Chapter 1

Introduction

1.1 Automatic Speech Recognition

Speech recognition technology has been deployed in a diversity of applications such as learning aids, medical transcription, commercial products and the telecommunication industry. The last decade has witnessed significant improvements in speech recognition technology. High performance algorithms and systems which are implemented in the laboratory have started to move into commercial deployment.

Robustness in speech recognition systems has improved in recent years with respect to both speaker and acoustical variability. A recognition system shows robustness if its recognition rate does not degrade substantially under mismatched conditions. The need for greater robustness in recognition has become more apparent as these technologies are being transferred to real-world applications. The present recognition systems are embedded in a diversity of applications such as mobile devices, automotive vehicles, industrial and military devices. These environments have significant level of background noise and distortion. Speech is degraded or corrupted due to the presence of additive and convolutional noise, reverberant noise and stressed speech. The performance of even the best state-of-art systems tends to deteriorate when the signal to
noise ratio is extremely low or under mismatched conditions between training and testing.

1.2 Speech Recognition Applications

Recent advances in speech recognition have enabled the deployment of speech recognition technology in diverse applications. Major applications are related to communication industry, commercial and educational tools, medical field and many other areas.

Communication: Speech recognition technology has been embedded in several cellular applications such as personal digital assistants (PDAs), mobile communication devices and voice enabled media player. The use of speech interface will ease the use of these devices. Speech technology enables users to operate the devices without having to look at the screen or halt their actions in order to input the functions. This eliminates the need for palm pen and small keyboard which are often cumbersome and slow.

Speech recognition systems have also been deployed in the automobile industry. Automobiles are embedded with car-navigation systems using advanced speech recognition and text-to-speech capabilities. The system would identify the intended route and guide the user to the destination. In addition, the system would help in handling hands-free functions for automobile accessories.

Commercial and Educational Tools: Commercial applications embedded with speech recognition technology facilitate users in the office environment. The users are able to command and control the desktop using their voices. Functions such as editing and formatting documents, launching applications and managing files are convenient with a speech interface. The dictation system also helps to improve submission of reports in
time, and offers effective communication and accessibility. The speech recognition technology has also emerged as a translation and learning tool for educational purposes. Any person can deliver a talk and the recognizer compiles the matter into textual notes that are accessible by public.

**Medical and Healthcare:** Speech recognition technology has been an effective tool for medical transcription. Tedium tasks such as the electronic medical record (EMR) management is more effective with speech technology. It saves time to perform searches, queries, form filling, data entry and record management by simply using voice in place of conventional keyboard. In addition, users with hearing impairments also benefit from speech recognition technology. The system enables users to communicate efficiently because conversation can be converted into visual text.

**1.3 Speech Recognition Limitations**

The difficulty of the speech recognition task depends on the nature of the input speech. Some of the characteristics that define the complexity of a speech recognition system are now discussed:

**Amount of Speakers:** The complexity of the recognition task depends on the variability of the input signals. A single user system (speaker dependent) has to cope only with the variability of the speech induced by this single speaker (intra-speaker variability). In multi-user systems (multi speaker), this variability increases due to inter-speaker variability. The inter-speaker variability is due to the fact that people do not have the same physiological characteristics and the same manner of speaking resulting in different types of speech signals. The complexity of the recognition task can vary from a single
user system (speaker dependent) to systems that are defined to be used by any user (speaker independent).

**Speaking Rate:** The complexity of the recognition task increases as the speech style becomes more spontaneous.

Some examples of possible speaking styles are listed below:

- **Isolated Word Recognition:** The words are surrounded by periods of silence allowing each word to be separately processed.

- **Continuous Speech Recognition:** The word boundaries are not well defined. Hence the recognition system has to deal with co-articulation effects as the pronunciation of a word is affected by the phonetic context.

- **Spontaneous Continuous Speech Recognition:** The complexity of the task can dramatically increase due to hesitations, grammatical errors, disfluencies, etc.

- **Keyword Spotting:** Some predefined words are detected in any utterances while it rejects those words which are out of vocabulary.

**Vocabulary Size and Complexity:** The size of the vocabulary and its complexity are also important factors of the recognition task. The complexity of the task increases not only with the size of the vocabulary but also when the words of the vocabulary are confusing even if the vocabulary is small.

**Computation and latency:** Any recognition system should have minimum computation and maximum throughput, hence the time taken for feature extraction and computing should not be relatively long.

**Difficulty of the Operating Conditions:** In real world conditions, a lot of perturbations can significantly affect the performance of speech recognizers such as stationary or non-
stationary environmental noises, correlated noises like echo and reverberation, linear and non-linear distortions, channel distortion and cross-talk.

1.4 Speech Recognition in Real World Conditions

Speech recognition systems are generally trained using data that are obtained under controlled conditions. The data is acquired in noise-free environments using high quality microphones. In real world operating conditions, however several factors can degrade the quality of the speech signal as it is shown in Figure 1.1 and therefore reduce the performance of a speech recognition system.

![Figure 1.1 Sources of distortion in real world system](image)

**Distortion due to the Environment:** In real world conditions, the nature of the environment can degrade the quality of the speech signal. Several factors that may occur simultaneously are as follows:

- **Stationary or Non-Stationary Noise:** An ambient noise can be present and, therefore, this noise is added to the speech signal. This noise can be stationary (e.g. the noise of a fan) or non-stationary (e.g. keyboard noise).
- **Reverberation and Echo:** The acoustics of the room can introduce reverberations and echo in the signal.
Acquisition of the Signal: In real operating conditions, the microphones used for the acquisition of the speech signal can introduce several distortions:

- **Linear and Non-linear Distortions:** Low quality microphones can suffer from linear and nonlinear distortions that degrade the quality of the signal.

- **Distance:** The distance from the speaker to the microphone can vary with time (in a hands-free system for example) and this results in a distortion of the amplitude of the signal.

- **Direction:** The direction of the microphone may induce distortion in the speech signal.

Transmission: The transmission of the speech signal can also introduce perturbations as explained below:

- **Filtering:** In order to increase the number of channels that can be transmitted in a fixed hardware solution, the bandwidth of the signal can be limited (e.g., in public switched telephone networks the bandwidth of speech signal is limited from 300 to 3300 Hz).

- **Channel Distortion:** The transmission channel can introduce distortions in the speech signal.

- **Noise:** Some noise can be added to the transmitted signal due to electrical perturbations of the transmission lines.

- **Cross-talk:** When several channels are present in the same transmission medium, they can influence each other.

In this thesis, the main focus is to study the effect of environmental noise in speech recognition so as to develop various solutions for speech based applications.
1.5 Speech Recognition in Noise

The degradation of the performance of a speech recognition system when operated in noisy conditions is due to the mismatch that exists between the training conditions which are generally clean and the noisy operating conditions. Figure 1.2 represents the temporal and spectro-temporal representation of a speech signal under clean and noisy conditions.

The noise modifies significantly the distribution of the value representing the signal, both in the temporal and the spectral domain. Several approaches have been developed to reduce this mismatch between training and recognition. They can be classified in three categories [1]:

![Figure 1.2 Temporal and spectro-temporal representation of (a) clean and (b) noisy speech signal](image-url)
1. **Speech Enhancement**: Before the extraction of the relevant features, the influence of the noise on the speech signal is modeled and the distortion introduced by the ambient noise is reduced.

2. **Robust Feature Extraction**: The features representing the speech signal are designed so as to be less sensitive to the noise conditions. This is achieved by analyzing the influence of the noise on the speech signal and deriving feature extraction methods that reduce the influence of the noise.

3. **Model Compensation**: The noise introduces a mismatch between the training and the recognition. The aim of the model compensation approach is to determine the influence of the noise on the distributions of the speech features and to modify the models used in the recognition to take into account the influence of the noise.

The work carried out in this thesis aims to improve the robustness of the automatic speech recognition system by robust feature extraction. To achieve this, an autocorrelation domain processing is used to derive noise robust features from speech degraded by noise. The autocorrelation function of a signal is related to the signal’s power spectrum through the Fourier transform [2] and it has the following attractive properties:

- **Additivity Property**: If the two signals are uncorrelated, the autocorrelation function of their sum is equal to the sum of their autocorrelation functions [3].

- **Robustness Property**: The autocorrelation function of a white random noise signal is zero everywhere except for the zero time-lag [4].

For broadband noise signals, it is mainly confined to lower time-lags and is very small or zero for higher time-lags [5]. As a result the additive noise does not affect the
higher lags of the autocorrelation function.

The commonly used source-system model of speech production views the speech signal as the output of a linear, time-varying system (or filter) excited by either a white noise source (for unvoiced speech) or a periodic pulse train source (for voiced speech) [6]. The autocorrelation function of the voiced speech signal is also periodic with the same period as the speech signal. Each period of this autocorrelation function contains information about the vocal tract system. Therefore, if a speech signal is degraded by an uncorrelated additive noise that has the aforementioned characteristics, the speech and noise contributions can be easily separated in the autocorrelation domain. It is because of this attractive property, autocorrelation domain processing has been used in this thesis for spectral estimation of the noisy signals [5,7-15].

1.6 Objectives of the Thesis

The aim of the thesis is to increase the robustness of speech recognition systems when operated in adverse conditions.

The main goals of this thesis are:

- To develop techniques that extract relevant speech features to achieve robustness in speech recognition under environmental distortion.
- To improve the existing feature extraction techniques for accuracy and efficiency.
- To achieve high accuracy and recognition rate under different stationary and non-stationary noises.

The baseline speech recognition system is trained under clean (or noise-free) conditions. The degradation of the performance of speech recognition system in adverse conditions is explained by the mismatch that exists between the training and the
recognition conditions. The scope of this thesis is to propose new techniques to extract robust features from distorted speech signal which are less sensitive to ambient noise.

1.7 Major Contributions

The contributions made by the present thesis are aimed at improving the performance of machine speech recognition in the presence of ambient acoustic noise and the quality of speech perceived by human listeners in the same conditions. The proposed techniques are based on processing the degraded speech signal in the autocorrelation domain. It is found that these techniques were able to suppress particular classes of stationary and non-stationary noise which conventional methods fail to handle. Several original contributions have resulted from the research reported in this thesis. These contributions are summarized as follows:

1. A robust feature extraction technique using differentiated relative autocorrelation sequence spectrum (DRASS) has been developed and discussed in Chapter 3. This technique introduces a novel representation of speech for the cases where the speech signal is corrupted by additive noises. In this the robust speech features are computed by reducing additive noise effects via an initial filtering stage followed by the extraction of autocorrelation spectrum peaks.

2. An improved technique of feature extraction based on differentiated relative higher order autocorrelation sequence spectrum (DRHOASS) has been developed and explained in Chapter 4. In the extraction process, initially the lower order coefficients of the noisy speech autocorrelation sequence are discarded and then the effect of noise is further suppressed using a high pass filtering in autocorrelation domain. Finally, the feature vector set of the speech signal is found using the spectral peaks of the filtered autocorrelation sequence. By making the spectral estimate in this
way, an increased robustness to noise is achieved.

3. Comparison of proposed features using differentiated relative higher order autocorrelation sequence spectrum (DRHOASS) with other classical methods in autocorrelation domain is highlighted and is given in Chapter 5. New range of autocorrelation coefficients show an upper hand in recognition rate as compared to other methods.

4. A technique for feature extraction based on channel adaptive relative autocorrelation sequence (CARAS) has been developed and elaborated in Chapter 6. The features extracted by this technique have reduced the effect of additive and convolutional distortions in the speech signal by two stage filtering.

1.8 Thesis Organization

The thesis is organized in seven chapters:

• Chapter 1 gives the brief introduction to the main purpose, structure and major contributions of this thesis.

• Chapter 2 provides a review of the foundations of current statistical automatic speech recognition (ASR) system. The chapter begins by looking at the human speech communication process and highlights the aspects that have contributed to improve the performance of speech signal processing algorithms. This leads to a review of major component of the current statistical ASR system, i.e., process of feature extraction. Then the prevalent techniques applied at the feature extraction stage and a review of noise robust ASR system is given.

• The first major contribution of the thesis is described in Chapter 3. The chapter highlights the autocorrelation domain as possessing favorable properties that are
relevant to noise robust speech processing. In this chapter, a novel feature extraction technique has been proposed that captures characteristics of the vocal-tract response while minimizing contributions from additive uncorrelated noise sources.

- In Chapter 4 performance of a newly proposed feature extraction technique for noise robust speech recognition has been introduced. The novel technique derived in this chapter is based on higher-order autocorrelation based spectral estimation. The new feature vector set based on differentiated relative higher order autocorrelation sequence spectrum (DRHOASS-MFCC) is evaluated and tested on a isolated word recognition task using a range of noises at various signal-to-noise ratios.

- Chapter 5 highlights performance comparisons of DRHOASS-MFCC feature vector set with the existing techniques which are based on the autocorrelation domain processing such as the linear prediction cepstral coefficients (LPCC) method, short-time modified coherence (SMC) method and the one-sided autocorrelation linear prediction coding (OSALPC) method on the basis of autocorrelation order range.

- A new technique for feature extraction which is robust for additive as well as convolutional noise is proposed in Chapter 6. Two stage filtering has been suggested in this technique, one in the temporal domain and the other in the spectral domain to take care of both types of noises i.e. additive stationary noise and convolutional channel noise. New feature vector set named channel adaptive relative autocorrelation sequence (CARAS) has been derived and tested on the isolated word corpus.

- Chapter 7 summarizes the contributions made in this thesis. The future scope of research work in this area has also been suggested.