CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Speech technology provides significant revenue opportunities through the introduction of new speech processing services. Speech is the most natural way for users to communicate from person-to-person and in the future, from person-to-system (Westall 1997). In Automatic Speech Recognition (ASR), much more fundamental knowledge is needed to allow machines to be made to approach the capability of humans.

The emerging applications of speech communication are demanding increasing levels of performance in noise adverse environments. The external factors contribute some degree of variability to the speech signal. These sources of variability must be carefully considered when developing applications based on speech recognition technology, as it is these characteristics that will ultimately determine whether a speech recognizer will work well in the real world. The contamination on the speech not only affects its audio quality but also precludes many further high level speech processing tasks such as speech signal coding and recognition (Ramirez et al 2005).

For human-to-human interaction, or for human to machine platforms, single channel noise suppression of speech signals offers a computationally efficient option for improving the effectiveness of communication in noisy environments.
The speech signal not only conveys the linguistic information (the message) but also a lot of information about the speaker himself: gender, age, social, and regional origin, health and emotional state and, with a rather strong reliability.

In many applications, the received speech signal contains periodic or non-periodic form of noise component. Depending on the amount of the noise, the quality of the received speech signal can range from being slightly degraded to being annoying to listen to, and finally to being totally unintelligible. It is necessary to enhance the quality of speech that has been degraded by background noise discussed by Ergun Ercelebi (2004).

In a natural environment, noise is inevitable and ubiquitous, speech signals are generally immersed in noise and can seldom be acquired and processed in pure form. Noise can profoundly affect human-to-human and human-to-machine communications, including changing a talker’s speaking pattern, modifying the characteristics of the speech signal, degrading speech quality and intelligibility, and affecting the listener’s perception and machine’s processing of the recorded speech. The objective of noise reduction is to restore the original clean speech from the mixed signal.

Speech quality is a major contributor to the telecommunication user’s perception of quality of service. As communications networks become more heterogeneous, identifying the root cause of voice quality problems can be a challenging task. The evaluation and assurance of speech quality has, consequently, become critically important for telephone service providers.

Traditionally, user opinion is measured offline using slow and costly subjective listening tests. In the most common test, listeners rate the speech they just heard on a five-point opinion scale, ranging from “bad” to “excellent.” The ratings are assigned integer scores ranging from 1 for “bad”
to 5 for “excellent.” The average of these scores, termed Mean Opinion Score (MOS), is widely used to characterize the quality of telephony equipment and services. As an alternative to subjective measurement, machine-automated “objective” measurement provides a rapid and economical means to estimate user opinion and makes it possible to perform real-time speech quality measurement on a network wide scale. Objective measurement can be performed either intrusively or nonintrusively.

The speech signals coming from distant sound sources are weakly received by microphones of the devices. In this case, Signal to Noise Ratios (SNRs) are seriously degraded compared to those under close-talking conditions. Thus, environmental noises have to be removed in order to achieve high-performance speech recognition accuracy.

1.2 CHARACTERISTICS OF SPEECH

Speech signals are composed of sequence of sounds. Sounds can be classified into following three distinct classes according to mode of excitation (Rabiner 2004):

i) Voiced sounds are produced by forcing air through glottis with the tension of vocal cords adjusted so that they vibrate in a relaxation oscillation, thereby producing a quasi-periodic pulse of air which vibrates the vocal tract;

ii) Fricative or unvoiced sounds are regenerated by forming a constriction at some point in vocal tract and forcing air through constriction at a high enough velocity to produce turbulence; and

iii) Plosive sounds result from making a complete closure and abruptly releasing it.
The speech organs are divided into three main groups: the lungs, larynx and vocal tract. The lungs act as a power supply and provide air flow to the larynx stage of the speech production mechanism. The larynx modulates air flow from the lungs and provides either a periodic puff-like or a noisy air flow source to the third organ group, the vocal tract. The vocal tract consists of oral, nasal and pharynx cavities, giving the modulated air flow its “color” by spectrally shaping the source.

Sound sources can also be generated by constrictions and boundaries, that are made within the vocal tract itself, yielding in addition to noisy and periodic sources, an impulsive air flow source. Following the spectral coloring of the source by the vocal tract, the variation of air pressure at the lips results in the traveling sound wave that the listener perceives a speech.

There are then three general categories of the sources for speech sounds: periodic, noisy and impulsive, although combinations of these sources are often present (Lawrence Rabiner 1993).

Such distinguishable speech sounds are determined not only by the source but by different vocal tract configurations and how these shape combines with periodic, noisy and impulsive sources. These more refined speech sound classes are referred to as phonemes, the study of which is called phonemics. A specific phoneme class provides a certain meaning in a word, but within a phoneme class, as we will see in a moment, there exist many sound variations that provide the same meaning. The study of these sound variations is called phonetics. Phonemes, the basic building blocks of a language, are concatenated, more or less, as discrete elements into words, according to a certain phonemic and grammatical rules.
1.3 SPEECH RECOGNITION IN ADVERSE ENVIRONMENTS

The performance of most speech recognition systems, whose designs are predicated on assumptions about the ambient conditions, such as low noise background, degrades rapidly in the presence of noise and distortion. It is therefore important to consider the problem of speech recognition in adverse environments that are inevitable in real-world applications (Rabiner 1993). The deployment of automatic speech recognizer often requires that the machine have the capability to adapt to new application environments. Acoustic as well as electric noises are inevitable that resist signal contamination and undesirable interference is not only analytically interesting but practically important.

Hands-free operation is a very important feature for speech activated systems. Speech detection provides a way to solve the problem for isolated word speech recognition. The endpoint detection improves the isolated word recognition accuracy. Good performance has been reported for moderate noise conditions such as SNRs above 5 dB. However, for severe noise conditions such as SNRs down to -10 dB in some application environments, no reliable detection algorithm has yet been reported discussed by Wu et al (1999).

A speech recognizer often encounters three main causes of adverse conditions: noise, distortion, and (human) articulation effects.

1.3.1 Noise

Acoustic ambient noise is usually considered additive, meaning that the recorded signal is a sum of the speech signal and the ambient noise. High levels of ambient noise are one of the primary concerns for a speech recognizer. Sources of acoustic ambient noise are abundant. For example, in
an office environment, sources of noise include office machinery such as type
writer or printer, Personal Computers (PC), or workstations, which are usually
equipped with moving components like disks and fans, telephone ringing and
background conversation of other people. These noise sources often provide
enough acoustic noise to cause severe performance degradation of a speech
recognizer.

1.3.2 Distortion

Aside from the additive contamination due to noise like signals, the
uttered signal inevitably undergoes a series of spectral distortions before
being recorded and processed for speech recognition. The room in which the
speech recognizer unit is deployed almost certainly has a varying degree of
reverberation that can alter the signal spectrum. The microphone transducer,
depending on its type and mounting position, also can significantly distort the
speech spectrum. When the transducer configuration used in testing is
different from the one used during training of the reference patterns, the
mismatch in spectral distortion becomes one of the major problems.

1.3.3 Articulation Effects

Many factors affect the manner of speaking of each individual
talker. Even the psychological act of communicating with a speech
recognition machine could make the talker produce a noticeable difference in
his or her sound formats and rhythmic pattern. Characteristic changes in
articulation due to environmental influence known as the Lombard effect, can
also be dramatic. When a talker speaks in an environment with a masking
noise of 90 dB, the first format of a vowel often increase while the second
format decreases, resulting in a potential shift in the vowel space.
1.4 PRE-PROCESSING METHOD FOR SPEECH RECOGNITION

Signal processing provides the theory, the methods and the tools for such purposes as the analysis and modeling of signals, classification and recognition of patterns, extraction of information from signals, synthesis and morphing of signals. Morphing is creating a new voice or image out of the existing samples.

One of the most natural communication tools used by humans is their voice. It is hence natural that a lot of research has been devoted to analyzing and understanding human uttered speech for various applications. The most obvious one is Automatic Speech Recognition (ASR), where the goal is to transcribe a recorded speech utterance into its corresponding sequence of words.

ASR in noisy environments remains a challenging problem since there are many possible types of environmental distortion, and it is difficult to compensate for all of these distortions accurately. The primary reason for poor recognition performance in noise is the mismatch between training and test conditions.

The overall performance of a speech recognizer highly depends upon the quality and robustness of the acoustic and features extracted from the speech wave as a front end of recognition algorithms. Speech parameterization is a challenging problem. As speech signal is highly variable, our knowledge about human speech perception is still limited, despite recent substantial advances, and most of the methods proposed so far are not very robust against variations in the speech signal. A speech recognizer is designed to recognize one word from a set of words or phrases
specified in a vocabulary. This is a difficult task, and not even a person is 100% accurate (Westall 1997).

The noisy speech signal is formed additively by speech and noise signals in which the noise is generated by environmental sources such as Airport, Babble, Car, etc. In real environments, complete noisy cancellation is not feasible as it is not possible to completely track varying noisy types and its characteristics change with time. When the adverse condition is due to additive noise alone, well established Voice Activity Detection (VAD) method or Speech Enhancement method are used to suppress the noise before applying the recognition algorithm.

The function of the speech recognizer is to convert spoken input into grammatically correct text as constrained by the recognizer vocabulary and grammar model. In theory, the speech input, to the recognizer, to be totally unconstrained, that is, natural language. The output of the speech recognizer is the text string that is most likely to have been spoken based on the recognizer’s vocabulary and grammar (Rabiner 1993). VAD is used to distinguish speech from noise and is required in a variety of speech communication systems. VAD is more critical for non stationary noise environments since it is needed to update the constantly varying noise statistics affecting a misclassification error strongly to the system performance.

An Automatic Speech Recognition (ASR) system is applied in noisy environments, Voice Activity Detection (VAD) is crucial to the performance of the overall system. The employment of the VAD for ASR will minimize physical distractions and make the system convenient to use. The VAD is a very important front-end processing in practical Automatic Speech Recognition (ASR) systems. A well-designed voice activity detector can greatly improve the performance of an ASR system in terms of accuracy and
speed. Moreover the employment of VAD for ASR on embedded mobile systems, such as PDA, smart phone, wireless car kits, will minimize physical distractions and make the system convenient to use (Bian Wu et al 2005).

The required characteristics of an ideal voice activity detector are: reliability, robustness, accuracy, adaptation, simplicity, real-time processing and no prior knowledge of the noise. Among them, robustness against noise environments has been the most difficult task to accomplish. In high SNR conditions, the VAD algorithms proposed to date show satisfactory performance; while in low SNR environments, all performances degrade to a certain extent. Robust VAD in noisy environments remains an unsolved problem.

An ASR system work in various environments with various noise types at various SNRs. So the VAD should be robust in all kinds of conditions. At the same time, the VAD algorithm should be of low complexity, which is mandatory for real-time. Therefore simplicity and robustness against noise are two essential characteristics of a practicable voice activity detector for an ASR system.

Enhancement of noisy speech improves the quality and intelligibility of voice communication in noisy environments such as for mobile and hands-free phones used in noisy public venues as in busy streets, moving cars and trains, noisy conference halls, cafeteria, noisy shops/markets, airports, factory floors (Qin yan et al 2008). In speech enhancement, initial information about speech and noise is needed. Exploiting the available a priori speech and noise information can further improve the performance of a speech enhancement system.

Speech enhancement methods that have been broadly grouped depending on whether a single or dual-channel approach is used. In single-
channel approach, noise characteristics must be eliminated during period of silence and it is assumed that noise characteristics are stationary during speech activity. The enhancement of noisy speech in a one microphone system is a long-standing problem, but the quality of the enhanced speech achieved so far is rather unsatisfactory. The main reason is the difficulty to provide an appropriate speech signal parameterization. The speech signal is a very complex one and no statistical models exist, which describe its properties precisely enough discussed by Marcin Kuropatwinski et al (2002).

1.5 MOTIVATION FOR THE RESEARCH

The objective of the thesis work is to maximize the Speech Recognition Accuracy in adverse noise conditions. The following observations are made based on the literature survey.

Speech recognition systems exhibit higher error rates for children due to variabilities in vocal tract length, formant frequency, pronunciation, and grammar. In the context of recognizing speech while children are reading out loud, these problems are compounded by speech production behaviors affected by difficulties in recognizing words that cause pauses, repeated syllables and other phenomena (Andreas Hagen et al 2007).

Noise power spectrum estimation is a fundamental component of speech enhancement and speech recognition systems. The robustness of such systems, particularly under low Signal to Noise Ratio (SNR) conditions and nonstationary noise environments, is greatly affected by the capability to reliably track fast variations in the statistics of the noise case discussed by Israel Cohen (2003).
A non-stationary signal is a signal with a quickly changing power spectrum over different frequency bands. For a segment of a non-stationary signal, it is a challenge to decide whether it is speech or noise.

There are many reasons why speech recognition is often quite difficult:

1. Natural speech is continuous; it often does not have pauses between the words. This makes it difficult to determine where the word boundaries are among other things.

2. Natural speech can also change with differences in global or local rates of speech, pronunciations of words within and across speakers and phonemes in different contexts.

3. Large vocabularies are often confusable.

4. Recorded speech is variable over room acoustics, channel characteristics, microphone characteristics and background noise.

Traditional noise estimation methods, which are based on voice activity detection, restrict the update of the estimate to periods of speech absence. Additionally, VADs are generally difficult to tune and their reliability severely deteriorates for weak speech components and low input SNR. So the update rate of the noise estimate is essentially moderate.

The SNR and noise powers estimations are important problems in speech processing. The SNR is a main measure of the speech quality index which is frequently used in the data collection and classification task and the local noise powers estimation is used in speech enhancement systems. In many applications, SNR and local noise power estimations of noisy speech are highly difficult because neither clean reference signal nor speech activity is given (Kazuya Takeda et al 2005).
Conventionally, the SNR and local noise power estimations are based on Voice Activity Detection (VAD). However, these methods work well only at high SNR conditions and therefore, its application is limited. The basic idea of this work is using the natural property of the speech signal, which always contains the silent duration to model and estimate them via a probabilistic mixture model. Furthermore the estimated distribution parameters are being used in the SNR and local noise power estimation without using VAD.

The objective of noise reduction in speech enhancement and in Automatic Speech Recognition differs. While the first tries to improve the intelligibility of the speech and/or to ease listener’s fatigue, the latter is effective if it succeeds in closing the gap between noisy and clean speech recognition accuracy. Nevertheless, a correlation can be expected between the improvements in speech quality on the one hand, and the improvement in recognition accuracy on the other hand.

One solution to mitigate the performance degradation of Automatic Speech Recognition systems in noisy operating environments is to enhance the observed speech prior to the recognizer preprocessing and decoding operations. The enhancement step can be performed independently of the recognition process by one of the manifold available speech enhancement algorithms.

1.6 FOCUS OF THE THESIS

The focus of this thesis is to develop a Robust Automatic Speech Recognition system with better Recognition Accuracy based on Voice Activity Detection Algorithm and / or Speech Enhancement Algorithm.
Since simplicity and robustness are two main challenges for VAD algorithms for ASR systems, the proposed VAD algorithm discussed in chapter 4 are devoted to designing a deliberate VAD algorithm suitable for ASR systems. This algorithm is not only of robustness in various noisy environments but also has a low computational complexity. In the proposed algorithm, a noise estimation is introduced to describe environmental noise signals which is used to update noise for every frame.

There have been numerous VAD algorithms proposed to date. In these algorithms different discriminating features are incorporated to cope with various environmental noises. Among them, the energy based features are most popular because of its simplicity and effectiveness. Moreover it is difficult to construct a good combination algorithm suitable for all environments. Some conventional features, such as zero-crossing rate, pitch based detection, and weak fricative detection are not used because they are not immunized against noise. The computation costs of these discriminating features are low in themselves, which makes them suitable for ASR.

Popular speech enhancement algorithms have been designed primarily to improve intelligibility and/or quality of the speech signal without consideration of what effect that may have on other speech processing systems.

The speech enhancement algorithms, even the most sophisticated ones, do not improve speech intelligibility. The reason is the fact that, it does not have a good estimate of the background noise spectrum, which is needed for the implementation of most algorithms. For that, accurate voice-activity detection algorithms are required. Much progress has been made in the design of Voice Activity Detection algorithms and noise-estimation algorithms which are capable of continuously tracking, at least, the mean of the noise spectrum. Noise-estimation algorithms are known to perform well in
stationary background noise environments. Evidence of this, a small improvement (<10 %) in intelligibility was observed with speech processed in car environments, but not in other environments (e.g., babble). The small improvement was attributed to the stationarity of the car noise, which allowed for accurate noise estimation (Hu and Loizou 2006).

The accurate noise estimation can contribute to improvement in intelligibility, but that alone cannot provide substantial improvements in intelligibility, to track accurately the spectrum of non-stationary noise. The absence of intelligibility improvement with existing speech enhancement algorithms is not entirely due to the lack of accurate estimates of the noise spectrum.

The most challenging case is the single channel speech enhancement where only noisy speech is available for recovering the clean speech. The aim is to minimize the effect of noise and to improve the performance of voice communication systems when input signals are corrupted by background noise.

1.7 ORGANIZATION OF THESIS

The thesis contains totally four contributing chapters which discuss about speech recognition accuracy.

Chapter 2 discusses the concept Hidden Markov Model, it is a type of model based on a statistical representation, and this helps the recognizer to cope with most of the variability in the way people speak. The recognizer aims to identify, from observing a sequence of speech features, which of the stored Markov models is most likely to have produced these observations. Implementation of this methods and results are discussed.
In Chapter 3, Speech Enhancement Algorithm based Automatic Speech are described to enhance the performance of Recognition Accuracy (RA). Implementation and results are discussed to prove the effectiveness of proposed method.

In speech recognition when a word or utterance begins or ends must be specified. VAD is used to disable speech recognition for silence segments. A novel method of VAD is proposed in the Chapter 4 to improve the RA. The method is implemented and its results are discussed.

Chapter 5 discusses about the hybrid algorithm for speech recognition. Hybrid algorithm combines the proposed speech enhancement and voice activity detection algorithm which are described and analyzed thoroughly with RA. Performance analysis of the both techniques are studied.

The concluding remarks and scope for the future extension of the proposed methods are discussed in the Chapter 6.

Finally the thesis is concluded with references and list of publications made out of the research work.