بسم الله الرحمن الرحيم

Sudan University of Science and Technology
College of Graduate Studies

Enhancement of Linear Predictive Coding by using Residual Signal

A thesis submitted to the College of Graduate Studies in partial fulfilment of the requirements for the degree of Master of Science in Electronic Engineering (Communications).

Nisreen Abdelrahim Suliman

Supervisor:

Dr. Jacqueline John George

December 2015

بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ

Abstract

Speech coding is an important aspect of modern telecommunications. The primary objective is to represent the speech signal with the fewest number of bits, and a sufficient level of quality. Most of the low bit rate speech coders employ linear predictive coding (LPC), which models the short-term spectral information as an all-pole filter. The LP coefficients are obtained from standard linear prediction analysis as a function of the input samples.

The problem with LPC is that it suffers from many limitations, although it provides intelligible reproduction at low bit-rate. However, only two kinds of excitation signals are used, which gives an artificial quality to the synthetic speech. The performance is further degraded in noisy environments declaring a frame as unvoiced even though it is voiced.

The aim of this thesis is to provide an enhanced version of linear prediction coding by making use of the residual signal. The research focuses on developing an algorithm that exploits the residual to reconstruct the signal. The LP analysis has been simulated; and its performance has been compared by changing the parameters (LP order, frame length, pitch estimation method). The obtained simulation results demonstrate that speech quality can be improved via exploiting the residual signal.

تجريدة

يعتبر ترميز الكلام جانبا هاما من جوانب الاتصالات الحديثة . الهدف الأساسي هو تمثيل إشارة الكلام بواسطة أقل عدد من الخانات الرقمية، و مستوى كاف من الجودة. معظم طرق ترميز الكلام ذات السرعات الرقمية الصغيرة تستعمل الترميز التوقعي. في هذه الحالة يمكن استخدام مرشح (كلي الاقطاب) لتمثيل المعلومات الطيفية على المدى القصير. يتم الحصول على معاملات المرشح بوصفها دالة رياضية تعتمد علي عينات الإدخال.

يعاني الترميز التوقعي من الكثير من القيود ، على الرغم من أنه يوفر فعالية واضحة عند معدل خانات رقمية منخفض. احدي الاسباب أنه يتم استخدام نوعين فقط من إشارات الإثارة ، مما يعطي جودة الاصطناعية إلى اشارة الكلام الاصطناعية, كذلك يتأثر الأداء في البيئات الصاخبة.

الهدف من هذا البحث هو تقديم نسخة محسنة من الترميز التوقعي, و ذلك من خلال الاستفادة من إلاشارة المتبقية. ويركز البحث على تطوير خوارزمية تستغل إلاشارة المتبقية لإعادة بناء الإشارة الاصلية . تمت محاكاة و تحليل اداء الترميز التوقعي وتمت مقارنة أدائها عن طريق دراسة المتغيرات التالية: معامل التوقع، طول الفريم ، طريقة إستخراج التردد. تبين نتائج المحاكاة التي تم الحصول عليها أن نوعية الكلام يمكن تحسينها عن طريق استغلال إشارة المتبقية.

Dedication

To the bright memory of my beloved brother

Mohammed

Acknowledgement

First and foremost, praises and thanks to Allah, the Almighty, for His showers of blessings throughout my research work to complete the research successfully.

Second, this project would not have been possible without the support of many people. I would like to express my deep gratitude for my supervisor Dr. Jacqueline J. George, for her massive support, reading my numerous revisions and helping me make some sense of the confusion. Her knowledge, patience and valuable advice did much to bring this work to a successful conclusion.

Also, special thanks go to Dr. Abuagla Babikir, the head of the school of electronic engineering, Sudan University of Science and Technology, for his great support and encouragement.

Finally, thanks to my family and friends who endured this long process with me, and have always been offering support, love and warm prayers.

Table of Contents

List of Figures

List of Tables

List of Mathematical Notations

$\tilde{(\cdot)}$	Indicates a quantized parameter
$\hat{(\cdot)}$	Indicates an estimated parameter
$\lvert \cdot \rvert$	Magnitude operator
$\lfloor \cdot \rfloor$	The floor function ,which returns the greatest integer less than or equal to the operand)
$\log(\cdot)$	Natural logarithm
$\mathcal{F}\{\cdot\}$	Fourier Transform
$\mathcal{F}^{-1}\{\cdot\}$	Inverse Fourier Transform
f_s	Sampling frequency
N	Number of samples or frame length
N_i	Number of incremental samples
T	Total time in seconds
T_i	Incremental time in seconds
L	Order of prediction filter
c	Speed of sound (340 m/s)
$x[n]$	Speech sample at time instant n
$\tilde{x}[n]$	Quantized speech sample at time instant n
$\hat{x}[n]$	estimated speech sample at time instant n
k_i	The i-th reflection coefficient
$\{k_i\}_n$	The set of reflection coefficient at time n
a_i	The i-th linear prediction coefficient
$\{a_i\}_n$	The set of LPCs at time n
$\tilde{a}_i$	The predicted i-th linear prediction coefficient
$e[n]$	Error signal at time instant n
$\overline{E_x}$	Signal energy, estimated from the sequence $\{x[n]\}$

E_e	Error energy, estimated from the sequence $\{e[n]\}$
P_e	Power of prediction error
$\mathrm{PG}[n]$	Prediction gain at time n
$H(\cdot)$	Generic filter transfer function
$G(\cdot)$	Pre-emphasis filter transfer function
$\mathcal{H}(\cdot)$	LPC synthesis filter transfer function
$\mathcal{A}(\cdot)$	LPC inverse filter transfer function
$\mathcal{P}(\cdot)$	Symmetric polynomial
$\mathcal{Q}(\cdot)$	Asymmetric polynomial
$r(\ell)$	the autocorrelation of lag ℓ
$c(a,b)$	covariance as a function of the lags a and b
$r(l,m)$	correlation as function of two lags l and m
d	quefrency index of the cepstrum signal
$cep(d)$	Cepstrum of a signal at index d
lar_i	Log area ratio
$\mathrm{MSF}[\cdot]$	Magnitude sum function
$\mathrm{MDF}[l,m]$	Magnitude difference function of two lags l and m
$\mathcal{N}(\mu,\sigma^2)$	Gaussian distribution with mean μ and variance σ^2

List of Abbreviations

ADC	Analogue to Digital Converter
ADPCM	Adaptive Differential Pulse Coded Modulation
ACF	Auto-Correlation Function
ACV	Auto-Covariance Function
CELP	Code-Excited Linear Prediction
DAC	Digital to Analogue Converter
DFT	Discrete Fourier Transform
DSP	Digital Signal Processing
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
IDFT	Inverse Discrete Fourier Transform
IFFT	Inverse Fast Fourier Transform
IIR	Infinite Impulse Response
LAR	Log Area Ratio
LP	Linear Prediction
LPA	Linear Prediction Analysis
LPC	Linear Prediction Coefficients
LSF	Line Spectral Frequencies
LSP	Line Spectrum Pairs
MELP	Mixed Excitation Linear Prediction
MDF	Magnitude Difference Function
MSF	Magnitude Sum Function
PCM	Pulse Code Modulation
PG	Prediction Gain
PSD	Power Spectral Density

SNR	Signal-to-noise ratio
VSELP	Vector Selectable Excited Linear Prediction

Chapter 1 : Introduction

1.1 Background

Historically, the speech coding technology has been dominated by coders based on linear prediction, where most speech coding standards depend on waveform approximation. The performance of such methods depends on various parameters such as the bit rate, coding delay and coding complexity. [1]

Mainly, there are two types of speech coders: waveform-following coders and model-base coders. Waveform following coders will exactly reproduce the original speech signal if no quantization errors occur. Model-based coders will never exactly reproduce the original speech signal, regardless of the presence of quantization errors, because they use a parametric model of speech production which involves encoding and transmitting the parameters not the signal. [2]

In particular, linear predictive coding (LPC) is defined as a digital method for encoding an analog signal in which a particular value is predicted by a linear function of the past values of the signal. It was first proposed as a method for encoding human speech by the United States Department of Defense in federal standard 1015, published in 1984. LPC vocoders are considered model-based coders, which means that LPC coding is lossy even if no quantization errors occur. [1] [2]

Furthermore, LP analysis can also be a method to remove the redundancy in the short-term correlation of adjacent samples. For decoding, LP synthesis is used to generate (or reconstruct) a signal, that has spectral

contents close to the original, depending on the LPCs along with the structure of the model. Displacing the frame samples by the LPCs makes the amount of bits required to carry the information lowered, therefore achieving the purpose of compression [2] [3].

1.2 Problem Statement

Linear predictive coding does not recover the entire signal, and faces many challenges, such as:

- The difficulty of accurately classifying a speech frame as voiced/unvoiced.
- The choices of excitation signal are either an impulse train or standard Gaussian noise, which does not always match the real time observations.
- Conventional LPC methods do not preserve all information of the original signal, such as the phase information of the original signal.

Therefore, LPC suffers from the aforementioned limitations and cannot recover the entire signal. Ideally, a vocoder is need to be designed, which satisfies the requirements of performance of vocoders. Such performance measures are minimal bit-rate, low processing delay, and a high output speech quality. Therefore the problem considered in this thesis is to tackle the aforementioned problems by design of an enhanced LPC method.

1.3 Proposed Solution

A valid proposed solution is to use the residual to recover the entire signal. The residual signal following linear prediction analysis contains peaks corresponding to the excitation events in voiced speech together with additional peaks due to the reverberant channel. Henceforth, an enhanced

LPC procedure is designed includes sending the residual from the transmitting side to be utilized by the decoder at the receiving side.

1.4 Methodology

In order to achieve the targeted objectives of this study, simulation software, Matlab, was used. First, a preliminary mathematical background for LPC would be collected, and then LPC speech analysis and synthesis algorithm would be developed and addressed using Matlab.

Furthermore, different methods for pitch estimation were investigated, namely the autocorrelation method and the cepstrcum method.
Finally, the impact of several parameters, such as the impact of prediction order and the frame length were investigated.

1.5 Thesis Outline

The outline of this thesis is as follows.

Chapter 2 provides a literature review about speech signal and speech coding classifications, concepts and attributes.

Chapter 3 focuses on linear prediction coding and the mathematical models associated with it. Principles of linear prediction coding are presented in this chapter beginning with the speech production model, followed by structure of the algorithm.

Chapter 4 illustrates the algorithm development in addition to simulation results and discussion.

Chapter 5 summarizes the work conducted in this thesis, along with further suggestions for future work.

.

Chapter 2 : Literature Review

2.1 Introduction

The speech is an acoustic pressure wave that is produced by the human vocal tract [4], as air is pushed from the lungs through the vocal tract, and vocal organs which is controlled by muscles like vocal chords, jaw, lips and tongue. The vocal tract refers to the pharyngeal and oral cavities grouped together. Therefore, the major factors affecting how speech sounds are the shape of the vocal tract and the excitation signal [2] [3].

The shape and dimensions of the vocal tract are changed continuously over time creating specific natural frequencies called resonant frequencies or formants frequencies and it can be implemented as an acoustic filter with time-varying frequency response. The formants frequencies shape the power spectrum of the speech sound [3].

The excitation signal is the source of energy to excite the resonant qualities of the vocal tract. It contains energy at many frequencies, and the relative strengths of these frequencies are altered as they travel through the vocal tract [2] [3].

The vocal chords vibrate, open and close rapidly during speech production forming pressure pulses near the glottis, which in turn, propagate towards the oral and nasal openings. The speed in which the chords open and close are unique for each individual, it defines the feature and personality of the voice. The time span between a particular point in the opening and closing of the vocal chords to that corresponding point in the next cycle is referred

to as the pitch period, and pitch is the frequency of the quasi-periodic excitation [2] [3] [5].

2.2 Speech Signal Classification

Speech signal are classified depending on how excitation is produced to; voiced sounds, unvoiced sounds and mixed excitation sounds [6] [5] [7].

2.2.1 Voiced sounds

Such sounds include vowel letters and some consonants letters. Voiced sounds are created when the vocal chords vibrate in such a way that the flow of air from the lungs is interrupted periodically (quasi-periodic) creating a sequence of pulses to excite the vocal tract [3]. Voiced sounds are characterized by strong periodicity present in the signal, with the fundamental frequency (pitch frequency). For men, pitch ranges from 50 to 250 Hz, while for women the range usually falls somewhere in the interval of 120 to 500 Hz [3].

2.2.2 Unvoiced sounds

Such sounds include the letters ‘s’ and ‘p’. Unvoiced sounds are pronounced without the aid of the vocal chords; they don't display any type of periodicity and are essentially random (noise-like turbulence) in nature. They are produced when air is forced through a constriction in the vocal tract and then spectrally shaped by passing through the remaining portion of the vocal tract [2] [3].

2.2.3 Mixed excitation sound

An example for such sounds is the letter ‘z’. This sound has a periodic excitation (a phonetic view), so it is considered to be voiced. But, to

represent it in a speech coder, both the periodic and noisy attributes are present [2].

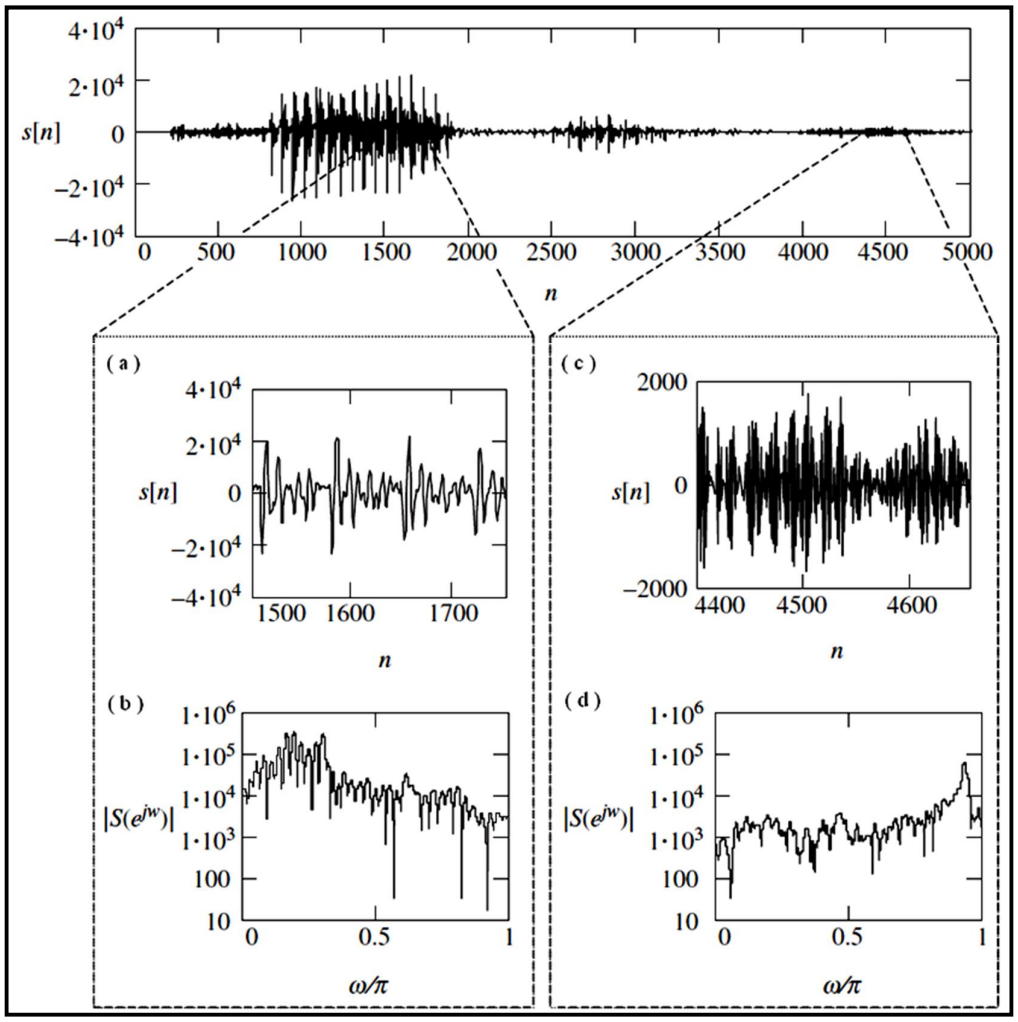

Figure 2.1: Speech Waveform. (a) Voiced frame. (b) Magnitude of the FFT of the voiced frame. (c) Unvoiced frame. (d) Magnitude of the FT of the unvoiced frame. [2]

The sounds is categorized as voiced or unvoiced based on the presence or absence of the periodic excitation, each is produced by different mechanisms at different places in the vocal tract, opening and closing of the vocal chords produces periodic voiced excitation and constriction on steady air flow, after the glottis causes the noisy turbulence of unvoiced excitation. However, many speech sounds have both periodic and noisy

components, since during transitions (voiced to unvoiced or vice versa) there will be randomness and quasi-periodicity that is difficult to judge as strictly voiced or strictly unvoiced; that the classification might not be absolutely clear for all frames [2] [3].

The expanded views of a voiced frame (a) and an unvoiced frame (c) are shown in figure 2.1, with the magnitude of the Fourier transform plotted as (b), (d). The shows an example of speech waveform uttered by a male subject about the word ‘‘problems’’ where both voiced and unvoiced signals are present in 256 samples in length. The non-stationary nature of speech signals can be noticed from the figure, where the signal changes constantly with time. The voiced frame is clearly periodic in time domain, where the signal repeats itself in a quasi-periodic pattern; and also in frequency domain, where a harmonic structure is observed [3].

The spectrum of voiced frame indicates dominant low-frequency contents, due to the low value of the pitch frequency. For the unvoiced frame, the signal is essentially random and there is a significant amount of high-frequency components, corresponding to rapidly changing signals [3].

2.4 Speech Coding

Speech coding is a procedure or method that reduces the amount of information needed to represent speech signals by representing them (or parameters of a speech production model) with few bits as possible and removing inherent redundancy from them; maintaining at the same time a reasonable level of speech quality. This can be a lossy coding scheme because a small amount of perceptible degradation is acceptable.

Speech coding can effectively reduce the storage space and the bit-rate of the speech signal that it is used widely in many applications, such as Real-

time applications (ex: telephone or cellular networks (transmission)) and storage applications, e.g., answering machines, voice mail systems, and multimedia.

In nature the speech signal is a continuous time signal (analogue) because it originates as sound pressure wave that exists at every instant of time. Digital speech signal is required, due to the inherently finite precision arithmetic capabilities of digital systems (especially speech coders) that are easily design and less costly implementation because of the advances in digital logic chips. Digital signals have many advantages like high noise immunity, adjustable precision, ease of developing secure communication systems, better reliability, less need for calibration and maintenance, ease of diagnosis and repair, easy to duplicate similar circuits and easily controllable by computer for all that the digital storage devices, transmission channels and DSP implementations are programmable and easily designed, tested and implemented.

The speech coding involves first converting the speech to a digital form and coding it producing a low-rate bit-stream by speech encoder, then transmits or stores it. At the receiver, the bit stream is decoded, and reconverted back to the original analogue signal by speech decoder. The encoder/decoder structure is known as a speech coder or codec. The term codec is a combination of 'coder-decoder'.

2.4.1 Basic Structure of Speech Coder

The basic structure in converting an analogue speech signal to an encoded digital one and vice versa is shown in figure 2.2.

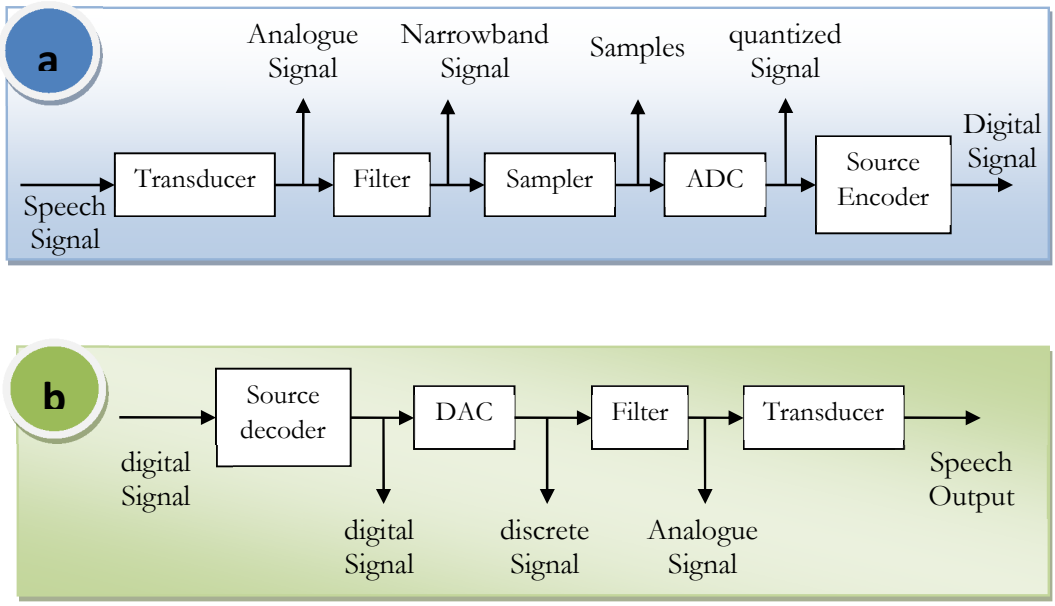

Figure 2.2 : Block Diagram of the Basic Structure for Speech (a) Encoder and (b) Decoder [3]

The input is sound pressure wave that is converted to an electrical speech signal (continuous time analogue signal) by microphone (transducer); then it passes through an anti-aliasing linear filter to eliminate frequency components that have frequencies out of the range between (300 – 3400) Hz, because this band-limited speech in preserves good intelligibility, speaker identity, and naturalness.

After filtering, the sampler receives the narrow-band speech (band-limited analog signal), and converts it to a discrete signal by taking set of amplitude values (samples) at certain time instances from it.

$$x[n] = x(nT), \qquad n = 0,1,2,3, \ldots \tag{2.1}$$

Where:

$x[n]$ Speech sample at time instant n

T Total time in seconds

The analog to digital convertor (ADC) converts the analogue decimal values into digital. The source encoder encodes the digital speech. The

output of the source encoder has substantially lower bit-rate than the input. Different types of speech coder give various bit-rates.

In the speech decoder the signal is passed to the source decoder generating the digital speech signal with the original rate. Then it is converted to continuous-time analogue signal through digital to analogue convertor (DAC) but the original analogue signal can be perfectly reconstructed by passing the output of the DAC through low-pass filter that rolls off to about 35 dB with the cut-off frequency equal to one-half of the sampling rate (4 kHz) in order to eliminate the aliasing artifacts caused by sampling. The speech can be heard when the analogue signal is passed through the speakers (transducer).

In almost all speech coders, the reconstructed signal differs from the original one since each speech coder has its desirable properties (attributes).

2.4.2 Speech Coder Attributes

Speech coder main goal is either to transmit a speech signal with low bit-rate as possible or to store it in a storage device with less memory as possible and in both cases the quality of the speech must be good and clear for the listener.

The appropriate bit-rate at which speech should be transmitted or stored depends on the cost of transmission or storage, the cost of coding (compressing) the digital speech signal, and the speech quality requirements. Speech coder desirable properties include low bit-rate, high speech quality, low coding delay, low computational complexity, robustness across different speakers/languages, robustness in the presence of channel errors, and good performance on non-speech signals.

a) Low Bit-Rate.

The degree of compression that the coder achieves can be measured by how much its bit rate is lowered from 64 Kbps (For telephone bandwidth speech is sampled at 8 KHz and digitized with an 8-bit quantizer, resulting in a bit rate of 64 Kbps). Lower bit-rate of the encoded bit-stream, requires less bandwidth for transmission, leading to a more efficient system. This requirement is in constant conflict with other good properties of the system, such as speech quality. Speech coders need not have a constant bit rate [3] [8].

b) High Speech Quality.

In almost all speech coders, the reconstructed signal differs from the original one. The decoded speech should have a quality acceptable for the target application, and can be determined by how the speech sounds to a listener. There are many dimensions in quality perception, including intelligibility, naturalness, pleasantness, and speaker recognisability. Speech coder quality ratings are determined by means of subjective listening tests; the absolute category rating (ACR) test is most often used. Quality can usually be improved by increasing bit rate or complexity, and sometimes by increasing delay [3] [8].

c) Low Coding Delay

One of the good attributes of a speech coder is measured by its coding delay. Delay is introduced in the process of speech encoding and decoding but excessive delay creates problems with real-time two-way conversations. The delay of the coder is more important for transmission than for storage applications. In large communication; delays of 300 ms or greater are particularly objectionable to users even if there are no echoes [3] [8].

Coding delay is given by the elapsed time from the instant a speech sample arrives at the encoder input to the instant when the same speech sample appears at the decoder output shown is figure 2.3, this definition does not consider exterior factors (such as communication distance or equipment) which are not controllable by the algorithm designer. Most low bit rate speech coders encode a block of speech, also known as a frame. The coding delay can be given by the sum of the next four major components shown in figure 2.4.

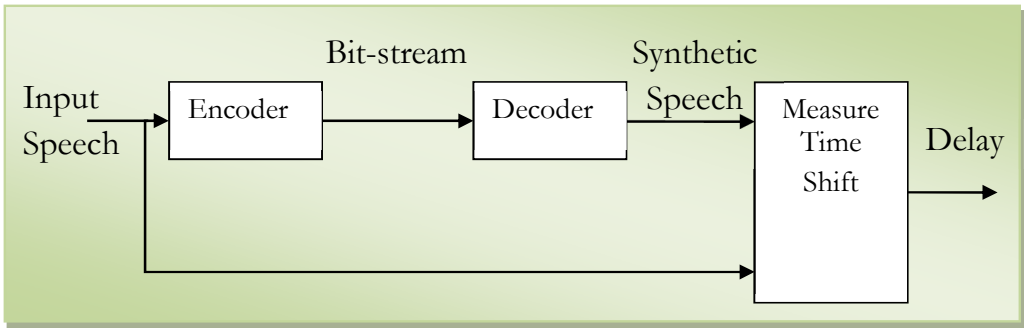

Figure 2.3: Delay Measurement System [2]

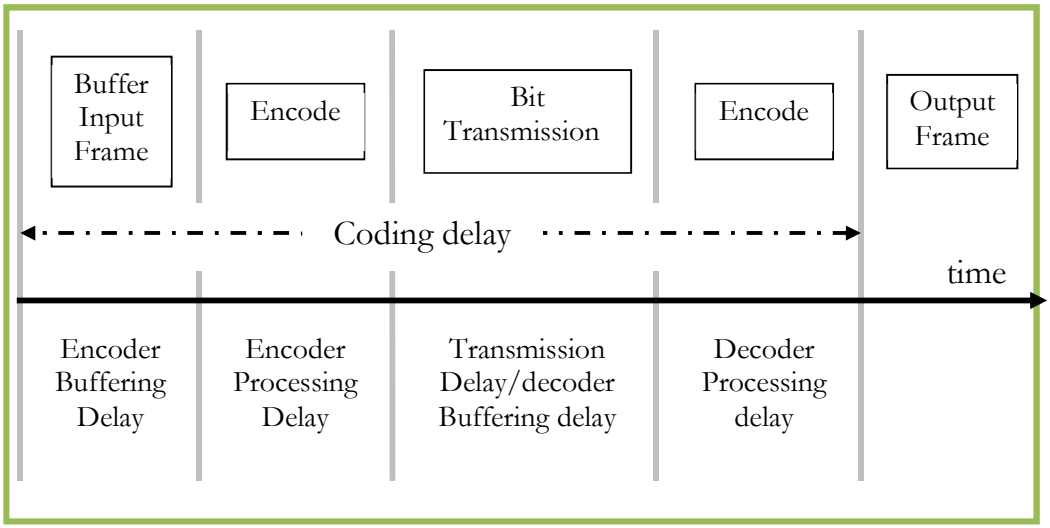

Figure 2.4: Illustration of the Components of Coding Delay[2]

1) Encoder Buffering Delay

   Many speech encoders require the collection of a certain number of samples before processing. The sum of frame size and other inherent delays call algorithmic delay [3] [8].

2) Encoder Processing Delay

   It is the amount of time required to process the buffered data to construct the bit-stream, it must be shorter than the buffering delay; otherwise the encoder will not be able to handle data from the next frame. It is dependent on the speed of the processor used [3] [8].

3) Transmission Delay

   Once the encoder finishes processing one frame of input samples, the resultant bits representing the compressed bit-stream are transmitted to the decoder with one of the transmission modes, this transmission delay is equal to encoder buffering delay in case of constant mode transmission or shorter than it in case of burst mode transmission. Transmission delay is also known as decoder buffering delay, since it is the amount of time that the decoder must wait in order to collect all bits related to a particular frame so as to start the decoding process [3].

4) Decoder Processing Delay

   This is the time required to decode the buffered bit stream to produce one frame of synthetic speech, its upper limit is given by the encoder buffering delay [3].

   In general, the encoder buffering delay has the greatest impact because it determines the upper limit for the rest of the delay components. Most low bit-rate coders often have high delay because long encoding buffer enables a more thorough evaluation of the signal properties, leading to higher coding efficiency and hence lower bit-rate. Thus, coding delay in most cases is a trade-off with respect to the achievable bit-rate. In practice, a reasonable estimate of the coding delay is to take 2.5 to 3

and 1.5 to 2.5 times the frame interval (encoder buffering delay) for constant mode transmission and burst mode transmission, respectively [3].

d) Low Computational Complexity

The degree of complexity is a determining factor in both the power consumption of a speech coder and the cost associated with its implementation that prefer to be low; these include the amount of memory needed to support its operation, as well as computational demand. Cost is always a factor in the selection of a speech coder for a given application.

e) Robustness across different speakers / languages

The technique used in the speech coder should be general enough to model different speakers (adult male, adult female, and children) and different languages adequately, since each voice signal has its unique characteristics.

f) Robustness in the presence of channel errors

This is crucial for digital communication systems where channel errors will have a negative impact on speech quality [3].

g) Good Performance on non-speech signals

In a typical telecommunication system, other signals might be present besides speech i.e., telephone signalling. Even though low bit-rate speech coders might not be able to reproduce all signals faithfully, it should not generate annoying artifacts when facing these alternate signals [3].

These attributes are pre-determined while trade-offs can be made among the others according to the application.

2.4.3 Classification of Speech Coders

It is common to classify the speech coders according to the bit-rate of the encoded bit-stream as shown in table 2.1. A given coder works fine at a certain bit-rate range, but the quality of the decoded speech will drop if it is decreased below a certain threshold. The minimum bit-rate that speech coders will achieve is limited by the information content of the speech signal [3].

Table 2.1: Classifications of speech coders according to bit-rate **[3]**

Category	Bit-rate range
High bit-rate	> 15 Kbps
Medium bit-rate	5 to 15 Kbps
Low bit-rate	2 to 5 Kbps
Very low bit-rate	< 2 Kbps

2.4.4 Speech Coder Types

The speech encoders are classified according to the coding technique used in the coder. Different coding techniques lead to different bit-rates, but all speech encoders are designed to reduce the reference bit-rate of 64 Kbps toward lower values. Classifying speech coders by coding techniques subdivides the speech coders into three types: waveform coders, model-based coders and hybrid coders.

2.4.4.1 Waveform Coders

Waveform coders (ex: pulse code modulation (PCM) and adaptive differential PCM ADPCM), attempt to preserve the original shape of the

signal waveform; if there were no quantization error, the original speech signal would be exactly reproduced. These coders are better suited for high bit-rate coding (practically at a bit-rate of 32 Kbps and higher), since performance drops sharply with decreasing bit-rate, their quality can be measured using signal-to-noise ratio (SNR) [3] [8].

Waveform coders encode the shape of the time-domain waveform. Basic waveform coding approaches often do not exploit the constraints imposed by the human vocal tract on the speech waveform. As such, waveform coders represent non-speech sounds (music, background noise) accurately, but do so at a higher bit rate than that achieved by efficient speech-specific encoders. There are various waveform coding approaches such as PCM, Non-uniform Pulse Code Modulation and differential waveform coding.

2.4.4.2 Model-based coders (or parametric coders)

They are based on parametric models of speech production. In the decoder the speech signal is generated from a model, which is controlled by some parameters whose values are estimated from the input speech signal during encoding and transmitted as the encoded bit-stream (only the values of the parameters are quantized); if there were no quantization error, the reproduced signal would not be the original speech. There are several proposed models, but the most successful is based on linear prediction, e.g., LPC and mixed excitation linear prediction (MELP) [6] [9] [10], and works well for low bit-rate. Increasing the bit-rate normally does not translate into better quality, since it is restricted by the chosen model. The quality cannot be measured by SNR because there is no attempt to preserve the original shape of the waveform; hence these coders have poor performance for non-speech signals. Perceptual quality of the decoded speech is directly related to the accuracy and sophistication of the underlying model [3] [8].

2.4.4.3 Hybrid Coders (or parametric coders)

Hybrid encoders, such as code-excited linear prediction (CELP) ([9] [10] [11] [12] [13]), combine the strength of waveform coders with that of parametric coders. This is achieved by depending on a speech production model, whose parameters are located during encoding, but additional parameters of the model are optimized in such a way that the decoded speech is as close as possible to the original waveform. This type works well for medium bit-rate coders, and attempts to match the original signal with the decoded signal in the time domain, by quantizing or representing the excitation signal to the speech production model and transmitted it as a part of the encoded bit-stream not like in parametric coder that achieves low bit-rate by discarding all detail information of the excitation signal; only coarse parameters are extracted. A hybrid coder tends to behave like a waveform coder for high bit-rate, and like a parametric coder at low bit-rate, with fair to good quality for medium bit-rate [3].

Chapter 3 : Linear Prediction Coding and Model

3.1 Introduction

There exist many different types of speech compression that make use of a variety of different techniques. There are many characteristics about speech production that can be exploited by speech coding algorithms. One fact that is often used is that period of silence take up greater than 50% of conversations. Most forms of speech compression are achieved by modelling the process of speech production as a linear digital filter. The digital filter and its slow changing parameters are usually encoded to achieve compression from the speech signal. LPC is one of the methods of compression that models the process of speech production. Specifically, LPC models this process as a linear sum of earlier samples using a digital filter inputting an excitement signal. An alternate explanation is that linear prediction filters attempt to predict future values of the input signal based on past signals.

3.2 Historical Background of LPC

The history of audio and music compression begin in the 1930s with research into PCM. Compression of digital audio was started in the 1960s by telephone companies who were concerned with the cost of transmission bandwidth. LPC origins begin in the 1970s with the development of the first LPC algorithms. In 1984, the United States Department of Defence produced federal standard 1015 which outlined the details of LPC. Extensions of LPC such as CELP and Vector Selectable Excited Linear Predictive (VSELP) ([1] [14] [15] [16]) were developed in the mid-1980s

and used commercially for audio music coding in the later part of that decade. The 1990s have seen improvements in these earlier algorithms and an increase in compression ratios at given audio quality levels.

3.3 Mathematical of LPC Model

In LPC, a particular value of the audio signal is predicted by a linear function of the past values of the signal. Human speech is produced in the vocal tract which can be approximated as a variable diameter tube [17] [18] [19].

The LPC model is based on a mathematical approximation of the vocal tract represented by this tube of a varying diameter. During encoding, LP analysis is applied to each frame individually to characterize the shape of the frame spectrum by computing the LP coefficients (LPCs), which are used to represent the frame in transmission or storage. Most LPC techniques are based on autocorrelation or covariance [20][21][22].

Let $x[n]$ be the speech samples at time instants$n = 0,1,\ldots N-1$. Thus, the objective is to find the values of the coefficients $\{a_i\}$, such that the objective function that is given by [18]

$$\sum_{q=0}^{N-1}\left(\sum_{i=1}^{L} a_i\,\tilde{x}[n+q-i]-x[n+q]\right)^2 \tag{3.1}$$

is minimized, where N denotes the length of the Frame, L is the order of the prediction filter and $\tilde{x}[n]$ denotes the quantized sample. The predicted value is given by

$$\hat{x}[n] = -\sum_{i=1}^{L} \tilde{a}_i\,\hat{x}[n-i] \tag{3.2}$$

3.4 LPC Filter Derivation

LPC is derived as a mathematical approximation to the vocal tract representation as a variable diameter tube [17] [23] [24]. The tube can be modelled as equal length sections of different diameter cylinders, where at the boundaries there will be some reflection of waves and forward propagation as well.

Mathematically, the reflection coefficients signify how much energy is reflected and how much is passed. These reflections cause spectral shaping of the excitation. This spectral shaping acts as a digital filter with the order of the system equal to the number of tube boundaries. The digital filter can be realized with a lattice structure with number of stages equivalent to the number of the tube sections. The reflection coefficients k_i are used as weights in the structure and the flow of the signals suggests the forward and backward wave propagation as shown in figure 3.1 [2].

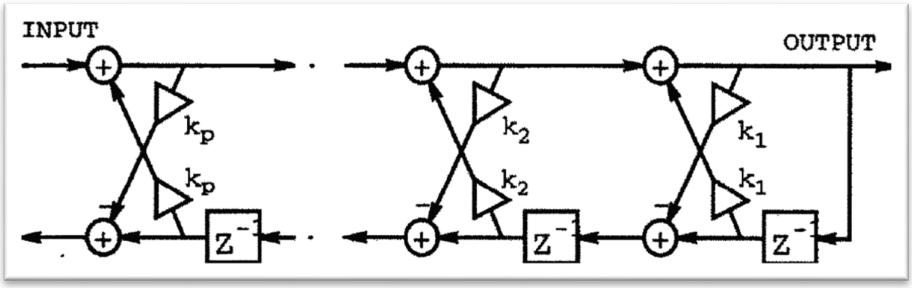

Figure 3.1: Lattice filter realization of multiple-tube model [2]

The white noise is used as an input to the filter instead of the excitation; and the output is the filtered excitation (speech). The time delay for each stage in the concatenated tube model is $\Delta x/c$ where c denotes the speed of sound. The lattice structure can be rearranged into the direct form of the standard all-pole filter model as shown in figure 3.2 [2].

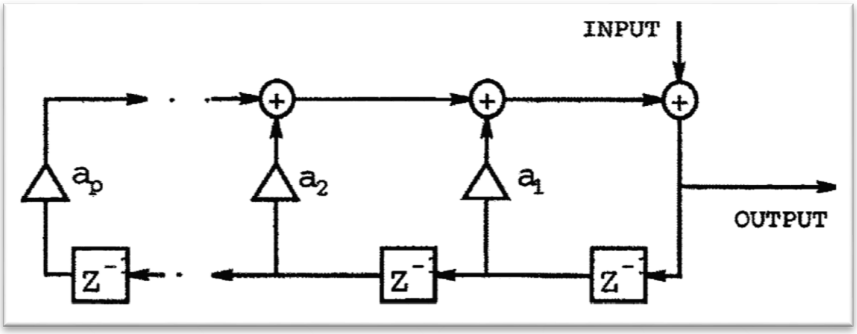

Figure 3.2: Direct form of all-pole filter representing vocal tract [2]

The predictor coefficients, a_i, of the digital filter delay the signal by a single time unit, Z^{-1}, and propagate a portion of the sample value. They are also represent the same information as reflection coefficients (k_i) in LPA, which is based on the all-pole filter. In z-domain notation, the transfer function of the filter is:

$$\mathcal{H}(z) = \frac{1}{\mathcal{A}(z)} \tag{3.3}$$

where

$$\mathcal{A}(z) = 1 - \sum_{i=1}^{L} a_i z^{-i} \tag{3.4}$$

3.5 Estimation of LP Coefficients

The LPCs $\{a_i\}$ are necessary to utilize the LP and must be estimated carefully to provide the closest approximation to the speech samples. The error between a predicted sample $\hat{x}[n]$ and the actual one $x[n]$ is given by

$$\begin{aligned} e[n] &= x[n] - \hat{x}[n] \\ &= x[n] - \sum_{i=1}^{L} a_i\, x[n-i] \end{aligned} \tag{3.5}$$

Hence, the values of $\{a_i\}$ can be computed by minimizing the mean-squared prediction error E_e over the segment, which is given by

$$E_e = \sum_n e^2(n) = \sum_n \left(x[n] - \sum_{i=1}^{p} a_i\, x[n-i] \right)^2 \tag{3.6}$$

To minimize the sum of the squared error, the partial derivatives of E_e with respect to the values of a_i will be set to zero [2] [25], i.e.,

$$\frac{\partial E_e}{\partial a_k} = 2 \sum_n x[n-k] \left(x(n) - \sum_{i=1}^{L} a_i\, x[n-i] \right) = 0, \qquad \text{for } k = 1,2,3, \dots L \tag{3.7}$$

and

$$a_1 \sum_n x[n-k]x[n-1] + a_2 \sum_n x[n-k]x[n-2] + \cdots + a_L \sum_n x[n-k]\, x[n-L] = \sum_n x[n-k]\, x[n] \qquad \text{for } k = 1,2,3, \dots, L \tag{3.8}$$

The autocorrelation and covariance are the methods used to solve these equations with L unknowns [2].

3.5.1 Autocorrelation Method

Given the seqeunce $[n] = \{x[0], x[1], \dots, x[N-1]\}$, the speech samples outside the predetermined boundaries are assumed to be zero. The equations for the LPCs $\{a_i\}$ are ordered in matrix form as

$$\mathbf{R}\,\mathbf{a} = -\mathbf{r}$$

$$\begin{bmatrix} r(0) & r(1) & \dots & r(p-1) \\ r(1) & r(0) & \cdots & r(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(p-1) & r(p-2) & \cdots & r(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} r(1) \\ r(2) \\ \vdots \\ r(p) \end{bmatrix} \tag{3.9}$$

where $r(\ell)$ is the autocorrelation of lag ℓ, which is given by

$$r(\ell) = \sum_{m=0}^{N-1-\ell} x[m]x[m+\ell] \tag{3.10}$$

Equation (3.9) is known as the normal equation, with a Toeplitz matrix. It can be solved by the Levinson-Durbin recursion, which is an algorithm for finding an all-pole IIR filter with a prescribed deterministic autocorrelation sequence. It has applications in filter design, coding, and spectral estimation, and produces the filter with minimum phase. When the error energy is minimized, we get

$$E^{(0)} = r(0) \tag{3.11}$$

$$a_i^{(i)} = k_i = \frac{r(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} r(i-j)}{E^{(i-1)}} \tag{3.12}$$

$$a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)} \quad \text{for } j = 1,2,\dots,i-1 \tag{3.13}$$

$$E^{(i)} = (1 - k_i^2)E^{(i-1)} \tag{3.14}$$

The order in the recursion is displayed between brackets as superscript, where i is the current order in the recursion, $1 \leq i \leq L$, and the i^{th} order coefficient ($a_i^{(i)} = k_i$) is the i^{th} reflection coefficient [2].

3.5.2 Covariance Method

Using covariances, the error minimization formulation results are given by

$$\begin{bmatrix} c(1,1)\, c(1,2) \cdots c(1,p) \\ c(2,1)\, c(2,2) \cdots c(2,p) \\ \vdots \quad\quad \vdots \quad\quad \ddots \quad\quad \vdots \\ c(p,1) c(p,2) \cdots c(p,p) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} c(1,0) \\ c(2,0) \\ \vdots \\ c(p,0) \end{bmatrix} \tag{3.15}$$

where the covariance c is given by

$$c(i,k) = \sum_{m=0}^{N-1} x[m-i]x[m-k] \tag{3.16}$$

Note that equation 3.15 is usually solved by efficient methods such as the Cholesky decomposition [26].

3.5.3 Prediction Gain

Prediction gain is defined as the ratio between the energy of the signal and the energy of the prediction error in dB is

$$\mathrm{PG}[n] = 10 \log_{10} \left(\frac{\sum_{m=n-N+1}^{n} x^2[m]}{\sum_{m=n-N+1}^{n} e^2[m]} \right) \tag{3.17}$$

Voiced frames on average achieve 3 dB or more in prediction gain than unvoiced frames, mainly due to the fact that periodicity implies higher correlation among samples, and thus more predictability. Unvoiced frames, on the other hand, are more random and therefore less predictable. For very low-amplitude frames, prediction gain is normally not calculated to avoid numerical problems; in this case, the frame can be assigned as unvoiced just by verifying the energy level.

3.6 Transformations of LP Parameters for Quantization

LP parameters are transformed by log area ratios or line spectral frequencies (LSFs) [27]. Both two transformations are proven to be useful for coding.

3.6.1 Log Area Ratios

The log area ratios (LARs) reduce the sensitivity to quantization noise when the value of the reflection coefficient is near 1. Let k_i be the i-th reflection coefficient of the filter. Each log area ratios is computed from the reflection coefficients $\{k_i\}$ as:

$$lar_i = \log\frac{1+k_i}{1-k_i} \tag{3.18}$$

The log area ratio can be converted back to the reflection coefficient by the inverse transform follows as:

$$k_i = \frac{1-\exp(lar_i)}{1+\exp(lar_i)} \tag{3.19}$$

3.6.2 Line Spectral Frequencies

The line spectral frequencies (LSFs), or line spectrum pairs (LSPs) [27] are an ordered set of parameters, particularly suited to efficient vector quantization. The LSFs are the roots of the symmetric and asymmetric polynomials $\mathcal{P}(z)$ and $\mathcal{Q}(z)$, which are defined as [17] [23]:

$$\mathcal{P}(z) = \mathcal{A}(z) + z^{-(L+1)}\mathcal{A}(z^{-1}) \tag{3.20}$$

$$\mathcal{Q}(z) = \mathcal{A}(z) - z^{-(L+1)}\mathcal{A}(z^{-1}) \tag{3.21}$$

where $\mathcal{A}(z)$ is the inverse LP filter of equation (3.4). The L roots, or zeros, of $\mathcal{P}(z)$ and $\mathcal{Q}(z)$ lie on the unit circle, in complex conjugate pairs. In addition, one root will be at $+1$, and one at -1. Their angle in the z-plane represents a frequency, and pairs, or groups of three, of these frequencies

are responsible for the formants in the LP spectrum. The bandwidth of the formant, i.e., the degree of sharpness of the formant peak, is determined by how close together the LSFs are for that formant. In practice, the zeros of the polynomials are found by numerical methods [2].

The LP coefficients $\{a_i\}$, can be recovered from the LSFs by multiplying out the terms of the roots of equations 3.20 and 3.21 (the LSFs) to obtain $\mathcal{P}(z)$ and $\mathcal{Q}(z)$. Henceforth, $\mathcal{A}(z)$ can be determined by noting that [1]

$$\mathcal{A}(z) = \frac{1}{2}[\mathcal{P}(z) + \mathcal{Q}(z)] \tag{3.22}$$

LSFs are the favoured format for the LP parameter representation in recent coder implementations due to two desirable properties. Closer LSFs produce a sharper formant peak. This property provides a useful, practical check for stability after the LSFs have been quantized. The LSFs can be checked for a minimum spacing, and separated slightly if necessary. The other property of the LSFs is the localized nature of their spectral impact. If one LSF is adversely altered by the quantization and coding process, that will only degrade the LP spectrum near that LSF frequency. Other representations of the LP information, e.g., reflection coefficients and log area ratios, are not localized in frequency [2].

3.7 Pitch Period Estimation/Encoding

One of the most important parameters in speech analysis, synthesis, and coding applications is the fundamental frequency, or pitch, of voiced speech. Pitch frequency is directly related to the speaker and sets the unique characteristic of a person. Voicing is generated when the airflow from the lungs is periodically interrupted by movements of the vocal cords. The time between successive vocal cord openings is called the fundamental period, or pitch period. For men, the possible pitch frequency range is usually

found somewhere between 50 and 250 Hz, while for women the range usually falls between 120 and 500 Hz. In terms of period, the range for a male is 4 to 20 ms, while for a female it is 2 to 8 ms.

3.7.1 The Autocorrelation Method

To perform the estimation on the signal $x[n]$, with n being the time index. Consider the frame ends at time instant m, where the length of the frame is equal to N . The autocorrelation value, which is given by,

$$r(l,m) = \sum_{n=m-N+1}^{m} x[n]x[n-l] \tag{3.23}$$

reflects the similarity between the frame $x[n]$, $n = m - N + 1$ to m, with respect to the time-shifted version $x[n - l]$, where l is a positive integer representing a time lag. The range of lag is selected so that it covers a wide range of pitch period values. By calculating the autocorrelation values for the entire range of lag, it is possible to find the value of lag associated with the highest autocorrelation representing the pitch period estimate, since, in theory, autocorrelation is maximized when the lag is equal to the pitch period.

It is important to mention that, in practice, the speech signal is often lowpass filtered before being used as input for pitch period estimation. Since the fundamental frequency associated with voicing is located in the low-frequency region (<500 Hz), lowpass filtering eliminates the interfering high-frequency components as well as out-of-band noise, leading to a more accurate estimate.

3.7.2 Magnitude Difference Function

One drawback of the autocorrelation method is the need for multiplication, which is relatively expensive for implementation, especially in those processors with limited functionality. To overcome this problem, the magnitude difference function is invented.

$$\mathrm{MDF}(l,m) = \sum_{n=m-N+1}^{m} x[n] - x[n-l] \tag{3.24}$$

For short segments of voiced speech it is reasonable to expect that $x[n] - x[n-l]$ is small for $l = 0, \pm T, \pm 2T, \ldots$, with T being the signal's period. Thus, by computing the magnitude difference function for the lag range of interest, one can estimate the period by locating the lag value associated with the minimum magnitude difference. Note that no products are needed for the implementation of the present method.

3.7.3 Fractional Pitch Period

The previous two methods can only find integer-valued pitch periods. That is, the resultant period values are multiples of the sampling period of 0.125 ms. In many applications, higher resolution is necessary to achieve good performance. In fact, pitch period of the original continuous-time (before sampling) signal is a real number; thus, integer periods are only approximations introducing errors that might have negative impact on system performance. Multi-rate signal processing techniques, such as interpolation, can be introduced to extend the resolution beyond the limits set by fixed sampling rate. One popular method is the Medan-Yair-Chazan algorithm [28].

3.7.4 Cepstral Pitch Extraction

When a periodic signal with fundamental frequency F_0 consists of many adjacent harmonics (as voiced speech signals do), the corresponding short-term spectrum exhibits a ripple due to its harmonic structure. The cepstrum of this signal will exhibit a strong peak at quefrency d equal to the period duration $1/F_0$. This can be used to determine if a speech segment is voiced or unvoiced and to determine the pitch period, $1/F_0$, if the segment is voiced [29].

The cepstrum is computed from the inverse discrete Fourier transform (IDFT) of the logarithm of the DFT of a given sequence. It is defined as

$$\text{cep}(d) = \mathcal{F}^{-1}\{\log_{10}|\mathcal{F}\{x[n]\}|\} \tag{3.25}$$

where the index d is the quefrency of the cepstrum signal. The quefrency is a type of time domain index. A peak in the cepstrum at quefrency d_0 corresponds to a periodic component in the original signal with period d_0 and frequency $1/d_0$.

The cepstrum extracts pitch information from a voiced speech signal because a voiced signal not only contains dominant spectral components at the fundamental frequency, but also contains harmonics of the pitch fundamental.

A Cepstral analysis of a short time segment of speech will produce a peak at the pitch period for voiced speech, but no prominent peaks for unvoiced speech.

3.8 The LPC Encoder

The components of a conventional LPC encoder are shown in figure 3.3. The input speech is first segmented into a number of non-overlapping frames. A pre-emphasis filter is used to adjust the spectrum of the input

signal. The voicing detector classifies the current frame as voiced or unvoiced and outputs one bit indicating the voicing state. The pre-emphasized signal is used for LP analysis, where ten LPCs are derived. These coefficients are quantized with the indices transmitted as information of the frame. The quantized LPCs are used to build the prediction-error filter, which filters the pre-emphasized speech to obtain the prediction-error signal at its output.

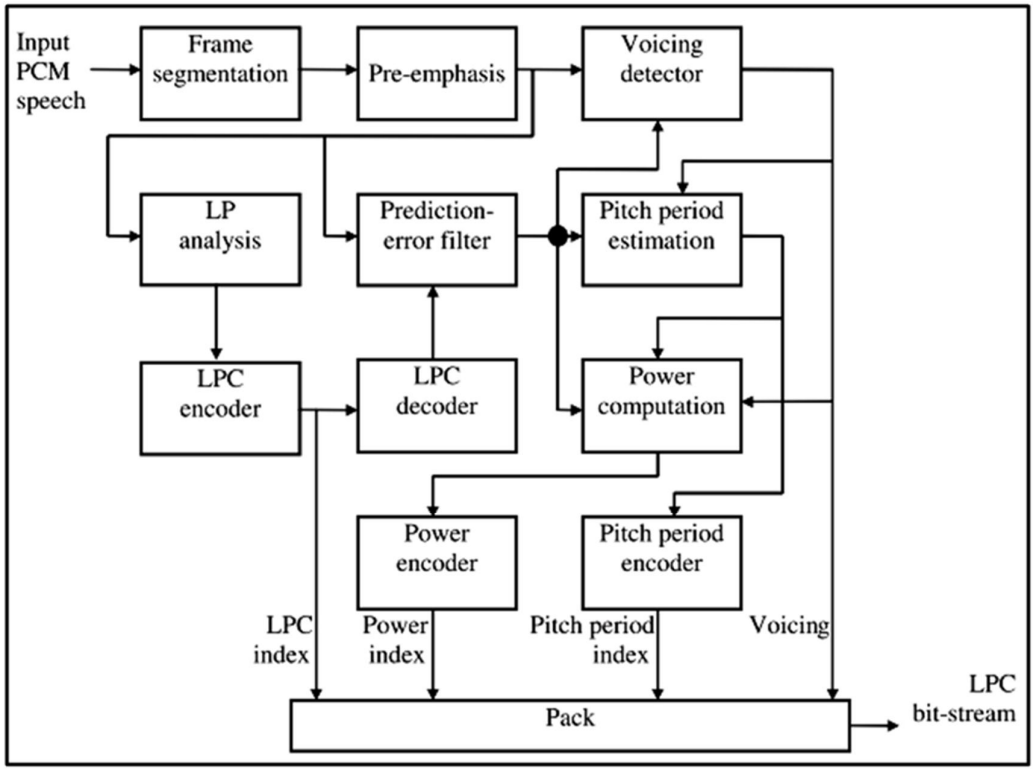

Figure 3.3: The LPC Encoder Block Diagram [2]

3.8.1 Pre-emphasis Filter

The typical spectral envelope of the speech signal has a high frequency roll-off due to radiation effects of the sound from the lips. Hence, high-frequency components have relatively low amplitude, which increases the dynamic range of the speech spectrum. As a result, LP analysis requires high computational precision to capture the features at the high end of the

spectrum. More importantly, when these features are very small, the correlation matrix can become ill-conditioned and even singular, leading to computational problems. One simple solution is to process the speech signal using the filter with system function

$$G(z) = 1 - \alpha Z^{-1} \tag{3.26}$$

which is high-pass in nature and represents the pre-emphasis filter. Denoting $y[n]$ as the input to the filter and $z[n]$ as the output, the value of α satisfies the difference equation $z[n] = y[n] - \alpha y[n-1]$.

3.8.2 Voicing Detector

The purpose of the voicing detector is to classify a given frame as voiced or unvoiced. In many instances, voiced/unvoiced classification can easily be accomplished by observing the waveform: a frame with clear periodicity is designated as voiced, and a frame with noise-like appearance is labelled as unvoiced. In other instances, however, the boundary between voiced and unvoiced is unclear; this happens for transition frames, where the signal goes from voiced to unvoiced or vice versa. The necessity to perform a strict voiced/unvoiced.

Energy is the most obvious and simple indicator of voicedness. Energy level of the output is controlled by the gain parameter. This is the most obvious and simple indicator of *voicedness*. Typically, voiced sounds are several orders of magnitude higher in energy than unvoiced signals. For the frame (of length N) ending at instant n, the energy is given by

$$E_x[n] = \sum_{m=n-N+1}^{n} x^2[m] \tag{3.27}$$

For simplicity, the magnitude sum function defined by [3] [30]:

$$\mathrm{MSF}[n] = \sum_{m=n-N+1}^{n} |x[m]| \qquad (3.28)$$

serves a similar purpose. Since voiced speech has energy concentrated in the low-frequency region, due to the relatively low value of the pitch frequency, better discrimination can be obtained by lowpass filtering the speech signal prior to energy calculation.

3.8.3 LP Analysis

Due to the dynamic nature of a speech signal, the LPCs must be calculated for every signal frame. Within a frame, one set of LPCs is determined and used to represent the signal's properties in that particular interval, with the underlying assumption that the statistics of the signal remain unchanged within the frame. The process of calculating the LPCs from signal data is called linear prediction analysis. This issue was discussed in details in section 3.2, specifically the autocorrelation method.

3.8.4 Prediction Filter Error

Different prediction schemes are used in various applications and are decided by system requirements. The prediction-error is given by

$$e[n] = x[n] - \hat{x}[n] \qquad (3.29)$$

The strategy is to minimize the average squared error $E_e[n]$ within a specific interval. which is to be minimized by the method of partial derivatives explained earlier in section 3.5.

3.8.5 Power Computation

The power of prediction error is computed for both voiced and unvoiced cases. For the unvoiced case, the power of the prediction error is computed as

$$P_e = \frac{1}{N}\sum_{n=0}^{N-1} e^2[n] \tag{3.30}$$

For the voiced case, power is calculated using an integer number of pitch periods:

$$P_e = \frac{1}{\left\lfloor \frac{N}{T} \right\rfloor T}\sum_{n=0}^{\left\lfloor \frac{N}{T} \right\rfloor N-1} e^2[n] \tag{3.31}$$

where $N > T$.

3.9 The LPC Decoder

The process of decoding a sequence of speech segments is the reverse of the encoding process. Each segment is decoded individually and the sequence of reproduced sound segments is joined together to represent the entire input speech signal. The building blocks of the LPC decoder are explained in the following subsections.

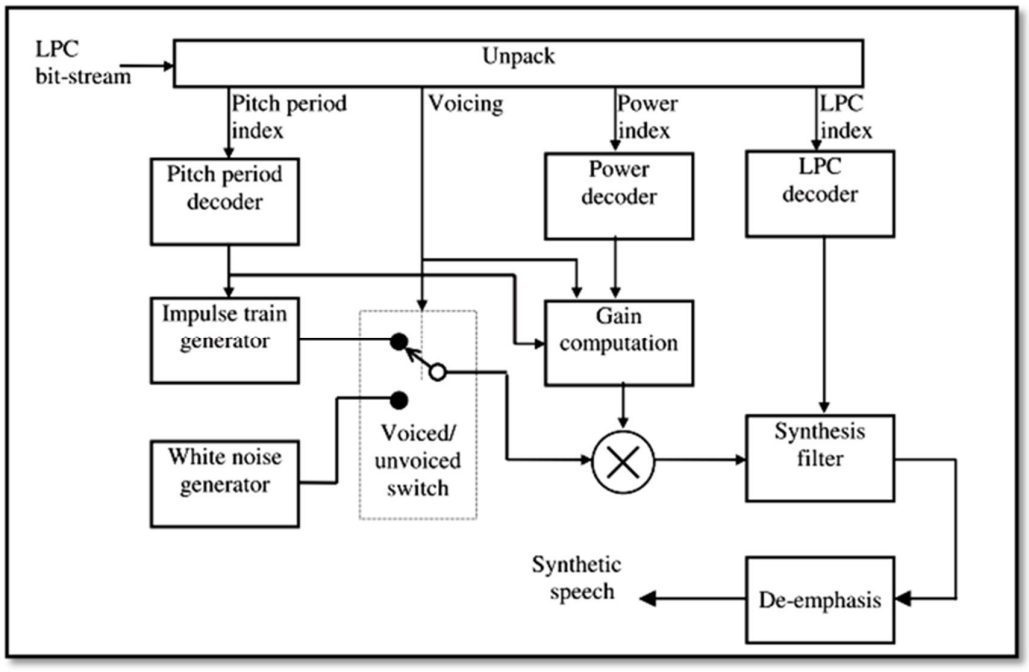

Figure 3.4: The LPC Decoder Block Diagram [2]

3.9.1 Impulse Train and White Noise Generators

An impulse train is used to model the excitation signal assuming the case of voiced speech. It is assumed that the output of the impulse train generator is comprised of a series of unit-amplitude impulses.

The white noise generator is assumed to yield a noise process with a Gaussian distribution, with zero mean and unit variance, i.e., $\mathcal{N}(0,1)$. Ransom noise is used to model the signal contributions for unvoiced speech. Therefore, depending on the voiced or unvoiced state of the signal, the switch is set to the proper location so that the appropriate input is selected.

3.9.2 Voice/Unvoiced Switch

In many instances, voiced/unvoiced classification can easily be accomplished by observing the waveform: a frame with clear periodicity is designated as voiced, and a frame with noise-like appearance is labelled as unvoiced. In other instances, however, the boundary between voiced and unvoiced is unclear; this happens for transition frames, where the signal goes from voiced to unvoiced or vice versa. The necessity to perform a strict voiced/unvoiced classification is indeed one of the fundamental limitations of the LPC model.

3.9.3 Gain Computation

Gain computation is performed as follows. For the unvoiced case, the power of the synthesis filter's input must be the same as the prediction error on the encoder side. Denoting the gain by g, where

$$g = \sqrt{p} \tag{3.32}$$

since the white noise generator has unit-variance output. For the voiced case, the power of the impulse train having an amplitude of g and a period of T, measured over an interval of length $\lfloor N/T \rfloor T$, must equal p. Carrying out the operation yields

$$g = \sqrt{Tp} \tag{3.33}$$

3.9.4 Synthesis Filter

Figure 3.8 depicts the synthesis filter. Mathematically, by exciting the synthesis filter with the system function $\mathcal{H}(z)$, defined by equation 3.3, and can be written as

$$\mathcal{H}(z) = \frac{1}{1 - \sum_{i=1}^{L} a_i z^{-i}} \tag{3.34}$$

then by using a white noise signal, the filter's output will have a PSD close to the original signal as long as the prediction order L is adequate.

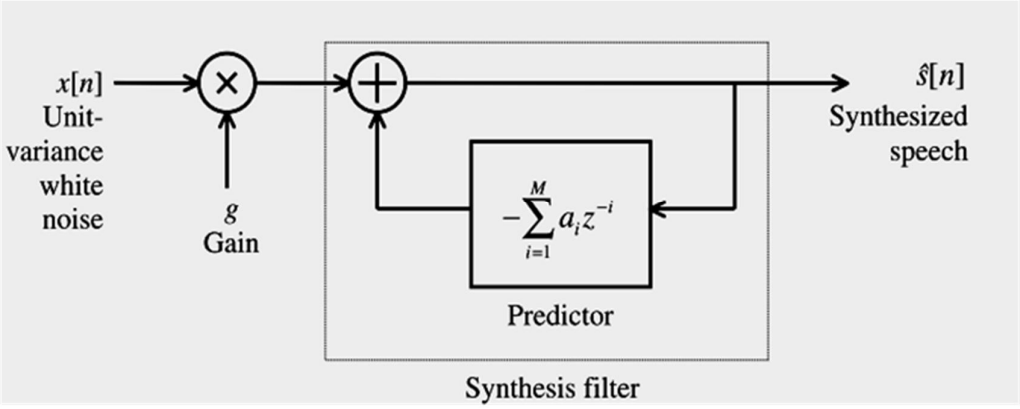

Figure 3.5: The Synthesis Filter Block Diagram [2]

The power spectral density of the original speech is captured by the synthesis filter. In fact, the combined spectral contributions of the glottal flow, the vocal tract, and the radiation of the lips are represented by the

synthesis filter. By using the LSF, stability of the synthesis filter can be restrained easily and the amount of distortion in a different frequency zone can be regulated.

3.9.5 De-emphasis Filter

Finally, the output of the synthesis filter is de-emphasized to yield the synthetic speech. To keep a similar spectral shape for the synthetic speech, it is filtered by the de-emphasis filter with system function

$$G(z) = \frac{1}{1 - \alpha z^{-1}} \tag{3.35}$$

at the decoder side, which is the inverse filter with respect to pre-emphasis which was addressed in section 3.8.1.

Chapter 4 : Simulation and Results

4.1 Introduction

This chapter is started with linear prediction analysis and synthesis simulation. Various choices of parameters for LP analysis are discussed, and the performance for different representations is compared in terms of prediction gain. The model is developed using Matlab, which is a powerful and flexible mathematical simulation platform. In this part, the LPC-based speech vocoder is simulated in three methods to examine the quality of synthesis speech signal. These methods are differed in the effect of pitch frequency, residual signal, and normalized Gaussian noise (the nature of the air from lungs). They are used to improve the spectrum of the prediction filter.

4.2 System Model Assumptions

In order to perform the LP analysis, some basic parameters must be chosen. The variation of these parameters results in a varying performance of the overall algorithm. The major parameters to consider are

- the design parameters of the pre-emphasis filter,
- Prediction order
- Sampling frequency, frame size and time.

The employed assumptions for the system model are shown in table 4.1. The autocorrelation method and Levinson-Durbin recursion are used to solve the LP equations to calculate the LP coefficients. The pitch frequency is calculated with autocorrelation method.

Table 4.1: Assumed System Parameters

System Parameter	Value
Pre-emphasis Filter	$G(z) = 1 - \alpha Z^{-1}$, $\alpha = 0.9375$
Prediction order	$L = 10$
Sampling frequency	$f_s = 8$ KHz
Frame size	$T = 30$ msec $N = 240$ samples
Frame time increment	$T_c = 20$ msec $N_c = 160$ samples

4.3 Model Flow Chart

The flow chart of the employed procedure is shown in two arts in Figure 4.1 and figure 4.2. The overall process can be summarized into the following steps:

1) First the signal is filtered and divided into segments
2) The linear prediction coefficients are calculated along with the prediction error.
3) If a specific frame is voiced, then assign pitch frequency $\leftarrow 0$, else the value of the pitch frequency is calculated with another subroutine.
4) The frame power is calculated.
5) The values are prepared:
    - a Gaussian process, which follows $\mathcal{N}(0,1)$,
    - and the residual signal is calculated.

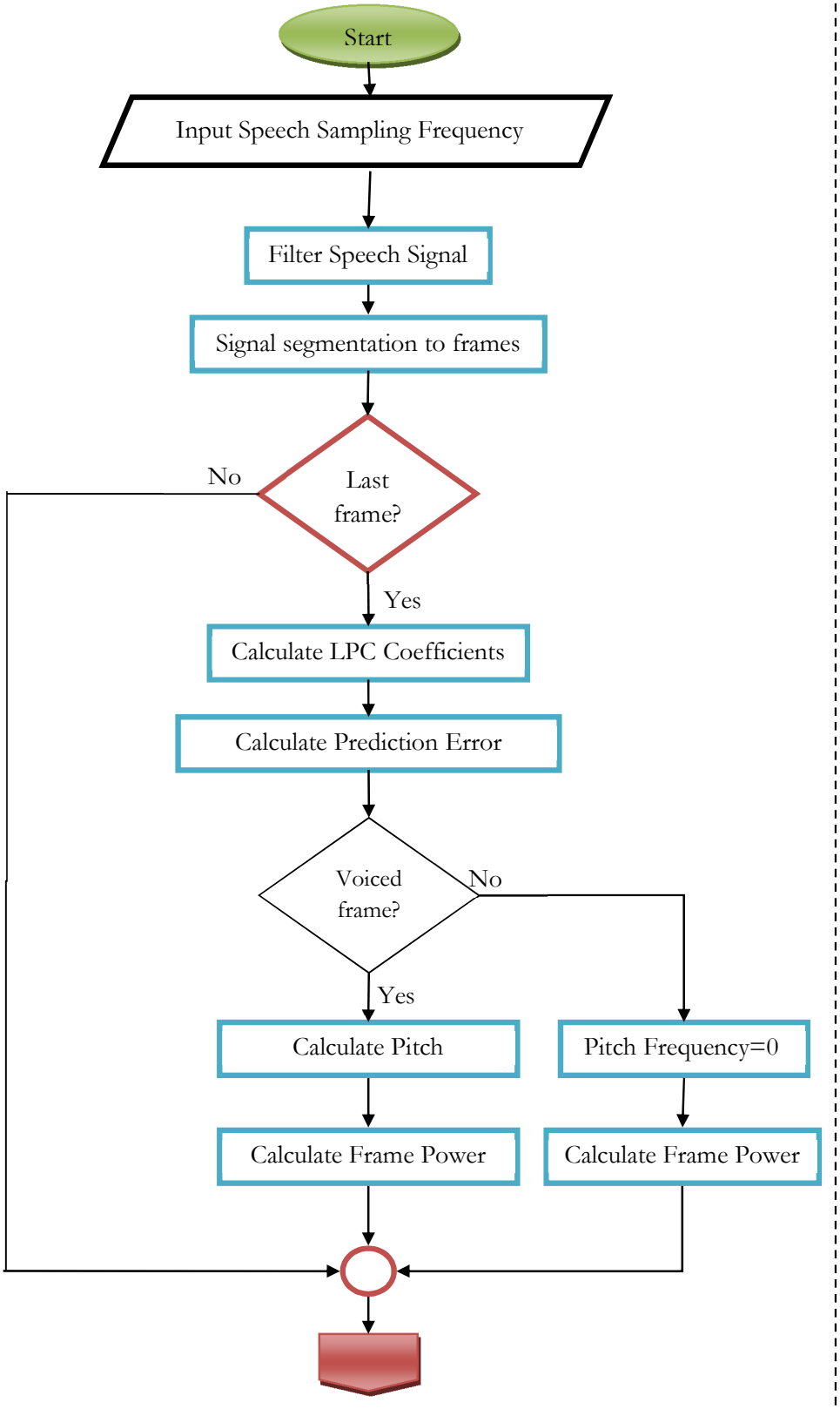

Figure 4.1 : Algorithm Flow Chart

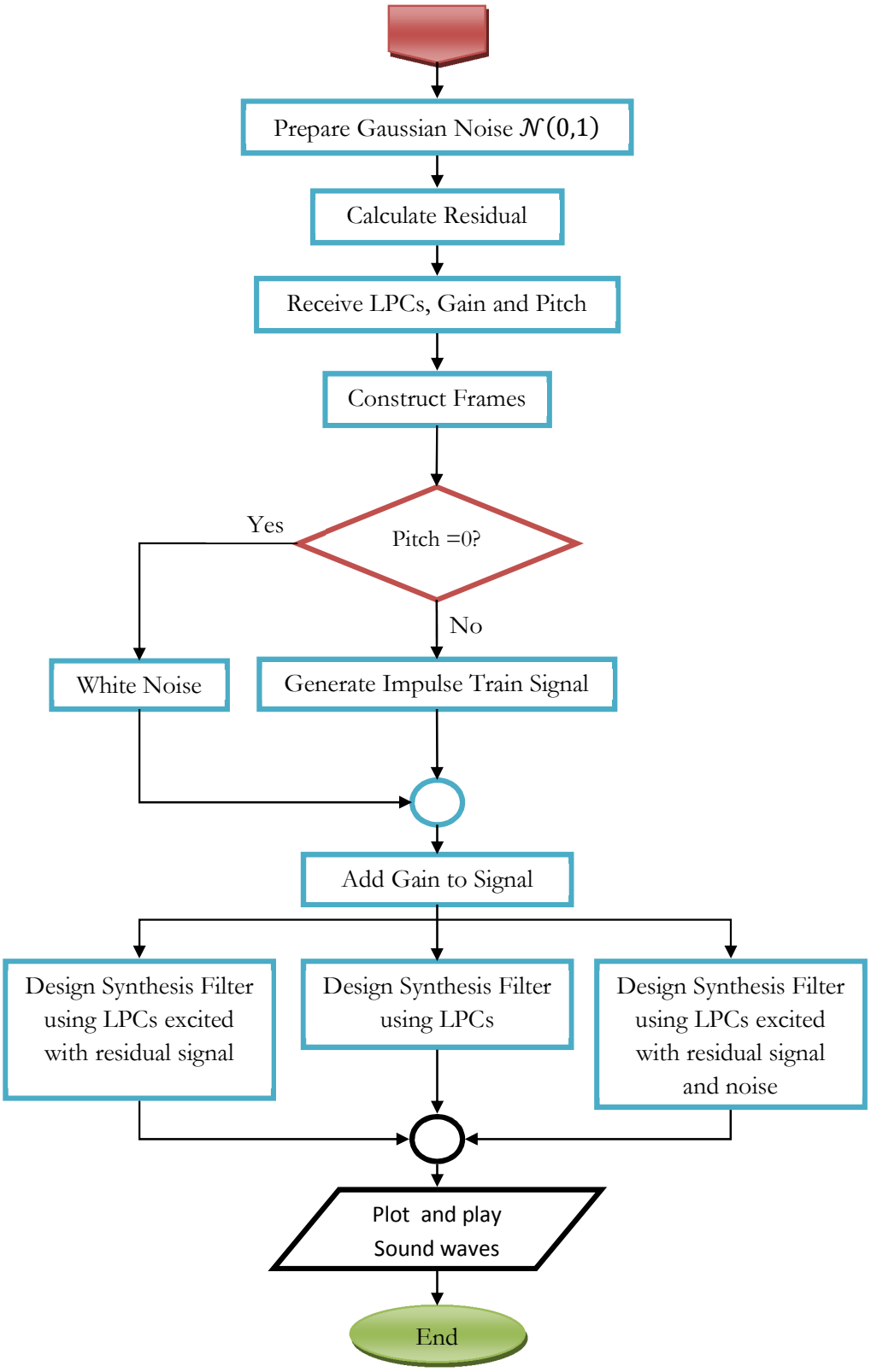

Figure 4.1 : Algorithm Flow Chart

6) The prediction gain, the LPCs and the pitch are processed at the receiving side.
7) The frame sequence is reconstructed.
8) Based on the pitch parameter, either an impulse train or a noise process is generated, and then the gain is added to the signal.
9) The synthesis filter is designed using three approaches
    - using LPCs only
    - using LPCs excited with residual signal
    - using LPCs excited with residual signal and noise
10) Finally, the output datasets are visualized and played.

4.4 Simulations Results and Discussion

4.4.1 Processing the Original Signal

Figure 4.3 depicts the original signal. The signal strength is plotted as a function of time. Figure 4.4 depicts the signal when the pitch frequency is estimated using the autocorrelation method, and figure 4.5 depicts the signal when the linear pitch frequency is estimated using the cepstrum method (explained in section 3.7.4). The figures are different in signal strength, however, when the sounds are played, the ear can pick differences. Figure 4.6 illustrates the output LPC compressed signal, while Figure 4.7 depicts the output voice-excited LPC compressed signal using residual and additive standard Gaussian noise.

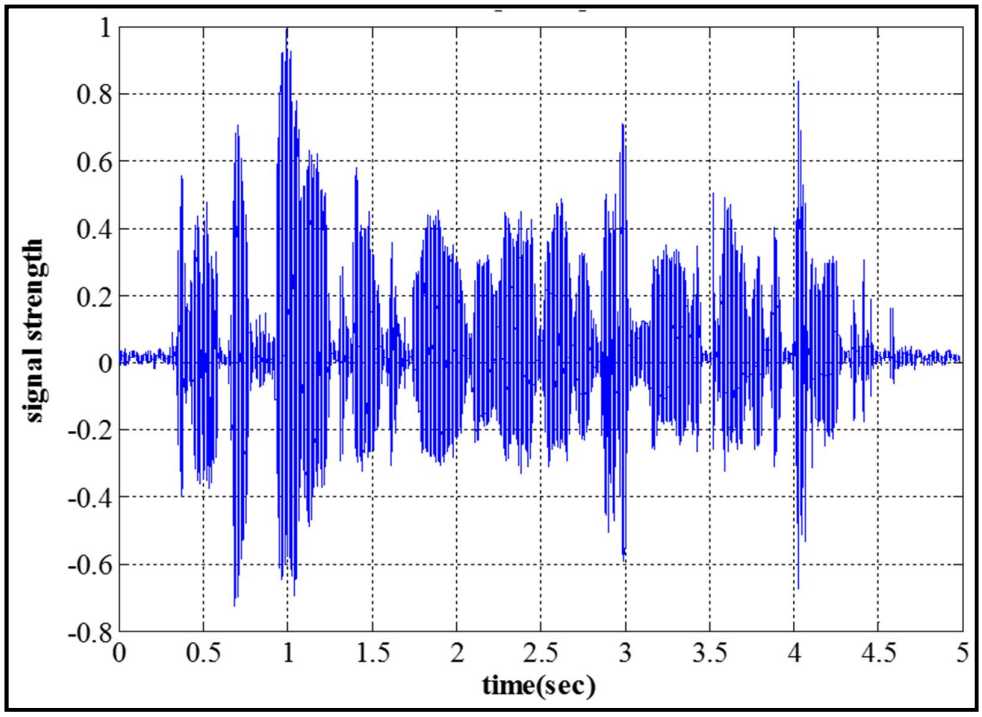

Figure 4.2: Original signal

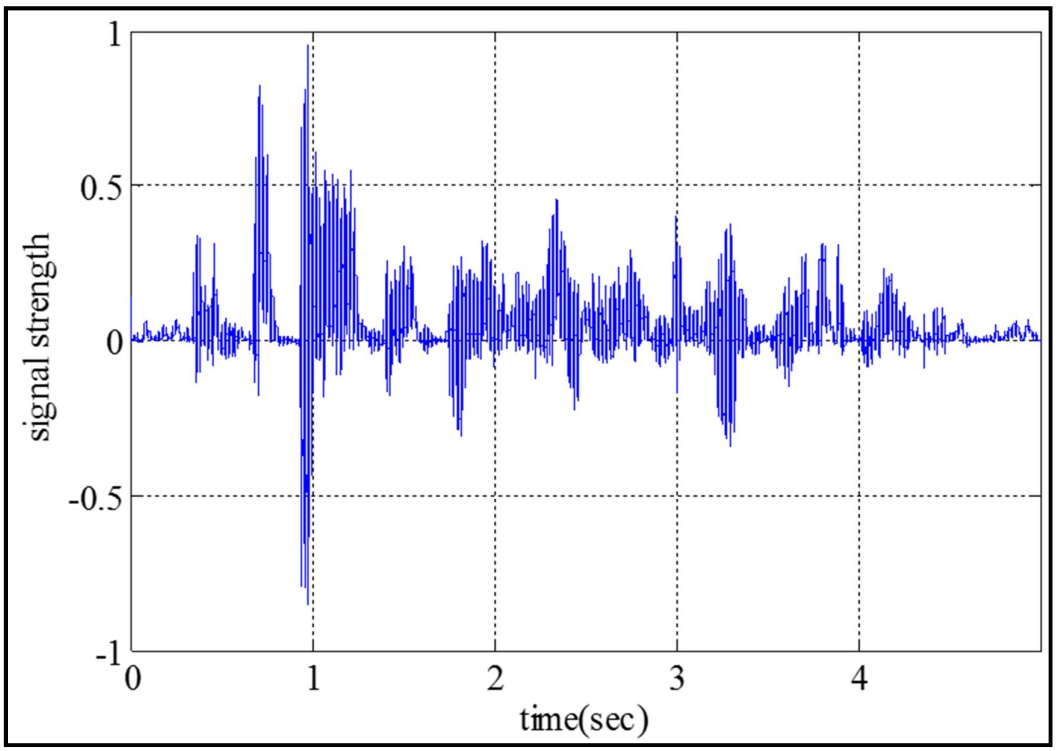

Figure 4.3 LPC compressed signal with Pitch Estimation via Auto-correlation Method

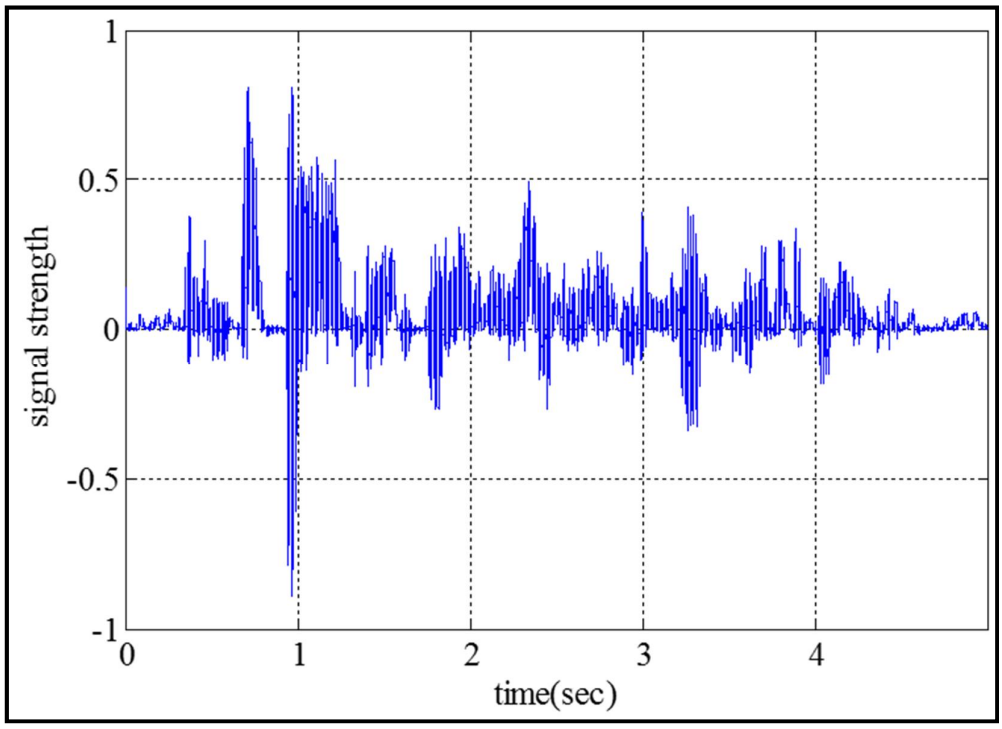

Figure 4.4: LPC compressed signal with Pitch Estimation via Cepstral Method

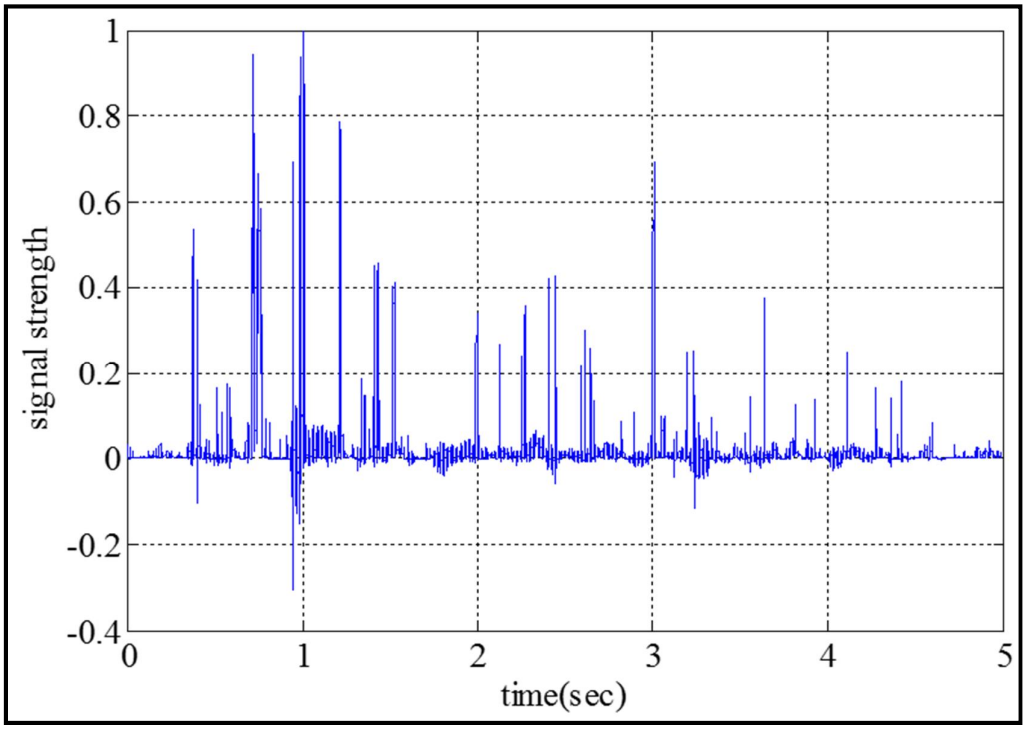

Figure 4.5: LPC compressed signal

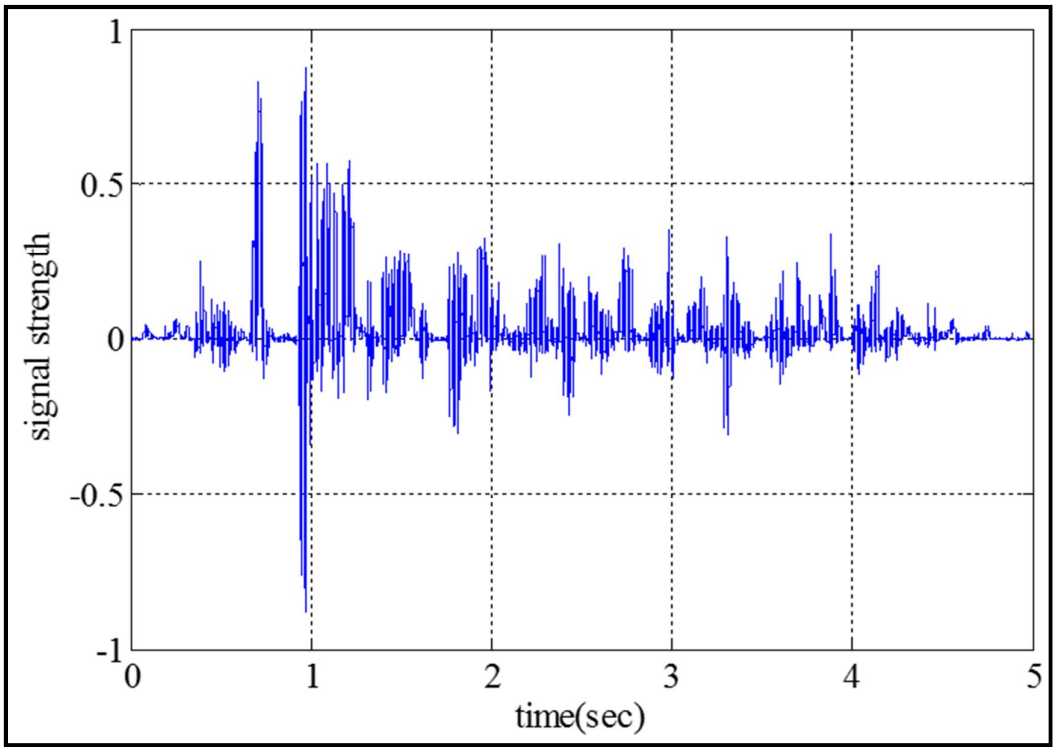

Figure 4.6: voice-excited LPC compressed signal using residual and standard Gaussian

4.4.2 Impact of Filter Order

It is necessary to find the minimum order of the LP analysis required to model the significant features of the speech. When the speech spectrum is modelled, the vocal tract resonances or formants are important. It has been shown previously in [20] that to model the vocal tract resonances the memory of filter $\mathcal{A}(z)$ must be at least twice the time required for the sound wave to travel from glottis to lips. This time interval is $2V/c$, where V is the length of the vocal tract (usually 17 cm) and c is the speed of the sound wave (340 m/s). So, the memory should be at least 1 ms. When the sampling frequency is 8 kHz, 1 ms memory means using 8 previous samples. Thus, the order of the filter should be at least 8.

4.4.3 Impact of Prediction Order on Prediction Gain

Figure 4.8 depicts the prediction gain as a function of the filter order. It can be seen that the prediction gain is increased as the filter order increases. However, as the filter order goes high, the prediction gain converges to a specific value. In fact, a high prediction gain implies that the LP filtering is likely to reflect the effect of the vocal tract more accurately so that the residual will be closer to the true excitation.

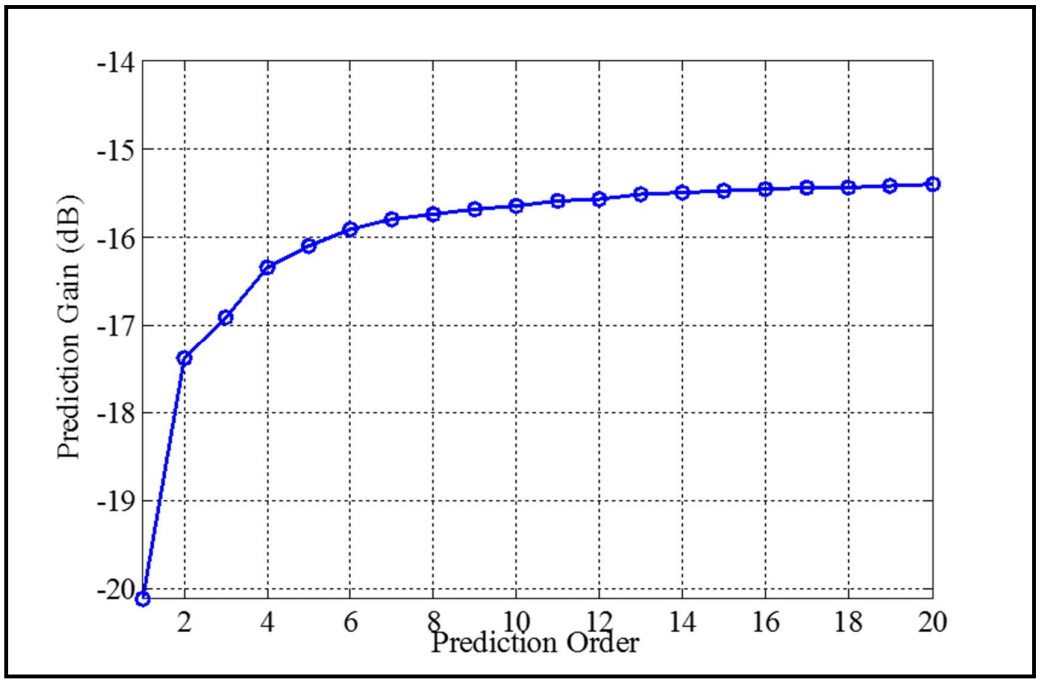

Figure 4.7: Prediction Gain versus the prediction order

Chapter 5 : Conclusion and Recommendation

5.1 Conclusions

This thesis focused on the problem of enhancement of linear predictive coding. First, preliminary background information about speech coding was presented. This included the properties of speech signals and the basic aspects of speech encoders. A review of linear prediction analysis of speech was presented. The resulting speech quality has been assessed objectively.

The results of study on the parameters of LP analysis; the prediction gain does not depend much on the window length. The position of the window with respect to the frame is defined by the window offset. This parameter affects the prediction gain. The highest prediction gain is obtained when the window and frame centers are aligned. The prediction gain is affected by the filter order. The prediction gain can get affected by the frame length, but it does not yield much difference in the quality of the heard audio.

5.2 Recommendations and Future Work

A lot of work has been put into this thesis but still there is room for improvement. For example, the following points can be considered:

- In this thesis, not all methods of pitch frequency estimation are considered, only the autocorrelation method and the cepstrum method are considered. Another method using the MDF function and the fractional pitch period Medan-Yair- Chazan algorithm [28] can be tried and analysed.

- In LPC estimation, the autocorrelation method is used. It is a fast method; other methods can be tried, such as the covariance, although it takes more time. Therefore, there is a trade-off between both methods that requires to be assessed.
- In this thesis, to apply LP transformation, only the log area ratios are used, they are explained in section 3.6.1. This method is used cause it is a function of the reflection coefficients which are a natural by-product of the previous steps. Therefore, LARs are chosen, but further work can be done by testing the LP transformation method using line spectral frequencies.

Bibliography

[1] J. Benesty, M. M. Sondhi and Y. Huang, Springer Handbook of Speech Processing, Springer Science & Business Media, 2007.

[2] R. G. Goldberg and L. Riek, A practical handbook of speech coders, CRC Press LLC, 2000.

[3] W. C. Chu, Speech Coding Algorithms: Foundation and evolution of standardized coders, John Wiley & Sons, Inc, 2003.

[4] I. Otung, Communication Engineering Principles, Palgrave Macmillan, 2001.

[5] W. Hess, Pitch Determination of Speech Signals: Algorithms and Devices: Algorithms and Devices, Springer Science & Business Media, 2012.

[6] A. Peinado and J. Segura, Speech Recognition Over Digital Channels: Robustness and Standards, John Wiley & Sons, 2006.

[7] J. D. Gibson, The Communications Handbook, CRC Press, 2002.

[8] V. Madisetti and D. Williams, Digital Signal Processing Handbook, CRC Press, 1999.

[9] A. Gatherer and E. Auslander, The Application of Programmable DSPs in Mobile Communications, John Wiley & Sons, 2002.

[10] M. Rajman, Speech and Language Engineering, EPFL Press, 2007.

[11] M. R. Schroeder and B. S. Atal, “Code-excited linear predictive (CELP): High quality speech at very low bit rates,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, Tampa, 1985.

[12] B. Atal, R. Cox and P. Kroon, “Spectral quantization and interpolation for CELP coders,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, Glasgow, 1989.

[13] M. Yong, “A new LPC interpolation technique for CELP coders,” *IEEE Trans. Commun.,* vol. 42, Jan 1994.

[14] I. A. Gerson and M. A. Jasuk, “Vector Sum Excited Linear Predictive Coding,” in *Advances in Speech Coding*, Springer Science & Business Media, 2012.

[15] K.-L. Du and M. N. S. Swamy, Wireless Communication Systems: From RF Subsystems to 4G Enabling Technologies, Cambridge University Press, 2010.

[16] J. D. Gibson, Digital Compression for Multimedia: Principles and Standards, Morgan Kaufmann, 1998.

[17] V. Madisetti, Video, Speech, and Audio Signal Processing and Associated Standards, CRC Press, 2009.

[18] N. Jayant, Signal Compression: Coding of Speech, Audio, Text, Image and Video, World Scientific, 1997.

[19] A. Akmajian, R. A. Demers and R. M. Harnish, Linguistics: An Introduction to Language and Communication, The MIT Press, 1984.

[20] J. D. Markel and A. H. Gray, Linear Prediction of Speech, Berlin: Springer-Verlag, 1976.

[21] L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, Englewood Cliffs: Prentice-Hall, 1978.

[22] J. LeRoux and C. Gueguen, "A fixed point computation of partial correlation coefficients," *IEEE Trans. ASSP,* no. ASSP-27, pp. 257-259, 1979.

[23] V. Madisetti, The Digital Signal Processing Handbook, CRC Press, 1997.

[24] R. D. Kent and R. Charles, The Acoustic Analysis of Speech, Singular/Thomson Learning, 2002.

[25] R. Linggard, Electronic Synthesis of Speech, CUP Archive, 1985.

[26] J. E. Gentle, Matrix Algebra: Theory, Computations, and Applications in Statistics, Springer Science & Business Media, 2007.

[27] F. K. Soong and H. S. Lee, "Line spectrum pair (LSP) and speech data compression," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, San Diego, 1984.

[28] Y. Medan, E. Yair and ,. D. Chazan, "Super resolution pitch determination of speech signals," *IEEE Transactions on Signal Prcessing,* pp. 40 - 48, 2002.

[29] A. Oppenheim and R. Schafer, Digital Signal Processing, Englewood Cliffs, NJ: Prentice Hall, 1975.

[30] G. Friedland and R. Jain, Multimedia Computing, Cambridge University Press, 2014.

[31] P. P. Vaidyanathan, The Theory of Linear Prediction, Morgan & Claypool Publishers, 2008.

[32] S. W. Smith, The Scientist and Engineer's Guide to Digital Signal Processing, California Technical Publishing, 1997.