A Fast and Low Bit rate Speech Coding Algorithm: Pulse Position Linear Predictive Coding

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

Master of Technology

in

Computer Science and Engineering

by

Prasanth G

Department of Computer Engineering
National Institute of technology, Calicut
Kerala - 673601
2005

Certificate

*This is to certify that the thesis entitled "***A Fast and Low Bit rate Speech Coding Algorithm: Pulse Position Linear Predictive Coding***" is a bonafide record of the work done by* **Prasanth G**, *(Roll no. Y3P214) under our supervision and guidance. The thesis has been submitted to the* **Department of Computer Engineering** *of* **National Institute of Technology, Calicut** *in partial fulfilment of the award of the Degree of* **Master of Technology** *in* **Computer Science and Engineering**.

Dr. V.K. Govindan
Professor & Head of Dept.
Dept.of Computer Engineering
NIT Calicut

Mr. K. A. Abdul Nazeer
Project Guide
Lecturer
Dept.of Computer Engineering
NIT Calicut

Abstract

The past decade has witnessed substantial progress towards the application of low bitrate speech coders to civilian and military communications as well as computer related voice applications. Central to this progress has been the development of new speech coders capable of producing high quality speech at low data rates. Most of these coders incorporate mechanisms to: synthesize speech using Linear Predictive Coding, represent the spectral properties of speech, provide speech waveform matching, and "optimize" the coder's performance for the human ear. A number of these coders have already been adopted in national and international cellular telephony standards.

The objective of speech is communication, whether face to face or cell phone to cell phone. Though intelligibility is the key factor in Speech Coding, naturalness of speech is desirable in most cases. There are vocoders like LD-CELP & MPELP which provide high quality and naturalness at medium bitrates. Although CELP has good quality at a low bitrate of 4.8 kbps, it is undesirable for real-time communication purposes due to it's highly complex algorithm. Therefore the search for a low complexity vocoder that can achieve reasonably good intelligibility and quality with low bit rate is still on. This thesis suggests a new vocoder to achieve low bit rate, acceptable quality and low complexity.

The main factor affecting natural sounding speech synthesis is the excitation signal used. This thesis work presents a new excitation signal and designs a vocoder based on the new excitation signal and a new parametric representation for the excitation signal. The new excitation signal is obtained by placing pulses at peaks and valleys of the residue signal. The new vocoder provides good quality at low codec delay and low bitrate.

Acknowledgements

First of all I would like to thank God Almighty for helping me successfully complete this thesis work.

I would like to thank **Dr. V. K. Govindan**, Professor & Head of the Department, CSED, NITC, for his constant support and encouragement throughout the project.

With a profound sense of gratitude, I would like to express my heartfelt thanks to my guide **Mr. K. A. Abdul Nazeer**, Lecturer, CSED, NITC for his expert guidance, cooperation and immense encouragement. I am highly indebted to him for his commendable support and evincing keen interest in this project.

I would also like to thank our Course & Project Coordinator **Ms. Priya Chandran**, Asst. Professor, CSED, NITC, for her help & support during the entire course. Last but not least, I also like extend my sincere thanks to the entire faculty and staff of the Dept. of Computer Engineering, NITC, for giving us all the necessary facilities and support throughout the course of my Master's Degree.

I needed the help and support from a lot of people to complete a Thesis work successfully. I am grateful to **Dr. Sathidevi**, Asst. Professor, Dept. of Electronics and Communication and **Mr. Kunjumoideen Kutty**, ISRO for their help during the project work. Without their help I would not have been able to complete this project in such a vast subject.

I also express my thanks to my loving parents, sister, classmates and friends, for their support and encouragement in the successful completion of this project work.

Prasanth G.

Contents

Chapter

1 Introduction 1

1.1 Historic Perspective 3

1.2 Organization of The Thesis 7

2 Modelling speech production 9

2.1 Human model of Speech Production 10

2.1.1 Anatomy 10

2.1.2 Role of Vocal Tract 13

2.1.3 Voice Production 17

2.2 Summary 19

3 Voice Coding strategies 20

3.1 Introduction 20

3.2 Speech Coder Classification 21

3.2.1 Waveform Coders 21

3.2.2 Vocoders(Voice Coders) 22

3.2.3 Hybrid Coders 32

3.3 Summary 32

4 Proposed Algorithm: Pulse Position Linear Prediction(PPLP) 33

4.1 Introduction 33

4.2 Previous Works 33
4.3 Approach in this Thesis 34
4.4 Analysis 36
4.4.1 Gain Computation 37
4.4.2 Pulse Position Analysis 37
4.4.3 Quantization 39
4.5 Synthesis 40
4.5.1 Excitation Signal Generation 40
4.5.2 LPC Synthesis 40
4.6 Summary 41

5 Computer Simulation and Discussion 42
5.1 Introduction 42
5.2 Performance Issues 42
5.2.1 Voice Quality 43
5.2.2 Bitrate 45
5.2.3 Algorithm Complexity 46
5.2.4 Coding Delay 46
5.3 Summary 46

6 Comparison 47
6.1 PPLP 47
6.2 LPC 10 48
6.3 MPLP 49
6.4 CELP 50
6.5 LD-CELP 50
6.6 Summary 50

7 Conclusion and Future Scope 51

7.1 Conclusion 51

7.2 Future Scope 51

Bibliography 53

Appendix

A Simulation Results 55

B Linear Predictive Coding 58

B.1 Introduction 58

B.2 Correlation Coefficients 58

B.3 Linear Prediction Model 59

B.4 Algorithm Used for LPC 62

B.4.1 Levinson-Durbin Algorithm 63

B.5 Summary 65

Tables

Table

4.1 Pulse Positions and corresponding bit vector codes 38

4.2 Bit allocation for LPC parameters per frame 39

4.3 Bit allocation per second 40

6.1 Comparison of various standard Vocoders and PPLP 49

Figures

Figure

1.1 Early attempts to synthesize voice : Schematic diagram of the Voder 6

1.2 Electrical vocal tract by H.K.Dunn 6

2.1 Human speech production mechanism 12

2.2 Human speech production Block diagram 12

2.3 Cross section of Larynx 14

2.4 Mathematical model of human sound production 14

2.5 Cross section of Larynx showing complete phonation cycle 17

3.1 Simplified tubular model of vocal tract used in LPC 22

3.2 A model for speech synthesis 23

3.3 Voiced/Unvoiced Decision 24

3.4 AMDF 26

4.1 Gain Envelope 41

5.1 M.O.S 44

A.1 Input Speech Signal 55

A.2 Output Speech synthesized by PPLP system 55

A.3 Input Speech Spectrogram 56

A.4 Output Speech Spectrogram 56

A.5 PPLP System: Analysis 57

A.6 PPLP System: Synthesis 57

B.1 Vocal tract model 58

B.2 Modelling of Speech Production 59

B.3 Linear Prediction (IIR) model of speech 60

B.4 Graphical representation of Voiced and Unvoiced Speech Production 60

Chapter 1

Introduction

Although with the emergence of optical fibers, bandwidth in wired communications has become inexpensive, there is a growing need for bandwidth conservation and enhanced privacy in wireless cellular and satellite communications. In particular, cellular communications have been enjoying a tremendous worldwide growth and there is a great deal of R&D activity geared towards establishing global portable communications through wireless personal communication networks (PCNs). On the other hand, there is a trend towards integrating voice related applications (e.g. voicemail) on desktop and portable personal computers often in the context of multimedia communications. Most of these applications require that the speech signal is in digital format so that it can be processed, stored, or transmitted under software control. Although digital speech brings flexibility and opportunities for encryption, it is also associated (when uncompressed) with a high data rate and hence high requirements of transmission bandwidth and storage. Speech Coding or Speech Compression is the field concerned with obtaining compact digital representations of voice signals for the purpose of efficient transmission or storage. Speech coding involves sampling and amplitude quantization. While the sampling is almost invariably done at a rate equal to or greater than twice the bandwidth of analog speech, there has been a great deal of variability among the proposed methods in the representation of the sampled waveform. The objective in speech coding is to represent speech with a minimum number of bits

while maintaining its perceptual quality. The quantization or binary representation can be direct or parametric. Direct quantization implies binary representation of the speech samples themselves while parametric quantization involves binary representation of speech model and/or spectral parameters. Vocoders based on Linear predictive coding represents the speech model using a binary representation of the parameters needed to reproduce the vocal tract filter and excitation signal.

Many of the Vocoders were accepted as Digital communication standards. Digital cellular standards that use 13, 8 and 6.7 kbits/s speech coding algorithms were established and significant progress has been made in getting high quality speech coders for the future half rate standards. In addition, high quality low delay coding at 16 kbits/s has also been achieved. But vocoders with low delay coding and lower bit rates are still a challenging problem. Current research is now focussed upon achieving high quality speech at and below 4.0 kbits/s. Improving the performance of low rate coders operating in noisy channels is also an open problem. Additionally there is a demand for robust lowrate coders that will accommodate signals other than speech, such as music. Finally, military oriented coding research is geared towards obtaining improved coders, 2.4 kbits/s coders, and also very low rate coders for applications that require increased tolerance to channel errors and low probability of intercept.

From the implementation standpoint, high quality speech at lower information rates always come at the expense of increased algorithmic complexity. Although the development of faster and highly integrated signal processing devices may partially solve this problem, low complexity and low sensitivity to machine precision will still be important for low power single chip implementations. This paper provides a robust algorithm that provides reasonable quality speech at low bit rates and yet with less algorithmic complexity.

1.1 Historic Perspective

Speech coding research started over fifty years ago with the pioneering work of Homer Dudley of the Bell Telephone Laboratories. The motivation for speech coding research at that time was to develop systems for transmission of speech over low bandwidth telegraph cables. Dudley practically demonstrated the redundancy in the speech signal and provided the first analysis synthesis method for speech coding. The basic idea behind Dudley's voice coder or vocoder was to analyze speech in terms of its pitch and spectrum and synthesize it by exciting a bank of ten analog band pass filters (representing the vocal tract) with periodic (buzz) or the random (hiss) excitation (for voiced and unvoiced sounds respectively). The channel vocoder received a great deal of attention during World War II because of its potential for efficient transmission of encrypted speech. Formant and pattern matching vocoders along with improved analog implementations of channel vocoders were reported through the fifties actions and sixties. In the formant vocoder the resonant characteristics of the filter bank track the movements of the formants. In the pattern matching vocoder the best match between the short time spectrum of speech and a set of stored frequency response patterns is determined and speech is produced by exciting the channel filter associated with the selected pattern. The pattern matching vocoder was essentially the first analysis synthesis system to implicitly employ vector quantization.

Although early vocoder implementations were based on analog speech representations, digital representations were rapidly gaining interest due to their promise for encryption and high fidelity transmission and storage. In particular, there had been a great deal of activity in Pulse Code Modulation (PCM) in the forties. PCM is a frequency domain straightforward method for discrete time, discrete amplitude approximation of analog waveforms and does not have any mechanism for redundancy removal. Quantization methods that exploit the signal correlation,

such as Differential PCM (DPCM), Delta Modulation (DM), and Adaptive DPCM were proposed later and speech coding with PCM at 64 kbits/s and with ADPCM at 32 kbits/s eventually became CCITT standards. The first all-electrical speech synthesizer known as "Voder" shown in figure 1.1 was demonstrated at World's Fair in 1939 . In 1950 H. K. Dunn developed a better model of electrical vocal tract (Figure 1.2).

With the flexibility offered by digital computers, there was a natural tendency to experiment with more sophisticated digital representations of speech. Initial efforts concentrated on the digital implementation of the vocoder. A great deal of activity, however, concentrated on the linear speech source system production model developed by Fant in the late fifties. This model consists of a linear slowly time varying system (for the vocal tract and the glottal model) excited by periodic impulse train excitation (for voiced speech) and random excitation (for unvoiced speech). Thus source system model became associated with Autoregressive (AR) time series methods. I was also associated with cases where the vocal tract filter is all pole and its parameters are obtained by Linear Prediction analysis, a process where the present speech sample is predicted by the linear combination of previous samples. Itakura and Saito and Atal and Schroeder were the first to apply Linear Prediction (LP) techniques to speech. Atal and Hanauer later reported an analysis synthesis system based on LP. Theoretical and practical aspects of linear predictive coding (LPC) were examined by Markel and Gray and the problem of spectral analysis of speech using linear prediction was addressed by Makhoul and Wolf.

However, LP is not the only method for source system analysis. Homomorphic analysis, a method that can be used for separating signals that have been combined by convolution, has also been used for speech analysis. Oppenheim and Schafer were strong proponents of this method. One of the inherent advantages

of homomorphic speech analysis is the availability of pitch information from the cepstrum.

The emergence of VLSI technologies along with advances in the theory of digital signal processing during the sixties and seventies provided even more incentives for getting new and low rate improved solutions to the speech coding problem. Analysis - Synthesis of speech using the Short Time Fourier Transform (STFT) was proposed by Flanagan and Golden in a paper entitled "Phase Vocoder". In addition, Schafer and Rabiner designed and simulated an analysis synthesis system based on the STFT, and Portnoff provided a theoretical basis ally for the time frequency analysis of speech using the STFT. In the mid to late seventies there was also a continued activity in linear prediction, transform coding, and subband coding. During the seventies, there were also parallel efforts for the application of linear prediction in military secure communications. A federal standard (FS1015) which is based on the LPC 10 algorithm, was developed in the early eighties.

Research efforts in the eighties and nineties have been focused upon developing robust lowrate speech coders capable of producing highquality speech for communications applications. Much of this work was driven by the need for narrowband and secure transmission in cellular and military communications. Competing methodologies promoted in the eighties included: sinusoidal on analysis synthesis of speech proposed by McAulay and Quatieri , multiband excitation vocoders proposed by Griffin and Lim, multipulse and vector excitation schemes for LPC proposed by Atal , and vector quantization (VQ) promoted by Gersho and Gray lysis and others. Vector quantization proved to be very useful in encoding LPC parameters. In particular, Atal and Schroeder proposed a linear prediction algorithm with stochastic vector excitation which they called "Code Excited Linear Prediction" (CELP). The stochastic excitation in CELP is determined using a perceptually weighted closed loop (analysis by synthesis) optimization. CELP

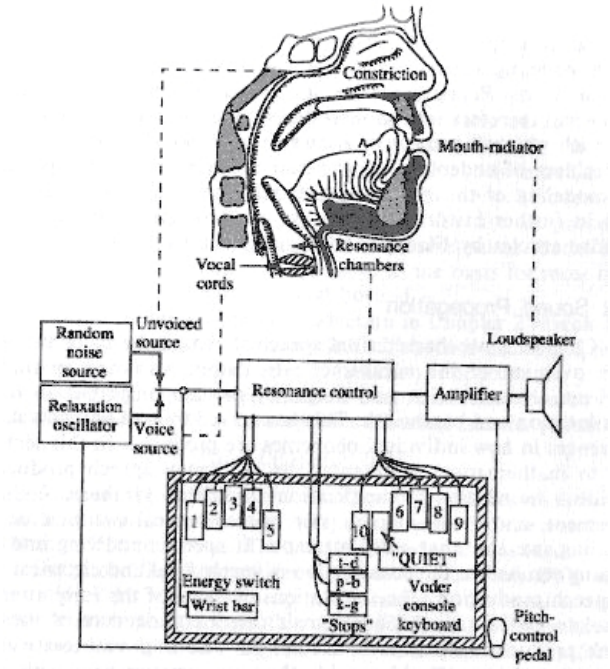

Figure 1.1: Early attempts to synthesize voice : Schematic diagram of the Voder

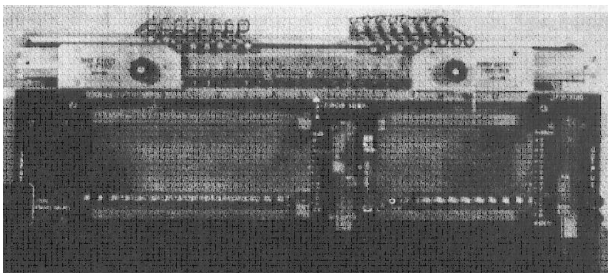

Figure 1.2: Electrical vocal tract by H.K.Dunn

coders are also called hybrid coders because they combine the features of traditional vocoders with the waveform matching features of waveform coders. Although the first paper on CELP addressed the feasibility of vector excitation coding, follow up work essentially demonstrated that CELP coders were capable of producing medium rate and even low rate speech adequate for communications applications. Although earlier it was an algorithm only feasible on Super Computers, Realtime implementation of hybrid coders became feasible with the development of highly structured codebooks.

Progress in speech coding, particularly in the late eighties, enabled recent adoptions of low rate algorithms for mobile telephony. An 8 kbits/s hybrid coder has already been selected for the North American digital cellular standard, and a similar algorithm has been selected for the 6.7 kbits/s Japanese digital cellular standard. In Europe, a standard that uses a 13 kbits/s regular pulse excitation algorithm has been completed and partially deployed by the "Group Speciale Mobile" (GSM). Parallel standardization efforts for secure military communications have resulted in the adoption of a 4.8 kbits/s hybrid algorithm for the Federal Standard 1016. In addition, a 6.4 kbits/s improved multiband excitation coder has been adopted for the International Maritime Satellite (INMARSATM) system and the Australian Satellite (AUSSAT) system. Finally we note that there are plans to increase the capacity of cellular networks by introducing half rate algorithms in the GSM, the Japanese, and the North American standards.

1.2 Organization of The Thesis

- Chapter 2 gives an overview of Human Speech Production Model.
- Chapter 3 explains the various Voice coding strategies used currently.
- Chapter 4 describes the newly proposed algorithm and excitation signal in

detail.

- Chapter 5 explains the Computer simulations done and discusses the results obtained from the simulations.
- Chapter 6 provides a comparison of PPLP vocoder with other present day vocoders.
- Chapter 7 concludes the work and discusses the future scope of the work.
- Appendix A gives the figures showing the results of the Computer Simulation.
- Appendix B describes Linear Predictive Coding, which is the heart of the present day speech coding technology and the algorithms involved in it.

Chapter 2

Modelling speech production

There are two types of Audio Coding Techniques: Source Modelling and Receiver Modelling. Receiver modelling tries to model the Receiver and will encode the signal based on the Receiver. So information that are not important at Receiver are always ignored. For example audio receiver modelling is done based on the working of Cochlea. Thus it is able to take advantage of the limitations of Cochlea. This is also called Psychoacoustic modelling, because it is related with how the acoustic details affects the brain. The Linear Predictive Coding approach to speech production is based on Source Modelling. In Source Modelling we model the source of the signal. Thus we can represent the actual signal by just certain parameters of the Source Model which represents the state of source at that instant. It is like one is able to visualize a football game by reading the commentary. This can be done in systems which are based on strict rules. Since speech production is done by a physical source which works based on some rules we can apply this technique here. So we represent the speech as few parameters representing the state of the vocal tract filter and the parameters needed to reproduce the excitation signal.

Linear Predictive Coding mathematically tries to simulate the actual process of Speech production in Human. Therefore for a complete study about Linear Predictive Coding, it is essential to know about the way Speech production actually happens in human beings. Here we present a small description of the Human Speech Production model. For more details refer [4]

2.1 Human model of Speech Production

Speech signal is in the form of a sound pressure wave that travels from the speaker's mouth to a listener's ears. Speech signals are composed of a sequence of sounds that serve as a symbolic representation for a thought that the speaker wishes to relay to the listener. The science that studies the characteristics of human sound production especially for the description, classification, and transcription of speech, is called phonetics.

Human speech production begins with an idea or thought that the speaker wants convey to a listener. The speaker conveys this thought through a series of neurological signals. To achieve this, a speaker forms an idea to convey, converts the idea into a linguistic structure by choosing appropriate words or phrases to represent that idea, orders the words or phrases based on learned grammatical rules associated with the particular language, and finally adds any additional local or global characteristics such as pitch intonation or stress to emphasize aspects important for overall meaning. Once this has taken place, the human brain produces a sequence of motor commands that move the various muscles of the vocal system to produce the desired sound pressure wave. This acoustic wave is received by the talker's auditory system and converted back to a sequence of neurological pulses that provide necessary feedback for proper speech production. This allows the talker to continuously monitor and control the vocal organs by receiving his or her own speech as feedback. The listener also convert speech into a sequence of neurological pulses which are then interpreted in the auditory cortex of the brain to determine what idea was received.

2.1.1 Anatomy

Speech waveform originates from the voluntary movements of anatomical structures which make up the human speech production system. Figure 2.1 por-

trays a diagram of the human speech production system in which we view the anatomy midway through the upper torso. The gross components of the system are the lungs, the trachea(windpipe), larynx (organ of voice production), pharyngeal cavity (throat), oral or buccal cavity (mouth), and nasal cavity (nose). In technical discussions, the pharyngeal and oral cavities are usually grouped into one unit referred to as the vocal tract, and the nasal cavity is often called the nasal tract. Accordingly, the vocal tract begins at the output of the larynx, and terminates at the input to the lips. The three main cavities of the speech production system (vocal plus nasal tracts) comprise the main acoustic filter. The filter is excited by the organs below it and is loaded at its main output by a radiation impedance due to the lips. The articulators, most of which are associated with the filter itself, are used to change the properties of the system, its form of excitation, and its output loading over time.

A simplified acoustic model illustrating these ideas is shown in Figure 2.2. In the average adult male(female), the total length of the vocal tract is about 17 (14) cm. The vocal tract of an average child is 10 cm. Repositioning of the vocal tract articulators causes the cross-sectional area of the vocal tract to vary along its length from zero (complete closure) to greater than 20 cm^2. The nasal tract constitutes an auxiliary path for the transmission of sound. By the opening of the velum it can get coupled with vocal tract and can substantially influence the frequency characteristics of the sound radiated from the mouth. Sounds produced like that are called nasal sounds.

Let us now focus on the larynx. From a technical point of view, the larynx has a simple, but highly significant, role in speech production. Its function is to provide a periodic excitation to the system for speech sounds that we will come to know as "voiced". Roughly speaking, the periodic vibration of the vocal folds is responsible for this voicing. From an anatomical point of view, however, the

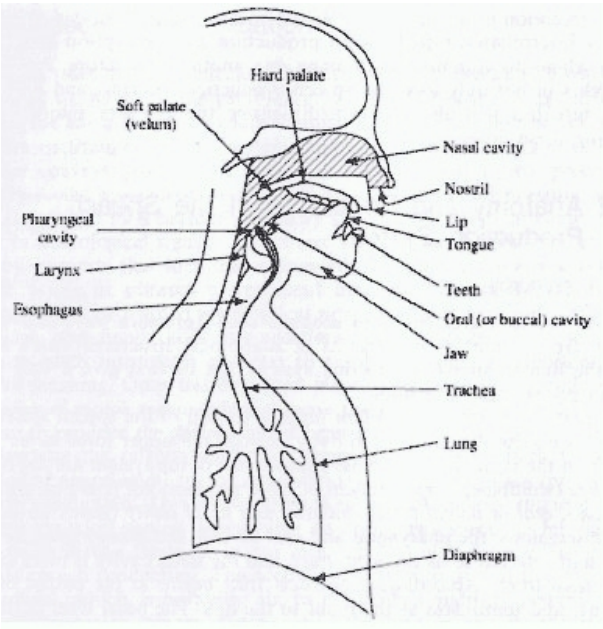

Figure 2.1: Human speech production mechanism

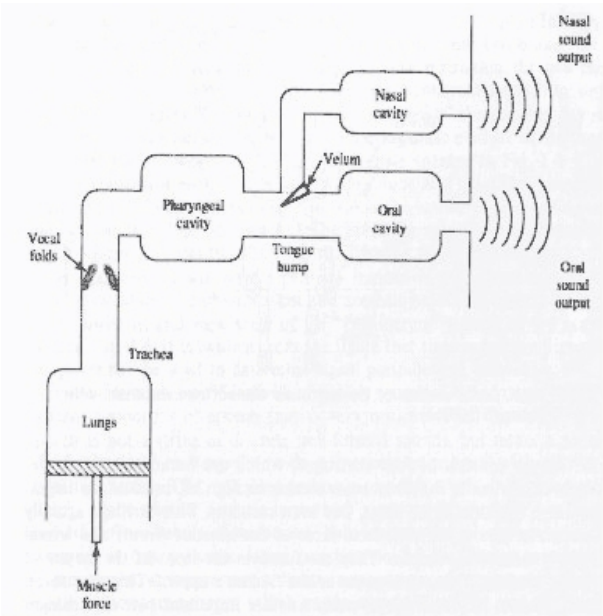

Figure 2.2: Human speech production Block diagram

larynx is an intricate and complex organ that has been studied extensively by anatomists and physiologists. Figure 2.3 shows the main features of Larynx. The main framework of the larynx consists of four cartilages, which are named in the figure. The thyroid cartilage actually consists of two plates that form most of the front and side walls of the larynx. They are fused in the front of the larynx at the projection commonly known as the "Adam's apple". The main cartilages of the larynx are held together by a network of ligaments and membranes that control their positioning during voicing and other functions of the organ. The vocal folds are a pair of elastic bands of muscle and mucous membrane that stretch from the thyroid cartilage in the front to the arytenoids in the rear.

2.1.2 Role of Vocal Tract

The spectral characteristics of the speech wave are time varying, since the physical system changes rapidly over time. As a result, speech can be divided into sound segments that possess similar acoustic properties over short periods of time. We note that speech sounds are typically partitioned into two broad categories: (1) vowels that contain no major airflow restriction through the vocal tract, and (2) consonants that involve a significant restriction through the vocal tract and are therefore weaker in amplitude and often "noisier" than vowels. The vowel portion is higher in amplitude and contains strong periodic structure.

One of the most important properties of speech that is very noticeable is that speech is not a string of discrete well-formed sounds, but rather a series of "steady-state" or "target" sounds with intermediate transitions. The preceding and/or succeeding sound in a string can grossly affect whether a target is reached completely, how long it is held, and other finer details of the sound. This interplay between sounds in an utterance is called **coarticulation**. Changes witnessed in the speech waveform are a direct consequence of movements of the speech system

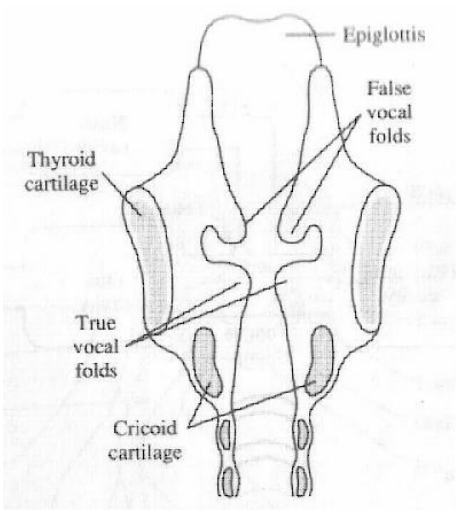

Figure 2.3: Cross section of Larynx

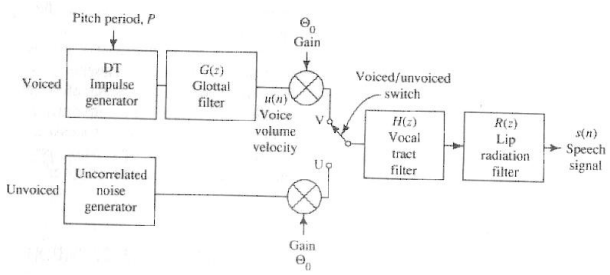

Figure 2.4: Mathematical model of human sound production

articulators, which rarely remain fixed for any sustained period of time. The inability of the speech production system to change instantaneously is due to the requirements of finite movement of the articulators to produce each sound. These articulators are human tissues and/or muscles which are moved from one position to another to produce the desired speech sounds. Unlike the auditory system which has evolved solely for the purpose of hearing, organs used in speech production are shared with other functions such as breathing, eating and smelling. The multiple role of these organs suggests that their present form may not be optimal for human communication. Due to the limitations of the organs for human speech production and the auditory system, typical human speech communication is limited to a bandwidth of 7 to 8 kHz.

If we look at a frequency domain magnitude spectra plot of an acoustic waveform we can notice that there are well defined regions of emphasis(loosely speaking resonances) and deemphasis (antiresonances) in the spectrum. These resonance are a consequence of the articulators having formed various acoustical cavities and subcavities out of the vocal tract cavities, much like concatenating different lengths of organ pipe in various orders. So the location of these resonances in the frequency domain depend upon the shape and physical dimensions of the vocal tract. Conversely, each vocal tract shape is characterized by a set of resonant frequencies. From a system modelling point of view, the articulators determine the properties of the speech system filter. Since these resonances tend to "form" the overall spectrum, speech scientists refer to them as **formants**. This term is often used to refer to the nominal center frequencies of the resonances, so "formant" may be used interchangeably with **formant frequency**. The formants in spectrum are usually denoted $F_1, F_2, F_3, ...$, beginning with the lowest frequency. In principle, there are infinite number of formants in a given sound, but in practice, we usually find 3-5 in Nyquist band after sampling.

Since the vocal tract shape is varied as a function of time to produce the desired speech sounds for communication, so must the spectral properties of the speech signal vary with time. In spectrogram of the waveform we can see the formant frequencies as dark horizontal bands in the spectrogram (Refer Figure A.3). Voiced speech is characterized as vertical striations due to the periodic nature of the glottal excitation(See Figure A.3). Unvoiced speech sounds, due to their noise like excitation, are characterized by rectangular dark patterns, with somewhat random occurrences of light spots due to sudden variations in energy. Spectrograms can only represent spectral amplitude, with no means of illustrating phase.

Types of Excitation One of the principal features of any speech sound is the manner of excitation. There are two element excitation types: (1) voiced and (2) unvoiced. Four other types of excitation, which are really just combinations of voiced, unvoiced and silence are usually delineated for modelling and classification purposes. These are: (3)mixed, (4)plosive, (5) whisper, and (6) silence. One or more of these excitation types may be blended in the excitation of a particular speech sound or class of sounds.

Voiced sounds are produced by forcing air through the glottis or an opening between the vocal folds. The tension of the vocal cords is adjusted so that they vibrate in oscillatory fashion. The periodic interruption of the subglottal airflow results in quasi-periodic puffs of air that excite the vocal tract. The sound produced by the larynx is called voice or phonation. An example of voiced sound is the vowel /I/ in the utterance of "six".

Unvoiced sounds are generated by forming a constriction at some point along the vocal tract, and forcing air through the constriction to produce turbulence. An example is the /s/ sound in "six".

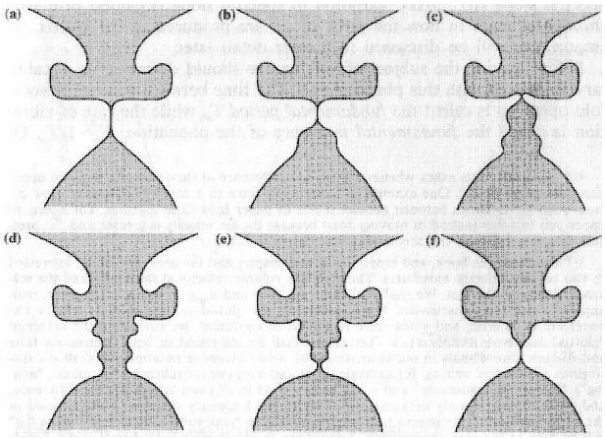

Figure 2.5: Cross section of Larynx showing complete phonation cycle

A sound may be simultaneously voiced and unvoiced (mixed). An example of this is the letter "z" (phonetic symbol /z/) in the phrase "three zebras". Further some speech sounds are composed of short region of silence, followed by region of voiced speech, unvoiced speech, or both. These plosive sounds are formed by making a complete closure (normally toward the front of the vocal tract), building air pressure behind the closure, and suddenly releasing it. A plosive example (silence+unvoiced) is the sound corresponding to /t/ in "pat". Another (silence+voiced) is the /b/ in "boot" in which the voicing for the following "oo" (/u/) vowel is turned on immediately after the release of pressure. The example of a plosive points out that for modelling purposes, it will be necessary to consider "silence" to be a form of "excitation". Whatever the source of excitation, the vocal tract acts as a filter amplifying certain frequencies while attenuating others.

2.1.3 Voice Production

Voicing is most interesting form of excitation in that it involves a special organ for its production. Voicing is accomplished when the abdominal muscles force the diaphragm up, pushing air out from the lungs into the trachea, then up to the glottis, where it is periodically interrupted by movement of the vocal folds.

The repeated opening and closing of the glottis is in response to the subglottal air pressure from trachea. Forces responsible for the glottal pulse cycle affect the shape of the glottal waveform and are ultimately related to its corresponding spectral characteristics. First, the subglottal air pressure below the true vocal folds builds up, pushing the vocal folds apart. The glottal slit begins to open, and air begins to rush out from the trachea through the glottis. The subglottal air pressure continues to force the glottis to open wider and outward, resulting in increased airflow through the glottis. By using the notion of conservation of energy, the kinetic energy is expressed as the square of the air velocity; whereas the potential energy is proportional to the air pressure. As the vocal folds spread apart, air velocity increases significantly through the narrow glottis, which causes a local drop in air pressure. Therefore, when the vocal folds are closed, air pressure and potential energy are high. As the glottis opens, air velocity and kinetic energy increase, while pressure and potential energy decrease. The glottis continues to open until the natural elastic tension of the vocal folds equals the separating force of the air pressure. At this point the glottal opening and the rate of airflow have reached their maxima. The kinetic energy that was received by the vocal folds gathers momentum and a suction effect caused by a Bernoulli force occurs when the glottis becomes narrow enough. Both the elastic restoring force and Bernoulli force act to close the vocal folds abruptly. The subglottal pressure and elastic restoring forces during closure causes the cycle to repeat. The time between successive vocal fold opening is called the fundamental period T_o, while the rate of vibration is called the fundamental frequency of the phonation, $F_o = \frac{1}{T_o}$. The fundamental period is dependant on the size and tension of the speaker's vocal fold at any given instant. This is the reason why men have lower fundamental frequency voice than females and children whose vocal fold have small average size and tension. Pitch is the perceived fundamental frequency of a sound, whether or

not that sound is actually present in the waveform. For men the possible pitch range is usually found somewhere between 50-250 Hz where as for women it lies between 120-500 Hz. Everyone has a habitual pitch level and it is shifted up or down to show stress, intonation and emotion.

2.2 Summary

In this chapter, we have explained "How Human Speech Production model works". This chapter explains our Speech production system. Here we portray it as an excitation source (Larynx) driving a Vocal tract physical filter. A good understanding of how speech is produced actually is needed to understand how we can code and synthesize speech.

Chapter 3

Voice Coding strategies

3.1 Introduction

Theoretically speaking, speech coding can be achieved based on two facts: redundancy in speech signals, and the perception properties of human ears. Speech signals contain a great deal of redundancy. The redundancies in speech are due to reasons give below:

(1) Speech signals have a strong correlation among adjacent samples.

(2) Voiced speech is quasi-periodic.

(3) The shape of the vocal tract is slowly time-varying.

(4) The distributions of transmission code values are not uniform.

The redundancy from the correlation between samples demonstrated by the uneven short term spectrum could be removed by appropriate filtering. The synthesized model in low bit speech coding is based on the second aspect of the voice being quasi-periodic. The third type of redundancy is used as the base for segmenting speech signals into frames. The fourth type of redundancy gets worked on mainly with various probability coding techniques.

"How humans hear" can be used in designing of the speech coder as well. We model the human cochlea for knowing about how human beings hear. It uncovers lots of facts useful for audio compression.:

- Limits of perception: Most speech ranges around 200-3400 Hz. The frequency resolution of the ear is also limited at about 2 Hz. Quantization can remove the redundant information to ears.

- Masking effects: Sometimes, an otherwise audible sound is masked with another one, e.g. a loud sound will cover up a weaker sound (intensity masking). Other factors also might cause masking: frequency, temporal separation of the sounds, tonality of the masker and so forth.

- Sensitivity of perception: Human ears are more sensitive to low frequency bands than higher ones. Voice pitches (the fundamental frequency of the vocal cord vibration) and formant structure are mainly located in the low frequency part. Another important property is that human ears are not sensitive to phase information of speech signals.

3.2 Speech Coder Classification

Voice coding strategies can be mainly classified into three categories namely Waveform coder, Vocoder and Hybrid Coder. More detailed information about different coders in use can be seen in [3] and [8].

3.2.1 Waveform Coders

Non speech specific coders or waveform coders are concerned with the faithful reconstruction of the time domain waveform and generally operate at medium rates. Though it will surely retain good quality the drawback is that it is too difficult to attain bit rates less than 32 kbps with it. But the advantage is that it is not speech specific and hence can be used for any kind of signals. This generally works on sample by sample basis. Since the waveform is preserved, objective quality measures can be used. The quality measure usually used is the Signal to

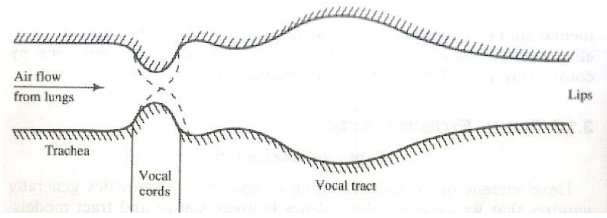

Figure 3.1: Simplified tubular model of vocal tract used in LPC

Noise Ratio (SNR). Examples of waveform coders are PCM 64 kbps, ADPCM 32 kbps, CVSDM 32 kbps

3.2.2 Vocoders(Voice Coders)

Speech specific coders or voice coders (vocoders) rely on speech models and are focussed upon producing perceptually intelligible speech without necessarily matching the waveform. Vocoders are capable of operating at very low rates but also tend to produce speech of synthetic quality. Vocoders work based on the simplified model of vocal tract as a simple tube as in Figure 3.1 with an excitation source at one end. The coding process in vocoders involve analyzing speech and extracting and transmitting model parameters. The decoding process uses model parameters to synthesize speech. This is a speech specific coder. It may not work well with other signals. The quality measures used are subjective. The quality measures used are MOS, DRT and DAM. Examples are LPC 10,CELP etc. This is also the area of work of this thesis.

The Linear Predictive Coder (Government Standard LPC-10)

LPC-10 is the most well-known among the many descendants of the channel vocoder. In LPC-10 the vocal tract is modelled as a single linear filter whose output y_n is related to the input ϵ_n by

$$y_n = \sum_{i=1}^{M} b_i y_{n-i} + G\epsilon_n \tag{3.1}$$

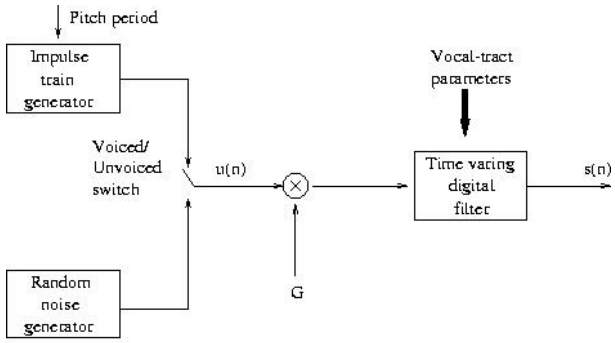

Figure 3.2: A model for speech synthesis

where G is called the gain of the filter. As in the case of the channel vocoder the input to the vocal tract filter is either the output of a random noise generator or a periodic pulse generator A block diagram of the LPC receiver is shown in Figure 3.2.

At the transmitter, a segment of speech is analyzed The parameters obtained include a decision as to whether the segment of speech is voiced or unvoiced, the pitch period if the segment is declared voiced, and the parameters of the vocal tract filter. The input speech is generally sampled at 8000 samples per second. In the LPC-10 standard, the speech is broken into 180 sample segments, corresponding to 22.5 milliseconds of speech per segment.

The Voiced/Unvoiced Decision

In Figure 3.3 we can notice that the samples of the voiced speech have larger amplitude; that is, there is more energy in the voiced speech. Also, the unvoiced speech contains higher frequencies. As both speech segments have average values close to zero, this means that the unvoiced speech waveform crosses the x=0 line more often than the voiced speech sample. Therefore, we can get a fairly good idea about whether the speech is voiced or unvoiced based on the energy in the

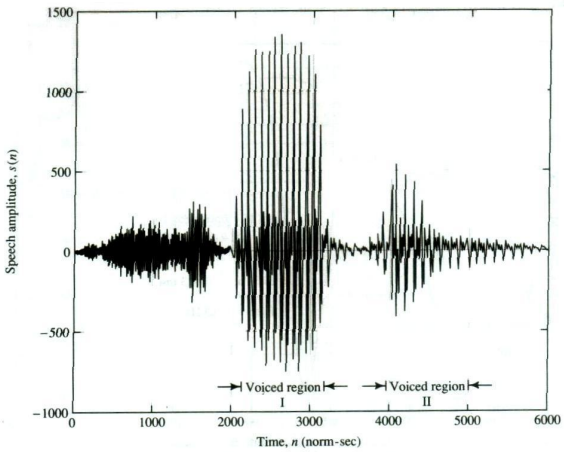

Figure 3.3: Voiced/Unvoiced Decision

segment relative to background noise and the number of zero crossings with in a specified window. In the LPC-10 algorithm, the speech segment is first low-pass filtered using a filter with a bandwidth of 1 kHz. The energy at the output relative to the background noise is used to obtain a tentative decision about whether the signal in the segment should be declared voiced or unvoiced. The estimate of the background noise is basically the energy in the unvoiced speech segments. This tentative decision is further refined by counting the number of zero crossings and checking the magnitude of the coefficients of the vocal tract filter. The voicing decisions of the neighboring frames are also consider in order to prevent a single voiced frame getting sandwiched between unvoiced frames.

Estimating the Pitch Period

Estimating the pitch period is one of the most computationally intensive steps of the analysis process. Pitch is the fundamental frequency of the speech segment. In voiced segments the sample values seem to repeat itself over a particular period, this is called the Pitch Period. However it is a difficult undertaking to extract the pitch from a short noisy segment, which may contain both voiced and unvoiced components.

Several algorithms make use of the fact that the autocorrelation of a periodic function $R_{xx}(k)$ will have a maximum when k is equal to the pitch period. Coupled with the fact that the estimation of the autocorrelation function generally leads to a smoothing out of the noise, this makes the autocorrelation function a useful tool for obtaining the pitch period. However, Voiced speech is not exactly periodic, which makes the maximum lower than we would expect from a periodic signal. Generally, a maximum is detected by checking the autocorrelation value against a threshold; if the value is greater than the threshold, a maximum is declared to have occurred. The problems with this approach are the difficulty to select a value, when there is uncertainty about the magnitude of maximum value. Another problem occurs because of the interference due to other resonances in the vocal tract.

Another algorithm for pitch estimation uses the Average Magnitude Difference Function (AMDF). The AMDF is defined as

$$AMDF(P) = \frac{1}{N} \sum_{i=k_o+1}^{k_o+N} |y_i - y_{i-P}| \tag{3.2}$$

If a sequence y_n is periodic with period P_o, the samples that are P_o apart in the y_n sequence will have values close to each other, and therefore the AMDF will have a minimum at P_o.(See Figure 3.4)

Not only do we have a minimum when P equals the pitch period, but any spurious minimums we may obtain in the unvoiced segments are very shallow; that is the difference between the minimum and average values is quite small. Therefore the AMDF can serve a dual purpose: It can be used to identify the pitch period as well as the voicing condition. AMDF is faster than autocorrelation method if a DSP chip is not used. However, autocorrelation is a better option if implemented in DSP chip. The pitch estimation assumes that human pitch lies between 20 and 60 samples(2.5 and 19.5 milliseconds)for 8000 samples/sec.

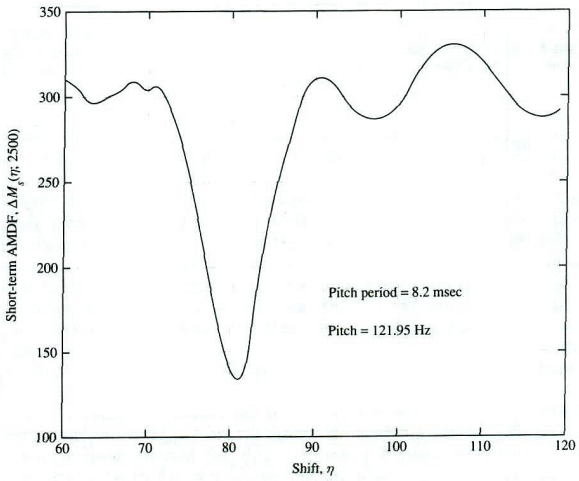

Figure 3.4: AMDF

Obtaining Vocal tract Filter parameters

Vocal tract filter parameters are obtained using LPC algorithms which are usually common in all vocoders. Hence they are explained in Appendix B.

Transmitting the Parameters

The parameters that need to be transmitted include the voicing decision, the pitch period and the vocal tract filter parameters. One bit suffices to transmit the voicing information. The pitch is quantized to 1 of 60 different values using a log-companded quantizer. The LPC-10 algorithm uses a 10th order filter for voiced speech and a 4th order filter for unvoiced speech. Thus, we have send 11 values (10 reflection coefficients and 1 gain) for voiced and 5 values for unvoiced speech. We need to use non uniform quantization for k1 and k2 because vocal tract filter is especially sensitive for errors in reflection coefficients that have magnitude close to one. The nonuniform quantization is implemented by first generating the coefficients (log area ratios)

$$g_i = \frac{1 + k_i}{1 - k_i} \tag{3.3}$$

which are then quantized using a 5 bit quantizer. The coefficients k3 and k4 are

both quantized using 5 bit uniform quantizer. In the voiced segments, coefficients k5 through k8 are quantized using a 4-bit uniform quantizer, k9 is quantized using 3-bit and k10 using 2-bit uniform quantizer. In unvoiced segments the unused 21 bits are used for error protection. The gain G is obtained by finding the root mean square(rms) value of the segment and quantized using 5-bit log-companded quantization. Including an additional bit for synchronization, we end up with a total of 54 bits per frame. Multiplying this by the total number of frames per second gives us the target rate of 2400 bits per second.

Voice Synthesis at Receiver

At the receiver, the voiced frames are generated by exciting the received vocal tract filter by a locally stored waveform. This waveform is 40 samples long. It is truncated or padded with zeros depending on the pitch period. If the frame is unvoiced, the vocal tract is excited by a pseudorandom number generator.

The LPC coder provides intelligible reproduction at 2.4 kbits. The use of only two kinds of excitation signals gives an artificial quality to the voice. This approach also suffers when used in noisy environments. The encoder can be fooled into declaring segments of speech unvoiced because of background noise. When this happens, the speech information gets lost.

Multipulse Linear Predictive Coding(MP-LPC)

In 1982 B.S.Atal and Remde introduced the idea of MP-LPC in which several pulses were used during each segment. The spacing of these pulses are determined by evaluating a number of different patterns from a codebook of patterns. In MP-LPC a code book of excitation patterns is constructed. Each entry in this codebook is an excitation sequence that consists of a few nonzero values separated by zeros. Given a segment from the speech sequence to be encoded, the encoder obtains the vocal tract filter using the LPC analysis described later. The encoder then excites the vocal tract filter with the entries of the codebook. The difference between the original speech segment and the synthesized speech is fed to a perceptual weighting filter, which weights the error using a perceptual weighting criterion. The codebook entry that generates the minimum average weighted error is declared to be the best match. The index of the best-match entry is sent to the receiver along with the parameters for the vocal tract filter.

Regular Pulse Excited Linear Prediction (RPE-LP)

This was developed in 1986 by Kroon, Deprettere and Sluyter as a modification of MP-LPC algorithm. Instead of using excitation vectors in which the nonzero values are separated by an arbitrary number of zero values, they forced the nonzero values to occur at regular spaced intervals. Furthermore, they allowed the nonzero values to take on a number of different values. They called this scheme regular pulse excitation (RPE) coding. A variation of RPE, called regular pulse excitation with long-term prediction (RPE-LTP) was adopted as a standard for digital cellular telephony by the Group Speciale Mobile (GSM) subcommittee of the European Telecommunications Standards Institute at the rate of 13kbps.

Code Excited Linear Prediction (CELP)

The MP-LPC was improved by B.S.Atal and Schroeder in 1984 with the introduction of the CELP system. In CELP, instead of having a codebook of pulse patterns, we allow a variety of excitation signals. For each segment the encoder finds the excitation vector that generates synthesized speech that best matches the speech segment being encoded. This approach is closed in strict sense to a waveform coding technique such as DPCM than to the analysis/synthesis schemes.

Federal Standard 1016(4.8kbps)

The Vocal tract filter used by the CELP coder in FS 1016 is given by

$$y_n = \sum_{i=1}^{10} b_i y_{n-i} + \beta y_{n-P} + G\epsilon_n \tag{3.4}$$

where P is the pitch period and the term βy_{n-P} is the contribution due to the pitch periodicity. The input speech is sampled at 8000 samples per second and divided into 30-millisecond frames containing 240 samples. Each frame is divided into four sub frames of length 7.5 milliseconds. The coefficients b_i for the 10th-order short-term filter are obtained using the autocorrelation method.

The pitch period P is calculated once every subframe. In order to reduce the computation load, the pitch value is assumed to lie between 20 and 147 every odd subframe. In every even subframe, the pitch value is assumed to lie within 32 samples of the pitch value in the previous frame.

The FS 1016 algorithm uses two codebooks, a stochastic codebook and an adaptive codebook. An excitation sequence is generated for each subframe by adding one scaled element from the stochastic codebook and one scaled element from the adaptive codebook. The scale factors and indices are selected to minimize the perceptual error between the input and synthesized speech.

The stochastic codebook contains 512 entries. These entries are generated using a Gaussian random number generator, the output of which is quantized to -1, 0 or 1. If the input is less that -1.2, it is quantized to -1; if it is greater than 1.2, it is quantized to 1; and if it lies between -1.2 and 1.2 it is quantized to 0. The codebook entries are adjusted so that each entry differs from the preceding entry in only two places. This structure helps reduce the search complexity.

The adaptive codebook consists of the excitation vectors from the previous frame. Each time a new excitation vector is obtained, it is added to the codebook. In this manner, the codebook adapts to local statistics.

The FS 1016 coder has been shown to provide excellent reproductions in both quiet and noisy environments at rates of 4.8 kbps and above. Because of the richness of the excitation signals, the reproduction does not suffer from the problem of sounding artificial. The lack of a voicing decision makes it more robust to background noise. The quality of the reproduction of this coder at 4.8 kbps has been shown to be equivalent to a delta modulator operating at 32 kbps. The price for this quality is much higher complexity and a much longer coder delay.

Low Delay CELP (LD-CELP):CCITT G.728 Speech Standard(16kbps)

The problem of longer delay of CELP was addressed with LD-CELP. By Coding delay we mean the time between when a speech sample is encoded to when it is decoded if the encoder and decoder were connected back-to-back. CCITT approved recommendation G.728, a CELP coder with a coder delay of 2 milliseconds operating at 16kbps. In order to lower the coding delay, the size of each segment has to be reduced significantly because the coding delay will be some multiple of the size of the segment. The G.728 specification recommendation uses a segment size of five samples. With five samples and a rate of 2 bits per sample, we only have 10 bits available to us. Using only 10 bits, it would be impossible to encode the parameters of the vocal tract filter as well as the excitation vector. Therefore, the algorithm obtains he vocal tract filter parameters in a backward adaptive manner; that is the vocal tract filter coefficients to be used to synthesize the current segment are obtained by analyzing the previous decoded segments. The CCITT requirements for G.728 included the requirement that the algorithm operate under noisy channel conditions. It would be extremely difficult to extract the pitch period from speech corrupted by channel errors. Therefore the G.728 algorithm does away with the pitch filter. Instead, the algorithm uses a 50th-order vocal tract filter. The order of the filter is large enough to model the pitch of most female speakers. Not being able to use pitch information for male speakers does not cause much degradation. The vocal tract filter is updated every fourth frame, which is once every 20 samples or 2.5 milliseconds. The autocorrelation method is used to obtain the vocal tract parameters.

As the vocal tract filter is completely determined in a backward adaptive manner, we have all 10 bits available to encode the excitation sequence. Ten bits would be able to index 1024 excitation sequences. However, to examine 1024 excitation sequences every 0.625 milliseconds is a rather large computational load.

In order to reduce this load, the G.728 algorithm uses a product codebook where each excitation sequence is represented by a normalized sequence and gain term. The final excitation sequence is a product of the normalized sequence and the gain. Of the 10 bits, 3 bits are used to encode the gain using a predictive encoding scheme, while the remaining 7 bits form the index to a codebook containing 127 sequences.

The low-delay CCITT G.728 CELP coder operating at 16 kbps provides reconstructed speech quality superior to the 32 kbps CCITT G.726 ADPCM algorithm.

3.2.3 Hybrid Coders

Hybrid coders as the name suggests, tries to incorporate the features of both waveform coders and vocoders. They can be classified as time domain and frequency domain coders. Frequency domain hybrid coders divide the signal into frequency bands and each band is quantized independently based on perceptual noise shaping. Examples of frequency domain hybrid coders are subband coding, transform coding and multi band excited vocoder (MBEV).Time domain hybrid coders are usually LPC based. Examples are Adaptive Predictive Coding (APC), Residual Excited Predictive Coding (RELP) and analysis-by-synthesis coders like Multi Pulse LPC (MPLPC), GSM coder (RPE-LTP) and Code Excited LPC (CELP).

3.3 Summary

In this chapter various Voice coding strategies were described. The different strategies are Waveform coder (which tries to reproduce the waveform correctly), Vocoder (which is based on Source modelling technique), and Hybrid coder (which incorporates the features of both waveform coder and vocoder). For more details about low bitrate speech coding refer [11], [3] and [8]

Chapter 4

Proposed Algorithm: Pulse Position Linear Prediction(PPLP)

4.1 Introduction

The Excitation signal used to excite the Linear Prediction Vocal tract filter is a decisive factor in both voice quality and data bitrate. The major advances made in the recent times in LPC vocoders were due to the invention of new excitation signals and new ways to generate and transmit them. It is a well known fact that the unnaturalness and artificiality of the vocoder outputs are mainly due to the use of poor and static natured excitation signal.

4.2 Previous Works

The first commonly used LPC based vocoder was LPC 10, which borrowed the concept of periodic impulses and noise excitation from channel vocoders used during WWII. Even though we were able to draw good intelligibility, this approach never gave good voice quality. Also it failed terribly during noisy environments. This was attributed mainly to the failure of the complex Pitch Detection and Voicing Decision algorithms. Again, the concept of one pulse per pitch period was proven wrong. The excitation for voiced speech should consist of several pulses in a pitch period rather than just one at the beginning of the period. Also there can be mixed frames which are partially voiced and partially unvoiced.

An approach to solve this was to try introducing multiple pulses per pitch period. A similar approach placing double and triple pulses per pitch period can be seen in [1]. An approach toward solving the problem of mixed frames was to determine pulse positions of weak periodicity in unvoiced frames. This approach is described in [9]. An approach taken to remove buzzy effect and to synthesize more natural sounding speech was to add baseband residual also as part of the excitation signal. This approach as seen in [10] makes the system robust to V/UV and pitch prediction errors.

Another approach is to find the pulse positions irrespective of voicing and pitch decisions. These approaches usually fall into the hybrid coders as they try to accurately reproduce waveform too. Examples of these algorithms are MPELP, CELP etc. These vocoders are highly robust against errors due to noise. They produce high quality sound at the cost of high codec delays.

4.3 Approach in this Thesis

PPLP algorithm uses an excitation signal that consists of pulses of amplitude equal to the root mean square energy of the frame (gain of frame). The way the position of pulses are determined is one of the important aspects of this proposed system. Many of the latest algorithms uses the idea of multiple pulses instead of the idea of periodic pulse. But the algorithms to choose the pulse position and amplitudes are often very complex like those used in MPELP and CELP. Also the need to send the amplitude of the pulses increases the bandwidth, making them a mid rate codec.

In this vocoder the pulse positions were chosen from the peak and valley positions in the residue obtained by the inverse filtering of the original speech. The idea is that in voiced speech the residue is essentially a set of steep pulses and valleys especially in the voiced region. The easiest way to represent it will be to

put impulses of the same value at their position. This idea will naturally bring randomness of values in the unvoiced region too. Thus this idea removes the need for detecting pitch and making voicing decision, by automatically achieving the desired results.

After further experiments it was discovered that the accurate amplitude of the pulses were not very much important for intelligibility of speech. Thus a low bit rate codec was formed by pulse positioning. By using latest vector quantization techniques or backward prediction we can achieve a very low bit rate voice coder which gives reasonably good quality this way. The outputs of experiments done showed acceptable quality, intelligibility and naturalness for a variety of sound samples. The Signal obtained by inverse filtering has got very little correlation between samples. Pulses at peaks and valleys give us a low correlation excitation signal having pitch & voicing information. The excitation signal also sounded similar to the residual signal and showed similarity in the spectrogram. Thus we are able to produce a good quality excitation signal using a less complex algorithm which requires low bit rate.

Accurate separation of speech into two classes, voiced and unvoiced, is difficult to achieve in practice. As we have seen earlier, there are more than two modes in which the vocal tract is excited and often these modes are mixed. Even when the speech waveform is clearly periodic, it is a gross simplification to assume that there is only one point of excitation in the entire pitch period. There is some evidence that, apart from the main excitation which occurs at glottal closure, there is secondary excitation, not only at glottal opening and during the open phase, but also after the closure. These results suggest that the excitation for voiced speech should consist of several pulses in a pitch period rather than just one at the beginning of the period.

The main difficulty so far in using a multi-pulse model has been our inability to develop a satisfactory procedure for determining these pulses. In this Algorithm an easy method for determination of pulse positions is mentioned. It was seen that this model is flexible enough to provide acceptable quality speech even at low bit rates. No a priori assumption about the nature of the excitation signal is made. The excitation consists of a sequence of pulses for all speech classes - including voiced and unvoiced speech. No attempt is made to make it periodic or non-periodic. But the objective is attained in the most natural way automatically.

4.4 Analysis

LPC analysis is done in the normal way itself. Levinson Durbin algorithm is used for this purpose. Refer Appendix B. But in the C program we used Lattice method in place of the normal autocorrelation method, because of the increased stability of the IIR filters generated by Lattice method. However, we sticked to the Levinson Durbin method in MATLAB. Before analysis the voice signal was passed through a Pre-emphasis Filter. Pre-emphasis is necessary because, in the spectrum of a human speech signal, the energy in the signal decreases as the frequency increases. Pre-emphasis increases the energy in parts of the signal by an amount inversely proportional to its frequency. Thus, as the frequency increases, pre-emphasis raises the energy of the speech signal by an increasing amount. This process therefore serves to flatten the signal so that the resulting spectrum consists of formants of similar heights. (Formants are the highly visible resonances or peaks in the spectrum of the speech signal, where most of the energy is concentrated.) The flatter spectrum allows the LPC analysis to more accurately model the speech segment. Without pre-emphasis, the linear prediction would incorrectly focus on the lower-frequency components of speech, losing important information

about certain sounds. After the pre-emphasis process the autocorrelation of the signal is calculated. The autocorrelation values are then used to find out the LPC coefficients. Ten LPC coefficients are calculated.

4.4.1 Gain Computation

Gain is the square root of the energy of the error signal. So it can be obtained as square root of the error energy E.

4.4.2 Pulse Position Analysis

The exact pulse position is very important in the steady-state voiced areas of speech where a small offset in the pulse position destroys the periodicity of the overall signal. But the exact pulse position is not perceptually important otherwise.

In voiced speech, the linear prediction residual signal has a pulse-like structure due to the slope discontinuity of the glottal pulse. A concept of peak as used in this paper is a positive valued position where the current value is greater than the previous value and the next value is lesser than or equal to the current value. The valley is a negative valued position where the current value is less than the previous value and the next value is greater than or equal to the current value. Pulses are located at positive peaks and negative valleys of the residue signal. Only peaks and valleys which have a minimum threshold value and have the corresponding signs (peaks are positive and valleys are negative) are chosen. The threshold was chosen such that it allows selection of only the major peaks and valleys. Otherwise it can result in a small static background noise and loss of natural voice quality. Since the average energy of signal varies from frame to frame, it is essential to vary threshold from frame to frame. So the threshold was chosen as a multiple of the mean of the displacement of signal samples from zero. The threshold level for qualifying for a pulse position can be controlled to experiment with their effect

Code	Pulse Position
0	No pulse
1	1st position
2	2nd position
3	3rd position
4	4th position
5	5th position
6	6th position
7	7th position

Table 4.1: Pulse Positions and corresponding bit vector codes

on the output quality. The peak and valley pulses are alternatively positioned starting with a positive peak pulse. The peak and valley positions are located in a serial fashion. Each pulse is to be represented with a triangular sawtooth pulse in the excitation signal, in order to reduce buzziness in the output. Each pulse is located apart by a minimum of 6 pulse positions.

Pulse position bitvector Quantization

All the pulses should be separated by at least 6 positions. So if we take any 7 consecutive locations there can be either a single pulse or no pulse in those locations. Therefore each 7 consecutive locations can be represented by 8 codes (3 bits) give in table above. Also, if the code of the current 7 positions is n then the next 7 position code will surely be in the set {0,n+1....7}. Due to this reason the number of possible series of these symbols is highly reduced. Therefore if we represent a series of these codes with an index from a codebook we can achieve more compression. Number of various possible series of codes and corresponding compression and codebook sizes are:

- If 7 codes taken together, Total No. of Possible Series=51284, Codebook-size=1076964 bits (1Mb), Compression ratio = 1.3125, Corresponding Codec bitrate=4kbps

Parameter	No. of bits
k1 and k2	8
k3 and k4	8
k5 to k10	18
Gain	5
Synchronization bit	1
Total	40 bits

Table 4.2: Bit allocation for LPC parameters per frame

- If 6 codes taken together, Total No. of Possible Series=11908, Codebook-size=214344(210Kb), Compression ratio = 1.2857, Correspond Codec bitrate=4.14kbps
- If 5 codes taken together, Total No. of Possible Series=2766, Codebook-size=41490(¡41Kb), Compression ratio = 1.25, Correspond Codec bitrate=4.2kbps

4.4.3 Quantization

The first two LPC coefficients were quantized after finding the log area ratios of the values. Alternatively sine inverse can also be used. This provides a non linear quantization. Rest is directly quantized with 5 to 3 bits. Thus resulting in 34 bits per frame for transmitting LPC coefficients. Extra 5 bits are needed for transmitting gain. 1 bit is used for synchronizing.

- log area ratios of k1 and k2 are 4-bit quantized
- k3 and k4 4-bit uniform quantized
- Other Reflection Coefficients are 3 bit encoded
- Gain is 5-bit uniformly encoded
- 1 bit for synchronization

Parameter	Allocation	No. of bits
LPC parameters	40*33.33	1334
Pulse Position Code	$\frac{(8000*\frac{3}{7})}{1.3125}$	2613
Total	3947 bps	$\approx$ 4kbps

Table 4.3: Bit allocation per second

4.5 Synthesis

A time varying filter is maintained by feeding it with the reflection coefficient received from the sender. The excitation signal is used to drive this LPC filter to produce voice. Let us see how the excitation signal is generated.

4.5.1 Excitation Signal Generation

Excitation signal is generated by using the bit vector and the gain. Firstly these values are decompressed. Then a signal is formed with triangular saw tooth pulses at the pulse positions mentioned in the bit vector. We put positive and negative saw tooth pulses alternatively starting with a positive pulse. We can experiment with various pulse shapes, as the shape is important in reducing buzziness. Related work can be seen in [7]. This pulse sequence will have the pitch and voicing information inbuilt in it. Then to give the correct energy level to the output, the excitation signal is multiplied by the gain value. The gain is actually an envelope (See figure 4.1)which coarsely reshapes the excitation signal to nearly the shape of the residue. The resulting pulse sequence is seen to posses spectrogram similar to that of the residue. It also sounds very much similar to the residue.

4.5.2 LPC Synthesis

LPC synthesis is done by feeding the excitation signal into the time varying IIR synthesis filter prepared by feeding the corresponding LPC reflection coefficients. The LPC synthesis filter works on the basis of the Levinson Durbin algo-

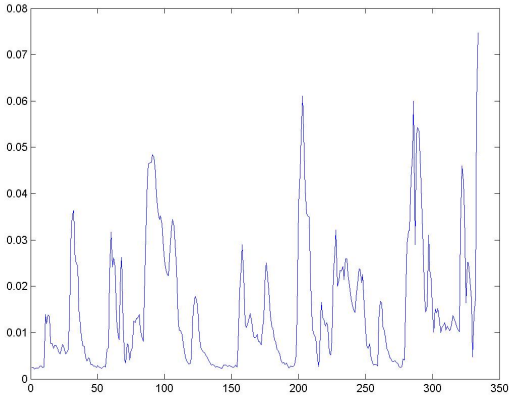

Figure 4.1: Gain Envelope

rithm explained earlier and the output is obtained. The output obtained showed considerable similarity in waveform and spectrogram. However, the waveform was only partially similar. The difference was visible mainly at the higher resolution. But the output sounded natural and maintained the voice quality to a certain extent. Intelligibility and robustness of the coder was particularly noteworthy. It worked well even in highly noisy environments. It even reproduced the music partially and especially the vocal in certain music files, though the output sometimes was noisy.

4.6 Summary

In this chapter the new Excitation signal and the method of coding and generating that signal is explained. Also this chapter describes the working of a Vocoder based on the new system.

Chapter 5

Computer Simulation and Discussion

5.1 Introduction

The Algorithm was coded in Visual C++ using MFC for user interface. The user can open any 8bit, 8000 samples per second wave file in the program. The program will encode the sound, decode it, synthesis the speech and play the output. Thus the system consisted of the coder and decoder connected back to back. Again this system was simulated in MATLAB 6.5 Simulink. The model diagram is provided towards the end of the report. Small modifications were made in the model figure given there for better clarity. The model was used to study the frequency and waveform related influences of various coding decisions made on the coder output. Input, output and their corresponding spectrograms are provided at the end of this report. On closer examination of the spectrogram we can see that the Vioced area (vertical striations (slight ridges)) and formants (horizontal bands) were reproduced well in the output. The output waveform also shows considerable similarity with the input waveform.

5.2 Performance Issues

Performance of a speech coder can be evaluated on the following issues: speech quality, coding rate, complexity of the algorithm and delay. In practice, coding strategies are usually the trade-off among them.

5.2.1 Voice Quality

The measures for evaluating speech quality can be divided into two categories: objective and subjective. Objective measures tell us how well the exact waveform is reproduced. Objective measures are often sensitive to both gain variations and delays. These measures are applicable for waveform codecs. Example of Objective tests is SNR (Signal to Noise Ratio). This is the ratio of the energy of the pure Signal to the energy of the Noise introduced by the codec in decibel. But this test is not suited for Vocoders as they do not reproduce the exact waveform. But the tests needed for these codecs should test how much the output is perceptually equal to the original sound. It should tests the factors like intelligibility and voice Quality. So we use Subjective testing. Subjective test procedures such as the Diagnostic Rhyme Test (DRT), the Diagnostic Acceptability Measure (DAM), and the Mean Opinion Score (MOS) are based on listener ratings. The DRT is an intelligibility measure where the subject's task is to recognize one of two possible words in a set of rhyming pairs (e.g. meat, heat). The rhyme test was originally introduced by Fairbanks speech and the modified rhyme test was developed later by House. The DAM scores are based on results of test methods evaluating the quality of a communication system based on the acceptability of speech as perceived by a trained normative listener.

The MOS is a measure which is widely used to quantify coded speech quality. The MOS usually involves 12 to 24 listeners (formal CCITT and TIA tests typically involve 32 to 64 listeners) who are instructed to rate phonetically balanced records according to a 5 level quality scale (Bad, Poor, Fair, Good, Excellent). Excellent speech quality implies that coded speech is indistinguishable from the original and without perceptible noise. On the other hand, bad (unacceptable) quality implies the presence of extremely annoying noise and artifacts in the coded speech. In MOS tests listeners are "calibrated" in the sense that they

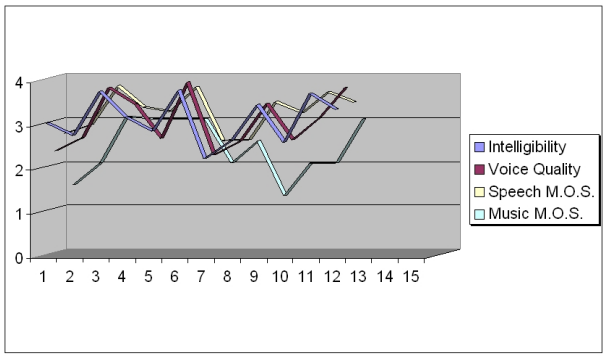

Figure 5.1: M.O.S

are familiarized with the listening NR conditions and the range of speech quality they will encounter. Ratings are obtained by averaging numerical scores over several hundreds of speech records. The MOS range relates to speech quality as follows: a MOS of 4 to 4.5 implies network quality, scores between 3.5 and 4 imply communications quality, and a MOS between 2.5 and 3.5 implies synthetic quality. It is to be noted that MOS ratings may differ significantly from test to test and hence they are not absolute measures for the comparison of different coders.

The codec was tested mainly with informal listening tests. Mean Opinion Score was collected based on the input from 12 listeners. An informal rhyming test was also done. PPLP algorithm seemed to do well in the informal listening tests, M.O.S. tests and the informal rhyming tests.

Mean Opinion Score:

A Mean Opinion Score test was conducted on an experimental basis. The test environment had a background noise of 30 to 33 decibels. It was conducted with 20 different samples. Samples were chosen carefully. Samples were in English, Malayalam and Hindi. Some of the listeners and files were not having intelligibility even in the original files. Such listeners were to be removed either fully or partially

from the rating. Unintelligible files were removed from the tests. Listeners as expected gave a wide range of Intelligibility and Voice Quality rating. The scores ranged from 2.6 to 3.8. Listeners chosen were of age group 20 to 35. A Mean Opinion Score of 3.31 was obtained. This is near to the M.O.S. of CELP which is 3.2. Even though M.O.S. cannot be used for comparison purposes, it gives an idea that the codec provides good enough intelligibility and voice quality. The Listeners were also asked to give separate score for Intelligibility. The coder received M.O.S. scores of 3.22 and 3.15 for intelligibility and Voice quality respectively. The greatest surprise came with an M.O.S. score of 2.38 for music. The score is given for the vocal reproduction mainly, although the background music reproduction was also given a small consideration. Even though the tests were carried out in a constrained and noisy environment, the results were very much acceptable. The voice quality and intelligibility loss can be mainly attributed to the clicking sounds generated by the coder. These sounds are said to be produced in RELP also, as was noticed in literature survey. Clicking sounds are increasingly observed when multiple speakers attempt to talk together. Clicking sounds and static noise were reduced considerably by using a moving average smoother on the generated excitation signal before synthesis. The moving average smoother also flattened the spectrum of the excitation signal. The coder with smoother got M.O.S. score near 4 compared to 3.31 for the one without it. The M.O.S scores are always subject to human error and the audibility and intelligibility of the listeners. But the overall results of M.O.S. tests were in favor of the codec.

5.2.2 Bitrate

The proposed codec has a codec bitrate of 4kbps, which makes it a low bitrate codec. It gives a good compression ratio of 16 times, which is more than that of CELP with a bitrate of 4.8kbps.

5.2.3 Algorithm Complexity

All the vocoders are based on Linear Prediction. Therefore the difference mainly comes in the choice of excitation signal and quantization. The proposed algorithm uses a very simple serial algorithm of the order of O(n) for excitation signal generation. Other similar codecs uses algorithms of higher complexities. The quantization method used also are very simple and does not consume much time. So the codec does very well with respect to algorithmic complexity.

5.2.4 Coding Delay

Since the algorithms used are of lower complexity, the codec delay is very less. It takes a coding delay of 9ms per frame which is much less than other complex algorithms.

5.3 Summary

In this chapter we have explained the way in which the simulation and studies were conducted. It describes the work done and elaborates on the results of the simulation.

Chapter 6

Comparison

6.1 PPLP

PPLP algorithm do not need calculation of pitch or voicing decision. Thus it escapes the possible degradation due to wrong calculation of pitch and voicing. It also does not need to run complex algorithm needed in pitch detection and voicing decision. Since the algorithm does not classify speech frames into voiced/unvoiced, it works well with transition frames and mixed frames. The pulse positions are entirely based on peak and valley positions. Therefore it automatically gets the correct and useful pitch and voicing information, which is visible in the output spectrogram. Also the algorithm takes much less execution time. Its excitation signal generation algorithm is $O(n)$ while other codecs use $O(n^2)$ and $O(n^3)$ algorithms. The difference is visible in the execution time measurements made during experiments. An LPC 10e implementation in MATLAB took 900 milliseconds for analysis and synthesis of 1 second of a sample file. RELP took 220 milliseconds for analysis and synthesis of same data. PPLP took only 300 milliseconds. Usually for a segment containing 20 milliseconds worth of data, the coding delay would be approximately somewhere between 40 to 60 milliseconds. This kind of delay is not acceptable for most realtime applications. RELP took 7.2 milliseconds per frame (each frame is 30 milliseconds, ie 240 samples) with out compression. It took 15 milliseconds per frame with compression. PPLP coder has an average coder delay

of 9 milliseconds per frame operating at 4kbps. LD-CELP as per a literature has a coder delay of 2 milliseconds. The pulse position algorithm, which is the core of excitation generation, took on an average around 5 to 14 ms for 1 second voice sample on a 2GHz PC.

6.2 LPC 10

LPC 10 assumes that the excitation signal for the vocal tract filter consists of two, voiced and unvoiced sources, and synthesize the excitation. Although this type of vocoder can compress speech with high intelligibility at a bitrate as low as 2.4kbps, it fails to produce a high quality of speech even before the quantization of the relevant parameters. The limited performance is mainly due to its simple excitation source model, especially for the transition segments such as onsets and irregular glottal pulses in voiced speech. In steady-state voiced speech, misclassification of a voiced frame as unvoiced one causes two kinds of distortion. First, the decoder fails to generate the pitch pulses in the misclassified frame and the overall periodicity is destroyed, which results in an artifact in the synthesized speech. The worse distortion is caused by the large amount of noise that corresponds to the total energy of the pitch pulses. Transition segments usually show weak periodicity, and they are often classified as unvoiced. In this case the distortion generated by decoder is the major source of distortion because there is little periodicity to be destroyed. On the contrary, when it is classified as a voiced frame and the excitation is synthesized as usual periodic voiced sounds, we found that there is no or only slight degradation of perceived quality as long as the estimated pitch is not extremely high or low. The synthesis method based on estimated pitch is not robust as the pitch is not defined and the estimated pitch is usually random in transition segments. Algorithmic complexity of pitch detection and voicing decision algorithms are high, resulting in high codec delays.

Algorithm	Bitrate (bits/sec)	M.O.S.	MIPS
PCM (G.711)	64k	4.3	0.01
ADPCM (G.721)	32k	4.1	2
LD-CELP (G.728)	16k	4.0	19
RPE-LTP (GSM)	13k	3.47	6
Skyphone-MPLP	9.6k	3.4	11
VSELP (IS-54)	8k	3.45	13.5
CELP (FS1016)	4.8k	3.2	16
STC-1	4.8k	3.52	13
IMBE	4.15k	3.4	3
STC-2	2.4k	2.9	13
LPC-10e	2.4k	2.3	7
PPLP (this paper)	4k	3.3	-

Table 6.1: Comparison of various standard Vocoders and PPLP

Comparison of various standard Vocoders and PPLP

6.3 MPLP

In the MPELP coder, the frame that shows a low correlation but a high peakiness value in the residual signal is set to the zittery voiced mode where the aperiodic pulses are synthesized with a maximum of 25% random offsets from their periodic positions. The excitation signal synthesized by the aperiodic pulses produces a natural quality of speech in transition segment, although it differs from the original one in the actual pulse positions and the number of pulses. This implies that whether the excitation is pulse-like or noise-like is perceptually important. However, the analysis/synthesis method of the MPELP coder for the transition segment is not directly related to the actual pulses in the original residual signal. Although strong pulses yield a high peakiness value, a high peakiness value does not necessarily indicate the existence of pulses. In addition, the synthesis method based on the estimated pitch is not robust, because the pitch is not defined and the estimated pitch is usually random in transition segments. Another major problem with MPLP is the algorithmic complexity. This results in very high codec delays.

This codec also requires high bitrate of 16kbps.

6.4 CELP

CELP method requires calculation of pitch and all the problems associated with the pitch extraction algorithms used in LPC-10 are present here also. Along with this are the problems caused by the use of highly complex algorithms. This results in high codec delay which makes it less attractive for real time applications

6.5 LD-CELP

LD-CELP is a low complexity version of CELP. It does away with pitch extraction and uses very small frames for reducing codec delay. It is very much preferred for real time applications due to its very low codec delay. But the negative side is its high bitrate. This is a fast mid rate codec.

6.6 Summary

In this chapter we compare the new algorithm with other similar algorithms on the basis of performance, voice quality and intelligibility.

Chapter 7

Conclusion and Future Scope

7.1 Conclusion

One of the most important and fast growing need in the present day is the need for efficient communication. It is very important in shrinking the earth into a global electronic village. The most important factor for making modern communication means accessible to common man is to provide better quality at low costs. A fast and low complexity voice codec can result in low cost VoIP devices and thus achieve this goal. The algorithm suggested in this paper provides acceptable speech quality and intelligibility at low complexity and low bit rate. The quality of the coder was tested with informal listening tests and Mean Opinion Scores.

7.2 Future Scope

Future scope for this thesis work are :

- Using the pulse periodicity and Vector Quantization to further compress the pulse position bit vector
- Using algorithms similar to those used in MPELP to find the best amplitudes for the pulses.

- Using backward linear prediction to remove need for sending LPC parameters or Vector Quantizing LPC parameters.

Bibliography

[1] Pearson W.A Alsaka Y.A. A new lpc excitation signal. In Southeastcon '88, Tokyo, Japan, 1988. IEEE Conference Proceedings.

[2] Bishnu. S. Atal and Joel R. Remde. A new model of lpc excitation for producing natural-sounding speech at low bit rates. In Acoustics, Speech and Signal Processing, volume 7, pages 614 – 617, Bell Laboratories, Murray Hill, NJ, USA, May 1968. IEEE International Conference on ICASSP '82.

[3] Sadaoki Furui and M. Mohan Sondhi. Advances in Speech Signal Processing. Marcel Dekker, Inc, 1992.

[4] John H. L. Hansen John R. Deller Jr. and John G. Proakis. Discrete-Time Processing of Speech Signals: An IEEE Press Classic Reissue. IEEE PRESS, 1993.

[5] J. G. Proakis and D. G. Manolakis. Digital Signal Processing: Principles, Algorithms,and Applications. Prentice-Hall, 1996.

[6] Atal B. S. and M. R. Schroeder. Predictive coding of speech signals. In Y. Konasi ed., Report of 6th International Congress on Acoustics, Tokyo, Japan, 1968. International Congress of Acoustics.

[7] Rabiner L. Sambur M., Rosenberg A. and McGonegal C. On reducing the buzz in lpc synthesis. In Acoustics, Speech, and Signal Processing, volume 2, pages 401 – 404, a, May 1977. IEEE International Conference on ICASSP '77.

[8] Khalid Sayood. Introduction to Data Compression Second Edition. Morgan Kaufmann publishers, 2000.

[9] Jongseo Sohn and Wonyong Sung. A low resolution pulse position coding method for improved excitation modeling of speech transition. In Acoustics, Speech and Signal Processing, volume 1, pages 265 – 268, Bell Laboratories, Murray Hill, NJ, USA, March 1999. IEEE International Conference on ICASSP '99. Proceedings.

[10] Chong Un and Won Sung. A 4800 bps lpc vocoder with improved excitation. In Acoustics, Speech and Signal Processing, volume 5, pages 142 – 145. IEEE International Conference on ICASSP '80., April 1980.

[11] Ming Yang. Low bit rate speech coding. In Potentials, volume 23, pages 32 – 36, a, Oct-Nov 2004. IEEE.

Appendix A

Simulation Results

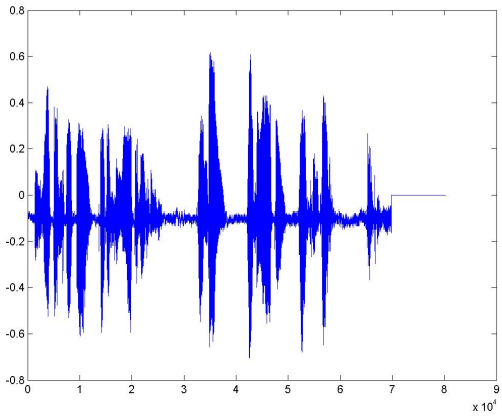

Figure A.1: Input Speech Signal

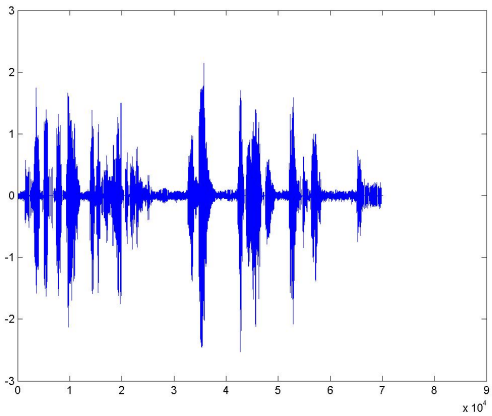

Figure A.2: Output Speech synthesized by PPLP system

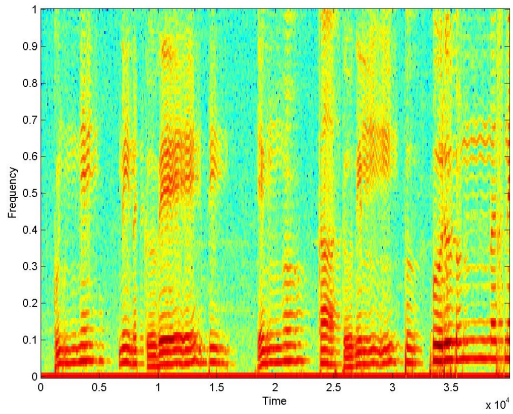

Figure A.3: Input Speech Spectrogram

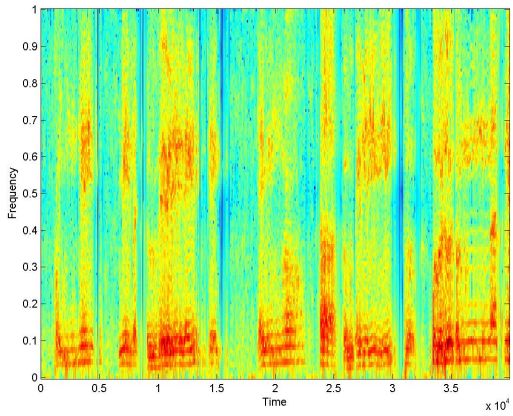

Figure A.4: Output Speech Spectrogram

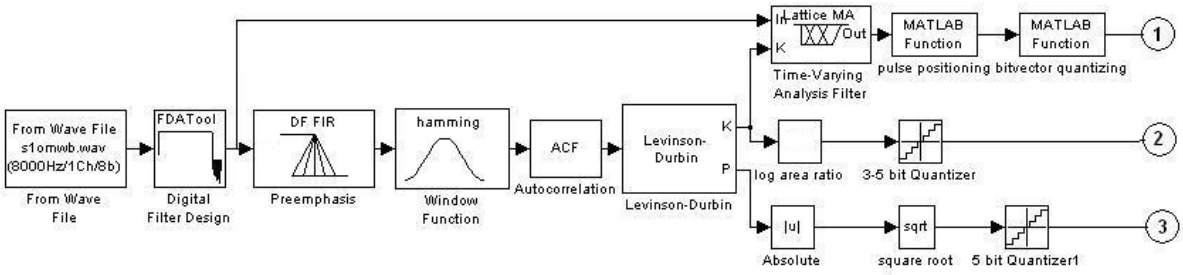

Figure A.5: PPLP System: Analysis

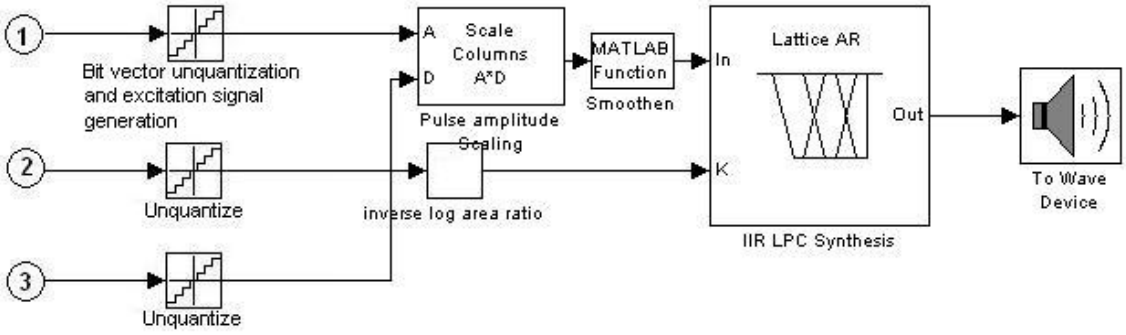

Figure A.6: PPLP System: Synthesis

Appendix B

Linear Predictive Coding

B.1 Introduction

Linear predictive coding is the heart of almost all the voice compression algorithms. Linear predictive coding (LPC) is a popular technique for speech compression, analysis and synthesis. It is one of the most important technology for the modern speech technology, for many of the practical and theoretical results in speech processing will be impossible without it. Speech analysis and synthesis with Linear Predictive Coding (LPC) exploit the predictable nature of speech signals. Cross-correlation, autocorrelation and autocovariance provide the mathematical tools to determine this predictability.

B.2 Correlation Coefficients

Correlation, a measure of similarity between two signals, is frequently used

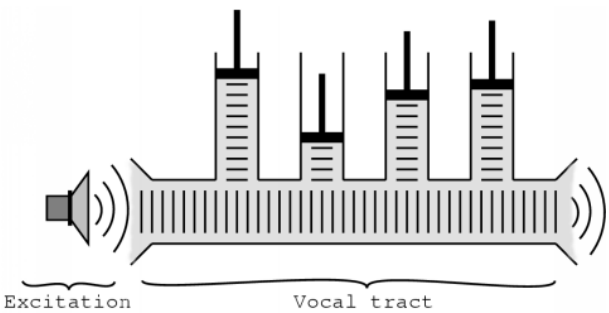

Figure B.1: Vocal tract model

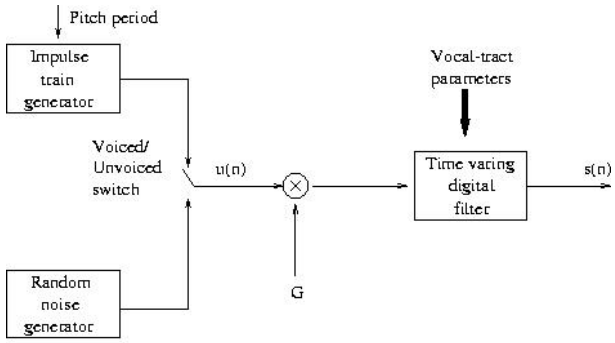

Figure B.2: Modelling of Speech Production

in the analysis of speech and other signals. The cross-correlation between two discrete-time signals x [n] and y [n] is defined as

$$r_{xy}[l] = \sum_{n=-\infty}^{\infty} (x[n]y[n-l]) \tag{B.1}$$

where n is the sample index and l is the lag or the time shift between the two signals. Since speech signals are not stationary, we are typically interested in the similarities between the signals only over a short time duration ($< 30ms$). So the cross-correlation is computed only over a window of time samples and for only a few time delays l=0,1,...,P. The autocorrelation sequence $r_{ss}[l]$, describes the redundancy in the signal s[n]

$$r_{ss}[l] = (\frac{1}{N} \sum_{n=0}^{N-1} (s[n]s[n-l])) \tag{B.2}$$

where s[n] n=-P,(-P)+1,...,N-1 are the known samples and the $\frac{1}{N}$ is a normalizing factor. Another related method of measuring the redundancy in a signal is to compute its auto-covariance

$$r_{ss}[l] = (\frac{1}{N-1} \sum_{n=l}^{N-1} (s[n]s[n-l])) \tag{B.3}$$

where the summation is over $N-l$ products (the samples {s[-P],...,s[-1]} are ignored).

B.3 Linear Prediction Model

Linear prediction is a good tool for analysis of speech signals. Linear

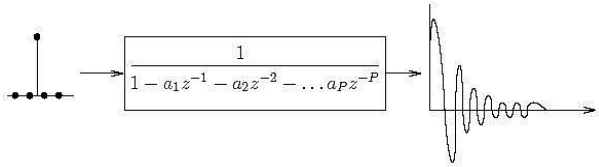

Figure B.3: Linear Prediction (IIR) model of speech

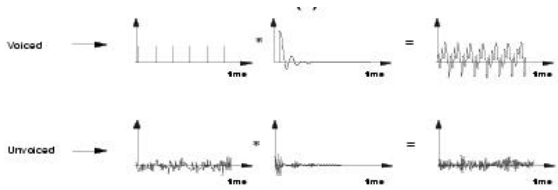

Figure B.4: Graphical representation of Voiced and Unvoiced Speech Production

prediction models the human vocal tract as an infinite impulse response (IIR) system that produces the speech signal. For vowel sounds and other voiced regions of speech, which have a resonant structure and high degree of similarity over time shifts that are multiples of their pitch period, this modelling produces an efficient representation of the sound. Figure B.3 shows how the resonant structure of a vowel could be captured by an IIR system. Figure B.4 shows how the output speech is synthesized from periodic pulse and noise excitation signal.

The linear prediction problem can be stated as finding the coefficients a_k which result in the best prediction (which minimizes mean-squared prediction error) of the speech sample s[n] in terms of the past samples s[n-k], k= {1,...,P}. The predicted sample $\hat{s}[n]$ is given by equation (B.4)

$$\hat{s}[n] = \sum_{k=1}^{P}(a_k s[n-k]) \tag{B.4}$$

where P is the number of past samples of s[n] which we wish to examine and $\hat{s}[n]$ is the predicted signal. The mean-squared prediction error $\tilde{e}^2(n)$ is given by

$$\mathcal{L}\{\tilde{e}^2(n)|a(0)=1\} = \mathcal{L}\{[\sum_{i=0}^{M} a[i]s[n-i] - e[n]]^2|a(0)=1\} \tag{B.5}$$

In order to find the values of $a(\eta)$ for which mean-squared prediction error is minimum, we differentiate the right side with respect to $a(\eta) = 1, 2, ..., M$ and set the result to zero

$$2\mathcal{L}\{[\sum_{i=0}^{M} a(i)s(n-i) - e(n)]s(n-\eta)|a(0)=1\} = 0 \tag{B.6}$$

or

$$r_s(n) + \sum_{i=1}^{M} a(i)rs(\eta - i) - r_{es}(\eta) = 0, \quad \eta = 1, 2, ..., M \tag{B.7}$$

where $r_{es} = \mathcal{L}\{e(n)s(n-\eta)\}$ denotes the temporal cross-correlation of the sequences e(n) and s(n), and $r_s = \mathcal{L}\{s(n)s(n-\eta)\}$, the temporal autocorrelation of s(n). Now, if we assume $\underline{e}$ to be a unity variance orthogonal random process,

meaning $r_e(n) = (L)\{e(n)e(n-\eta)\} = \delta(\eta)$. $r_{es}(\eta) = \acute{\Theta}_o\delta(\eta)$, $\eta \geq 0$, which is therefore zero for all positive η. Finally we have

$$\sum_{i=1}^{M} a(i)r_s(\eta - i) = -r_s(\eta), \quad \eta = 1, 2, ..., M \tag{B.8}$$

and as $\hat{a}(i) = -a(i)$ for i=1,2,...,M, we have

$$\sum_{i=1}^{M} \hat{a}(i)r_s(\eta - i) = r_s(\eta), \quad \eta = 1, 2, ..., M \tag{B.9}$$

These M equations are used to compute the conventional LP parameters where M is the order of the LP system. These equations are sometimes called the normal equations. This equation can be packed in the vector-matrix equation form as

$$\begin{pmatrix} r_s(0) & r_s(1) & ... & r_s(M-1) \\ r_s(1) & r_s(0) & ... & r_s(M-2) \\ r_s(2) & r_s(1) & ... & r_s(M-3) \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ r_s(M-1) & r_s(M-2) & ... & r_s(0) \end{pmatrix} \times \begin{pmatrix} \hat{a}(1) \\ \hat{a}(2) \\ \hat{a}(3) \\ . \\ . \\ . \\ \hat{a}(M) \end{pmatrix} = \begin{pmatrix} r_s(1) \\ r_s(2) \\ r_s(3) \\ . \\ . \\ . \\ r_s(M) \end{pmatrix}$$

Linear Prediction Coefficients are obtained by solving these linear equations.

B.4 Algorithm Used for LPC

The algorithm used for LPC analysis was Levinson-Durbin algorithm. For the C program Lattice method for LPC analysis is used, which provides increased stability for IIR synthesis filter. Other options are Shur's Polynomial and the Cholesky Decomposition using covariance method.

B.4.1 Levinson-Durbin Algorithm

In 1947, Levinson published an algorithm for solving the problem **Ax=b** in which **A** is a **Toeplitz**, symmetric, and positive definite, and b is arbitrary. Autocorrelation equations are exactly of this form, with b having a special relationship to the elements of A. In 1960, Durbin published a slightly more efficient algorithm for this special case. This algorithm is often referred to as **Levinson-Durbin** (L-D) **recursion** algorithm. An alternative is Lattice AR method which can produce more stable filters.

The L-D recursion is a recursive solution for the auto-correlation equations. It means the solution for the desired order-M model is successively built up from the lower-order models, beginning with the "0th order predictor," which is no predictor at all.

Following is the **Levinson-Durbin recursion algorithm** applied to the window n $\epsilon[m-N+1, m]$.

Initialization: l=0

ξ^0(m) = scaled total energy in the "error" from an "order 0" predictor,

. = average energy in the speech frame $(n; m) = s(n)w(m-n)$

. $= r_s(0; m)$

Recursion: for **l**=1,2,...,**M**,

(1) Compute the **l**th reflection coefficient,

$$k(l;m) = \frac{1}{\xi^{l-1}(m)}\{r_s(l;m) - \sum_{i=1}^{l-1}\hat{a}^{l-1}(i;m)r_s(l-i;m)\}.$$

(2) Generate the order-l set of LP parameters,
$\hat{a}(l;m) = k(l;m)$
$\hat{a}(i;m) = \hat{a}^{l-1}(i;m) - k(l;m)\hat{a}^{l-1}(l-i;m), \quad i = 1, ..., l-1.$

(3) Compute the error energy associated with the order-l solution,
$\xi^l(m) = \xi^{l-1}(m)\{1 - [k(l;m)]^2\}$.

(4) Return to Step 1 with l replaced by $l+1$ if $l < M$.

LPC Coefficient Representation

LPC Coefficients have to be tolerant to transmission errors. Transmission of the filter coefficients directly is undesirable, since they are very sensitive to errors. In other words, a very small error can distort the whole spectrum, or worse, a small error might make the prediction filter unstable. There are advanced representations such as log area ratios (LAR), Inverse Sine, line spectrum pairs (LSP) decomposition and reflection coefficients

Log Area Ratios: The ratio of the areas, A_i, of the sections in the lossless tube may be directly computed from the reflection coefficients:

$$\frac{A_{i+1}}{A_i} = \frac{1-k_i}{1+k_i} \tag{B.10}$$

Hence, assuming constant areas at the glottis the total cross section of the vocal tract could be calculated:

- The method is subject to all the LP assumptions.
- better normalization methods need to be employed so as to get realistic cross sections.
- with further fiddling this methods can be useful - for example as an aid to the deaf.

Alternatively this provides the reflection coefficients if we knew the areas.

Inverse Sine: In this method we take the sine inverse of the coefficient before quantizing. Sine inverse gives more variation when the value comes near one. So the coefficients are error tolerant.

Reflection Coefficients: Even if there is a little error in the reflection coefficient the filter constructed using the LPC coefficients obtained from the reflection coefficient will always be stable.

B.5 Summary

In this chapter Linear Predictive Coding Theory and some essential practical information related to use of LPC in speech technology are explained. Also common algorithms used for LPC, like the Levinson Durbin algorithm etc are explained.