SPEECH CODING TECHNIQUES IN MOBILE COMMUNICATION SYSTEMS

A MASTER THESIS

IN

ELECTRICAL AND ELECTRONIC ENGINEERING

ATILIM UNIVERSITY

BY

SALEH ABULGASEM

JANUARY-2017

SPEECH CODING TECHNIQUES IN MOBILE COMMUNICATION SYSTEMS

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

OF

ATILIM UNIVERSITY

BY

SALEH ABULGASEM

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

ELECTRICAL AND ELECTRONICS ENGINEERING DEPARTMENT

JANUARY-2017

Approval of the Graduate School of Natural and Applied Sciences, Atılım University.

_____

Prof. Dr. İbrahim AKMAN

Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

_____

Asst. Prof. Dr. K. Efe ESELLER

Head of Department

This is to certify that we have read the thesis "**Speech Coding Techniques in Mobile Communication Systems**" submitted by "Saleh Abulgasem" and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

_____

Asst. Prof. Dr. İ. Baran USLU

Supervisor

Examining Committee Members

Prof. Dr. H. Gökhan İLK _____

Prof. Dr. Ali KARA _____

Asst. Prof. Dr. Kutluk Bilge ARIKAN _____

Asst. Prof. Dr. Hakan TORA _____

Asst. Prof. Dr. İ. Baran USLU _____

Date: 24.01.2017

I would like to declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: SALEH ABULGASEM

Signature:

ABSTRACT

SPEECH CODING TECHNIQUES IN MOBILE COMMUNICATION SYSTEMS

Saleh Abulgasem

M.Sc. Electrical and Electronic Engineering

Supervisor: Asst. Prof. Dr. İ. Baran USLU

January 2017, 67 Pages

This thesis compares the speech quality, bit-rates and complexity for basic Linear Prediction Coding (LPC-10), Code Excited Linear Prediction (CELP, Federal Standard 1016) speech coder, and Analysis by Synthesis (AbS) method. The speech qualities of these three speech coding techniques are compared by performing subjective listening tests namely MOS (Mean Opinion Score) tests. All the simulations are implemented in Matlab. The test signals used are pure signals (i.e. there is no intentional noise creation). Results indicate that for speech quality of all tested signals the CELP standard performs the best among the other used methods. While for bit-rate and processing time (complexity) the LPC-10 coder gives the best result.

KEYWORDS: Speech compression, speech coding, LPC, AbS, CELP, mobile communication, MOS.

ÖZ

GEZGİN İLETİŞİM SİSTEMLERİNDE

KONUŞMA KODLAMA TEKNİKLERİ

Saleh Abulgasem

Yüksek Lisans Tezi. Elektrik ve Elektronik Mühendisliği Bölümü

Tez Yöneticisi: Asst. Prof. Dr. İ. Baran USLU

Ocak 2017, 67 sayfa

Bu çalışmada, Doğrusal Öngörümlü Kodlama-10 (Linear Predictive Coding – LPC-10), Kod Uyarımlı Doğrusal Öngörüm (Code Excited Linear Prediction - CELP) ve Sentez ile Analiz (Analysis by Synthesis - AbS) konuşma kodlama yöntemleri, ses kalitesi, bit oran ve karmaşıklık açısından karşılaştırılmaktadır. Konuşma kalitesi açısından başarım karşılaştırması öznel dinleme testleriyle, yani MOS (Ortalama Görüş Skoru) testi ile yapılmıştır. Tüm benzetimler Matlab programında gerçekleştirilmiştir. Testlerde kullanılan konuşma sinyalleri temizdir (herhangi bir arkaplan gürültüsü içermemektedir). Sonuçlar göstermektedir ki, ses kalitesi açısından yöntemler içerisinde en başarılısı CELP yöntemidir. Diğer yandan; bit oranı ve işlem süresi (karmaşıklık) açısından LPC-10 yöntemi en iyi sonuçları vermektedir.

ANAHTAR KELİMELER: Konuşma sıkıştırma, konuşma kodlama, LPC, AbS, CELP, gezgin iletişim, MOS.

ACKNOWLEDGEMENTS

First and foremost I would like to thank Dr. Baran Uslu for being my advisor and guiding me throughout the course of my graduate studies. Also I would like to thank everyone who gave a helping hand.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AaS	Analysis and Synthesis
AbS	Analysis by Synthesis
ACELP	Algebraic Code-excited Linear Prediction
AMR	Adaptive Multi-Rate
BSC	Base Station Controller
BSS	Base-Station Subsystem
BTS	Base Transceiver Station
CELP	Code-excited Linear Prediction
CS-CELP	Conjugate Structure Code-excited Linear Prediction
FIR	Finite Impulse Response
GSM	Global System for Mobile communication
IIR	Infinite Impulse Response
LD-CELP	Low Delay Code-excited Linear Prediction
LPC	Linear Prediction Coding
LSP	Line Spectral Pairs
LTI	Linear Time Invariant
LTP	Long Term Prediction
MELP	Multi-pulse Excitation Linear Prediction
MOS	Mean Opinion Score
MSC	Mobile Switching Center
MSE	Mean Square Error
NSS	Network Switching System
PSTN	Public Switched Telephone Network
RPE-LTP	Regular Pulse Excitation Long Term Prediction
SNR	Signal to Noise Ratio
STP	Short Term Prediction
VSELP	Vector Sum Excited Linear Prediction

CHAPTER 1: INTRODUCTION

The research of speech processing has involved in a lot of helpful applications as a result of the natural sense of speech communication. On this account speech processing is fundamental to the operation of these electronic digital cellular communication networks such as Global System for Mobile communication used in Europe, and of course many other networks throughout the world, beside rising in Voice over Internet Protocol (VoIP) applications [1].

The effective digital depiction of the speech signal allows user to make use of bandwidth efficiently in the transmitting of the signal or even storage space. In cellular communication systems this type of bandwidth efficiency is of critical significance, simply because in mobile communication the channel bandwidth is restricted [2].

In the past decade, the applications of speech communication have completely advanced significantly. Study into new coding techniques and improvement of pre-existing methods has proceeded at a speedy pace, in addition with the current market needs for better coders. Digital cellular, and Internet voice communications are most of the notable day-to-day applications that are attracting the marketplace demand. The quest here is higher quality speech at a reduced transmission bandwidth. The need will certainly keep on growing with the development of remote oral communication.

Literature Review

Digital encoding of speech started long time ago, many developments have been made in speech processing technology, specifically in speech coding, speech manipulation is now widely used in international, digital mobile and wireless networks. But still exist demands for lower bit rates and more good quality coding techniques for many network applications.

A. Kondoz (1987) *Digital Encoding of Speech Signal*, gives many aspects of digital speech coding. Developing of speech coding has been gone into many directions, the

most important one is type of excitation signal which is derived from pitch period (such RELP Regular, and MELP Mixed Excited Linear Prediction …) also MP-LPC Multi-Pulse and RPE Regular Pulse Excitation coding, later RPE-LTP Regular Pulse Excitation Long Term Prediction was used as GSM standard.

The other path is by quantization, using many proposed techniques such as vector quantization, Sumesh. Kaul (1984) *Vector quantization techniques for speech coding*, Sophisticated coders from Analysis-by-Synthesis (AbS) family like CELP gets benefits from residual signal beside mentioned excitation signal.

Karim Abboud (1992) *Wideband CELP Speech Coding,* was to study the coding of wideband speech and to improve on Code-Excited Linear Prediction (CELP) coders in terms of speech quality and bit rate by using vector quantization.

Hans Engström & Johan Ross (2008) *Voice Codec for Floating Point Processor*. This work aimed to implement a functional speech decoder adapted to the floating point DSP. Also neural network can get involved in this field.

Robert Zopf (1995) *Real-Time Implementation of A variable Rate CELP Speech Codec*, he claim that to get a high quality and average bit rate CELP it is necessary to make a tuning of the output bit rate according to an analysis of input speech signal. Shum, Ellen (1998) *Optimisation techniques for low bit rate speech coding*, made a comparison between half-rate and full-rate speech coding used in GSM networks, the author says that the half-rate algorithm is more efficient speech compression than full-rate. Although GSM full-rate codec operates at 13 kbps and half-rate at 5.6 kbps, they both give a good toll speech quality.

In more sophisticated codec called Adaptive Multi Rate AMR, M. Goudarzi *(2008) Evaluation of Voice Quality in 3G Mobile Networks*, investigates the performance of AMR-GSM by Using objective measurement tools, as well as the effect of other parameters such as the gender of the speaker by over 200 speech samples. AMR codec is based on an A-CELP technology (Algebraic Code Excited Linear Prediction).

M. Tandel, V. Shah, B. Patel (2011) *Implementation of CELP CODER and to evaluate the performance in terms of bit rate, coding delay and Quality of speech*, a paper discusses the implementation of CELP CODEC and its performance and

resulted that the CELP codec output is quite similar to the input signal, the CELP gives high quality speech at bit rates between 4.8 and 9.6 Kbps.

Even though the revolution of speech communication tends toward cheaper applications such as Voice over Internet Protocols VoIP applications there is no much difference in using of speech coding techniques by these applications.

This thesis studies the three coding techniques LPC-10, AbS and CELP of speech signal, concentrating on using pitch period to implement excitation signal and used in LPC-10 to make equivalent classification decision, it compares all coders in speech quality, bit-rate stream and complexity.

Hypothesis

The main purpose of this thesis was to carry out a detailed analysis of the performance differences between LPC-10, AbS, and CELP techniques. Synthetic output speech, which is the result of LPC-10 (implemented in MATLAB) speech processing and the same speech signals processed by the AbS and CELP method (both implemented in MATLAB) are used as test signals. Comprehensive subjective listening tests were conducted to test the quality of speech codecs.

Organization of the Thesis

The first chapter starts with an introduction and gives a general concepts of speech processing as well as literature review. The second chapter about GSM networks, speech codecs and standards. the third chapter details the basics of speech and beside the details of basic coder LPC-10 (FS1015) also lists out the various domains of speech and their specific characteristics. The fourth chapter describes the Analysis-by-Synthesis and Code-excited Linear Prediction algorithms, this chapter describes both the AbS and Federal Standard CELP (FS1016) speech compression technique in detail. The fifth chapter discusses the experiments and simulation results and last chapter states the conclusion derived from these results.

1.1 Speech Processing

Speech processing is the exploration of speech signals as well as processing techniques of these signals. The signals can certainly be highly refined in a digital depiction; therefore speech processing could be thought to be an exceptional case of digital signal processing, used for speech signal. The different components of speech processing consist of the reconstruction, manipulation, storage, analysis and synthesis of digital speech signals as well. It can also be linked to natural language processing.

1.2 Digital Speech Processing

Thinking about the application of digital signal processing methods to speech communication difficulties, it is really useful to pay attention to three major topics:

The representation of speech signals in digital form, the execution of processing methods reach from the very simple pulse code modulation technique (PCM) to sophisticated vocoders, as well as the classes of applications which depend greatly on digital processing. The representation of speech signals in digital form is, obviously, of basic matter. Within this regard we have been directed by the famous sampling theorem which affirms that a band restricted signal could be represented by samples taken occasionally over time given that the samples are taken at a very high rate. Hence, the technique of sampling dominates all the theory and application of digital speech processing. There are lots of potential for discrete representations of speech signals, these types of representations could be categorized into two groups, specifically waveform representations and parametric representations [3].

1.3 Speech Processing Applications

The fundamental domains of speech processing are:

1.3.1 Speech Coding

Generally concept, speech coding is a technique to represent a digitized speech signal using couple of bits as they possibly can, keeping up simultaneously a realistic degree of speech intelligibility and clarity, Speech coding continues to be and still a big issue in the area of digital speech processing [4].

In the real truth, an having access to limitless volume of bandwidth is impossible. For that reason, there is certainly a necessity to code and compress signals. Speech compression is needed in transmission over long channel or long distance communication, high quality speech storage space. Take for instance, in digital cellular technology lots of customers really need to share the exact same rate of bandwidth portion. Using speech compression gives you a lot more customers to share the available system.

1.3.2 Speech Recognition

Speech recognition is the method of translation of voiced words into readable file format, which evidently implies into explicit language.

1.3.3 Speaker Recognition

To recognize or to know the speaker; this accomplished by knowing personality characteristics of that speaker such as pitch period, this can be done by comparing a parameter of particular input signal to many other similar type parameters stored in machine or system. Similar is the case word recognition, in which the words voiced by the spokesman are compared to the readymade saved data. These kinds of systems are generally put to use as security systems or perhaps recognition systems at which real recognition of individuals is absolutely essential.

1.3.4 Speech Synthesis

Speech synthesis is the man-made imitation of human speech. A system designed for this function is known as a speech synthesizer, which can be used in software package or even hardware.

A text-to-speech system transforms a verbal representations such as phonetic transcriptions into an oral representation. Synthesized speech could be made by concatenating letters/words of recorded speech which are saved in databases, so simply you can build your own speech synthesizer with a collected of recorded words or letters of yourself [5].

Systems vary in the size of the saved speech units; a system that saves phones or diaphones offers the biggest output range. For particular usage domains, the storage space of the whole words or sentences provides for good quality output. On the other

hand, a synthesizer can easily incorporate a model of the vocal path as well as other human voice features to create a completely and totally "synthetic" voice output.

1.3.5 Speech Enhancement

In many cases, speech signals are affected by noise or even echoes in ways in which restrict their performance for communication. In this kind of instances digital signal processing methods can be utilized to enhance the speech good quality by eliminate that noise or echo from the speech signal.

1.4 Discrete-Time Speech Signal

In most cases of affairs concerning signal processing, it is really normal to get started with a representation of the signal as a continually varying pattern. The sound wave produced in human speech is most definitely of this character. It is mathematically hassle-free to symbolize such continually different patterns as functions of a variable t, which signifies time. Within this study we will make use of notation of the type x(t) to represent continually varying (or analog) time signals.

As we will see, it can also be easy to symbolize the speech signal as a series of figures. Generally the notation of the form s(n) has been used to represent sequences. In the event that, as is the case for sampled speech signals, a sequence could be regarded as a sequence of samples of an analog signal taken occasionally with sampling period, T, in that case we might find it helpful to clearly show this through the use of the notation, s(nT). Figure 1 displays a good example of a speech signal symbolized both as an analog signal as well as a sequence of *samples* at a sampling rate of 8-16 kHz. In the subsequent numbers, comfort in plotting might determine the usage of the analog representation (i.e. continual functions) even if the discrete representation has been taken into consideration. In this kind of instances, the continuous curve can easily be thought of as the envelope of the sequence of samples.

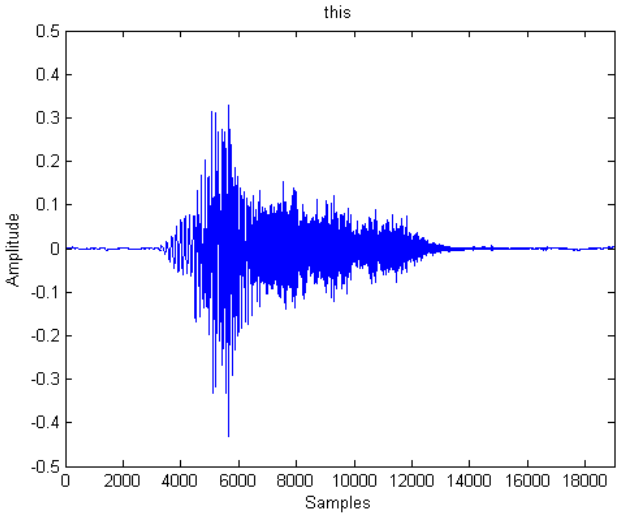


Figure 1 An Example of Discrete Signal

1.5 Z-Transform

The z-transform offers a helpful frequency representation to examine the spectral features of a poles and zeros system of discrete signal, and as the more common illustration of the Fourier transform. The z-transform is known as:

$$S(z) = \sum_{n=-\infty}^{\infty} s(n) z^{-n} \quad (1)$$

The inverse transform of S(z) to achieve s(n) is:

$$s(n) = \frac{1}{2\pi j} \oint S(z) z^{n-1} dz \quad (2)$$

Region of Convergence (RoC)

RoC is an essential in z-transform because it gives the definition of region where z-transform can take place. To determine RoC, it is good to formulate the Z-Transform as:

$$S(z) = \frac{Q(z)}{P(z)}$$

Where the zeros and poles of the equation are represented by P(z) and Q(z). The RoC of the Z-transform as mentioned depends on the convergence of the polynomials P(z) and Q(z) [6].

CHAPTER 2: MOBILE COMMUNICATION SYSTEMS

The growth of cellular phone market and cellular subscribers to the networks (more than 2 billion) has led to many developing stages of speech codecs, because the speech communication is the major issue in cellular mobile field. In same way the demand of low-bit rate speech coder has increased to make as much as possible users using the same network channel. Without using speech codecs the bandwidth would be wider.

Focusing on GSM, this chapter describes the speech coding standards, and most popular worldwide used mobile networks, also this chapter gives an overview of speech codecs in those networks.

The general block diagram of digital communication system is shown in Figure 2. With different purposes and different amount of redundancies, source coding aims to improve the bit rate so it removes some information, but channel coding adds small amount of information for controlling and error correction, finally the modulator makes a waveform suitable for transmission. Demodulator, channel and source coders do the opposite operations what have done in transmitter side [7].

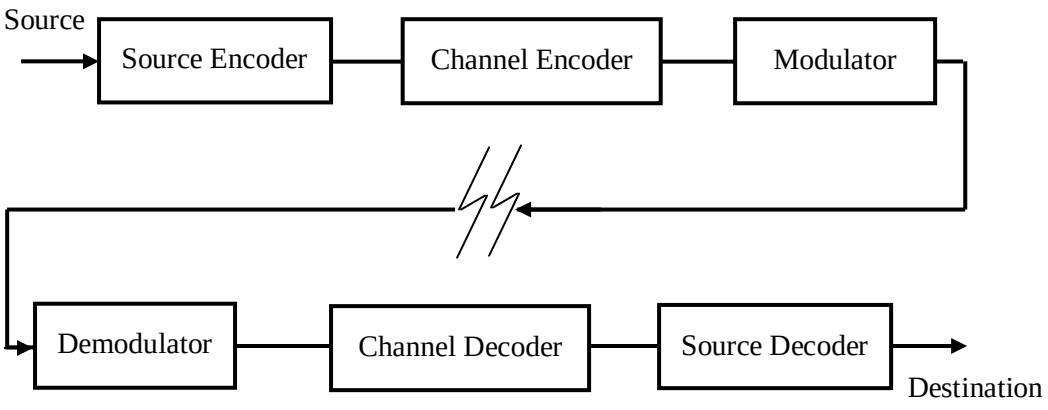


Figure 2 Block Diagram of Digital Communication System

2.1 Speech Coding Standards

Simply standard means manipulation in a particular way extended to generalized solution designed for a specific problem or task, it permits some factory to produce a speech coding equipment similar in processing with another.

Speech coding is performed according to a sequence of specified steps known as an algorithm; it's a set of rules followed by calculations can be elaborated by computer program (for instance Matlab) or a machine. An algorithm in general is any predefined procedure that takes some values as input and produces some values as output. A lot of signal processing problems such as finding pitch period in speech coding can be formulated as a unambiguous computational problem. In basic, an algorithm is specified with a set of instructions, can the machine understand.

Next organizations and their standards are listed, and briefly reviewed [8]:

- **ITU-T Speech Coding Standard**

The Telecommunications Standardization ITU-T Sector of the International Telecommunication Union (ITU) is focused on creating speech coding standards for network telephony for both wired and wireless networks.

The ITU-T, in the past exactly in 1993 the name was Telephone and Telegraph Institute Consultative Committee (CCITT) has standardized speech coding methods mainly for PSTN telephony with 3.4 kHz input speech band-width and 8 kHz sampling frequency.

- **European Digital Cellular Telephony Standards**

European Telecommunications Standards Institute (ETSI) independent organization has memberships from European countries and companies, whose purpose is to produce and create telecommunication equipments and standards. ETSI is organized by application; the most important group in speech coding is originally named the Group Special Mobile – (GSM) after being able to work worldwide under GSM networks, the meaning was changed to Global System for Mobile Communications [9].

The number of subscribers of GSM networks at end of 2009 reached 4 billion and still growing rapidly [10].

2.2 Global System for Mobile Communications

2.2.1 Second Generation Cellular System:

Global System Mobile by far controls the world today in over a hundred countries. These networks operate at 9.6 Kbps and are based on international standards defined by the European Telecommunications Standards Institute (ETSI).

2.2.2 GSM Network Architecture:

The architecture of GSM network defines several interfaces for multiple suppliers. A simple architecture is shown in Figure 3

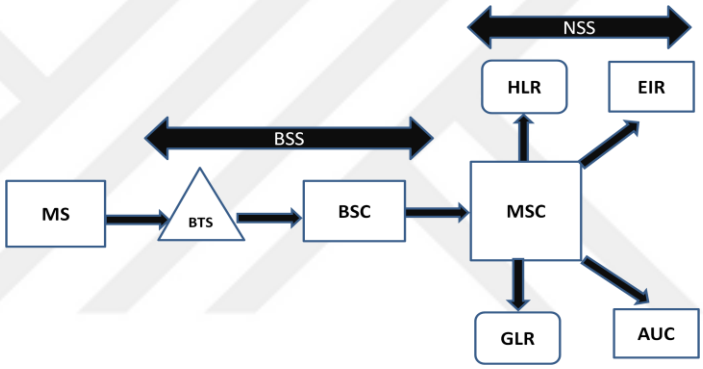


Figure 3 A Simple Architecture of System Mobile Network

Mobile station (MS): any physical equipment that can communicate with the network such as cell phone, or Personal Data Assistant-PDA. A Mobile Station (ME) includes a wireless transceiver, a digital signal processor, and a subscriber identity module (SIM). SIM part can decide the device is MS or not. Speech coding and coding take place in this unit.

Base Station Subsystem (BSS): is responsible for transmitting, receiving and managing signal traffic, it is composed of a base station controller (BSC) and base transceiver stations (BTS) these two parts can communicate together by Abis standardization. Each BTS defines a single cell and includes radio antenna, radio transceiver and a link to a base station controller BSC as shown in Figure 4. BSC

reserves radio frequencies, manages handover of mobile units from one cell to another within BSS, and controls paging service.

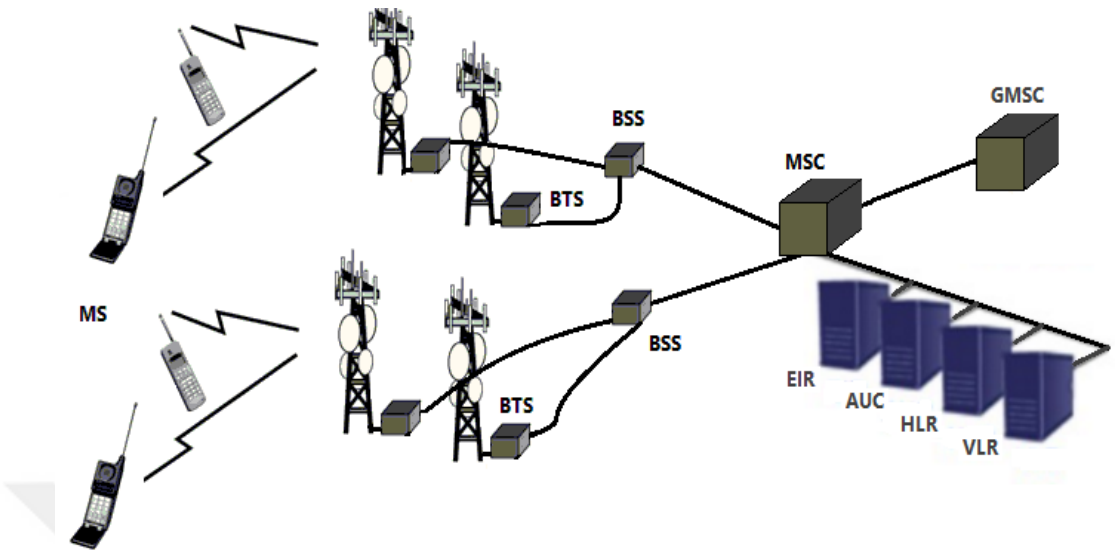


Figure 4 A Detailed Architecture of System Mobile Network

Mobile Switching Center (MSC): It is the heart of GSM network. It is the main part of Network Switching System (NSS). It connects cellular networks with Public Switched Telecommunications Networks (PSTN).

Specifically, an MSC controls handoffs between cells in different BSSs, authenticates users and validates accounts, and enables worldwide roaming of mobile users. To support these features, an MSC consists of the following databases:

- Home location register (HLR) database – stores information about each subscriber that belongs to it.
- Visitor location register (VLR) database – keeps the information about subscribers physically, more precisely current position.
- Authentication center database (AUC) – for authentication permission activities and keeps encryption keys
- Equipment identity register database (EIR) – keeps known of the type of devices that exists at the mobile station.

2.2.3 GSM concepts

Handover

Also called handoff is process of moving user in MS from point to another without losing the connection, there are many types of handover in GSM.

Types of GSM handover [10]:

- ***Intra-BTS handover:*** when the interferences of changing place happen, then it is required to change channel used by a mobile. In this type of handover, the mobile continues connected to the same BTS, but changes the channel.
- ***Inter-BTS Intra BSC handover:*** when the mobile station device changes in position out of the one BTS area into another but inside same BSC.
- ***Inter-BSC handover:*** when the mobile's position changes out of the range of BTS controlled by one BSC, now the area of moving is getting bigger, handing over can be from one BSC to another. This kind of the handover is controlled by one MSC.
- ***Inter-MSC handover:*** when changing occurs between networks.

These types of handover controlling can be performed only in GSM. The process it seems simple but it's very complicated.

1- North American Digital Cellular Telephony Standards

Telecommunication Industries Association (TIA) of the Electronic Industries Association (EIA) North America, has developed standard for mobile communication based on Code Division Multiple Access (CDMA) and Time Division Multiple Access (TDMA) technologies.

2- Secure Communication Telephony

Standardization has been organized by the Department of Defense (DoD) in the USA. Some of speech coding standards has been created by DoD, known as U.S. Federal standards has developed for military applications.

Table-1 [4, 7] shows the standard acronym usually followed by number in the first column the organization coming second and its associated bit rate and application coming third and fourth, respectively.

Table- 1 Summary of Organizations and its Standard

Standard	Organization	Bit rate kb/s	Application
G.711 PCM	ITU-T	64	General
FS-1015	DoD	2.4	Secure
GSM 6.10 RPE-LTP	ETSI	13	Mobile radio
G.726 ADPCM	ITU-T	16	General
VSELP	TIA	8	Digital cellular
GSM 6.2 VSELP	ETSI	5.6	GSM cellular
FS 1016	DoD	4.8	Secure
G.728 LD-CELP	ITU-T	16	General
MELP	DoD	2.4	Secure
G.723 ACELP	ITU-T	5.3	Multimedia
G.729 CS-ACELP	ITU-T	8	General
GSM ACELP	ETSI	12.2	General
AMR-ACELP	ETSI	2.4	Secure

2.3 Speech Codecs

- Waveform coders

- ✓ Waveform codecs try to reproduce similar signal to the original signal without using of any kind of knowledge of just how the signal to be encoded was directed respect to the other types of coders. In general they are low intricacy codec's which generate top quality speech at rates above 16 Kbits/s, Whenever the data level is reduced below this stage the reconstructed speech quality that could be received degrades swiftly, an example of this coder is Pulse Code Modulation.

- Source coders

- ✓ Unlike waveform coders, source coders function using a model of just how the source was produced, and try to extorted, from the signal being coded, the parameters of the model. These types of model parameters are sent to the decoder.

- Hybrid coders

- ✓ From the name is hybrid between the two basic coders, hybrid codec seek to benefit from advantages of both coders waveform as well as source codec.

Example of codec used in mobile communication system CELP (Code-Excited Linear Prediction).

Source and hybrid codecs will be studied in more details.

2.4 Speech Coders Attributes

Bit-Rate: is the number of bits per second (bps) which is required to encode the speech into a data stream.

Speech Quality: some indicators are required in order to compare the performance of two speech codecs, for example, the complexity or say the clarity and quality of the speech produced by each coder. The term clarity tells us whether the decoded speech is easily understandable, while the term quality is an indicator if the speech is closer to natural speech sounds or not. It is possible for a coder to yield highly clear speech that is low quality in that the speech may sound very robotic and the speaker is not recognized. Moreover, it is not likely that unclear speech would be called high quality, but there are cases in which perceptually speech does not have high clarity. We discuss the common measures of quality used in formal tests of speech codecs.

MOS: the Mean Opinion Score (MOS) is well-known performance measure. To find a MOS value for a codec, many listeners (≥ 20 people) are asked to judge the quality of the decoded speech in one of five categories:

Table- 2 Mean Opinion Score - MOS

Quality	MOS
excellent	5
good	4
fair	3
poor	2
bad	1

The numbers shown in the Table-2 are used to give a numerical value (cardinal number no fraction) to the subjective evaluations, and the numerical ratings of all listeners are taken as average value of all records to produce a MOS value for the codec. A MOS between 4.0 and 4.5 usually denotes a high quality.

Complexity: The computational complexity of codec (usually measured by Instruction per Seconds) tends to give good quality despite the availability of increasing cost of hardware implementation and other requirement for instance consuming power. Invariably, coders with high complexity can comparably reduce the bit rate required for higher quality.

Memory: some of coders specially sophisticated coders need memory while do processing, so the memory term is related to complexity concept. Template based coders require memory with fast response of large amounts of instructions to deal and store processing history.

Delay: some processing delay is unavoidable in a speech coder. This delay is coming from being complexity of the algorithmic and needs time for computation, also the buffering requirements of the algorithm. For real-time processing long delay is not accepted in speech coders, in order to get acceptable levels of performance the coding delay must be minimized to a level that not detectable by listener's ear.

Error Sensitivity*:* a complicated coders have ability to process more complex algorithms to attain lower bit rates, often add more bits in order to control channel errors. This may distinct difference in the form of noise.

Bandwidth: referred to by the acronym BW, bandwidth of codec is not taken as 'abstract' word but 'respect to' which means it is determined by some other attributes, like speech intelligibility for example in military and quality in mobile telecommunication.

2.5 Performance

Bit Rate vs. Quality

The designer has to determine the demanded level of quality of the speech reconstruction. The codec will attempt to balance the quality of the synthesized

speech with the bit rate of the encoding and decoding. Generally, the quality of the speech declines as the bit rate decreases, so high bit rate leads to high quality, in this case the speech processor contains a sophisticated algorithm that is computationally hard with long encoding delays, i.e., needs using of a DSP board or special audio processor device. So far as it goes, the cost restricts our choices. See Figure 5 [12].

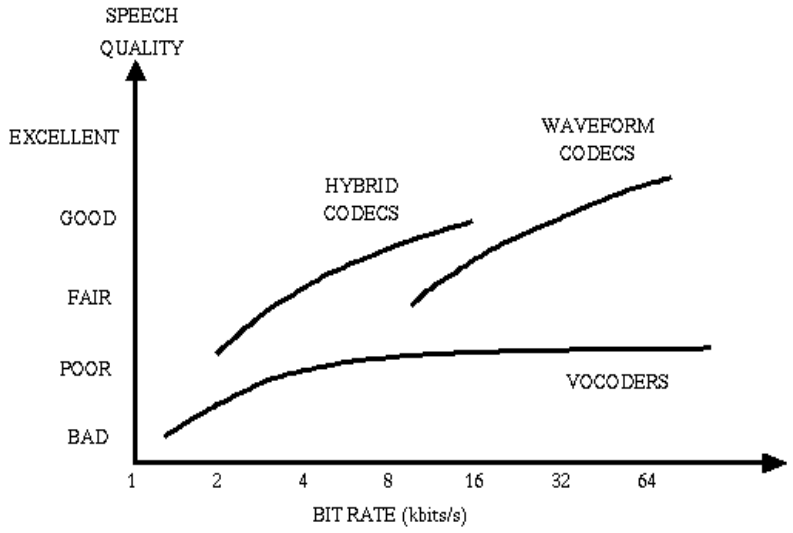


Figure 5 Speech Quality vs. Bit Rate Trade-off

CHAPTER 3: SPEECH CODING BASICS

A speech is an auditory wave that passes on information and facts from a spokesman to an audience. Whenever two persons are at remoteness from one another there needs to be a medium to transmit.

Sound waves are longitudinal waves, they also called compression waves because they produce compression and rarefaction when traveling through a medium they depend upon the activity of atoms which make up the medium (air, water, and so on) for their engendering as opposed to electromagnetic waves that make up their very own means (e.g., photos with regards to lighting), of sound waves for outer physical media. In this case, when there is no means of transmitting, you will have no sound.

3.1 Speech through Air

3.1.1 Speech Production

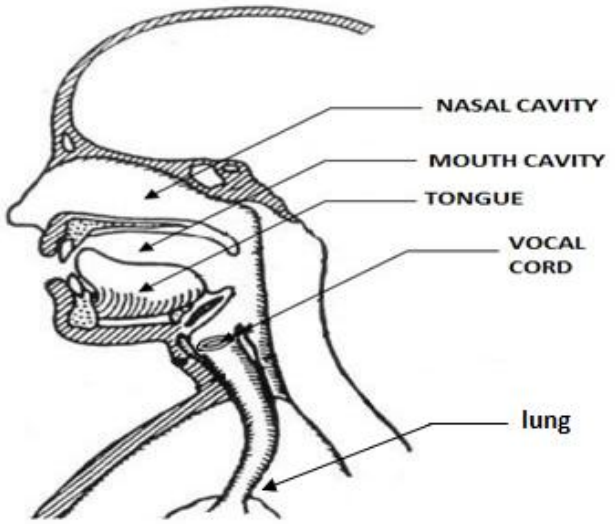


Figure 6 Human Speech Production

The first task in speeches generation is the contraction of the lung tissue which happens to be functioning as an energy source with the aid of the diaphragm as

shown in Figure 6 [13], generating a stream of air via the vocal system. The glottis is the very first organ the stream of air passes through. There, based on the audio uttered, the stream can be created a periodic signal or noise. In case it is periodical, in that case the utterance is considered to be voiced. It follows that, the basis rate of frequency of the signal is known as the pitch. It is connected to the general view of the speaker's tone of voice. It will likely be generally larger in rate of frequency for children, women and therefore much lower for men. Usual qualities for the pitch range between 100-400 Hertz [14].

In case the stream of air is left unscathed, the utterance is considered to be unvoiced. In general unvoiced part comes from sounds without aid of vocal tracts. Furthermore, there is also what called "transition state" which is transition between voiced and unvoiced sounds, in varied ratios, this is an issue itself and has its effects on the quality of speech acquisition [15].

Subsequent the stream extends to the velum. When it is open, sound coupling with the nasal cavity takes place. Once the coupling takes place, the utterance is considered to be nasal: for example, think about the sound of the consonants 'm' and 'n' in the English vocabulary.

Lastly, the tongue, as well as the lips gives fine forming of the air flow so as to generate an extensive variety of sounds [14].

3.1.2 Speech Perception and Detection

The ear comprises of about three major parts, the outer, the middle as well as the inner ears. The outer ear includes the ear lobe (pinna) and the external auditory canal. The work of the ear lobe is always to direct sounds into the ear and also help out with the localization of sounds. The external auditory canals direct the sound into the middle ear. The canal is around 2.7 cm in length and closed off by the ear drum. Therefore, it could be seen as a sound tube that resonates at 3 kHz [16][17].

The ear drum is a hard membrane, around 0.1 mm viscous, which happens to be flexible at the edge (just like the diaphragm of a spokesman). Any time a sound wave hits this membrane it vibrates. This vibration is thereby transmitted to the three bone structure in the middle ear and also from that point to the inner ear. These types of rawhide bones work as a transformer and match up the acoustic impedance of the

inner ear with that of air. Muscle tissues connected to these types of bones reduce the vibration when it is too hostile and so safeguard the inner ear. This proper protection only works best for sounds below 2 kHz but it is not going to work with impulsive sounds. The Eustachian tube links the middle ear to the vocal canal and also eliminates any kind of static pressure difference between the middle ear and the outer ear. If perhaps a substantial pressure difference is found then the Eustachian tube opens and the difference is taken out [16].

The inner ear comprises the Semicircular canals, the Cochlea, as well as auditory nerve terminations. The functionality of the Semicircular canals is always to control balance. The Cochlea is liquid packed and also helical in shape (it looks like the protective covering of a snail). Inside the Cochlea there exists a hair-lined membrane known as the Basilar membrane. This membrane transforms the mechanical sign into a neural sign. Several frequencies are loved by various parts of this membrane allowing a rate of recurrence evaluation of the signal to be performed. As a result the ear is basically a array analyzer that reacts to the significance of the signal. The rate of recurrence decision is most effective at lower frequencies.

Whenever the nerve and skin stimulation appear in the human brain through the acoustic nerve, the substantial neural action already taking place is reinforced by the nerve stimulation from the ear. This customization of human brain action results in identification and understanding of the speaker's statement.

3.1.3 Speech Conversation and Communication

Knowing a spoken utterance believes that the hearing organ of listener is good enough which in turn the spoken sounds are effectively regarded as phonemes of English or regardless of what word is spoken. Phonemes are the most compact units of spoken language which make a big difference to meaning – matching roughly to the letters in a word (e.g., the sounds that a, r, and t produce in the phrase ***art***). Auditory processing of language likewise consists of the ability to integrate the standalone sounds of a phrase into the viewpoint on a thoughtful word and of sequences of significant words.

Knowledge of Word Organization

- *Syntax* (or sentence structure) represents the guidelines that govern the organization of words in a sentence or perhaps utterance. Understanding an utterance demands a prowess to discover the meaning implicit in the organization of words. For instance, a sentence having different meanings in spite of having the exact same words.
- *Morphology and lexical* (a part of grammar) represents guidelines that govern meaning found in the structure of the words by themselves. Alterations within words (e.g., adding an ***'s'*** to ***lion*** to get plural ***lions***, or derive past simple by adding an ***'ed'*** to ***ask*** to get ***asked***) influences meaning. Understanding an utterance demands a prowess to discover the meaning related to this kind of alterations of the words.

For speech coding needs, the physiological generation of speech patterns is broken into two steps. The very first one is excitation, which can either comprise of white noise, according to non-voiced sounds, or a periodical signal at a sufficient rate of frequency, in accordance with voiced sounds, or a blend of both. The other one is the vocal canal, which can be regarded like a filter transforming the excitation. For the needs of vocoding, the vocal canal could be regarded as one time varying filtration system.

Hence encoding will include in taking out the three basic parts of info from the speech : the voicing info, showing the pertinent quantity of voiced and unvoiced signals in the excitation, the pitch info, in case the voiced signal exists in the excitation, and the vocal canal filtration parameters.

3.2 Speech through Channel

Characterize the signal via the channel differs from the signal through air, it has own environment and standards which change from coder to coder. Let’s have a look of existing coders.

3.2.1 Forms of Speech Coders

Waveform Coders

An example of waveform coder is Pulse Code Modulation (PCM), in this type many samples are taken directly from original signal and converted them to a digital bit stream to reproduce an encoded signal, as a result waveform codecs don't study features or have a knowledge of that signal.

Another examples of this class is Adaptive Differential Pulse Code Modulation ADPCM, the block diagram of ADPCM is shown below [18]:

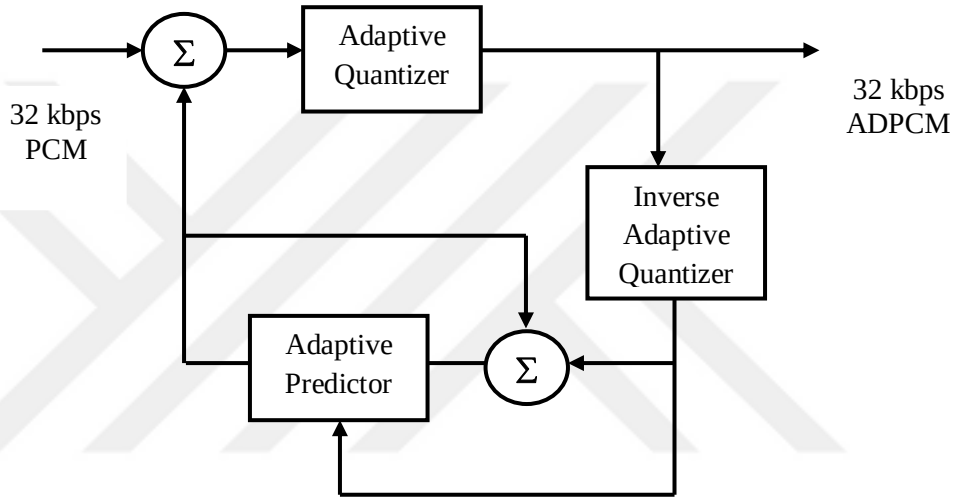


Figure 7 ADPCM Encoder

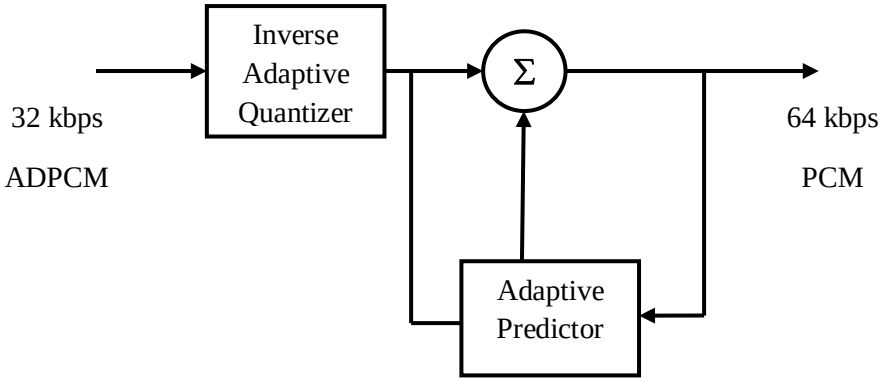


Figure 8 ADPCM Decoder

These coders are more suitable for high bit-rate coding, due to performance drops noticeably with decreasing bit-rate as shown in Figure 5, performance of these coders is getting better as the bit-rate greater than 32 kbps.

Parametric Coders

Focusing on the characteristics of the signal, studying it and find the best parameters during encoding which can represent the original signal at receiver side. The goal of this type is not preserve the original signal but to preserve the bandwidth as much as possible so that users can use bandwidth portion efficiently. The alternative name of this coder is vocoder, this name came from that was being designed for a voice (human sound) signal and is not suitable for other signals, see Figure 5**.**

An example of this type of coders is LPC-10. The block diagram is shown below:

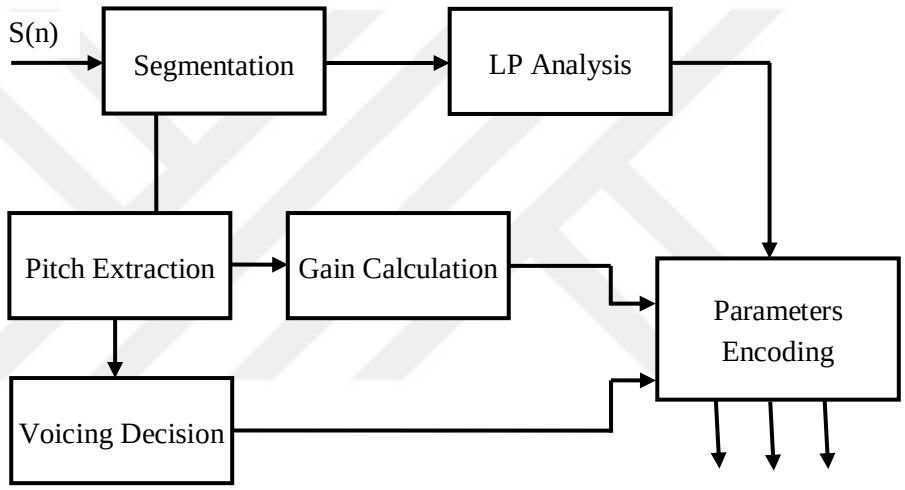


Figure 9 Block Diagram of LPC Encoder

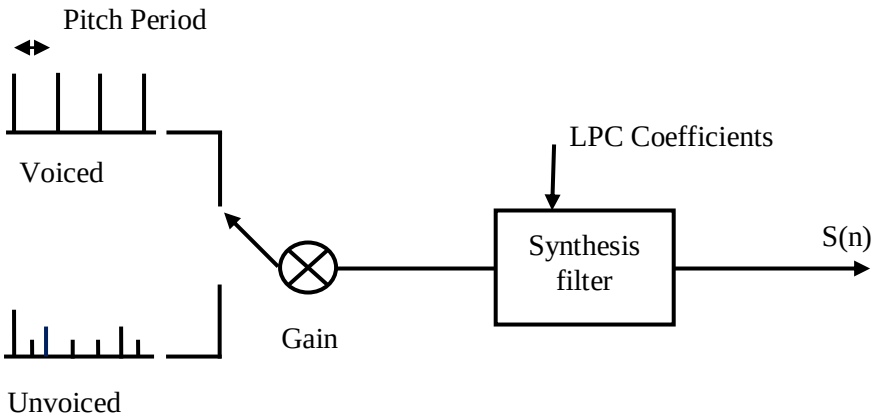


Figure 10 Block Diagram of Decoder

The vocal channel is considered as a time-varying filtration which is excited with either a white sound source, for unvoiced speech segments, or perhaps a train of pulses split up by the pitch period for voiced speech, thus, as opposed to the waveform coding the info which have to be delivered to the decoder is the filter parameters, a voiced/unvoiced flag, the necessary variation of the excitation signal derived from pitch period for voiced and noise for unvoiced part of speech.

Hybrid Coders

A hybrid coder is combination of a waveform coder with that of a parametric coder. Like a parametric coder, it strongly depends on how can human produce speech?. Additional parameters of the production model are optimized in such a way that the decoded speech is as close to the original waveform as possible by using a loop with perceptual weighted error minimization.

From Figure 5, this class at medium bit-rate coders, have best quality among other codecs.

The code-excited linear prediction (CELP) algorithm as example is hybrid codec, the big difference between a hybrid coder and a parametric coder is that the hybrid one tries to represent the residual signal to the speech production model, which is transmitted as part of the encoded bit-stream. The latter, however, achieves low bit-rate by discarding all detail information of the excitation signal; only coarse parameters are extracted.

At high bit-rate a hybrid coder behavior goes like a waveform coder, but like a vocoder at low bit-rate, with good quality for medium bit-rate.

Though other models of hybrid codec really exist, the most prosperous and widely used are time domain Analysis-by-Synthesis (Abs) codec, and also Code-Excited Linear Prediction AbS family member is considered as hybrid coder.

The block diagram of AbS codec is shown in Figure 11 [18]:

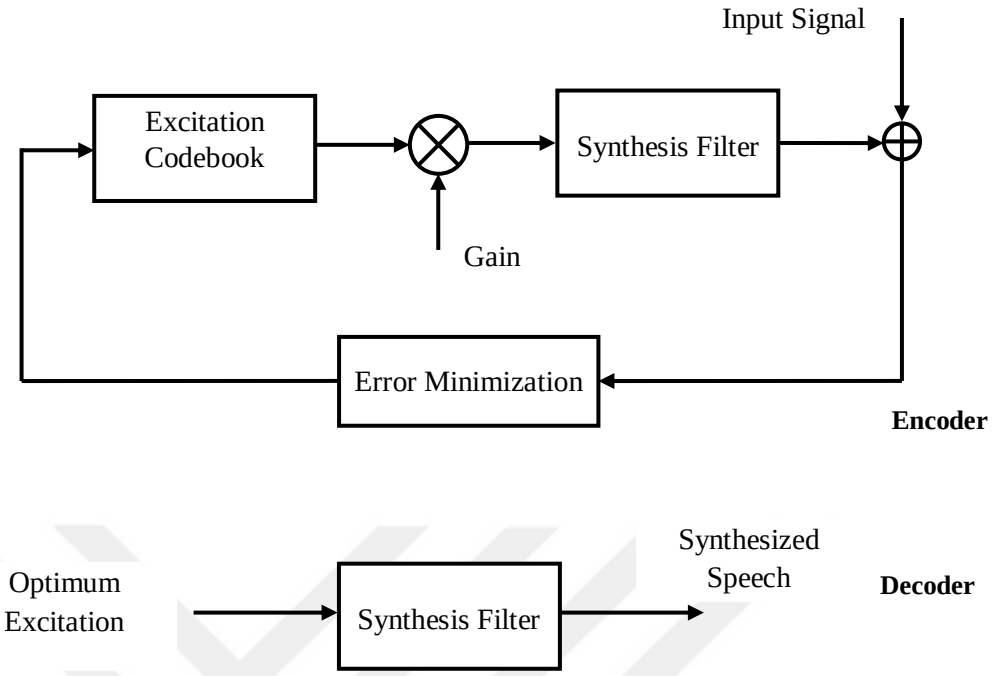


Figure 11 Block Diagram of AbS Codec

3.3 Linear Prediction

Linear prediction is a method of time series analysis, that develops from the scrutiny of linear systems. Utilizing linear prediction, the structures of a such a system can be resolute by analyzing the systems outputs and inputs.

A linear system acquires its output as a linear amalgamation of its previous and current inputs. It can be labeled as time-invariant if the system parameters don't change as time goes by. Statistically, LTI (linear time-invariant) systems can be symbolized by the subsequent equation [19]:

$$y(n) = \sum_{j=0}^{q} b_j s(n-j) - \sum_{k=1}^{p} a_k y(n-k) \qquad (3)$$

This is the common difference equation for all linear systems, with the input signal s and output signal y, and scalars a_k and b_j, fork = 1 . . . p and j = 1 . . . q where the max of q and p is the order of the structure.

By arranging equation 1 in another format and transforming to the Z-domain, It is possible for us to reveal the transfer utility H(z) of such system:

$$y(n) + \sum_{k=1}^{p} a_k\, y(n-k) = \sum_{j=0}^{q} b_j\, s(n-j) \qquad (4)$$

$$\sum_{k=1}^{p} a_k\, y(n-k) = \sum_{j=0}^{q} b_j\, s(n-j) \qquad where\ a_0 = 1$$

By taking z-transform:

$$\sum_{k=1}^{p} a_k\, z^{-k} Y(z) = \sum_{j=0}^{q} b_j\, z^{-j}\, S(z) \qquad (5)$$

$$H(z) = \frac{Y(z)}{S(z)} = \frac{\sum_{j=0}^{q} b_j\, z^{-j}}{\sum_{k=1}^{p} a_k\, z^{-k}} \qquad (6)$$

The coefficients of the output and input signal examples in equation (3) disclose the zeros and poles of the transfer function.

Linear estimate follows logically from the general arithmetic of linear schemes. As the output of the system is defined as a linear grouping of past samples, the output of the system's future can be anticipated if the scaling coefficients a_k and b_j are known. These scalars are known also as the analysis coefficients of the system.

From equation (3), you can verbalize the equations required to regulate the parameters of an all-pole linear system, which is also called the linear estimated normal equations. First, using the all-pole method, a linear prediction appraisal of ˆy at illustration number n for the output sign of y and a p^{th} order estimate filter which can be classified by:

3.3.1 Speech Production as a Linear System

Glottal Source:

The source sign is among two claims: a pulse train of a particular basic frequency for voiced sounds and also white noise for unvoiced sounds Figure 12. This two-state source suits fairly well with genuine glottal actions, even though a moment of blended excitation is never represented very well.

Vocal Tract Filter:

The vocal canal is parameterized by its resonances, which can be referred to as *formants*. Every sound tube possesses normal resonances, the variables which are element of its shape.

In spite of the fact that the vocal canal changes its shape, therefore it is resonances, persistently with running speech, it is not irrational to accept it fixed over short-time intervals of the order of 20 milliseconds. In this manner, speech patterns generation could be seen as a LTI system and linear prediction can be applied to it.

The exact source-filter model used for LPC-10 is referred to as the linear predictive coding model. They have two fundamental parts: evaluation or encoding and synthesis or decoding. The encoding section of LPC-10 entails checking the speech sign as well as cracking it into sections. Every section is then checked further to find the solutions to many key inquiries: Is the section voiced or unvoiced? What exactly is the pitch of the section? What parameters are required to build a filtration system that models the vocal canal for the present section?

LPC-10 encoding is normally carried out by a sender who replies these types of questions and in most cases send all these answers onto a receiver. The recipient carries out LPC-10 synthesis through the use of the replies obtained to develop a filtration that once supplied the right input source can correctly recreate the original speech sign. Basically, LPC-10 synthesis attempts to mimic human being speech generation. Figure 13 shows exactly what parts of the receiver correspond to what components in the human body structure. This diagram is made for an overall tone of voice or speech coder which is not exact to linear predictive coding. Almost all tone of voice coders are inclined to model two things : excitation and articulation.

Excitation is the form of sound which is passed into the filtration or vocal canal while articulation is the modification of the excitation sign into speech.

3.3.2 Speech Production Model

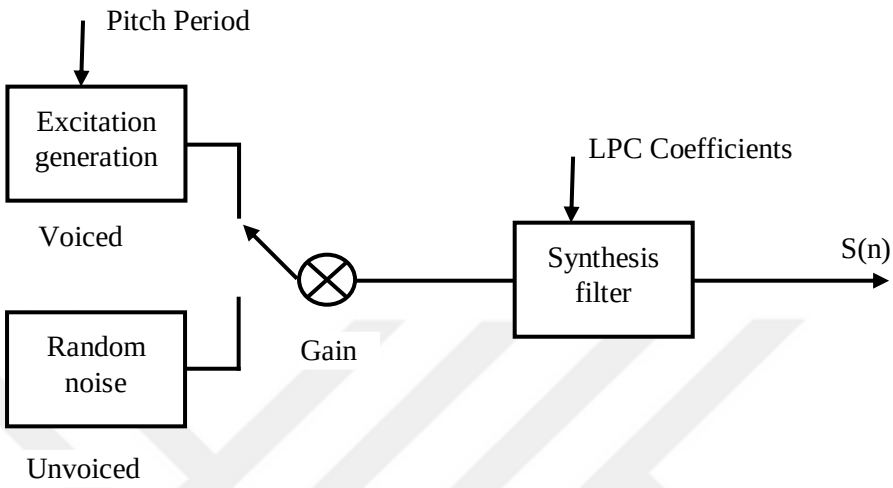


Figure 12 Block Diagram of Speech Production Model

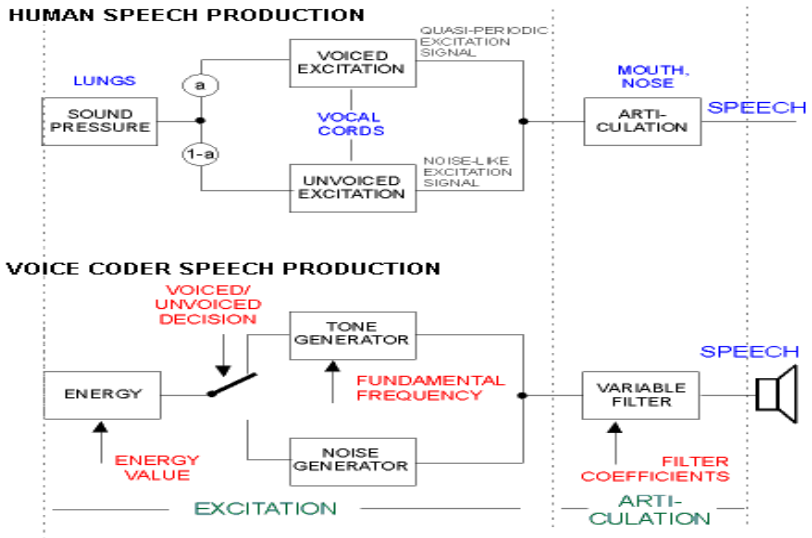


Figure 13 Speech Production vs. Human Production Model

3.3.3 Linear Predictive Coding of Speech Signal

Speech Analysis

We try to calculate model parameters α_k that reduce the prediction error e(n), this can be obtained by what called Least Square minimization method that intends to minimize the total residual energy $e^2(n)$. There are two strictly relevant methods are normally used :

1- Autocorrelation method
2- Covariance method

In comparison, each of these methods has its own particular qualities and shortcomings; figuring out which is more beneficial to use is incredibly determined by the signal being dissected.

Short-Term Processing

The speech signal in nature is non-stationary, therefore, in order to make the analysis applicable; the signal is divided into frames of length 10-30ms this called short-term this is a very important term before going further to any kind of analysis, it is actually present parameters of entire speech signal not in one value but many values. Furthermore, in sophisticated coders they divide each frame into subframes in length of 5ms.

Autocorrelation Method

The expected signal $\hat{s}(n)$ could be depicted below [14]:

$$\hat{s}(n) = -\sum_{k=1}^{P} \alpha_k . s(n-k) \qquad (7)$$

In which α_k's are the linear prediction coefficients and s(n) is the windowed speech series received by multiplying short-period speech frame with a hamming or perhaps comparable form of window that is provided by:

$$s(n) = x(n).w(n) \qquad (8)$$

In which w(n) is the windowing series. The predictive error e(n) is estimated by the difference between original sample s(n) and the expected sample $\hat{s}(n)$ that is given by :

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{P} \alpha_k . s(n-k) \qquad (9)$$

$$e(n) = s(n) + \sum_{k=1}^{P} \alpha_k . s(n-k) \qquad (10)$$

In reality the error e(n) is considered to be zero and the coefficients corresponds to a_k, so we trying to find α_k to be as much as possible closer to a_k which means e(n) is closer to zero.

The widely used method of calculating the LP coefficients by least squares auto correlation approach. This is done by reducing the overall prediction error. The overall prediction error could be depicted below:

$$E = \sum_{n=-\infty}^{\infty} e^2(n) \qquad (11)$$

This can easily be broadened using the formula provided below:

$$E = \sum_{n=-\infty}^{\infty} [s(n) + \sum_{k=1}^{P} \alpha_k . s(n-k)]^2 \qquad (12)$$

The rates of α_k which reduce the overall error E could be calculated by finding

$$\frac{\partial E}{\partial a_k}$$

And then make it zero for k=0, 1, 2...p.

$$\frac{\partial \mathrm{E}}{\partial \mathrm{a_k}} = 0$$

The solution which lead to find the LP coefficients. This is often depicted below:

$$\frac{\partial \mathrm{E}}{\partial \mathrm{a_k}} = \frac{\partial}{\partial \mathrm{a_k}} . \sum_{\mathrm{n}=-\infty}^{\infty} [\mathrm{s(n)} + \sum_{\mathrm{k}=1}^{\mathrm{P}} \alpha_{\mathrm{k}} . \mathrm{s(n-k)}]^2 = 0$$

The yielded differential equation could be formulated in the form of:

$$\sum_{n=-\infty}^{\infty} s(n-i).s(n) = \sum_{k=1}^{P} \alpha_k \sum_{n=-\infty}^{\infty} s(n-i).s(n-k)$$

Where i=1, 2, 3...p.

This formula could be adjusted to be a related of autocorrelation series R(i) depicted below :

$$\sum_{k=1}^{P} \alpha_k R(i-k) = R(i)$$

For i=1, 2, 3...p.

In which the autocorrelation series used in earlier formula could be written given below

$$R(i) = \sum_{n=i}^{N-1} s(n)s(n-i) \qquad (13)$$

For i= 1, 2, 3...p and N denotes to a particular frame in a period of time series. This is often depicted in the matrix form given below,

$$R.A=-r$$

In which R is the p×p symmetric matrix of components *R(i, k) = R(|i-k|), (1<=i, k<=p)*, r is a column vector with components *(R(1),R(2),...R(P))* and lastly A is the column vector of LPC coefficients (a(1), a(2),....a(p)). It could show R is Toeplitz matrix where the diagonal does not change, which could be depicted as:

$$R = \begin{bmatrix} R(1) & R(2) & R(3) & & R(P) \\ R(2) & R(1) & R(2) & & R(P-1) \\ R(3) & R(2) & R(1) & \dots & R(P-2) \\ \vdots & \vdots & \vdots & & \vdots \\ R(P) & R(P-1) & R(P-2) & & R(1) \end{bmatrix}$$

Finally, the LP coefficients could be calculated as outlined:

$$\mathbf{A} = -\mathbf{R}^{-1}.\mathbf{r} \qquad (14)$$

There is certainly an algorithm used to fix this type of matrix and it is known as Levinson-Durbin algorithm.

Pitch Period Estimation

Speech signal can be classified into voiced, unvoiced and it can be also silence regions. The periodic part assigned to the voice part of speech contain the pitch period which is required to build excitation for the reproduction of voiced speech at receiver side. The noise-like excitation is present for unvoiced speech. There is no excitation during silence region so can be ignored or perhaps removed.

For precise decision its preferable to find pitch period before classification step. But according to classification of voiced and unvoiced even the silent region each of them has own characteristics, one of these characteristics is the pitch period of voiced part is larger compared to unvoiced part, from here pitch period can be considered as classifier.

The periodicity associated with each segment or frame is defined as 'pitch period ' in the time domain and 'Pitch frequency or Fundamental Frequency F_o' in the frequency domain.

There are many estimators used to find the fundamental frequency or pitch period:

- Time domain – Autocorrelation

  It's a convolution between one frame and itself, once the convolution has been done the output is symmetric signal with max value at origin and other few max values. The pitch period is the lag at where the first maximum occurs.

- Frequency domain – Cepstrum

  The cepstrum of speech is defined as the inverse Fourier transform of the log magnitude spectrum.

Classification

Analyzing Speech Signal

The voiced speech section is described by the quasi-periodic, high energy, less figure of zero crossings and a more connection among succeeding samples. The existence of formants configuration evident that the section is voiced in the frequency domain,. Furthermore, the spectrum is going to have more energy, naturally, within the formants area. The autocorrelation of a section of voiced speech is going to have a resilient peak at the pitch period. The great energy can be taken into consideration in terms of high values for voiced part. However, only energy level can't choose the voicing info. Energy and explicit periodicity together are indisputable evidence in identifying the voiced part. Likewise the reasonably low zero-crossings could be indirectly visualized as a smooth variation among series of sample values. Figure 14 [word 'this'] shows an example of voiced/unvoiced parts. This Figure shows the distinctive nature and should be carefully observed so as to point whether a section of the speech is voiced or not, also other case of mixing or transforming from voice to unvoice (and vice versa) called transition.

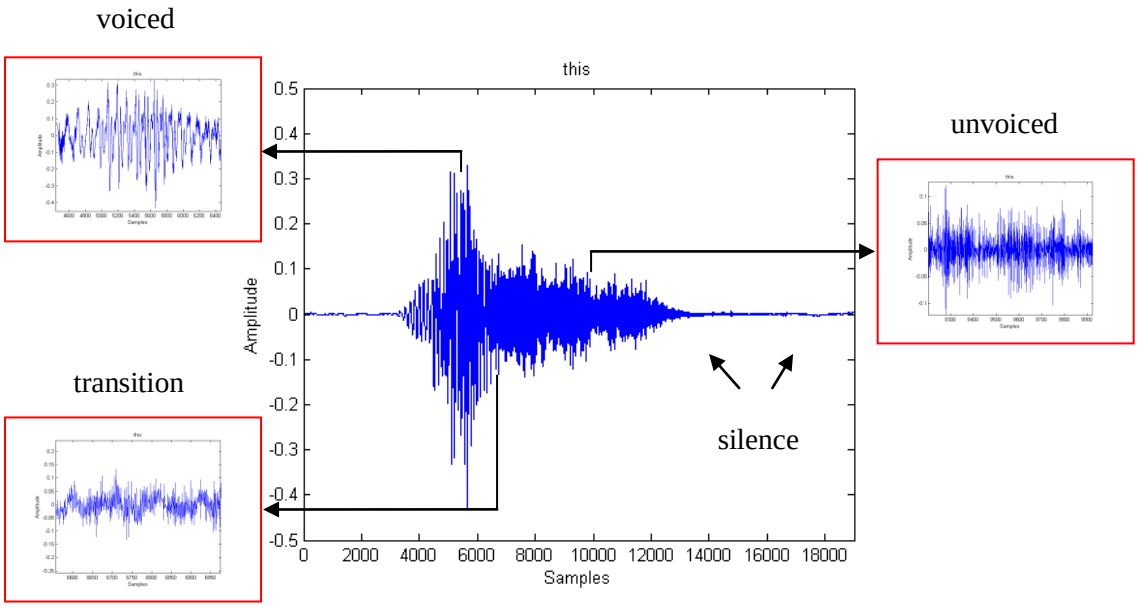


Figure 14 Classification of Speech Signal

At same time, the unvoiced speech section is characterized by being noise-like and aperiodic, fairly low energy likened to voiced speech, a lot of zero crossings and

somewhat less correspondence among succeeding samples. The unvoiced part could be known by visualizing the speech signal within the time domain because of its noise similarity and aperiodicity features, . It is good to mention that in the frequency domain, the nonexistence of formants is evident that the section is unvoiced. Furthermore, the spectrum is going to have an upward trend beginning from zero rate of frequency and surging upwards. The autocorrelation of a section of unvoiced part is going to have components of random noise beside some other information. The low-trend energy is going to be visualized via low values for unvoiced section. However, with only energy you cannot come to a decision of unvoicing data. The figure of zero-crossings remain necessary with the energy to recognize the unvoiced section. The high zero-crossings could be indirectly perceived as rapid deviations among a series of sample values. Figure 14 shows an example of unvoiced part of speech signal.

Moreover, the silence region is assorted by the nonexistence of any signal or lowest energy likened to voiced or unvoiced speech sections. The silence part can be easily recognized by visualizing the waveform within the time domain because of the nonappearance of signals. The autocorrelation of a section of silence part won't have any info. However, energy seems to be a good pointer for recognizing the silence parts. It is shown in Figure 14.

CHAPTER 4: ANALYSIS BY SYNTHESIS AND CODE-EXCITED LINEAR PREDICTION

The LPC-10 codec (discussed in Chapter 3) have no mechanism to check for an improvement on the estimated parameters before sending them to the receiver side, because of that the quality of such coders suffer in quality and intelligibility tends to be robotic or monotone sounding. In the same way Analysis and Synthesis goes, which mean the estimated parameters may not work properly.

Analysis-and-Synthesis AaS is a combination of two processes together named encoding and decoding performed in closed loop in order to find the optimum parameters [15]. At the minimum error the parameters are sent, as a result AbS gives a higher quality compared to LPC-10 and AaS codecs.

4.1 Analysis by Synthesis – (AbS)

The main difference between basic form of Linear Predictive Coding or alternative LPC-10 form AaS Analysis and Synthesis, and AbS can be expressed as a sentence "just send I'm right " in AbS "I'm wrong don’t send". These implies meaning that in LPC-10 the calculated parameters found by particular method are correct, so send them to receiver side, but AbS don't send any parameter until making sure that this parameter is the best available to send.

4.1.1 Block Diagram of AbS Codec

The block diagram of AbS encoder and decoder are shown in Figure 15 and Figure 16, respectively.

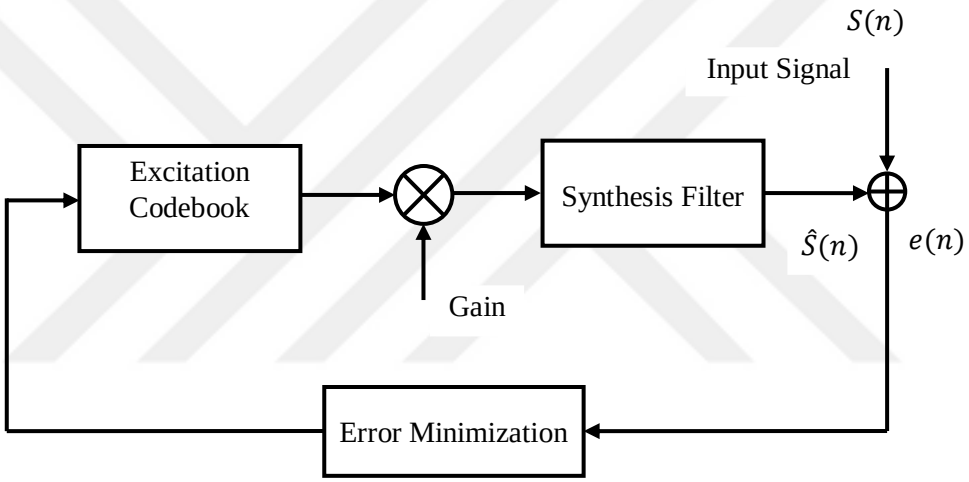


Figure 15 Block Diagram of AbS Encoder

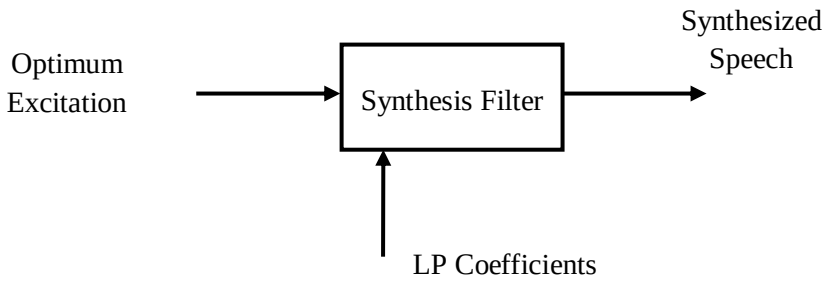


Figure16 Block Diagram of AbS Decoder

AbS algorithm

1- Input speech signal is segmented into frames of length 20-30 ms.
2- Apply hamming widow for each frame.
3- Find LPC-10 coefficients for each frame, then convert them to reflection coefficients for transmission purpose.
4- For every frame and with MSE Mean Square Error criterion find the best gain and pitch parameters by searching inside closed loop.

LPC analysis

The spectral envelope of speech is found by LPC analysis filter which represented by all-poles filter of usually block of signal or frame. This is can be achieved by:

$$\hat{s}(n) = \sum_{i=1}^{p} \alpha_i . s(n-i) \tag{15}$$

$$A(z) = 1 + \sum_{i=1}^{p} \alpha_i z^{-1} \tag{16}$$

Where A(z) is transfer function of analysis filter also called FIR - Finite Impulse Response filter.

LPC Synthesis

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 + \sum_{i=1}^{p} \alpha_i z^{-1}} \tag{17}$$

Where H(z) is transfer function of synthesis filter also called IIR filter - Infinite Impulse Response filter.

4.1.2 The Prediction Error Signal

The difference between original produced signal by human and synthesized signal usually not zero, this difference is e(n) prediction error. All codecs aim to make e(n) go to zero, good approximation is happening when e(n) goes to zero.

All parameters found in closed loop with MSE criterion.

4.1.3 Pitch Period Search

The pitch period has been found by implementing a codebook which contains an expected pitch periods of any speaker covered inside range.

4.2 Code-excited Linear Prediction – (CELP)

CELP Algorithm:

1- Input speech signal is segmented into frames of length 20-30 ms, in order to get closer to stationary form.
2- Find LPC-10 coefficients for each frame, then converted them to reflection coefficients.
3- Subdivide each frame into blocks of 5ms, so that each frame will be a group of 4 blocks.
4- For each subframe:
    a- Initialize the contents of filters.
    b- Find the lag (pitch period) that minimize the difference (e) between original and synthesized signal by testing all possibilities of pitch inside range [16 160] samples this is called adaptive codebook.
    c- Find the best index *k* of Gaussian codebook that also minimize the difference (e).
    d- Filter the codebook vector corresponding to the index *k* found in step b by pitch periods found in step a.
5- Filter the vector resulted in step c with LPC-10 coefficients found in step 2.
6- Repeat steps from 2 – 5 for each frame.

4.2.1 Block Diagram of CELP Codec

CELP Encoder

See Figure 17, it shows the path that the signal must go through, although some figures don't display an analysis stage, there is an analysis stage .

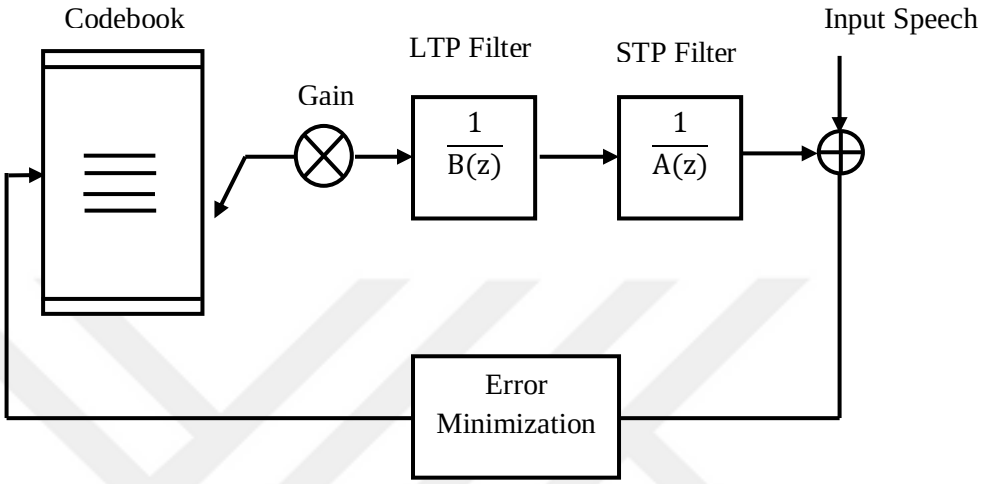


Figure 17 Block Diagram of CELP Encoder

CELP Decoder

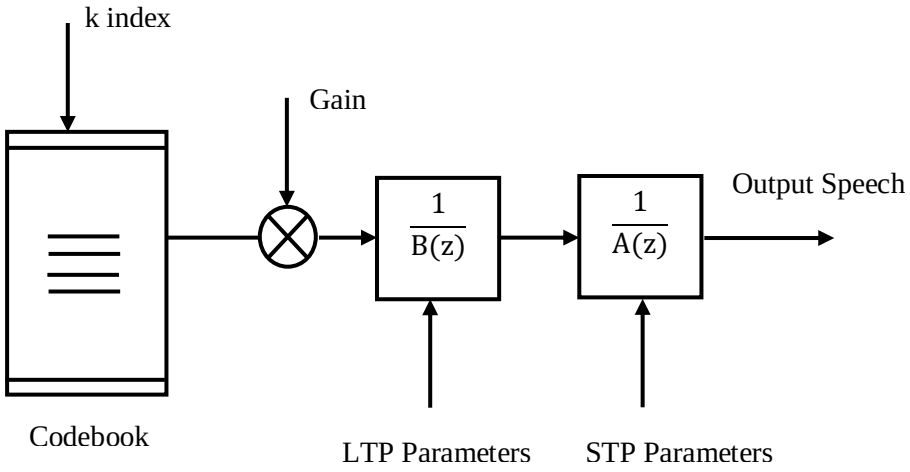


Figure 18 Block Diagram of CELP Decoder

The CELP decoder is represented by the block diagram shown in Figure 18, it is simpler than the encoder.

The prediction error or residual signal behavior has random components beside other information like pitch period, and pitch gain, the following stages aim to remove all information and match the remaining random components with one of codebook vectors:

4.2.2 Short-Term Prediction (STP)

The STP stage can be divided into two steps:

LPC analysis and LPC Synthesis, both are similar what have been done in LPC-10 and AbS.

STP aims to find LP coefficients from original signal, then filtering the candidate codebook vector by these coefficient.

4.2.3 Long-Term Prediction (LTP)

The CELP codec from fact that by removing any information from interest signal we will get a white noise this noise can be exist somewhere inside codebook or at least best match so it depends on size of used codebook, a solution to deal with this signal by removing it's information first by LTP and then by STP at node (d) in Figure 19 of original and synthesized difference signal, the result is depending on how efficiently processed is a white noise.

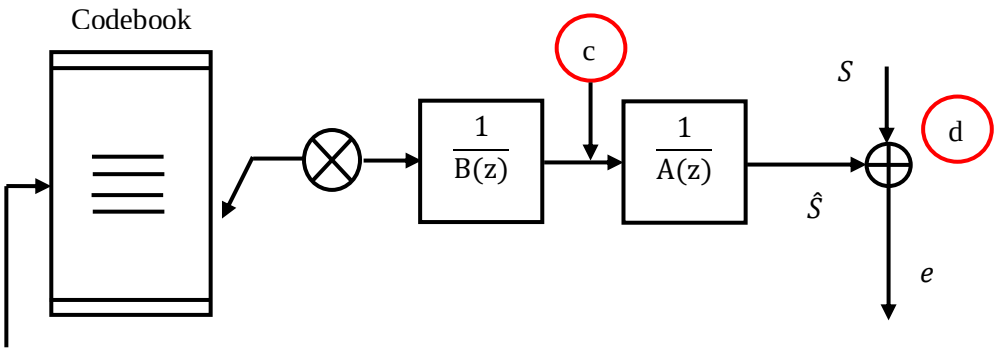


Figure 19 Processing Direction of Two Cascade Filters

At node c in Figure 19 the excitation e(n) will be the sum of the pitch predictor and the signal c(n) taken from codebook amplified by gain G, the final excitation is given by:

$$e(n)=g.c(n)+b.e(n-P) \qquad (18)$$

where g is gain and b 'pitch gain' coefficients which reflect the amount of correlation between the distant samples [7].

The mean-squared error between an LPC residual (containing pitch), and the reconstructed pitch signal resulting from the analysis is used to find pitch and pitch gain.

$$b = \frac{\sum_{n=0}^{N-1} e_S\,(n) e_S\,(n-P)}{\sum_{n=0}^{N-1} e_{STP}^2\,(n-P)} \qquad (19)$$

let's state the situation, from point a toward point b we actually rebuild a signal $\hat{s}$ similar to the original signal s by passing a random signal through cascade of two synthesized filters.

The total excitation e is derived from two excitation one is coming from a codebook called adaptive codebook containing all possibilities of pitch period in range [16 160], so there is no actually pitch estimation by mean estimation like what have done in LPC-10 vocoder.

The other excitation is from a random codebook. Moreover, the excitation is performed by:

- Adaptive codebook: the parameters are found by cross correlation between input signal and the codebook of pitch period.
- Gaussian random codebook: the needed parameters are found by cross correlation between excitation derived from adaptive codebook and Gaussian codebook.

So, the block diagram of encoder shown in Figure 17 can be converted to an equivalent block diagram shown in Figure 20 and Figure 21 [20]:

4.2.4 Codebook Search

Figure 17 shows the coding procedure of the encoder. First performs a linear predictive analysis of the input speech signal, we find the linear prediction coefficient. This will converted into the LSP Line Spectral Pairs or reflection

coefficients for reliable transmission. Next pitch analysis, carried out a codebook, the search for gain codebook turn, finds the index as speech synthesized by the prediction coefficients quantized is closest to the input signal.

Components selection are performed in a closed loop to minimize the error signal between the synthetic and original signals

One can ask, which operation is performed before STP or LTP?

The answer of this question is LTP, why?

Because the STB does not affect on candidate codebook vector found by LTP.

Adaptive Codebook Search

First performs a linear predictive analysis of the input speech signal, we find the linear prediction coefficient, the weighted LPC it has to be computed as well.

By doing cross correlation between weighted input signal and filtered signal with pre-selected pitch range [16 160], we find the parameters b, P and these values are determined to minimize the error energy.

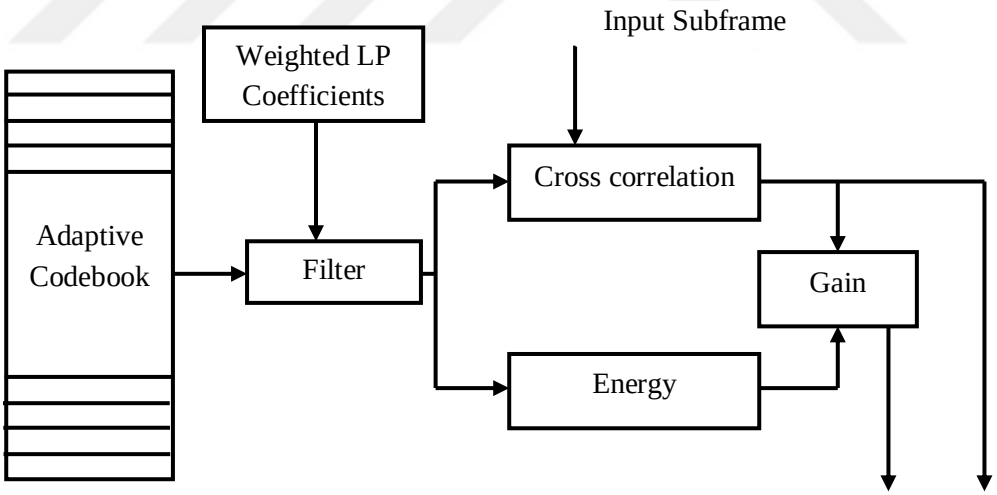


Figure 20 Adaptive Codebook Search Method

Stochastic Codebook Search

Codebook can be either “deterministic” or “stochastic” codebook-10 bits are common, where deterministic codebooks are derived from a training set of vectors and stochastic codebooks the residual from the long-term predictor roughly is Gaussian.

Once the parameters has been determined then we find the signal e(n-P) based on the computed b pitch gain and P pitch lag.

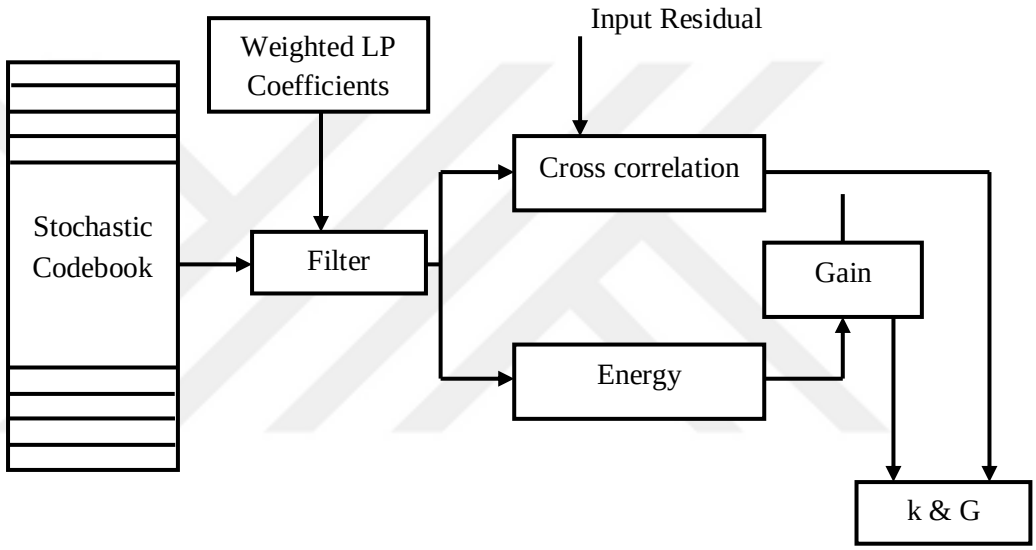


Figure 21 Stochastic Codebook Search Method

From above Figure, the other parameters k-index and g are determined to minimize the error energy between the filtered signal "e1" and Gaussian codebook

4.2.5 Perceptual Weighting Filter

The Perceptual Weighted Filter is used to reduce the spread of speech formants in general to reduce noise in lower and higher frequencies [15].

The difference between the original $s(n)$ and the synthesized signal $\hat{s}(n)$ is the error signal $e(n)$ in Figure 19. that will be minimized by improving or modifying the excitation signal. In the way of minimization, error signal is spectrally weighted to support the important frequencies.

One method weighting filter, it is giving by:

$$W(z) = \frac{A(z)}{A(z/\gamma)} \qquad (20)$$

$$= \frac{1 - \sum_{i=1}^{p} a_i z^{-i}}{1 - \sum_{i=1}^{p} a_i \gamma^i z^{-i}} \qquad where\ 0 \le \gamma \le 1$$

The coefficients of the filter $A(z/\gamma)$ are $a_i\gamma^i$ which can be seen from

$$A(z/\gamma) = 1 - a_1(z/\gamma)^{-1} - \cdots - a_M(z/\gamma)^{-M}$$

$$= 1 - (a_1\gamma)z^{-1} - \cdots - (a_M\gamma^M)z^{-M} \qquad (21)$$

The effect of the factor γ does not change the centre formant frequencies but just broadens the bandwidth of the formants by : [21]

$$\Delta f = -\frac{f_s}{\pi} \ln \gamma \qquad (22)$$

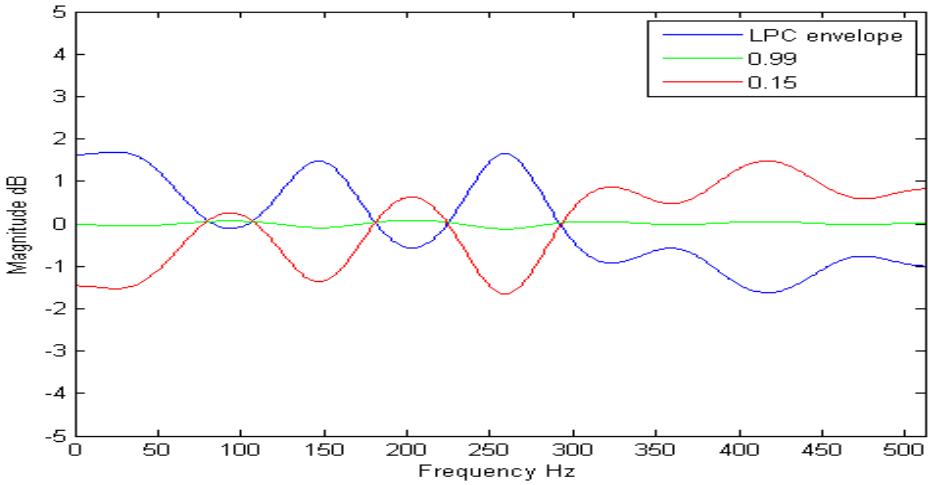


Figure 22 Weighting Filter Spectrum (γ=0.99,0.15) Compared with the Original

CHAPTER 5: SIMULATIONS

In this chapter we explain the details of the algorithms of the three speech coding techniques namely; LPC-10, AbS and CELP with flowcharts and give the simulation results.

5.1 Part1: Linear Prediction Coding (LPC-10)

Implemented Matlab code is obeying the block diagrams in Figure 23 and Figure 24:

Encoder

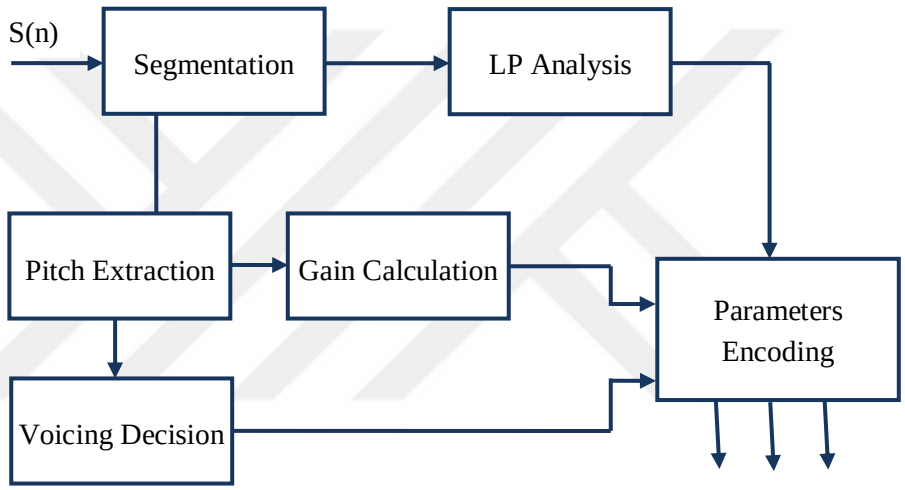


Figure 23 Block diagram of used LPC encoder

Decoder

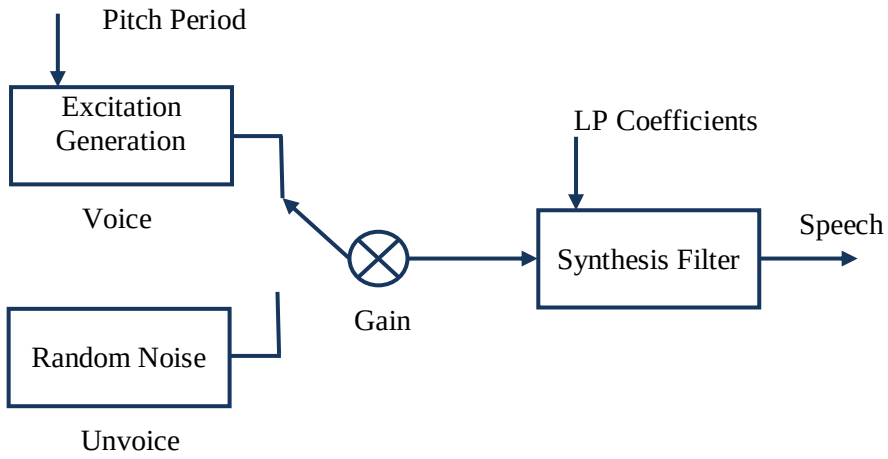


Figure 24 Block Diagram of Used LPC Decoder

LPC-10 - Flow Chart

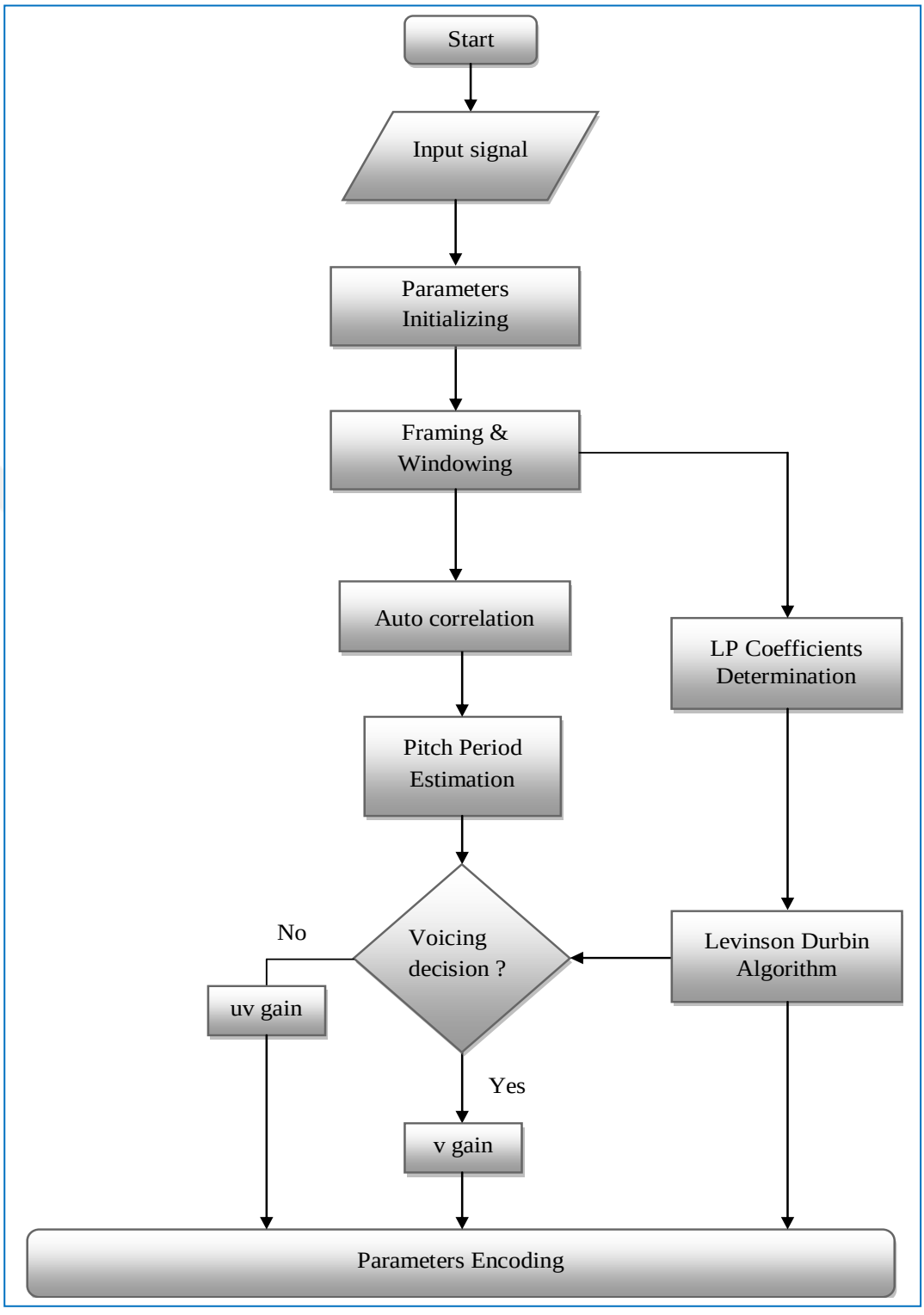


Figure 25 LPC Flow Chart

Input Signal

Speech signal wideband .wav format sampled at 16kHz is chosen.

Segmentation

This stage includes framing, windowing LPC analysis, it's general for all codecs.

Pitch Period Estimation

There are many methods used to find pitch period such as:

- Autocorrelation method
- AMDF Average Magnitude Difference Function

Autocorrelation method is actually most wide used to estimate the pitch periods of short parts of speech signal, where the processing is done in time domain.

Classification

Many parameters are denoted to classification voice and unvoice process:

- **Autocorrelation coefficient at unit sample delay**

Briefly the correlation between neighbored speech samples, it is known that the neighbored samples of voiced speech waveform are highly correlated and this make the parameter C1 as shown in Figure 26 to be close to 1. On the other hand, the correlation parameter is closed to zero for unvoiced speech.

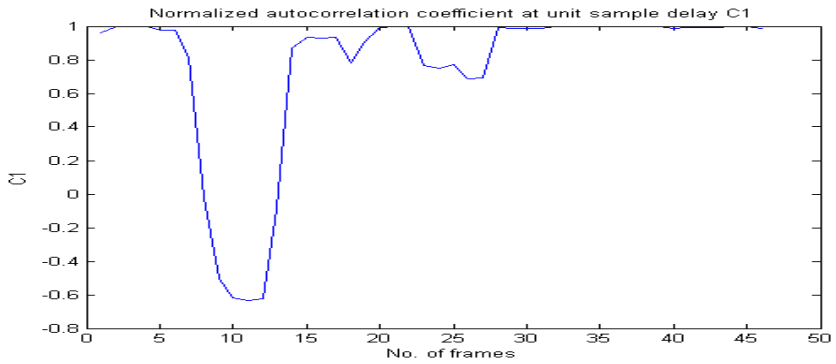


Figure 26 Autocorrelation Coefficient at Unit Sample Delay

Figure 26 shows the process of classification by taking a threshold level where the frame is voiced if the threshold value is greater than a determined value and unvoiced if less than.

- **Energy**

Energy for voiced part is obviously greater than unvoiced part, by visualizing the speech you can easily recognize the regions of voiced speech.

- **Zero-crossing**

Denoted by ZC rate, in case of unvoiced sounds the ZCR value is significantly high compared to the region of voiced sounds. Unvoiced speech has in general, higher zero-crossing rate than voiced speech, so we can use this for discriminating voiced from unvoiced part.

- **Pitch Period**

Among many parameters in speech analysis, synthesis, and coding applications, one parameter is essential called the fundamental frequency, or pitch of voiced speech. The pitch period is also can be used to detect voiced and unvoiced parts of speech, this is what have been done in this thesis for LPC-10 Matlab implementation.

Gain Calculation

The gain can be found by Square root of multiplication the pitch period with energy of prediction error for voice and Square root of just prediction error energy for the unvoiced part.

function [ar, vi, kappa] = levinson_durbin(r, p)

The function solves the Toeplitz system of equations, or sometimes called Yule-Walker AR equations by using the Levinson-Durbin recursion. Where r is a vector of autocorrelation coefficients, and p is the order of the recursion.

The prediction error energies for the pth-order solution are returned in the vector vi, and the p estimated reflection coefficients in the vector kappa.

For voiced frame the gain is calculated by:

$$G(n) = \sqrt{(P(n) \times vi(n))} \qquad (22)$$

Where P is the pitch period of voiced frame, vi is the prediction error vector.

For unvoiced frame:

$$G(n) = \sqrt{vi(n)} \qquad (23)$$

Where n denotes to a particular frame. Figure 27 shows the calculated gain for both voiced and unvoiced parts of signal.

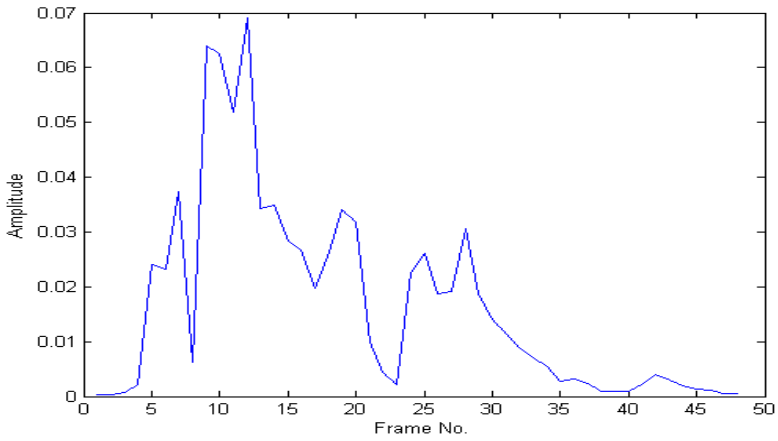
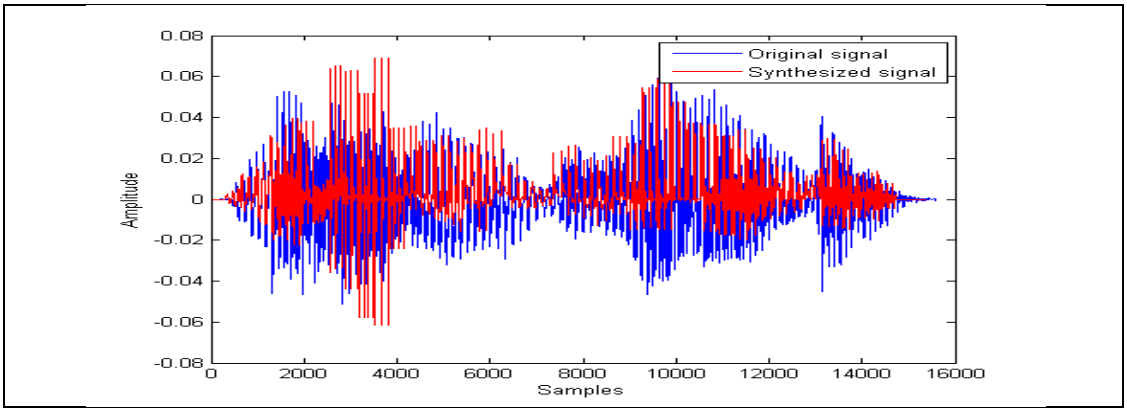


Figure 27 Calculated Gain for Voiced and Unvoiced Part



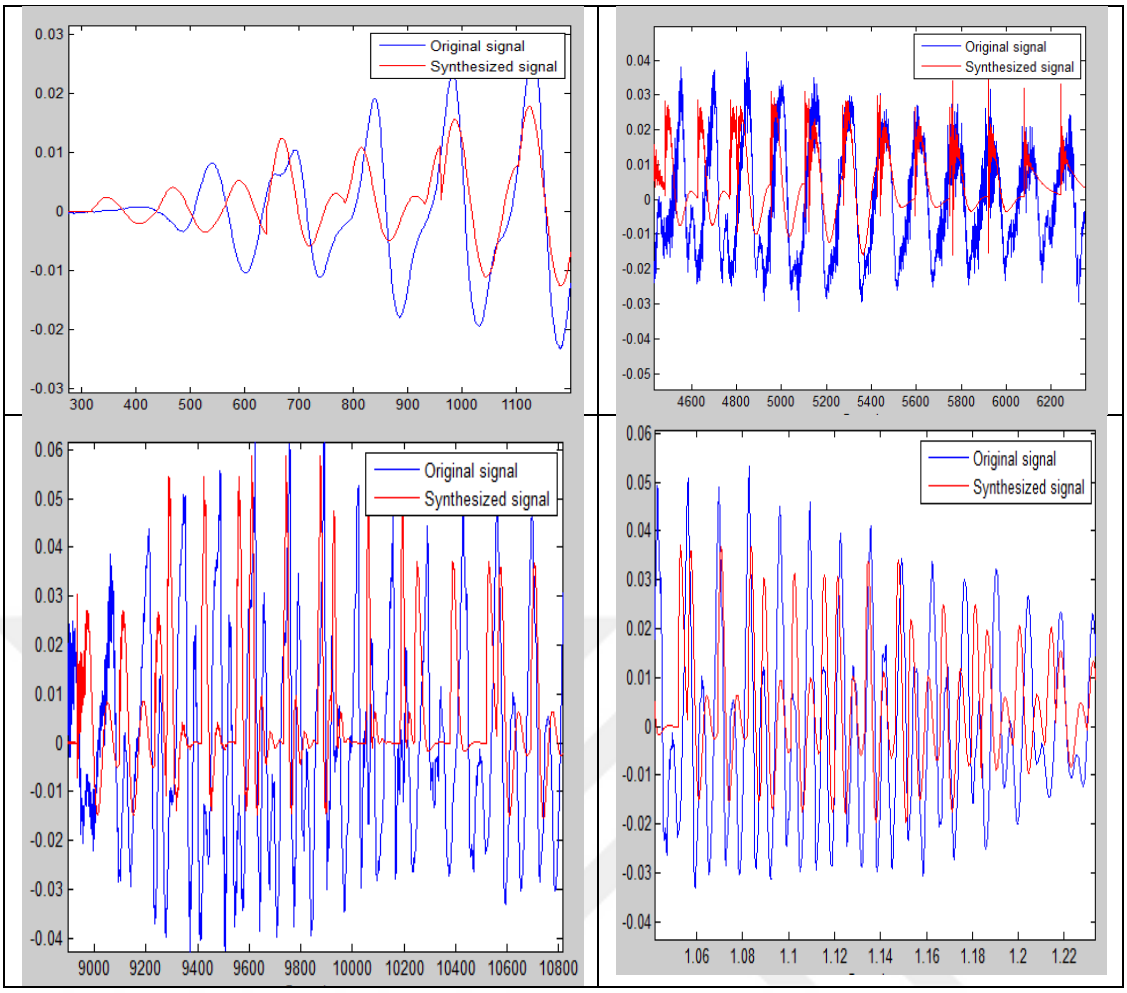


Figure 28 Original and Synthesis Frames

5.2 Part2: Analysis by Synthesis (AbS)

LPC Analysis & Synthesis

Similar to what have done in Part1.

Parameters Calculation

Finding the parameters such as pitch and gain is done by exhaustive search, let's start with the basic AbS code that have been implemented :

One frame from a wav file has a multiple pitch periods as shown in Figure 29:

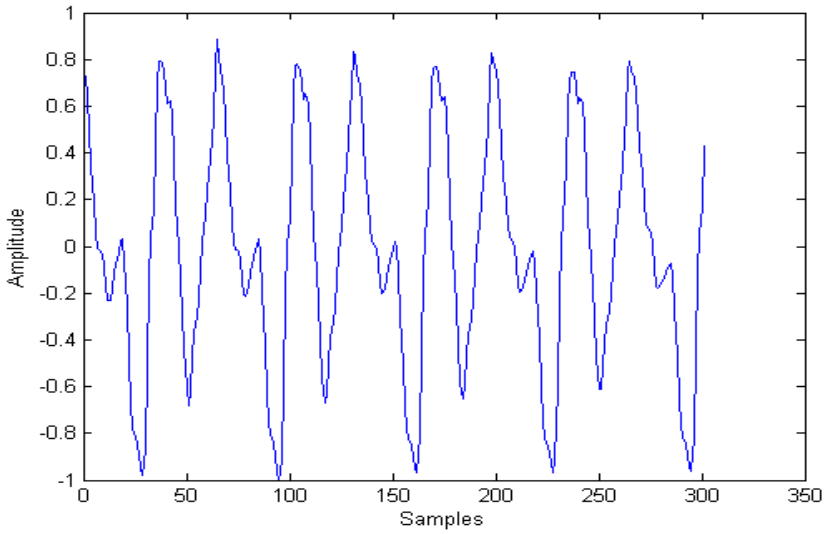


Figure 29 One Frame with Multiple Pitch Periods

By applying loop to find the pulse position with the minimum Mean Squared Error criterion the result was:

Frame size =300 samples

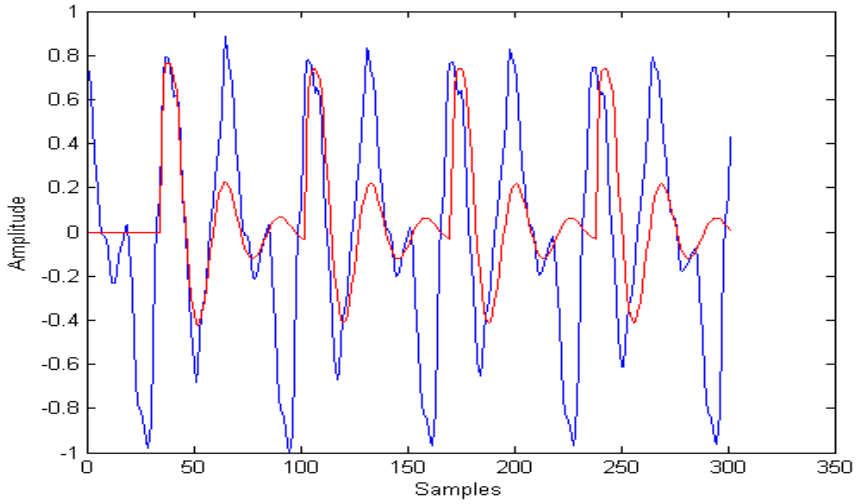


Figure 30 Original (blue) and Synthesis (red) Signals

As shown from Figure 30 the values of synthesized signal from the beginning are zeros that because the first impulse of excitation signal at sample 35, see Figure 31.

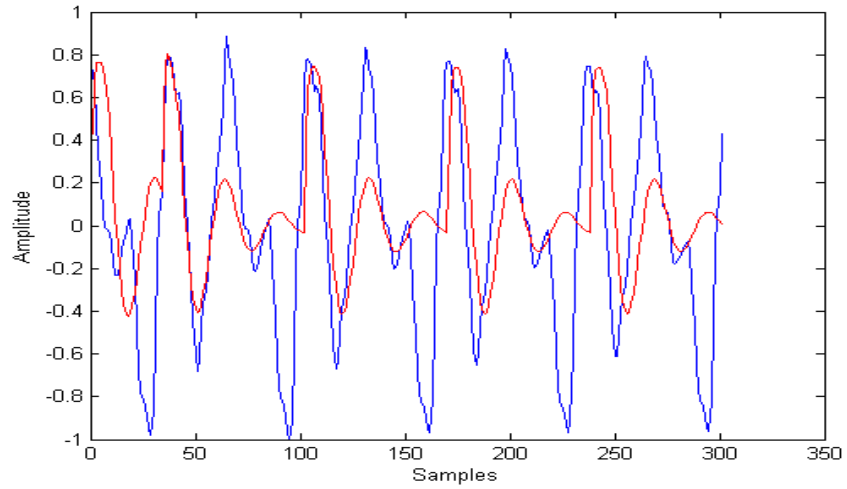


Figure 31 Synthesis Signal with First Impulse

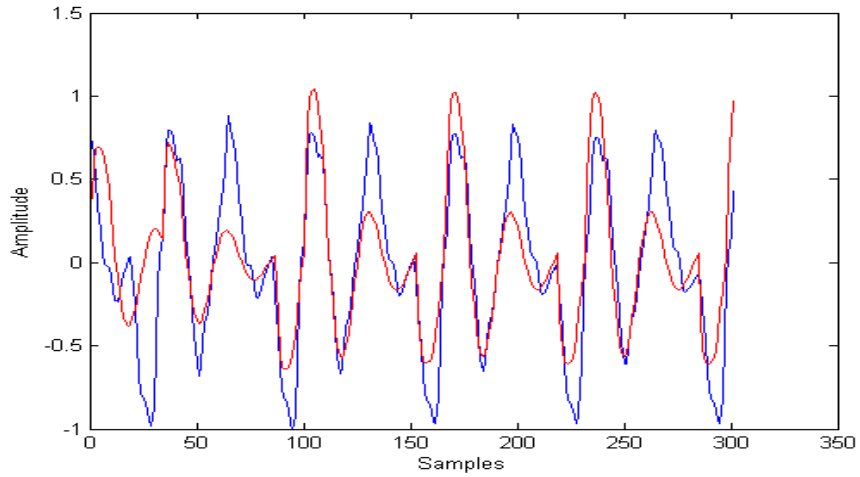


Figure 32 Synthesis Signal with Negative Impulses

Next finding the position of the excitation and gain by MSE automatically:

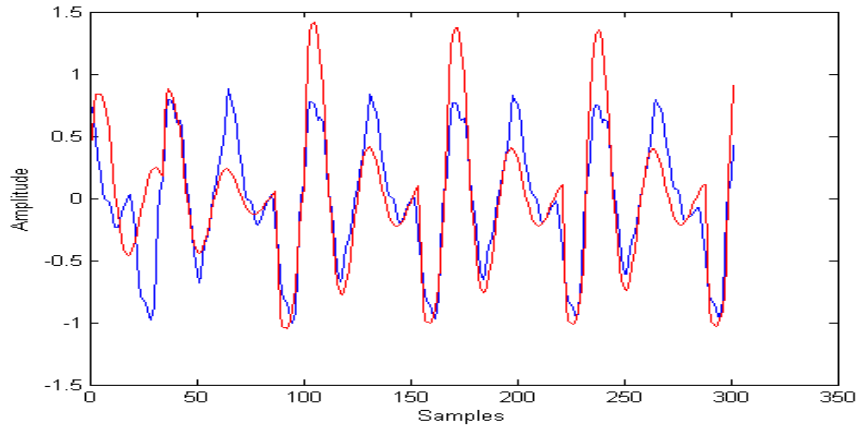


Figure 33 All Parameters Achieved Automatically

With same criterion (MSE) finding the parameters automatically, these parameters are:

- Pulse location
- Pulse repetition
- Gain

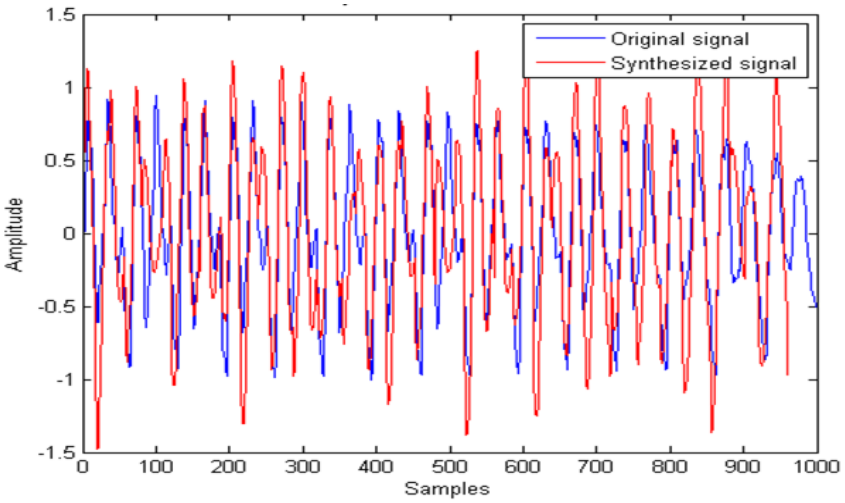


Figure 34 Finding All Parameters Automatcally for Three Frames

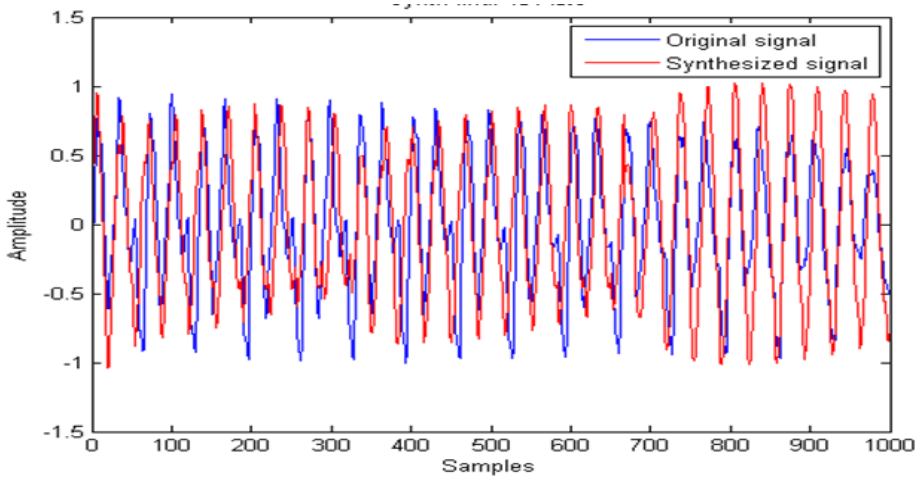


Figure 35 Finding the Pulse Position

On the same way by using MSE criterion a codebook of all expected pitch period values has been implemented to find the optimum pitch period.

5.3 Part3: Code Excited Linear Prediction (CELP)

CELP codec block diagrams see Figures 17, 18, 20 and 21.

Flow chart of Celp coder

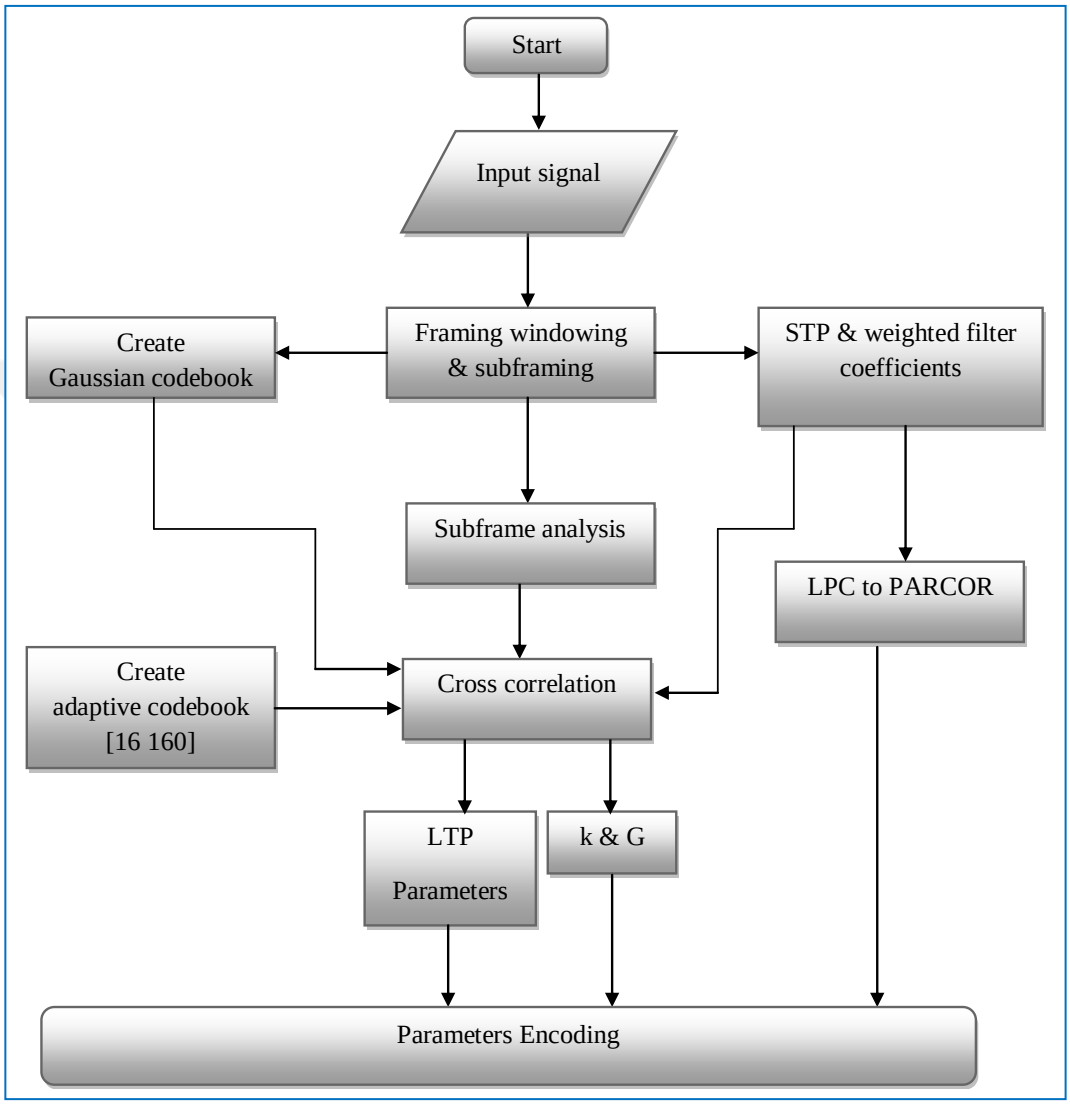


Figure 36 CELP Flow Chart

LPC Analysis

In CELP this analysis called STP- Short Term Prediction and it's quite similar to what have done in Part1

Excitation Signal

It's known that the CELP coder uses codebook to excite the LP synthesis filter.

As shown in Figure 19, there are three stages has been placed after codebook gain, LPC envelope and pitch prediction.

We generated a Gaussian codebook containing 1024 sequences of length 40 by using following commands:

codebook_G = randn(40,1024);

To ensure that the values don't change every time and don't change at receiver side, this command has been added:

randn('state',0);

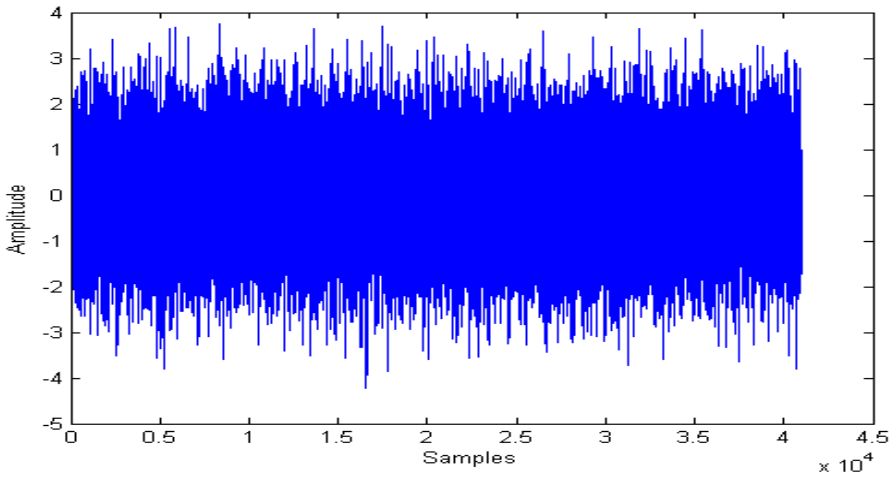


Figure 37 Stochastic Codebook

By determining the excitation parameters k (index which is L×K where L subframe length usually 40 samples and 1< k > K), amplified by gain, and filtered with P, and b (long-term filter coefficients), they used to generate the excitation sequence, e(n), for the speech block x(n) of length L.

The parameters P and b in the pitch synthesis filter are estimated in the pre-determined range of pitch and b should be less than 1.4-1.5.

CHAPTER 6: RESULTS AND CONCLUSION

We give the results and our comments here, in the last chapter of the thesis. The results of the MOS tests carried out, as well as, the performance comparison in terms of quality, complexity and bit rate are given in this part.

6.1 Results

6.1.1 Quality

Subjective Quality

All listeners were asked to do this test by giving a number of five categories (Table-2) characterized by numerical values as shown in Table-3, where 4 records were taken by experts.

Table- 3 Subjective Testing Records

man1.wav	"In the course of a December tour in Yorkshire, I rode for a long distance in one of the public coaches"																		
Tester No.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
LPC man1_LPC.wav	**2**	**2**	**2**	**2**	**3**	**2**	**2**	**3**	**1**	**1**	**3**	**1**	**1**	**2**	**2**	**2**	**3**	**1**	**2**	**1**
AbS man1_AbS.wav	**3**	**3**	**3**	**3**	**2**	**2**	**3**	**3**	**1**	**2**	**2**	**1**	**1**	**1**	**2**	**3**	**2**	**3**	**1**	**2**
CELP man1_CELP.wav	**4**	**5**	**5**	**5**	**4**	**3**	**4**	**4**	**4**	**4**	**4**	**4**	**3**	**4**	**4**	**4**	**4**	**4**	**4**	**5**
woman1.wav	"This is example for using in matlab"																			
Tester No.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
LPC woman1_LPC.wav	**2**	**1**	**2**	**3**	**3**	**1**	**2**	**3**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**2**	**2**	**1**	**2**	**3**
AbS woman1_AbS.wav	**3**	**2**	**2**	**3**	**2**	**2**	**3**	**3**	**1**	**2**	**1**	**1**	**1**	**2**	**2**	**3**	**3**	**2**	**2**	**1**

woman1_CELP.wav	4	4	4	5	4	3	4	4	3	5	3	4	2	3	3	4	4	3	5	4
woman2.wav	"To administer medicine to animals is frequently a very difficult matter"																			
Tester No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
woman2_LPC.wav	2	2	2	2	3	2	2	2	1	1	3	1	2	2	2	1	2	1	2	1
woman2_AbS.wav	3	3	3	3	2	2	3	3	2	2	2	1	1	2	2	3	1	2	2	2
woman2_CELP.wav	4	5	5	5	4	3	4	4	3	3	4	4	3	4	3	4	4	3	5	5
woman3.wav	"What movies have you seen recently"																			
Tester No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
woman3_LPC.wav	1	2	2	4	3	2	3	2	1	1	3	1	1	4	3	1	3	2	2	1
woman3_AbS.wav	1	3	3	2	2	1	4	2	1	2	2	2	2	2	4	3	2	1	2	2
woman3_CELP.wav	5	5	5	5	4	4	5	5	4	5	4	4	4	4	4	4	5	4	4	5

The mean value of opinion scores are listed in Table- 4

Table- 4 Average Values of Records

File name	LPC-10	AbS	CELP
man1.wav	1.9	2.15	4.1*
woman1.wav	1.75	2.05	3.75
woman2.wav	1.8	2.2*	3.95
woman3.wav	2.1*	2.15	4.45

Objective Quality

Signal-to-Noise Ratio SNR

The SNR formula used could be depicted below:

$$SNR = 10\log_{10}(sum(x)^2/sum(x-y)^2) \qquad (24)$$

Where, x: input signal , y: output signal

Table- 5 Signal to Noise Ratio of Processed Files

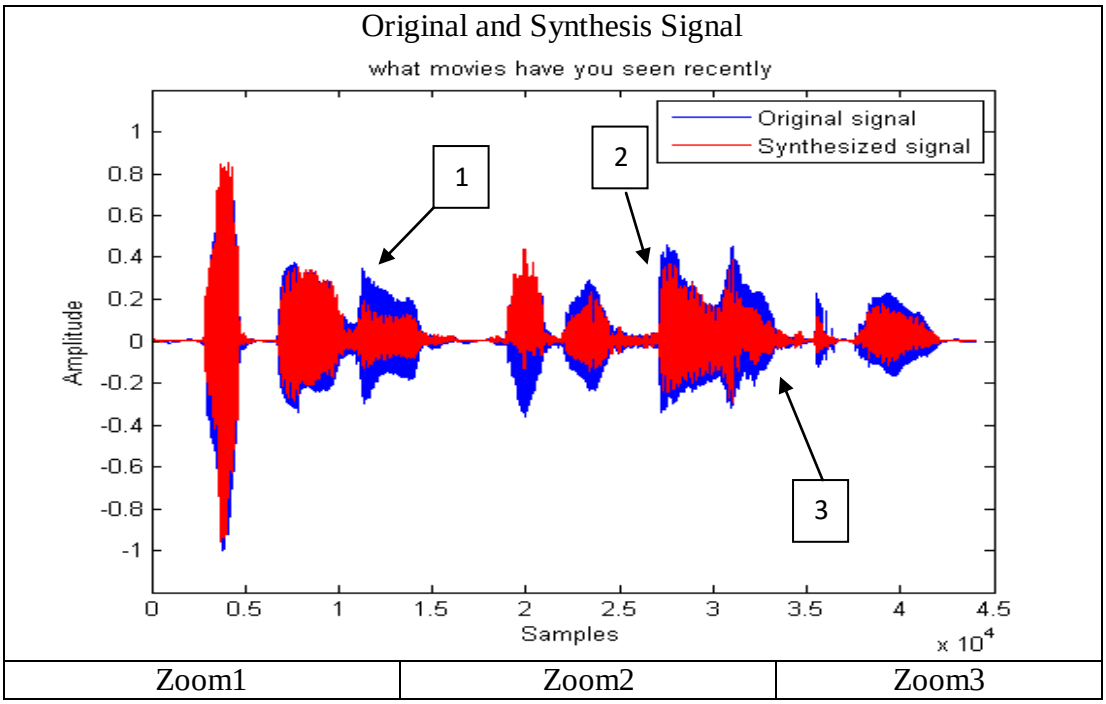
Codec Type	SNR in dB			
	Man1	Woman1	Woman2	Woman3
LPC	-36.9171	-36.0534	-36.0104	-25.5097
AbS	-27.6317	-29.6984	-5.3355	-8.1233
CELP	-2.7077	12.1303	0.9355	11.7255

Signal Matching

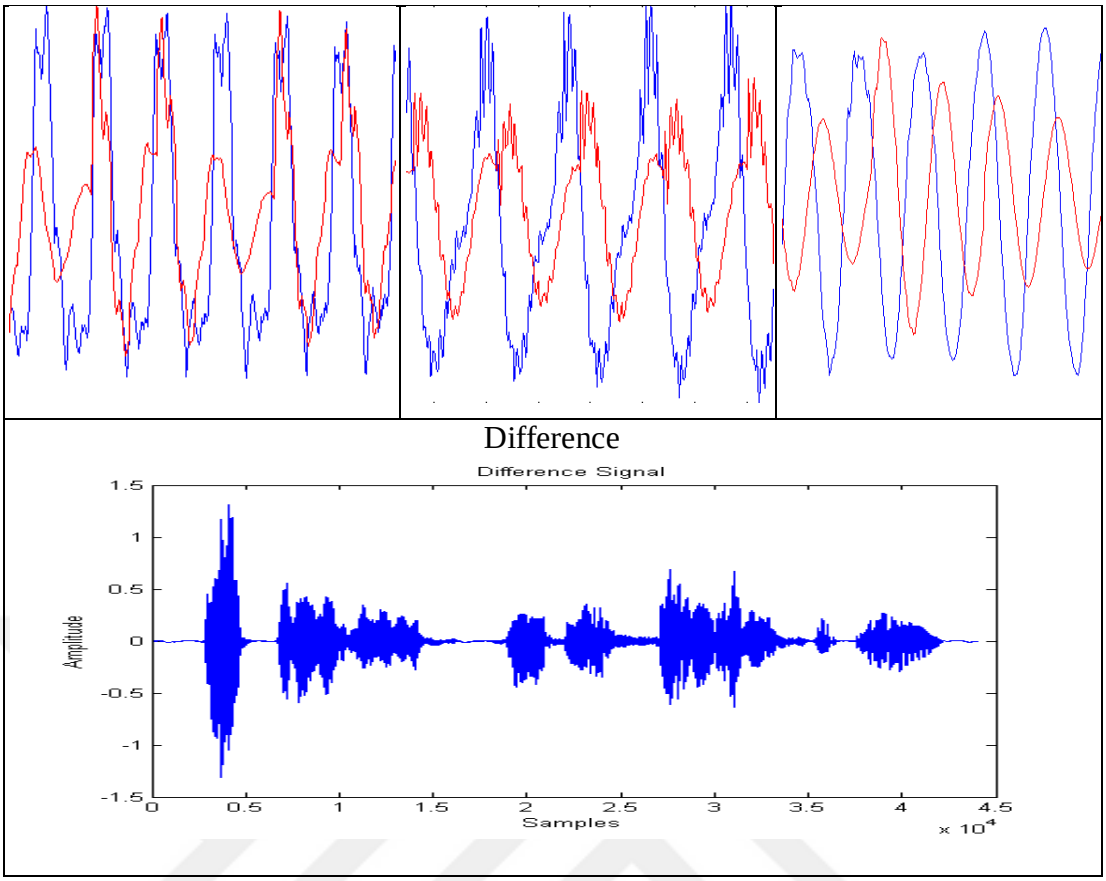
Matching means how is the synthesized signal closer to original signal by graphs, the best MOS candidate from each coder has been chosen.

Three signals are taken from Table- 4 (*)

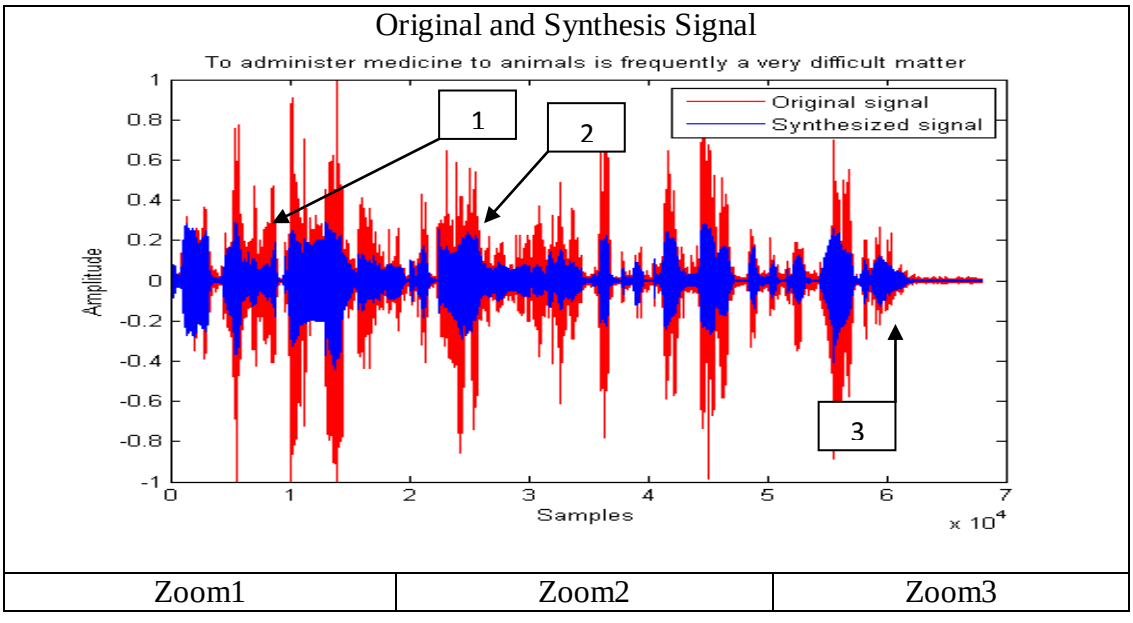
LPC-10



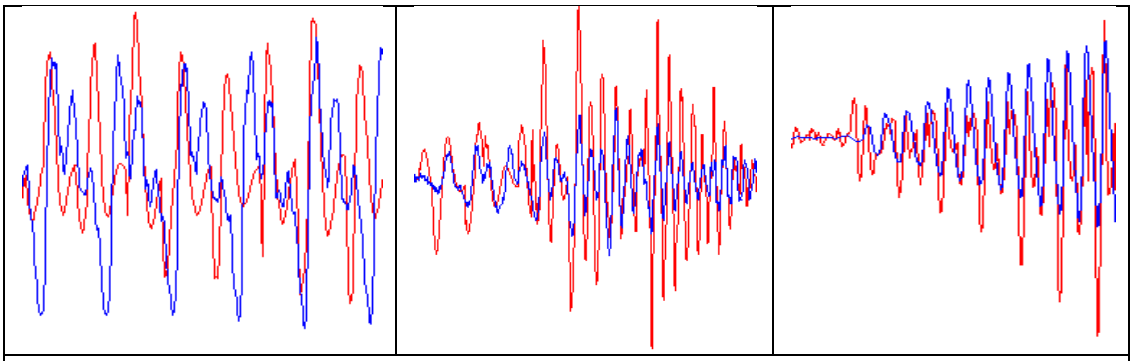
Zoom1	Zoom2	Zoom3



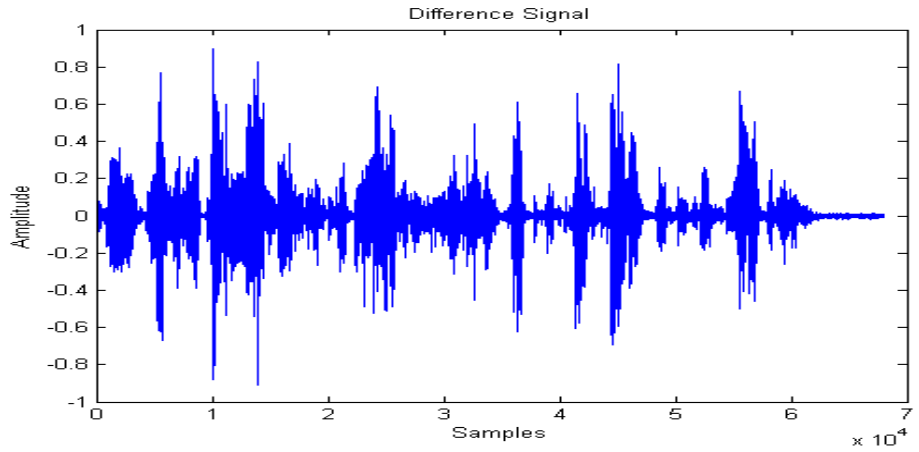
AbS



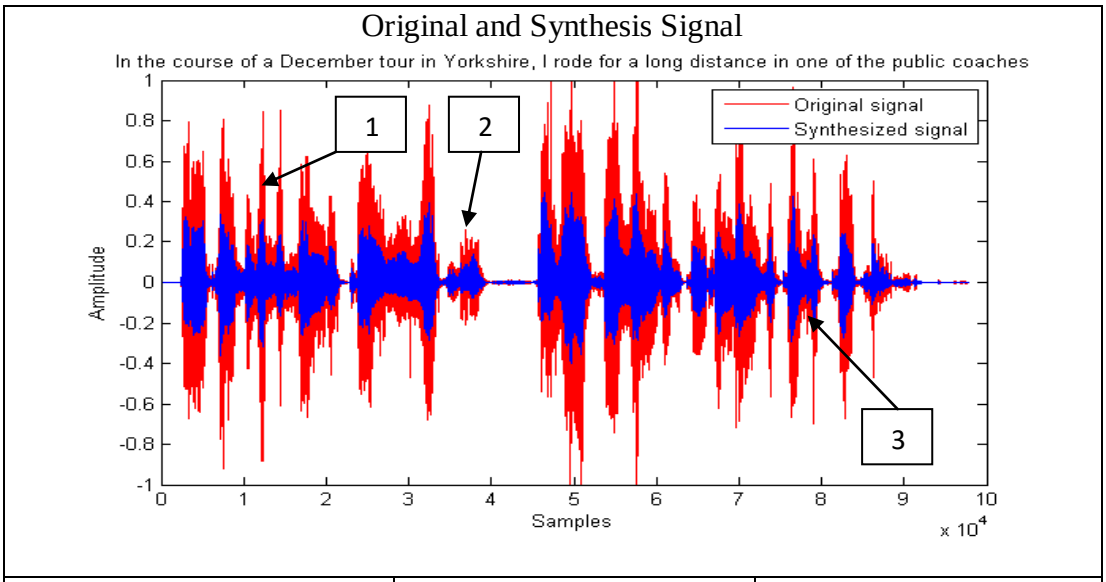
Zoom1	Zoom2	Zoom3



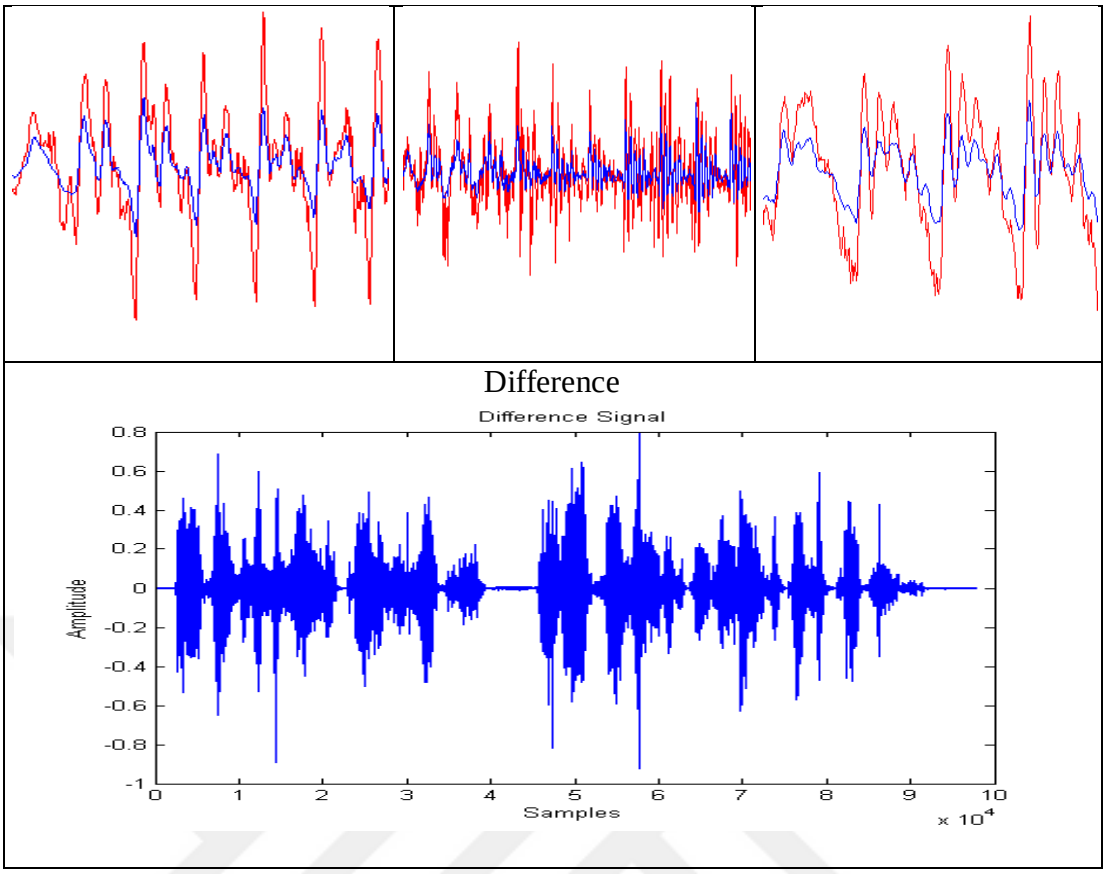
Difference



CELP



Zoom1	Zoom2	Zoom3



6.1.2 Bits Allocation

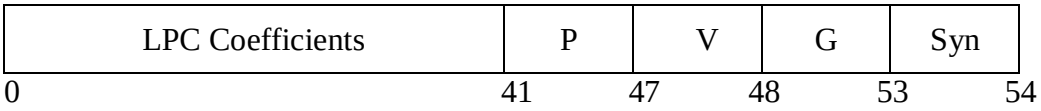
- Sampling frequency Fs = 16000 Hz
- Window length (frame): 20 ms; which corresponds to 160 samples per frame by the given sampling rate Fs
- Overlapping: no overlapping.
- Number of predictor coefficients of the LPC-10 model = 10
- Normalized input signal for all coders

Table- 6 Bit Rate for LPC-10 Codec

	Number of bits per frame
LP coefficients	41
Gain	5
Pitch period	6
Voiced/unvoiced switch	1
Synchronization	1
Total	54
Bit rate	**100×54=5400 bits/second***

The detailed bit allocation is performed as follows :

The 54-bit Frame



54 bits TOTAL BITS PER FRAME

Sample rate = 8000 samples/seconds for narrowband and 16000 samples/seconds for wideband

Samples per segment = 160 samples/segment

Segment rate = Sample Rate/ Samples per Segment

= (16000 samples/second)/(160 samples/second) = 100 segments/second

Segment size = 54 bits/segment

Bit rate = Segment size * Segment rate

— = (54 bits/segment) * (100 segments/second) = 5.4 kbits/second

Table- 7 Bit Rate for AbS Codec

	Number of bits per frame
LP coefficients	10 * 6 = 60
Gain	5
Pitch period	7
Total	72
Bit rate	**72*100=7200 bits / seconds**

Table- 8 Bit Rate for CELP Codec

	Number of bits per frame
LP coefficients	10 * 6 = 60
Gain	7
Pitch period	8
Codebook index k	8
Pitch gain	8
Total	91
Bit rate	**91×100=9100 bits/seconds**

6.2 Discussion of Results

6.2.1 Quality Performance

- Subjective Quality

Table- 9 Mean Opinion Score of Output Signals (MOScore for 20 people)

File Name	LPC-10	AbS	CELP
man1.wav	1.9	2.15	4.1
woman1.wav	1.75	2.05	3.75
woman2.wav	1.8	2.2	3.95
woman3.wav	2.1	2.15	4.45

A comparison of the original speech signal against the LPC-10, AbS and CELP reconstructed speech have been studied. The outputs have a lower quality than the original input speech. All of the synthesized speech signals have a portion of hearable noise with the unclarity of the words themselves, where the output of vocoder LPC-10 being approximately unintelligible. The noise in background is very strong in LPC-10 with slightly improvement in AbS but much better in CELP. The AbS reconstructed speech sounds more articulated and less mechanic and the words are recognizable. Resulting that the speech was reconstructed using CELP sounded the best.

- Objective Quality

The results shown in Table-5 that have been achieved are reflect the results that obtained by subjective quality, where high SNR were recorded for CELP codec and low SNR for LPC-10 codec with negative values for all output.

Finding accurate pitch period is critical issue because it carries personality feature.

Figure 38 gives the pitch period found by Autocorrelation method

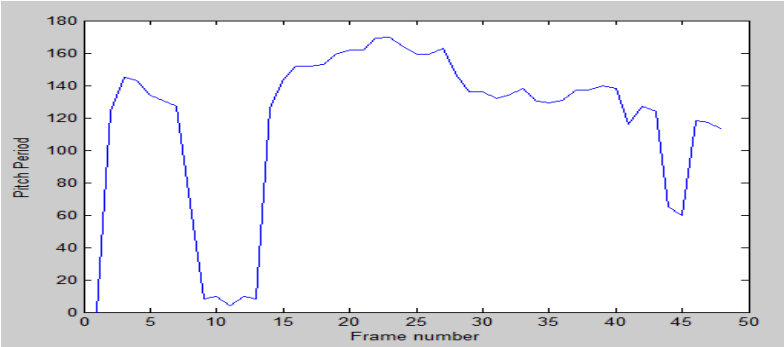


Figure 38 Pitch Periods by Autocorrelation

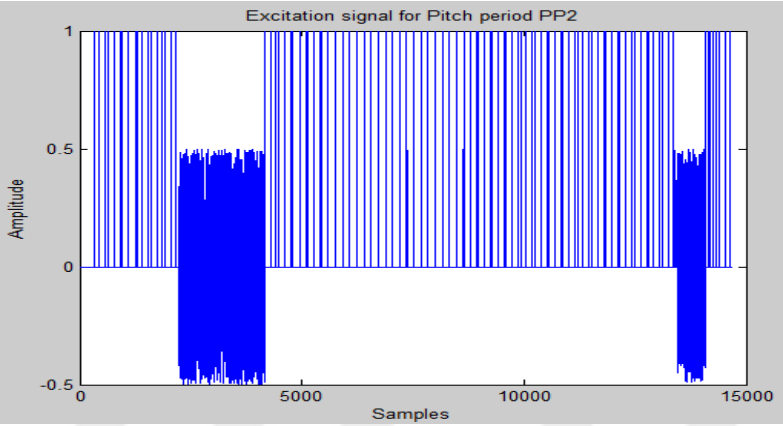


Figure 39 Excitation Signal Derived from Pitch Period

As shown in excitation signal Figure 39 it shows voiced , unvoiced and silence parts.

Speech quality term implies accurate parameters that can be found by Mean Square Error MSE or Signal to Noise Ratio SNR. Take for an example the above pitch period can be found from this point of view. After applying MSE and SNR without knowing the pitch period for every individual frame these periods have been got at Maximum SNR at same, Minimum MSE. See Figure 40

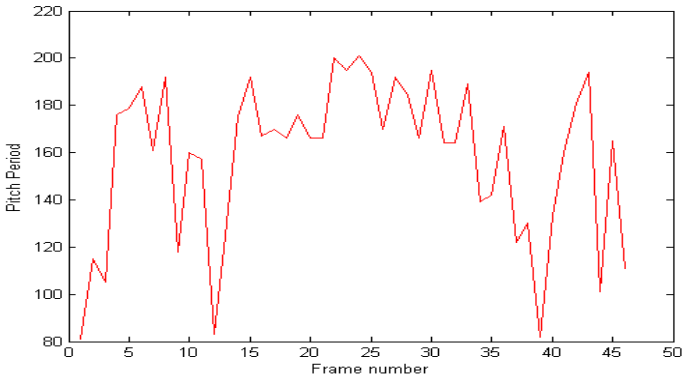


Figure 40 Pitch Periods Found by MSE

6.2.2 Delay and Complexity

The measured time of processing actually denotes to both delay and complexity, the processed time is accomplished by commands inside implemented Matlab code at the beginning of code and end, then we take the difference of them as shown in Table 9, this method implies computational complexity; where the faster in time the lower in complexity and vice versa.

Table- 10 Processed Time of LPC-10, AbS, and CELP Comparisons

Coder Type	Time in Seconds			
	Man 1	Woman 1	Woman 2	Woman 3
LPC-10	0.7334	0.375	0.5625	0.3906
AbS	19.1094	5.78125	14.2969	6.6719
CELP	24.1563	7.90625	18.9219	13.6875

A delay noticed with processing of Figures and sounds in implemented Matlab codes; to be closer to the real time processing these commands have been omitted.

6.2.3 Bit-Rate Performance

The bit stream is different from coder to coder we have to decide a particular features and specify the parameters to be encoded for each coder separately Tables 5, 6, and 7 show these parameters. The bit rate that obtained from each coder LPC-10, AbS and CELP in general is low with most increasing difference toward CELP. For instance in the LPC-10 vocoder the parameters required at receiver side to excite the filter with these calculated parameters (LP coefficients, pitch and gain) gave low bit rate but led to results almost unintelligible words. The one thing can be focusing on it is the pitch period, I think absence the accuracy of pitch values OR can be because ignoring other components of one pitch period and the produced excitation result in a decline in the quality which much similar to robotic. Moving to the AbS technique, is not far away from LPC-10 but it considers the pitch components in its implementation which means AbS codec does not rely on one pitch period in one frame.

Unlike LPC-10, and AbS, the CELP added more parameters which are positively affected on quality and negatively on complexity.

6.3 Conclusion

The key word of this thesis is the residual signal, the results achieved from the CELP are much intelligible compared to other codecs as expected. On the other hand, the basic LPC-10 results are the worst. While AbS stay in the middle. One can notice that the LPC-10 uses pitch estimator and does not use the residual signal , the AbS uses more precise values for pitch period and either does not use the residual signal, but the CELP uses both.

Implementation of LPC-10 comprehensively provides the information on how generally a vocoder works, but from point view of quality the other coders by exploiting some features has led to a result is far good what could be achieved using LPC-10 technique.

To sum up, one codec attribute is defined and then tradeoff between others, for example if we selected quality at same time the bandwidth slice was limited then the delay of the system would be long.

Future Work

More concentration on excitation and pitch. The most dispute point in all references is excitation, weather is derived from pitch or other kind of excitation. One solution we can save the minimum error signal found by Analysis by Synthesis - CELP and use it instead of purely random noise. Other solution filter the residual signal by more than 1-tap pitch predictor and then use it as excitation signal next times, by this way the number of excitation vector will be much less than the excitation from Gaussian random noise vectors.

LabView is an interactive software tool can be used to teach the various modules of speech coders to undergraduate and graduate students [22].

REFERENCES

[1] Chen, J. C. and T. Zhang. IP-Based Next-Generation Wireless Networks: Systems, Architectures, and Protocols, Wiley (2004).

[2] Deller, J. R. J., et al.. Discrete-Time Processing of Speech Signals, Wiley (2000).

[3] Sigmund, M. X. Voice Recognition by Computer, Tectum-Verlag (2004).

[4] Chu, W. C. Speech Coding Algorithms: Foundation and Evolution of Standardized Coders, Wiley (2004).

[5] Pieraccini, R. and L. Rabiner. The Voice in the Machine: Building Computers That Understand Speech, MIT Press (2012).

[6] Ingle, V. K. and J. G. Proakis. Digital Signal Processing Using MATLAB, Cengage Learning (2011).

[7] Simon Haykin, Communication Systems 4th Edition, Wiley & Songs (2000).

[8] Kondoz, A. M. Digital Speech: Coding for Low Bit Rate Communication Systems, Wiley (2005).

[9] Ahmad, A. Wireless and Mobile Data Networks, Wiley (2005).

[10] Saily, M., et al. GSM/EDGE: Evolution and Performance, Wiley (2011).

[11] Agrawal, D. P. and Q. A. Zeng. Introduction to Wireless and Mobile Systems, Cengage Learning (2015).

[12] Garg, V. Wireless Communications & Networking, Elsevier Science (2010).

[13] Rabiner, L. R. and B. H. Juang. Fundamentals of Speech Recognition, PTR Prentice Hall. (1993).

[14] Bertrand M. T. MSC thesis "Joint source-channel coding: application to speech coding" ASTON university

[15] Goldberg, R. G. A Practical Handbook of Speech Coders Ed. Randy Goldberg Boca Raton CRC Press (2000).

[16] https://en.wikipedia.org/wiki/Ear

[17] https://en.wikipedia.org/wiki/Sound_localization

[18] Hanzo, L., et al. Voice and Audio Compression for Wireless Communications, Wiley (2008).

[19] O'Cinneide, A., Dorran, D. & Gainza, M. The Problem and its Solution and Application to Speech. DIT Internal Technical Report, (2008).

[20] Ramamurthy, K. N. and A. Spanias. MATLAB Software for the Code

Excited Linear Prediction Algorithm: The Federal Standard-1016, Morgan & Claypool Publishers (2010).

[21] Atti, V. and A. Spanias. Algorithms and Software for Predictive and Perceptual Modeling of Speech, Morgan & Claypool (2011).

[22] Karthikeyan N. Ramamurthy, Jayaraman J. Thiagarajan and Andreas Spanias. An interactive speech coding tool using LabView Conference Paper January (2011).