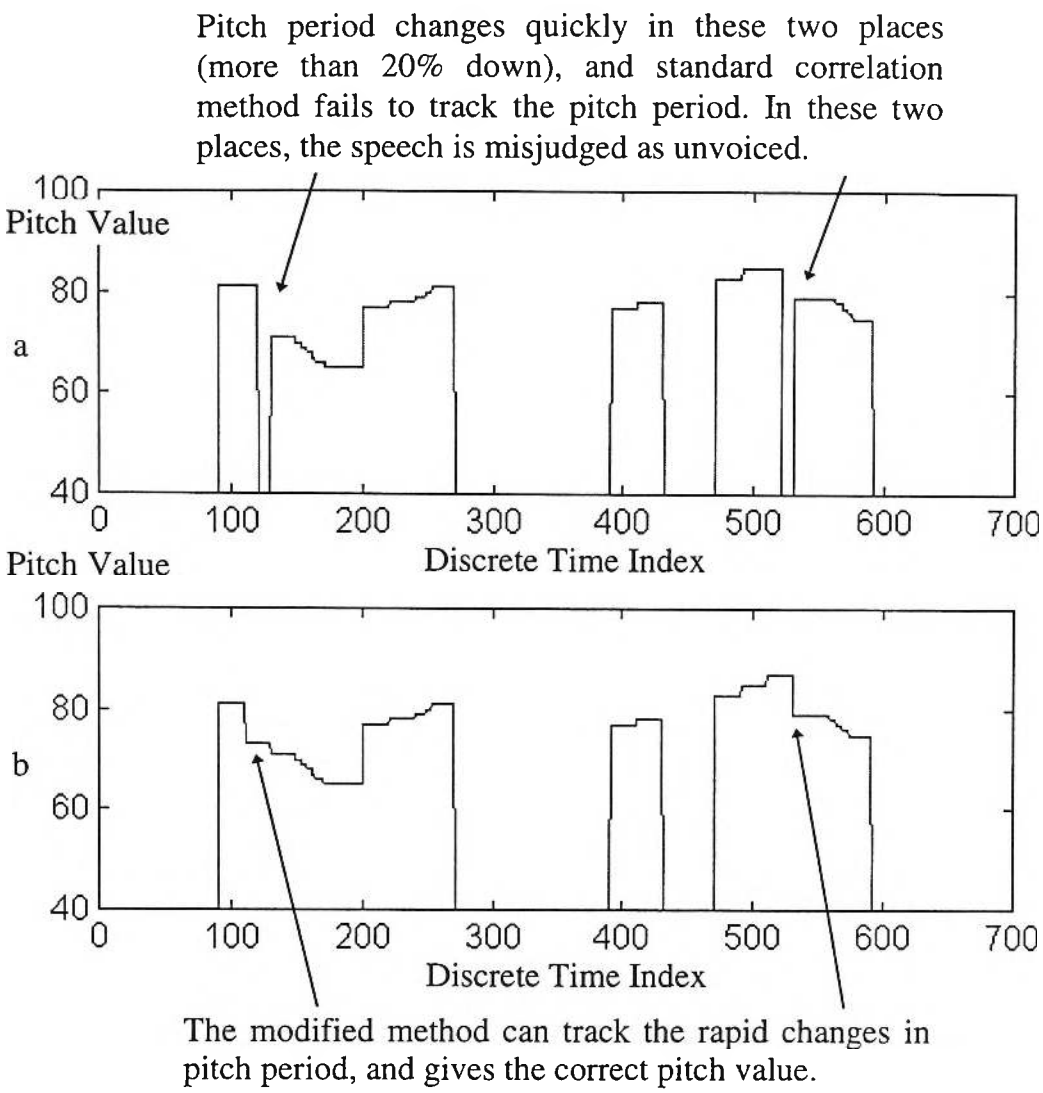


Figure 3.1: (a) Pitch contour obtain by standard autocorrelation method; (b) Pitch contour obtained by the proposed method using composition function. (The pitch is sampled at 400Hz.)

3.3.2 Pitch Multiple Checking

Once the pitch estimate P has been found, a pitch multiple check procedure is performed. A set of the integer sub-multiples of P which are greater than 20, $\{P/2, P/3, \cdots, P/n\}$ is considered. Starting from the largest of the these sub-multiples, every sub-multiple is checked against the thresholds defined in (3.6), (3.7) and (3.8).

$$R(P) > 1.0 \quad \textit{and} \quad \frac{R(P)}{R(\frac{P}{n})} < 2.5 \qquad \ldots\ldots \qquad (3.6)$$

$$R(P) > 0.9 \quad \textit{and} \quad \frac{R(P)}{R(\frac{P}{n})} < 1.5 \qquad \ldots\ldots \qquad (3.7)$$

$$\frac{R(P)}{R(\frac{P}{n})} < 1.35 \qquad \ldots\ldots \qquad (3.8)$$

where $R(P)$ and $R(\frac{P}{n})$ are the correlation values of the pitch and its sub-multiple. If a sub-multiple satisfies the threshold, it will replace the original estimate. The reason for using different thresholds is that the pitch detector is more likely to make a pitch multiple error when the speech is highly periodic [25].

Furthermore, a pitch tracking method is used to improve the pitch estimate. Pitch usually changes slowly, and, thus the pitch estimates of the past frames can help to justify the pitch of the current frame [25]. Let P_{-1} and P_{-2} denote the pitch estimates of the previous two speech frames. If:

$$\left|P_{-1} - P_{-2}\right| < 0.1 \times P_{-2} \quad \text{and} \quad \left|P - P_{-2}\right| > 0.3 \times P_{-2} \qquad \ldots\ldots \qquad (3.9)$$

Then the current pitch estimate P will be replaced by P/n, which:

$$\left|\frac{P}{n} - P_{-2}\right| = \min \qquad \ldots\ldots \qquad (3.10)$$

From Figure 3.2, we see that the pitch multiple checking procedure successfully adjusts the doubled pitches.

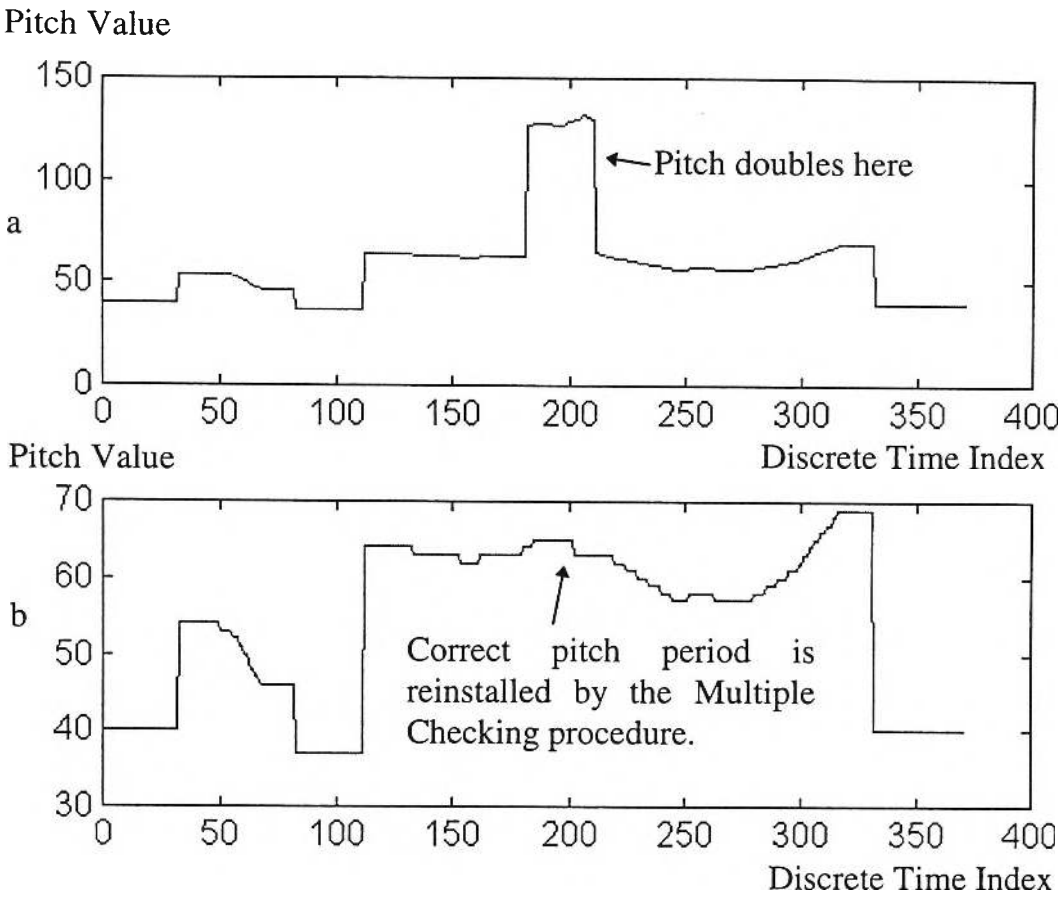


Figure 3.2: (a) Pitch contour before the multiple checking; (b) Pitch contour after the multiple checking.

3.3.3 Pitch Interpolation

During the WI encoding and decoding procedures, the pitch period is interpolated between succeeding frames. As the pitch value may change abruptly, interpolation across these changes will make the waveform extraction procedure fail and cause degradation in the reconstructed speech. So, instead of direct interpolation of the pitch with current and nearby frames, the interpolation is performed between current pitch and $P_{nearby} \times Int(\frac{P_{current}}{P_{nearby}})$, where $P_{current}$ and P_{nearby} are the pitch of the current and

nearby frame respectively [33]. Pitch quantization uses 7-bits. For 8000Hz sampling rate, the pitch value ranges from 20 to 146, corresponding to pitch frequency from 400Hz to 55Hz. A pitch value of 147 represents an unvoiced frame.

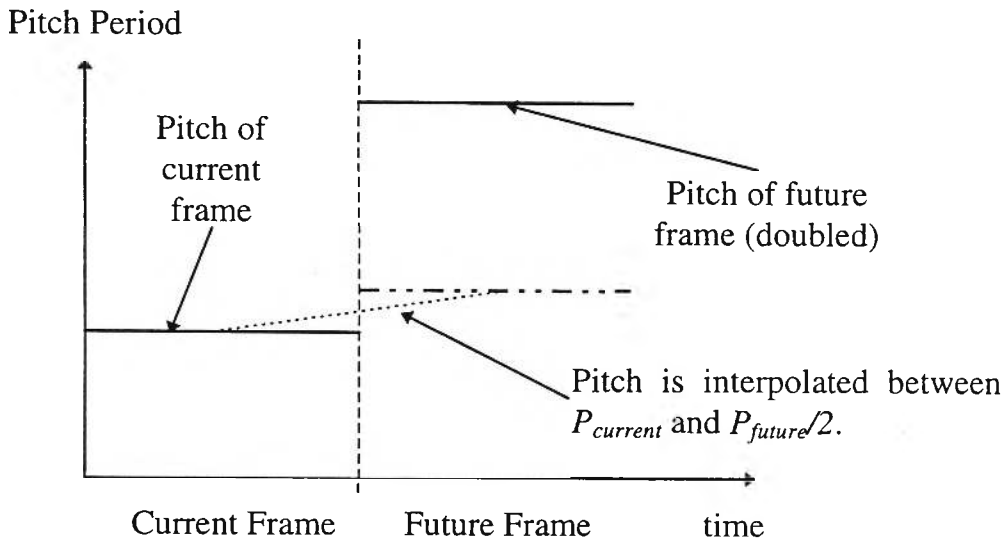


Figure 3.3: An example of the pitch interpolation operation.

3.4 SEW/REW Decomposition

In Chapter 2, the Characteristic Waveforms are roughly decomposed where the SEW is defined as the mean prototype of the analysis frame, and the REW is equal to the incoming prototype minus the SEW. Here, a 21-tap FIR lowpass filter is used to improve the decomposition accuracy. This FIR filter will result in a one frame delay (ten prototypes). Similar to the alignment procedure, the DFT prototypes are padded with zeros or truncated at the end to have the same length before passing into the filter. If the pitch doubles in the successive frames, a procedure similar to that described in Section 2.3.2.2 is performed to force the prototypes fed into the FIR filter to contain the same number of pitch cycle waveforms. For best performance, the

filtering operation is performed on the unnormalized discrete CW surface [33], which emphasizes the waveforms of loud regions.

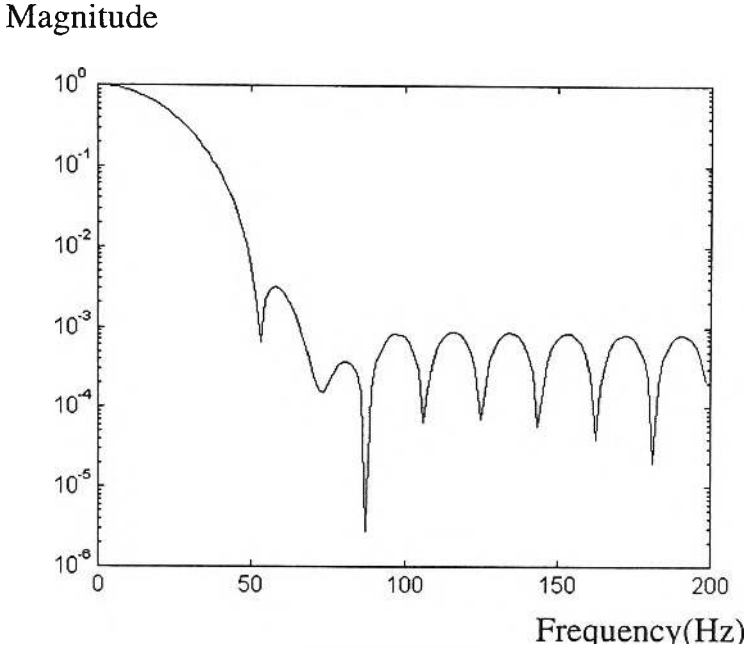


Figure 3.4: Frequency response of lowpass FIR filter (corner frequency is 20 HZ)

The sampling rate of the SEW is 40Hz (one SEW per frame). As the perception of vowels will be affected if the lowpass frequency is lower than 16Hz [33], the cut-off frequency of the FIR lowpass filter is chosen to be 20 Hz. Figure 3.4 gives the frequency response of the filter. Compared with the decomposition method introduced in Chapter 2, use of the FIR filter gives a smoother SEW surface. The FIR lowpass filter offers 8.75dB attenuation in signal amplitude at half of the SEW sampling frequency (20Hz). To increase the attenuation and hence reduce aliasing, the length of the FIR filter needs to be increased. A 41-tap FIR filter with corner frequency of 18Hz gives 14.0dB

amplitude attenuation. However, increasing the filter length will also result in a significant increase in the computational load and the delay of the coder.

3.5 SEW Quantization

SEW quantization is important for the performance of the WI coder. In the SEW quantization mechanism described here, the SEW phase and magnitude spectrum are separated. The magnitude spectrum is quantized by a 7-bit codebook and transmitted, while the phase spectrum is not transmitted, it is derived from the transmitted pitch information [34].

3.5.1 SEW Phase Quantization

For unvoiced speech (classified as a quantized pitch value of 147), the phase spectrum of the SEW is a uniformly distributed random signal, representing a spread-out waveform. While for voiced speech (pitch value 20~146), the SEW phase spectrum is a typical pulse phase spectrum that is extracted from real speech (see Figure 3.5) [34].

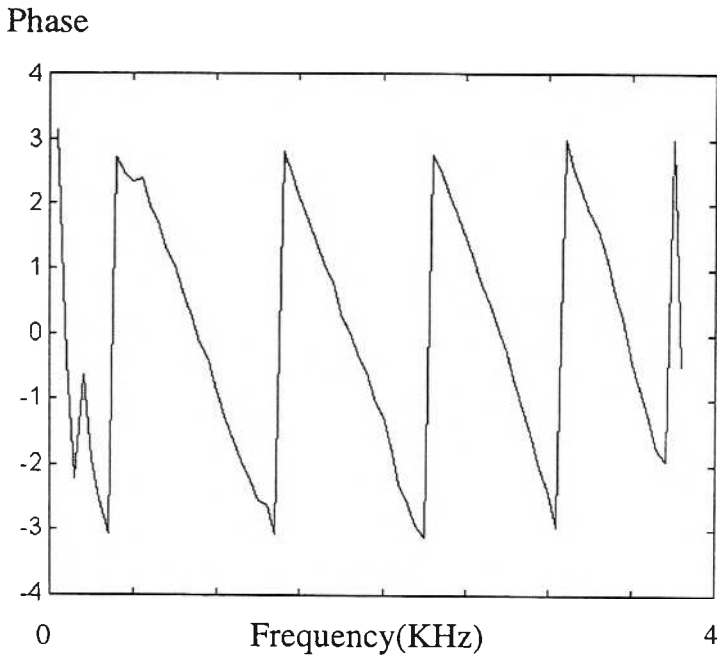


Figure 3.5: Typical pulse phase spectrum

Two methods used to make the voiced/unvoiced decision are considered. One is based on the normalized correlation function *R(p)*.

$R(p) >= 0.5$	The speech is judged as voiced.
$R(p) < 0.5$	The speech is judged as unvoiced.

The other method is based on the shape of the extracted prototypes in the time domain. If the prototype is flat, it is judged to be from a voiced segment. If the prototype contains a pulse, it is judged as unvoiced. First, the average gain of the prototype $\overline{A}$ is calculated:

$$\overline{A} = \frac{1}{N}\sum_{t=0}^{N-1}\left|U(t_i,t)\right| \qquad \ldots\ldots \qquad (3.11)$$

where N is the prototype length, and $U(t_i,t)$ is the time-domain prototype at time interval t_i.

Then the biggest absolute value of the time-domain prototype samples A_{max} is found:

$$A_{max} = \max\{|U(t_i,0)|,|U(t_i,1)|,\cdots,|U(t_i,N)|\} \quad \ldots\ldots \quad (3.12)$$

Finally, the voiced/unvoiced decision is made according to:

$A_{max} > 3.65 \times \overline{A}$ The prototype is judged as voiced.

$A_{max} < 3.65 \times \overline{A}$ The prototype is judged as unvoiced.

The later method which is based on the time domain prototype shape makes a better voiced/unvoiced decision during the informal listening test. The tonal effects in the output speech are reduced.

3.5.2 SEW Magnitude Quantization

For the SEW magnitude quantization, the SEW magnitude above 800 Hz, which is less important in terms of perception, is inferred from the REW magnitude. As the LP residual signal has a flat power spectrum, the magnitude spectrum of the SEW can approximated by [5], [34]:

$$|SEW(f)| = 1 - |REW(f)| \qquad f > 800\text{Hz} \quad \ldots\ldots \quad (3.13)$$

For the SEW magnitude below 800Hz (which is more important perceptually), a 7-bit eight dimensional codebook describes the spectral behaviour. Each dimension represents a frequency bin, covering a 100Hz spectral region. During the SEW codebook search, the candidate SEW is derived from the codebook by:

$$|SEW_{cand}(k)| = |SEW_{codebook}(n)|$$
$$n = \text{int}(80 * k \,/\, pitch) \qquad \ldots\ldots \qquad (3.14)$$

The original SEW and the SEW candidature are converted to the speech domain through the LP filter:

$$|SEW'(k)| = \frac{|SEW(k)|}{|A(k)|}$$
$$|SEW_{cand}'| = \frac{|SEW_{cand}(k)|}{|A(k)|} \qquad \ldots\ldots \qquad (3.15)$$

where $A(k)$ is the LP analysis filter, $|SEW'(k)|$ and $|SEW_{cand}'|$ are the magnitude spectrum of the speech domain SEW and SEW candidature. The SEW codebook selection is performed in the speech domain by using the mean squared error (MSE) criteria.

Figure 3.6 shows an example of the original and quantized SEW magnitude. It can be seen that below 800Hz, the SEW magnitude is accurately quantized by the 7-bit SEW codebook. Above 800Hz, the SEW magnitude is derived from the REW and is only roughly quantized.

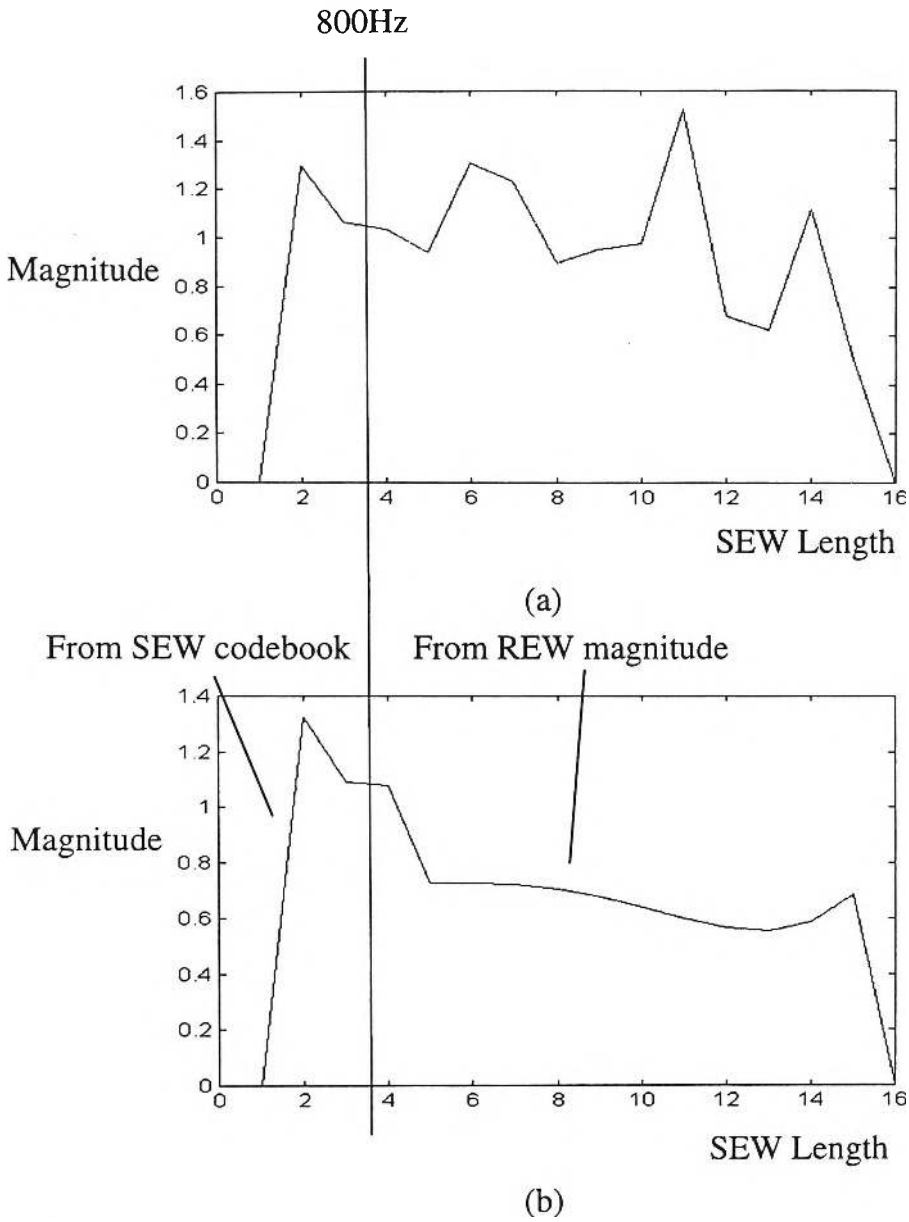


Figure 3.6: (a) Original SEW magnitude; (b) Decoded SEW magnitude. (The length of this example SEW is 16.)

3.6 REW Quantization

3.6.1 REW Phase Quantization

In Chapter 2, the REW phase spectrum is approximated by a uniform distributed Gaussian random spectrum. Another REW phase representation method is tested here.

It has been found that for unvoiced speech, the residual signal can be replaced by white noise with the power contour and the spectral power envelope preserved [33]. Therefore, a random white noise is generated and transformed to the DFT domain. The REW is then reconstructed by weighting the white noise with the transmitted REW magnitude in the DFT domain. This method gives good reconstructed speech quality but is computationally too complex.

3.6.2 REW Magnitude Quantization

Owing to the complexity problem, a polynomial representation of the REW magnitude is proposed. The Chebyshev polynomials are historically the oldest of various sets of orthogonal polynomials [48]. Five shifted Chebyshev polynomials represent the REW magnitude spectrum. The first five shifted Chebyshev polynomials are defined as:

$$
\begin{aligned}
&T_0(x) = 1 \\
&T_1(x) = 2x - 1 \\
&T_2(x) = 8x^2 - 8x + 1 \qquad\qquad 0<x<1 \qquad\qquad \ldots\ldots \qquad (3.16) \\
&T_3(x) = 32x^3 - 48x^2 + 18x - 1 \\
&T_4(x) = 128x^4 - 256x^3 + 160x^2 - 32x + 1
\end{aligned}
$$

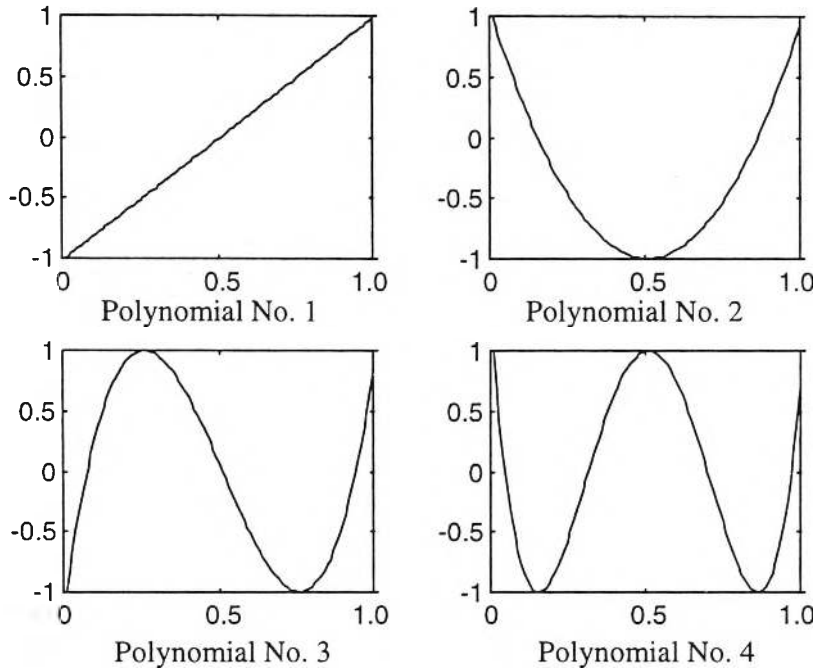


Figure 3.7: Shapes of shifted Chebyshev polynomials

The REW magnitude spectrum can be described by Chebyshev polynomial expansions:

$$REW(k) = \sum_{n=0}^{4} a_n T_n(\frac{k}{K}) \qquad \ldots\ldots \qquad (3.17)$$

$$a_0 = \frac{1}{K}\sum_{k=0}^{K-1} REW(k)$$

$$a_n = \frac{2}{K \cdot \pi}\sum_{k=0}^{K-1} \frac{REW(k) \cdot T_n(\frac{k}{K})}{\sqrt{\frac{k}{K}(1-\frac{k}{K})}} \qquad \text{n=1,2,,4}$$

where K is the prototype length, a_n are the coefficients of the polynomial expansions. The REW magnitude is quantized using a 3-bit vector codebook of sets of polynomial coefficients. Let a_n^* represent a set of polynomial coefficients in the REW codebook, the error criterion is then:

$$E = \sum_{k=0}^{K-1}(REW(k) - \sum_{n=0}^{4} a_n^* T_n(\frac{k}{K}))^2 \qquad \ldots\ldots \qquad (3.18)$$

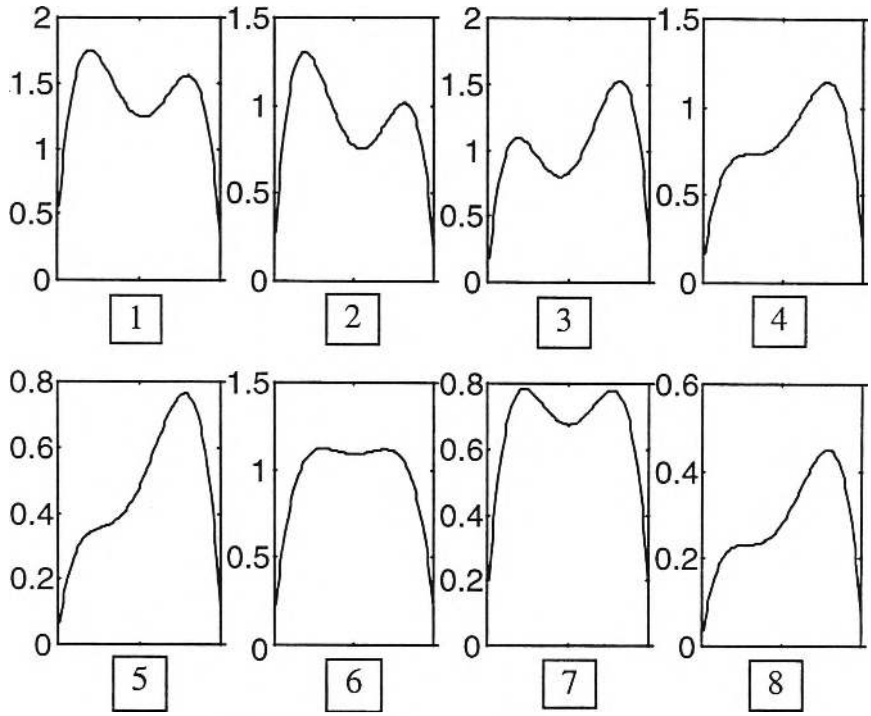


Figure 3.8: Eight shapes in the REW magnitude codebook

Figure 3.8 shows the shapes in the REW codebook. Shapes 1 and 2 can represent the REW which most of the signal power is in the low frequency region, while shapes 3, 4, 5 and 8 represent REWs with more energy in the high frequency region. Shapes 6 and 7 represent REWs with flat magnitude spectrums. These eight shapes cover almost all kinds of REW magnitude spectrum. As the REW is only represented by low order Chebyshev polynomials, there are peaks in the REW codebook shapes. Since the real REW has a relatively flat magnitude, these peaks in the REW magnitude spectrum may introduce tonal effects in the output speech which are undesirable. At low rates, these peaks in the REW spectrum have little effect on the output speech quality, however, it is worthwhile considering high-order polynomials which give more accurate representation for higher bit rate transmission.

3.7 Coder Performance

The performance of the new 2.4kb/s WI coder was tested using a two-step testing procedure. In the initial step, a WI coder which uses unquantized parameters was tested. This coder incorporated the new pitch detection and SEW/REW decomposition mechanisms introduced in this chapter. Figure 3.9 (b) shows one frame of the reproduced speech of the "unquantized" coder. The reconstructed speech approached "transparent" speech quality. The coded speech scored 3.71 in MOS tests. This result indicates that, by using the new pitch detection and SEW/REW decomposition algorithms, the speech waveform are successfully extracted, decomposed and reconstructed.

In the second step, the parameter quantization is examined. A new 2.4kb/s fully-quantized WI coder is tested. The pitch, gain, LSFs, SEW and REW quantization are included. Figure 3.9 (c) show a segment of the coded speech. The speech achieves perceptually good quality. The LSF, SEW and REW quantization mechanisms which are introduced in this chapter were proved to be more superior then those used in the baseline coder. The SEW/REW quantization was found to be the key element to the coder performance. To obtain transparent speech quality, the SEW and REW have to be well quantized.

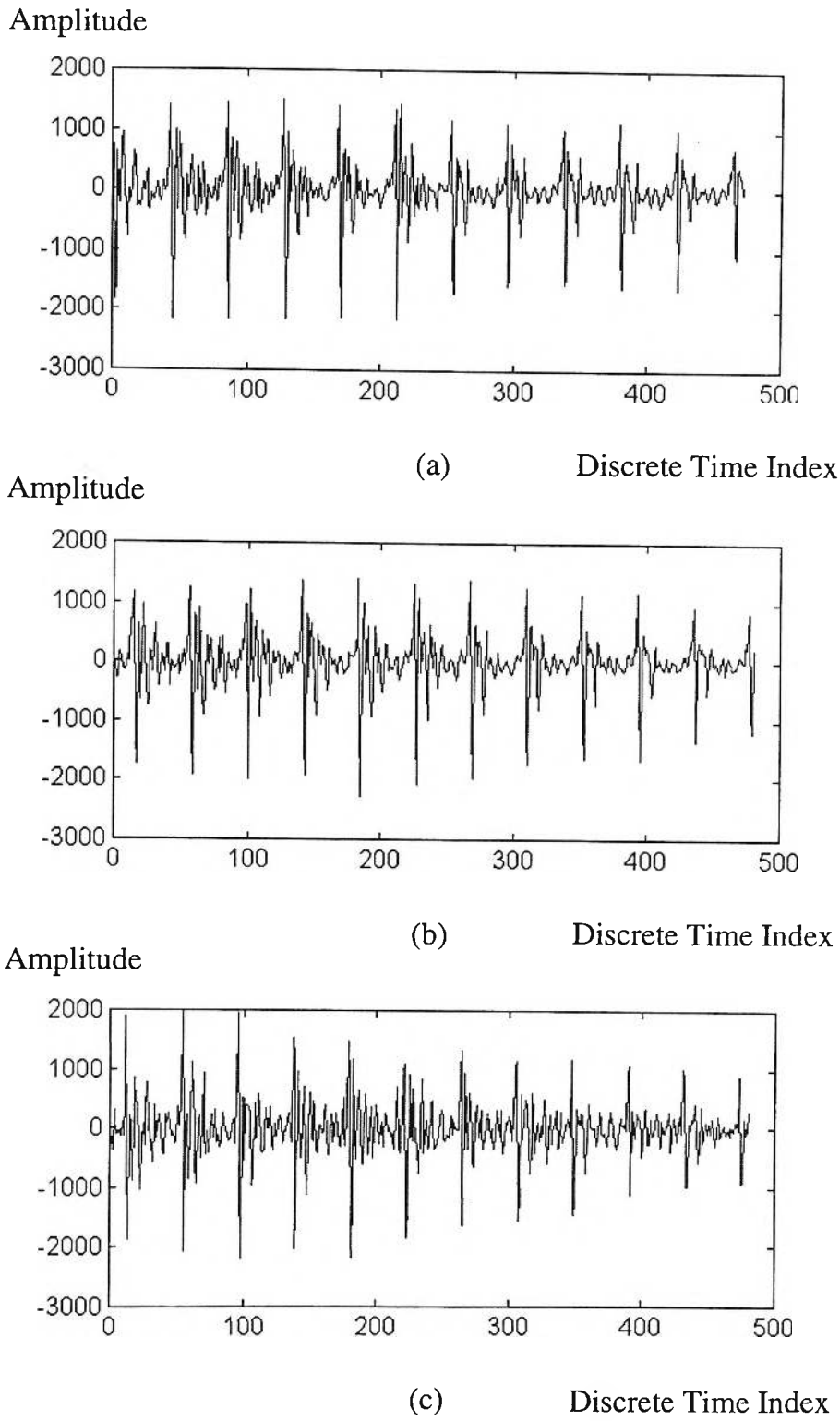


Figure 3.9: (a) One frame of original speech; (b) The reproduced speech of the WI coder with all the parameters unquantized; (c) The reproduced speech of the improved 2.4kb/s WI coder.

Informal listening tests were conducted to evaluate the performance of the new 2.4kb/s fully quantized WI coder. It was found that the new coder performs better than the baseline coder introduced in Chapter 2. Among the 16 listeners, 87.5% (14 listeners) preferred the speech quality of the new coder, while only 12.5% (2 listeners) preferred the baseline coder. The output speech of the new coder was judged sound clearer, more natural and less noisy.

3.8 Conclusions

This chapter introduces some techniques to improve the performance of the WI coder. The pitch detection, LSF quantization, SEW/REW decomposition and SEW/REW quantization mechanisms are improved. Results show that this WI coder reproduces almost transparent speech using unquantized parameters. The fully-quantized 2.4kb/s WI coder works well in terms of perceptual quality.

It was also found that the SEW/REW decomposition and quantization are essential to the speech reconstruction quality. In the next Chapter, the analysis-by-synthesis mechanism is considered for the SEW/REW quantization.

CHAPTER 4

WAVEFORM INTERPOLATION AND ANALYSIS-BY-SYNTHESIS

4.1 Introduction

Analysis-by-synthesis (A-by-S) is one of the key features to the success of the CELP class speech coder. The analysis-by-synthesis mechanism integrates the decoder (synthesis) into the encoder (analysis) loop. The coder parameters are found by minimizing the mean squared error (MSE) between the original and synthesized speech signal. This error signal is perceptually weighted by a filter *W(z)*. Figure 4.1 shows the diagram of the analysis-by-synthesis technique.

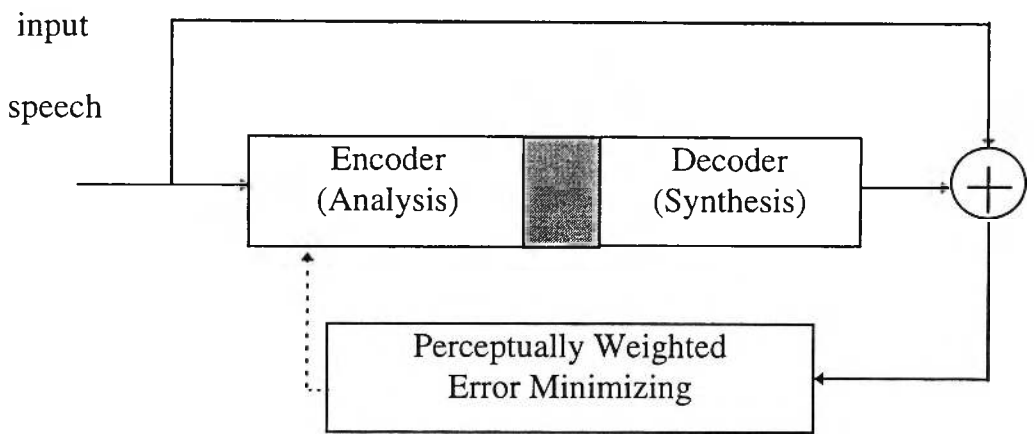


Figure 4.1: Analysis-by-synthesis mechanism principle

The perceptual weighting filter *W(z)* increases the noise in the formant regions and reduces it in between formant regions. *W(z)* is given by [3]:

$$W(z) = \frac{1 + \sum_{k=1}^{p} a_k z^{-k}}{1 + \sum_{k=1}^{p} \alpha^k a_k z^{-k}} \qquad \ldots\ldots \qquad (4.1)$$

where a_k is the coefficient of the pth order LP filter. α controls the increase in the noise power in the formant regions. For a sampling rate of 8000Hz, α is typically chosen to be 0.8 .

Waveform Interpolation coders have been found to be successful at low bit rates [33], [34]. However, Waveform Interpolation coders do not incorporate the analysis-by-synthesis mechanism. Instead, Waveform Interpolation uses open-loop quantization of Characteristic Waveform parameters. A closed-loop WI coder which uses an altered analysis-by-synthesis mechanism is proposed here. This technique operates on a prototype-by-prototype basis, optimizing a codebook search within each frame.

This chapter is organized as follows. Section 4.2 discusses how to adapt analysis-by-synthesis mechanisms to the WI coder. Section 4.3 presents approaches to analysis-by-synthesis in closed-loop WI coding. Section 4.4 discusses the incorporation of the perceptual weighting filter in analysis-by-synthesis architecture. Section 4.5 presents the results. Finally, Section 4.6 concludes this chapter.

4.2 Adapting A-by-S to WI

The fundamental problem when considering incorporation of the analysis-by-synthesis technique in Waveform Interpolation is that the reproduced speech of a WI coder is generally not synchronous with the original speech. As a result, the mismatches in time alignment of the original and reproduced speech will introduce a significant

increase in error signal energy which is perceptually irrelevant. This prevents the immediate adoption of A-by-S techniques in WI coding.

To overcome this weakness, a generalized analysis-by-synthesis paradigm is proposed [47]. The concept of this new paradigm is shown in Figure 4.2. The original speech is modified so that it optimally matches the speech produced by the decoded speech. The error minimizing procedure is based on the modified input speech and the speech produced by the decoder.

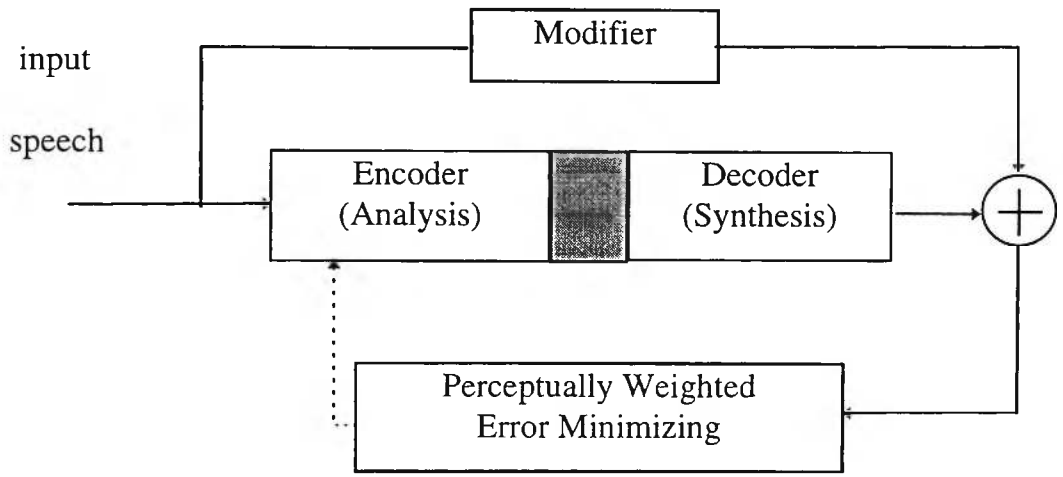


Figure 4.2: Generalized analysis-by-synthesis paradigm

Closed-loop Waveform Interpolation is an example of the implementation of the generalized analysis-by-synthesis technique. Instead of direct sample-by-sample comparison of the input and output speech signal, a set of unquantized prototypes (Characteristic Waveforms) are used to represent the modified input speech. This series of prototypes is compared with the synthesized prototypes and the speech encoded by minimizing the perceptually weighted error between the original and synthesized prototypes. Closed-loop WI coders operate on a prototype-by-prototype

basis. If each prototype is accurately quantized, an accurate representation of the input speech will be achieved [7].

4.3 Approaches to A-by-S in WI

A series of restrictions are placed on the incorporation of analysis-by-synthesis mechanisms in Waveform Interpolation coding by the low rate parameter transmission. In WI coders, the prototypes are generally described by a Fourier-series, and at low bit rates, the phase information of the prototype is discarded. Thus, both magnitude and magnitude/phase closed-loop searching is investigated in this Chapter.

In Waveform Interpolation coding, the prototype (Characteristic Waveform) is decomposed into the slowly-evolving-waveform (SEW) and rapidly-evolving-waveform (REW) components. A direct prototype analysis-by-synthesis search can be achieved by joint optimization of the SEW and REW vectors. However, this one-stage search requires high computation. For a 7-bit SEW and 3-bit REW codebook, the one-stage search needs 128*8 times of SEW/REW selection operation. Instead, a two-stage sub-optimum search is used to reduce the computational load. The SEW and REW vectors are then selected sequentially and each codebook search attempts to find the vector which minimizes the quantization error. The two-stage search needs only 128 times of SEW plus 8 times of REW selection operation.

In this thesis, the SEW magnitude below 800Hz is quantized, while the magnitude response of the SEW above 800Hz is approximated by $1-|REW|$. For each of the ten

prototypes in a frame, the mean squared error between a candidate SEW vector and the prototype is computed. This operation is performed in the speech domain. The error computation is performed as:

$$E = \sum_{k} \left| \frac{|SEW_{cnad}(k)|}{|A(k)|} - \frac{|U(k)|}{|A(k)|} \right| \qquad \ldots\ldots \qquad (4.2)$$

$k=1,2,\ldots, (K_m/10)$

where K_m is the interpolated pitch value (prototype length), $SEW_{cand}(k)$ is the candidate SEW vector, $U(k)$ is the incoming prototype and $A(k)$ is the LP synthesis filter.

For the analysis-by-synthesis search of the REW vector, the correct level of REW must be established. As the REW represents the noise component of speech, the REW can simply be computed as the extracted prototypes following removal of the mean of the ten prototypes of that frame.

A more accurate REW search is described below. First, the SEW vector is selected as described above. Then, the REW vector search is performed upon adjusted incoming prototypes, with the quantized SEW contribution subtracted. To complete this subtraction, the SEW phase information is needed. The SEW phase spectrum can be considered to be identical to the incoming prototype or to be the fixed SEW phase used at the decoder. In the latter case, the SEW and the incoming prototype should be time-aligned before the subtraction operation. These two methods offer similar performance, but the latter one which requires the alignment procedure is

computationally more complex. The REW search is then performed by computing aggregate mean squared errors between the adjusted prototypes and the REW in the speech domain.

4.4 Perceptual Weighting Filter

The incorporation of the analysis-by-synthesis architecture in Waveform Interpolation coding allows for the exploitation of perceptual weighting techniques. The search process is identical to that discussed in Section 4.3, apart from the addition of the perceptual weighting filter. In the SEW vector search, the weighted mean squared error is computed as:

$$E_W = \sum_k \left| \frac{|SEW_{cnad}(k)|}{|A(k)|} - \frac{|U(k)|}{|A(k)|} \right| W(k)$$
$$= \sum_k \left| \frac{|SEW_{cnad}(k)|}{|A(k)|} - \frac{|U(k)|}{|A(k)|} \right| \frac{A(k)}{A(k/\alpha)} \qquad \ldots\ldots \qquad (4.3)$$

$k=1,2,\ldots, (K_m/10)$

where $W(k)$ is the perceptual weighting filter. To reduce the computational load, the perceptual weighting filter is moved into the synthesis procedure:

$$E_W = \sum_k \left| \frac{|SEW_{cnad}(k)|}{|A(k)|} - \frac{|U(k)|}{|A(k)|} \right| \frac{A(k)}{A(k/\alpha)}$$
$$= \sum_k \left| \frac{|SEW_{cnad}(k)|}{|A(k/\alpha)|} - \frac{|U(k)|}{|A(k/\alpha)|} \right| \qquad \ldots\ldots \qquad (4.4)$$

$k=1,2,\ldots, (K_m/10)$

This method for complexity reduction is similar to that used in the CELP algorithm [16], [55].

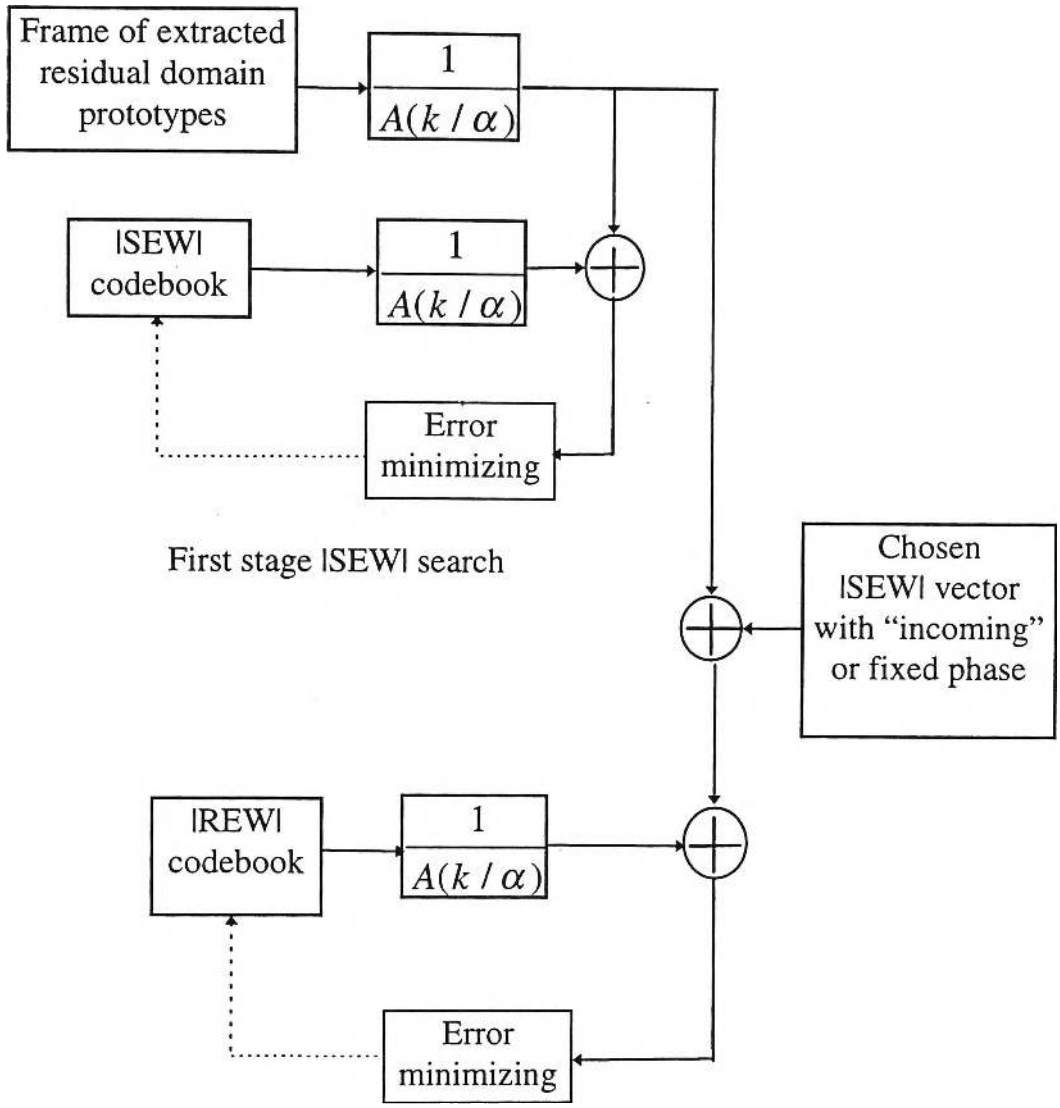


Figure 4.3: closed-loop SEW/REW search mechanism

The closed-loop REW search can be modified in a similar way to incorporate the perceptual weighting process:

$$E_W = \sum_k \left| \frac{|REW_{cnad}(k)|}{|A(k)|} - \frac{|U^*(k)|}{|A(k)|} \right| \frac{A(k)}{A(k/\alpha)}$$
$$= \sum_k \left| \frac{|REW_{cnad}(k)|}{|A(k/\alpha)|} - \frac{|U^*(k)|}{|A(k/\alpha)|} \right| \quad \ldots\ldots \quad (4.5)$$

$k=1,2,\ldots, (K_m/10)$

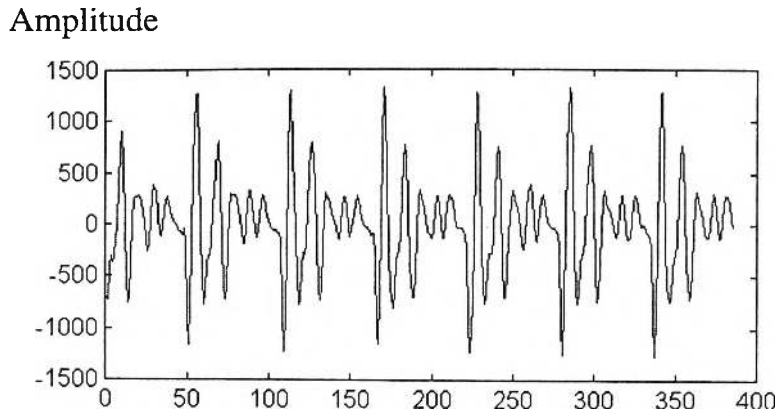
where $U^*(k)$ is the incoming prototype adjusted by the chosen SEW vector.

The complete analysis-by-synthesis search process of closed-loop WI coding is shown in Figure 4.3.

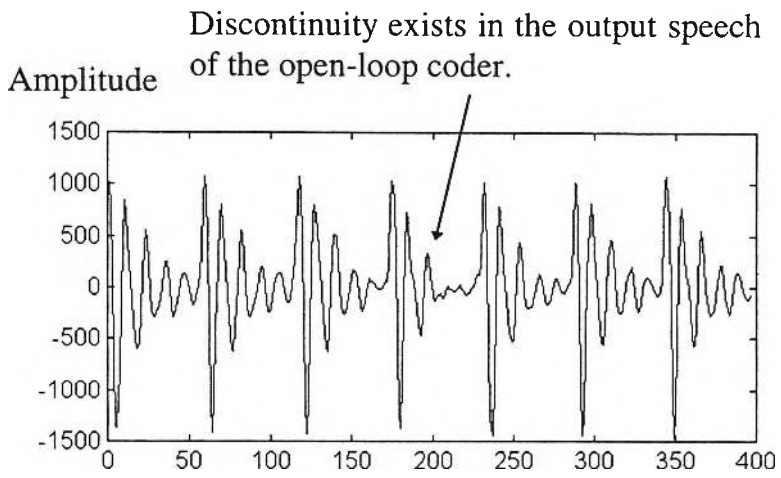
4.5 Results

The performance of the closed-loop WI coder which uses analysis-by-synthesis techniques and ordinary open-loop WI coder is examined. Informal listening tests show that analysis-by-synthesis WI coders achieve equivalent speech quality to the standard open-loop WI coder. However, the closed-loop WI coder which uses perceptually weighted analysis-by-synthesis techniques was preferred by a significant majority of listeners. Compared with the open-loop WI coder, it produces clearer and smoother speech with an appropriate SEW/REW level being established.

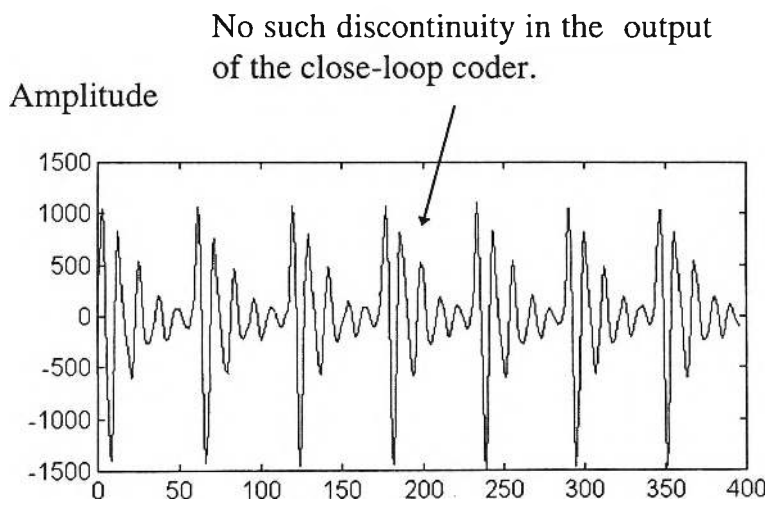
Among the 16 listeners, 75% (12 listeners) favored the closed-loop coder using perceptual weighting A-by-S technique, 12.5% (2 listeners) gave no preference, and 12.5% (2 listeners) preferred the open-loop coder.



(a) Discrete Time Index



(b) Discrete Time Index



(c) Discrete Time Index

Figure 4.4: (a) Input speech signal; (b) The reconstructed speech by the open-loop WI coder; (c) The reconstructed speech by the closed-loop WI coder.

Figure 4.4 gives an example of the open-loop coded (b) and closed-loop coded (c) speech signal. It can be seen that the speech of the closed-loop coder evolves smoothly, while the speech generated by the open-loop coder has a certain degree of discontinuity in some parts of the waveform. (notice the area where the arrows point to).

The closed-loop WI coder also surpasses the open-loop coder in terms of delay and complexity. In the open-loop WI coder, the REW/SEW are decomposed by highpass/lowpass filtering. The SEW/REW filtering is a complex operation (for a pitch period of 40, the SEW/REW filtering needs more than 16,000 multiply/adds per frame) and generates at least one frame of delay. However, in the analysis-by-synthesis WI coder, the SEW and REW search is performed directly upon the incoming prototype, the highpass/lowpass filtering decomposition procedure is eliminated, resulting in a simpler encoding architecture.

4.6 Conclusions

This chapter presents an altered analysis-by-synthesis mechanism which overcomes the non-synchronous nature of the input/output speech of WI coding. The proposed architecture operates on a prototype-by-prototype basis. A two-stage sub-optimum SEW/REW vector search is used. The CELP style perceptual weighting techniques are exploited in both the SEW and REW search. In conclusion, the results indicate that

the incorporation of perceptually weighted analysis-by-synthesis mechanisms into Waveform Interpolation improves the coder performance.

CHAPTER 5

WAVEFORM INTERPOLATION AT BIT RATES ABOVE 2.4 KBITS/S AND LOW COMPLEXITY WI CODER

5.1 Introduction

One of the distinguishing advantages of WI coders over other low rate algorithms is that they offer scalability to higher rates [7]. Waveform Interpolation coders encode input speech on a prototype (Characteristic Waveform) basis. The information in the prototypes is quantized and transmitted with the WI decoder reconstructing speech by interpolation of the received prototypes. By increasing the update rate and/or quantization accuracy of the speech prototypes, scalability to higher bit rates can be achieved. This chapter utilises this fact to produce WI coders at bit rates between 2.4kbits/s and 3.6 kbits/s.

It is known that WI coders can reproduce transparent speech given that all the parameters are unquantized (see Chapter 3 and [7]). This suggests the possibility of improving the performance of the WI coder at higher bit rates, where parameters are quantized more accurately than at 2.4kb/s. This chapter tests the performance of both open and closed-loop A-by-S WI coding mechanisms at higher bit rates. Using the 2.4kb/s coders described in Chapter 3 (open-loop) and Chapter 4 (closed-loop) as a basis, the improvement in speech quality attained by allocating further bits to each individual coder parameter is investigated. Efficient allocation of bits among the different quantized parameters can thus be achieved at a variety of higher rates.

Although WI coders can provide high quality speech, the primary disadvantage is the high computational load associated with the waveform extraction and quantization. Techniques have been developed to reduce the coder complexity with no or very little

degradation in the perceptual quality of the reconstructed speech. Such techniques are described in this Chapter.

This Chapter is organized as follows. Section 5.2 presents the effect of higher bit rates for each of the parameters. Section 5.3 gives the bit allocation of coders operating between 2.4kb/s and 3.6kb/s and examines the coders performance. Section 5.4 discusses the motivation for low-complexity WI coding. Section 5.5 presents the low-complexity SEW/REW decomposition, analysis and quantization. Section 5.6 presents the low-complexity Waveform Interpolation coding architecture. Finally, Section 5.7 concludes this chapter.

5.2 The Effect of Higher Bit Rates for Each Parameter

Firstly, the effect of higher bit rates for each individual parameter used in WI coding is examined. The Waveform Interpolation algorithm codes speech using the LSFs, pitch, gain, SEW and REW parameters. Given extra bits for each of these parameters, either the size of the codebook, or the update rate or both can be increased. Each of these possibilities and the consequences of the choice in perceptual terms is considered. As the SEW and REW are quantized by significantly different mechanisms in the open-loop and closed-loop Waveform Interpolation coders, the performance of the two coders for varying SEW and REW update and coding rates might be expected to differ substantially. Hence, the SEW and REW quantization at high bit rates are investigated separately for open-loop and closed-loop WI coders.

5.2.1 LSF and Pitch

30-bit Split-VQ LSF transmission, which is used in the 2.4kb/s WI coder, results in <1dB distortion and is generally considered to be transparent [45]. A more accurate representation will not introduce significant perceptual improvement. Furthermore, the codebook size can be reduced to 26-bits by using multi-stage LSF codebook quantization [44], [45]. The update rate of once per 25ms frame is adequate and while an increase improves perceptual quality, the significant bit-rate increase is unjustified.

For the pitch, 7-bit integer representation of the pitch value(20~147 for 8000Hz sampling rate) is adequate. This is particularly the case in WI where minor variations between input pitch and integer, and quantized pitch will be substantially catered for by the continuous interpolation techniques used during synthesis. The transmission rate of one pitch per frame is thus adequate.

5.2.2 Gain

Increased resolution in the gain codebook can give significant improvements in perceptual quality. At 2.4kb/s a 4-bit differential gain codebook fails to adequately track rapid changes in input speech energy and, overall, output synthesized speech suffers some loss in gain resolution. When using a 5-bit codebook, however, this loss of resolution is substantially removed, resulting in clearer speech. A 6-bit gain codebook was tested and found to offer similar performance, indicating that further

increases in gain codebook size were unnecessary. As gain is coded using differential quantization techniques (incorporating a step capability to track rapid speech energy changes), which means the gain is lowpass filtered, an update rate of two gain indexes per frame is adequate.

5.2.3 SEW

In the SEW quantization, an eight-dimension codebook describes the SEW magnitude spectrum below 800Hz. Increasing the size of the SEW codebook from 7-bits to 9-bits gives marginal improvements in speech quality and spectrum behavior. During informal listening tests the speech was reported as sounding smoother and more natural. Improvements for closed-loop WI coders are less significant and this can be explained by the improved selection mechanism resulting from a closed loop technique. Further, in a closed loop system the complexity penalties of using larger SEW codebooks do not appear to warrant the perceptual improvement. A 10-bit SEW codebook was also tested for an open-loop coder, results indicate similar performance to a 9-bit codebook. While further increasing the codebook size for quantizing the SEW below 800Hz gives little improvement, a 16-dimension SEW codebook which covers the SEW spectrum below 1600Hz was considered a possibility, 9-bit and 10-bit SEW codebooks (16-dimension) were tested, but were found to offer no significant improvement. This can be explained by the fact that the frequency resolution of the human ear decreases rapidly with increasing frequency [49].

In open-loop WI coding, the SEW is obtained by lowpass filtering the CW surface by a FIR filter with a corner frequency of 20Hz. According to sampling theory, the update rate of one SEW per frame (40Hz) is adequate for open-loop coding (assuming the filter is ideal). In closed-loop WI coding, increased update rates e.g. of 2 SEWs per frame offers minor improvements. This is in accordance with the concept of the SEW as the slowly-evolving, underlying waveform component.

5.2.4 REW

A 3-bit REW codebook of sets of the first five Chebyshev polynomial coefficients is used in the 2.4kb/s coder. An increase in the REW codebook size from 3-bits to 5-bits will give better quality for both the open-loop and closed-loop coders. The speech sounds clearer, especially in terms of high frequency content (this is particularly noticeable in fricatives). The reasoning behind this is complicated by the interaction between the REW and SEW magnitudes. A 5-bit REW codebook will, clearly, incorporate a wider variety of REW shapes, however, as the high frequency part of the SEW is also derived from the REW, the SEW will also be better represented. Further increases in REW codebook size give no significant improvement. There is little perceptual difference between a REW codebook with a set of five Chebyshev polynomials and a REW codebook with seven polynomials for the same codebook size.

In both types of coders, the required update rate for the REW was found to be at least 4 times per frame (corresponding to time resolution of 6.25 ms). This is in accordance

with the fact that the power contour and the spectral-power envelope of unvoiced speech should be preserved with a time resolution of about 5 ms [33]. Reduction of this update rate introduces a harsh, mechanical feel to the reproduced speech. Beyond six updates per frame, little improvement in perceptual quality was noted. In particular no clear preference was shown between speech encoded with ten REWs per frame and that using just five.

Parameter	LSF	pitch	gain	SEW	REW
bits/frame (2.72kb/s)	30	7	5*2=10	9	3*4=12
bits/frame (3.24kb/s)	30	7	5*2=10	9	5*5=25
bits/frame (3.60kb/s)	30	7	5*2=10	9*2=18	5*5=25

Table 5.1: Bit allocation per frame for different bit rates.

5.3 Configuration and Coder Performance

Based on the results of Section 5.2, bit allocations for three bit rates were established. The bit allocations of these coders are shown in Table 5.1. At 2.72kb/s, priority is given to the codebook size of the gain and the SEW (the latter primarily when using open-loop encoding). The sizes of the gain and SEW codebooks are increased to 5-bit and 9-bit respectively. At 3.24 kb/s, the extra bits were given to the REW quantization. Both the codebook size and updating rate of REW were adjusted to 5-bits and 5 updates per frame. As transmitting two SEWs per frame also gives minor

improvements, in a 3.6kb/s coder, the extra bits were used to transmit two 9-bit SEWs for each frame.

Informal listening tests and spectrogram comparisons found that for both the open-loop WI and closed-loop WI coder, clear improvements are apparent between the 2.4kb/s coder and 2.72kb/s, and subsequently between the 2.72kb/s and 3.24kb/s coders. However, the perceptual quality of the 3.6kb/s coder and 3.24kb/s coder is very similar (see Table 5.2). At 3.6kb/s WI approaches Toll quality, however formal MOS testing will be required to substantiate initial results.

Bit Rate Increasing	**Percentage of Listener Acknowledging Quality Improvement**	**Percentage of Listener Acknowledging No Quality Improvement**
From 2.4kb/s to 2.72kb/s	50%	50%
From 2.72kb/s to 3.24kb/s	87.5%	12.5%
From 3.24kb/s to 3.6kb/s	43.25%	56.75%

Table 5.2: Result of informal listening test (16 listeners) of WI coders at bit rates above 2.4kb/s.

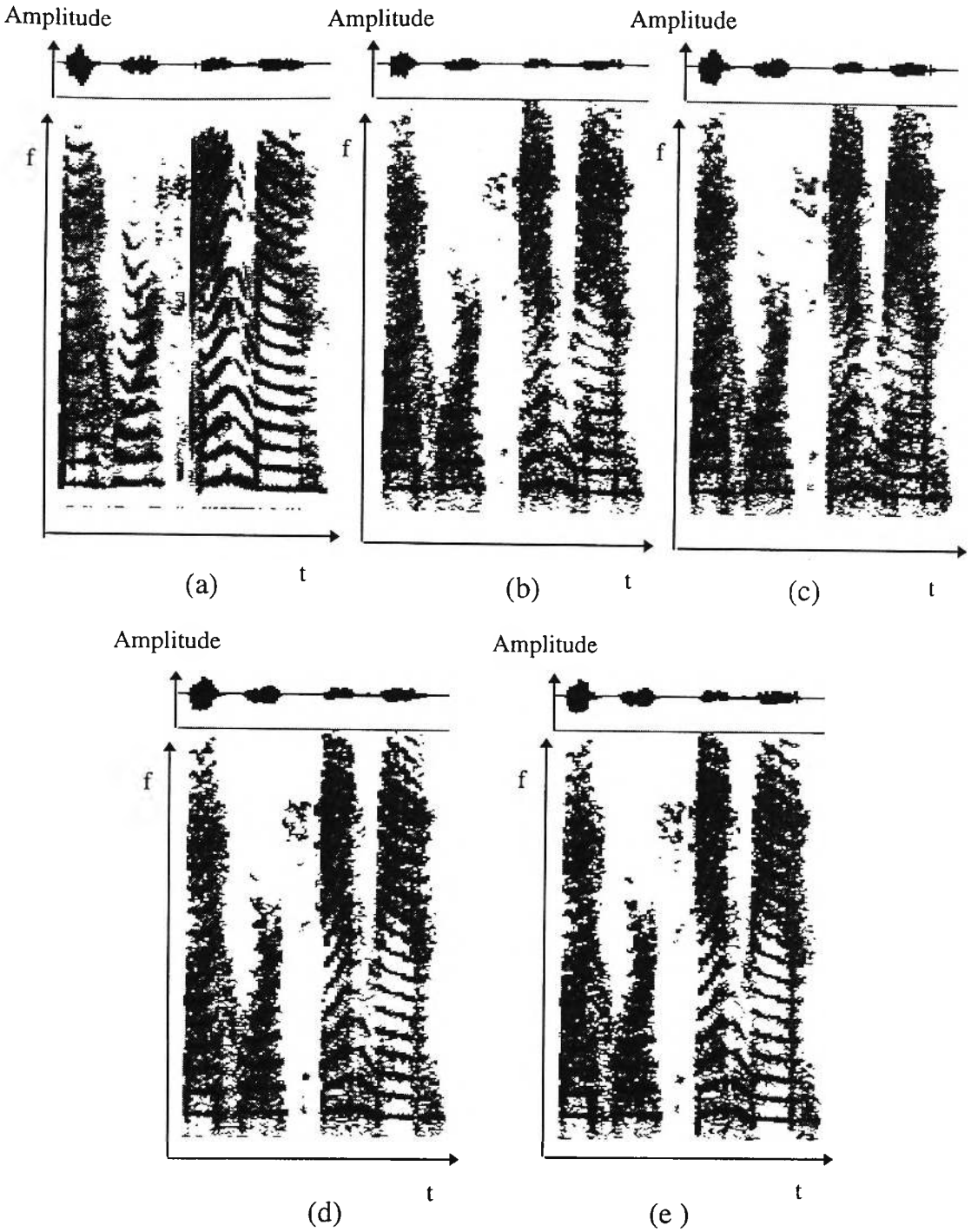


Figure 5.1: (a) The waveform (the upper part of the picture)and spectrogram (the lower part) of the original speech; (b) Coded speech of the 2.4kb/s coder; (c) Coded speech of the 2.72kb/s coder; (d) Coded speech of the 3.24 kb/s coder; (e) Coded speech of the 3.60kb/s coder.

Figure 5.1 illustrates the performance of these coders. In the 2.4kb/s coder, we can see from the power contour of the speech waveform that there is gain loss in the coded speech. Also, the spectrum of the coded speech is smeared, the pitch harmonic disperses. In the 2.72kb/s coder, with the improvement in the gain quantization, the loss in the power contour of the output speech is removed. In the 3.24kb/s coder, as the SEW/REW quantization is more accurate, the harmonic dispersion effect in the spectrum is greatly reduced. The spectrum of the coded speech is less distorted and has a clearer harmonic structure. The spectral distortion is further reduced in the 3.6kb/s coder.

5.4 Low Complexity Waveform Interpolation Coding

Waveform Interpolation coding paradigm performs well in terms of perceptual quality, speaker recognizability and robustness against channel errors [36]. However, the complexity of the WI code is very high. The waveform extraction procedure, including the intense DFT operations and time alignment operation, and the SEW/REW filtering procedure are enormously complex [34], [50]. This Chapter proposes approaches to low-complexity WI coding which greatly simplify the coding procedures.

The proposed low-complexity WI coder is based on the following considerations. At 2.4kb/s, the bit budget is so small (less than 0.1 bit per spectral component) that the SEW and REW are only poorly represented. The phase spectrum is discarded. The

REW magnitude is represented by five Chebyshev polynomials. For the SEW, only the magnitude spectrum below 800Hz is quantized and the speech quality is totally dominated by the quantizer. There is no need to generate high-resolution sequences of the SEW and REW. Therefore, the high-complexity waveform extraction and decomposition operations can be significant simplified.

5.5 Low-Complexity SEW/REW Decomposition and Quantization

In standard Waveform Interpolation coding, high resolution REW/SEW decomposition is performed by highpass/lowpass filtering the aligned Characteristic Waveforms (prototypes). However, at low bit rates, the REW and SEW can be obtained by a simpler procedure. In the low-complexity WI coder, the noise-like REW component can be defined as the difference between the normalized present and previous pitch-cycle prototypes [50]. The SEW is thus defined as the mean prototype of the current analysis frame [7], [50]. The complex highpass/lowpass filtering operation is thus removed. The low-complexity SEW and REW analysis and quantization is described in the following sections.

5.5.1 REW Quantization

Investigation of many of the REW spectrums found that most of the spectrum shapes are almost monotonically increasing with frequency in the region below 3500Hz and

decreasing in the region above 3500Hz for speech sampled at 8000Hz. Eight shapes are thus selected to form the 3-bit REW codebook. Figure 5.2 shows the shapes in the REW magnitude codebook.

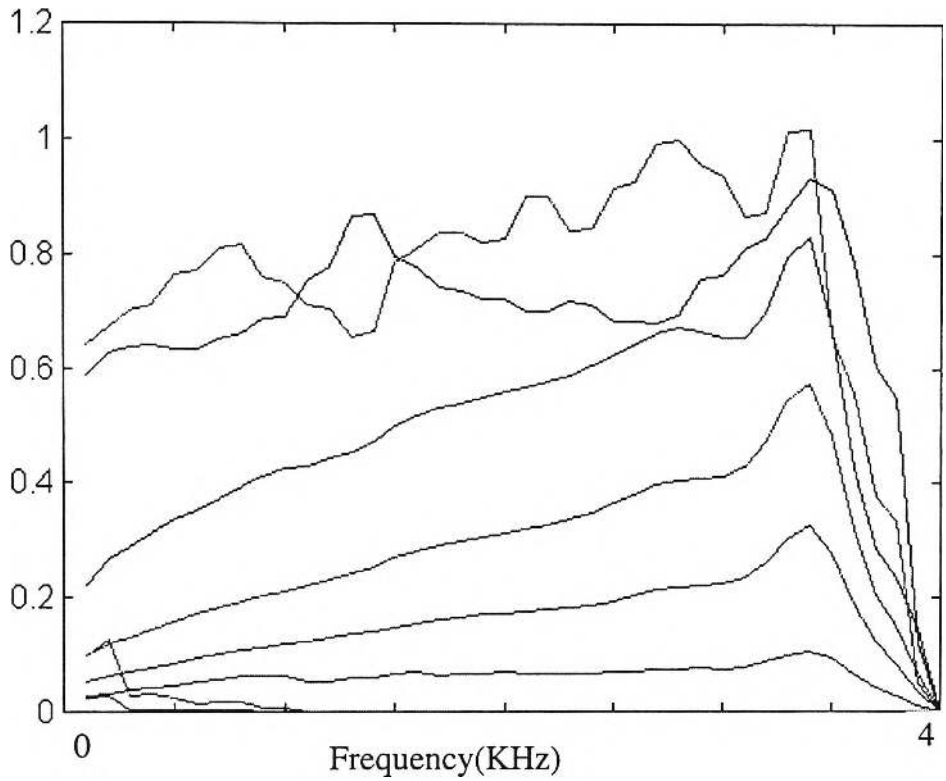


Figure 5.2: Eight shapes of the REW codebook

Coding the REW magnitude spectrum requires a curve fitting calculation. However, a simplified REW search procedure is proposed [50]. As shown in Figure 6.2, the eight REW codebook vectors have different levels of energy. So, the indices of the REW codebook vectors are made to correspond to their energies. Therefore, the REW codebook search can be performed by calculating the energy of the REW spectrum.

As the REW is defined as the difference between the present and previous prototype, the energy of the REW spectrum is approximately proportional to a factor:

$$u = 1 - R(P) \qquad \ldots\ldots \qquad (5.1)$$

where P is the pitch length, $R(.)$ is the standard normalized correlation function and $R(P)$ is the correlation between the present and previous prototypes. If the parameter u has a small value, the previous and present prototypes must be highly correlated, indicating a low level of REW energy. Alternatively, a large value of u indicates a high REW energy. The parameter u is then mapped into an index ranging from 0 to 7 which points to the REW codebook as:

$$REW_{indice} = map\ (\ u\ ) \qquad \ldots\ldots \qquad (5.2)$$

where $map(.)$ is the mapping function.

The REW magnitude is represented by a forty-dimensional 3-bit codebook. Each dimension represents a frequency bin which covers a 100Hz spectrum region. Compared with the Chebyshev polynomials representation, this method reduces the complexity in the REW decoding procedure.

This approach dramatically reduces the complexity of the REW analysis procedure. Firstly, the time alignment procedure is removed. Secondly, no highpass filtering is needed. And thirdly, the REW is obtained by time domain operations (correlation function), such that the DFT calculation is not required. Finally, the polynomial expansion analysis of the REW spectrum is not used.

5.5.2 SEW Quantization

The SEW is now defined as the average spectrum of the prototype in the current analysis frame. Given the pitch period P for the current frame, an integer M is determined which is equal to the number of the pitch-length prototypes in one frame (the frame size is 200).

$$M = \text{int}\,(\,200/P\,) \qquad \ldots\ldots \qquad (5.3)$$

The SEW is obtained by calculating the average DFT spectrum of the M pitch-length prototypes. Alternatively, to reduce the DFT complexity, a 256-point FFT can be applied. The size of the analysis frame is first extended to $M^{`}*P$.

$$M^{`} = int\,(256/P) \qquad \ldots\ldots \qquad (5.4)$$

The $M^{`}*P$ size signal sequence is padded with zeros at the end to a length of 256. Then, the FFT coefficients of the signal $\varpi(.)$ are calculated. The FFT spectrum has peaks at the pitch harmonic places. The magnitude of the pitch harmonics are thus extracted from the FFT spectrum by:

$$S(K) = \left|\omega(K\frac{256}{P})\right| \qquad K=0,1,\ldots,P \qquad \ldots\ldots \qquad (5.5)$$

where $|\omega(.)|$ is the magnitude of the FFT sequence, $S(K)$ is the Kth pitch harmonic. The pitch harmonic sequence $S(K)$ is equivalent to the unnormalized SEW. A gain-scaling procedure is then performed to this sequence. The lower 800Hz of the normalized SEW magnitude is quantized by an eight-dimension 7-bit codebook, while the remainder of the SEW spectrum is derived from the REW.

This SEW search procedure is much simpler than that used in the standard WI coder. No explicit prototype needs to be generated, and the time-alignment and lowpass filtering operation is removed. The high complexity, intense, DFT operation is replaced by applying FFT calculations directly to the residual sequence.

Tasks	Processing Time per Frame (ms)	Calls per Frame	Processing Time per call (ms)
Time Alignment	2.06	10	0.21
DFT	1.50	10	0.15
SEW/REW Filtering	0.33	10	0.033
Total	3.86		

Table 5.3: Computational complexity of SEW/REW decomposition in the standard WI coder.

Tasks	Processing Time per Frame (ms)	Calls per Frame	Processing Time per call (ms)
Polynomials Calculation	0.06	4	0.015
REW Codebook Search	0.03	1	0.03
Total	0.09		

Table 5.4: Computational complexity of REW quantization in the standard WI coder.

Tasks	Processing Time per Frame (ms)	Calls per Frame	Processing Time per call (ms)
SEW Codebook Search	0.14	1	0.14
Total	0.14		

Table 5.5: Computational complexity of SEW quantization in the standard WI coder.

Tasks	Processing Time per Frame (ms)	Calls per Frame	Processing Time per call (ms)
REW Energy Calculation	$2.2*10^{-3}$	10	$2.2*10^{-4}$
REW Codebook Search	$1.7*10^{-3}$	4	$4.3*10^{-4}$
Total	$3.9*10^{-3}$		

Table 5.6: Computational complexity of REW quantization in the low-complexity WI coder.

Tasks	Processing Time per Frame (ms)	Calls per Frame	Processing Time per call (ms)
FFT	0.31	1	0.31
SEW Codebook Search	0.14	1	0.14
Total	0.45		

Table 5.7: Computational complexity of SEW quantization in the low-complexity WI coder.

We have implemented a low-complexity and standard WI coder on a Pentium 166MHz personal computer using C language. For a pitch period of 5ms (40 samples), the standard SEW/REW decomposition and quantization requires processing time of 4.09ms, while the low-complexity SEW/REW analysis only needs 0.45ms processing

time. Tables 5.3~5.5 show the computational complexity of different tasks of SEW/REW quantization in the standard coder, and Tables 5.6~5.7 show the complexity of SEW/REW quantization in the low-complexity coder.

5.6 Low-complexity WI Coder

The overall low-complexity Waveform Interpolation coding architecture is described in the this section. At the encoder, the input speech is converted to the residual domain via an LP analysis filter. The pitch period is calculated from the residual signal once per frame. Ten pitch length prototypes are extracted. The gains of the ten prototypes are computed and differentially quantized. The time domain REW analysis (u-coefficients calculation) is performed four times per frame. Eight bits are used to transmit the REW information per frame, twice as a 3-bit index pointing to the REW codebook and twice as a binary decision between previous and future quantized REW. The SEW spectrum is computed once per frame. The analysis frame is extended to a length of 256 points and then an FFT operation is applied. The pitch harmonics are extracted from the FFT magnitude spectrum, and the SEW is obtained by gain-normalizing the pitch harmonic sequences.

The decoder is not changed. The residual domain prototype is constructed by the decoded SEW and REW. It is then converted to the speech domain by the synthesis filter. After gain-scaling, the output speech is obtained by continuous interpolation of the prototypes.

Figure 5.3 shows the diagram of the encoder of the low-complexity waveform algorithm. The decoder is same as the standard WI coder in Figure 2.4, and is thus not replete here.

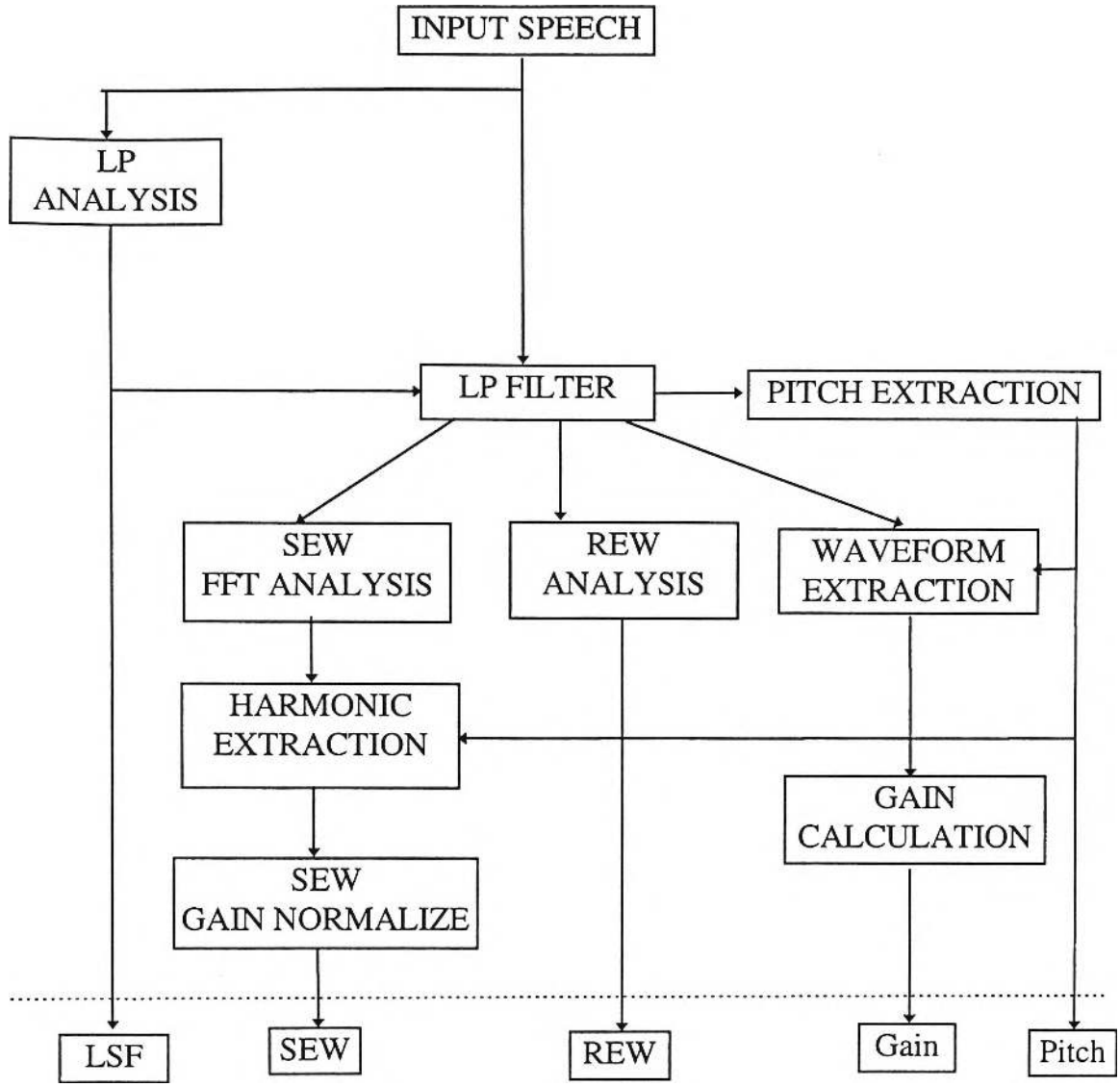


Figure 5.3: Diagram of Low-complexity WI encoder

For a speech file which contains 67 frames (frame size is 25ms), the standard WI coder needs 15.56 sec to encode the input speech, while the low-complexity WI coder only uses 6.33 sec. The execute time of the low-complexity coder is only 40.7% of that of the standard coder. Informal listening tests were conducted to test the performance of the low-complexity WI coder. Fifty percent of listeners could not find any degradation, while the remainder only recognized minor degradations in the speech quality.

5.7 Conclusions

This Chapter presented increased bit-rate Waveform Interpolation (WI) coders with rates between 2.4kb/s and 3.6kb/s. Both open-loop and closed-loop WI coders (the latter using an Analysis-by-Synthesis technique) were tested. In particular, the performance of the two types of WI coders at bit rates higher than 2.4kb/s is examined. At high bit rates, the codebook size and/or the update rate of the parameters used in WI coding can be increased. Results show that by increasing the bit rate, and with no or little change in the coder structure, the perceptual quality of speech produced by WI coders is substantially improved. The potential of higher bit rates for each of the quantized parameters was discussed, and suggested bit allocations for different increased bit rates are derived. Four coders, operating at bit rates of 2.4kb/s, 2.72kb/s, 3.24kb/s and 3.6kb/s, respectively were considered with informal listening tests show successive improvement in speech quality. At 3.6kb/s, WI approaches Toll quality. This indicates that Toll-quality coding, using WI, at 4kb/s is feasible. It is suggested that the remaining bit-allocation (between 3.6 to 4kb/s) might be used to improve the robustness of WI coded speech.

The low-complexity Waveform Interpolation coding paradigm at 2.4kb/s is also described in this Chapter. Since the bit budget for the SEW and REW is very small at this low bit rate, the complex high-resolution SEW/REW decomposition and quantization procedure was replaced by a simpler signal coding method in this scheme. The REW search is purely performed on the time domain, while the SEW analysis is primarily an FFT calculation. The highly complex time alignment operation, lowpass/highpass filtering operation and the intense DFT operation are not required. The computational load at the encoder is thus greatly reduced. The proposed low-complexity WI coding algorithm is between two and three times faster than the high-complexity WI coders for the encoder. Informal listening tests show that the quality of the low-complexity WI coder is equal or close to the standard WI coder.

Chapter 6

Conclusions and Suggestions for Further Research

6.1 Conclusions

Low-rate speech coding has advanced rapidly in the past decade to encompass such opportunities as cellular and satellite communications as well as computer-related voice applications. Central to this progress has been the research and development of a family of techniques described as Code-Excited Linear Prediction (CELP) coding, proposed firstly by B. S. Atal (1985). As the CELP class coders produce high-quality speech in the range of 4 to 16 kb/s, the research front has moved towards bit rates below 4kb/s. Several algorithms have been proposed to operate at this bit range including: mixed-excitation linear prediction (MELP) coding, multi-band excitation (MBE) coding, harmonic coding and so on. The Waveform Interpolation coding algorithm has been emerging as a promising approach in recent years that offers perceptually good quality in the neighbourhood of 2.4kb/s. The Waveform Interpolation coding mechanisms have been investigated and developed by many researchers such as Kleijn (1993), Burnett and Bradley (1995), Kleijn and Haagen (1995), Burnett and Pham (1996), Kleijn and Shoham (1996), Shoham (1997). It has been found that Waveform Interpolation coders perform better than other state-of-the-art coders in terms of speech quality and robustness against channel errors and background noise.

This thesis has dealed with the Waveform Interpolation coding algorithm, and has focused on new signal decomposition and quantization techniques to improve the

compression quality and reduce coding complexity. Chapter 2 gives a brief review and technical outline of the CELP algorithms. The principle of Waveform Interpolation coding is also introduced. The WI coders treat the incoming speech as a concatenation of evolving pitch-length prototypes (Characteristic Waveforms). The Characteristic Waveform surface is decomposed into a slowly-evolving waveform (SEW) and a rapidly-evolving waveform (REW). The SEW and REW are quantized differently as they have different perceptual properties. The signal decomposition and quantization procedures are performed in the linear prediction (LP) residual domain.

In Chapter 3, the details of a 2.4kb/s Waveform Interpolation coder are described. The quantization of the LSFs is improved by using a weighted MSE criteria. As pitch estimation is crucial to the WI coder, a new pitch estimation method is proposed. This pitch estimator gives reliable pitch values even when the pitch period is changing rapidly. Only the SEW magnitude spectrum below 800HZ is quantized, which reduces the quantization complexity. The REW magnitude spectrum is quantized with Chebyshev polynomials.

Analysis-by-synthesis (A-by-S) mechanisms have found favour in low-rate speech coders. However, the WI class of coders which is based on open-loop quantization of the Characteristic Waveforms, does not explicitly incorporate this technique. A closed-loop WI coder which utilise the A-by-S mechanisms is presented and investigated in Chapter 4. As the output speech of the WI coder is generally not synchronous with the input speech, a modified A-by-S technique is proposed to operate on a prototype-by-prototype basis rather than a sample-by-sample basis.

Furthermore, perceptual weighting techniques can be incorporated in this A-by-S architecture. Results show that the performance of the WI coder is improved. The output speech sounds clearer and smoother.

The Waveform Interpolation coding structure also offers scalability to work at different bit rates. The transmission rate of the WI coder can be changed simply by using different update rates and/or codebook size of the Characteristic Waveform parameters. At higher bit rates, the CW will be quantized better, suggesting the possibility of improving the coder performance by increasing the bit rate. The performance of WI coders at bit rates higher than 2.4kb/s is tested in chapter 5. Firstly, the effect of higher bit rates for each of the coder parameters was investigated. Bit allocations of coders working between 2.4kb/s to 3.6kb/s were then proposed. Informal listening tests indicated successive improvement in the speech quality.

Although Waveform Interpolation coders produce good-quality speech, the computational complexity is relatively high. Chapter 5 introduces a low-complexity WI coder operating at 2.4kb/s. A simplified SEW/REW decomposition and quantization mechanism is proposed. The encoder works 2 to 3 times faster than the high complexity one, while the speech quality is maintained.

6.2 Suggestions for Further Research

Waveform Interpolation coding is a very active area of research and development, yet there are many challenges ahead for researchers. The following suggestion are made for further research.

6.2.1 CW Decomposition

The Characteristic Waveform (CW) decomposition is, perhaps, the most distinguishing feature of WI coding. The aim of the decomposition procedure is to separate the unvoiced component (REW) and the voiced component (SEW). The decomposition is achieved by highpass/lowpass filtering of the CW surface. However, the filtering operation does not provide a thorough separation of the voiced and unvoiced speech. If the $\overline{\text{REW}}$ contains some voiced signal, the reproduced speech will sound rough, and the mixed unvoiced signal in SEW may introduce tonal effect. A more accurate decomposition needs to be considered to improve the quantization quality.

6.2.2 SEW Magnitude Quantization

In this thesis, only the baseband containing the lower 800Hz of the SEW magnitude spectrum is coded. A full-spectrum quantization should improve the SEW representation. A 1600Hz baseband SEW quantization was tested in Chapter 5, but results failed to verify such improvement. It may be explained by the properties of

auditory perception that the higher frequency components of the spectral envelope are perceptually less important. A more sophisticated perceptual weighting technique which emphasizes the lower frequency part is required for the full-spectrum SEW quantization [49].

6.2.3 REW Magnitude Quantization

Examination of many REW magnitude spectrum has revealed that most of them are increasing smoothly in the frequency region below 3500Hz and decreasing quickly above 3500Hz. The shifted Chebyshev polynomials do not match this shape very well. This means a more efficient polynomial expansion analysis needs to be considered. The solutions that may be taken are as follows:

- As the high frequency part of the REW is perceptually less important, the REW above 3500Hz can be ignored in the polynomial expansion analysis. The smooth lower spectrum of REW will be better fitted by the Chebyshev polynomials.
- Use of other polynomials which are able to match the REW shape may improve the curve fitting.
- Cubic cardinal splines [24], [56], [57] have been used by some researchers (Kleijn) in the Waveform Interpolation procedure [34]. It is possible that the spline representation will approximate the REW shape better.

6.2.4 SEW/REW Phase Quantization

In low-rate WI coders, the phase spectrum of the SEW and REW is not transmitted but approximated by a random spectrum or a spectrum representing a pulse shape. There are, however, possible advantages to transmitting the phase spectrum at higher bit rates. Since generally the SEW/REW decomposition procedure does not separate the voiced and unvoiced components of the speech thoroughly, the phase approximation of the SEW and REW will introduce some distortion. With the phase spectrum being transmitted, the Characteristic Waveforms will be better represented. Furthermore, given the SEW phase spectrum, the voiced/unvoiced decision is not needed. The coder will be more robust against the pitch error and the background noise. As the phase quantization is a complex procedure, the efficient phase quantization is still an unresolved problem which needs further consideration.

6.2.5 Variable Rate WI Coder

The Waveform Interpolation coding algorithm offers scalability to work at different bit rates. With no or little modification in the coder structure, the transmission rate of the WI coder can be changed resulting in a variable rate coder. Further research is required to design an embedded coding structure and codebooks for variable WI coding.

6.2.6 Low Complexity WI Coder

A simplified signal decomposition technique has been proposed to reduce the complexity in the WI encoder. The computational load of the decoder can also be reduced. By using the spline representation and FFT operation, the complex DFT operation in the decoder will be removed.

Author's Publications

Ni, Jun., Burnett, I.S., 'Waveform Interpolation at bit rates above 2.4kb/s', T*enCom97, IEEE*, Brisbane, Australia, 1997.

Burnett, I.S., Ni, Jun., 'Waveform Interpolation and analysis-by-synthesis - a good match ?', *Speech coding workshop, IEEE*, 1997.

References

[1] Atal, B.S., 'Predictive coding of speech at low bit rates', *IEEE Transactions on Communications*, Vol. COM-30, No.4, April 1982, p600-614.

[2] Atal, B.S. and Schroeder, 'Stochastic coding of speech signal at very low bit rates', *Proc. Int. Conf. On Communications*, May 1984, pp1610-1613.

[3] Atal, B.S., 'Code-excited Linear Prediction (CELP): High-quality Speech At Very Low Bit Rates', *Proc. Int. Conf. On Acoustics, Speech and Signal Processing, IEEE*, March 1985, pp937-940.

[4] Atal, B.S. and Remde, J.R., 'A New Model of Excitation for Producing Natural Sounding Speech At Low Bit Rates', *Proc. Int. Conf. On Acoustics, Speech and Signal Processing, IEEE*, 1982, pp614-617.

[5] Burnett, I.S. and Holbeche, R.J., 'A Mixed Prototype Waveform/CELP Coder for Sub 3kb/s', *Proc. Int. Conf. Acoustics Speech Signal Processing, IEEE*, 1993, pp II175-II178.

[6] Burnett, I.S. and Bradley, G.J., 'New Techniques for Multi-Prototype Waveform Coding at 2.84kb/s', *Proc. Int. Conf. Acoustics Speech Signal Processing, IEEE*, 1995, pp261-264.

[7] Burnett, I.S. and Phan, D.H., 'Multi-Prototype Waveform Coding using Frame-by-Frame Analysis-by-Synthesis', *Proc. Int. Conf. Acoustics Speech Signal Processing, IEEE*, 1996, pp212-215.

[8] Chen, J.H. and Gersho, A., 'Gain-Adaptive Vector Quantization with Application to Speech Coding', *IEEE Transactions on Communications*, Vol. COM-35, No. 9, September 1987, pp918-930.

[9] Chen, J.H. and Cox, R.V., 'LD-CELP: A High Quality 16kB/s Speech Coder With Low Delay', *Conf. Rec. IEEE*, VOL.1, 1990, pp528-532.

[10] Chen, J.H., Cox, R.V. and Lin, Y.C., 'A Low-Delay CELP Coder for the CCITT 16kb/s Speech Coding Standard', *IEEE Journal on Selected Areas in Communications*, Vol. 10, No. 5, June 1992, pp830-849.

[11] Chen, J.H. and Gersho, A., 'Adaptive Postfiltering for Quality Enhancement of Coded Speech', *Proc. IEEE On Speech and Audio Processing*, VOL.3, NO.1, January 1995.

[12] Cheng, Y.M. and O'Shaughnessy, D., 'Automatic and Reliable Estimation of Glottal Closure Instant and Period', *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 37, No. 12, December 1989, pp1805-1815.

[13] Cuperman, V. and Gersho, A., 'Vector Predictive Coding of Speech at 16kbits/s', *IEEE Transactions on Communications*, Vol COM-33, No. 7, July 1985, pp685-696.

[14] Cox, R.V., Kroon, P., Chen, J.H., Thorkildsen, R., O'Dell, K.M. and Isenberg, D.S., 'Speech Coders: From Idea to Product', *AT&T Technical Journal*, March/April, 1995, pp14-21.

[15] Das, A., Rao, A.V. and Gersho, A., 'Variable - Dimension Vector Quantization', *IEEE Signal Processing Letters*, Vol. 3, No. 7, July 1996, pp200-202.

[16] Davidson, G. and Gersho, A., 'Complexity Reduction Methods For Vector Excitation Coding', *Proc. Int. Conf. On Acoustics, Speech, and Signal Processing, IEEE*, 1986, pp3055-3058.

[17] Details To Assist In Implementation Federal Standard 1016 CELP, *NCS Technical Information Bulletin 92-1*, 1992.

[18] Draft Recommendation G.729, *ITU-U Sector*, June, 1995.

[19] Eriksson, T. and Sjoberg, J., 'Dynamic Bit Allocation in CELP Excitation Coding', *Proc. Int. Conf. Acoustics Speech Signal Processing*, 1993, pp II171-II174.

[20] Gersho, A., 'Advances in Speech and Audio Compression', *Proceedings of the IEEE,* Vol. 82, No. 6, June, 1994, pp900-918.

[21] Gerson, I. and Jasiuk, M., 'Vector Sum Excited Linear Prediction (VSELP) Speech Coding at 8kb/s', *Proc. Int. Conf. On Acoustics, Speech, and Signal Processing, IEEE*, VOL.1, April 1990, pp461-464.

[22] Granzow, W., Atal, B.S., Paliwal, K.K. and Schroeter, J., 'Speech Coding at 4kb/s and Lower using Single-Pulse and Stochastic Models of LPC Excitation', *Proc. Int. Conf. Acoustics Speech Signal Processing*, 1991, pp217-220.

[23] Griffin, D.W. and Lim, J.S., 'Multi-band excitation vocoder', *Proc. Int. Conf. Acoustics Speech Signal Processing, IEEE*, 1988, pp1223-1235.

[24] Hou, H.S. and Andrews, H.C., 'Cubic Splines for Image Interpolation and Digital Filtering', *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-26, No. 6, December 1978, pp508-517.

[25] INMARSAT M Voice Codec, *Digital Voice System Inc.*, August 1991, pp1-143.

[26] Kang, G.S. and Fransen, L.J., 'Application Of Line-Spectrum Pairs To Low-Bit-Rate Speech Encoders', *Proc. Int. Conf. On Acoustics, Speech, and Signal Processing, IEEE* , 1985, pp244-247.

[27] Kataoka, A., Moriya, T. and Hayashi, S., 'An 8-kbit/s Speech Coder Based on Conjugate Structure CELP', *Proc. Int. Conf. Acoustics Speech Signal Processing*, 1993, pp II592-II595.

[28] Kleijn, W.B., 'Speech Coding Below 4kb/s Using Waveform Interpolation', *Globecom*, 1991, pp. 1879-1883.

[29] Kleijn,W.B., 'Continuous Representations in Linear Predictive Coding', *Proc. Int. Conf. Acoustics Speech Signal Processing, IEEE* , May 1991, pp201-204.

[30] Kleijn, W.B., Kroon, P., Cellario, L. and Sereno, D., 'A 5.85 kb/s CELP Algorithm for Cellular Applications', *Proc. Int. Conf. Acoustics Speech Signal Processing, IEEE*, 1993, pp. II596-II598.

[31] Kleijn, W.B., 'Encoding Speech Using Prototype Waveforms', *IEEE Trans. Speech Audio Processing*, Vol.1, 1993, pp 386-399.

[32] Kleijn, W.B. and Haagen, J., 'A Speech Coder Based on Decomposition of Characteristic Waveforms', *Proc. Int. Conf. Acoustics Speech Signal Processing, IEEE*, 1995, pp 508-511.

[33] Kleijn, W.B. and Haagen, J., 'Waveform Interpolation for Speech Coding and Synthesis', *Speech Coding and Synthesis* (Kleijn, W.B. and Paliwal, K.K., eds.), Elsevier Science Publishers, 1995, pp175-208.

[34] Kleijn, W.B., Shoham Y., Sen, D. and HageR, J., 'A Low-complexity Waveforms Interpolation Coder', *Proc. Int. Conf. Acoustics Speech Signal Processing, IEEE*, 1996, pp 212-215.

[35] Kleijn, W.B and Hagen, R., 'On Memoryless Quantization In Speech Coding', *IEEE Signal Processing Letters*, Vol. 3, No. 8, August 1996, pp228-230.

[36] Kohler, M.A., Supplee, L.M. and Tremain, T.E., 'Progress Towards a New Government Standard 2400 bps Voice Coder', *Proc. Int. Conf. Acoustics Speech Signal Processing, IEEE*, 1995, pp488-491.

[37] Kroon, P., Deprettere, E.F. and Sluyter R.J., 'Regular-pulse excitation: a novel approach to effective and efficient multi-pulse coding of speech', *Proc. Int. Conf. Acoustics Speech Signal Processing, IEEE*, 1986, pp1054-1063.

[38] Laroia, R., Phamdo, N. and Farvardin, N., 'Robust and efficient quantization of speech LSP parameters using structured vector quantizers', *Proc. Int. Conf. Acoustics Speech Signal Processing, IEEE* , 1991, pp641-644.

[39] Lupini, P., Hassanein, H. and Cuperman, V., 'A 2.4kb/s CELP Speech Codec with Class-Dependent Structure', *Proc. Int. Conf. Acoustics Speech Signal Processing, IEEE* , 1993, pp II143-II146.

[40] McAulay, R.J. and Quatieri, T.F., 'Speech analysis/synthesis based on a sinusoidal representation', *Proc. Int. Conf. Acoustics Speech Signal Processing, IEEE*, 1986, pp744-754.

[41] Marques, J.S., 'Harmonic coding at 4.8kb/s', *Proc. Int. Conf. Acoustics Speech Signal Processing, IEEE* , 1990, pp17-20.

[42] McCree, A.V. and Barnwell, T.P., 'A New Mixed Excitation LPC Vocoder', *Proc. Int. Conf. Acoustics Speech Signal Processing, IEEE* , 1991, pp593-596.

[43] McCree, A.V. and Barnwell, T.P., 'Implementation and Evaluation of a 2400 bps Mixed Excitation LPC Vocoder', *Proc. Int. Conf. Acoustics Speech Signal Processing, IEEE* , 1993, pp159-162.

[44] Paliwal, K.K. and Atal, B.S., 'Efficient vector quantization of LPC parameters at 24bits/frame', *Proc. Int. Conf. Acoustics Speech Signal Processing, IEEE* , 1991, pp661-664.

[45] Paliwal, K.K., 'Quantization of LPC parameters', *Speech Coding and Synthesis* (Kleijn, W.B. and Paliwal, K.K., eds.), Elsevier Science Publishers, 1995, pp433-466.

[46] Rabiner, L.R., 'Toward Vision 2001: Voice and Audio Processing Consideration,' *AT&T Technical Journal*, March/April 1995, pp4-13.

[47] Ramachandran, R.P. and Mammone R., 'Modern methods of speech processing', *Kluwer Academic Publishers*, 1995.

[48] Rivlin, T.J., 'The Chebyshev polynomials', *A Wiley-Interscience Publication*, 1990.

[49] Sen, D., Irving, D.H. and Holmes, W.H., 'Use of an Auditory Model to Improve Speech Coders', *Proc. Int. Conf. Acoustics Speech Signal Processing*, 1993, pp II411-II414.

[50] Shoham, Y., 'Very Low-complexity Interpolative Speech Coding at 1.2 to 2.4 kbps', *Proc. Int. Conf. Acoustics Speech Signal Processing, IEEE*, 1997, pp1599-1602.

[51] Singhal, S. and Atal, B.S., 'Improving Performance of Multi-Pulse LPC Coders at Low Bit Rates', *Proc. Int. Conf. Acoustics Speech Signal Processing*, Vol. 1, No. 1.3, March 1984.

[52] Spanias, A.S., 'Speech Coding: A Tutorial Review', *Proceedings of the IEEE*, Vol.82, No.10, October, 1994, pp1541-1582.

[53] Software Tools for Speech and Audio Coding Standardization, *Telecommunication Standardization Sector of ITU*, November 1996.

[54] Saito, S., 'Fundamentals of speech signal processing', *Academic Press*, 1985.

[55] Trancoso, I.M. and Atal, B.S., 'Efficient Procedure For Finding The Optimum Innovation In Stochastic Coders', *Proc. Int. Conf. Acoustics Speech Signal Processing*, 1986, pp2375-2378.

[56] Unser, M., Aldroubi, A. and Eden, M., 'B-Spline Signal Processing: Part I - Theory', *IEEE Transactions on Signal Processing*, Vol. 41, No. 2, February 1993, pp821-833.

[57] Unser, M., Aldroubi, A. and Eden, M., 'B-Spline Signal Processing: Part II - Efficient Design and Applications', *IEEE Transactions on Signal Processing*, Vol. 41, No. 2, February 1993, pp834-848.

Appendix I: Testing of WI Coder in Chapter Three

In Chapter Three, informal listening tests were taken to compare the performance of the new 2.4kb/s WI coder with the baseline coder described in Chapter two. Sixteen listeners were selected They were the students of University of Wollongong. They are all native English speakers. Their age was between 20 to 26 years old. Among the listeners, 11 are male, 5 are female.

The test material was chosen from a voice database provided by University of Oregon, USA. The test material included 6 sentences spoken at normal speed. Each sentence has duration of about 10 seconds. Three different male speakers spoke three of the sentences; three female speakers spoke the other three sentences.

The tests were organized in such way: First, the original speech was played, then the synthesis speech of the two coders were played in a random order. The listeners were asked to give preference of the synthesis speech compared with the original speech. Each listener was required to test all the six sentences.

The hardware used in these tests included a Pentium166 Personal Computer, a 16bit sound card with digital output and a digital tape recorder, the software was the *Goldwave* shareware. The test material was edited using the *Goldwave*. The speech was also played by the *Goldwave*, outputted by the sound card, and recorded into a digital tape. During the tests, the digital tape, which contains the test material, was replayed and the listeners listen to the speech through headphone.

Appendix II: The Complexity of the WI Coder in Chapter Five

A low complexity WI coder was presented in Chapter Five. This coder greatly simplified the SEW/REW decomposition and quntization procedures. A detailed analysis of the complexity reduction by means of orders of operations is given below.

In a standard WI coder, the encoding process includes LP analysis and LSF quantization, pitch detection, waveform extraction, SEW/REW decomposition and quantization, and gain quantization. All together approximately $1.4*10^6$ multiples and $1.4*10^6$ adds are needs for one frame of speech (40 frames equals to one second), e.g., 10^2MIPS (million instruction per second).

Tasks	Multiple Operations (per frame)	Add Operations (per frame)
LP Analysis and LSF Quantization	$1*10^5$	$1*10^5$
Pitch Detection	$2*10^5$	$2*10^5$
Waveform Extraction	$1*10^4$	$1*10^4$
SEW/REW Decompostion	$1*10^6$	$1*10^6$
SEW/REW Quantization	$3*10^4$	$3*10^4$
Gain Quantization	$8*10^3$	$8*10^3$
Total	$1.4*10^6$	$1.4*10^6$

Table A.1: Computational complexity of the standard WI coder

In low complexity WI coder, the computation load of SEW/REW decomposition and quantization is reduced. It will take approximately $3*10^5$ multiples and adds to encode one frame speech, which is about 20MIPS.

Tasks	Multiple Operations (per frame)	Add Operations (per frame)
LP Analysis and LSF Quantization	$1*10^5$	$1*10^5$
Pitch Detection	$2*10^5$	$2*10^5$
Waveform Extraction	$1*10^4$	$1*10^4$
SEW/REW Decomposition and Quantization	$6*10^4$	$6*10^4$
Gain Quantization	$8*10^3$	$8*10^3$
Total	$3*10^5$	$3*10^5$

Table A.2: Computational complexity of the low complexity WI coder

The result shown that the low complexity WI coder reduce the operations dramatically. What needs to be mentioned is that both the standard and low-complexity WI coders were programmed in C code and were not optimized. An optimized DSP code should be able to reduce the computation load further more.