Chapter 1

INTRODUCTION TO SPEECH CODING

1.1. OVERVIEW OF SPEECH CODING

Speech is a very special type of signal for different reasons. The most preliminary of these is the fact that speech is a non stationary signal. This makes the speech signal hard to analyze and model. The second reason is that factors like intelligibility, coherence and other such characteristics play a vital role in the analysis of the speech signals. The third reason in communication point of view is that the number of discrete values required to describe one second of speech signal corresponds to 8000 samples (at the minimum). As bandwidth is the parameter which affects the cost of processing, speech signals are subjected to compression before transmission.

Speech coding or compression is a process of obtaining a compact representation for the speech signals, for the purpose of efficient transmission over band limited wired or wireless channels and also for efficient storage. In recent day’s speech coders became the essential components for telecommunications and multimedia as the utilization of the bandwidth affects the cost of transmission. The goal of speech coding is to represent the samples of a speech signal with a minimum number of bits without any reduction in the perceptual quality. Speech coding helps a telephone company to carry out more voice calls on a single fiber link or cable. Speech coding is very important in Mobile and Cellular communications where the data rates for a
particular user are limited, as lower the data rates for a voice call more services can be accommodated[1-2]. Speech coding is also useful for Voice over IP, Video conferencing and Multimedia applications to reduce the bandwidth requirement over internet [3]. In addition, most of the speech applications require minimum coding delay because long coding delays hinder the flow of the speech conversation [4].

A speech coder is one which converts a digitized speech signal into a coded representation and transmits it in the form of frames. At the receiving end the speech decoder receives the coded frames and performs synthesis to reconstruct the speech signal. The speech coders differ primarily in bit-rate, complexity, delay and perceptual quality of the synthesized speech [5]. There exist two types of coding techniques, narrowband speech coding and wideband speech coding. Narrowband speech coding refers to coding of the speech signals whose bandwidth is between 300 to 3400 Hz (with 8 KHz sampling rate), while wideband speech coding refers to coding of the speech signals whose bandwidth is less than 50 to 7000 Hz (with 14 –16 KHz sampling rate). Narrowband speech coding is more common than wideband speech coding because of the narrowband nature of the telephone channel lines (whose bandwidth lies between 300 to 3400 Hz) [6]. In recent days there is an increase in demand for wideband speech coding techniques in applications like video conferencing.
In this work the experiment is carried out using a standard TIMIT database and involves the following steps:

- **Speech analysis:** It involves Framing, windowing, overlapping of frames, Calculation of linear predictive coefficients, Estimation of pitch of a frame, Conversion of linear predictive coefficients (LPC) to line spectral frequencies parameters (LSF).

- **Vector Quantization:** It involves the design of Codebooks using Linde, Buzo, Gray algorithm, Design of various vector quantizers, Implementation of Vector Quantization on line spectral frequencies.

- **Speech Synthesis:** It involves the conversion of line spectral frequencies to linear predictive coefficients and performs synthesis using pitch, gain, and quantized LPC parameters.

The other steps involve the computation of spectral distortion, outliers, and unstable frames.

The motivation behind this work is to develop an efficient vector quantizer having low bit-rate, complexity and memory requirements when compared to the existing techniques and to best suit it for applications involving speech coding.

### 1.2. SPEECH CODING SYSTEM

In a speech coding system, initially the input speech signal which is analog in nature is digitized using a filter, sampler and analog to digital converter (A/D) circuits. The filter used is an anti aliasing filter which is a low pass filter used before a sampler to remove all signal
frequencies that are above the nyquist frequency. The filtering is done to avoid the problem of aliasing. If the speech signal sampling frequency is less than twice the bandwidth of a sampled speech signal the problem of aliasing occurs. The best solution to aliasing is to make the sampling frequency greater by 2.5 times the bandwidth of the analog speech signal. According to nyquist theorem the sampling frequency must be at least twice the bandwidth of the continuous-time signal in order to avoid aliasing. A value of 8 KHz is commonly selected as the standard sampling frequency for the telephone speech signals, since the telephone speech signal frequency range is between 300 to 3400 Hz [5].

Later the sampler converts the analog speech signal into a discrete form and will be given as an input to A/D converter whose output is a digitized speech signal. Most speech coding systems were designed to support telecommunication applications, by limiting the frequency contents between 300 and 3400 Hz. To convert the analog speech signal to digital format, to maintain the perceptual quality and to make the digital speech signal indistinguishable from the input it is necessary to sample the speech signal with more than 8 bits per sample. The block diagram of a speech coding system is shown in Fig 1.1. Throughout the thesis the parameters considered for the digital speech signal are sampling frequency of 8 KHz and 8 bits per sample. Hence the input speech signal taken will have a bit-rate equal to 64 Kbps.
The function of a source encoder is to reduce the bit-rate of the input speech signal below 64 Kbps. Any bit-rate below 64 Kbps is treated as compression and the output of the source encoder is an encoded speech signal having a bit-rate less than 64 Kbps. The coding algorithm used in this thesis for the reduction of the bit-rate is the linear predictive coding (LPC) algorithm. Using this algorithm the bit-rate is reduced to 1 Kbps, i.e., a reduction in bit-rate by 64 times with respect to the input. The output of the source encoder is given as an input to the channel encoder which provides error protection to the bit stream transmitted over the communication channel where noise and interference can reduce the reliability of the transmitted data.

Fig 1.1 Speech coding system

At the receiving end the channel decoder recovers the encoded data from the error protected data and will be fed to the source decoder so as to recover the original speech signal. Later the speech signal is fed to a digital to analog (D/A) converter to convert the speech signal from
digital to analog format. Finally, the analog speech signal is fed to an 
anti aliasing filter to prevent aliasing during the reconstruction of 
continuous speech signal from the speech samples, that again 
requires perfect stop-band rejection to guarantee zero aliasing [5,7-9].

1.3. ATTRIBUTES OF SPEECH CODERS

The aim of speech coding is to enhance the quality of a speech 
signal at a particular bit-rate or to minimize the bit-rate at a given 
quality. The bit-rate at which the speech is to be transmitted or stored 
depends on the rate of transmission or storage, the computation of 
coding the digital speech signal and the quality of the speech signal 
required. Hence the desirable properties of the speech coder include 
[5, 10] and are explained below:

- Low bit-rate
- High speech quality
- Robustness to different speakers/languages
- Channel errors
- Low memory requirements
- Less computational complexity
- Low coding delay

Low Bit-Rate: The lower the bit-rate of an encoded bit stream the 
lesser is the bandwidth required for transmission. But any reduction 
in the bit-rate results in a loss in the quality of the speech signal 
which is undesirable. So a trade off must be maintained between
reduction in the bit-rate and the quality of speech signal depending on the intended application.

**High Speech Quality:** The decoded speech signal must have high quality and must be suitable for the intended application. The quality of the speech signal depends on factors like intelligibility, naturalness, pleasantness and speaker recognizability.

**Robustness to Different Speakers and Languages:** The speech coder must be general enough, so that it is used for any type of speakers like male, female and children, it can also be used for any type of language. But it is not a trivial task because every voice signal has its own characteristics.

**Robustness to Channel Errors:** This is crucial for digital communication systems where channel errors have a negative impact on the quality of the speech signal.

**Low Memory Size and Computational Complexity:** In order to have good marketability for the speech coder the costs associated with its implementation must be as low as possible. The cost of a product depends on the memory required to support its operations and the computational complexity. So the speech coder must have lower memory requirements and computational complexities to have better marketability.

**Low Coding Delay:** Coding delay is the time elapse from the time the speech sample arrives at the encoder input to the time the speech sample appears at the decoder output. An excessive delay creates
problems in real time two way communications. So the coding delay must be as low as possible.

1.4. SPEECH CODING

The objective of speech coding is to compress the speech signal by reducing the number of bits per sample, such that the decoded speech is audibly indistinguishable from the original speech signal. Specifically speech coding methods achieve the following gains [11]:

- Reduction in bit-rate or equivalently the bandwidth.
- Reduction in memory requirements which decreases in a proportionate manner with respect to bit-rate.
- Reduction in the transmission power required, since compressed speech signal has less number of bits per second to transmit.
- Immunity to noise, some of the saved bits per sample can be used as protective error control bits to the speech parameters.

1.4.1 Speech Coding Methods

Speech coding methods are broadly classified as lossless and lossy coding methods. In lossless coding the speech signal reconstructed at the decoder end can have exactly the same shape as the input speech signal wherein the lossy coding techniques have the reconstructed speech signal perceptually indistinguishable from the original speech signal. In lossy coding techniques though the reconstructed speech signal waveform differs from the original speech signal waveform,
majority of the speech coding techniques are based on the lossy coding techniques and removes the information which is irrelevant from the perceptual quality point of view. Modern speech coding methods is divided into waveform coding methods and parametric coding methods.

1.4.2 Classification of Speech Coders

Speech coders are classified based on the bit-rate at which they produce output with reasonable quality and on the type of coding techniques used for coding the speech signal [5].

1.4.2.1 Classification by Bit-Rate

The classification of speech coders based on the bit-rate is shown in Table 1.1. Speech coders are classified as High bit-rate coders, Medium bit-rate coders, Low bit-rate coders and Very Low bit-rate coders depending on the bit-rate range at which the speech coders produce reasonable qual

<table>
<thead>
<tr>
<th>Type of coder</th>
<th>Bit-Rate Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>High bit-rate coders</td>
<td>&gt;15 Kbps</td>
</tr>
<tr>
<td>Medium bit-rate coders</td>
<td>5 to 15 Kbps</td>
</tr>
<tr>
<td>Low bit-rate coders</td>
<td>2 to 5 Kbps</td>
</tr>
<tr>
<td>Very Low bit-rate coders</td>
<td>&lt;2 Kbps</td>
</tr>
</tbody>
</table>
1.4.2.2 Classification by Coding Techniques

Based on the type of coding technique used speech coders are classified into three types and are explained below. They are:

- Waveform coders
- Parametric coders
- Hybrid coders

1.4.2.2.1 Waveform Coders

Waveform coders digitize the speech signal on a sample-by-sample basis. Its main goal is to make the output waveform to resemble the input waveform. So waveform coders retain good quality speech. Waveform coders are low complexity coders, which produce high quality speech at data rates above and around 16 Kbps. When the data rate is lowered below this value the reconstructed speech signal quality gets degraded. Waveform coders are not specific to speech signals and can be used for any type of signals. The two types of waveform coders are Time domain Waveform Coders and Frequency domain Waveform Coders. Time domain waveform coders utilize the digitization schemes based on the Time domain properties of the speech signal. Some of the examples of Time domain waveform coding techniques are Pulse Code Modulation (PCM), Differential Pulse Code Modulation (DPCM), Adaptive Differential Pulse Code Modulation (ADPCM) and Delta Modulation. Frequency domain waveform coders segment the speech signal into small frequency bands and each band is coded separately using waveform coders. In these coders, the accuracy of encoding is altered dynamically between the bands to suit
the requirements of a speech waveform. Examples of frequency domain waveform coders are Subband coders and Adaptive transform coders [12-13].

1.4.2.2.2 Parametric Coders

In parametric coders the speech signal is assumed to be generated from a model controlled by some speech parameters. In these coders the speech signal is modeled using a limited number of parameters corresponding to the speech production mechanism. These parameters are obtained by analyzing the speech signal before transmission. At the receiving end the decoder uses these parameters to reconstruct the original speech signal. In these coders the output speech signal is not an exact replica of the input speech signal and the resemblance of the input speech signal is lost. However, the output speech signal will utter the same as the input speech signal. Using parametric coders it is possible to obtain very low bit-rates at a reasonable quality. To retain the quality one has to use waveform coders taking advantage of the properties of both speech production and auditory mechanisms, so that the resulting quality is good at the cost of increased bit-rates compared to parametric coders. At lower bit-rates, the quality attained by the waveform coders is less compared to parametric coders. Examples of parametric coders are Linear Predictive Coders (LPC) and Mixed Excitation Linear Predictive (MELP) coders. This class of coders works well for low bit-rates. In these coders increasing the bit-rate does not result in an increase in the quality as it is restricted by the coder chosen [14-16].
1.4.2.2.3 Hybrid Coders

Hybrid coders try to fill the gap between waveform coders and parametric coders. Hybrid coders operate at medium bit-rates between those of waveform coders and parametric coders and produce high quality speech than parametric coders. There are a number of hybrid encoding schemes, which differ in the way the excitation parameters are generated. Some of these techniques quantize the residual signal directly, while others substitute approximated quantized waveforms selected from an available set of waveforms. A hybrid coder is a combination of both waveform coder and parametric coder. Like parametric coders, hybrid coders relies on the speech production model, as in waveform coders an attempt is made to match the original signal with the decoded signal in time domain. An example of hybrid coder is the Code Excited Linear Predictive (CELP) coders [15-16].

1.5. LINEAR PREDICTIVE CODING

Linear predictive coding (LPC) is defined as a method for encoding an analog signal, in which the value of a current speech sample is estimated using the past few speech sample values of a speech signal. It was proposed by the Department of defense in federal standard 1015, USA, published in 1984. Under normal conditions speech is sampled at 8,000 samples per second with 8 bits per sample. This makes the bit-rate of 64,000 bits per second. Using linear predictive coding, the bit-rate is reduced to 2,400 bits per second. In this thesis
the bit-rate has been reduced to 1000 bits per second. At this reduced bit-rate there is a small amount of loss in the quality of speech signal. Hence using linear predictive coding there is a loss in the quality of the speech signal and hence it is considered as a lossy compression technique [17-20].

1.6. SPEECH PRODUCTION AND MODELING

Speech signals are composed of a sequence of sounds. The speech sounds and the transitions between them serve as a symbolic representation of information. The arrangement of these sounds is governed by the rules of the language. The study of these rules and their implementation in human communication is called linguistics. The study and classification of the sounds of speech is called phonetics.

1.6.1 The Mechanism of Speech Production

The speech waveform is a sound pressure wave produced by the control movements of anatomical structures, which will make up the human speech production system. The human speech production mechanism is shown in Fig 1.2 [21]. Regardless of the language spoken, all people use relatively the same anatomy to produce sounds. The output produced by each human’s anatomy is governed by the laws of physics. The process of speech production in humans is summarized as air being pushed from the lungs, through the vocal tract and out through the mouth to generate speech. In this type of
description the lungs is thought of as the source of sound and the vocal tract as a filter that produces various types of sounds which will make up the human speech.

In order to understand how the vocal tract turns the air from the lungs into sound it is important to understand several key definitions. Phonemes are defined as a limited set of individual sounds. There are two types of phonemes they are voiced and unvoiced sounds which are considered by the linear predictive coder when analyzing and synthesizing speech signals.

Fig 1.2 Human speech production model

Examples of voiced sounds are vowel sounds. Voiced sounds have greater energy levels and distinct formant or resonant frequencies.
Voiced sounds are produced when air from the lungs vibrate the vocal cords in a periodic manner, producing a sequence of air pulses (glottal pulses). The rate of vibration of the vocal cords gives the information about the pitch of the speech sound produced. These air pulses created by the vibrations finally pass through the rest of the vocal tract where some frequencies resonate. During the production of voiced sounds, it is known that women and children produce sounds with greater pitch than men due to faster rate of vibration of the vocal cords. It is therefore important to include information regarding pitch period in the analysis and synthesis of speech signals, if the final output is expected to accurately represent the original input signal.

Unvoiced sounds are more often consonant sounds they have high resonant frequencies and lesser amount of energy than voiced sounds. Unvoiced sounds are produced when there is a turbulent flow of air through the vocal tract. During this period the vocal cords do not vibrate, instead they remain open until the sound is produced. Pitch is an unimportant attribute for unvoiced speech as there is no vibration of the vocal cords and hence no glottal pulses [22].

The categorization of sounds as voiced or unvoiced is an important consideration in the analysis and synthesis process. In fact, the vibration of the vocal cords or lack of vibration is one of the key components in the production of different types of sound. Another component that influences speech production is the shape of the vocal tract. Different vocal tract shapes will produce different sounds or resonant frequencies. The vocal tract consists of the throat, tongue,
nose, mouth and lips. It is defined as a path for producing speech using the vocal organs and this path shapes the frequencies of the vibrating air traveling through it. When a person speaks, the vocal tract constantly changes its shape at a very slow rate to produce different sounds, which flow together to create words.

A final component that affects the production of speech in humans is the amount of air that originates from the lungs. The air flowing from the lungs is considered as a source for the vocal tract. The vocal tract acts as a filter and produces speech by taking from the source. The greater the volume of air that comes out of the lungs the louder the sound produced [23].

In the source filter model of speech production the air from the lungs acts as a source and the vocal tract as a filter and it is used in the linear predictive coding of speech signals. It is based on the idea of separating the source from the filter in the production of speech. This model is used in both encoding and decoding of the LPC and is obtained from the mathematical representation of the vocal tract as a tube of varying diameter. The excitation of the air traveling through the vocal tract is the source. This air is periodic, when producing voiced sounds by vibrating the vocal cords, or it is turbulent and random when producing unvoiced sounds. The encoding process of LPC involves determining a set of parameters for modeling the vocal tract during the production of speech sounds. The decoding process uses the parameters acquired during encoding for building up a synthesized version of the original speech signal. LPC never transmits
any estimates of the speech to the receiver. It only sends the model to
produce the speech and some indications about what type of sound is
to be produced. Sound waves are pressure variations that propagate
through air (or any other medium) by the vibrations of the air
particles. Modeling these waves and their propagation through the
vocal tract provides a framework for characterizing how the vocal tract
shapes the frequency content of the excitation signal.

Fig 1.3 [24] shows how nasal sounds and vocal sounds are
produced from the human body. Two important components of speech
production are velum and larynx and are explained below:

**Velum:** Velum is used to control the acoustic coupling between nasal
and vocal tracts. When the velum is lowered the nasal tract is
acoustically coupled with the vocal tract to produce nasal sounds.
When the velum is drawn up tightly, the nasal tract is effectively
sealed off and the oral tract is coupled to the vocal tract to produce
oral sounds.

**Larynx:** Larynx is used to produce periodic excitations (for voiced
sound) when air is pushed from the lungs through the vocal tract and
comes out of the mouth as speech. The main source of energy is from
the lungs with a diaphragm. When a person speaks air is pushed from
the lungs, passes through the vocal cords, larynx into the vocal tract
and comes out from the oral and nasal cavities as speech.

Glottis is the V-shaped opening between the vocal cords and is the
most important source of sound in the speech production system.
During speech production the vocal cords vibrate, modulate the air
flowing through it and produce a vibrant sound from which voiced and unvoiced sounds are produced. The rate of vibration of the vocal cords gives the information about the pitch of the sound produced. In general, children and women have high pitch (vocal cords vibrate fast) while male adult have low pitch (vocal cords vibrate slowly).

Fig 1.3 Functional components of speech production system

During unvoiced sounds production the vocal cords remain continuously opened and do not vibrate. The vocal tract shape determines the type of sound produced. When a person speaks the
vocal tract will continuously change its shape in a slow manner (in the range 10msec to 100msec) to produce different types of sounds. During the production of voiced sounds, the lungs force the air through the epiglottis thereby the vocal cords vibrate and interrupts the air stream through it and produces quasi periodic pulses of air. So, for voiced sounds the excitation of the vocal tract produces a quasi periodic signal.

1.7 LPC MODEL

The source filter model used in LPC is also known as the linear predictive coding model [22]. It has two key components LPC analysis (encoding) and LPC synthesis (decoding). In LPC analysis the speech signal is examined and is segmented into blocks called frames. Each frame is examined to find:

- Whether the frame is voiced or unvoiced.
- The pitch of each frame.
- Parameters needed to build up a filter that models the vocal tract.

The goal of LPC analysis is to estimate whether the speech signal is voiced or unvoiced, to find the pitch of each frame and to the parameters needed to build the source filter model. These parameters is transmitted to the receiver and the receiver will carry out LPC synthesis using the received parameters and builds a source filter model, that when provided a correct input, will accurately reconstruct the original speech signal.
1.8 MATHEMATICAL MODEL OF SPEECH PRODUCTION

The mathematical model of speech production is shown in Fig 1.4.

In LPC analysis for each frame a decision-making process is made to conclude whether a frame is voiced or unvoiced. If the frame is voiced an impulse train is used to represent it with non zero taps occurring at intervals of the pitch period. In this thesis the method used for the estimation of the pitch period is the autocorrelation method. If the frame is considered as unvoiced the frame is represented using white noise and the pitch period is zero as the vocal cords do not vibrate. So the excitation to LPC filter is an impulse train or white noise. It is important to emphasize that the parameters pitch, gain and filter coefficients will vary with time from one frame to the other [22]. The important points are summarized below:

- The mathematical model of speech production is often called as the LPC Model.
- The model says the input to the digital filter is an impulse train or white noise and the output is a digital speech signal.
- Impulse train is used to produce voiced sounds and the white noise sequence is used to produce unvoiced sounds.

The relation between the mathematical and physical models of speech production is given by

\[
\begin{align*}
H(z) \ (\text{LPC Filter}) & \quad \leftrightarrow \quad \text{Vocal Tract} \\
u(n) \ (\text{Innovations}) & \quad \leftrightarrow \quad \text{Air} \\
V \ (\text{Voiced}) & \quad \leftrightarrow \quad \text{Vocal Cord Vibration} \\
T \ (\text{Pitch Period}) & \quad \leftrightarrow \quad \text{Vocal Cord Vibration Period} \\
UV \ (\text{unvoiced}) & \quad \leftrightarrow \quad \text{Fricatives and Plosives} \\
G \ (\text{gain}) & \quad \leftrightarrow \quad \text{Air volume}
\end{align*}
\]