Chapter 1
INTRODUCTION

Speech is the most important form of human communication. Though communication languages were first conceived over the ages and different ways of communication techniques have been developed and implemented to convey appropriate information across a distance. In the recent years telephone has revolutionized this process and therefore the demand for efficient communication and data storage is continuously increasing.

In recent communication technology developments like speech, image and video signals play a vital role in the Cellular and Mobile Communication. In digital communication world, Signal representation is one of the important parameter. Demand for cellular, Mobile and convenient forms of communication have been an explosion in the use of Cellular and Satellite Telephony, both of which have significant capacity constraints. The main purpose of Speech coding research is to address the problem of accommodating more users over such limited capacity by coding speech before transmitting it across a network.

1.1 MOTIVATION:

Speech processing [52] is used widely in every day’s applications that most people take it for granted, such as Wireless Networks, Cellular Telephony, Telephony system and Telephone answering machines.

Speech coding [3] is used to reduce the amount of information needed to represent a speech signal. Speech coding has become the most effective area of research particularly in the field of Signal processing for the past few years.

Due to the development in signal processing technology and powerful ideas, the research had been started in the 1980’s. Speech coding [90] has provided a reasonable solution in handling large volumes of information that has to be carried out from one point to another point. This often leads to the increase in
telecommunications links even with the enormous channel capacities of various transmission systems. Due to more number of users and less available bandwidth, engineers and researchers are trying for various approaches and methods in compressing the speech signals.

Due to its importance and increase in demand, researchers are trying different approaches to improve the Quality of the speech signal after receiving the compressed Noisy speech signal through the channel.

One of the basic approaches for improving the quality of speech signal is to adopt different methods of filtering techniques. Thus, the researchers have started introducing enhancement techniques like Spectral Subtraction and Kalman filtering.

For the past few years Spectral Subtraction and Kalman filtering methods have become popular in the research field of navigation because of its better estimation characteristics. Since then, electrical engineers and researchers are trying to manipulate its advantages to useful purpose in speech processing.

Consequently, today it had become a popular filtering technique for estimating and resolving redundant errors containing in speech. The objective of this thesis is to enhance the compressed noisy speech signal using Spectral Subtraction and Kalman filter techniques and is compared with the performance of newly proposed Recursive filter. Design of the new filter is based on Kalman filter concepts.

1.2 THE ACOUSTICAL ENVIRONMENT:

The acoustical environment [46] is defined as a set of translations which influence the speech signal until it is in digital form. There are two main sources of distortion: additive noise and channel distortion.

Additive noise can be a fan running in the background, a door slam, a conversation among people, vehicles noise etc. the additive noise can be stationary or non-stationary. Stationary noise is one which is generally coming out from a computer fan or air conditioning machine.
A non-stationary noise, caused by door slam, radio, Television has statistical properties that change over time. A signal received with speaker very close to the mouth then microphone will receive fewer amounts of noise and reverberation. However, if the microphone is far away from the speaker then it can pick up a lot of noise and/or reverberation.

Reverberation is generally caused by the reflection of acoustic waves on the walls and other objects which can dramatically alter the quality of speech signal.

1.3 SPEECH:

Speech is used to communicate information from one person to another person. The important concepts of the speech communication are speech production and perception.

Speech is the process associated with the production and perception of the noises used in the spoken language. However there are large numbers of disciplines and studies made on speech and the speech sounds, including Acoustic, Psychology, Speech pathology, Linguistic, Cognitive science and Computer science.

1.4 SPEECH PERCEPTION:

Speech perception is an ability of human being who can able to interpret, understand and recognize the sounds used in the language. The study of the speech perception is related to the “phonetic field and phonology”.

Speech perception is also defined as how best the human beings can recognize the speech sounds and make use of this information to understand the spoken language. The research on speech, have applications in building of computer systems which can recognize the speech, as well as to improve the recognition for hearing impaired persons.

There are lots of Biological and Psychological factors which can affect the speech: disorders with the lungs, vocal cords, respiratory affections among others.
1.5 SPEECH COMMUNICATION:

Speech is the most important form of human communication. For that reason, there is a big trend to increase the quality of speech signal and improve telecommunications limitations. However most of the people use the communication devices like Telephones, Mobiles, Internet etc., as their primary goods and the customers demand for more coverage and Quality.

*However, the background noise is an important handicap. If it is joined with other distortions, it can seriously damage the service Quality.*

However, still today the computers have a lack of human abilities like speaking, listening, understanding and learning. As J. Benesty (2005)[26] described that most of the people live in a noisy environments and all the applications (telecommunications, hands-free communications, recording, human-machine interfaces etc) require at least one microphone, where there is every Possibility of Signal to be corrupted by background noise and reverberation.

Speech signal corrupted by background noise can be “cleaned” with digital signal processing tools before it is played out, transmitted, or stored. Speech processing is the study of speech signals and the signals are usually processed in a digital representation and thereby speech processing can be observed as the intersection of “digital signal processing and natural language processing”.

1.6 SPEECH CODING:

Speech coding is a method of reducing the amount of information needed to represent a speech signal [1]. For the past few years, Speech coding has become a major active area of research. Due to the development in digital signal processing technology and prevailing ideas, the subject had been started in the 1980’s. Speech coding can provide a solution for handling huge volume of information which has to be communicated from one point to another point. This often leads to the increase in wide communications links even with the enormous channel capacities of various transmission systems.
However, in the recent years of Wireless communication, the use of speech coding techniques are essentially for less transmission bandwidth and also voice storage and various multimedia applications and that requires digital speech coding [14].

The advantages with coded speech signals are:

- The compressed signals are less sensitive to channel noise.
- Coded signals are easier to error-protect, encrypt, multiplex and packetize.
- Coded speech signals are more efficient transmission over bandwidth constrained channels due to lower bit rate.

Speech coding [28] [29] is for compression of speech (into coded samples) for transmission with speech codec’s that use audio signal processing and speech processing techniques.

1.7 SPEECH ENHANCEMENT:

Enhancement is defined as improvement in the value or quality of something or it is simply means that improvement in intelligibility and/or quality of a degraded speech signals by using Digital signal processing tools.

Speech enhancement [27] deals not only to reduce the noise but also to dereverberation and separation of independent signals. Since this field is fundamental for many researchers and there is a great interest to the industry which is always looking for new solutions that are both effective and practical. This is a very difficult problem for two reasons:

First, the nature and characteristics of the noise signals can change dramatically in time and between applications. It is also difficult to find algorithms that really work in different practical environments.

Second, the performance measure can also be defined differently for each application. Generally there are two criteria’s to measure the performance: Quality and Intelligibility of the signal. It is very hard to satisfy both at the same time.
“Speech enhancement is a critical area of speech processing” and the goal is to improve the Quality, Intelligibility and/or Pleasantness of a speech signal. The most common approach in speech enhancement is noise removal, where by estimation of noise characteristics, it is possible to cancel noise components and retain only the clean speech signal.

1.8 ATTRIBUTES OF SPEECH CODERS:

Many speech coding algorithms are discussed in the literature. The performance of the algorithms depends on the type of application and also depends on scenario and the constraints of that application. The main goal of a speech coding algorithms is to minimize the bit rate.

Coding of the speech signals has major importance in many applications. It is not the only attribute of importance and indeed other attributes may have high impact on the applications. The main attributes [4] [5] [7] [11] of a speech coder are:

**Subjective Quality:** This is referred as Quality of the reconstructed speech signal at the receiver. Generally the reconstructed speech signal may not necessarily correlate to objective measures such as the Signal to Noise Ratio. Quality of the speech signal may be further divided into intelligibility and naturalness of the reconstructed speech signal. Quality of the speech signal is also referred as the ability of the spoken word to be understood; the latter refers to the “human-like" rather than “Robotic" or “Synthetic" characteristic of many current low bit rate coders.

**Complexity:** The computational complexity is one of the basic requirements of the algorithm design. More over coders that are used to reduce the bit rate require greater algorithmic complexity.

**Memory:** The memory storage requirements are also related to the algorithmic complexity. Template-based coders require large amounts of fast memory to store algorithm coefficients and waveform prototypes.
**Bit rate:** This is defined as number of bits per second (bps) that is required to encode the speech samples into a data stream.

**Delay:** Most of the communication systems provide a delay which is inevitable in a speech coding. It is not only due to the algorithmic design complexity, but also due to the some of the basic process requirements of the specific algorithm. Performance of the real-time speech coders depends on the coding delay and hence all these must be minimized in order to achieve acceptable levels of performance.

**Error Sensitivity:** Complex algorithms that are designed to achieve lower bit rates will often produce bit streams which are more susceptible to channel or storage errors. Hence they may manifest themselves in the form of noise bursts or other artifacts.

**Bandwidth:** It is the range of frequencies by which the coder is able to faithfully reproduce the signal. Applications such as mobile and cellular communications are able to accept a lower bandwidth and the possibility of compromising with the speech intelligibility.

**1.9 OBJECTIVES OF THE THESIS:**

The main Objectives of this thesis are:

1. To achieve High perceived quality (how well a human perceives the audio signals)
2. To achieve High measured intelligibility (The message is understood)
3. To achieve Low bit rate (bits per second of speech)
4. To achieve less Spectral distortion (dB), Computational complexities (K Flops) and Memory requirements (Floats).
5. To achieve Robustness to transmission errors (e.g. intermittent cuts in the channel)
6. To analyze the unconstrained vector quantization with respect to spectral distortion (dB), Computational complexities (K Flops) and Memory requirements (Floats).
7. To analyze Multi stage vector quantization (MSVQ) with respect to spectral distortion (dB), Computational complexities (K Flops) and Memory requirements (Floats).

8. To analyze and compare the performance of unconstrained and MSVQ techniques with respect to spectral distortion (dB), Computational complexities (K Flops) and Memory requirements (Floats).


10. To analyze the Performance of speech signal corrupted by White Gaussian Noise using the method of “First Enhancement and then compression” using Spectral Subtraction and Kalman filter at 2 dB, 5 dB, 10 dB, 15 dB, 20 dB, 25 dB and 30 dB. The performance and quality is measured in terms of Pitch, Formants, Signal to Noise Ratio (SNR in dB) and Mean Opinion Score (MOS). After looking into the performance of these filters, it is further decided to incorporate Real World Noise signals which have a high impact on the listener. So the research has been further continued.

11. To analyze the Performance of speech signal corrupted by Real World background noises like “Factory noise, Fire Engine noise, Machine Gun noise, Vehicle noise, Ambulance noise, pink noise, traffic noise and Volvo bus noise etc.,” using the method of “First compression and then enhancement”. The performance and quality is measured in terms of Pitch, Formants, Signal to Noise Ratio (SNR in dB) and Mean Opinion Score (MOS).

12. To analyze the Performance of speech signal corrupted by Real world background noises as stated above using the method of “First enhancement and then compression”. The performance and quality are compared with reference to Pitch, Formants, Signal to Noise Ratio (SNR in dB) and Mean Opinion Score (MOS).
13. Based on the concepts of Wiener filter and Kalman filter, it has been proposed a new filter called “RECURSIVE FILTER” and analyze the performance of compressed noisy speech signal with the proposed filter.

14. To apply the proposed filter for above said two methods with white Gaussian noise at various dB and with Real World Noise signals.

15. To calculate the performance of the recursive filter in terms of pitch, formants, SNR and MOS.

16. To draw final conclusions from the output tables, graphs, figures and store the audio files.

1.10 LITERATURE SURVEY:

Over the past few decades, most of the attention was given to system modeling of speech signals and has been incorporated in several commercial speech coding standards.

First scalar quantization is introduced and has come out with some of the disadvantages like space-filling, the shape advantage, and the memory advantage. Most of the researchers have proposed that the quantization technique used should have less computational and memory requirements and it should not result in suboptimal quantization performance of intelligibility. Speech coders operating at low bit rates demand efficient encoding of Linear Predictive Coding (LPC) coefficients. Line Spectral Frequencies (LSF) parameters are currently one of the most efficient choices of transmission parameters for the LPC coefficients [8].

Multi Stage Vector Quantization (MSVQ) [97] can achieve very low encoding and storage complexity when compared to unconstrained vector quantization. This thesis has proposed a new technology to design the decoder codebook, which is different from the encoder codebook to optimize the overall performance. The improvement in performance is achieved without affecting the encoding complexity, but with a considerable increase in storage complexity of decoder.

As proposed by Dr. M. Satya Sai Ram [1-3] a 3-stage 2-switch 3-part Multi Switched Split Vector Quantization (MSSVQ) has provided better performance in
term of spectral distortion, complexity and memory requirements is less when compared with Split Vector Quantization (SVQ) and Multi Stage Vector Quantization (MSVQ). Whereas MSVQ has better spectral distortion but complexity and memory requirements are more but the quality of the signal is less. So in order to enhance the performance of MSVQ to get better quality of the signal without losing the information, speech enhancement techniques are proposed before transmitting the signal.

Vector Quantization is the process of approximating continuous-amplitude signals by digital (discrete-amplitude) signals [11], is an important aspect of data compression or coding. Speech coding is concerned with the reduction of the number of bits necessary to transmit or store analog data which is subjected to distortion or fidelity criterion.

Generally to synthesize speech signal [72] [73] autoregressive system is excited by one of the two excitation sources: train of impulses is used for voiced sounds, and random noise is used for unvoiced sounds. The parameters of autoregressive system are obtained by Linear Prediction (LP) analysis and are represented by an all-pole synthesis filter which is analyzed by overlapping segments of the speech signal.

Many researchers have given a detailed review of Linear predictive coding (LPC) which can be found in [1] [9] [4] [14]. A major issue in LPC is the quantization of LP parameters [5] [10]. The performance of the Coding systems is significantly better if they operate on sets of symbols or vectors rather than on individual symbols. The concept of representing in vectors has resulted in the development of vector coding techniques, and specifically, Vector Quantization [2] [11].

In Vector Quantization [76-78], a vector is formed when a signal is encoded by the index of its best match in a codebook of prototype vectors (also called code vectors). The decoder then uses this index to retrieve the best match code vector from the codebook. The retrieved code vector is therefore used to reconstruct the original vector. The concepts of Vector Quantization are extensively used in several areas like speech, image, and video coding. Hence, the use of vector quantization at source
based system model has influenced coding of speech signals [82] and enabled to get high quality speech coding at very low bit rates.

An optimal vector quantizer [15] usually operates with a single large codebook without any constraints which are imposed on its structure. Generally to design VQ codec for transparent (high quality) speech coding then the vector dimensions and codebook sizes required to implement are very large. Typically, a vector of 8–12 Linear Prediction parameters are derived from 10–20 ms segment of appropriately windowed speech and these are coded with at least 24 bits so that quality of the reconstructed segment is maintained [1] [4] [5].

“Unconstrained vector quantizer” [4] will require sample vectors in its codebook to encode these Linear predictive parameters. It is observed that encoding complexity and the memory requirements are large and hence many researchers have proposed structurally constrained vector quantization techniques which can reduce the complexity. These proposed techniques are implemented and resulted in achieving the signal with degradation in the reconstruction Quality when compared to the optimal Vector Quantization technique.

As per Dr. M. Satya Sai Ram [1-4], implementation of various VQ of LP parameters are discussed. In multistage vector quantization (MSVQ), a vector of LP parameters is encoded by Multiple Stage VQ encoders. Generally encoders are arranged in a cascaded structure so that each stage encoder encodes the error between the original vector and the reconstruction generated by all preceding VQ stage encoders [2]. The sub optimality of a MSVQ arises from

i) The use of multiple codebooks to generate the reconstruction,

ii) Stage by Stage (sequential) search procedure for encoding the vectors and

iii) The use of the traditional sequential design algorithm for generating the stage codebooks.

To obtain the optimal criterion joint full search is preferred rather than the sequential search so that its complexity would be similar to that of the unconstrained
vector quantization. Keeping in view of system constraints and computing capability, sequential search may be replaced by other improved techniques.

Speech enhancement is used to improve the quality of the speech signal by using various enhancement techniques and enhancement algorithms [17]. For the past few years there has been considerable attention on enhancing the speech signal degraded by additive background noise.

Additive noise suppression has various applications depending on the nature of the environment. General applications like using the mobile in a noisy environments conditions like street vehicle noise, in trains, near factory machines, machine gun, fire engine noise or in a moving car [25] etc is an obvious application and removing the background noise while transmitting speech from the cockpit of an airplane to the ground station or to the cabin which may also be the part of enhancement.

The spectral subtractive algorithm is one of the first algorithms proposed for enhancing the noisy speech signal degraded by additive background noise and it has gone through many modifications from time to time [22][24]. In this thesis the main objective is to analyze of the Spectral Subtraction technique that have been proposed for enhancement of noisy speech signal degraded by additive background noise and acoustic Noise. In the past few decades most of the researchers [21] [23] has introduced Speech enhancement techniques and explained basic Spectral Subtraction technique and also given various modified versions of Spectral Subtraction till date.

Further the researchers have continued to focus on filtering techniques and has come out with second type of enhancement technique called Kalman filter method for speech enhancement and was presented by Paliwal (1987) [19].

As per Paliwal [19] this method is best suitable for reduction of white noise to comply with Kalman filter assumption. So to derive Kalman filter equations it is normally assumed that the process noise (the additive noise that is observed in the observation vector) is uncorrelated and has a normal distribution [25]. This assumption leads to whiteness character of this noise.
However, there are different methods developed by the researchers to fit the Kalman filter approach to noisy environments like in streets vehicle noise, in trains, near factory machines or in a moving car is an obvious application. The ultimate task is to remove the background noise [18] before transmitting speech signal from one place to another place or from cockpit of an airplane to the ground station. In this research the simulated vocoder (LPC) using Matlab was implemented and simulated for attaining the compression [26]. The results obtained from unconstrained VQ was compared with other implemented speech compression using MSVQ and measured in terms of spectral distortion, complexity and memory requirements.

In this thesis it has been proposed to analyze Spectral Subtraction and Kalman filter methods for enhancement of the compressed noisy speech signal and the result of one method was compared with other and further continued with newly proposed filter approach of speech enhancement.

1.1 SPEECH PRODUCTION PHENOMENON:

The above Figure 1.1 shows functional components of vocal system which shows how the nasal sounds and vocal sounds are produced from the human body.
Two important components for speech production are:

**Velum**: To control acoustic coupling between nasal and vocal tracts. When the vocal tract is lowered, nasal tract is acoustically coupled with vocal tract to produce nasal sounds. When it is drawn up tightly, nasal tract is effectively sealed off for the production of non-nasal sounds.

**Larynx**: To provide a periodic excitation (for voiced sound)

Air is pushed from the lung through the vocal tract and comes out of the mouth as speech. The diaphragm is the source of energy to the lungs. During the speech, the airflow is forced to move through the glottis and in between the vocal cords and then to the larynx to the three main cavities of the vocal tract, the pharynx and the oral and nasal cavities.

The air passes through the oral and nasal cavities and exits through the nose and mouth, respectively. The V-shaped opening between the vocal cords is called the glottis and this is the most important form of source in the vocal system. The vocal cords can vibrate in many ways during speech. The main purpose of the vocal cord is to modulate the airflow by rapidly opening and closing, which produces buzzing sound that can even produce vowels and voiced consonants.

For certain voiced sound the rate at which the vocal cords vibrate determines the pitch of your voice. More often women and young children tend to have high pitch (fast vibration) while adult males tend to have low pitch (slow vibration). Vocal cords do not vibrate but remain constantly opened for certain fricatives and plosive (or unvoiced) sounds. The shape of the vocal tract determines the sound. As the human speaks with different sound, the vocal tract changes its shape accordingly. The shape of the vocal tract changes relatively slowly (on the scale of 10 msec to 100 msec). The amount of air coming from the lung determines the loudness of the voice. Figure 1.2 represents the block diagram of the human speech production. To produce voiced sounds, the lungs press the air through the epiglottis and the vocal cords vibrates. These vocal cord vibrations will interrupt the air stream and produce a quasi-periodic
pressure wave. So for voiced sounds, the excitation of the vocal tract is a Quasi-periodic signal.

Figure 1.2 Block diagram of Human Speech Production

The excitation of the vocal tract is noisier in the case of unvoiced sounds. These two signals are passed through mouth or nose to produce the speech.