Advanced Pre-and-Post Processing Techniques for Speech Coding

HASSAN FARSI

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

Centre for Communication Systems Research (CCSR) School of Electronics and Physical Sciences University of Surrey Guildford, Surrey GU2 7XH, UK

August 2003

© H. Farsi 2003
Summary

Advances in digital technology in the last decade have motivated the development of very efficient and high quality speech compression algorithms. While in the early low bit rate coding systems, the main target was the production of intelligible speech at low bit rates, expansion of new applications such as mobile satellite systems increased the demand for reducing the transmission bandwidth and achieving higher speech quality. This resulted in the development of efficient parametric models for speech production system. These models were the basis of powerful speech compression algorithms such as CELP, MBE, MELP and WI.

The performance of a speech coder not only depends on the speech production model employed but also on the accurate estimation of speech parameters. Periodicity, also known as pitch, is one of the speech parameters that greatly affect the synthesised speech quality. Thus, the subject of pitch determination has attracted much research in the area of low bit rate coding. In these studies it is assumed that for a short segment of speech, called frame, the pitch is fixed or smoothly evolving. The pitch estimation algorithms generally fail to determine irregular variations, which can occur at onset and offset speech segments. In order to overcome this problem, a novel preprocessing method, which detects irregular pitch variations and modifies the speech signal such as to improve the accuracy of the pitch estimation, is proposed. This method results in more regular speech while maintaining perceptual speech quality.

The perceptual quality of the synthesised speech may also be improved using postfiltering techniques. Conventional postfiltering methods generally consider the enhancement of the whole speech spectrum. This may result in the broadening of the first formant, which leads to the increase of quantisation noise for this formant. A new postfiltering technique, which is based on factorising the linear prediction synthesis filter, is proposed. This provides more control over the formant bandwidth and attenuation of spectral speech valleys.

Key words: Pitch smoothing, speech pre-processor, postfiltering.
Acknowledgments

I would like to take this opportunity to express my sincere thanks to my supervisor Prof. Ahmet Kondoz whose support and guidance throughout my research was most helpful and much appreciated. Special thanks to Dr. Stephane Villette for his friendly advice and help. I am also grateful to my colleagues in the Multimedia Research Group, especially Dr. Khaldoon Taha Al-naimi, Stewart Worrall, Sertaç Eminsoy and Christian Sturt for their friendship and help.

I am greatly indebted to my parents and my family whose constant love and support enabled me to reach this stage of education. As a small token of my appreciation, I would like to dedicate this work to them. Last but not least I would like to acknowledge the support of the Ministry of Culture and Higher Education of the Islamic Republic of Iran for providing the sponsorship throughout this research.
A List of publications ................................................................. 163
Bibliography ......................................................................... 164
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D-to-1D</td>
<td>Two Dimensional to one Dimensional</td>
</tr>
<tr>
<td>A/D</td>
<td>Analogue to Digital</td>
</tr>
<tr>
<td>ACELP</td>
<td>Algebraic Code Excitation Linear Prediction</td>
</tr>
<tr>
<td>ACR</td>
<td>Absolute Category Rating</td>
</tr>
<tr>
<td>ADPCM</td>
<td>Adaptive Difference Pulse Code Modulation</td>
</tr>
<tr>
<td>AMDF</td>
<td>Average Magnitude Difference Function</td>
</tr>
<tr>
<td>ARMA</td>
<td>Autogressive Moving Average</td>
</tr>
<tr>
<td>ASRC</td>
<td>Arcsine of Reflection Coefficients</td>
</tr>
<tr>
<td>ATC</td>
<td>Adaptive Transform Coding</td>
</tr>
<tr>
<td>CCITT</td>
<td>Consultative Committee International Telegraph and Telephone</td>
</tr>
<tr>
<td>CDMA</td>
<td>Code Division Multiple Access</td>
</tr>
<tr>
<td>CELP</td>
<td>Code Excitation Linear Prediction</td>
</tr>
<tr>
<td>CS-ACELP</td>
<td>Conjugate Structure Algebraic-Code-Excited Linear-Prediction</td>
</tr>
<tr>
<td>CW</td>
<td>Characteristic Waveform</td>
</tr>
<tr>
<td>DCT</td>
<td>Discrete Cosine Transform</td>
</tr>
<tr>
<td>DFT</td>
<td>Discrete Fourier Transform</td>
</tr>
<tr>
<td>DSP</td>
<td>Digital Signal Processing</td>
</tr>
<tr>
<td>DTFS</td>
<td>Discrete Time Fourier Series</td>
</tr>
<tr>
<td>EVRC</td>
<td>Enhanced Variable Rate Coder</td>
</tr>
<tr>
<td>FFT</td>
<td>Fast Fourier Transform</td>
</tr>
<tr>
<td>GCI</td>
<td>Glottal Closure Instant</td>
</tr>
<tr>
<td>Acronym</td>
<td>Definition</td>
</tr>
<tr>
<td>-----------</td>
<td>------------------------------------------------------</td>
</tr>
<tr>
<td>GCIDS</td>
<td>Glottal Closure Instant Determination Signal</td>
</tr>
<tr>
<td>GSM</td>
<td>Global System for Mobile Communication</td>
</tr>
<tr>
<td>HEWLPR</td>
<td>Hilbert Envelope of Windowed Linear Prediction Residual</td>
</tr>
<tr>
<td>IMBE</td>
<td>Improved Multi Band Excitation</td>
</tr>
<tr>
<td>ITU-T</td>
<td>International Telecommunication Union – Telecommunication</td>
</tr>
<tr>
<td>LAR</td>
<td>Log Area Ration</td>
</tr>
<tr>
<td>LD-CELP</td>
<td>Low Delay Code Excitation Linear Prediction</td>
</tr>
<tr>
<td>LP</td>
<td>Linear Prediction</td>
</tr>
<tr>
<td>LPC</td>
<td>Linear Prediction Coding</td>
</tr>
<tr>
<td>LSF</td>
<td>Line Spectral Frequency</td>
</tr>
<tr>
<td>LSP</td>
<td>Line Spectral Pairs</td>
</tr>
<tr>
<td>MBE</td>
<td>Multi Band Excitation</td>
</tr>
<tr>
<td>MELP</td>
<td>Mixed Excitation Linear Prediction</td>
</tr>
<tr>
<td>MIPS</td>
<td>Millions of Instructions Per Second</td>
</tr>
<tr>
<td>MLED</td>
<td>Maximum Likelihood Epoch Determination</td>
</tr>
<tr>
<td>MOS</td>
<td>Mean Opinion Score</td>
</tr>
<tr>
<td>PCM</td>
<td>Pulse Code Modulation</td>
</tr>
<tr>
<td>PCS</td>
<td>Personal Communication Systems</td>
</tr>
<tr>
<td>PPG</td>
<td>Pitch Prediction Gain</td>
</tr>
<tr>
<td>PWI</td>
<td>Prototype Waveform Interpolation</td>
</tr>
<tr>
<td>RAM</td>
<td>Random Access Memory</td>
</tr>
<tr>
<td>REW</td>
<td>Rapidly Evolving Waveform</td>
</tr>
<tr>
<td>RPE-LTP</td>
<td>Regular Pulse Excitation-Long-Term Prediction</td>
</tr>
<tr>
<td>SDM</td>
<td>Spectral Distortion of Modified</td>
</tr>
<tr>
<td>SDO</td>
<td>Spectral Distortion of Original</td>
</tr>
<tr>
<td>SEW</td>
<td>Slowly Evolving Waveform</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal to Noise Ratio</td>
</tr>
<tr>
<td>SSMMO</td>
<td>Synthetic Spectral Matching of Modified to Original</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>TIA</td>
<td>Telecommunication Industry Association</td>
</tr>
<tr>
<td>V/UV</td>
<td>Voiced/Unvoiced</td>
</tr>
<tr>
<td>VBR</td>
<td>Variable Bit Rate</td>
</tr>
<tr>
<td>VLSI</td>
<td>Very Large Scale Integration</td>
</tr>
<tr>
<td>VSELP</td>
<td>Vector Sum Excited Linear Prediction</td>
</tr>
<tr>
<td>WI</td>
<td>Waveform Interpolation</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Background

Speech communication is possibly the most important interface between humans, and it is now becoming an increasingly important interface between human and machine. As such, speech represents a central component of digital communication systems and constitutes a major driver of telecommunications technology.

With the increasing demand for telecommunication services (e.g. digital cellular telephony and mobile satellite systems), speech coding has become a fundamental element of digital communication systems. The original Pulse Code Modulation (PCM) required a sampling frequency of 8kHz and 8 bits per sample when using logarithmic companding [1]. This sets the bit rate of digital speech at 64 kb/s. In the 1970’s, adaptive quantisation techniques such as ADPCM were introduced, which reduced the bit rate to 32 kb/s while maintaining high quality speech [2]. This was acceptable on trunk telephone links, where large bandwidths are available.

However, since the mid 1980’s there has been great interest in mobile telecommunications systems, in particular satellite systems and cellular telephony. In these applications the bandwidth is very limited, due to the characteristics of the communication channel used and the growing number of users. This requires to use low bit rate, reliable, high quality speech coders. The need to save bandwidth in wireless networks and to reduce the memory requirements in voice storage systems are two of the many reasons for the high activity in low bit rate speech coding research and development.

Speech coding algorithms have been developed, operating at various bit rates, from very low bit rates (below 1 kb/s) producing synthetic speech to high bit rates (8-16 kb/s) producing toll quality, i.e. speech sounding like the original. The fast growing mobile telephony industry and the emergence of new multimedia products are demanding even lower bit rates while maintaining high quality speech. Military mobile applications, wireless Personal Communication Systems (PCS) and voice-related computer applications (e.g., message storage, speech and audio over internet) also contribute to that demand.

The synthesised speech quality of a low bit rate speech-coding algorithm depends on accurate determination of the speech model parameters such as pitch and voicing. Due to existing non-stationary characteristics of speech signals, especially when moving from unvoiced speech
Chapter 1. Introduction

segments to voiced ones, the conventional methods for speech parameter determinations may fail.

The work presented in this thesis focuses on the development of low bit rate speech coders through improving the estimation of the important speech parameters such as pitch and voicing. A pre-processor is proposed to facilitate efficient estimation and coding of pitch and voicing. Since the perceptual speech quality is maintained, the proposed pre-processor can be used in combination with any low bit rate speech coder.

The perceptual speech quality may also be improved using postfiltering techniques in synthesis stage and thus a new postfiltering method is proposed.

The performance of the proposed pre-processor and postfilter is evaluated in combination with the standard Mixed Excitation Linear Prediction (MELP) 2.4 Kb/s [3] and the proposed 8.25 Kb/s speech coder [4] which is based on Algebraic Code-Excited Linear Prediction (ACELP) [5] structure, respectively.

1.2 Outline of thesis

The objectives of this thesis are twofold. The first is related to the improved pitch and voicing estimations through a pre-processing method. The second objective is to enhance the synthesised speech quality using a postfiltering technique.

Chapter 2 starts with an introduction to the major issues in speech coding. Followed by a brief description of speech coder classification and the main attributes affecting speech coders design.

Chapter 3 introduces basic tools employed by speech coders. This includes a description of the well-known Linear Prediction (LP) model of speech, speech classification and pitch determination. In the pitch determination part, different techniques are presented in time and frequency domain. Next, the major features used in voiced/unvoiced (V/UV) classification of speech are discussed. Binary V/UV classification has its limitations arising when a speech segment as a mixture of voiced and unvoiced. Voicing-level is introduced to overcome this problem. Two popular techniques employed in Multi-Band Exciting (MBE) [6] and Mixed Excitation Linear Prediction (MELP) [7], are then described.

Chapter 4 briefly introduces Waveform Interpolation (WI) and the standard MELP coders.

In first part of Chapter 5, the effect of irregular pitch variations on pitch estimation is presented. The second part addresses inaccuracy of the estimated pitch on WI coder. This
includes Characteristic Waveform (CW) extraction and decomposition in analysis stage and speech reconstruction in synthesis stage. In the third part, the effect of irregular pitch variation on the voicing level strengths in the standard MELP coder is investigated.

Using a pre-processor placed before speech encoder enhances the accuracy of the estimated parameters. In Chapter 6, we propose a general pre-processor, which results in more regular speech and smoother pitch evolutions and maintains perceptual speech quality. This is performed by moving the irregular pitch pulses. The pre-processor not only modifies irregular pitch variations it also increases long-term correlation for the more regular pitch cycles. This is performed by constructing high-correlated target signal and modifying low-correlated pitch cycles. The proposed pre-processor provides a more reliable pitch and voicing level estimation. This is demonstrated for the standard MELP coder. As a result, the pre-processor in combination with the standard MELP provides better perceptual speech quality in comparison with the MELP alone.

Another approach for enhancing the synthesised speech quality is through the use of postfiltering techniques. In Chapter 7, we propose a new postfilter technique based on the LP synthesis filter factorisation, which provides better perceptual speech quality than the conventional postfilters.

The summary of the work covered in this thesis is reviewed in chapter 8. This chapter also includes future works, which can be followed to either improve the performance of the presented pre-processor and postfilter methods or exploiting them for other applications in the area of speech coding and processing.

1.3 Original contributions

In summary, the original work covered in this thesis is as follows:

- Investigating the effect of irregular pitch variations on the speech parameter determinations, for instance pitch and voicing-level. This is addressed for the case of a WI and MELP coders.

- Design of the new pre-processor to modify irregular pitch variations such that pitch evolves smoothly during a voiced speech frame.
• Highly correlated target signal is used to enhance the low-correlated pitch cycles. This results in more regular speech such that the pitch and the voicing level can be estimated more accurately.

• Design of a new postfilter to provide better speech quality than the conventional postfilters due to more control over formant bandwidths and spectral valleys attenuation.

• Improving the performance of the standard MELP 2.4 Kb/s and ACELP without increasing the encoding rate.
Chapter 2

Digital Speech Coding

2.1 Introduction

Digital speech coding can be considered as an application of digital signal processing techniques to represent the speech signal. In the first step, it is required that the analogue speech signal to be sampled and digitised using an analogue-to-digital (A/D) converter. The output of the A/D converter is a linear Pulse-Code-Modulated (PCM) signal. Introduction of PCM in 1938 was the beginning of an era in speech communication, which rapidly progressed in the following years. Digital transmission of speech demonstrates several advantages over analogue transmission. The binary nature of the digital signal enables development of algorithms to recover it when the transmission channel is erroneous. Furthermore, it can be easily manipulated by software developed for different purposes such as compression and storage.

When a digital speech signal is transmitted, the bandwidth required is a function of the bit rate. Similarly, when a digital signal is stored, the bit rate determines the space required on the storage medium. Thus, system cost is often a function of the bit rate of digital speech signal. The purpose of speech coders is to reduce the bit rate of digital speech signals. Because of advances in mobile and satellite communications in the last two decades, the demand for coding the digital speech at lower rates with high quality was ever increasing. Such demand motivated wide spread research works which resulted in development of highly efficient coding algorithms amongst which Linear Prediction Coding (LPC), Mixed Excitation Linear Prediction (MELP), Multi-Band Excitation (MBE) and Waveform Interpolation (WI) can be seen as milestones. Using these algorithms high quality speech could be achieved at bit rates as low as 2.4 kb/s. The recent advances of the Very Large Scale Integration (VLSI) technology have allowed the implementation of highly complex speech compression algorithms in real time applications.

This chapter reviews the basic techniques used in low bit rate speech coding. In section 2.2 the speech production mechanism and speech properties are briefly described. Section 2.3 introduces the characteristics of the perceptual human auditory system. The different classes of speech coders are reviewed in section 2.4. In section 2.5, the required attributes to select a
speech coding system, are introduced. Subjective listening test is the most important measure to evaluate performance of a speech coding system. The popular listening tests are described in sections 2.6. In section 2.7, some of speech coder standards with their applications are introduced.

2.2 Speech production and properties

Removing predictable, redundant or pre-determined information from speech signal is used by many speech coders to reduce the bit rate consumptions. In order to search better speech coding algorithms, it is therefore important to have a good understanding of the production of human speech and the properties of speech signals. As shown in Fig. 2.1-a, human speech is produced when air is exhaled from the lungs, through the vocal folds and the vocal tract to the mouth opening. From the signal processing point of view, as shown in Fig. 2.1-b, this speech production mechanism can be modelled as an excitation signal exciting a time-varying filter (the vocal tract), which amplifies or attenuates certain sound frequencies in the excitation.

![Vocal Tract Cavity](image)

**Figure 2.1:** (a) The speech organs [8], Linear separable equivalent DSP model of the speech production mechanism.
The vocal tract consists of a combination of the throat, mouth, the tongue, the lip, and the nose. Since these change shape during generation of speech, the vocal tract is modelled as a time-varying system. The properties of the excitation signal highly depend on the type of speech sounds, either voiced or unvoiced. Examples of voiced speech are vowels (/a/, /i/, /o/, /u/) while fricatives such as /p/ and /k/ are examples of unvoiced sounds. The excitation for voiced speech is a quasi-periodic signal generated by the periodical abduction and adduction of the vocal folds where the airflow from the lungs is intercepted. Since the opening between the vocal folds is called the glottis, this excitation is sometimes referred as a glottal excitation. Generally, the vocal tract filter is considered as a linear system and therefore, not able to alter the periodicity of the glottal excitation. Hence, in time domain, voiced sounds are quasi-periodic in nature as well. For unvoiced speech, the vocal folds are widely open. The excitation is formed as the air is forced through a narrow constriction at some point in the vocal tract and creates a turbulence. In time domain, the unvoiced speech and its excitation signal both tend to be noise-like and lower in energy as compared to the voiced case. Figure 2.2a illustrates an example of both unvoiced and voiced speech segment in time domain. In spectral domain, due to the quasi-periodicity of voiced speech segments, its spectrum possesses a major harmonic line structure as depicted in Figure 2.2c. The spacing between the harmonics is called the fundamental frequency. The envelope of the spectrum, also known as the formant structure, is characterized by a set of peaks, each of which is called a formant. The formant structure is primarily attributed to the shape of the vocal tract. Thus, by moving the tongue, jaw or lips, the structure would be changed correspondingly. Also, the envelope falls at about -6 dB/octave due to the radiation from the lips and the nature of the glottal excitation [8]. Figure 2.2b shows the power spectrum of the unvoiced segment. As opposed to the voiced spectrum, there is relatively less useful spectral information embedded in an unvoiced segment. It does not have any distinctive harmonics and it is broadband and flat.

2.3 Human auditory perception

In order to reach maximal performance in a speech coder, it is also essential to take advantage of the human auditory system. The performance of a speech coder can be generally improved by exploiting the perceptual properties of the ear. This avoids major audible degradation in low bit rate speech coders. Auditory masking is one of the well-known properties of the auditory system [8]. It has a strong effect on the perceptibility of one signal in the presence of another. Noise is less likely to be heard at frequencies of strong speech energy (e.g., formants) and more likely to be heard at frequencies of low speech energy (e.g., valleys). Spectral masking is a popular technique that takes advantage of this perceptual limitation by
concentrating most of the noise (resulting from compression) in high-energy spectral regions where it is least audible. It was reported that humans recognise voiced and unvoiced sounds differently [8]. For voiced signals, the correct degree of periodicity and the temporal continuity in voiced segments [9][10][11] affect the human perception (although excessive periodicity would lead to reverberation and buzziness). In frequency domain, the amplitudes and the locations of the first three formants (usually below 3 kHz) and the spacing between the harmonics are important [12].

Figure 2.2: Time and frequency representations of a voiced and unvoiced speech segment. (a) A speech segment consists of an unvoiced and voiced segment in time domain. (b),(c) The speech spectrums for a 32 ms unvoiced and voiced segments starting at 10 ms and 150 ms, respectively.

For unvoiced signals, it has been shown in [13] that the unvoiced speech segments can be modelled by a noise-like signal with a similar spectral envelope without a drop in the perceived quality of the speech signal. In both voiced and unvoiced cases, the time envelope of the speech signal contributes to speech intelligibility and naturalness [14][15].
2.4 Classes of speech coders

Traditionally, speech-coding algorithms are divided into two distinct classes: waveform-approximating coders and parametric coders [16]. Waveform coders are not highly influenced by speech production models; as a result, they are simpler to implement. The objective with this class of coders is to produce a reconstructed signal that matches the original signal as accurately as possible. In other words, the reconstructed signal converges towards the original signal with increasing bit rate and thus Signal-to-Noise Ratio (SNR) is used as a quantitative measure for speech coder performance [16].

However, parametric coders rely on speech production models. They extract the model parameters from the speech signal and code them. The quality of these speech coders is limited due to the inaccuracy of the employed speech model, which results in the synthetic reconstructed signal. However, as seen in Fig. 2.3, they provide superior performance for lower bit rates. Recently, a new class of speech coders, named hybrid coders, is introduced which combines advantages of both waveform and parametric coders, and gives good quality speech at intermediate bit rate.

![Figure 2.3: Quality versus Bit Rate for different Speech Coding classes [16].](image)

2.4.1 Waveform-approximating coders

Since the goal of waveform coders is to match the original signal sample for sample, this class of coders is more robust to different types of input. Pulse Code Modulation (PCM) is the
Chapter 2. Digital Speech Coding

simplest type of coder, using a fixed quantiser for each sample of the speech signal. Given the
non-uniform distribution of speech sample amplitudes and the logarithmic sensitivity of the
human auditory system, a non-uniform quantiser yields better quality than a uniform quantiser
with the same bit rate. Thus, the CCITT standardized G.711 in 1988, a 64 kb/s logarithmic
PCM toll quality speech coder for telephone bandwidth speech [17].

In exchange for higher complexity, toll quality speech can be obtained at much lower bit
rates. With Adaptive Differential Pulse Code Modulation (ADPCM), the current speech
sample is predicted from previous speech samples; the error in the prediction is then
quantised. Both the predictor and the quantiser can be adapted to improve performance.
G.727, standardized in 1990, is an example of a toll quality ADPCM system, which operates
at 32 kb/s. Another possibility is to convert the speech signal into another domain by a
Discrete Cosine Transform (DCT) or another suitable transform [18]. The transformation
compacts the energy into a few coefficients, which can be quantised efficiently. In Adaptive
Transform Coding (ATC), the quantiser is adapted according to the characteristics of the
signal [18].

2.4.2 Parametric coders

The performance of parametric coders, also known as source coders or vocoders, highly
depends on accurate speech production models and speech excitation parameters. These
coders are typically designed for low bit rate applications (such as military or satellite
communications) and are primarily intended to maintain the intelligibility of the speech. Most
efficient parametric coders are based on Linear Predictive Coding (LPC). With LPC, each
frame of speech is modelled as the output of the LP filter representing the vocal tract, to an
excitation signal. Parameters for this filter and its excitation are then coded and transmitted.
Pitch and intensity parameters are typically used to code the excitation and various filter
representations are used for the linear system [19]. Communications quality speech can
currently be achieved at rates below 2 kb/s with vocoders based on LPC [8].

2.4.3 Hybrid coders

The speech quality of waveform coders drop rapidly for bit rates lower than 16 kb/s. On the
other hand, there is a negligible improvement in the quality of parametric coders at rates
above 4 kb/s. Hybrid coders are used to bridge this gap, providing good quality at medium bit
rate. These coders tend to be more computationally demanding. However, due to advances in
Digital Signal Processing (DSP) chip technology in the recent years, the computational
complexity has not been an obstacle in the deployment of hybrid coders. Considering the
demand to reduce the bit rate and to improve the perceptual quality of the synthesized speech,
future generations of the speech coders are likely to belong to this category. Assuming an
error free transmission medium, the perceptual quality of the synthesized speech depends
upon the following factors:
- Accuracy of the synthesis model and its estimated parameters.
- Accuracy of the excitation estimate.
- Approximations introduced by the encoding process (quantisation errors).
Code-Excitation Linear Prediction (CELP) coders have received a lot of attention recently
and are the basis for most speech coding algorithms currently used in wireless technology. In
CELP coders, standard LPC analysis is used to obtain the excitation signal. Pitch modelling is
used to efficiently code the excitation signal. Many variations of CELP coders have been
standardized, G.729 operating at 8 kbps [5], G.728 a low delay coder operating at 16 kbps
[20], and all the digital mobile telephony encoding standards including [21][22][23] GSM,IS-
54,IS-95, and IS-136.

2.5 Speech-coder attributes

Since the extensive research has been done in the area of speech coding, there is a variety of
existing speech coding algorithms. In selecting a speech coding system, the following
attributes are typically considered:

Complexity: This includes the memory requirements and computational complexity of the
algorithm. In virtually all applications, real-time coding and decoding of speech is required.
To reduce costs and minimise power consumption, speech-coding algorithms are usually
implemented on DSP chips. Thus, the computational requirements of speech coding
algorithms have tracked the performance of the hardware. Speed is the most commonly
measured as the number of Millions of instructions Per Second (MIPS) necessary for the real
time implementation of the speech-coding algorithm. 16-bit fixed-point DSPs are most
commonly used for low cost implementations. Thus, complexity is often specified in terms of
fixed-point MIPS and the number of 16-bit words of random access memory (RAM) needed
for an implementation.

Delay: The overall one-way delay of a speech coding system is the time between when the
talker emits a sound and when the listener first hears it. This delay contains the algorithmic
delay, the computational delay, the multiplexing delay and the transmission delay. The
algorithmic delay is the sum of the look-ahead and buffer length used in the speech-coding
algorithm. The computational delay is associated with the time required for processing the
speech. In many transmission systems, a block of bits is assembled by encoder and decoder for channel coding purposes. The delay associated with these assembling operations is called multiplexing delay. Finally, the transmission delay is a result of the finite speed of electromagnetic waves in any given medium. Under these circumstances, a one-way delay of 150 ms is perceivable during highly interactive conversations [24]. Thus, the speech coder must be chosen accordingly, with low-delay coders being employed in environments.

**Bit rate:** The bandwidth available in a system determines the upper limit for the bit rate of the speech coder. However, a system designer can select from fixed-rate or Variable-Bit-Rate coders (VBR). A VBR coder allocates the minimum number of bits required to maintain sufficient speech quality for a given segment of speech. However, they are more complex in design compared to the fixed rate coders. VBR coders are particularly advantageous for voice storage, Code Division Multiple Access (CDMA) [25] wireless networks, packet switched networks, and statistical multiplexing of speech for multi-channel communications.

**Quality:** The quality of a speech coder can be evaluated using extensive testing with human subjects. This is a very tedious process and thus objective distortion measures are frequently used to estimate the subjective quality. The following categories are commonly used to compare the quality of speech coders:

1. Commentary quality describes wide-bandwidth speech with no perceptible degradations;
2. toll quality speech refers to the type of speech obtained over the public switched telephone network;
3. communications quality speech is completely intelligible but with noticeable distortion; and,
4. synthetic quality speech is characterized by its 'machine-like' nature, lacking speaker identification and being slightly unintelligible.

**Robustness:** In some applications, robustness to background noise and/or channel errors is essential. Typically, the speech being coded is distorted by various kinds of acoustic noise in urban environments; this noise can be quite excessive for cellular communications. The speech coder should still maintain its performance under these circumstances. Random or burst errors are frequently encountered in wireless systems with limited bandwidth. It is best to integrate the robustness against random channel errors in the design of the quantisers of the speech coder [16]. In the burst-error class, error detection schemes are used to classify each frame of received bits as usable or not usable. If a frame of received bits is deemed unusable, the decoder enters into a special mode and the signal characteristic is made to converge slowly to a white noise signal.

**Bandwidth:** Speech signals in the public switched telephone network are band-limited to 300 Hz - 3400 Hz. Most speech coders use a sampling rate of 8 kHz, providing a maximum signal bandwidth of 4 kHz. However, to achieve higher quality speech for video conferencing applications, larger signal bandwidths must be used. Wideband speech signals have a
bandwidth of 500-7000 Hz and are sampled at 16 kHz. The wider bandwidth improves naturalness and intelligibility of speech signal.

2.6 Performance evaluation

One of the major difficulties in designing and testing various speech coders is the lack of an objective quality measure to reflect the perceptual synthesised speech quality. The most commonly used objective criteria such as SNR, segmental SNR, log spectral distance [26] are sensitive to gain variations and misalignment between the original and coded speech. In addition, low bit rate speech coders do not preserve the waveform similarity and the SNR based quality measures become meaningless. A number of objective methods based on human auditory perception models have been proposed (Schroeder et al. 1979 [27], Wang et al. 1992 [28], Paillard et al. 1992 [29], ITU-T Recommendation P.861 [30]), but none has yet eliminated the necessity of subjective testing.

The most commonly performed subjective tests are Absolute Category Rating (ACR) tests [31][32] of which one example is the Mean Opinion Score (MOS) test [33]. In the MOS test a number of listeners are asked to evaluate the quality of recorded speech according to a five-level scale. For narrow-band speech, a score of 4 to 4.5 implies toll quality and a score between 3.5 and 4 indicates communications quality. Scores below 3.5 mean that the reconstructed speech is of poor quality; synthetic speech often scores in the range 2.5 to 3.5. The MOS scores can differ from one test to another significantly, often due to cultural and/or linguistic biases, and therefore are not an absolute comparison between coders.

Table 2.1: Quality rating scale for an A/B comparative test.

<table>
<thead>
<tr>
<th>Description</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>A better than B</td>
<td>-2</td>
</tr>
<tr>
<td>A Slightly better than B</td>
<td>-1</td>
</tr>
<tr>
<td>No Preference</td>
<td>0</td>
</tr>
<tr>
<td>B Slightly better than A</td>
<td>+1</td>
</tr>
<tr>
<td>B better than A</td>
<td>+2</td>
</tr>
</tbody>
</table>

A versus B test is a comparison test, in which the listeners identify the quality if the second stimulus relative to the first using a two sided rating scale. Table 2.1 gives possible responses after two stimuli have been presented.
Subjective testing in general is time consuming and therefore expensive. Many proposed coders have not been subjected to rigid testing and the reported results are difficult to calibrate.

2.7 Speech coding standards

In the past decade, more speech coding standards have been created than in all the previous years. The reasons for this are the maturity of speech coding technology and the need to satisfy a growing demand for new speech communication technologies. This also allows equipment manufacturers to combine their research efforts, and made competition between them possible, thereby lowering the prices. The standardisation procedure identifies and defines the speech coding requirements for the next generation communication products. This has become the main driving force in speech coding research. Table 2.2 shows the major speech coding standards adopted in the last 30 years.

2.8 Conclusions

In this chapter three classes of speech coders (waveform coders, vocoders and hybrid coders) were reviewed. Among these classes, hybrid speech coders produce high speech quality at low bit rates and therefore they are subject of more interest. The main factors, which needed to be addressed during the design of a speech coder, have been presented. These factors are generally contradictory, and therefore design trade-offs should be made when tailoring a speech coder to a given application. Subjective listening tests are known as the best criterion to evaluate the quality of speech synthesised by a speech coder. Typically, MOS and A vs. B tests have been described. The final section of the chapter presents a review of the existing telephone band speech coding standards.
Table 2.2: Major Speech Coding Standards adopted in the last 30 years.

<table>
<thead>
<tr>
<th>Bit Rate (Kb/s)</th>
<th>Standard</th>
<th>Coder Type</th>
<th>Application</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>CCITT G.711</td>
<td>Companded PCM</td>
<td>PSTN</td>
<td>1972</td>
</tr>
<tr>
<td>16</td>
<td>CCITT G.721</td>
<td>ADPCM</td>
<td>PSTN</td>
<td>1984</td>
</tr>
<tr>
<td>2.4</td>
<td>US-FS1015</td>
<td>LPC</td>
<td>Secure Voice</td>
<td>1984</td>
</tr>
<tr>
<td>24</td>
<td>CCITT G.723</td>
<td>ADPCM</td>
<td>DCME</td>
<td>1986</td>
</tr>
<tr>
<td>13</td>
<td>ETSI-GSM</td>
<td>RPE-LTP</td>
<td>Digital cellular telephony</td>
<td>1987</td>
</tr>
<tr>
<td>32</td>
<td>CCITT G.726</td>
<td>ADPCM</td>
<td>DCME</td>
<td>1988</td>
</tr>
<tr>
<td>64, 56, 48</td>
<td>CCITT G.722</td>
<td>Subband Coder</td>
<td>ISDN (wideband)</td>
<td>1988</td>
</tr>
<tr>
<td>7.95</td>
<td>TIA IS54</td>
<td>VSELP</td>
<td>TDMA digital telephony, North America</td>
<td>1989</td>
</tr>
<tr>
<td>6.7</td>
<td>Full-rate PDC</td>
<td>VSELP</td>
<td>Japanese digital cellular telephony</td>
<td>1990</td>
</tr>
<tr>
<td>4.15</td>
<td>Inmarsat</td>
<td>IMBE</td>
<td>Mobile telephony via satellite</td>
<td>1990</td>
</tr>
<tr>
<td>40</td>
<td>CCITT G.727</td>
<td>ADPCM</td>
<td>Packet circuit multiplex equipment</td>
<td>1990</td>
</tr>
<tr>
<td>4.8</td>
<td>FS1016</td>
<td>CELP</td>
<td>Secure telephone</td>
<td>1991</td>
</tr>
<tr>
<td>16</td>
<td>ITU-T G.728</td>
<td>LD-CELP</td>
<td>Videophone, Circuit multiplication equipment</td>
<td>1992</td>
</tr>
<tr>
<td>8.5, 4, 2, 0.8</td>
<td>TIA IS96</td>
<td>QCELP</td>
<td>CDMA digital cellular telephony</td>
<td>1993</td>
</tr>
<tr>
<td>3.45</td>
<td>Half-rate PDC</td>
<td>VSELP</td>
<td>Japanese TDMA personal digital cellular</td>
<td>1993</td>
</tr>
<tr>
<td>5.6</td>
<td>Half-rate TCH-HS</td>
<td>VCELP</td>
<td>GSM cellular system</td>
<td>1994</td>
</tr>
<tr>
<td>5.3, 6.3</td>
<td>ITU-T G.723.1</td>
<td>MP-LPQ, ACELP</td>
<td>Video conferencing</td>
<td>1995</td>
</tr>
<tr>
<td>8</td>
<td>G. 729A</td>
<td>CS-ACELP</td>
<td>Telephony over packet networks</td>
<td>1998</td>
</tr>
<tr>
<td>6.6-23.85</td>
<td>ITU-T G.722.2</td>
<td>ACELP</td>
<td>3rd generation mobile telephony</td>
<td>2001</td>
</tr>
<tr>
<td></td>
<td>GSM-AMR-WB</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chapter 3

Basic Tools in Speech Coding

3.1 Introduction

In order to compress speech at low bit rates, the speech characteristics are required to be considered. Linear Prediction Coding (LPC) is a widely used technique to present the short-term correlation characteristic of speech signal. The LPC exploits the redundancies of the speech signal by modelling the speech signal as a linear prediction filter, excited by a signal called the excitation signal. The excitation signal is also called the residual signal. In order to remove redundant information from speech more features should be exploited. Speech classification as voiced and unvoiced has been widely employed in low bit rate coding algorithms. In this approach each speech frame is identified as either voiced or unvoiced. Periodicity of the speech signal plays the main role on the classification. The voiced frames are usually identified by their quasi-periodic characteristics, whereas the unvoiced segments are of noise-like nature. Furthermore, the voiced segments contain higher energy. These and the other features are usually employed to classify the input speech frame as voiced or unvoiced. In [6], it is shown that considering speech signal spectrum as a mixture of voiced and unvoiced spectrums improves the perceptual speech quality. Therefore, some speech coders such as MBE, MELP divide the speech spectrum into a number of bands and for each band perform voiced/unvoiced classification. A voicing level is used to present the quasi-periodicity of voiced segments.

The periodicity is an important characteristic of voiced speech segment. This feature can be well described by its period known as pitch period in time domain or fundamental frequency in frequency domain and plays a major role in low bit rate speech coders.

This chapter discusses these issues as follows: In section 3.2 the LP coding is discussed. This contains calculation and different representation of LP coefficients. Section 3.3 details different techniques employed for pitch estimation. Binary V/UV classification and voicing-level estimation are discussed in sections 3.4 and 3.5, respectively.
Chapter 3. Basic Tools in Speech Coding

3.2 Linear prediction in low bit rate coding

LPC analysis is an essential component in most speech coding algorithms. It removes the short-term correlation (redundancy) in a speech signal by employing a time-varying linear prediction (LP) filter. The filter coefficients are known as LP coefficients and the filter output is called an excitation signal or residual signal. These LP coefficients characterise the spectral envelope of the speech signal approximating the human vocal tract while the residual describes the glottal excitation. LPC analysis decomposes speech into two highly independent components, vocal tract parameters (LP coefficients) and the glottal excitation (LP excitation). These two components have very different quantisation requirements. As a result, separate analysis and quantisation scheme can be applied to each to enhance coding efficiency.

In the most general case, LPC consists of a pole-zero model (also known as an autoregressive moving average, or ARMA, model) for \( H(z) \) given by:

\[
H(z) = \frac{S(z)}{E(z)} = G \cdot \frac{1 + \sum_{i=1}^{q} b_i z^{-i}}{1 - \sum_{j=1}^{p} a_j z^{-j}} = \frac{B(z)}{A(z)}
\]

(3.1)

Where

\[
A(z) = 1 - \sum_{j=1}^{p} a_j z^{-j}
\]

(3.2)

\[
B(z) = 1 - \sum_{i=1}^{q} b_i z^{-i}
\]

and \( G \) is the filter gain.

Thus, the speech sample \( s[n] \) is a linear combination of the \( p \) previous output samples \( s[n-1], \ldots, s[n-p] \) and the \( q+1 \) previous input samples \( e[n], \ldots, e[n-q] \). This is expressed mathematically in the following difference equation:

\[
s(n) = \sum_{j=1}^{p} a_j s(n-j) + G \sum_{i=0}^{q} b_i e(n-i) \quad b_0 = 1
\]

(3.3)

Nasals and fricatives, which contain spectral nulls, can be modelled accurately with the zeros in this ARMA model whereas the poles are crucial in representing the spectral resonances. However, due to its analytical simplicity and to avoid solving a set of non-linear equations [34], all-pole models (also known as autoregressive, or AR, models) are extensively used in
real-time systems with constraints on computational complexity. Using an AR model for 

\[ H(z) \], Eq. (3.3) can be written as:

\[ s(n) = G \cdot e(n) + \sum_{j=1}^{p} a_j s(n - j) \]  

(3.4)

Equation 3.4 states that the current sample of speech can be obtained by summing the 
weighted current excitation sample and \( p \) previous speech samples. The scaling term \( G \) is 
usually considered to be unity. The objective is now to calculate the LP coefficients \( a_j \) such 
that the best prediction for the speech segment under analysis is obtained. To achieve this, the 
predicted and error signals \( \hat{s} \) and \( e \) are defined as:

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

(3.5)

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

The optimum prediction coefficients can be obtained by minimising the estimation error \( E \) 
given by Equation 3.6.

\[ E = \sum_{n} e^2(n) = \sum_{n} [s(n) - \hat{s}(n)]^2 = \sum_{n} [s(n) - \sum_{j=1}^{p} a_j s(n - j)]^2 \]  

(3.6)

Where \( N \) is the length of the analysis frame. \( E \) can be minimised by setting its partial 
derivatives to zero:

\[ \frac{\partial E}{\partial a_j} = 0 \]  

(3.7)

This equation can be rewritten as [1]:

\[ \sum a_j \phi(i, j) = \phi(i, 0) \]  

(3.8)

Where

\[ \phi(i, j) = \sum_{n} s(n - i)s(n - j) \]  

(3.9)
Thus, to obtain the LP coefficients, the values of $\phi(i, j)$ are calculated followed by solving Equation 3.8. Autocorrelation and covariance methods are the most common approaches to calculate the LP coefficients. In the following section, only the autocorrelation method is detailed and a description of covariance and other method can be found in [1].

### 3.2.1 Autocorrelation method

In this method, the speech signal $s(n)$ is first multiplied by an analysis window $w(n)$ of finite length $N$ to obtain a windowed speech segment $s_w(n)$.

$$s_w(n) = s(n)w(n)$$ (3.10)

The simplest analysis window is a rectangular window. A rectangular window has an abrupt discontinuity at the edge in the time domain. As a result there are large side lobes and undesirable ringing effects [19] in the frequency domain representation of the rectangular window. To discard the large oscillations, it is required that to employ a window without abrupt discontinuities in the time domain. Thus, the window $w(n)$ is typically chosen to be a Hamming window to minimise the side-lobe energy and is defined as:

$$w(n) = \begin{cases} 0.54 - 0.46 \cos \left( \frac{2\pi n}{N-1} \right) & \text{for } 0 \leq n \leq N \\ 0, & \text{Otherwise} \end{cases}$$ (3.11)

The above assumption (samples being zero outside the analysis frame) converts the correlation function in equation 3.9 to:

$$\phi(i, j) = \sum_{n=0}^{N+p-1} s_w(n-i)s_w(n-j) \quad 1 \leq i \leq p \quad , \quad 1 \leq j \leq p$$ (3.12)

In [1], it is shown that the equation 3.12 can be rewritten as autocorrelation function given in 3.13.

$$\phi(i, j) = R(|i-j|) \quad 1 \leq i \leq p \quad , \quad 1 \leq j \leq p$$ (3.13)

Where the autocorrelation function $R$ is defined as:
\[ R(k) = \sum_{n=0}^{N-1-k} s_w(n)s_w(n + k) \]  

(3.14)

in this case equation 3.8 can be rewritten as:

\[ \sum_{j=1}^{p} a_j R(|i-j|) = R(i), \quad 1 \leq i \leq p \]  

(3.15)

In the matrix form the above equation can be rewritten as:

\[
\begin{bmatrix}
  R(0) & R(1) & \cdots & R(p-1) \\
  R(1) & \ddots & \ddots & \vdots \\
  \vdots & \ddots & \ddots & \vdots \\
  R(p-1) & \cdots & R(0)
\end{bmatrix}
\begin{bmatrix}
  a_1 \\
  a_2 \\
  \vdots \\
  a_p
\end{bmatrix}
= 
\begin{bmatrix}
  R(1) \\
  R(2) \\
  \vdots \\
  R(p)
\end{bmatrix}
\]

(3.16)

Since the matrix in Eq. (3.16) has a Toeplitz structure, the coefficients \( \{a_k\} \) can be obtained by Levinson-Durbin recursion [26]. In addition, the Toeplitz structure can guarantee the poles of the resulting LP synthesis filter to be located inside the unit circle and hence, the filter stability is always satisfied [1].

Figure 3.1: Speech spectrum (solid lines) with the corresponding spectral envelope (dash lines).
3.2.2 The LP model properties

LP analysis results in approximate estimate of the speech spectrum envelope shape. Since this analysis is based on an all pole filter for modelling vocal tract, the resulting estimation is more accurate at the peaks of amplitude spectrum called *formants*. Figure 3.1 demonstrates the spectra of a voiced speech frame and resulting LP filter with $p = 10$.

The residual signal can be computed by filtering speech signal through the inverse LP filter. It has a significantly lower dynamic range compared to the original speech signal and therefore can be encoded using less number of bits. In the frequency domain, inverse filtering is equivalent to removing the spectral envelope and therefore the residual signal has a rather flat spectrum.

Figures 3.2 and 3.3 illustrate the performance of inverse filtering on a segment of voiced speech. Lower dynamic range and flat spectrum of the residual signal can be observed in Figures 3.2 (b) and 3.3 (b), respectively.

The LP analysis is performed on a segment of speech under this assumption that the speech signal is stationary. This limits the length of the segment to 20-30 ms which is the common frame lengths in low bit rate speech coding algorithms. In each frame, up to five formants are usually considered [19]. Since each formant can be identified with at least a pair of conjugate poles [35], the order of the LP filter should be equal or greater than ten. However, in order to estimate the spectrum more accurately, most speech coders use $10^{th}$ order filter in the analysis [19]. The performance of the LP model is usually evaluated by a parameter called *prediction gain*, which is the ratio of the original speech energy and the LP residual energy. A high prediction gain means that the LP model has been successful in removing most of the short-term correlation from the speech.

3.2.3 Representations of the LP filter

In speech-coding applications, it is necessary to quantise the LP filter coefficients, $\{a_k\}$, with as little distortion as possible. Also, it is required that the all-pole filter remains stable after quantisation of the LP coefficients. Direct scalar quantisation of the LPC coefficients [16] is usually not performed as small quantisation errors in the individual LPC coefficients can produce relatively large spectral errors and can also result in instability of the synthesis filter. As a result, it is necessary to use a relatively large number of bits to perform transparent quantisation of LPC parameters by quantisation of the LP coefficients $\{a_k\}$ themselves.
Figure 3.2: Speech signal (a) with the corresponding LP residual (b).

Figure 3.3: Speech spectrum with spectral envelope (a) and the resulting residual spectrum (b).
Thus, other superior parametric representations have been formulated. Some of these representations are: Reflection coefficients or PARCORs [8], Log-Area Ratios (LARs) [36], Arcsine of Reflection Coefficients (ASRC) [8] and Line Spectral Frequencies (LSF) or equivalently, Line Spectral Pairs (LSP) [8]. Despite its computational complexity, LSF is widely used as it provides easy stability checking procedure, spectral manipulations and perceptual quantisation for coding [37]. The details of LSF representation are described in the following section.

3.2.3.1 Line Spectral Frequencies (LSF)

The LSFs or LSPs were introduced by Itakura [37]. They present the phase angles of an ordered set of poles on the unit circle that describes the spectral shape of the LP analysis filter $A(z)$ defined in Eq. (3.2). Conversion of the LP coefficients $\{a_k\}$ to the LSF domain relies on $A(z)$. Given $A(z)$, the corresponding LSFs are defined to be zeros of the polynomials $P(z)$ and $Q(z)$ defined as:

$$P(z) = A(z) + z^{-(p+1)}A(z^{-1})$$

$$Q(z) = A(z) - z^{-(p+1)}A(z^{-1})$$

It directly follows that:

$$A(z) = \frac{1}{2} [P(z) + Q(z)]$$

Soong and Juang [38] have shown that if $A(z)$ is minimum phase (corresponding to a stable $H(z)$), all the roots of $P(z)$ and $Q(z)$ lie on the unit circle, alternating between the two polynomials with increasing frequency, $\omega$. The roots occur in complex-conjugate pairs and hence there are $p$ LSFs lying between 0 and $\pi$. The process produces two fixed zeros at $\omega=0$ and $\omega=\pi$ which can be ignored. It has also been shown [8] that if the $p$ line spectral frequencies $\omega_i$ have an ascending ordering property and are unique, then the LP analysis filter $A(z)$ is guaranteed to have minimum phase (stable corresponding synthesis filter):

$$0 < \omega_1 < \omega_2 < \cdots < \omega_p < \pi \quad \text{[radians/sec]}. \tag{3.19}$$

If $\omega_0$ and $\omega_p$ represent the angles of roots of $P(z)$ and $Q(z)$, the LSPs are expressed as:

$$LSP(2i) = \cos(\omega_i) \quad i = 0, 1, \cdots, p-1$$

$$LSP(2i+1) = \cos(\omega_0)$$

$$LSP(2i+1) = \cos(\omega_p)$$
The LSFs are the frequencies of the above roots and thus can be obtained by:

\[ LSF(i) = \frac{\cos^{-1}[LSP(i)]}{2\pi T} \]  

(3.21)

In which \( T \) is the sampling rate. Several approaches for solving for the roots of \( P(z) \) and \( Q(z) \) have been proposed such as the Ratio Filter method [39], the DFT method [40] and the Chebyshev series method [41].

The intra-frame correlation of the LSFs is the most advantageous characteristic of these parameters. This characteristic can be employed efficiently in quantisation of the LP coefficients. The LSFs are also related to the peaks of magnitude spectrum, i.e. the formants. If two neighbouring LSFs are close in frequency, it is likely that they correspond to a narrow bandwidth spectral resonance in that frequency region. In addition, they usually contribute to the overall tilt of the spectrum. Figure 3.4 shows the relation between formants and LSFs for a speech frame.

![Figure 3.4: Relation of LSFs and formants](image-url)
3.2.4 Interpolation of linear prediction parameters

Linear prediction based speech coders describe the envelope of the speech spectrum using an autoregressive model. The LP based models may vary considerably in consecutive frames or during a frame especially transition frames. In order to follow the changes in spectra or to smooth the spectral transitions, the LP coefficients should be updated more frequently, which amounts to reduce the frame length. However, this increases the encoding bit rate. To avoid the augmentation in bit rate, interpolation of LP coefficients can be used in the consecutive analysis frame. Thus, undesired transients due to a large change in the LP based model at adjacent frames are avoided in the synthesised speech signal. Usually, a frame is divided into several equally spaced time intervals called subframes, and the interpolation is performed at these subframes. Since stability of the synthesis filter resulting from the interpolated LSFs can easily be checked, most speech coders perform the interpolation of LP coefficients in LSF domain. A linear or piecewise linear function is employed to perform the interpolation between the adjacent LSFs. In order to study the effect of the LSF interpolation on the prediction gain, we use a linear function to update LP coefficients for each subframe and also change the subframe length. The speech signal is firstly windowed using a hamming window centred at the end of current frame and then the LP coefficients are calculated using autocorrelation method and converted to LSF's. The length of input frame and hamming widow are 20ms and 25 ms, respectively. If the number of subframes is M, the interpolated LSF of subframe \( k^{th} \) is given by:

\[
\Omega_{m,k} = \frac{M - k}{M} \Omega_{m-1} + \frac{k}{M} \Omega_{m}
\]  

(3.22)

Where \( \Omega_{m-1} \) and \( \Omega_{m} \) are the estimated LSFs of previous and current frame. Next, the interpolated LSF's are converted to the LP coefficients and the residual signal is computed using the resulting inverse LP filtering.

<table>
<thead>
<tr>
<th>Prediction Gain type</th>
<th>Subframe Length (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20</td>
</tr>
<tr>
<td>Short-term</td>
<td>11.31</td>
</tr>
<tr>
<td>Long-term</td>
<td>4.82</td>
</tr>
</tbody>
</table>
Short-term and long-term prediction gains [42] are then calculated using the resulting residual and input speech signals. Results are shown in Table 3.1. A high average prediction gain means that the LP filtering more accurately reflects the effect of the vocal tract so that the residual may be closer to a true excitation [42]. In addition, the prediction gain is almost saturated for subframes shorter than 4 ms.

3.3 Pitch determination

Most low bit rate speech coders invest in pitch estimation for efficient encoding of voiced speech. Consequently, the subject of pitch determination has attracted much research in the area of low bit rate coding. In spite of numerous research works on the development of pitch-determination algorithms, the issue of accurate pitch for all speech materials has still remained unsolved. This is mainly contributed to the weak stationary assumption made within the analysis window. In other words, these methods try to find the pitch period for a signal which is fundamentally quasi-periodic. The result is inaccuracy in the calculated pitch period. Nevertheless, since employing the pitch period is an essential tool in low bit rate speech coding, these methods have found wide application in speech compression algorithms. Therefore, there is scope for improving the efficiency of the current pitch determination algorithms. Pitch determination algorithms can be performed in time or frequency domain.

3.3.1 Time domain algorithms

In time domain pitch determination algorithms, similarity between a segment of voiced speech with its shifted version is used as a measure. The shift value, which maximises the similarity between the two signals, is declared as the pitch period. On the other hand, the difference between the two signals can also be used to indicate the similarity of these signals. In this case, the shift value leading to the minimum of the difference function is considered as the pitch period. The autocorrelation and average magnitude difference function (AMDF) are the most common measures used to specify the pitch period in time domain.

3.3.1.1 Autocorrelation method

Maximisation of the correlation between the original speech and its shifted version is the most popular technique used for searching the pitch value. In autocorrelation method, this is performed by computing the normalised autocorrelation of the signal for a given lag value \( \tau \) as given by Equation (3.23).
Chapter 3. Basic Tools in Speech Coding

3.3.1.2 AMDF method

The autocorrelation method is a straightforward and fairly accurate method for pitch determination.
Figure 3.6: The resulting AMDF for first frame of speech depicted in Fig. 3.5 (a).

However, it requires a considerable amount of computational work, which may prohibit its application where a limited computational capacity is available. In these cases, the AMDF technique is a suitable alternative. The AMDF introduces another error function as:

$$E(\tau) = \frac{1}{N} \sum_{n=0}^{N-1} |s(n) - s(n + \tau)|$$  \hspace{1cm} (3.25)

Figure 3.6 shows the AMDF technique applied to the same speech material of Figure 3.5 (a). The main advantage of the AMDF function is that it only requires additions and subtractions, making it very suitable for hardware implementation. However, current DSP processors normally offer a one-cycle multiply-add instruction, and therefore making this irrelevant. The AMDF technique can detect the multiple of the pitch instead of the pitch, as autocorrelation method and so a sub-multiple algorithm is required to alleviate this problem. In addition, the performance of the AMDF function is relatively poor whenever there is a sudden change in the amplitude and/or pitch within a voiced frame.

### 3.3.2 Frequency domain algorithms

Fundamental frequency can be determined by analysing the voiced frame in the frequency domain. The main feature of the voiced speech spectrum is the harmonic structure, i.e., the
distance between successive harmonics being the pitch frequency. This voiced speech spectrum characteristic is relied upon in all frequency domain based pitch detection algorithms. Some of these techniques are discussed in the following sections.

3.3.2.1 Harmonic peak detection

In this method, all the peak frequencies are firstly detected and the fundamental frequency is calculated by finding the spacing between these frequencies. The spacing between the harmonics can be found by adding the samples of the amplitude spectrum at equal trial frequencies and detecting the pitch frequency corresponding to the highest sum. In order to apply this method, the spectrum of speech signal is sampled using a comb function $C(\omega, \omega_0)$, which is given by Equation 3.26.

$$
C(\omega, \omega_0) = \sum_{k=-\infty}^{\infty} \delta(\omega - k\omega_0)
$$ (3.26)
Where $\Omega_m$ is the maximum frequency in the speech spectrum. The sum of the sampled spectrum at equal intervals of $\omega_s$ is then obtained as:

$$A(\omega_0) = \frac{\Omega_0}{\Omega_m} \sum_{k=1}^{\Omega_0/\Omega_m} S(k\omega_0)C(k\omega_0)$$  \hspace{1cm} (3.27)

Where $S(\omega)$ represents the frequency spectrum. The pitch frequency $\omega_0$ is the frequency which maximises function $A(.)$ in Eq. 3.27. Figure 3.7 (a) shows function $A(.)$ for $\omega = 50 - 400$ Hz and Figure 3.7 (b) indicates the case where the fundamental frequency is selected by this method.

### 3.3.2.2 Synthetic spectral matching

This method proposed by Griffin [6] is based on the similarity of the speech spectrum to a synthetic voiced spectrum generated for each candidate of fundamental frequency, $\omega_0$. An error criterion is defined as:

$$E(\omega_0) = \sum_{m=0}^{N_f-1} |S_w(m) - \hat{S}_w(m,\omega_0)|^2$$  \hspace{1cm} (3.28)

The fundamental frequency $\omega_0$ minimises $E(\omega_0)$. $S(m)$ and $\hat{S}_w(m,\omega_0)$ are the windowed speech and synthetic spectrum signals, respectively. In order to reduce the computational complexity, the error minimisation process is divided into non-overlapping frequency bands centred at the harmonic frequencies of $\omega_0$, and when synthesising $\hat{S}_w(m,\omega_0)$ for a particular harmonic, the spectral leakage from the other harmonics is assumed to be negligible.

The simplified error criterion for the $k^{th}$ harmonic is given by:

$$E_k = \sum_{m=n_k}^{b_k} |S_w(m) - A_k(\omega_0)W(\frac{2\pi}{N_f}m - k\omega_0)|^2 \hspace{1cm} \text{for } k = 1, 2, \cdots, K(\omega_0)$$  \hspace{1cm} (3.29)

Where $N_f$ is the length of FFT and $K(\omega_0) = \left\lfloor \frac{\pi}{\omega_0} \right\rfloor$. 
The harmonic boundaries $a_k$ and $b_k$ are defined as:

$$a_k = \left\lfloor \frac{N_f}{2\pi}(k - \frac{1}{2})\omega_0 \right\rfloor \quad b_k = \left\lfloor \frac{N_f}{2\pi}(k + \frac{1}{2})\omega_0 \right\rfloor$$

(3.30)

and $A_k(\omega_0)$ is given by:

$$A_k(\omega_0) = \frac{\sum_{m=a_k}^{b_k} S_w(m) W\left(\frac{2\pi}{N_f}m - k\omega_0\right)}{\sum_{m=a_k}^{b_k} \left| W\left(\frac{2\pi}{N_f}m - k\omega_0\right) \right|^2}$$

(3.31)

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

$$E'(\omega_0) = \sum_{k=1}^{K(\omega_0)} E_k$$

(3.32)
Figure 3.8 shows an original speech spectrum and the synthesised spectrum for the selected $\omega_0$. A small deviation in $\omega_0$ results in large deviation at the higher harmonics and large error values, $E(\omega_0)$. Therefore the fundamental frequency can be determined with high precision. However, since a full search at fractional resolutions is computationally intensive, the synthetic spectral matching is normally used for refining the initial estimated pitch obtained through another method.

### 3.3.3 Comparison of different methods

Among the various pitch determination algorithms, the time domain technique, which relies on the autocorrelation method is the most commonly used method. The reason is that its main arithmetical requirement simply involves multiply and adds which can be found as a single instruction in most of the new generations of the DSP chips. On the other hand, all frequency domain algorithms require a Fourier transform of speech, which is more computationally demanding. Although developments of fast algorithms such as FFT and DSP processors mean that more efficient transform methods are available, the frequency domain algorithms are still more complex than the time domain techniques. However, frequency domain algorithms are better suited to the coding schemes, which involve Fourier transform for other functions in the coder. The accuracy of the time domain methods depends on the sampling rate. This may be deduced by:

$$\frac{(f_s / \hat{f}_s)}{P} \times 100$$

Where $f_s$ and $\hat{f}_s$ are Nyquist and upsampling rates, respectively and $P$ is the pitch period in samples. For instance, assuming an 8 kHz sampling frequency, the accuracy of these methods varies from 0.6 to 5 percent for fundamental frequencies of 50 and 400 Hz, respectively. For a higher accuracy, the speech signal may be up-sampled. On the other hand, the accuracy of the frequency domain algorithms depends on the FFT size and also on the method employed.

Thus, employing a time-domain technique for initially pitch estimation and refining the estimated pitch using a frequency-domain algorithm seems to be a good strategy for pitch estimation. For instance, in the multiband coding scheme, the autocorrelation method is used to initially predict the pitch period and a spectrum matching is used to increase the accuracy of the initial pitch estimation.
3.3.4 Pitch interpolation

In most low bit rate speech coders, the pitch is estimated only once per frame and, when needed, the intermediate pitch values are obtained by interpolation between two adjacent pitch values. This can be performed by using a conventional linear interpolation technique. If \( P(n_1) \) and \( P(n_2) \) are the pitch values of previous and current frames, then the pitch can be linearly interpolated by:

\[
P(n) = \frac{(n_2 - n)P(n_1) + (n - n_1)P(n_2)}{n_2 - n_1} \quad n_1 \leq n \leq n_2
\]  

(3.34)

where \( n_2 - n_1 = N \) is the frame length.

Nevertheless, in natural speech especially at the beginning and end of a voiced segment, the pitch value occasionally doubles/triples/halves [44]. In addition, pitch estimators suffer from frequent errors in which the estimated pitch is an integer multiple of the actual pitch. In [44], Kleijn shows if no special attention is paid and the linear interpolation is performed across these changes, the reconstructed speech by WI coder would result in audible chirps.

To correct this problem, the intermediate pitch values are computed in the following way:

For the case where \( P(n_1) < P(n_2) \):

\[
P(n) = \begin{cases}
C(n_2 - n)P(n_1) + (n - n_1)P(n_2) \\
C(n_2 - n_1) \\
C(n_2 - n_1)P(n_1) + (n - n_1)P(n_2) \\
C(n_2 - n_1)
\end{cases} \quad \text{for } n_1 \leq n \leq \frac{n_1 + n_2}{2}
\]

(3.35)

where \( C \) is defined to be the ratio of \( P(n_2) \) to \( P(n_1) \) rounded to the nearest integer.

For \( P(n_1) > P(n_2) \):

\[
P(n) = \begin{cases}
(n_2 - n)P(n_1) + C(n - n_1)P(n_2) \\
C(n_2 - n_1) \\
(n_2 - n)P(n_1) + C(n - n_1)P(n_2) \\
C(n_2 - n_1)
\end{cases} \quad \text{for } \frac{n_1 + n_2}{2} \leq n \leq n_2
\]

(3.36)

where \( C \) is the nearest integer ratio of \( P(n_1) \) to \( P(n_2) \).

The factor \( C \) can be considered as an indicator showing whether the pitch has (sub)multiplied. When \( C \) is unity, it indicates that there is no pitch doubling or tripling and the above two formulations become the linear interpolation equation (3.34). On the other hand, when \( C \) is greater than one, it implies that the pitch has (sub)multiplied and the interpolation described by (3.35, 3.36) is performed in such a way that the pitch changes.
discontinuously at the midpoint by the factor $C$. Figure 3.9 illustrates an example of such interpolation in the case of pitch doubling and halving.

Figure 3.9: Interpolation of pitch in the case of pitch doubling. The left diagram interpolates between 30 and 70 using (3.34) and (3.35) for linear and piece-wise linear interpolation respectively, and the right diagram does the vice-versa using (3.34) and (3.36).

3.4 Voiced/Unvoiced classification

As described in section 2.1, the voiced and unvoiced speech segments differ in the way the excitation is generated. The quasi-periodic nature of voiced segments is the main feature for speech classification as voiced and unvoiced. In the following sections some of well known features for discriminating voiced from unvoiced speech are described.

3.4.1 Periodicity

Periodicity is the main feature for separating voiced speech from unvoiced. Due to the quasi-periodic characteristic of the voiced speech, the pitch is usually well defined. In contrast, since the unvoiced speech is of a random-like nature, there is no repetition of the signal for any pitch period. Thus, the normalised autocorrelation function can be used as a measure for V/UV classification. This is achieved through computing the normalised autocorrelation function at the estimated pitch period as follows:

$$R(T) = \frac{\sum_{n=1}^{N} s(n)s(n+T)}{\sqrt{\sum_{n=1}^{N} s^2(n)} \sqrt{\sum_{n=1}^{N} s^2(n+T)}}$$

(3.37)
where \( N \) is the length over which \( R(T) \) is computed and \( T \) is the computed pitch period. \( R(T) \) has a value between \(-1\) and \(+1\). A value of \( R(T) \) close to \(+1\) indicates high correlation between the speech and its shifted version which is a quasi-periodic signal (voiced speech) whereas a value of negative or close to zero leads to an uncorrelated or low similarity signal which is a random-like signal (unvoiced speech). Figure 3.10 shows speech signal with its corresponding \( R(T) \) values. Since the pitch usually evolves smoothly during a frame, it is more efficient to compute the normalised autocorrelation function at range of \([T - 2, T + 2]\) and to consider the maximum of the resulting values for voicing decision. This procedure is used in the MELP coder for voicing decision of the speech subbands [3].

![Figure 3.10: Original speech and the corresponding normalised autocorrelation R(T) (for clarity weighted and shifted up) and possible voicing threshold at 0.62.](image)

**3.4.2 Peakiness**

The energy of the voiced speech is usually concentrated around the pitch pulse locations and little in other regions. This can be more obvious in the residual domain where the short-term correlation is removed. On the other hand, the energy of the unvoiced speech due to its random-like nature is spread and normally there is no concentrated energy. As a result, the existence of regular peaks indicates voiced speech whereas their absence indicates unvoiced speech. The ratio of the L2 norm to the L1 norm of the residual signal can be used as a
possible measure for checking the presence of the pitch pulses. In case of voiced speech, this ratio will be high, typically higher than 1.5, and low for unvoiced speech [43].

![Figure 3.11: Original speech and the corresponding peakinesses ($P_k$ in solid and $\hat{P}_k$ in dash line) weighted and shifted up for clarity with possible voicing threshold at 1.44 (horizontal dashed line).](image)

This ratio is called peakiness and is given by:

$$P_k = \sqrt{\frac{1}{N} \sum_{n=1}^{N} r^2(n)}$$

(3.38)

Where $r(.)$ is the residual signal and $N$ is the length over which $P_k$ is computed.

There is occasionally a single spike in an unvoiced speech which results in a high value of $P_k$. To overcome this problem, a second peakiness, $\hat{P}_k$, is employed. In this method, a certain number of samples around the pitch pulse are excluded from the summations of Eq. (3.38). $P_k$ and $\hat{P}_k$ are usually close to each other for voiced or unvoiced speech without any spike. However, in case of unvoiced speech with a single spike, $\hat{P}_k$ is much lower than $P_k$. 
Thus, the difference between $P_k$ and $\hat{P}_k$ is used to enhance peakiness measure based for voicing decision. Figure 3.11 depicts a speech signal with its corresponding $P_k$ and $\hat{P}_k$.

### 3.4.3 Zero crossing

Unvoiced speech is characterised as random-like in nature. As a result, the number of times which the signal crosses the zero line, is usually high.

![Zero crossing rate graph](image)

Figure 3.12: Original speech and corresponding zero crossing rate (for clarity weighted and shifted up) with possible voicing threshold at 0.56.

In contrast, this zero-crossing rate is normally lower for voiced speech. The zero-crossing rate is given by:

$$Z_e = \frac{1}{N} \sum_{i=1}^{N} \delta(i)$$  \hspace{1cm} (3.39)

Where

$$\delta(i) = \begin{cases} 
1 & \text{if } s(i)s(i+1) \leq 0 \\
0 & \text{Otherwise} 
\end{cases}$$  \hspace{1cm} (3.40)
The $Z_c$ is a value between 0 and 1. A value close to 1 indicates unvoiced speech whereas a value close to zero indicates voiced speech. Figure 3.12 depicts a speech signal with its corresponding zero-crossing rate.

### 3.4.4 Low to full (LF) band spectrum energy ratio

One of the characteristic of voiced speech spectrum is that most of the energy is concentrated in the lower part of the spectrum. In contrast, the energy of the unvoiced speech is concentrated in higher frequencies or spread over the whole spectrum. Therefore, the ratio between the energy of the low frequency band (e.g. 0-2 kHz) to the energy of the original signal can be used as a measure for V/UV decision. In order to compute the energy of the low frequency band, the original speech signal can be filtered using a low pass filter with cutoff frequency of 2 kHz. The LF is normally high, close to unity, for voiced speech whereas is low, usually close to 0.5, for unvoiced speech. Figure 3.13 shows a speech signal with its computed LF.

![Figure 3.13: Original speech and corresponding LF (for clarity weighted and shifted up) with possible voicing threshold at 0.6.](image_url)
3.4.5 Linear prediction gain

The successive samples of the voiced speech are highly correlated with each other whereas they show low correlation in the case of unvoiced speech. As a result, LP modelling is more efficient for voiced speech and the prediction gain of the LP filter will be much higher. The prediction gain can be computed by:

\[
G_p = \frac{\sum_{n=1}^{N} s^2(n)}{\sum_{n=1}^{N} r^2(n)}
\]  

(3.41)

Where \( s \) and \( r \) are the original speech and the corresponding LP residual signal. Since this ratio gives values \( \geq 1 \), its inverse, \( G_p^{-1} \), is usually used as a measure for V/UV decision. \( G_p^{-1} \) a value between 0 and 1, a value close to 1 corresponding to a completely random signal. A plot of \( G_p^{-1} \) with speech signal is given in Figure 3.14.

Figure 3.14: Original speech and the corresponding inverse of prediction gain (for clarity weighted and shifted up) with possible voicing threshold at 0.38.
3.4.6 Pre-emphasis energy ratio

The high correlation present in voiced speech can also be exploited through the normalised pre-emphasis energy. In this method, a first order LP filter is used instead of the tenth order. The pre-emphasis energy ratio is given by:

$$P_r = \frac{\sum_{n=1}^{N} |s(n) - s(n-1)|}{\sum_{n=1}^{N} |s(n)|}$$

(3.42)

Since there is a high correlation between successive samples of the voiced speech, the first-order normalised correlation is usually high, close to 0.85, whereas it is close to zero for unvoiced speech. As a result, $P_r$ is close to zero for voiced speech and around 1 for unvoiced speech. Figure 3.15 shows a speech signal with its corresponding $P_r$.

Figure 3.15: Original speech and corresponding pre-emphasis, $P_r$ (for clarity weighted and shifted up) with possible voicing threshold at 0.92.
3.4.7 Decision process

A good V/UV speech classification can be provided using either one of the methods described above. In most cases, all the methods will result in the same voicing decision. However, one or more of the methods may give the wrong speech identification. In order to minimise the inaccurate decisions, it is more logical that the final decision is performed based on a combination of the results of each method. This implies to normalise all results to have the same range, typically between 0 to 1, and setting decision thresholds for each of them. The final decision can be obtained by a majority voting, or sum of the appropriate weighted results and comparing with a threshold. There is no arithmetic solution for computing the thresholds and weighting factors. These parameters are obtained through listening to the synthesised speech quality and adjusting the values accordingly.

3.5 Voicing-level estimation

The objective of speech classification is to decide whether a speech frame is voiced or unvoiced. However, there are two problems with such classification, i.e., the decision about transition frames (the frames which include both voiced and unvoiced parts), and, speech frames containing sounds, which include both periodic and noisy components. In such cases, a binary V/UV decision is unable to indicate the proper phonetic state of the frame. As a result, some tonal artefacts appear as the output of vocoders, which use such a binary decision. One simple solution, which was first introduced in the Multi-Band Excitation (MBE) vocoder, is to consider each frame as a combination of both voiced and unvoiced components. The MBE technique uses the frequency domain for V/UV analysis and decides which frequency band should be declared as voiced or unvoiced. MBE based coders separate the speech spectrum into harmonic bands, and check the matching each band with a reconstructed all voiced synthetic spectrum [6]. Good matching bands are declared voiced and the rest unvoiced. The MELP coder separates the spectrum into 5 subbands, and determines the voicing of each band using the periodic similarity of the signal [7].

3.5.1 Voicing-level in MBE coding

The voicing decision based on the MBE vocoder is performed by the construction of a speech synthetic spectrum assumed to be all voiced and comparing it with the original speech spectrum. The voiced regions of the speech spectrum show high similarity with the
corresponding synthetic spectrum whereas the unvoiced areas have larger dissimilarity. Therefore, this similarity or dissimilarity measure is used to make the V/UV decision for each frequency band by comparing it against a predetermined threshold.

The MBE splits the speech spectrum in a number of harmonic bands and the normalised error for each band is calculated as:

\[
D_k = \frac{\sum_{m=a_k}^{b_k} |S(m) - \hat{S}(m, \omega_0)|^2}{\sum_{m=a_k}^{b_k} |S(m)|^2}
\]

where \( \omega_0 \) is the fundamental frequency, \( S(m) \) is the original spectrum, \( \hat{S}(m, \omega_0) \), \( a_k \) and \( b_k \) are the synthetic spectrum, the first and the last harmonics in the \( k^{th} \) band, respectively, which are computed through the procedure described in section 3.3.2.2.

The voicing decision is made by comparing the normalised error, \( D_k \), with an adaptive threshold function for each band. If \( D_k \) is less than the threshold of the corresponding band it is declared voiced; otherwise it is declared unvoiced.

The adaptive threshold function, \( \Delta_k(\omega_0) \), is defined as [19]:

\[
\Delta_k(\omega_0) = (\alpha + \beta \omega_0)[1.0 - \varepsilon(k-1)\omega_0]M(\omega_0, \omega_{avg}, \omega_{min}, \omega_{max})
\]

where \( \alpha = 0.35, \beta = 0.557, \varepsilon = 0.4775 \) are the factors that give good subjective quality, and

\[
M(\omega_0, \omega_{avg}, \omega_{min}, \omega_{max}) = \begin{cases} 
0.5 & \omega_{avg} < 200 \\
(\omega_0 + \omega_{min})(2\omega_0 + \omega_{max}) & \omega_{avg} \geq 200, \omega_{min} < \mu \omega_{max} \\
1.0 & \text{Otherwise}
\end{cases}
\]

is the adaptation factor that controls the decision threshold for voicing decision and \( \mu = 0.0075 \). The parameter \( \omega_0 \) is the average energy of the current speech frame. The parameters \( \omega_{avg}, \omega_{max} \) and \( \omega_{min} \) roughly correspond to the local average energy, the local maximum energy and the local minimum energy, respectively. These three parameters are updated every speech frame according to [19].
Chapter 3. Basic Tools in Speech Coding

\[ \zeta_{\text{avg}} (j) = 0.7\zeta_{\text{avg}} (j - 1) + 0.3\zeta_0 \] (3.46)

\[ \zeta_{\text{max}} (j) = \begin{cases} 
0.5\zeta_{\text{max}} (j - 1) + 0.5\zeta_0 & \text{if } \zeta_0 > \zeta_{\text{max}} (j - 1) \\
0.99\zeta_{\text{max}} (j - 1) + 0.01\zeta_0 & \text{Otherwise}
\end{cases} \] (3.47)

\[ \zeta_{\text{min}} (j) = \begin{cases} 
0.5\zeta_{\text{min}} (j - 1) + 0.5\zeta_0 & \text{if } \zeta_0 \leq \zeta_{\text{min}} (j - 1) \\
0.975\zeta_{\text{min}} (j - 1) + 0.025\zeta_0 & \text{if } \zeta_{\text{min}} (j - 1) \leq \zeta_0 < 2\zeta_{\text{min}} (j - 1) \\
1.025\zeta_{\text{min}} (j - 1) & \text{Otherwise}
\end{cases} \] (3.48)

Indices (j) and (j-1) are used to indicate the parameters of the current and previous frames, respectively.

3.5.2 Voicing-level in MELP

The voicing decision in MELP is performed by computing the normalised autocorrelation function on bandpass filtered versions of the original speech at the estimated pitch value and comparing the results against a threshold. The voicing decision analysis begins by filtering the original speech into five frequency bands. These filters are 6th order Butterworth filters, with pass-bands of 0-500, 500-1000, 1000-2000, 2000-3000 and 3000-4000 Hz [3].

A refined pitch measurement is made using the first band filter output signal. This measurement is performed by computing the normalised autocorrelation over lags of [P-5, P+5] where P is the initial pitch estimate. The maximum of the resulting normalised autocorrelation values is used as the bandpass voicing strength for the first band and the correspond pitch lag is selected as the refined pitch value which is used in the computation of the band-pass voicing for the remaining bands.

For the other bands, it is more efficient that the normalised autocorrelations are computed on the envelopes of the bandpass filtered speech signals. The time envelope for each band is computed as:

\[ y_k (n) = |x_k (n)| - |x_k (n - 1)| + C_1 y_k (n - 1) + C_2 y_k (n - 2) \] (3.49)

Where \( x_k (n) \) and \( y_k (n) \) are the bandpass filter signal of k\(^{th}\) band and the corresponding its envelope, respectively and \( C_1 = 1.9266 \) and \( C_2 = -0.9409 \).
Chapter 3. Basic Tools in Speech Coding

The band-pass voicing strength for the remaining bands is computed using the normalised autocorrelation at the refined pitch on the resulting time envelopes, and also on the bandpass filtered signals. The maximum of these two autocorrelations is then considered as the bandpass voicing strength of that band.

The resulting bandpass voicing strengths, $V_{bp_i}, i = 1, 2, ..., 5$, are then refined using the peakiness of the signal. If the peakiness exceeds 1.34, then the lowest band voicing strength, $V_{bp_1}$, is forced to 1.0. If the peakiness exceeds 1.6, then the lowest three bands voicing strengths, $V_{bp_i}, i = 1, 2, 3$, are all forced to 1.0.

The voicing decision for each band is made by comparing the corresponding voicing strength with a threshold:

- If $V_{bp_1} < 0.6$, all bands are declared unvoiced
- If $V_{bp_1} > 0.6$, the first band is declared voiced, and the remaining voicing strengths are then compared with the same threshold (≥0.6). For each band, if the voicing strength exceeds 0.6, the band is declared voiced; otherwise it is declared unvoiced.
- There is one exception. If the second, third and the fourth bands are declared unvoiced and the fifth band is declared voiced, the last band is forced to be unvoiced.

In spite of simplicity of the method employed in the MELP coder for voicing strengths estimation, it has some shortcomings. Since the bandwidth of each band is fixed, a spectral harmonic may be located at the boundary of two bands. This can result in inaccurate voicing strength estimation, especially in the case of a highly energetic harmonic. Moreover, in the case of a low fundamental frequency, due to large bandwidth, a large number of spectral harmonics may be located in a certain band. This leads to crude voicing strength estimation for the corresponding band. However, since the quantisation of the voicing decisions require only 5 bits, the MELP coder has shown high efficiency in low bit rate speech coding [3].

3.6 Conclusions

In this chapter three basic tools in low bit rate speech coding techniques namely LP analysis, pitch and voicing estimation were discussed.

It was addressed that the LP technique is a powerful tool in modelling the short-term correlation of speech. In addition, different representations of LP coefficients were discussed. Due to the LSF properties, the case in which the stability of the LP synthesis filter may be
checked easily and good quantisation performance, most speech coders use this representation for LP coefficients transmission.

Different pitch estimation algorithms were also discussed both in time and frequency domain. Among the various pitch determination algorithms, the time domain techniques, due to having low complexity, are commonly used in low bit rate speech coders. Although frequency domain algorithms, in comparison with time domain techniques, are more complex and are usually used for pitch refinement purposes, they are better suited to frequency domain based speech coders where FFT is used for other parameter estimations such as pitch harmonics and voicing-level estimations.

The main speech features used for binary V/UV classification and voicing-level estimations were described. Due to existence of mixed voiced and unvoiced speech parts in some cases and also transient frames, a binary V/UV decision is unable to indicate the proper phonetic state of the frame. This problem was overcome through using voicing-level estimation, which considers frame as a combination of both voiced and unvoiced components. The MELP and MBE were discussed as speech coders, which employ voicing-level estimation instead of binary V/UV classification.

Pitch and voicing estimations are open problems, but it is possible to make speech signal more regular such that these parameters can be effectively estimated through pre-processing the speech frames. This will be discussed in chapters 6.
Chapter 4

Overview of WI and MELP Coders

4.1 Introduction

WI and MELP coders have shown high speech quality at bit rates lower than 2.8 Kb/s. WI was first introduced by W. B. Kleijn [45] and the first version was called Prototype Waveform Interpolation (PWI). PWI encoded voiced segments only and therefore, it was used in combination with other schemes such as CELP for coding unvoiced segments. In 1994, it was further refined to become WI, which is capable of encoding both voiced and unvoiced speech [46]. Similar to the principles of PWI, WI represents a speech signal with a sequence of evolving waveforms. For voiced speech, these waveforms are simply pitch-cycles whereas for unvoiced speech and for background noise, the waveforms are of varying lengths and contain mostly noise-like signals. Since the evolving waveforms are not limited to pitch-cycles anymore, it is not appropriate to use the terms pitch-cycle or prototype waveform to describe the evolving waveform. Instead, the term Characteristic Waveform is adopted, which will be abbreviated to CW from here on. The WI decomposes the CW into a smoothly evolving waveform (SEW) and a rapidly evolving waveform (REW). The SEW represents the quasi-periodic component of the speech signal while the REW represents the remaining non-periodic and noise components in the signal. Since the two waveforms have very different perceptual requirements, they can be quantised separately to enhance coding efficiency. High quality WI coding at 2.8 Kb/s was reported by O. Gottesman and A. Gersho in [47][48].

The MELP is based on the traditional LPC parametric model, but also includes four additional features [7]. These are: mixed excitation, aperiodic pulses, adaptive spectral enhancement, and pulse dispersion. The mixed excitation is implemented using a multi-band mixing model. This model can simulate frequency-dependent voicing strength using an adaptive filtering structure implemented with a fixed filter bank. The primary effect of this mixed excitation is to reduce the buzzy quality usually associated with LPC vocoders. When the input speech is voiced, the MELP coder can synthesise using either periodic or aperiodic pulses. Aperiodic pulses are used most often during transition regions between voiced and unvoiced segments. This feature enables the decoder to reproduce erratic glottal pulses.
without introducing tonal sounds. The adaptive spectral enhancement filter is based on the poles of the linear prediction synthesis filter. Its use enhances the formant structure of the synthetic speech and improves the match between the synthetic and natural bandpass waveforms. It also gives the synthetic speech a more natural quality. Pulse dispersion is implemented using a fixed filter based on a spectrally flattened triangle pulse. This filter spreads the excitation energy within a pitch period, reducing some of the harsh quality of the synthetic speech.

In [3], it was shown that the performance of the 2400 b/s MELP vocoder is close to that of the government standard 4800 b/s CELP coder. The MELP vocoder is the new 2400 b/s Federal Standard speech coder. It was selected by the United States Department of Defence Digital Voice Processing Consortium (DDVPC). MELP was selected as the best of the seven candidates and even better the FS1016 4800 b/s voice coder, a coder with twice the bit rate. The MELP is robust in difficult background noise environments such as those frequently encountered in commercial and military communication systems.

This chapter is organised as two main sections. First section describes WI coder, including how the parameter are estimated at WI encoder and how the speech is synthesised at WI decoder. In the second section, the MELP is detailed.

### 4.2 Waveform Interpolation (WI) Coding

There are two approaches for WI coding. In the first approach (i.e. Prototype Waveform Interpolation (PWI) coder [45]) the pitch-cycle waveform is extracted once per frame (only for voiced speech) and the other pitch-cycles are reconstructed by interpolation at the decoder. In this approach a V/UV classifier is required at the encoder. On the other hand, in the second approach, called Multi-prototype waveform coding [44][49], CWs are extracted at 2ms intervals and so no distinction is necessary between voiced and unvoiced. Figure 4.1 provides a block diagram of such a WI coder.

The first step in coding process is the linear prediction analysis, which is performed at the frame rate. The LP parameters are conventionally interpolated to a sub-frame update rate. Then, CWs are extracted from LP residual signal using pitch information. Interpolation is used to obtain better estimation of the LP parameters and pitch period for each sub-frame. Also, once the CW is extracted from the residual speech signal, the smoothness of the surface $U(t, \phi)$ in the $t$-direction (CW extraction time) must be maximised. This can be accomplished by alignment in $\phi$ of CW with the previously extracted CW. Then, CWs are normalised such that the magnitude spectrum has unit magnitude and are then decomposed into two components by filtering along the time axis. High-pass filtering results in a Rapidly Evolving
Waveform (REW), representing the noise-like component of speech. Low-pass filtering results in a Slowly Evolving Waveform (SEW), representing the nearly periodic component of speech [44][50][51].

The receiver reconstructs and then adds the evolving REW and SEWS waveforms to reconstruct CW. This waveform is scaled to have the proper signal power. The CW is upsampled by linear interpolation between the prototype waveforms to construct two-dimensional CW surface. This surface is converted to the reconstructed residual using phase track information. The residual signal is passed through the synthesis LP filter to construct synthesised speech signal.

4.2.1 WI encoder

In the following subsections, the procedures employed in a WI encoder are discussed. These procedures consist of CW extraction, alignment, decomposition of two-dimensional surface, \( U(n, \phi) \), into the SEW and REW surfaces. Note that due to using discrete time, \( t \)-index indicating continuous time is substituted by \( n \)-index.
4.2.1.1 CW extraction

CW extraction rate can be performed either at a fixed or time-varying rate [44]. The fixed extraction rate leads to an essentially quadratic increase of the computational effort with the pitch period. In contrast, time varying extraction rate significantly decreases the worst-case computational requirement. It has been shown that the bandwidth of $U(n, \varphi)$ along the $n$-direction is $\frac{1}{2P(n)}$ and so a waveform-sampling rate of at least $\frac{1}{P(n)}$ should prevent any aliasing along $n$-axis [44]. Thus, the number of extracted CWs with frame rate $f_r$ is equal to:

$$CW\ Extraction\ Rate = \frac{1}{P(n).f_r}$$

(4.1)

CW extraction can be performed in speech domain [51][52] or in residual domain. Since the extraction can be more efficiently performed on a linear prediction residual signal [44], the entire WI procedure operates on the LP residual. The first stage of CW extraction is to determine the pitch period. The pitch period can be calculated by time or frequency domain algorithms. The pitch estimator provides an estimation once per frame. However, WI requires the pitch period at every extraction point. To solve this problem while maintaining the same level of computational complexity, a pitch interpolator is employed to calculate the intermediate pitch values. It has been shown that if the conventional linear interpolation technique is applied, the resulting reconstructed speech would result in audible chirps [44]. To correct this problem, the pitch interpolator described in section 3.3.4 is used.

After pitch determination and interpolation, the CW is extracted. The extraction starts with a segment of the residual signal of length one pitch period. Ideally, each segment is centred on the extraction time. However, the extraction location can be moved slightly to minimize discontinuities resulting from the periodic extension. In the modification of the extracted prototype [52], samples around the prototype boundaries are substituted by interpolation between their original and predicted amplitudes. Then, the segment is time scaled to a length $2\pi$ and periodically extended in $\phi$-direction. Thus, the periodic function of $\phi$ can be considered by a finite (time-dependent) Fourier series [44].

4.2.1.2 CW alignment

The extraction procedure provides a Discrete Time Fourier Series (DTFS) description for every extracted CW. In general, these CWs are not phase aligned which indicates an
inaccurate waveform evolutions for voiced speech. In order to obtain an accurate description of the evolving CWs, an alignment of the CWs must be performed. This procedure aligns the current CW with the previous CW by introducing a circular time shift to the current one.

Suppose now a circular time shift of $T$ samples (to the right) is applied to the current CW, $U(n_1,m)$ then becomes:

$$U(n_1,m-T) = \sum_{k=1}^{P(n_0)} \left[ A_k(n_1) \cos \left( \frac{2\pi k (m-T)}{P(n_1)} \right) + B_k(n_1) \sin \left( \frac{2\pi k (m-T)}{P(n_1)} \right) \right]$$ (4.2)

From a DSP point, a circular shift $T$ in the time-domain is equivalent to adding a linear phase $\frac{2\pi T}{P(n_1)}$ in the DTFS domain. Next, in order to determine the amount of time shifting $T$ required for aligning $U(n_1,m-T)$ with $U(n_0,m)$, the following criterion is maximised:

$$T = \arg \max_{0 \leq T' < P(n_1)} [U(n_1,m-T'),U(n_0,m)]$$ (4.3)

$n_0$ and $n_1$ indicate previous and current extraction time, respectively. Since $U(n_0,m)$ and $U(n_1,m)$ may have different dimension, a spectral zero padding is applied to the shorter CW to match the length of the longer CW. Eq. (4.3) can be rewritten in terms of DTFS coefficients as follows:

$$T = \arg \max_{0 \leq T' < P(n_1)} \sum_{k=1}^{M} \left[ A_k(n_0)A_k(n_1) + B_k(n_0)B_k(n_1) \cos \left( \frac{2\pi k T'}{P(n_1)} \right) \right] + \left[ B_k(n_0)A_k(n_1) - B_k(n_1)A_k(n_0) \sin \left( \frac{2\pi k T'}{P(n_1)} \right) \right]$$ (4.4)

Where

$$M = \left[ \frac{\max(P(n_0), P(n_1))}{2} \right]$$ (4.5)

The right-hand side of (4.4) is the cross-correlation between the two CWs expressed in terms of the DTFS coefficients. A detailed proof of this can be found in [44][51]. Equation (4.4) can be also expressed in terms of a normalised time shift $\tau$ :

$$\tau = \frac{2\pi T'}{P(n_1)}$$ (4.6)

By substituting $\tau$ into (4.4), we can obtain:
\[ \tau = \arg \max_{\tau} \sum_{k=1}^{M} \left\{ \left[ A_k(n_0)A_k(n_1) + B_k(n_0)B_k(n_1) \right] \cos(k\tau') + \left[ B_k(n_0)A_k(n_1) - B_k(n_1)A_k(n_0) \right] \sin(k\tau') \right\} \]

(4.7)

The advantage of performing the alignment in the DTFS domain is that it allows fractional alignment without any extra computations and the conventional up-sampling and down-sampling procedures are avoided. Also, this fractional alignment can be at any desired resolution \( \tau \) can be any value between 0 and \( 2\pi \). It was reported [51] that \( \tau \) with a resolution of \( \frac{1}{4} \) of a sample gives sufficiently good perceptual results.

Next, the time shift \( \tau \) corresponding to the maximum cross-correlation is incorporated into DTFS representation of the current CW. This can be performed by expanding the \( \sin(.) \) and \( \cos(.) \) terms in (4.2) using fundamental trigonometric identities. By grouping the relevant terms together in the resulting expansion, we would have a new set of DTFS coefficients:

\[
A'_k(n_1) = A_k(n_1)\cos(k\tau) - B_k(n_1)\sin(k\tau) \quad k=1,2,...,M
\]

(4.8)

\[
B'_k(n_1) = A_k(n_1)\sin(k\tau) + B_k(n_1)\cos(k\tau)
\]

and

\[
\bar{U}(n_1, m - T) = \sum_{k=1}^{M} \left[ A'_k(n_1) \cos \left( \frac{2\pi km}{P(n_1)} \right) + B'_k(n_1) \sin \left( \frac{2\pi km}{P(n_1)} \right) \right]
\]

(4.9)

### 4.2.1.3 Decomposition of CW surface

After the CW extraction and alignment, the CWs power is then normalised. Thus, two-dimensional surface of aligned, normalized CWs can be constructed. This surface has to be transmitted to WI decoder. At first sight, it appears that an accurate representation of the CWs may require a very high transmission rate. Fortunately, not all of the information contained in the surface is perceptually relevant to human ears. The CW representation is particularly convenient for the separation into voiced and unvoiced components. For the voiced component of the speech signal, the CW evolves slowly as a function of CW-extraction time (SEW). In contrast, for the unvoiced component of speech, CW evolves rapidly (REW). These two components must sum to the entire evolving waveform:

\[
U(n, \phi) = U_{REW}(n, \phi) + U_{SEW}(n, \phi)
\]

(4.10)

The separation of \( U(n, \phi) \) into \( U_{REW}(n, \phi) \) and \( U_{SEW}(n, \phi) \) is accomplished with a simple filtering operation. The SEW is formed by low-pass filtering the CW evolving along the time
axis (n-axis) and the REW can be found by subtracting the SEW from the CW surface. For voiced speech, the SEW and the REW represent respectively a shaped pulse-like waveform and a noise component. Due to the presence of the periodicity in voiced regions, the SEW generally has a much higher energy level than the REW. In contrast, for unvoiced speech where the signal evolves quite rapidly and exhibits no apparent periodicity, the decomposition distributes most of the CW energy to the REW. Figure 4.2 illustrates an example of a SEW and a REW surfaces decomposed from the CW surface.

![Figure 4.2: Decomposition of the CW surface into the SEW and the REW.](image)

Since the DTFS operation is a linear transformation, low-pass filtering the CWs in the time domain is equivalent to low-pass filtering their DTFS coefficients. In practice, the SEW is obtained by low-pass filtering the time sequence $A_k(n_i), A_k(n_{i+1}), A_k(n_{i+2}), ...$ and also $B_k(n_i), B_k(n_{i+1}), B_k(n_{i+2}), ...$ for each Fourier series coefficient and for all $n$ [10]. For perceptually meaningful results, the cut-off frequency of low-pass filter is set to 25 Hz. As a result the SEW can be down-sampled to 50 Hz [44] (assuming an ideal low-pass filter).
4.2.1.4 WI encoder parameters

The WI encoder decomposes a speech segment into five parameters; pitch, LSFs (or LPC), power of CWs, SEW and REW.

4.2.3 WI decoder

The speech signal can be reconstructed by using the LSFs, pitch, power, SEW and REW surfaces. This is discussed in the following sections.

4.2.3.1 CW Power de-normalization and re-alignment

As a first step, the power of each incoming CW is de-normalised. This de-normalizer uses the pitch information supplied by the interpolator to determine the length of the CW. Due to the quantisation of the CWs, the decoded CWs may be misaligned and therefore, they are required to be realigned. Obviously, if the coder is operating without quantisation and since the CWs have already been aligned previously at the encoder, this realignment is unnecessary.

4.2.3.2 Instantaneous pitch and CW generation

In the WI paradigm, a CW and a pitch length at every sample point are required to reconstruct the one-dimensional residual signal. The instantaneous pitch periods can be obtained by equations (3.35) and (3.36) (see section 3.3.4, Chapter 3). In order to up-sample the CWs linear interpolation can be used. When the up-sampling is performed between two CWs of the same dimension, straightforward interpolation can be applied. However, in general, the pitch will change over the intervals and the CWs will have different lengths (different number of coefficients \( \{ A_k, B_k \} \)). To facilitate the interpolation in this case, one can time-scale the shorter CW to the length of the longer one prior to the interpolation. Such a time-scale operation is equivalent to padding zero harmonics to its DTFS description. Suppose \( n_0 \) and \( n_1 \) to be the time instants at the boundaries of an interpolation interval, then the instantaneous CW, \( U(n,m) \), at index \( n \) can be computed by interpolating between \( U(n_0,m) \) and \( U(n_1,m) \).

In time domain, this interpolation can be expressed as:

\[
U(n,m) = \left( \frac{n_1 - n}{n_1 - n_0} \right) U(n_0,m) + \left( \frac{n - n_0}{n_1 - n_0} \right) U(n_1,m) \quad n_0 \leq n \leq n_1, 0 \leq m \leq M \quad (4.11)
\]
and in DTFS domain:

\[
\begin{align*}
A_k(n) &= \left( \frac{n_1-n}{n_1-n_0} \right) A_k(n_0) + \left( \frac{n-n_0}{n_1-n_0} \right) A_k(n_1) \\
B_k(n) &= \left( \frac{n_1-n}{n_1-n_0} \right) B_k(n_0) + \left( \frac{n-n_0}{n_1-n_0} \right) B_k(n_1)
\end{align*}
\]

\[k=1,2,\ldots,[M/2] \]  

(4.12)

Where \(M\) is given by Eq. (4.5).

### 4.2.3.3 Phase track estimation

Phase track, \(\varphi(.)\), is used to transform the two-dimensional CW surface back into the one-dimensional residual signal. The phase track is constructed by the pitch values. The phase contour at each sample point can be updated as:

\[
\varphi(n) = \varphi(n-1) + \int_{n-1}^{n} \frac{2\pi}{P(n')} dn'
\]

(4.13)

where \(\varphi(n)\) and \(\varphi(n-1)\) are the current and previous phase values. Since the integral in Eq. (4.13) results in a positive value, \(\varphi(n)\) is an increasing function. In order to prevent \(\varphi(n)\) from overflowing, a subtraction off a multiple of \(2\pi\) is used. Since pitch per sample normally evolves smoothly, the integral in equation (4.13) is approximated as:

\[
\varphi(n) = \varphi(n-1) + \pi \left( \frac{1}{P(n-1)} + \frac{1}{P(n)} \right)
\]

(4.14)

The initial phase, \(\varphi(0)\), can be set to an arbitrary number at the beginning of a speech signal. The initial phase offset value does not affect the perceptual quality of the reconstructed speech [44].

### 4.2.3.4 2D-to-1D conversion

After 2-dimensional CW surface construction, this surface should be converted to the one-dimensional residual signal, \(r(n)\). This conversion is performed based on sample-by-sample transformation using phase track information. Figure 4.3 demonstrates the reconstruction process for a residual segment. The interpolated phase track corresponding to the segment is
Chapter 4. Overview of WI and MELP Coders

illustrated in Fig. 4.3a. Figure 4.3b shows the interpolated CW surface where each CW is normalized to a length of $2\pi$. In order to reconstruct $n^{th}$ sample of the residual signal, $r(n)$, index of $n$ selects $n^{th}$ CW and $\phi(n)$ specifies sample position of the selected CW. This can be presented by:

$$r(n) = U(n, \phi(n)) = \sum_{k=1}^{P(n)/2} [A_k(n) \cos(k \phi(n)) + B_k(n) \sin(k \phi(n))] \quad 0 \leq \phi(n) < 2\pi \quad (4.15)$$

The reconstructed residual signal is used to excite the LP synthesis filter to obtain the final speech signal. The filter coefficients are computed as a result of the LSF interpolation.

![Diagram of CW surface and residual signal](image)

Fig. 4.3 Transformation from a CW surface to a residual signal. (a) An interpolated phase track for a voiced segment, which has a constant pitch at 40. (b) The interpolated pitch track superimposed on the interpolated CW surface [51].

4.3 Transmission rate of WI parameters

In [44] a particular 2.4 kb/s WI coder was reported. The coder operates on frames of 25 ms in length (that is equal to 200 samples for 8 KHz sampling rate) and uses the Fourier-series representation of the CWs. The pitch period and LPCs (or LSFs) are calculated, interpolated and transmitted once per frame. Therefore, the transmission rate of the pitch period and the LSFs is 40 Hz. The CWs are extracted at rate of 480 Hz (equal to 12 CWs per frame). After CW alignment and normalisation to unit power, the CW surface is constructed. The following subsections describe how the CW surface is transmitted.
4.3.1 CWs power transmission

As is well known that logarithm of the signal power is perceptually more relevant than the signal power itself [44]. Therefore, the CW power values are transformed to the logarithmic domain. They are then down-sampled from a rate of 40Hz to 80Hz (twice per frame) [44]. At the decoder, the signal power is up-sampled to a rate of 480Hz by interpolation. This interpolation is linear and is performed directly on the logarithmic power value [44][51]. Once the log power contour is up-sampled, the signal power can be obtained by the exponential operation.

4.3.2 REW surface transmission

It is well known that little degradation in speech quality is heard if the phase of the DTFS of the REW is replaced by a random phase spectrum and also if the magnitude spectrum of the DTFS of the REW is transmitted within a 5ms interval [44]. Therefore, in WI encoder, the REW surface is down-sampled to a rate of 160Hz [44]. Each down-sampled REW is then converted to its polar notation where the phase spectrum is discarded and the amplitude spectrum is only transmitted. At the decoder, the REW spectra are up-sampled at a rate of 480Hz. Each of the up-sampled REW amplitude spectrums is combined with a random phase spectrum and then transformed back to its rectangular coordinates. The values of the random phase spectra are independent and uniformly distributed in $[0, 2\pi]$. 

4.3.3 SEW surface transmission

Since the cut off frequency of the decomposition filter is only about 20Hz, the SEW has a very small evolution bandwidth. Thus the SEW surface can be down-sampled to a lower rate, i.e. 40 Hz. The down-sampled SEW is converted to its polar notation where the phase spectrum is discarded and the magnitude spectrum is transmitted. At the decoder, the SEW magnitude spectrum is attached to a fixed phase spectrum and converted back to the rectangular coordinates. The SEWs are up-sampled by a factor 12, which is a rate of 480Hz. In table 1, the transmission rate of the WI parameters is shown.

After reconstructing REW and SEW surface, the CW surface can be reconstructed. The residual signal $r(.)$ is reconstructed by 2D-to-1D conversion and excites the LPC synthesis filter to obtain the speech signal. As shown in [44], the WI coder provides a good speech modelling.
Table 4.1. Transmission rate of the 2.4 kb/s WI coder [44].

<table>
<thead>
<tr>
<th>Parameters Update (Hz)</th>
<th>LPC Pitch</th>
<th>Power</th>
<th>SEW(Magnitude)</th>
<th>SEW(Phase)</th>
<th>REW(Magnitude)</th>
<th>REW(Phase)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>40</td>
<td>40</td>
<td>80</td>
<td>40</td>
<td>0</td>
<td>160</td>
</tr>
</tbody>
</table>

4.4 Mixed-Excitation Linear Prediction (MELP) coder

Traditional pitch-excited LPC vocoders use either a periodic pulse train or white noise as the excitation for an all-pole synthesis filter. These vocoders produce intelligible speech at very low bit rates, but they sometimes sound mechanical or buzzy and are prone to annoying thumps and tonal noises. These problems arise from the inability of a simple pulse train to reproduce all kinds of voiced speech.

The MELP vocoder uses a mixed-excitation model to produce more natural sounding speech because it can represent a richer ensemble of possible speech characteristics. It is also robust in background noise environments. The 2400 bps MELP vocoder is Federal Standard speech coder. It was selected by the United States Department of Defence Digital Voice Processing Consortium (DDVPC) after a multi-year extensive testing program [53][54].

4.4.1 The MELP vocoder algorithm

The MELP vocoder includes four additional features in comparison with the traditional LPC parametric model. As shown in Fig. 4.4, the synthesiser has the following added capabilities:

1. Mixed pulse and noise excitation
2. Periodic or aperiodic pulses
3. Adaptive spectral enhancement
4. Pulse dispersion filter.

These features allow the MELP to mimic more of the characteristics of natural human speech.
4.4.1.1 Mixed excitation

The most important feature of the MELP is the mixed pulse and noise excitation. As shown in Fig. 4.4, the pulse train and noise sequence are each passed through time-variant spectral shaping filters and then added together to give a full-band excitation [7].

The frequency-shaping filter coefficients are computed by a weighted sum of fixed bandpass filters for each frame. The pulse filter is calculated as the sum of each of the bandpass filters weighted by the voicing strength in that band. The noise filter is obtained by a similar weighted sum. The weights are set to keep the total pulse and noise powers constant in each frequency band. These two frequency-shaping filters combined give a spectrally flat excitation signal with a staircase approximation to any desired noise spectrum [7]. As an example, Figure 4.5 shows speech spectrum with corresponding voicing strengths and the resulting frequency-shaping filters, where frequency bands 0-500, 500-1000 and 2000-3000 Hz are voiced and, 1000-2000 and 3000-4000 Hz are unvoiced. The synthesised pulse train and noise sequence of the corresponding residual signal are shown in Figures 4.6 and 4.7 in time and frequency domain, respectively. The filter bank is designed such that the sum of all the bandpass filter responses is a digital impulse. This guarantees that if all bands are fully voiced, the fullband excitation will be an undistorted pulse.

Figure 4.5: Speech spectrum and voicing strengths (a) with corresponding pulse and noise shaping filters (b,c).
Figure 4.6: The original residual speech signal (a) and the corresponding pulse and noisy parts (b,c) reconstructed by the MELP decoder.

In [7], this is performed by implementing FIR filters, designed by windowing the ideal bandpass filter impulse responses with a Hamming window. In order to use the mixed-excitation characteristics correctly, it is required that the desired spectral shaping mixture is accurately estimated for each frame. The relative pulse and noise power in each frequency band is determined by an estimate of the voicing strength at that frequency in the input speech. For lower bands, this is performed using the strength of normalised autocorrelation function around the pitch lag. Using this method at higher frequencies results in a slightly whispered quality synthetic speech [7]. In order to solve this problem, the envelopes of the bandpass filtered speech are computed as follows:

\[ y_k(n) = |x_k(n)| - |x_k(n-1)| + C_1 y_k(n-1) + C_2 y_k(n-2) \]  \hspace{1cm} (4.16)

Where \( x_k(n) \) and \( y_k(n) \) are the bandpass filter signal of \( k^{th} \) band and the corresponding envelope, respectively and \( C_1 = 1.9266 \) and \( C_2 = -0.9409 \).

Next, the autocorrelation analysis is performed on these bandpass filtered envelopes to determine the amount of pitch periodicity.
Figure 4.7: The original residual spectrum (a) of the speech signal depicted in Figure 4.5 with the reconstructed spectrum of the pulse and noise shaping filters (b, c respectively) and the synthesised residual spectrum (d).

The overall voicing strength in each frequency band is chosen as the largest of the correlation of the bandpass filtered input speech and the correlation of the envelope of the bandpass filtered speech.

4.4.1.2 Aperiodic pulses

The MELP can remove the buzzy quality from the LPC speech output. However, due to the presence of short, isolated tones another distortion occasionally appears. An effective solution is to destroy the periodicity in the voiced excitation by varying each pitch period length with a pulse position jitter uniformly distributed up to ±25%. Figure 4.8 shows a residual speech signal whose pitch period is 68 samples with the synthesised frequency-shaping filter outputs where the synthesised jitter is −17%. In other words, the period of each synthesised pitch cycle is changed to 57 samples.
Chapter 4. Overview of WI and MELP Coders

4.4.1.3 Adaptive spectral enhancement

The third feature of the MELP is adaptive spectral enhancement. This provides a good matching between the bandpass filtered synthetic speech and natural speech waveforms in the formant regions. The adaptive filter is achieved as a pole/zero filter whose poles are computed by a bandwidths expanded version of the LPC synthesis filter, with $\beta = 0.8$ [7]. Since this all-pole filter introduces a disturbing lowpass filtering effect by increasing the spectral tilt, a weaker all-zero filter calculated with $\alpha = 0.5$ is used to decrease the tilt of the overall filter.
without reducing the formant enhancement. In addition a first-order highpass filter is used to further reduce the lowpass muffling effect [7].

**4.4.1.4 Pulse dispersion filter**

The pulse dispersion filter improves the match between the bandpass filtered synthetic speech and natural speech waveforms in frequency bands which do not contain formant resonances. At these frequencies, the synthetic speech often decays to a very small value between the pitch pulses. In these cases, the bandpass filtered natural speech has a smaller peak-to-valley ratio than the synthetic speech. The pulse dispersion employed in the MELP [7], is a fixed FIR filter, based on a spectrally flattened synthetic glottal pulse, which introduces time-domain spread to the synthetic speech. This is performed using a fixed triangle pulse, based on a typical male pitch period, but first removes the lowpass characteristics from its frequency response. The filter coefficients are generated by taking a DFT of the triangle pulse, setting the magnitudes to unity, and taking the inverse DFT. The dispersion filter is applied to the entire synthetic speech signal to avoid introducing delay to the excitation signal prior to synthesis.

**4.4.2 The MELP encoder parameters**

The employed MELP coder is the 2400 bps Federal Standard speech coder. The input speech signal is high-pass filtered using a 4th-order Chebychev filter, having a cutoff frequency of 60 Hz and a stop-band rejection of 30 dB, to remove any low frequency energy. The output of this filter will be referred to as the input speech signal throughout the following description. The MELP coder takes the input signal and segments it into frames of 180 samples (22.5 ms). For each frame, the coder extracts 10 LP coefficients, 2 gain factors, 1 pitch value, 5 bandpass voicing strength values, first 10 Fourier magnitudes and an aperiodic flag. The following few sections briefly describe how MELP extracts these parameters and use them to reconstruct speech.

**4.4.2.1 LP coefficients**

A 10th order linear prediction analysis is performed on the input speech signal using a 200 samples (25ms) Hamming window centered on the last sample in the current frame. The traditional autocorrelation analysis procedure is implemented using the Levinson-Durbin recursion. In addition, a bandwidth expansion coefficient of 0.994 (15 Hz) is applied to the
prediction coefficients, \( a_i, i=1, 2, \ldots, 10 \), where each coefficient is multiplied by 0.994. Next, the linear prediction residual signal is calculated by filtering the input speech signal with the prediction filter whose coefficients were determined by linear prediction analysis.

### 4.4.2.2 Pitch values

There are several steps before a final value is assigned to the pitch. First the coder finds the integer pitch value, followed by a fractional pitch refinement. The integer pitch is calculated using the input speech signal. The signal is processed using a 6th-order Butterworth lowpass filter with cutoff frequency 1 kHz. The integer pitch value, \( P_1 \), is computed as the maximum of the signal autocorrelation and can take on values of lag ranging from 40-160 samples [55].

A refined pitch measurement is made using the 0-500 Hz filtered output signal. Two pitch candidates are considered in this refinement, namely the integer pitch values (\( P_1 \)) from the current and previous frames. For each candidate, the normalised autocorrelation is used to perform an integer pitch search over lags from 5 samples shorter to 5 samples longer than the candidate, and a fractional pitch refinement is performed around the optimum integer pitch lag. This produces two fractional pitch candidates and their corresponding normalised autocorrelation values. The candidate having the higher normalised autocorrelation is selected as the fractional pitch, \( P_2 \).

For computation of the final pitch, \( P_3 \), an integer pitch search is performed over lags of 5 samples shorter to 5 samples longer than \( P_2 \), rounded to the closest integer. Before assigning \( P_3 \) the final pitch value, fractional pitch refinement and pitch doubling checks are made.

### 4.4.2.3 Bandpass voicing strength values

Five bandpass voicing strengths, \( V_{bp_i}, i=1, 2, \ldots, 5 \), are determined based on five 6th order Butterworth filters with passbands of 0-500, 500-1000, 1000-2000, 2000-3000 and 3000-4000 Hz, respectively. The normalised autocorrelation value that results from \( P_2 \) is saved at the lowest bandpass voicing strength, \( V_{bp_1} \). For each remaining band, the bandpass voicing strength is the larger of the autocorrelation value, computed at \( P_2 \), determined for the corresponding bandpass signal and its time envelope.

The peakiness of the residual signal is computed as the ratio of the L2 norm to the L1 norm of the residual signal to bias the initial voicing strengths estimation. This was detailed in section 3.5.2 (see Chapter 3).
4.4.2.4 Gain factors

The input speech signal gain is measured twice per frame using a pitch-adaptive window length. This length is identical for both gain measurements and is determined as follows. When $V_{bpi} > 0.6$, the window length is the shortest multiple of $P$ which is longer than 120 samples. If this length exceeds 320 samples, it is divided by 2. When $V_{bpi} \leq 0.6$, the window length is 120 samples. The gain calculation for the first window produces $G_1$ and is centred 90 samples before the last sample in the current frame. The calculation for the second window produces $G_2$ and is centred on the last sample in the current frame. The gain is the RMS value, measured in dB, of the signal in the window, $s_n$:

$$G_i = 10 \log_{10} \left( 0.01 + \frac{1}{L} \sum_{n=1}^{L} s_n^2 \right)$$  \hspace{1cm} (4.16)

Where $L$ is the window length. If a gain measure is less than 0.0, it is clamped to 0.0.

4.4.2.5 Fourier magnitude calculation

This analysis measures the Fourier magnitudes of the first 10 pitch harmonics of the prediction residual generated by the quantised prediction coefficients. It uses a 512-point FFT of a 200-sample window centered at the end of frame. First, a set of quantised prediction coefficients is calculated from the quantised LSF vector. Then the residual signal is generated using the quantised prediction coefficients. Next, a 200-sample Hamming window is applied, the signal is zero-padded to 512 points, and a complex FFT is performed. Finally, the magnitudes of the complex FFT output are computed, and the harmonics are searched using a spectral peak-picking algorithm.

4.4.2.6 Aperiodic flag

The aperiodic flag is set to 1 if $V_{bpi} < 0.5$ and set to 0 otherwise. When set, this flag tells the decoder that the pulse component of the excitation should be aperiodic, rather than periodic. The above parameters are quantised by using scalar and vector quantisers and transmitted at 2400 bps. Table 4.3 shows the bit allocation for the MELP parameters where the frame length is 22.5 ms.
Table 3: Bit allocation of speech parameters
in the standard 2400 bps MELP coder [55].

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Bit allocation</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSF’s</td>
<td>25</td>
</tr>
<tr>
<td>Fourier Magnitudes</td>
<td>8</td>
</tr>
<tr>
<td>Gain(2 per frame)</td>
<td>8</td>
</tr>
<tr>
<td>Pitch</td>
<td>7</td>
</tr>
<tr>
<td>Bandpass voicing</td>
<td>4</td>
</tr>
<tr>
<td>Aperiodic flag</td>
<td>1</td>
</tr>
<tr>
<td>Sync bit</td>
<td>1</td>
</tr>
<tr>
<td>Total bits</td>
<td>54</td>
</tr>
</tbody>
</table>

4.4.3 The MELP decoder

The received bits are first unpacked and assembled into parameter codewords. Pitch is decoded first, as it dictates reception of either voiced or unvoiced speech segment at the decoder. The decoder reconstructs speech signal for each pitch cycle. The MELP synthesis parameters are interpolated pitch-synchronously for each synthesised pitch period. These parameters include the gain, LSF’s, pitch, jitter, Fourier magnitudes, pulse and noise coefficients for mixed excitation, and spectral tilt coefficient for the adaptive spectral enhancement filter. A number of Fourier magnitudes required to obtain pulse excitation should be half of a pitch period. Since only the first ten Fourier magnitudes are estimated using the normalised residual signal and provided in the decoder, the remaining magnitudes are considered as unity. Then, inverse DFT is applied to the resulting Fourier magnitudes to obtain pulse excitation. On the other hand, a uniform number generator is used to obtain the noise excitation pattern. The pulse and noise excitation signals are then filtered and summed to obtain the mixed excitation signal. Prior to synthesising speech, the mixed excitation signal is spectrally enhanced. The filter coefficients of this enhancement stage are generated by bandwidth expanded LP filter transfer function. Synthesised speech is obtained by filtering the spectrally enhanced excitation signal with the LP synthesis filter (whose coefficients are generated from the transmitted LSFs), gain modification and filtering with a pulse dispersion filter [7][55].
4.3 Conclusions

The fundamentals of the WI and the MELP coders have been presented in this chapter. The required parameters for speech modelling and the suitable techniques to estimate these parameters were discussed for both coders. The accuracy of the estimated pitch plays a main role in estimating the other parameters in these coders. For instance, the accurate pitch is required to extract the CW in WI encoder and reconstruct phase track in WI decoder. In MELP, the voicing level strengths are estimated using the estimated pitch value. Thus, inaccuracy of the estimated pitch can lead to incorrect estimation of the other parameters. This can affect the performance of these coders. This is shown as WI and MELP limitations in Chapter 5 and the proposed solution is described as a pre-processor in Chapter 6.
Chapter 5

Limitations of WI and MELP Coders

5.1 Introduction

Accurate estimation of speech model parameters plays a major role in the performance of a speech coder. Most speech coders assume that pitch evolves smoothly within a frame. Thus, the pitch is usually estimated and transmitted once per frame and in some speech coders, for instance WI, the intermediate pitch values are obtained by interpolation between two adjacent pitch periods. However, due to the non-stationary character of speech signals, the pitch occasionally has irregular variations which leads to inaccurate pitch estimates. Since the estimated pitch is usually used for determination of the remaining speech parameters, e.g. voicing strengths computation in a MELP coder and CW extraction in a WI coder, inaccurate estimation of pitch corresponds to inaccurate estimation of these parameters. The aim of this chapter is to present the limitations of WI and MELP coders.

5.2 WI Limits

Although the pitch usually evolves smoothly during a frame, it sometimes has non-linear variations, which can affect the accuracy of speech parameters estimation. Obviously, the inaccurate estimation of the speech parameters can degrade the perceptual quality of the synthesised speech, despite having an accurate model for speech generation and small-added quantisation noise. In the following sections, the effect of irregular pitch variations on WI coder is presented, where WI coder parameters are unquantised.

5.2.1 WI encoder limits

The CW extraction procedure was described in section 4.2.1.1. In order to show limitation of WI, the speech signal depicted in Fig. 5.1 is applied to the WI encoder. The first step involves that the residual signal is obtained from the speech signal using the LP analysis filter. The
corresponding residual signal is shown in Fig. 5.2 (a). Next, the pitch period is estimated and the intermediate pitch values are obtained by interpolation of two adjacent pitch periods as described in section 3.3.4. In Fig. 5.2 (b) the corresponding pitch values are shown. It is observed that when the pitch has non-linear variations the estimated pitch values differ from the real ones. Next, the CWs are extracted and aligned to construct a CW surface. The CW extraction is performed every 2 ms.

![Speech signal with irregular pitch variations. Dash lines show frame boundaries.](image1)

**Figure 5.1:** Speech signal with irregular pitch variations. Dash lines show frame boundaries.

![Residual Signal](image2)

**Figure 5.2:** (a) The last four frames of the resulting residual signal of Fig. 5.1, (b) The estimated pitch values (dash-dot lines) in comparison with the real pitch values (solid lines).
Fig. 5.3 shows some of the extracted CWs. For proper operation, the extracted CW is expected to contain only one pitch pulse (Fig. 5.3-a). However, due to incorrect intermediate pitch values, when the interpolated pitch is longer than the real pitch value, there is an extra pitch pulse within the extracted CW (Fig. 5.3-b). On the other hand, when the interpolated pitch is shorter than the real one, the CW does not contain any pitch pulses (Fig. 5.3-c).

After the alignment and the normalisation procedures on the extracted CWs, the CW surface is constructed. Fig. 5.4 shows the resulting CW surface for the second and third frames. For voiced speech, the resulting CW surface is expected to have maximum smoothness in CW-extraction time direction (index direction)[44]. However, due to incorrect pitch estimations and corresponding CWs, some of the pitch pulses are not phase aligned. Next, the CW surface is decomposed into SEW and REW by low-pass and high-pass filtering with a cut-off frequency of 20Hz. It is expected that the voiced speech is transferred to the SEW surface, but the misalignment of the pitch pulses causes the bandwidth of the evolution spectrum to exceed 20Hz.
Figure 5.4: The resulting CW surface for the second (a) and third (b) frames of the residual signal depicted in Figure 5.2 (a).
As a result, a part of the voiced speech is transferred into the REW surface and the decomposition is performed incorrectly. Fig. 5.5 and 5.6 show the resulting SEW and REW surfaces for the second and third frames, respectively.

As described in section 4.3.3, for the WI coder operating at low bit rates, the SEW is transmitted only once per frame. Due to misaligned pitch-pulse locations, the resulting SEW
surface is unsuitable for down sampling [44] and this procedure is performed incorrectly. This can affect the reconstructed CW surface and also the synthesised speech at the WI decoder as described in the following section.

Figure 5.6 : The resulting SEW (a) and the REW (b) surfaces of the third frame.
5.2.2 WI decoder limits

As described in the last section, due to existing misaligned pitch-pulse locations in the SEW surface, the shape of the down sampled (transmitted) SEW depends on the time-index selection. Fig. 5.7 shows two different shapes of the transmitted SEW as:

a) The SEW including one pitch-pulse,

b) The SEW including two pitch-pulses.

![Figure 5.7: Two possible SEWs, (a) The SEW containing two pitch pulses, (b) The SEW containing one pitch pulse.](image)

After up-sampling the SEW and the REW surfaces, the CW surface is reconstructed. In Fig.5.8, the resulting CW surface is shown for two different shapes of the SEW where the CW surface has been up-sampled by factor 4 for clarity. It is observed that in the area where there is no pitch pulse in the analysis stage, some pitch pulses are created. On the other hand, in the area where there is a pitch pulse in the analysis stage, the pitch pulse has disappeared. Next, in order to convert two-dimensional surface to a one-dimensional signal, the phase track is reconstructed using the pitch information. Fig.5.9 shows the reconstructed residual signal caused by two different shapes of the SEW for the second frame.

It is observed that the reconstructed residual signal in comparison with the original one includes an additional pitch pulse in the area where the estimated pitch value is larger than the
real pitch. In contrast, the pitch pulse has disappeared in the area where the estimated pitch value is shorter than the real one.

Figure 5.8: Two resulting CW surfaces using two different shapes of the SEW. (a) The resulting CW surface using the SEW including two pitch pulses, (b) The resulting CW surface using the SEW including one pitch pulse.

Figure 5.9: The reconstructed residual signal using two different shapes of the SEW in comparison with the original residual signal, (a) The original residual signal, (b) The reconstructed residual signal using the SEW containing two pitch pulses, (c) The reconstructed residual signal using the SEW containing one pitch pulse.
5.3 MELP coder limitations

In the other experiment, the effect of irregular pitch variations on the standard MELP is studied. First, parameter estimation errors, due to inaccurate pitch, are presented in the analysis stage and then the effect of these errors on reconstructed speech is shown in the synthesis stage.

5.3.1 Pitch estimation error

A MELP frame interval is 22.5 ms in duration. For integer pitch calculation, the input speech signal is first processed with a 1KHz, 6th order Butterworth low pass filter. The integer pitch value, \( P_i \), is the value of \( x \), where \( x = 40, 41, \ldots, 160 \), for which the normalised autocorrelation function, \( r(x) \), is maximised. This function is defined by [55]:

\[
r(x) = \frac{C_x(0,x)}{\sqrt{C_x(0,0)C_x(x,x)}} \tag{5.1}
\]

Where

\[
C_x(m,n) = \sum_{k=-\lceil x/2 \rceil}^{\lceil x/2 \rceil} s_{k+m}s_{k+n} \tag{5.2}
\]

and \( \lceil x/2 \rceil \) represents the truncation to an integer value. The centre of the pitch analysis window is at sample \( s_0 \) in Eq. (5.2). This window is centred on the last sample of the current frame. Next, the fractional pitch refinement is applied. This procedure utilizes an interpolation formula to increase the accuracy of an input pitch value. This value is first rounded to the nearest integer, \( T \) samples. The interpolation formula presumes that \( r(x) \) has a maximum between lags of \( T \) and \( T + 1 \). Hence, \( C_x(0,T-1) \) and \( C_x(0,T+1) \) are computed and compared to determine if the maximum is more likely to fall between \( T \) and \( T + 1 \) or between \( T - 1 \) and \( T \).

If \( C_x(0,T-1) > C_x(0,T+1) \), then the maximum probably falls between \( T - 1 \), \( T \) and the pitch, \( x \), is decremented by one prior to interpolation. The fractional offset, \( \Delta \), is then computed using the interpolation equation:

\[
\Delta = \frac{C_x(0,T+1)C_x(T,T) - C_x(0,T)C_x(T,T+1)}{C_x(0,T+1)[C_x(T,T) - C_x(T,T+1)] + C_x(0,T)[C_x(T+1,T+1) - C_x(T,T+1)]} \tag{5.3}
\]
Chapter 5. Limitations of WI and MELP Coders

Where $C_T(m,n)$ is defined by Eq. (5.1). In some cases, this formula produces an offset outside the range of 0.0 to 1.0, so the offset is clamped between $-1$ and $2$. The fractional pitch is $T + \Delta$ and is clamped between 20 and 160.

In order to present the effect of irregular pitch variations on the pitch estimation, the speech signal depicted in Fig. 5.1 is applied to the MELP coder. In order to evaluate the performance of coder, we focus on three frames, where the first one has regular pitch variations and the others have the irregular variations. Fig. 5.10 shows these frames with the local pitch values. The estimated pitch values are 73.8, 73.2 and 71.3 samples respectively. Then, the normalised autocorrelation function, $r(\tau)$, for the first and second lowpass filtered speech frames is calculated for $\tau = 40, 41, \ldots, 160$. The resulting functions are shown in Fig. 5.11.

Due to smooth pitch variations of the first frame, the maximum of the function $r(\tau)$, is much higher than the other values and so the estimated pitch value is close to the local pitch value. In the second frame, although the function $r(\tau)$ is maximised at $\tau = 73$, the other local maximum placed at $\tau = 103$ may be a candidate for the estimated pitch value. Fig. 5.12 shows the spectrum of the second lowpass filtered speech frame. For clarity, only the frequency samples under 2KHz are shown. It is observed that the fundamental frequency is actually 78 Hz, which means that the pitch value is equal to 103 samples.

![Figure 5.10: Three speech frames with local pitch values; the first frame has smooth pitch evolution and the second and third have irregular pitch variation.](image)

---

Where $C_T(m,n)$ is defined by Eq. (5.1). In some cases, this formula produces an offset outside the range of 0.0 to 1.0, so the offset is clamped between $-1$ and $2$. The fractional pitch is $T + \Delta$ and is clamped between 20 and 160.

In order to present the effect of irregular pitch variations on the pitch estimation, the speech signal depicted in Fig. 5.1 is applied to the MELP coder. In order to evaluate the performance of coder, we focus on three frames, where the first one has regular pitch variations and the others have the irregular variations. Fig. 5.10 shows these frames with the local pitch values. The estimated pitch values are 73.8, 73.2 and 71.3 samples respectively. Then, the normalised autocorrelation function, $r(\tau)$, for the first and second lowpass filtered speech frames is calculated for $\tau = 40, 41, \ldots, 160$. The resulting functions are shown in Fig. 5.11.

Due to smooth pitch variations of the first frame, the maximum of the function $r(\tau)$, is much higher than the other values and so the estimated pitch value is close to the local pitch value. In the second frame, although the function $r(\tau)$ is maximised at $\tau = 73$, the other local maximum placed at $\tau = 103$ may be a candidate for the estimated pitch value. Fig. 5.12 shows the spectrum of the second lowpass filtered speech frame. For clarity, only the frequency samples under 2KHz are shown. It is observed that the fundamental frequency is actually 78 Hz, which means that the pitch value is equal to 103 samples.

![Figure 5.10: Three speech frames with local pitch values; the first frame has smooth pitch evolution and the second and third have irregular pitch variation.](image)
However, because of irregular local pitch variations, the local maximum of the function $r(\tau)$ can be close to the absolute maximum and the decision based on the absolute maximum may lead to incorrect pitch estimations.

![Figure 5.11: The normalised autocorrelation function for the first (a) and second (b) frames.](image1)

![Figure 5.12: The spectrum of the second lowpass filtered speech frame.](image2)
5.3.2 Voicing strengths error

The voicing strengths estimation procedure was described in section 3.5.2. Now, the effect of irregular pitch variation on the voicing strength estimations is studied. Fig. 5.13 and 5.14 show five frequency bands of the first and second speech frames of Fig. 5.10 with the corresponding voicing decisions. As shown in Fig. 5.10, despite having strongly voiced speech for the second frame, only the first band is considered as voiced whereas all frequency bands of the first frame are estimated as voiced. This especially can be observed in Figure 5.14-b, where despite the existing harmonics in the second band, this band is declared unvoiced. In order to find out the correct voicing strength for the second band, we calculate the normalised autocorrelation function of the first and second bands for the first and second frames when the pitch changes from 40 to 160 samples. Results are shown in Fig. 5.15. For the first frame, the maximum of the normalised autocorrelation function for the first and second bands are placed at the same pitch lag, which actually is the estimated pitch. However, for the second frame, firstly, because the maximum of the normalised autocorrelation function of the second band is higher than the voicing strength threshold, this band is declared voiced. Secondly, due to irregular pitch variations, the maximum values of the normalised autocorrelation function of the first and second bands for the second frame are placed in different pitch lags. As shown in Fig. 5.15-b, the maximums are placed in pitch lag of 74 and 103 samples for the first and second bands, respectively. The second band is considered as incorrectly unvoiced because the corresponding normalised autocorrelation value at the estimated pitch period (which is 73.2 samples) is less than the voicing strength threshold. In the other experiment, the voicing strength of the second band is forced as voiced. The mixed excitation signals reconstructed by the decoder, for the real (the voicing strength of the second band is unvoiced) and manual cases are obtained. Next, the reconstructed spectrum is compared with the original spectrum for each band based on the voicing measure used in MBE. Results are shown in table 5.1.

<table>
<thead>
<tr>
<th>Table 5.1: Voicing measure used in MBE for evaluation of the estimated voicing strengths.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
</tr>
<tr>
<td>Frame 1.</td>
</tr>
<tr>
<td>Frame 2 (Real case)</td>
</tr>
<tr>
<td>Frame 2 (Manual Case)</td>
</tr>
</tbody>
</table>

Dₖ : Voicing measure for kᵗʰ band.
Chapter 5. Limitations of WI and MELP Coders

Figure 5.13: The five frequency bands of the first frame.

Figure 5.14: The five frequency bands of the second frame.
Figure 5.15: The normalised autocorrelation function of the first and second frequency bands for the first (a) and second (b) frames.
It is observed that in the first frame, where the pitch evolves smoothly, the voicing measure is lower in comparison with the second frame. In addition, in the second frame, when the second band is forced as voiced, the voicing measure for each band reduces but it is still higher than the first frame one. The reason is that both the state of voiced/unvoiced for each band and the estimated pitch are used in the voicing measure based on the MBE coder. In other words, as shown in Fig. 5.11-b, unreliable normalised autocorrelation results in the inaccurate pitch estimated, which affects the voicing measure.

In the following section the effect of estimation errors on the reconstructed speech is presented.

![Figure 5.16: The original and the reconstructed first band speech in time domain for the first (a: original, b: reconstructed) and the third (c: original, d: reconstructed) frames.](image)

**5.3.3 Reconstruction errors**

During the synthesis stage, the estimated parameters such as the voicing strengths and the pitch value are used to reconstruct speech signal. We present the effect of the inaccurate estimated pitch value and voicing strengths on the synthesised speech.
Chapter 5. Limitations of WI and MELP Coders

Since the MELP decoder reconstructs speech signal based on pitch cycle synthesis, the inaccurate estimated pitch may result in additional or absence of pitch pulse in case of the estimated pitch being shorter or larger than the real one, respectively. For instance, Fig. 5.16 shows the original and synthesised speech signals. It is observed that one pitch pulse has disappeared from the third synthesised speech frame. Since the length of the synthesised frame is 180 samples and the estimated pitch value for the third frame is 114 samples, the third synthesised frame can only contain two pitch pulses.

Next, the effect of inaccurate voicing strengths on the reconstructed speech is studied. As can be seen in Fig. 5.17-b, the periodicity of the second and the third synthesised speech frames has been considerably affected whereas the periodicity of the first frame has been well maintained. For the second and the third frames, the estimated pitch values are 73 and 114 samples, respectively and only the first frequency band is declared voiced whereas all frequency bands of the first frame are declared voiced. Fig. 5.18 shows the original and the synthesised speech spectrums of the first and third frames. As shown in Fig. 5.18-b, due to the accurate voicing strengths estimation of the first frame, the pitch harmonics appear in all frequency bands as in the original spectrum.
Figure 5.18: The original and the reconstructed speech spectrums of the first and third frames.
However, for the third frame, since a noise sequence filtered by the noise-shaping filter represents the residual spectrum for frequency band of 500-4000Hz, the pitch harmonics only appear in the frequency band of 0-500Hz. This can be observed in Fig. 5.18-d, where despite having pitch harmonics in the original spectrum, especially in frequency band of 500-1000 Hz, no harmonics appears in the synthesised frequency band of 500-4000 Hz.

5.4 Conclusions

This chapter has presented the limitations of WI and MELP coders with regards to estimation of speech model parameters. These parameters are pitch (on WI and MELP) and voicing strengths (on MELP). Irregular pitch variations were introduced as main factor causing these limitations. Thus, pitch pre-processor can be applied before a speech coder to improve pitch regularity such that the essential parameters can be more effectively estimated from the pre-proced speech signal. This is achieved in the pre-processor described in Chapter 6.
Chapter 6

Pitch Modification

6.1 Introduction

Most low bit rate speech coders rely on the pitch estimate in determination of other speech model parameters [6][7]. For instance, in a MELP coder, voicing strengths are estimated using the normalised autocorrelation values computed at the estimated pitch lag. In a sinusoidal coder, the speech waveform is represented by the sum of sine waves, the frequency of sinusoids is computed using the estimated pitch value. In a WI coder, the estimated pitch value is used to determine the length of the CWs during the CW extraction procedure in the analysis stage and to construct a phase track, which is used to convert a two-dimensional surface to a one-dimensional signal during synthesis stage.

The basic assumptions for pitch estimation algorithms are:

- Voiced speech samples are correlated at specific time intervals called pitch period. Although usually these samples are usually highly correlated, they sometimes have low correlation and as a result the resulting normalised autocorrelation is unreliable for pitch estimation.

- Pitch evolves smoothly during a voiced frame. Although this usually happens, as discussed in Chapter 5, pitch occasionally has irregular variations, which can lead to inaccurate pitch estimation. The irregular pitch variations usually happen in transition frames where speech characteristics change from quasi-periodic to random like signal or vice versa and therefore the long-term correlation of speech signal is affected.

In order to overcome these problems, a methodology that provides more accurate estimation of the parameters is required. This methodology can be performed either inside a speech coder using alternative algorithms or outside as a pre-processor. In the first case, no modification is performed on the speech signal and alternative factors such as history and future of a parameter, spectral matching and etc. are used to find a better estimation. Whereas making more regular speech is the objective of the second method. This enables the algorithms used in speech coder to a more accurate pitch. This technique is called speech pre-processing since the speech signal is modified before passing to speech coder.
Chapter 6. Pitch Modification

This chapter is organised as follows: In the next section the existing pre-processors and techniques for pitch modification are presented. In section 6.3, we propose and describe a new pre-processor, which leads to smooth pitch evolutions and provides more regular speech such that the required parameters can be effectively estimated in a speech encoder. In section 6.4, the problem caused by misalignment between the LP filter and its residual is described and in order to overcome this problem, we propose a new methodology as pre-analysis and post-processing. In section 6.5, the effect of the new pre-processor is evaluated in i) pitch estimation ii) voicing level estimation iii) subjective listening tests. This is performed by comparing the case which the pre-processor is in combination with the standard MELP 2.4 Kb/s and the MELP alone. Finally, the new pre-processor is evaluated in background noise in section 6.6. The Chapter ends with section 6.7 where the conclusions are drawn.

6.2 Existing Pre-processors and pitch modification techniques

In the following section we address current pre-processors employed by speech coders and current techniques for pitch modification.

6.2.1 Existing pre-processors

Existing pre-processors have been designed for special coders. For instance, in [56], Kleijn introduces a pre-processor in combination with a block-DFT based WI coder. This coding structure maintains the advantages of earlier WI coders and adds the asymptotically perfect reconstruction property. Since the alignment procedure employed by a WI encoder causes the relative phase loss of the CW, the WI coder does not give perfect reconstruction [56]. The alignment procedure is included in as a part of the pre-processor performed outside the WI coder. This is performed by moving a pitch pulse to the centre of a CW such that the modified segment is maximally correlated with the previous cycle and thus the employed alignment procedure in earlier WI coders is not required.

The pre-processor employed in [57] performs high-pass and adaptive noise suppression filtering before estimation of speech parameters. After speech parameter estimation, the residual signal is modified to generate a target residual for the fixed codebook search in a CELP coder. A shifted target residual is generated using the past-modified residual and the delay contour of the current frame. This shifted residual is used as a target for shifting the residual of the current subframe. All pitch pulses in the original residual are shifted individually to match the delay contour of the modified target residual.
6.2.2 Existing pitch modification techniques

Existing pitch-modification techniques can be classified into two groups. The aim of the first group is rendering speech at an arbitrary rate different from the original rate. This can increase the ease-of-use and the efficiency of speech reproduction equipment. A number of algorithms for high-quality time and pitch scaling have been reported in [16]. In fact, the common character of these algorithms is to change the duration of all pitch cycles by an arbitrary factor. In other words, all pitch cycles are scaled in the same way and thus in the existing techniques the irregular pitch variations present in the modified speech.

In the second group, the alternative techniques are applied to the original speech before further parameter analysis is performed. For instance, in [58][59], a time warper is applied to the original speech to enhance the stationarity of voiced speech segments. The main feature of the employed time warper is to remove the part of the frequency variation, which progresses linearly with time, without changing the time duration of that segment. As a result, irregular pitch-pulse locations remain in the modified speech and time-domain algorithms employed for pitch and voicing level estimation may fail. In addition, the quality of the modified speech is not necessarily maintained perceptually.

We propose a new pre-processor, which enhances the regularity of speech signal and also maintains the perceptual speech quality and thus can be used in combination with any speech coder.

6.3 Pre-processor description

The proposed pre-processing algorithm modifies the residual signal such that it is more convenient for coding. During the modification, the lengths of the pitch cycles in a frame are altered so as to evolve more smoothly. The modification is based on the local pitch estimation. The local pitch values are only used in the encoder and only the refined pitch value is transmitted to the decoder once per frame. The input to the pre-processor is the linear prediction residual of the speech signal and an associated pitch track. The pitch period is estimated once in a frame using conventional autocorrelation based methods and the resulting estimate is then linearly interpolated for each pitch cycle. The output is a modified linear prediction residual, which is constructed by concatenation of modified/unmodified pitch cycles of the residual signal.

If the frame is unvoiced, no modifications are performed to the pitch value. During voiced sections the main task of the pre-processor is to smooth the pitch values of the pitch cycles while keeping long-term correlation of the speech signal and also maintaining the perceptual
speech quality identical to the original. In the following sections, the operation of the pre­
processor on a step-by-step basis is discussed.

6.3.1 Local pitch estimations

A simple method based on the energy of the LPC residual is employed by the Telecommunication Industry Association (TIA) Enhanced Variable Rate Coder (EVRC) [57],
to detect the pitch pulses. The EVRC computes the pitch pulse locations by searching for a maximum in a five-sample moving energy window within a region one and half times the pitch period, and then finds the rest of the pitch pulses by searching recursively at a separation of one pitch period. In [60], the performance of the residual energy based pitch pulse location is improved using the Hilbert Envelope of Windowed LP Residual (HEWLPR). A robust pitch pulse detection algorithm based on the group delay of the phase spectrum has also been reported [61].

The proposed pre-processor requires a pitch pulse detection algorithm, which can detect the pulses at stationary voiced segments with high accuracy. Therefore, an improved pitch pulse detection algorithm as the algorithms used in the EVRC energy based pitch pulse detection and the HEWLPR is proposed for the pre-processor. In fact, the EVRC and the HEWLPR algorithms compute all possible pitch pulse locations and if there is a difference between the determined locations, the corresponding pitch pulses are modified during the first stage of pitch modification (section 6.3.2). Thus, after applying these two algorithms, two vectors are created. One containing the valid pitch pulse locations, \( V_{val} \), and the other one having the invalid pitch pulse locations, \( V_{inv} \). These vectors may be updated during the pitch contour construction. However, pitch pulse refinement (section 6.3.2) and pitch cycle modification (section 6.3.3) are performed using \( V_{inv} \) and \( V_{val} \) information.

In [62], it is proposed to use localised energy and adaptive threshold for pitch pulse locations. Initially all the possible pitch pulse locations are determined by using the localised energy of the residual signal, \( r(n) \), and an adaptive threshold, \( t(n) \). The localised energy, \( e(n) \), is computed by moving a rectangular window with length of five samples across \( r(n) \), and is given by equation 6.1:

\[
e(n) = \frac{1}{5} \sum_{i=-2}^{2} |r(n+i)| \quad 2 \leq n < N - 2
\] (6.1)
Where $N$ is sum of the current frame length, which is 180 samples and the buffered samples of the previous and the next frame. The length of this buffer is considered to be 20 samples. The adaptive threshold function $t(n)$ is updated for each half pitch period by taking 0.65 of the maximum of $e(n)$ in the pitch period symmetrically centred around the half pitch period chosen to calculate $t(n)$, and $t(n)$ is given by equation 6.2.

$$
t(n_k - T_{1/4} + n_{T/2}) = 0.65 \max [e(n_k - T_{1/2} + n_T)]
$$

for $0 \leq n_T < T_1$ and $0 \leq n_{T/2} < T_{1/2}$  

(6.2)

Where $T_i = \left\lfloor \frac{T + \frac{1}{2}}{2} \right\rfloor$, $T_{1/2} = \left\lceil \frac{T}{2} + \frac{1}{2} \right\rceil$, $T_{1/4} = \left\lfloor \frac{T}{4} + \frac{1}{2} \right\rfloor$, $n_k = kT_{1/2}$ for $1 \leq k < \left\lfloor \frac{2N}{T} \right\rfloor$, and $T$ is the pitch period.

For frame boundaries, the previous and next frame samples are used for localised energy. The initial amounts of the threshold function, $t(n)$, is given by equation 6.3.

$$
t(m) = 0.65 \max [e(T_{1/2})] \text{ for } 0 \leq m < T_{1/4}
$$

(6.3)

The samples locations, for where $e(n) > t(n)$, are considered as regions which may contain pitch pulses if these regions are part of the current frame. If $e(n) > t(n)$, for more than eight consecutive samples, those regions are ignored, since in those regions the residual energy is smeared, which is not a feature of pitch pulses. The maximum amplitude of each remaining region is considered as a possible pitch pulse location. If any of the two candidate locations are closer than 16 samples (which is considered the minimum pitch period) from each other only one, which has the higher local energy, is taken and the other one is considered as invalid pitch pulse and so a member of vector $V_{inv}$.

Applying an adaptive threshold to estimate the pitch pulse locations from the localised energy is advantageous, especially for the segments where the energy of the LPC residual varies rapidly. Figure 6.1 shows this event for two occasions, (a) a male offset and (b) a female onset. The male speech frame has a pitch period of about 76 samples and two high-energy invalid pitch pulses. The female speech frame has a pitch of about 78 samples, which also consist of one high-energy invalid pulse. The energy function $e(n)$ and the threshold function $t(n)$ are also depicted in Figure 6.1, both $e(n)$ and $t(n)$ are shifted upwards for clarity. The figures also show that $e(n)$ at the invalid pulses may be higher than $e(n)$ at the valid pitch
pulses. Therefore selecting the highest $e(n)$ to detect a pitch pulse location as in [57] may lead to errors. However, since $e(n) > t(n)$, for some of the irregular pulses as well as for valid pitch pulses, further refinements are required.

Due to reducing the competitions, the refinements are applied if the standard deviation of the vector including the estimated pitch-pulse locations is more than ten percent of the vector mean.

The proposed pitch pulse locations refinement in [62] relies on the accuracy of the estimated pitch value, but as it was demonstrated in Chapter 5, in the case of irregular pitch variations, the estimated pitch value can be incorrect. Therefore, in order to obtain the valid pitch pulse locations two-stage refinement is proposed.

![Figure 6.1: (a) Male offset speech frame including two high-energy irregular pitch pulses, (b) Female onset speech frame including one high-energy irregular pitch pulse.](image)

6.3.1.1 Pitch-pulse location refinement: First stage

As shown in the previous section, both the valid and invalid pitch pulses are detected using concentrated energy measure and adaptive threshold. The HEWLPR with adaptive threshold is used to separate the invalid pitch pulses from the valid ones.

In [60], it is assumed that the speech signal within a pitch period is induced by a pulse at one epoch or an event. This epoch is defined as a representation of the Glottal Closure Instant (GCI), because the GCI induces the sound vibration and introduces most of the energy within each pitch period. An epoch occurs when the conditional probability density, or likelihood function, of the epoch is maximised. This can also be performed through the maximisation of
Chapter 6. Pitch Modification

a function called Maximum-Likelihood Epoch Determination (MLED) signal which is defined by equation 6.4 [60].

\[ \hat{f}(n_0) = \sum_{n=0}^{L-1} s(n + n_0) \hat{s}(n) \quad \text{for} \quad 0 \leq n_0 < N \]  (6.4)

Where \( N \) and \( L \) are the frame size and the length of the wavelet due to an epoch, \( \hat{s}(n) \).

Assuming that speech production can be modelled as an all-pole linear system, z-transform of the wavelet due to an epoch can be expressed by the equation 6.5.

\[ \hat{S}(z) = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-1}} \]  (6.5)

Where \( a_i \) and \( p \) are the coefficients and the order of the polynomial respectively. The MLED creates not only a strong and sharp epoch pulse, but also a set of weaker pulses representing the suboptimal epoch candidates within a pitch period. The energy ratio between the valid epoch pulse and the sub-pulses varies largely which results in ambiguity for pitch pulse location determination. This was overcome using a selection signal, \( g(n) \), superimposed on the GCI’s:

\[ g(n_0) = \left[ \hat{f}^2(n_0) + \hat{f}_H^2(n_0) \right]^{1/2} \]  (6.6)

Where \( \hat{f}_H(.) \) is the Hilbert transform of \( \hat{f}(.) \) which can be identified as a filter with the transfer function:

\[ H(\omega) = \begin{cases} -j & 0 < \omega < \pi \\ 0 & \omega = 0, \pi \\ j & -\pi < \omega < 0 \end{cases} \]  (6.7)

Finally, the GCI Determination Signal (GCIDS) is the MLED signal \( \hat{f}(n_0) \) multiplied by the selection signal \( \hat{g}(n_0) \):

\[ \theta(n_0) = \hat{f}(n_0) \cdot \hat{g}(n_0) \]  (6.8)
Chapter 6. Pitch Modification

Where

\[
\hat{g}(n_0) = \begin{cases} 
  g(n_0) - \bar{g}(n_0) & \text{if } g(n_0) \geq \bar{g}(n_0) \\
  0 & \text{if } g(n_0) < \bar{g}(n_0)
\end{cases}
\]  

(6.9)

and

\[
\bar{g}(n_0) = \frac{\sum_{n_0=0}^{N-1} g(n_0)}{N}
\]

(6.10)

The GCIDS compared with the MLED provides stronger and sharper epoch pulse and weaker sub-pulses, which is more suitable for more pitch pulse location determination. The following steps are performed to the GCIDS.

1. \( \hat{s}(n) \) is computed using 4ms of an impulse-response of the LP inverse filter.
2. The MLED signal is calculated through Eq. (6.4).
3. The Hilbert transform of the selection signal is computed using 256-points FFT of the MLED signal multiplied by \( H(\omega) \) and then taking the resulting FFT.
4. The selection signal is obtained using Eq. (6.6) and subtracted from its mean. The resulting signal is rectified so that the positive values are maintained.
5. The GCIDS is computed by multiplication of the MLED and the rectified selection signals.

Fig. 6.2 shows the residual signal and the corresponding GCIDS in comparison with localised energy variations. The experiments show that the valid pitch pulse locations appear in the GCIDS and the invalid pitch pulse locations are mostly removed and in few cases still appear with very low energy. Thus, the searching pitch-pulse procedure using the adaptive threshold is performed on the GCIDS signal. Consider \( n_p \) to be the pitch-pulse location found by the algorithm described in section 6.3.1. The maximum energy of GCIDS at interval \([n_p-2, n_p+2]\) is compared with the corresponding threshold. If it is higher than 95 percent of the threshold, the found pitch-pulse at \( n_p \) is considered as a valid pitch pulse. Otherwise, this is considered as an invalid pitch pulse and is refined by procedure described in section 6.3.2.

Assuming \( V_{evrc} \) and \( V_{hevlp} \) are the vectors which include pitch pulse locations obtained in section 6.3.1 and the HEWLPR refinement, we compute vectors \( V_{val} \) and \( V_{inv} \) as follows:

\[
V_{val} = V_{evrc} \cap V_{hevlp}
\]

\[
V_{inv} = V_{evrc} - V_{hevlp}
\]

(6.11)
6.3.1.2 Pitch contour construction: second stage of pitch pulse locations refinement

As it was described in Chapter 5, the pitch contour based on the estimated pitch values can be inaccurate. Therefore we construct the pitch contour based on the local pitch values in $V_{val}$. The aim of the pitch contour construction is to change the irregular pitch values such that the pitch evolves smoothly during a frame. This is performed using the history and the future of the pitch variations. In order to have the history and the future of the pitch variation, we keep the local pitch information of the previous frame and also estimate the valid local pitch values of the next frame. If $V_{val-prv}$, $V_{val-curr}$ and $V_{val-next}$ are the valid pitch pulse locations of the previous, current, and next frame respectively, the vector $V_{valid}$ is defined by equation 6.12.
\[
V_{\text{valid}}(k) = \begin{cases} 
V_{\text{val-prv}}(k)-N & 1 \leq k \leq L_{\text{val-prv}} \\
V_{\text{val-cur}}(k-L_{\text{val-prv}}) & 1 \leq k - L_{\text{val-prv}} \leq L_{\text{val-cur}} \\
V_{\text{val-next}}(k-L_{\text{val-prv}}-L_{\text{val-cur}})+N & 1 \leq k-L_{\text{val-prv}}-L_{\text{val-cur}} \leq L_{\text{val-next}} 
\end{cases}
\] (6.12)

Where \( L_{\text{val-prv}} \), \( L_{\text{val-cur}} \) and \( L_{\text{val-next}} \) are the dimension of the vectors \( V_{\text{val-prv}} \), \( V_{\text{val-cur}} \) and \( V_{\text{val-next}} \), respectively and \( N \) is the frame size. There are two exceptions for onset and offset frames. Due to the non existence of the previous and the next pitch pulse locations for onset and offset frames, respectively, the locations of two next and two previous frames are used to construct \( V_{\text{valid}} \). Next, the difference between each two successive elements of the vector \( V_{\text{valid}} \) is calculated. The resulting vector, \( V_{\text{local}} \), indicates the local pitch variations where each local pitch cycle is started and ended by two successive pitch pulses. The irregular pitch values appear as maximum or minimum values in the vector \( V_{\text{local}} \). Figure 6.3 shows a residual signal with the corresponding vector \( V_{\text{local}} \).

The pitch contour is constructed by regular pitch values and therefore it is required that these are separated from irregular values. The separation is performed based on the pitch average. In order to obtain the pitch average, the maximum and minimum values (if existing) are ignored in calculation of the mean of the vector \( V_{\text{local}} \). The maximum and minimum values are defined as the points that have more than ten percent variation with their neighbouring points. If the vector length resulting from ignoring the maximum and minimum values is more than four, searching and ignoring these points is performed again. Next, the pitch average for each frame is obtained by calculating the mean of the resultant vector \( V_{\text{local}} \). If there is not enough information (at least two points required) to calculate the pitch average of the current frame, \( P_{\text{avg-cur}} \), it is obtained by the arithmetic mean of the previous and the next pitch average. The regular and irregular pitch values of the current frame are evaluated by the measure given in Eq. 6.13.

\[
R(k) = \frac{|V_{\text{local}}(k) - P_{\text{avg-cur}}|}{P_{\text{avg-cur}}} \quad L_1 \leq k \leq L_2
\] (6.13)

Where \( L_1 \) and \( L_2 \) select the elements of \( V_{\text{local}} \) within the current frame. If \( R(k) \) is less than ten percent, the corresponding local pitch is considered as a regular pitch value. Otherwise, pitch doubling and pitch halving is checked. In order to prevent pitch doubling, the relative \( R(k) \) is checked. If it is between 95 - 105 percent there is a possibility of pitch doubling. This
may happen when a low-energy pitch pulse is placed between two high-energy pitch pulses. Thus, maximum local energy of samples placed between 45-55 percent of the relative local pitch cycle is compared against the energies of both sides of a pitch pulse. If it is more than fifty percent of the minimum pitch-pulse energies, we consider it as being a pitch pulse in this area and therefore the vectors $V_{val}$ and $V_{local}$ are updated.

Half-pitch errors may occur if the relative $R(k)$ is between 45-55 percent. This may happen when a high-energy-irregular pitch pulse is placed between two regular pitch pulses and appeared in GICIDS. In order to avoid such half pitch errors we calculate the addition of the corresponding and the next local pitch values in $V_{local}$. If the result is close to pitch average (between 95-105 percent of pitch average), the pitch halving has happened and therefore the relative pitch pulse is discarded from $V_{val}$ and considered as an invalid pitch pulse in $V_{inv}$ and then $V_{local}$ is updated.

If the corresponding pitch is neither pitch doubling nor pitch halving, it is declared as an irregular pitch and the corresponding pitch-cycle is modified through the described procedure.

Figure 6.3: (a) The residual speech signal including irregular pitch variations, (b) The corresponding local pitch variations.
in section 6.3.3. Next, the pitch contour is constructed. Consider $V_{\text{local}}(i)$ as the irregular pitch value placed between the regular pitch values $V_{\text{local}}(m)$ and $V_{\text{local}}(n)$, $(m < i < n)$, the modified irregular pitch value is given by Eq. 6.14.

$$V_{\text{local}}(i) = V_{\text{local}}(m) + \frac{V_{\text{local}}(n) - V_{\text{local}}(m)}{n - m} (i - m)$$ (6.14)

Some exceptions may exist especially at onset and offset frames. If the irregular pitch value is the first element of the vector $V_{\text{local}}$, at onset frames or the last element at offset frames, the first and the last two regular pitch values before and after the irregular pitch are considered respectively for irregular pitch-value modification given in Eq. 6.14. Figure 6.4 shows the modified pitch values in comparison with the original ones for the residual speech depicted in Fig. 6.3(a).

![Figure 6.4: The original local pitch variations in comparison with the modified one. Stars and squares are the original and the modified pitch values respectively.](image)

Thus, at the end of this stage, the following information is provided:

- Updating the invalid pitch pulse locations in vector $V_{\text{inv}}$. This information is used in the pitch pulse refinement described in section 6.3.2.
Chapter 6. Pitch Modification

- Specifying the regular and irregular pitch values and the corresponding local pitch cycles in vector $V_{local}$.
- Estimation of the modified pitch values instead of the irregular ones based on the smooth pitch evolutions.

The last two items are used in the pitch cycle modification described in section 6.3.3.

6.3.2 Pitch pulse refinement

The aim of this stage is to refine high-energy invalid pitch-pulses to make more regular speech. The inputs are the residual speech signal and valid and invalid pitch pulse locations given by $V_{val}$ and $V_{inv}$, respectively. The proposed refinement is performed based on local-energy variations in the first regular local pitch-cycle before and after the current cycle. Therefore, firstly, the pitch cycle including the invalid pitch pulse and the previous and next regular pitch-cycles are founded by using the information of $V_{inv}$ and $V_{val}$. If $n_i \in V_{inv}$ is the invalid pitch-pulse location, the cycle $c_i$ is centred by $n_i$ with length of $L_i$. $L_i$ is selected such that no other pitch pulse is included. Next, two cycles from the regular pitch cycles are searched to have maximum correlation with $c_i$ and the same length. In order to reduce the computations, searching is performed in $[n_i-2, n_i+2]$. Assuming $c_m$ and $c_n$ are the resulting cycles, we consider a moving window with length of five samples to compute the local energy for $c_m$, $c_n$ and $c_i$ excluding the invalid pitch pulse energy. The assumption that refines the invalid pitch-pulse is that the local energy of the samples around the pitch-pulse (excluding the pitch-pulse) changes as a fraction of the local energy of the corresponding samples in the previous and the next cycle. Thus, we define the factor $\beta_1$ and $\beta_2$ as given by equation 6.15.

$$
\beta_1 = \frac{1}{M} \sum_{k=1}^{M} e_i(k) \\
\beta_2 = \frac{1}{M} \sum_{k=1}^{M} e_n(k) \\
M = \left\lfloor \frac{L_i}{5} \right\rfloor
$$

(6.15)

Where $e_i(.)$, $e_m(.)$ and $e_n(.)$ are the local energies of the cycles $c_i$, $c_m$ and $c_n$, respectively. Assuming $E_{ii}$, $E_{mi}$ and $E_{ni}$ as the invalid pitch-pulse energy and the local energy of the
corresponding samples in \( c_m \) and \( c_n \), the irregular pitch pulse and the two samples around it are normalised by a factor \( \lambda \) such that the conditions given in equation 6.16 are satisfied.

\[
\frac{E_{ii}^m}{E_{mi}} = \beta_1
\]

\[
\frac{E_{mi}}{E_{ii}} = \beta_2
\]

(6.16)

Where \( E_{ii} \) is the energy of the modified invalid pitch-pulse:

\[
E_{ii}^m = \frac{E_{ii}}{\lambda^2}
\]

(6.17)

Since the conditions given in Eq. 6.16 may not be satisfied together using the factor \( \lambda \), we define the function \( \xi(\cdot) \) as follows:

\[
\xi(\lambda) = \left| \frac{E_{ii}^m - \beta_1}{E_{mi}} \right| + \left| \frac{E_{mi} - \beta_2}{E_{ii}} \right|
\]

(6.18)

\[
= \frac{1}{\lambda^2} \cdot E_{ii}^m - \beta_1 + \lambda^2 \cdot E_{mi}^m - \beta_2
\]

The optimum normalisation factor \( \lambda_{opt} \) is obtained through partial differentiating \( \xi(\cdot) \) with respect to \( \lambda \):

\[
\frac{\partial \xi(\lambda)}{\partial (\lambda)} = 0
\]

(6.19)

This implies that:

\[
\lambda_{opt} = \sqrt{\frac{E_{ii}}{E_{mi}^m \cdot E_{mi}^m}}
\]

(6.20)

As an example, two invalid pitch-pulses in the speech signal shown in figure 6.5(a), are refined using the described algorithm. The synthesised speech is shown in figure 6.5 (b). In order to study the effect of the refinement procedure on the pitch estimation, the original and the modified speech signals are filtered using a low-pass filter and then the normalised autocorrelation is computed for pitch lags between 40-160 samples.
The results depicted in figure 6.6 show that the maximum of the normalised autocorrelation (which occurs at a pitch lag of 78 samples) of the modified speech is around 0.1 higher than the original one. In addition, the maximum occurred at the submultiple of the pitch (at a pitch lag of 117 samples) is more attenuated. This leads to a more reliable pitch estimate.

6.3.3 Pitch cycle modification

The aim of this stage is to modify irregular pitch cycles such that the pitch evolves smoothly during a frame while maintaining perceptual speech quality and making more regular speech. This ensures that the pitch can be estimated correctly.

The inputs are the residual speech signal, the regular and irregular pitch values and the corresponding pitch cycles and the modified pitch values obtained in section 6.3.1.2. The output is the modified residual signal for which pitch evolves smoothly.

In our modification, we define local pitch cycle as samples placed between two pitch pulse locations. In other words, a pitch cycle starts with a pitch pulse and ends with the next one. The proposed modification is performed based on stretching or compressing the irregular pitch cycle by inserting a segment from the regular pitch cycle or discarding a segment from
the irregular pitch cycle. A segment with length of 11 samples and centred by a pitch pulse is
maintained without any change for every pitch pulse.

Figure 6.6: The normalised autocorrelation function computed for both original (Stars) and the
modified (Squares) speech.

Suppose $C_k(n)$ with length $P_k$ is the irregular local pitch cycle, which needed to be modified
and the modified pitch value computed in section 6.3.1.2 is $\bar{P}_k$. Also, $C_{k-1}(n)$ with length $P_{k-1}$
is the previous regular local pitch cycle.

If $P_k < P_{k-1}$ a segment, $r_m$, from the regular pitch cycle is inserted into the irregular pitch
cycle at the same position. The modified pitch cycle is calculated by equation 6.21.

$$
\tilde{C}_k(n) = \begin{cases} 
C_k(n) & 0 \leq n < N_p \\
\delta r_m(n-N_p) & N_p \leq n < L+N_p \\
C_k(n-L) & L+N_p \leq n < P_k 
\end{cases}
$$

(6.21)
Chapter 6. Pitch Modification

The segment $r_m$ is described by its start point, $N_p$ and with length $L$ given by equation 6.22.

$$L = \left| P_k - P_t \right|$$  (6.22)

$\delta$ is a factor that changes the energy of the segment $r_m$ based on the irregular pitch cycle energy and is given by equation 6.23.

$$\delta = \sqrt{\frac{E_k}{E_{k-1}}}$$  (6.23)

Where $E_k$ and $E_{k-1}$ are the energy of the pitch cycles $c_k$ and $c_{k-1}$ respectively.

If $P_k > P_{k-1}$, the segment $r_m$ from the irregular pitch cycle is discarded. The modified pitch cycle is calculated by equation 6.24.

$$C_k(n) = \begin{cases} 
C_k(n) & 0 \leq n < N_p \\
C_k(n+L) & N_p \leq n < P_k 
\end{cases}$$  (6.24)

In order to find the optimum segment $r_m$, we define a two-variable function $\Psi(\alpha, \Delta)$ as follows:

$$\Psi(\alpha, \Delta) = \alpha(c_{k_x}, c_k) + \Delta(r_m, \tilde{c}_k)$$  (6.25)

$\alpha$ indicates the normalised correlation between the previous pitch cycle $C_{k-1}$ and the modified pitch cycle $\tilde{C}_k$. $\Delta$ is proportional to the reverse of discontinuity energies at the connection points. Since $|\alpha| < 1$, we normalise the discontinuity energies to maximum energy such that their reverse is fitted at the same range with $\alpha$ as follows:

$$\Delta(r_m, \tilde{C}_k) = -\frac{1}{M} E(N_p, \tilde{C}_k) + 1$$  (6.26)

Where $E(.)$ is the normalised discontinuity energy and $M$ depends on the number of connection points:
Chapter 6. Pitch Modification

\[
M = \begin{cases} 
1 & \text{if } r_m \text{ is discarded} \\
2 & \text{if } r_m \text{ is inserted}
\end{cases} \tag{6.27}
\]

With maximisation of the function \( \Psi(.) \) in terms of the segment \( r_m \), the optimum segment \( r_m \) is extracted and the modified local pitch cycle \( C_k \) is constructed by (6.21) or (6.24). This procedure is performed for all of the irregular pitch cycles.

6.3.3.1 Smoothing techniques

To smooth the modified cycle at the connection points, various techniques can be utilised. Overlap and add method has been widely used for this purposes.

To apply this method to the modified pitch cycle at the connection point \( (n_0) \), three samples are considered on both sides as given in equation 6.28.

\[
w_1(i) = \tilde{C}_k(n_0 - N - i) \quad w_1(N + i + 1) = w_1(i) \\
w_2(i) = \tilde{C}_k(n_0 + i) \quad w_2(N + i + 1) = w_2(i) \tag{6.28}
\]

\[0 \leq i \leq N\]

Where \( N=3 \) is the number of the contributed samples.

The contribution of the amplitude of the samples decreases with distance to the connection point. Thus, the samples around the connection point are substituted by the modified samples given by the following formula:

\[
\tilde{C}_k(n_0 - N + i) = \frac{2N - i}{2N + 1} w_1(i) + \frac{i + 1}{2N + 1} w_2(i) \quad 0 \leq i \leq 2N - 1 \tag{6.29}
\]

Fig. 6.7 shows the original and the modified residual signals with the local pitch variations. In order to show the effect of modification, the pitch values are estimated by using the algorithm used in the standard MELP 2.4 Kb/s. The estimated pitch values are 72 and 73 samples for the second original and the modified frames, respectively. Fig. 6.8 shows the normalised autocorrelation function used for estimating the pitch values.
Figure 6.7: (a) The original speech including irregular pitch variations, (b) The modified speech including smoothly pitch evolution.

Figure 6.8: The normalised autocorrelation function computed for the second frame of the original (stars and solid line) and the modified (square and dash-dot line) speech.
Chapter 6. Pitch Modification

Obviously, due to having high correlation at a pitch lag of 73 samples and low correlation for other pitch lag values, especially at a pitch lag of around 100 samples, the modified speech signal will produce a more reliable pitch estimate. Although the estimated pitch values are close to each other, due to the presence of the irregular pitch variations the other speech parameter estimations can be affected, as described in Chapter 5. The effect of the modification on the voicing-level estimations is studied in section 6.6.

6.3.4 Smoothing pitch-pulse evolution

Linear prediction analysis represents the speech signal as a set of coefficients that estimate the vocal tract shape and a residual signal which approximates the glottal excitation signal. During voiced speech, the articulators in the vocal tract usually move slowly leading to the smooth evolution of the speech power spectrum. Moreover, for these sounds the excitation consists of a series of pitch pulses that usually change shape slowly with time. However, there are factors besides the change in the vocal tract shape that contribute to the frame-to-frame and also inter frame (onset and offset frames) variation of LP parameters. The variations of the LP coefficients contribute to the changes in the pitch pulses shape for the pulses located in adjacent frames and also subframes. Since efficient coding of the pitch pulses relies on the similarity of successive pitch waveforms (e.g. WI coding), the performance of this coding stage is affected. The most common approach for reducing the fluctuation in the LP coefficients is to interpolate them at intervals of 5 to 10 ms between update instants. However, since this is accomplished independently of the evolving residual waveform, non-smoothly evolving pitch pulses may result.

In order to make sure that the pitch pulses evolve smoothly during a frame and also frame-to-frame, we propose a method that increases the correlation of the pitch cycles based on a highly correlated target signal. Thus, the aim of this stage is to make more regular speech. This is performed for the voiced speech frames regardless of whether pitch is of smooth or irregular evolution. The inputs are the modified/unmodified residual signal and the predetermined local pitch locations. The output is the modified residual signal for which the pitch cycles evolve smoothly.

6.3.4.1 Target correlation concept

In the following sections, a pitch cycle is centred by a pitch pulse with length of the interpolated pitch value. The basic idea of the target correlation approach is to modify low correlated pitch cycles of the residual signal. Thus, it is required that the target contains highly correlated pitch cycles. The low correlated cycles are searched by computing the
normalised cross-correlation between the pitch cycles of residual and the relative pitch cycles of the target signal and comparing it against a threshold (Figure 6.9).

Figure 6.9: A block diagram of target correlating based for pitch pulse evolution smoothing.

These cycles are modified using the high correlated cycles (section 6.3.4.3). Thus, in the first step, the requirement is to construct the target signal.

6.3.4.2 Target construction

Ideally, the target signal should be as close as possible to the excitation of the vocal tract. Since the signal is unknown, the original residual signal can be used to design the target. The construction algorithm attempts to remove artefacts introduced by the standard LP method from the residual waveform. This is accomplished by guaranteeing a smooth evolution of pitch pulse shapes during voiced speech segments. The production of an individual target cycle is first described and is then followed with the algorithm that constructs a target frame.

6.3.4.2.1 Target cycle construction

In the first step, the pitch cycles are extracted using the estimated pitch value, $P$. Each pitch cycle contains one centred pitch pulse to minimize boundaries energy. Consider $L$ consecutive pitch cycles in the LP residual signal. After normalisation to unit energy and appropriate alignment we obtain the set of cycles $y_0, y_1, \ldots, y_{L-1}$.

The target cycle, $x$, is considered to have maximum correlation with vectors $y_0, y_1, \ldots, y_{L-1}$ [63].

$$x = \arg \max \sum_i \|y_i^T x\|^2$$

$$\|x\| = 1$$

(6.30)

The operator $\|\|^2$ denotes the 2-norm, i.e. $\|x\|^2 = x^T x$.

Equation 6.30 can be rewritten as:
\[
\sum_i \|y_i^T x\|^2 = (Y^T x)^T (Y^T x) = x^T Y Y^T x
\] (6.31)

Thus \(x\) is obtained by:

\[
\frac{\partial \Phi(x)}{\partial x} = 0, \quad \Phi(x) = x^T Y Y^T x
\] (6.32)

Where the \(P \times L\) matrix \(Y\) \((L < P)\) is given by:

\[
Y = [y_0, y_1, \ldots, y_{L-1}]
\] (6.33)

The derivation vector of \(\Phi(x)\) with respect to the vector \(x\) leads to:

\[
2x^T Y Y^T = 0
\] (6.34)

or equivalently:

\[
YY^T x = 0
\] (6.35)

Since the rank of the \(P \times P\) matrix \(YY^T\) cannot exceed the rank of \(Y\) or \(Y^T\) (which is \(L\)) [64] and \(L < P\), \(YY^T\) is not a full rank matrix and therefore there is a non-zero solution. Equation 6.35 is solved by Singular Value Decomposition (SVD) method [65]. In this method, a matrix \(A\) can be rewritten as the product of a column-orthogonal matrix \(U\), a diagonal matrix \(W\) with positive or zero elements (the singular values), and the transpose of an orthogonal matrix \(V\). Any column of \(V\) whose corresponding \(w_j\) of \(W\) is zero yields a solution. The vector, which maximizes Equation 6.31, is the target cycle \(x\).

Since the dimension of the matrix \(YY^T\) is \(P \times P\), searching the target cycle \(x\) may require high computation, especially for the large pitch values. In order to reduce the cost of computation, we restrict the length of the cycles \(y_0, y_1, \ldots, y_{L-1}\) to the minimum pitch value (2.5 ms equal to 20 samples for sampling frequency of 8 kHz). Since high-energy part of the cycles (which is pitch pulse region) has the main contribution in the cross-correlation value, the cycles \(y_0, y_1, \ldots, y_{L-1}\) are centred at pitch pulse location with length of \(P = 20\).

### 6.3.4.2.2 Target frame construction

The target frame is constructed through the target pitch cycles. Each pitch cycle is constructed using the procedure described in the previous section, considering the past target cycles, the current cycle, and possibly some cycles in the future. The current and the future cycles can be
obtained from the original LP residual. The individual pitch cycles are extracted and normalised to have unity energy by using pitch pulse location information and the interpolated pitch values. They are then zero padded in time domain to have the same length and circularly aligned such that the cross-correlation between each cycle and the previous one is maximised. Each target cycle is constructed using the \( n_1 \) target cycles from the past and the \( n_2 \) current and future cycles. In other words, for any cycle \( y_i \) the relative target cycle is given by:

\[
Y = [x_{i-n_1} \ldots x_{i-1} y_i \ldots y_{i+n_1}]
\]

\[
x_i = \arg \max_{x_i} \frac{\|Y^T x_i\|}{\|x_i\|} \quad (6.36)
\]

The resulting cycles are then rescaled and realigned with the original ones before replacing them in the residual waveform. Figure 6.10 illustrates the original residual and the constructed target signal for \( n_1 = n_2 = 2 \). Compared to the original LP residual, the smooth evolution in the target cycles is clearly noticeable.

Figure 6.10: (a) The original residual signal including non-smoothly pitch cycles evolution, (b) the resulting target signal.
Chapter 6. Pitch Modification

6.3.4.3 Pitch cycle evolution smoothing

After constructing the target signal, the normalised cross-correlation is computed between the cycles of the residual signal and the relative target cycles. Next, the ratio of the minimum to the maximum of the resulting values is computed. If the ratio is higher than a threshold (threshold = 0.85), no modification is performed. Otherwise, the low-correlated cycles are replaced based on linear interpolation between the high-correlated cycles as given in Eq. 6.37.

\[
\begin{align*}
y(i) &= \frac{M + 1 - k}{M + 1} C_{b0}(i) + \frac{k}{M + 1} C_{bl}(i) \quad 1 \leq k \leq M \\
\hat{C}_k(i) &= \mu.y(i) \quad 0 \leq i \leq L
\end{align*}
\]  

In this formula, \(C_{b0}\) and \(C_{bl}\) are the high correlated cycles before and after low correlated cycle \(C_k\), \(M\) indicates the number of the low-correlated cycles placed between \(C_{b0}\) and \(C_{bl}\).
Figure 6.12: (a, b) The original and the modified residual signals, (c) the normalised cross-correlation between the target cycles and the original/modified residual cycles, (d) the normalised autocorrelation of the modified and original speech signals.
\( \mu \) controls the resulting cycle energy to be identical to the original one and is given by Equation 6.38.

\[
\mu = \sqrt{\frac{E_k}{E_y}} \tag{6.38}
\]

Where \( E_k \) and \( E_y \) are the energy of the original cycle and the reshaped cycle \( y \) given in Eq. 6.37.

In order to reduce the discontinuity energy, the overlap-and-add method described in section 6.2.3.1 is applied at the connection points. Fig. 6.12 shows the effect of the smoothing pitch cycles on the cross-correlation between successive residual cycles and also the normalised autocorrelation function. Obviously, due to higher correlation at a pitch lag of 45 samples and lower correlation at other pitch lags, especially at a pitch lag of 68 samples, a reliable pitch can be estimated.

### 6.4 Resulting misalignment

The LP coefficients are usually estimated once in a frame and are interpolated for intermediate sub-frames. The interpolation is performed in the LSF domain.

The resulting time-variant LP filter is thus constant over a sub-frame. After pitch cycle modification and pitch-pulse refinement, the speech signal is constructed by passing the modified residual signal through a time-varying synthesis filter. Although the proposed algorithm causes the pitch to be estimated more reliably and to evolve smoothly during a frame, it results in a loss of synchronization between the LP filter and the modified residual signal. This may cause severe artefacts in the synthesised speech signal.

Suppose the LP filter in the \( k^{th} \) intermediate sub-frame is defined by:

\[
A(z, t) = A_k(z) = 1 - a_{1,k}z^{-1} - \ldots - a_{n,k}z^{-n} \tag{6.39}
\]

Where \( z^{-1} \) is the unit delay operator, \( z^{-1}s(t) = s(t-1) \), and \( a_{1,k}, a_{2,k}, \ldots, a_{n,k} \) are the LP coefficients. The degree \( n \) of the polynomial is assumed to be constant. The cycle index \( k = 0,1,\ldots \) is related to discrete time instant \( t = 0,1,\ldots \) through \( k = \lfloor t/m \rfloor \), where \( m \) is the length of the cycle. The operator \( \lfloor . \rfloor \) denotes rounding towards the nearest integer. The residual signal \( r(t) \) is obtained from the speech signal \( s(t) \) through the analysis filter,

\[
r(t) = A(z, t).s(t) \tag{6.40}
\]
In a simple case, the modified signal, $e(t)$, is the same as the original and the only difference is a non-linear delay, $p(t)$:

$$e(t) = r(t - p(t)) \quad (6.41)$$

The phase difference in samples introduced by the pitch modification is denoted by $\phi(t)$, and it is typically a function of time. The synthesised speech signal $\bar{s}(t)$ is reconstructed from the modified residual signal as:

$$\bar{s}(t) = \frac{1}{A(z,t)} e(t) = \frac{1}{A(z,t)} r(t - p(t)). \quad (6.42)$$

To assess the coding distortion introduced by the mere phase displacement between $e(t)$ and $r(t)$, it is assumed that the modified residual signal is ideal except for the phase difference. By substituting (6.40) into (6.42), an expression for the coding distortion $d(t)$ is computed [66]:

$$d(t) = s(t) - \bar{s}(t + p(t)) = \frac{A(z,t + p(t)) - A(z,t)}{A(z,t + p(t))} s(t) \quad (6.43)$$

Obviously, when the LP filters at a phase difference $p(t)$ are identical for all $t$, then $\bar{s}(t - p(t))$ is equal to $s(t)$. Generally, the mere phase shift in the synthesised signal is not audible if $p(t)$ has been formed reasonably. However, equation (6.43) implies that the time modification using the residual signal distorts the synthesised speech. The magnitude of the distortion depends on the time-variant error filter as given in equation 6.44.

$$F(z,t,p(t)) = \frac{A(z,t + p(t)) - A(z,t)}{A(z,t + p(t))} \quad (6.44)$$

As an example, we examine the case where an artefact occurs in the proposed algorithm described in section 6.3. Figure 6.13 depicts the original and modified residual signals along with the corresponding synthesised signals.
For the third sub-frame in the original residual signal, there is no pitch pulse. However, after modification, one pitch pulse is shifted from the fourth sub-frame to the third sub-frame. The resulting synthesised speech signal, Fig. 6.13(d) comprises a severe, audible artefact as a result of the misalignment between the analysis and synthesis filters. Fig. 6.14 shows the magnitude responses of the synthesis filters in the third and the fourth sub-frame. The gain of the third sub-frame filter is around 8 dB lower at low frequencies and hence the response caused by the shifted pitch pulse is attenuated. In the next section a method, which re-obtains the residual signal and synthesises the modified residual signal correctly, is introduced. In other words, the proposed algorithm plays a pre-analysis-and-post-processing role for the modified residual signal.

### 6.4.1 Pre-analysis and post-processing method

Usually, in most pitch modification methods, the residual signal is processed if the input frame is voiced. Hence, in the proposed algorithm, if the frame is unvoiced no processing is required and speech signal is reconstructed by passing the residual signal through the synthesis filter.

During voiced sections, the main task of the algorithm is that the correct update of the inverse filter coefficients such that the shifted pitch pulses are synthesised without resulting audible artefacts.

In order to remove the coding distortion obtained in equation (6.43), it is necessary to make the magnitude response of the error filter in equation (6.44) zero:
This means that the filter coefficients used at time $t + p(t)$, $A(z, t + p(t))$, in the synthesis stage are required to be identical to the coefficients at time $t$, $A(z, t)$, in the analysis stage. However, since $A(z, t + p(t))$ are calculated by the interpolation between the adjacent LP coefficients, this condition cannot be satisfied.

The experiments show that the perceptual speech quality is affected where a pitch pulse is reconstructed inaccurately. This occurs when the pitch pulse location after modification falls into another sub-frame (see Fig. 6.13a,c) and the LP coefficients of previous or the next sub-frame are inaccurately used in the synthesis filter. This leads to attenuating or amplifying the pitch pulses which can affect the synthesised speech quality and thus there is a need to control the synthesis LP filter gain while updating the filter coefficients. In the next section, we propose an algorithm, which segments the speech signal depending on pitch pulse locations in the analysis stage and updates the LP coefficients properly in the synthesis stage. The new algorithm is divided into two parts, before and after the pitch modification of the residual signal as pre-analysis and post-processing (Figure 6.15).
6.4.1.1 The Pre-analysis stage

The speech signal is usually divided into sub-frames of equal lengths. However, after pitch modification of the residual signal, both the pitch pulse locations and sub-frame lengths may have changed. It is evident in Fig. 6.13(d) that applying the LP coefficients incorrectly in the synthesis stage results in the pitch pulses being either attenuated or amplified. As shown in Fig. 6.16, it is required that the LP coefficients obtained in the analysis stage are aligned in the synthesis stage depending on the pitch pulse locations. In the first stage of the algorithm, the residual signal is re-computed before it is pitch modified using the new LP coefficients. The inputs to the pre-processor are the linear prediction residual of the speech signal and the associated pitch track. The pitch period is estimated once in a frame using conventional autocorrelation based methods, and the resulting estimate is then linearly interpolated for the sub-frames.

The outputs are a re-computed linear prediction residual, which is constructed by the concatenation of the residual pitch cycles, and a set of LP coefficients per each pitch cycle. The initial residual signal and the intermediate pitch values are used to search the pitch pulse locations. The search is based on the concentrated energy of residual samples and on the interpolated pitch period estimate. It is useful to search the pitch pulses for an up-sampled residual signal [16]. Next, we divide the input speech into pitch cycles instead of sub-frames.

If \( p_1, p_2, ..., p_M \) are pitch pulse locations, \( k^{th} \) pitch cycle, \( c_k(n) \), is considered as follows:

\[
c_k(n) = r(n) \quad \left[ \frac{(p_k + p_{k+1})}{2} \right] \leq n \leq \left[ \frac{(p_k + p_{k+1})}{2} \right]
\]  

(6.46)

Thus, the number of pitch cycles varies for every speech frame, depending on the number of pitch pulse locations. Next, a Hamming window with a length dependent on the local pitch values (up to 160 samples) centred on the pitch pulse location, is applied for each pitch cycle.
Chapter 6. Pitch Modification

This is followed by the computation of the LP coefficients. In fact, the computation of the LP coefficients for each cycle plays the role of the interpolation of the LP coefficients performed by traditional methods. The residual pitch cycle is constructed by passing the pitch cycle through the resulting LP filter. By concatenation of the residual pitch cycles, the new residual signal is computed and is applied as an input to the proposed pitch modification algorithm described in section 6.3. Fig. 6.17 shows a block diagram of the pre-analysis stage in combination with the proposed pitch-smoothing algorithm.

Figure 6.16: The LP coefficients alignment in synthesis stage based on pitch pulse locations.

Figure 6.17: A block diagram of the pre-processing stage in combination with the pitch modification of the residual.
6.4.1.2 The Post-processing stage

The second stage of the proposed algorithm is performed after the pitch scaling. The inputs to the post-processing stage are the pitch scaled residual of the speech signal, an associated pitch track and the LP coefficients computed in the analysis stage for each cycle. The output is the pitch scaled speech signal.

![A block diagram of the post-processing stage.](image)

**Figure 6.18 A block diagram of the post-processing stage.**

![The original (a) and the pitch smoothed speech signals with (c) and without (b) the proposed pre-analysis and post-processing.](image)

**Figure 6.19: The original (a) and the pitch smoothed speech signals with (c) and without (b) the proposed pre-analysis and post-processing. The vertical solid and dashed lines in (a) show the speech segmentation respectively with the traditional and new method in analysis stage.**
In order to align the predetermined LP coefficients correctly, in the first step the pitch pulse locations are searched within the pitch scaled residual signal or may be obtained from the pitch scaling processor. Next, the residual signal is divided into the pitch cycles. Each pitch cycle contains a pitch pulse and is computed through equation (6.33) using the pitch pulse locations. Each pitch cycle is then passed through the synthesis LP filter to reconstruct the corresponding speech cycle. The LP filter coefficients are updated for each pitch cycle. The modified speech signal is reconstructed by concatenation of the resulting speech cycles. An overlap-add method described in section 6.2 is used at the connection points to minimize discontinuities. Fig. 6.18 shows a block diagram of the post-processing stage. Fig. 6.19 shows the resulting pitch modified speech signals with and without using the proposed pre-and-post processing. It is observed that the pitch pulse of the fourth pitch cycle is correctly reconstructed using the proposed algorithm. In contrast, this pitch pulse is highly attenuated without using the algorithm.

6.5 Pre-processor evaluation

In order to evaluate the proposed pre-processor, we construct a database including words and short sentences (for both male and female speakers), where pitch evolves irregularly, from the NTT speech database [67]. The new database is used for evaluating the pre-processor in the following areas:

- Pitch estimation
- Voicing level estimation
- Subjective listening tests.

6.5.1 The effect of the pre-processor on pitch estimation

The algorithm applied in the standard MELP is used to estimate the pitch values of the original and the modified speech. In [19] it is shown that more accurate estimated pitch leads to higher Pitch Prediction Gain (PPG), \( \alpha \), given by:

\[
\alpha = \frac{R(T)}{R(0)}
\]  

(6.47)

Where \( T \) is the estimated pitch and \( R(T) \) refers to the autocorrelation of the speech signal with time lag of \( T \) samples.
Chapter 6. Pitch Modification

The PPG can be used as a measure to indicate the accuracy of the estimated pitch. We therefore compute the PPG for the original and modified speech. Due to being fractional pitch values, the PPG is computed from the up-sampled speech signals with an accuracy of 0.25 samples. Fig. 6.20 shows the original speech with the resulting PPGs of the modified and the
original speech signals. It is observed that the PPG of the modified speech increases in comparison with the original speech. Since the frame whose pitch evolves irregularly affects the PPG of the previous and the next frames, we only compare the resulting PPG of these original frames with the relative modified frames. If the difference between the PPG of the original frame and the relative modified one is less than a set threshold, we assume there is no preference between the estimated pitch values of the original and modified frames. Otherwise the estimated pitch value corresponding to the higher PPG is considered as the correct pitch value. Figure 6.21 indicates the results of comparison of the PPGs of the original and the modified speech for two different thresholds 0.1 and 0.2. It is observed that the pre-processor greatly improves the PPG of the frames containing irregular pitch variations and thus the pitch can be estimated more accurately for these frames.

As mentioned in section 6.3.4, the proposed pre-processor also provides smooth pitch cycle evolution to produce more regular speech. This was performed based on the target correlation signal. In the next experiment, the effect of pitch cycle smoothing on the pitch estimation is studied. In this step, we use frames containing smooth pitch variations and compute the PPGs of the original and the modified speech. Figure 6.22 shows the results of comparison of the PPGs of the original and the modified speech for two different thresholds, 0.1 and 0.2. It is observed that the proposed method for pitch cycle smoothing also improves the PPG of speech signal and consequently provides more reliable autocorrelation for pitch estimation.

![Pitch Predictor Gain (PPG) of modified and original speech](image)

Figure 6.21: The resulting PPGs of modified speech in comparison with the original one using two different thresholds, threshold = 0.1 (top) and 0.2 (bottom).
Chapter 6. Pitch Modification

Figure 6.22: The resulting PPGs of modified speech in comparison with the original one using two different thresholds, threshold = 0.1 (top) and 0.2 (bottom).

Figure 6.23: (a) The original LP residual, (b) the modified LP residual, (c), (d) the corresponding original and modified LP residual spectrums, respectively.
In fact, PPG shows the effect of the pre-processor in the time domain. The pre-processor also provides more suitable speech spectrum for pitch estimation. Figure 6.23 shows the original and the modified speech frame along with their corresponding frequency spectrums. Since speech coders usually estimate the pitch from the low pass filtered speech signal with a cut-off frequency around 1KHz [55], as observed in Figure 6.23 (d), the modified residual spectrum is more harmonic in comparison with the original residual spectrum for frequencies between 0 to 1KHz. Thus, the pitch can be more accurately estimated from the modified speech signal.

In order to show the effect of the pre-processor in the frequency domain the synthetic spectral matching method detailed in chapter 3, section 3.3.2.2, is used. We define the ratio of Synthetic Spectral Matching of the Modified to the Original, SSMMO, given in 6.48.

\[
SSMMO(n) = \frac{\tilde{E}(\tilde{\omega}_0, n)}{E(\omega_0, n)}
\]  

(6.48)

\[\tilde{E}(\cdot)\] and \(E(\cdot)\) indicate the synthetic spectral matching distortion of the modified and the original speech signals.

\[
\tilde{E}(\tilde{\omega}_0, n) = \sum_{m=0}^{N_f-1} |\tilde{S}_w(m, n) - \hat{S}_w(m, \tilde{\omega}_0, n)|^2
\]

\[E(\omega_0, n) = \sum_{m=0}^{N_f-1} |S_w(m, n) - \hat{S}_w(m, \omega_0, n)|^2
\]  

(6.49)

Where \(\tilde{S}_w(n), \hat{S}_w(n), S_w(n)\) and \(\hat{S}_w(n)\) are the spectrums of the modified, the synthetic modified, the original and the synthetic of the \(n^{th}\) frame speech, respectively. \(N_f, \omega_0\) and \(\tilde{\omega}_0\) respectively are the FFT length, the estimated pitch frequency of the original and the modified signals. This is measured for original frames including irregular pitch variations and the corresponding modified frames. Figure 6.24 shows the original and the modified speech signals and the corresponding synthetic spectral matching distortions. In Figure 6.24-c, it is observed that for the frames including irregular pitch evolutions, the resulting distortion is significantly higher for the original compared to the modified one and consequently the SSMMO decreases (Figure 6.24-d). When SSMMO<1, this indicates the estimated pitch of the modified speech is more accurate than the original one and vice versa. Thus, the accuracy of the estimated pitch is measured as:

\[\text{Accuracy of Modified estimated pitch > the original} \quad \text{if} \quad SSMMO < 1 - TH\]
\[\text{Same accuracy} \quad \text{if} \quad 1 - TH \leq SSMMO \leq 1 + TH\]
\[\text{Accuracy of Original estimated pitch > the Modified} \quad \text{if} \quad SSMMO > 1 + TH\]  

(6.50)
Figure 6.24: (a) The original speech, (b) The modified speech, (c) The corresponding synthetic spectral matching distortion of the original and the modified, (d) SSMMO function with threshold TH=0.1.

Figure 6.25: Accuracy of the estimated pitch using the SSMMO function with two different thresholds. (Top) TH = 0.1, (Bottom) TH=0.2.
Figure 6.26: (a) The original speech, (b) and (c), the estimated pitch values of the original and modified male speech, respectively.

Figure 6.27: (a) The original speech, (b) and (c), the estimated pitch values of the original and modified female speech, respectively.
Chapter 6. Pitch Modification

The TH is considered as a guard region to prevent small differences between the resulting distortions in our measurements. Figure 6.25 shows the accuracy of the estimated pitch of the modified speech in comparison with the original one using the SSMMO with TH=0.1 and TH=0.2. This shows that the pre-processor modifies frames containing irregular pitch evolution such that the pitch can be more accurately estimated.

In the next experiment, the effect of the pre-processor on the frame-to-frame pitch variations is presented. Since the modification is based on smoothing the pitch contours, it is expected that the pitch will evolve smoothly from frame-to-frame as compared to the original counterpart. Figures 6.26 and 6.27 show the pitch variations of the original and the modified male and female speech signals. The circled segments show where the estimated pitch values of the original speech result in high variations whereas the modified ones lead to smooth evolutions.

6.5.2 Effect of the pre-processor on voicing level estimation

As detailed in the last section, the pre-processor provides more regular speech such that the pitch can be more accurately estimated. This also affects the accuracy of the other parameters estimation, e.g., voicing level. In the following, we show that the pre-processor can also provide more accurate voicing level estimation. We apply both the original and the modified speech separately as input to the standard MELP 2.4 Kb/s.

![Figure 6.28: (Left) The original speech signal, (right) the corresponding modified speech.](image)
Figures 6.28 and 6.29 show the original and the modified speech signals and the corresponding normalised autocorrelation function computed for pitch lags between 20-160 samples. The voicing decisions of five bands are computed by comparing the normalised autocorrelation value for the estimated pitch with a voicing threshold. Figure 6.30 shows the modified and the original speech spectrums in the five frequency bands along with the estimated voicing level for each band.

Figure 6.29: The corresponding normalised autocorrelation function of the original (a), and the modified (b) first speech frame shown in Fig. 6.28, for five frequency bands.
It is observed that in spite of being strongly voiced speech (Fig. 6.28-left), only the first band of the original speech is estimated as voiced, whereas the first four bands of the modified are estimated as voiced. In order to provide a quantitative measure of the accuracy of the estimated voicing level, we compute the spectral distortion given in Eq. 6.49, where $\hat{S}_n(n), \hat{S}_m(n)$ are substituted by the original and the modified synthesised spectrums. This distortion is only computed for frames containing irregularly pitch variations and also one frame before and after these frames. Next, the SSMMO is calculated for the last four bands (for male speech only) and also for the full band (for both male and female speech). The accuracy of the voicing decision for each band is given by Eq. 6.50.

Figure 6.31 shows the accuracy of the voicing decision of the modified speech in comparison with the original one by using the spectral distortion with TH=0.1 and TH=0.2.
It has to be mentioned that the spectral distortion depends on two parameters, voicing decision for each band and the estimated pitch. As observed in Fig. 6.31, due to more accurate voicing and pitch estimations, the resulting spectral distortions of the modified speech is lower than the original for the bands 2-5. For the first band, the voicing decision is the same. However, due to improved pitch value the spectral distortion is even lower.

6.5.3 Subjective listening tests

For testing purposes, the pre-processor detailed in section 6.2 was applied on short sentences and words (for both male and female) for which the pre-processor was activated. An AB-test with 15 (4 trained and 11 untrained) listeners was carried out on the original and the processed speech files (including 8 speech sentences, 4 male and 4 female, and 10 single words for both male and female speech) to investigate the perceptual speech quality of the
modified before applying to a speech coder. In the next experiment, the effect of the pre-processor on a speech coder is evaluated. This was performed by applying the original and the modified speech as the inputs to standard MELP 2.4 kbps. An A vs. B comparison test was carried out on the synthesised speech files. The results are shown in Table 6.1. The results obtained indicate that there is no statistical difference in perceptual quality between the original and the modified speech. However, the pre-processor in combination with the MELP provides significantly better perceptual speech quality than the standard MELP.

Table 6.1: Modified speech vs. original one and MELP + pre-processor vs. MELP.

<table>
<thead>
<tr>
<th>Speech Type</th>
<th>Better</th>
<th>Slightly better</th>
<th>Same</th>
<th>Slightly worse</th>
<th>Worse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modified speech vs. Original speech</td>
<td>0.2</td>
<td>11.2</td>
<td>76.3</td>
<td>12.3</td>
<td>0</td>
</tr>
<tr>
<td>Pre-processor + MELP 2.4kbps vs. MELP 2.4kbps</td>
<td>4.2</td>
<td>51.4</td>
<td>30.1</td>
<td>12.2</td>
<td>2.1</td>
</tr>
</tbody>
</table>

6.6 Pre-processor in noisy speech

Robustness to background noise is an important factor for any practical speech-coding algorithm. The speech coders designed for mobile and military communication applications are frequently encountered with acoustic noise. The background noise may be suppressed before the encoding process using a noise suppressor [57]. However, this involves additional complexity and delay, which may not be desirable for mobile communication applications. Therefore, the speech coding algorithms are expected to produce intelligible synthesised speech even at the presence of the background noise.

The difficulties specific to the new pre-processor are the robustness of the V/UV classification, initial pitch estimation, searching the pitch pulse locations and the employed pitch-cycle modification algorithm. In the following section, the robustness of these algorithms is tested under the background noise and the performance of the pre-processor is evaluated.

6.6.1 Performance under background noise

The employed V/UV classification and the pitch estimation algorithms are those used in standard the MELP 2.4 Kbps. Although the robustness of the standard MELP was reported in
and since our propose pre-processor is independent of the speech coder, increasing the robustness of the V/UV and pitch estimation algorithms may be achieved by investing more robust and complex techniques.

Figure 6.32: The effect of SNR on local energy variations of the residual signal and GCIDS. In each column: (Top): the residual signal, (Middle) and (Bottom) the corresponding local energy variations and the GCIDS, respectively.

The pitch contour construction and the pitch-cycle modification algorithms are tested using 32 seconds of male and female speech containing irregular pitch variations contaminated with vehicle noise. The SNRs of the noisy speech is between 10 and 25 dB.

The pitch contour construction has the main role in the pre-processor. This construction is performed based on searching pitch pulse locations. Thus, in order to test the robustness of the pitch contour construction the employed algorithms for searching pitch pulse locations are examined. As described in section 6.2.1 and 6.2.2, the pitch pulse locations are searched based on the concentrated energy and the refined based on the Glottal Closure Instant
Determination Signal (GCIDS). Although, as reported in [60], GCIDS is robust under white noise, the experiments show that the robustness of the GCIDS under vehicle noise may not be guaranteed for SNRs lower than 15 dB. Figure 6.32 shows the energy variations of the residual speech and the resulting GCIDS for noisy female speech with SNR 10-30 dB in comparison with the clean speech. As it is observed, for SNRs lower than 15 dB, in some cases where the noise power is higher than the original speech, searching pitch pulse locations fails.

On design of the proposed pitch-cycle modification algorithm two main speech characteristics are considered. The first one is the correlation between the successive pitch cycles (long-term correlation) and the second is minimisation of the speech discontinuities resulting from inserting/discardng a segment during the pitch cycle modification. These characters are independent of whether speech is noisy or clean and is provided by pitch-cycle modification. Due to robustness of the employed tools for SNR higher than 15 dB, the robustness of the pre-processor is expected for these SNRs. In the following section, we demonstrate the robustness of the pre-processor on the pitch and voicing level estimations as well as speech quality.

6.6.1.1 Robustness on the pitch and voicing level estimations

The block diagram depicted in Figure 6.33 was used for testing purposes.

![Diagram](image)

**Figure 6.33:** An employed block diagram for testing the pre-processor under background noise.

The database used in the previous section was contaminated with vehicle noise for SNRs between 10 to 25 dB. The noisy speech was then applied to the pre-processor.
Figure 6.34: The accuracy of the estimated pitch values resulting from clean/noisy original/modified speech signals using pitch prediction gain (PPG) and spectral distortion (SD) with thresholds 0.1 and 0.2.
Next, the employed algorithm in the standard MELP is used for both $S(n)$ and the $	ilde{S}(n)$ signals to estimate pitch values. In order to investigate the accuracy of the estimated pitch values, the pitch prediction gain and the spectral distortion detailed in the section 6.4.1 are computed. Results are shown in Figure 6.34.

![Figure 6.34: The effect of the background noise on the pitch variations. (a) Original male speech, (b) and (c) the pitch evolution of the original and the modified clean speech, respectively, (d, e, f and g) the estimated pitch values of the modified speech under background noise with SNR 25, 20, 15 and 10 dB respectively.](image)

It is observed that the accuracy of the estimated pitch values is affected by decreasing SNR. As shown in the previous section, the reason is that searching accurately the pitch pulse locations may fail. This leads to inaccurate pitch contour construction.

Next, the performance of the pre-processor on the frame-to-frame pitch evolution is tested. Figures 6.35 and 6.36 show the pitch variations for noisy-modified male and female speech, respectively in comparison with clean original and modified ones. The circled regions show
where the variations of the estimated pitch values are affected by decreasing the SNR. The high variations are observed for SNRs lower than 15 dB.

Figure 6.36: The effect of the background noise on the pitch variations. (a) Original female speech, (b) and (c) the estimated pitch values of the original and the modified clean speech, respectively. (d, e, f and g) the pitch evolution of the modified speech under background noise with SNR 25, 20, 15 and 10 dB respectively.

In order to investigate the effect of the pre-processor on the voicing level estimation under background noise, the noisy original and modified speech signals were applied to the standard MELP 2.4 kb/s. Next, the accuracy of the estimated voicing strengths was evaluated using the spectral distortion for each band and also for full band. Results are shown in Figure 6.37. It is observed that the resulting spectral distortion of the modified speech is lower than the original one for all frequency bands which means more accurate voicing and pitch estimations. However, the spectral distortion of the modified speech increases as SNR decreases. This means that the accuracy of the voicing strength and also pitch estimation is affected more.
Figure 6.37: Accuracy of the voicing level estimation using spectral distortion of the modified speech in comparison with the original one under background noise with threshold 0.2.
6.6.1.2 Subjective listening test

The database used for clean speech was used under background noise and applied to the pre-processor. An AB-test with 15 (4 trained and 11 untrained) listeners was carried out on the noise-suppressed original and the processed speech files to investigate the perceptual speech quality of the modified speech before applying to a speech coder. Results are shown in Table 6.2. The results obtained indicate that there is no difference in perceptual quality between the original and the modified speech under background noise for SNRs higher than 15 dB.

**Table 6.2: Modified speech vs. original speech**

<table>
<thead>
<tr>
<th>Speech Type</th>
<th>Better</th>
<th>Slightly better</th>
<th>Same</th>
<th>Slightly worse</th>
<th>Worse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean speech</td>
<td>0.2</td>
<td>11.2</td>
<td>76.3</td>
<td>12.3</td>
<td>0</td>
</tr>
<tr>
<td>25 dB SNR</td>
<td>0.2</td>
<td>10.8</td>
<td>75.9</td>
<td>12.9</td>
<td>0.2</td>
</tr>
<tr>
<td>20 dB SNR</td>
<td>0.1</td>
<td>10.2</td>
<td>74.6</td>
<td>14.7</td>
<td>0.4</td>
</tr>
<tr>
<td>15 dB SNR</td>
<td>0.1</td>
<td>9.3</td>
<td>71.6</td>
<td>17.8</td>
<td>1.2</td>
</tr>
<tr>
<td>10 dB SNR</td>
<td>0</td>
<td>7.8</td>
<td>66.4</td>
<td>23.4</td>
<td>2.4</td>
</tr>
</tbody>
</table>

**Table 6.3: Modified speech + MELP vs. MELP alone**

<table>
<thead>
<tr>
<th>Speech Type</th>
<th>Better</th>
<th>Slightly better</th>
<th>Same</th>
<th>Slightly worse</th>
<th>Worse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean speech</td>
<td>4.2</td>
<td>51.4</td>
<td>30.1</td>
<td>12.2</td>
<td>2.1</td>
</tr>
<tr>
<td>25 dB SNR</td>
<td>3.4</td>
<td>50.2</td>
<td>31.2</td>
<td>12.9</td>
<td>2.3</td>
</tr>
<tr>
<td>20 dB SNR</td>
<td>2.8</td>
<td>47.3</td>
<td>32.6</td>
<td>14.2</td>
<td>3.1</td>
</tr>
<tr>
<td>15 dB SNR</td>
<td>2.2</td>
<td>43.1</td>
<td>33.6</td>
<td>17.3</td>
<td>3.8</td>
</tr>
<tr>
<td>10 dB SNR</td>
<td>0.8</td>
<td>30.3</td>
<td>36.2</td>
<td>25.9</td>
<td>6.8</td>
</tr>
</tbody>
</table>

In the next experiment, the noisy speech signals were applied as the inputs to standard MELP 2.4 kbps and an A vs. B comparison test was carried out on the synthesised speech files. The results are shown in Table 6.3. This shows that the pre-processor in combination with the MELP provides significantly better perceptual speech quality than the standard MELP at SNRs higher than 15 dB.
6.7 Conclusions

In this chapter the current pre-processors and techniques for pitch-estimation improvement were addressed. We proposed a new pre-processor, which modifies the residual signal to make more regular speech. The modification is performed on the cycles containing irregular pitch variation where these cycles are marked using a smooth pitch contour. This procedure is based on inserting or discarding a segment. The optimum segment is searched based on increasing the correlation between the modified and the previous cycle and minimisation of the resulting discontinuities at connection points. The pitch prediction gain and spectral distortion were used as quantitative measures to demonstrate that the benefits of applying the pre-processor are to reduce the inaccuracy of the pitch and voicing level estimation in speech coders. The subjective listening tests show that the pre-processor maintains the original speech quality and can therefore be used in combination with any speech coder. In addition, it was shown that the proposed pre-processor in combination with the standard MELP 2.4Kbps provides significantly better quality than MELP alone. The performance of the pre-processor under background noise was tested. Vehicle noise with SNRs of 10 to 25 dB was considered as the background noise. Results of these experiments show the robustness of the pre-processor at SNRs higher than 15 dB.
Chapter 7

Postfiltering

7.1 Introduction

For speech perception, speech formants regions are perceptually much more important than spectral valleys regions. Since it is impossible to push the quantisation noise under an inaudible threshold in both formant and valley regions at low encoding rates [68] a good strategy is to allow more quantisation noise, yet lower than the inaudible threshold, at the formant regions whereas sacrificing the amount of quantisation noise introduced in the valleys regions. This technique is performed in analysis-by-synthesis coders by incorporating subjective weighting factors into the error minimisation procedure [68]. This causes the noise components in some of the valley regions to exceed the threshold. However, these noise components can be made inaudible by attenuating them with a postfilter at the decoder. The idea of filtering speech with a formant-equalised frequency response, or even the idea of enhancing noisy speech with a filter having a speech-like frequency response, was proposed in 1965 [69] by Schroeder. In [68], a postfilter consisting of a long-term postfilter section in cascade with a short-term postfilter section was proposed. The long-term section emphasises pitch harmonics and attenuates the spectral valleys between pitch harmonics. On the other hand, the short-term postfilter maintains formant information and attenuates the spectral valleys between the formants. The following sections present the short-term postfilter and thus we use 'postfilter' for simplicity.

The postfilter is designed based on the quantised speech spectrum. Due to variations of the local characteristics of the speech spectrum, use of an adaptive postfilter is preferred to the fixed one. The adaptive postfiltering attempts to suppress the noise components by attenuating the valleys and not altering the formant information. In performing such attenuation the speech component in the valley region will also be attenuated. Generally, the intensity of the spectral valley can be attenuated by as much as 10 dB without any audible effect [70]. Therefore, by attenuating the components in the spectral valleys, the postfilter introduces a minimum perceptual distortion in the speech signal. However, it also achieves a substantial noise reduction.
This chapter is organised as follows: In the next section the conventional postfiltering is overviewed. In section 7.3, we demonstrate the shortcomings of the current conventional postfilter. These problems are resolved using the proposed modified conventional postfilter. The new scheme uses the synthesis LP filter factorisation which is based on formant locations and searching the optimum shaping constants for each formant. This procedure is performed for each frame and thus the optimum constants are updated frame-by-frame. Section 7.4 details the new postfilter design procedure. Finally, in section 7.5, the new and the conventional postfilters are evaluated in combination with the full-rate GSM-AMR proposed standard based on ACELP coder operating at bit rates between 8.25 to 13.65 kb/s [4].

7.2 Conventional postfilter overview

From above, the frequency response of an ideal short-term postfilter should follow the peaks and valleys of the spectral envelope of speech. Since the LPC spectrum contains the spectral envelope information, it is natural to derive the short-term postfilter from the LPC predictor. A conventionally used short-term postfilter is given by [5][68]:

\[
P(z)=\frac{1-\sum_{k=1}^{p} \alpha_k a_k z^{-k}}{1-\sum_{k=1}^{p} \beta_k a_k z^{-k}}(1-\mu z^{-1}) = \frac{1-A(z/\alpha)}{1-A(z/\beta)}(1-\mu z^{-1})
\]  

(7.1)

Where the modified synthesis filter \( \frac{1}{1-A(z/\beta)} \) attempts to model the speech spectral envelope through emphasising the formant regions and achieving the quantisation noise reduction in the valley regions. The second filter structure \( 1-A(z/\alpha) \) tries to flatten the spectrum and remove the low pass spectral tilt introduced by the synthesis filter for voiced speech. This tilt has the side effect of making the postfiltered speech sound muffled. The difference between \( \alpha \) and \( \beta \) determines the filtering effect.

Subjectively, a large difference will give a disturbing deep voice effect with muffling of the higher frequencies. If the difference is not large enough then there is no subjective improvement in the speech quality. For instance, Fig. 7.1 shows three different choice of \( (\alpha = 0.6, \beta = 0.7), (\alpha = 0.6, \beta = 0.8) \) and \( (\alpha = 0.6, \beta = 0.9) \) where the enhanced spectrum is normalised such that the first formant magnitude is maintained identical to original one. The optimum values for \( \alpha \) and \( \beta \) are highly dependent on the bit rate and the type of speech coder and are typically determined using subjective listening tests [68].
Chapter 7. Postfiltering

7.3 Motivation for optimum constants

For the adaptive postfilter proposed in [68], the shape of the postfilter is adapted using the LP spectrum, yet the constants $\alpha$ and $\beta$ are fixed for all the synthesised speech frames.

![Figure 7.1: The original and the modified LP spectrums using the conventional postfiltering with three different choices of constants.](image)

As shown in Fig. 7.1, although the higher difference between these constants ($\alpha = 0.6, \beta = 0.9$) provides more attenuation in valleys regions, it also causes more attenuation in the other formants. On the other hand, using the lower difference between these constants ($\alpha = 0.6, \beta = 0.7$) maintains the formant information, however, slightly attenuates valleys regions. In fact, using the same constants for all of the formants causes the formants to be weighted in the same way. Since low-frequency formants play a more important role in perceptual speech quality than the other formants, it is required that these formants and the relative valleys are shaped differently from the other formants and valleys. In the following section we design a new postfilter, which factorises the LP synthesis filter to multi-polynomials based on the formant information. Each formant is represented by a polynomial and the constants $\alpha$ and $\beta$ are then searched for each formant (each polynomial) such that the formant information are maintained while attenuating the valleys regions.
Figure 7.2: (a) The original and the corresponding quantised LP spectrums, (b) The corresponding first two LSF’s, (c) The corresponding pole locations.
As shown in Figure 7.2(a), a possible side effect of the LPC parameters quantisation is broadening of the formants where the LP/LSF coefficients are quantised using a split vector quantiser [71]. This occurs when LSFs representing the relative formant become further apart than the original one (Fig 7.2.b) or equivalently, the radii of the poles representing the formant decrease (Fig. 7.2.c). In fact, increasing the bandwidth undesirably amplifies quantisation noise around the main formant and therefore affects speech quality.

7.4 New postfilter description

From above, the conventional postfilters encounter two problems. Firstly, weighting the formants by the same constants without considering the role of the main formant on perceptual speech quality, secondly amplifying quantisation noise around the formant frequencies. Thus, the proposed postfilter performs the following tasks:

- Producing a narrower bandwidth for the main formant,
- Shaping formants and attenuating valley regions leading to better perceptual speech quality.

The new postfilter is actually a time domain filter which uses the modified LP synthesis, analysis, and high/low pass filters that are driven from an LPC spectrum and are configured by the constants: $\alpha$ (for modified LP synthesis filter), $\beta$ (for modified LP analysis filter) and $\mu$ (for high pass filter). The new postfilter uses the pole information in the LPC spectrum and finds the relation between poles and formants and then computes the optimum constants for shaping each formant. In fact, the difference between the new postfilter and the conventional one is that the constants are adapted both intra-frame and frame-to-frame.

The inputs to the postfilter are the synthesised speech, the decoded LPC and voicing information and the output is the enhanced speech signal. The LPC and voicing information can be obtained from the decoder directly or in case of unavailability can be computed from the synthesised speech. In case of unvoiced speech, the conventional postfiltering can be applied. Thus, the new postfiltering is performed for voiced speech only.

In the following sections, the new postfilter design approach is detailed.

7.4.1 LP filter factorisation

In this section, we propose a novel postfiltering method, which is based on LP synthesis filter factorisation. In order to control the shape of each formant, we calculate a polynomial representing the formant. This is performed using the relation between the poles of the synthesis filter and the formants. Since the formants frequencies appear as local
maximums in the speech spectrum and considering flat residual spectrum, we assume that pole angles in a LPC spectrum contain information about the formant locations.

![Figure 7.3: (a) The LP spectrum and the corresponding poles, (b) The pole locations.](image-url)
Chapter 7. Postfiltering

Given that a LPC spectrum is defined as:

\[ H(z) = \frac{1}{1 - A(z)} \]  \hspace{1cm} (7.2)

Where

\[ A(z) = \sum_{i=0}^{M} a_i z^{-i} \]  \hspace{1cm} (7.3)

and M is the order of the LPC predictor, the poles can be found by solving for the roots of \( 1 - A(z) \). Since most speech coders use 10th order LPC filter, we consider M=10. Eigenvalue method is used in solving for the roots of \( 1 - A(z) \)[65]. In this method, the coefficients \( \{a_i\} \quad i=0,1,\cdots,M \quad a_0 = 1 \) are placed into a companion matrix whose eigenvalues are the roots of \( 1 - A(z) \).

Since the coefficients of \( 1 - A(z) \) are real, its roots are in complex conjugate pairs, although two real poles might exist. If two real poles exist, they always have an angle of 0 and \( \pi \). Noting this symmetrical property, the poles can be divided into a group of positive angles and a group of negative angles. In the positive angle group, the angles of the poles are arranged in ascending order whereas for the other group the arrangement is based on descending order. With this arrangement \( r_1 \) and \( r_6 \) have the same radius and occur in conjugate angles when there are no real roots.

To analyse the relation between poles and the formants, a typical LPC spectrum is plotted with the pole angles located on the normalised frequency axis as shown in Fig. 7.3. In this figure, the locations of poles 1 through 5 are noted by \( P_1 \) through \( P_5 \). Poles \( P_3 \), \( P_4 \) and \( P_5 \) indicate the exact locations of the formant peaks. However, the poles are not always located at the peaks as shown in the example. In general, a wide formant bandwidth has two or three poles that are close together. This can be observed in Fig. 7.3 where the bandwidth of the first formant is wider than the second formant. The first formant has poles \( P_1 \) and \( P_2 \) that are close together whereas the other formants only have a single pole. With knowledge of the poles, estimation of the formants and the valleys regions can begin and the poles representing the formants are computed.

In order to estimate formant and valley regions, the magnitude response of the synthesis filter for any given angle, \( \omega \) is computed using pole locations as:
Postfiltering

\[ H(k) = \prod_{i=1}^{M} \left( 1 + r_i^2 - 2r_i \cos \left( \theta_i - \frac{k \pi}{N} \right) \right)^{\frac{1}{2}} \quad k = 0, 1, \ldots, N - 1 \]  

(7.4)

Where \( r_i \) and \( \theta_i \) are the radius and the angle of pole \( P_i \), respectively and \( M \) is the order of the synthesis filter and \( N \) indicates frequency resolution. Next, the formants and the valleys are searched from the resulting magnitude response using the property that the formants and the valleys appear as local maximum and minimum, respectively. Each pole is then allocated to the nearest formant based on L1 norm as follows:

\[ P_i \in \omega_k \quad \text{if} \quad |\theta_i - \omega_k| < |\theta_i - \omega_m| \quad L \leq m \leq F, m \neq k \]

(7.5)

Where \( \omega_k \) is \( k \)th formant, \( L \) and \( F \) indicate the minimum and maximum number of formants. In our experiment \( L = 2 \) and \( F = 5 \) were considered. The exceptions are for real poles where the poles with angles 0 and \( \pi \) are allocated to the first and last formant, respectively.

![Figure 7.4: The LP synthesis filter factorisation based on the formants and pole locations.](image-url)
Each formant can be represented by a polynomial using the allocated poles. The order of the polynomial can be between 2 to 6 depending the number of the allocated poles. As an example, figure 7.4 shows the factorisation of the synthesis filter to the polynomials representing the formants. Thus, the synthesis filter is decomposed as:

\[
H(z) = \frac{1}{1-A(z)} = \frac{1}{1-\sum_{i=1}^{M} a_i z^{-i}} = \prod_{k=1}^{p} \frac{1}{Q_k(z)}
\]  

(7.6)

\(Q_k(z)\) is the polynomial, which represents the \(k^{th}\) formant given by:

\[
Q_k(z) = \prod_{i=m}^{n} (1 - P_{ik} Z^{-1}) \quad 1 \leq m, n \leq M
\]  

(7.7)

Where \(P_{ik}\) indicates the \(i^{th}\) pole allocated to the \(k^{th}\) formant.

In the following section, the new postfilter is designed using these polynomials.

### 7.4.2 Narrower bandwidth construction

In design procedure of the proposed postfilter, the higher energy formant is considered as the main formant. Although the first formant has usually higher energy for voiced speech, sometimes the second or even the third formant has more energy. This happens for weak voiced and female speech signals. The energy of the formant is computed using the LPC spectrum and the formant bandwidth as follows:

\[
E_k = \sum_{i=1}^{d_k} \left| \frac{H \left( \frac{N}{\pi} \omega_k + i \right)^2}{d_k - c_k} \right|
\]  

(7.8)

Where \(c_k\) and \(d_k\) are given by:

\[
c_k = \left\lfloor \frac{N}{\pi} \delta \omega_k - \frac{1}{2} \right\rfloor
\]

\[
d_k = \left\lfloor \frac{N}{\pi} \delta \omega_k + \frac{1}{2} \right\rfloor
\]  

(7.9)
The bandwidth of the $k^{th}$ formant, $2\delta b_k$, represents 3dB bandwidth of the relative polynomial, $1/Q_k(z)$. In case of presenting $k^{th}$ formant by two poles, or equivalently $Q_k(z)$ with a 2-order polynomial, and for $r_k$ close to unity $2\delta b_k$ is given by [72]:

$$2\delta b_k = 2(1 - r_k) \quad (7.10)$$

Figure 7.5: The effect of changing the radii of the poles representing the main formant, (a) the resulting main formant and (b) the LP spectrum, (c) the new and original pole locations.
For the voiced segments, since the first formant (or the first LSF parameters) plays the main role in speech quality [71], the main formant, \( \omega_m \), is computed based on the first formant using the formant energies as follows:

\[
\omega_m = \begin{cases} 
\omega_f & 1.2E_i \geq Max(E_k), k = 2, \ldots, F \\
\omega_j & \text{Otherwise}
\end{cases}
\]  

(7.11)

\( \omega_j \) indicates the formant with maximum energy.

Since the main formant is represented by a band pass filter, in DSP point, its bandwidth can be narrowed using the following techniques:

- Moving the relative poles of the main formant towards the unit circle,
- Changing the phase of the relative poles of the main formant,
- Changing the constants of the synthesis, inverse and high pass filters using polynomials representing the formants.

### 7.4.2.1 The poles movement

In the first technique the poles representing the main formant are moved towards the unit circle through an additive step size \( \delta r_i \), given by Equation (7.12), to radius of the poles.

\[
\delta r_i = \frac{1 - r_i}{S}
\]

(7.12)

Where \( r_i \) is the radius of the pole \( P_i \) and \( S \) indicates the number of steps (\( S=5 \)).

Figure 7.5 shows the effect of moving the pole in the LPC where the radius of the pole is changed from .86 to .95. In this technique, three conditions are considered: I) the main formant has to be maintained at the identical angle, II) stability of the synthesis filter, III) preventing undesirable formants.

If two poles represent the main formant, the order of the corresponding polynomial is two and so the main formant is located at the positive angle of the pole. For this case, the first condition is satisfied. Whereas for the main formant representing by more than two poles, the principal pole (the pole closer to unit circle) is moved towards the unit circle with a step size, \( \delta r_i \), and the other poles are then moved as well so that the new main formant, \( \tilde{\omega}_m \), is close enough to the main formant:

\[
|\omega_m - \tilde{\omega}_m| \leq \frac{2\pi}{N}
\]

(7.13)
Stability of the synthesis filter is naturally provided using the property that the roots of LPC filter are inside of unit circle. Although theoretically moving the poles towards the unit circle guarantees stability of the synthesis filter, our experiments show that poles too close to the unit circle, especially in case of high-energy residual, result in a saturated output. Thus, the maximum radius of the moving pole was restricted to 0.985, which was obtained experimentally.

Figure 7.6: The effect of inaccurate moving the poles representing the main formant on the formant (a) and the LP spectrum (b), (c) the new and original pole locations.
In case of more than two conjugate poles, moving the poles with the smaller radius may create an undesirable formant or even increases the main formant bandwidth. This is shown in Fig. 7.6 where moving the pole $P_2$ from 0.7 to 0.9 has created an undesirable formant. This situation can be prevented by checking a local maximum existence at the relative pole angle and the resulting bandwidth. In order to reduce the complexity of the design procedure, the pole with bigger radius, $r_1$, is moved towards unit circle and then the other pole radius, $r_2$, is multiplied by factor of $\lambda$:

$$\lambda = \frac{r_1}{r_2}$$ (7.14)

In fact, this causes the ratio of the new pole radii to be identical to the original one. Note that the principal pole plays the main role in the LPC spectrum and thus moving this pole is performed with higher resolution. In the design procedure, the additive step size given by Eq. (7.12) is modified as $\gamma \delta r_i$, where $r_i$ is the radius of the principle pole and $\gamma = 0.5$.

### 7.4.2.2 Changing the angles

In the second technique, the angles of the poles are modified to make them closer to each other and the radii of the poles are maintained identical with original ones. Obviously, this technique is applied when more than two poles represent the main formant. Figure 7.7 shows the effect of this technique on the first formant where poles $P_3$ and $P_2$ with angles of 0.31 and 0.52 rad presenting the first formant are modified to have same radii with angles of 0.32 and 0.45 rad. In this case, we also consider three conditions: I) the main formant has to be maintained at the identical angle, II) stability of the synthesis filter.

Assume that $\theta_i$ is the angle of one of the poles representing the main formant which is updated using Equation 7.15.

$$\tilde{\theta}_i = \theta_i + \delta \theta$$ (7.15)

$\delta \theta$ indicates step-size angle variation and given by:

$$\delta \theta = \frac{\omega_m - \theta_i}{L}$$ (7.16)

Where $\omega_m$ is the main formant frequency and $L=10$.

In our experiment, firstly the angle of the principle pole and then the angle of the other pole are updated so that the Eq. 7.13 is satisfied. It is mentioned that since the principle pole plays
the main role in the main formant shape, the angle of this pole is normally closer to $\omega_m$ and therefore it is updated by smaller step size.

Figure 7.7: The effect of changing the angles of the poles representing the main formant, (a) the resulting main formant and (b) the LP spectrum, (c) the new pole locations.
In addition, the closer the poles are to each other results a higher resonance for the corresponding formant. Since this can result in a saturated output, the modification of the angles continues until the difference between the updated angles is more than a threshold. This threshold was experimentally found to be \( \approx 0.02 \) rad.

In the proposed postfilter design procedure, two described techniques are used for enhancing the main formant. The first is for the case of a pair of conjugate poles presenting the main formant whereby only the radius is altered. Both modification techniques are used for the case, where more than two conjugate poles representing the main formant.

![Figure 7.8: The effect of changing the constants on, (a) the main formant and (b) the LP spectrum for two different choices, #1: \( \alpha_m = 0.7, \beta_m = 0.8 \), #2: \( \alpha_m = 0.5, \beta_m = 0.9 \).]

7.4.2.3 Changing the shaping constants

In the third technique the main formant bandwidth decreases using the constants of the synthesis, inverse and high pass filters. For instance, Figure 7.8 shows the effect of two different pairs \( \alpha_m = 0.7, \beta_m = 0.8 \) and \( \alpha_m = 0.5, \beta_m = 0.9 \) on the main formant and also LPC spectrum where these constants are equal to 0.8 and 0.7 for the other formants in both cases. In fact, the higher difference between \( \alpha_m \) and \( \beta_m \) more decreases the bandwidth. However, as shown in Fig. 7.8.b, this higher difference attenuates other formants and so affects speech
quality. Thus, maintaining the other formants information is a factor, which is used to control $\alpha_m, \beta_m$. This is performed using different constants for each formant. Unfortunately, due to the existing the interaction between the spectrums of the polynomials representing the formants, it is impossible to find the optimum constants as a close form and thus these are searched using a weighting function described in the following section.

### 7.4.3 Effect of shaping formants and attenuating valleys on speech quality

The second task of the proposed postfilter is to adjust attenuation and amplification levels for the spectral speech valleys and the formants magnitudes. Since there is no objective measure to find the optimum levels, a subjective listening test is carried out on the postfiltered speech signals. The postfiltering is performed on the original speech using the desirable postfilter. Figure 7.9 shows a block diagram of the employed system.

![Figure 7.9: A block diagram of the system used for effect of shaping formants and attenuating valleys on speech quality.](image)

Current frame (with length of 160 samples) with the buffered speech samples (including 20 samples from previous frame and 20 samples look ahead) are first windowed using a Hamming window of length 200 samples. Next, the LP analysis is performed on the windowed speech signal and the resulting LP parameters are used and to compute the residual speech signal and also to determine formants and valleys information. This information with the desirable attenuation and amplification levels are used to construct a target postfilter spectrum. This spectrum has a constant value in the spectral valleys and the formant bandwidths whereas for other frequencies its spectrum magnitudes are obtained through the linear interpolation between the constant values of the relative valley and formant. A moving average window with length of 5 samples is used to reduce the effect of the resulting discontinuities at edge points. Next, the original speech spectrum is multiplied by the target postfilter spectrum in frequency domain to produce the enhanced speech spectrum. The
Chapter 7. Postfiltering

resulting spectrum is used to compute a new set of LP parameters [73]. Figure 7.10 shows a
LP spectrum, the corresponding target postfilter magnitude and the resulting LP spectrum.
The original residual speech signal is then passed through the synthesis LP filter whose
coefficients computed from the enhanced speech spectrum. An overlap-and-add is used to
reduce the resulting discontinuities at frame boundaries.

Figure 7.10: (a) The synthesised LP spectrum, (b) the postfilter spectrum with attenuation 40%
in spectral valleys and 20% amplification of the first formant magnitude, (c) The synthesised and
the enhanced LP spectrums.

Since the gain of the enhanced LP filter is not necessarily identical to original LP filter, the
gain of the enhanced speech signal, $s_{pf}(n)$, can differ from the original speech, $s(n)$. An
adaptive gain control proposed in [74] is used to compensate for gain differences between
$s_{pf}(n)$ and $s(n)$. The gain scaling factor $G$ for the current frame is computed by:

$$G = \frac{\sum_{n=0}^{N_f-1} |s(n)|}{\sum_{n=0}^{N_f-1} |s_{pf}(n)|}$$  \hspace{1cm} (7.17)

Where $N_f$ is the length of the current frame.
Chapter 7. Postfiltering

The gain-scaled enhanced speech signal, \( \hat{s}_{pf}(n) \) is given by:

\[
\hat{s}_{pf}(n) = g(n)s_{pf}(n) \quad n = 0, 1, \ldots, N_f - 1
\] (7.18)

Where \( g(n) \) is updated on a sample-by-sample basis and given by:

\[
g(n) = 0.85g(n-1) + 0.15G \quad n = 0, 1, \ldots, N_f - 1
\] (7.19)

The initial value of \( g(-1) = 1.0 \) is used. Then for next frame, \( g(-1) \) is set equal to \( g(N_f - 1) \) of the previous frame. For testing purposes, this system was used to process 6 input speech sentences (3 male, 3 female) from the NTT database. A MOS test comprising 8 listeners (4 trained and 4 untrained) was carried out on the processed speech files. The results are shown in Figure 7.11. The results obtained indicate that the speech quality degrades when highly attenuating the spectral valleys and also highly attenuating/amplifying the formants. It is also observed that the perceptual speech quality can be improved using the optimal attenuation and amplification levels for spectral valleys and the main formant magnitude. Since for voiced speech segments, the main formant is usually the first one, there is a dominant peak (i.e. not a flat characteristics of MOS score), which indicates the first formant being the most perceptually important formant. However, the second formant is occasionally considered as the main formant which is why we observe a semi-flat sensitivity of the second formant’s MOS score shown in Fig. 7.11-b. Therefore, our expectation is that by focusing on amplifying the main formant whilst maintaining the other formants and attenuating the valleys, a near optimum postfiltering can be achieved.

7.4.4 Optimum poles and shaping constants search

The polynomials representing the formants are used to design the new postfilter. The new postfilter is given by:

\[
P(z) = \frac{1}{D} \Omega(z) h_p(z)
\]

\[
\Omega(z) = \frac{\bar{Q}_m(z/\alpha_m) \prod_{k=1}^{F} Q_k(z/\alpha_k)}{\bar{Q}_m(z/\beta_m) \prod_{k=1}^{F} Q_k(z/\beta_k)}
\]

\[
h_p(z) = (1 - \gamma \mu z^{-1})
\] (7.20)
Figure 7.11: The resulting MOS scores for different amplifications/attenuations of the first three formants ((a), (b) and (c) for the first, second and third formant, respectively) and valleys (d).
Chapter 7. Postfiltering

$h_p(z)$ is a first-order high-pass filter included for tilt compensation [5] and $\mu$ is given by:

$$\mu = \frac{R(1)}{R(0)} \tag{7.21}$$

where

$$R(\tau) = \sum_{n=0}^{L} p(n) \cdot p(n + \tau) \tag{7.22}$$

and $p(.)$ is the impulse response of filter $\Omega(z)$, $L$ is the length of the impulse response, which is considered to be of length 2.5ms, $F$ indicates the number of the polynomials (formants), $Q_k(.)$ and $\tilde{Q}_m(.)$ are the polynomials representing the $k^{th}$ and the main formants, respectively and $\gamma = 0.9$. The reason for isolation of $\tilde{Q}_m(.)$ from the other polynomials is that $\tilde{Q}_m(.)$ is a function of both constants ($\alpha_m$ and $\beta_m$) and also the modified poles resulting from the first and second techniques described in sections 7.3.2.1 and 7.3.2.2, whereas $Q_k(.)$ is only a function of the constants $\alpha_k$ and $\beta_k$.

Moving the poles and changing the constants causes the main formant gain to increase. This gain is controlled by a factor $D$ given in Eq. 7.23.

$$D = \frac{1}{N} \sum_{k=0}^{N-1} \left| \Omega(\omega_k) H_p(\omega_k) \right|^2 \tag{7.23}$$

Since the poles representing the formants are available, the magnitude spectrum of $\Omega(.)$ and $H_p(.)$ are computed using the pole locations as follows:

$$\Omega(\omega_k) = \prod_{i=1}^{l_i} \left( \frac{1}{\prod_{k=1}^{F} \left( \prod_{i=1}^{l_i} \frac{1}{\sqrt{1 + \alpha_i^2 r_{ml}^2 - 2 \alpha_i r_{ml} \cos(\theta_{ml} - \omega_k)}} \right)^{F} \right) \prod_{k=1}^{F} \left( \prod_{i=1}^{l_i} \frac{1}{\sqrt{1 + \alpha_i^2 r_{kl}^2 - 2 \alpha_i r_{kl} \cos(\theta_{kl} - \omega_k)}} \right) \tag{7.24}$$

$$H_p(\omega_k) = \sqrt{1 + \mu^2 - 2 \mu \cos(\omega_k - \theta_\mu)} \quad \omega_k = k \frac{\pi}{N} \quad k = 0, 1, \ldots, N-1$$

Where $l_i$ indicates the number of poles for the $i^{th}$ formant and $\theta_\mu$ is given by:

$$\theta_\mu = \begin{cases} 0 & \text{if } \mu \geq 0 \\ \pi & \text{otherwise} \end{cases} \tag{7.25}$$
Next, we find the optimum constants and the pole locations representing the main formant so that better perceptual speech quality is obtained. Each candidate of the shaping constants and the pole locations leads to different attenuation/amplifications of the formants and the valleys. Thus, the attenuation/amplification level is computed for each formant and valley using the synthesised, $H(\omega)$, and the enhanced LP spectrums, $H_{eq}(\omega) = H(\omega)P(\omega)$. The computed levels are used to estimate MOS score using curves plotted in Fig. 7.11 for each formant. Since the MOS scores of the first formant has higher variations compared to the second and the third formants, the average of the MOS scores estimated for each formant and valley is considered as the resulting MOS score for the enhanced LP spectrum.

$$Mos_k(\alpha_k, \beta_k, p_k) = \frac{1}{N_{\mu}} \sum_{j=1}^{N_{\mu}} Mos(\omega_j, j) \quad (7.26)$$

Where $N_{\mu}$ is the number of formants and valleys, $\omega_j$ is formant or valley frequency, $Mos(\omega_j, j)$ is the corresponding MOS score for $\omega_j$ and $Mos_k(.)$ is average of the MOS scores.

The optimum constants and the modified poles representing the main formant are obtained by maximisation of the function $Mos_k(.)$.

$$\left\{ \alpha_k, \beta_k, p_k \right\}_{k=1, \ldots, F} \left\{ \tilde{p}_{ml} \right\}_{l=1, \ldots, l_m} = \text{arg max } MOS(\alpha_i, \beta_i, \tilde{p}_{ml}) \quad (7.27)$$

When

$$0.5 \leq \alpha_i \leq 0.7, \quad \alpha_i + 0.1 \leq \beta_i \leq 0.95, \quad 1 \leq i \leq F, \quad i \neq m$$

$$0.4 \leq \alpha_m \leq 0.65, \quad \alpha_m + 0.3 \leq \beta_m \leq 0.95$$

$$\tilde{p}_{ml} \leq 0.98 \quad l = 1, \ldots, l_m \quad (7.28)$$

Note that index $m$ indicates the main formant parameters, $F$ is the number of the formants, $l_m$ is the number of the poles, $\tilde{p}_{ml}$, representing the main formant.

This maximisation results in the optimum constants of $\{\alpha_i, \beta_i\}_{i=1, \ldots, F}$ for each polynomial, also the optimum pole locations of the main formant. This information is used to construct the new postfilter given in Equation 7.20.
As an example, Figure 7.12 shows the enhanced LPC spectrum in comparison with the synthesised one along with the corresponding optimum constants for each formant.

![Figure 7.12: The resulting LP spectrums using the new and conventional postfilters in comparison with the synthesised LP spectrum.](image)

7.5 The postfilter evaluation

Due to non-existence of quantitative measures, the new postfilter is evaluated by subjective listening test. The ACELP coder operating at 8.25 kb/s [4], where the speech signal is synthesised using the interpolated LP coefficients for four-5ms sub-block, is used. The employed postfiltering is as follows:

\[
P(z) = \Omega(z)(1 - \mu z^{-1})
\]

\[
\Omega(z) = \frac{1 - A(z/\alpha)}{1 - A(z/\beta)}
\]

(7.35)
Chapter 7. Postfiltering

Where \( \alpha = 0.7, \beta = 0.8 \) and \( \mu \) is obtained using the same procedure employed in the new postfilter.

In order to measure the subjective performance between the new postfilter and the one employed in ACELP, two postfilters were separately used in the same bit-rate and an A vs. B test was conducted. In this test, 10 sentence pairs for 6 speakers (3 male and 3 female speakers) were processed by the two 8.25 kb/s ACELP coder. Fourteen listeners (4 trained and 10 untrained) performed this test. The results are shown in Table 7.1. The results obtained indicate that the new postfilter provides significantly better perceptual speech quality than the conventional one. As shown in Figure 7.12, the reason is that the new postfilter provides narrower main formant bandwidth and attenuates the valleys more compared with the conventional postfilter while maintaining the other formants information.

<table>
<thead>
<tr>
<th>Speech Type</th>
<th>Better</th>
<th>Slightly better</th>
<th>Same</th>
<th>Slightly worse</th>
<th>Worse</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACELP + new postfilter vs. ACELP + conventional postfilter</td>
<td>5.8</td>
<td>43.3</td>
<td>25.5</td>
<td>22.7</td>
<td>2.7</td>
</tr>
</tbody>
</table>

7.6 Conclusions

In this chapter, the new postfilter design approach was described. This approach was based on LPC filter factorisation depending on the formant locations. The proposed postfilter was designed based on two features. These features are: i) narrower main formant bandwidth, ii) shaping the formants and attenuating valleys corresponding to better perceptual speech quality. This was performed by LP synthesis filter factorisation and searching the optimum pole locations representing the main formant and the constants \( \alpha, \beta \) for each formant. Since the proposed postfilter gain is not necessarily unity gain and so causes power distortion on the enhanced speech, the power controller was employed in concatenation with the postfilter to maintain the enhanced speech power.

The new postfilter was evaluated in comparison with the conventional postfilter. For this purpose, these two postfilters were used in ACELP 8.25 kb/s speech coder [4]. The subjective listening tests show that the proposed postfilter in combination with the ACELP provides significantly better quality than the conventional filter with ACELP.
Chapter 8

Conclusions and Future work

8.1 Preamble

The objective of this thesis has been to improve perceptual speech quality for low bit rate speech coders. The work has focused on two main areas:

1. Pre-processing: Since the speech model used in any speech coder and also the final speech quality depends on accurate parameters estimation, the aim of the first area was to make more regular speech such that the main parameters such as pitch and voicing level can be more reliably estimated. This was performed using a new pre-processor. Since irregular pitch variations were addressed as a factor causing inaccurate estimations, the pre-processor modified the residual speech signal such that pitch evolves smoothly within a frame and from frame-to-frame. In addition, it provides more correlation between successive pitch cycles and therefore more regular speech. The pre-processor was evaluated in combination with MELP 2.4 kb/s coder and also the robustness was tested against background noise.

2. Postfiltering: A new postfilter was designed based on the LP filter factorisation and using formants and valleys information. The proposed postfilter was tested against a conventional postfilter in combination with the full-rate GSM-AMR proposed standard based on ACELP coder [4] operating at 8.25 kb/s.

8.2 Concluding overview

Chapter 2 presented a review of digital speech coding. The main speech coding paradigms, the design criteria, and the existing standards were discussed.
Chapter 3 described fundamental speech coding techniques. Linear Prediction Coding (LPC) used to model speech spectral envelope was first introduced. This was followed by LPC to LSF transformation, which makes manipulations easier. The pitch determination algorithms based on time domain and frequency domain techniques were also introduced and compared. The existing techniques for speech classification as voiced/unvoiced and voicing-level estimation were addressed.

Chapter 4 introduced the fundamentals of WI and MELP coders. The basic analysis at the encoder and synthesis model, and the parameter estimation techniques employed by these coders were discussed.

The limitations of WI and MELP coders were investigated in Chapter 5. This was performed for the case of irregular pitch variations effects on the estimated parameters such as pitch and voicing strengths. The inaccuracy of these parameters were studied for the other parameter estimations in analysis stage (or in encoder), and on the synthesis stage process.

In order to overcome the problems caused by irregularly pitch variations, a new pre-processor was proposed in Chapter 6. The new pre-processor follows two tasks: i) modifying the residual speech signal such that the pitch evolves smoothly, ii) modifying low-correlated pitch cycles such that smooth pitch-cycle evolutions are resulted. In both tasks, the perceptual speech quality was maintained. Due to moving the pitch pulses during the modification, a misalignment may be created between the modified residual signal and the synthesised LP filter. This problem was solved using a new pre-analysis and post-synthesis technique, which controls the energy of the LP filter based on pitch pulse locations. The effect of the pre-processor on the estimated pitch values and voicing strengths were measured using Pitch Prediction Gain (PPG) and synthetic spectral distortion. An A vs. B comparative test was carried out on the processed speech signals resulting from MELP 2.4 kb/s in combination with the pre-processor and the MELP alone. The subjective listening tests showed that the pre-processor with MELP provides better speech quality. The robustness of the pre-processor was tested in the presence of background noise. The subjective listening tests showed the robustness of the pre-processor for SNRs over 15 dB.

The synthesised speech quality can be perceptually enhanced using a postfilter. In Chapter 7 the shortcomings of the conventional postfilters were investigated. These limits were overcome using a new postfilter scheme. The new postfilter achieved two tasks: i) providing narrower main formant bandwidth, ii) Shaping the formants and attenuating valleys to provide better perceptual speech quality. This was based on the LP synthesis filter factorisation using
formant information and searching for optimum shaping constants for each formant and optimum pole locations representing the main formant. An A vs. B test was carried out on the processed speech signals resulting from the ACELP coder [4] with the proposed and conventional postfilter. The subjective listening tests showed that the new postfilter provides better speech quality.

8.4 Future work

The research work carried out and presented in this thesis was primarily focused on improving speech quality through increasing the accuracy of parameter estimation at speech encoder as pre-processor, and enhancing the perceptual synthesised speech quality using postfiltering. This section suggests possible future research directions, which may further improve on the parameter estimation and speech quality.

1. The employed criterion for irregular pitch variations was based on pitch pulse locations. Since the irregular variations also affect the frequency characteristics of the speech signal, a new criterion can be considered in frequency domain based on irregular pitch-harmonic variations or a hybrid pitch pulse location technique based on time domain and pitch-harmonic variations in frequency domain can be also considered.

2. The employed technique for pitch modification was performed in time domain. This modification can be also performed on the residual spectrum in frequency domain or a hybrid time and frequency domain. It seems that enhancing the characteristics of speech signal in time and frequency domain with respect to more regular speech might improve the performance of the proposed pre-processor.

3. The employed technique for postfiltering uses the conventional postfilter form to provide the desired enhanced spectrum. Other techniques can be introduced by defining a target synthesised speech spectrum and searching for an optimum pole-zero filter (postfilter) such that the resulting spectrum from cascading the LP synthesis filter and the new postfilter matches the target spectrum.
Appendix A

List of publications


Bibliography


