Chapter 1

INTRODUCTION

1.1 Background

Speech is the primary means of communication between human beings. Speech processing is the mathematical analysis and application of electric signals for information storage and retrieval of that results from the acoustic pressure waves gathered from human vocalization. The area of speech processing includes the natural operation of speech analysis, coding, enhancement, synthesis and recognition. Speech analysis is the learning of the speech production mechanism in order to generate a mathematical model of physical phenomena. Speech coding aims to store speech specific parameter information for subsequent recovery. The process of improving intelligibility and speech quality of noise corrupted speech signal using various algorithms is known as speech enhancement. Speech synthesis is the artificial production of human speech from coded instruction. Automatic Speech Recognition (ASR) is the process of synthesis in reverse defined as the ability of a machine or program to identify the linguistic contents involved in the speech signal.
Speech production is a severely complicated process. It can be defined in basic terms as follows. The normal breathing mechanism allows the air to enter into the lungs. As the air blow out from the lungs through the trachea, the tensed vocal cords within the larynx gets to vibrate by the air flow. The air flow is chopped into quasi-periodic pulses which are then tuned in frequency while passing through the pharynx, the mouth cavity and possibly nasal cavity. Depending upon the position of the various articulators different sounds are produced [Rabiner and Juang(1993)].

Speech recognition research has a history of more than 60 years. For reasons ranging from technological curiosity about the mechanisms for mechanical realization of human speech capabilities, to the desire to automate simple tasks naturally requiring human-machine interactions, research in ASR and speech synthesis by machine has attracted a great deal of attention over the past six decades [Rabiner and Juang(2003)]. The earliest attempts to build an ASR system were made in 1950s based on acoustics phonetics. These systems relied on spectral measurements, using spectrum analysis and pattern matching to make recognition decisions on tasks such as vowel recognition [J.W.Forgie and C.D.Forgie(1959)]. Filter bank analysis was also implemented in some systems to provide spectral information. In the 1960s several basic speech recognition ideas were emerged. Zero-Crossing Analysis (ZCA) and speech segmentation were used, and dynamic time aligning and tracking ideas were proposed [Reddy(1966)]. In the 1970s, speech recognition research achieved major milestones. Isolated word recognition systems became possible using Dynamic Time Warping (DTW). Linear Predictive Coding (LPC) was extended from speech coding into speech recogni-
tion systems based on LPC spectral parameters. IBM came out with the effort of large vocabulary speech recognition system in the 70s, which turned out to be highly successful and had a great impact in speech recognition research. AT & T Bell Labs also began to make truly speaker independent speech recognition systems by studying clustering algorithms for creating speaker independent patterns [L. R. Rabiner and Wilpon(1979)]. In the 1980’s, connected word recognition system were devised based on algorithms that concatenated isolated words for recognition. Hidden Markov Models (HMM) are widely used in almost all researches after mid-1980s. In the late 1980s, Neural Networks were also introduced to problems in speech recognition as a signal classification technique.

There have been lots of popular attempts carried out towards ASR which kept the research in this area vibrant. Generally a speech recognition system tries to identify the basic unit in any language, phonemes or words which can be compiled into text [Gold and Morgan(2000)]. The potential applications of ASR include computer speech to text dictation, automatic call routing and machine language translation. ASR is a multi disciplinary area that draws theoretical knowledge from mathematics, physics and engineering. Specific topics include signal processing, information theory, random processes, machine learning or pattern recognition, psycho-acoustics and linguistics.

1.2 Motivation

To design an intelligent machine that can recognize the spoken word by different speakers in different environments and comprehend its meaning is far from
achieving the desired goal on any language. As the speech recognition technology become more and more sophisticated, its uses become more and more widespread. For decades, AT & T Bell Labs, USA has been at the fore front of speech recognition and natural language technology research. They have invested more than one million research hours over the past few decades in Speech and Language Technology research. Recently it is reported that they have developed a core technology platform, which is a cloud-based system of services that not only identifies words but interprets meaning and context to deliver accurate result. The system is built on servers that model and compare speech to recorded voices. This system needs to get improved accuracy so as to use as a speaker independent continuous speech recognition and understanding system in English.

AT & T is not alone in its quest for developing more intelligent voice-activated technologies. IBM, Microsoft and Google have each invested heavily in this area for the past few years. Microsoft has already incorporated some speech recognition technology. Current trend shows that technology will advance with more reliable speech recognition tools in near future. Under these contexts in order to incorporate speech recognition and understanding capability in different regional languages a lot of works related to the signal processing and language technology is to be carried out in each language for generating the required know hows. In this circumstance we started a study on Vowel/Consonant-Vowel (V/CV) unit classification to build a Speech Recognition and Understanding System in Malayalam language. V/CV units occur repeatedly in normal speech and recognition of these units are important for development of any speech recognition system [Chandrashekhar and Yegnanarayana(2002)]. Furthermore CV units are natural
units of speech production in the sense that, typically most syllables are of CV type [Greenberg(1999)].

The present research work is motivated by the knowledge that a little attempts were reported for the automatic speech recognition of V/CV speech unit in Indian languages like Hindi, Tamil, Bengali, Marathi etc., but very few works have been found to be reported in the literature on Malayalam V/CV speech unit recognition, which is the principal language of South Indian state Kerala [Sunilkumar(2002)], [Prajith(2008)]. So more basic research works are essential in the area of Malayalam V/CV speech unit recognition.

Current advanced speech recognition systems use traditional basic speech features such as Linear Predictive Coding Coefficients (LPCC) and Mel Frequency Cepstral Coefficients (MFCC), which are switched linear model of the human speech production mechanism [Patil and Basu(2008)], [Raja and Dandapt(2010)]. One limitation of these models is the inability to extract the non-linear, non-stationary and higher-order characteristics of the speech production process. Researchers in this area have already suggested in the literature that there is affirmation on non-linear characteristics in both voiced and unvoiced speech patterns [Teager.S.M(1989)] [Casdagli(1991)] [Banbrook and McLaughlin(1994)] [Narayanan(1999)] and [Prajith(2008)]. To capture the non-linear, non-stationary and higher order characteristics of the V/CV speech production process an improved feature extraction technique is to be introduced. To extract the non-linear information of V/CV speech unit, we use the Reconstructed State Space (RSS) based State Space Point Distribution (SSPD) parameters. Another improved fea-
ture extraction technique based on Wavelet Transform, namely Normalized Wavelet Hybrid Feature is also studied for the classification of Malayalam V/CV speech units. The peculiarity of the proposed method lies in the fact that using wavelet transform based multi resolution approach, non-stationary nature of the speech signal can be accurately considered.

1.3 Outline of the Work and Main Results

The objective of chapter 2 is to establish a necessary background for the following chapters with a review on previous works. The former part of the chapter contains a brief report on the state-of-the-art speech recognition research with a review of reported methods and main investigations towards the feature extraction using traditional feature extraction methods, multi resolution analysis and non linear dynamics. The later part gives a brief survey on research findings reported in the application of connectionist approach and statistical learning algorithms.

In the first section of chapter 3 a brief overview of phonological description of speech including articulatory and acoustic phonetics has been presented. This provides a frame work for later discussions on speech production mechanism and the corresponding speech signal, especially concentrating on Vowel/Consonant-Vowel (V/CV) sounds and their characterization. In the second section data acquisition of Malayalam V/CV sounds is presented in detail. Almost all new computers possess a built in sound card. There are three types of sound cards built onto our computers motherboard, internal and external sound cards. A typical use of this sound card is to provide the audio component for multimedia applications.
such as music compositions, editing video or audio and entertainment. These cards can be modified for speech processing applications by incorporating an anti-aliasing pre-sampling Low Pass Filter (LPF) at the input appropriately. For this purpose, an 8th order low pass butterworth filter developed is connected prior to the Analog to Digital Converter (sound card) for removing the frequency component above 4 kHz. A speech database of V/CV sounds in Malayalam is created using the data acquisition system developed, for speech modeling and classification studies described in the following chapters.

A preliminary analysis on Malayalam V/CV speech unit recognition using traditional speech feature extraction techniques, namely Linear Predictive Coding (LPC) technique and Mel Frequency Cepstral Analysis (MFCA) technique are presented in chapter 4. These techniques are widely used in speech signal processing due to its simplicity and efficiency to model the speech production system. LPCC and MFCC parameters generally referred as the characteristics of the speakers vocal tract. These parameters contain a high degree of information about the speaker’s identity in speaker independent environment. To study the Malayalam V/CV speech signal characteristics, the LPCC parameters are extracted using Linear Predictive Coding (LPC) technique. Reflection Coefficients (RC) and Log Area Ratios (LAR) with the objective to determine the set of predictor coefficients that minimize the prediction error over the short segment of speech waveform are also extracted. Levinson Durbin Recursion (LDR) based Yule-Walker equation is used for this purpose. The LDR generates these three feature coefficients. Another feature extraction technique called MFCA technique, which is based upon a switched linear model of the human speech production mechanism is used to
extract the Mel Frequency Cepstral Coefficients (MFCC). This model describes human speech production as an excitation source and a linear time invariant filter representing the vocal tract. Cepstral analysis allows the excitation source energy to be separated from the frequency response characteristics of the vocal tract.

Chapter 5 focuses on speech modeling using Wavelet Transform, a tool for Multi Resolution Analysis (MRA), which can be used to represent the speech signal in time-frequency plane. Certain ideas of wavelet theory appeared quite a long time ago. Over the last decades wavelet analysis has turned to be a standard technique in the areas of geophysics, meteorology, audio signal processing and image compression. Wavelet transform can be defined as the transformation of the signal under analysis into another representation which presents the signal in a more useful form. The main advantage of wavelet transform over Fourier Transform (FT) is the multi resolution analysis of signals with localization on both time and frequency. In the present work we utilize these characteristics of wavelet transform using two major wavelet decomposition techniques namely Classical Wavelet Decomposition (CWD) and Wavelet Packet Decomposition (WPD) which are derived from wavelet transform. Normalized Wavelet Hybrid Feature (NWHF) vector is generated by combining CWD and WPD method and then feature normalizing using z-score normalization technique.

Chapter 6 presents a relatively novel and accurate time-domain approach to model and classify Malayalam V/CV speech units. In dynamical system approach, by embedding a signal into adequately high dimensional space, a topologically equivalent state space of the system is generated. This embedding is known as
Reconstructed State Space (RSS). It is typically constructed by mapping time-lagged copies of the original signal onto axes of the new high dimensional space. The time evolution within the RSS traces out a trajectory pattern referred to as its attractor which is a representation of the dynamics of the underlying system. Since the attractor of an RSS captures all the relevant information about the underlying system, it is an efficient choice for signal analysis, processing and classifications. The RSS approach studied here has the advantage of extracting both linear and non-linear aspects of the entire system. Here we used the feature extraction technique based on the statistical models of Reconstructed State Space (RSS).

Since conventional cepstral coefficients are frame-based, a feature vector is computed and used for statistical analysis every 10 milliseconds, where as RSS have a feature vector for the entire sample. Due to the amount of data, time complexity is far greater for RSS based approach to speech recognition. For making this RSS method useful, this issue must be solved. In this thesis we propose an accurate method to extract a feature vector from RSS, called State Space Point Distribution (SSPD) parameter. A trajectory matrix of embedding dimension $d=2$ is formed using RSS of each speech sound. After that, scatter plot of the row vector of the trajectory matrix named State Space Map (SSM) is generated for each speech sound with embedding dimension $d=2$ and time delay $\tau = 1$. A feature extraction method using RSS based State Space Point Distribution (SSPD) parameters are studied from state space diagram. The SSM and the corresponding SSPD plot obtained for different speaker shows the identity of the sound so that an efficient feature vector can be formed using SSPD. On analysis it is found that RSS based SSPD models are able to differentiate each Malayalam Vowel/consonant-
Vowel (V/CV) speech unit in speaker independent environment.

Chapter 7 deals with the recognition of Malayalam CV speech unit based on the above discussed features using neural network pattern classification algorithm. Pattern classification can be defined as the categorization of the input pattern into identifiable classes via significant feature extraction of the pattern from the background of irrelevant details. It can be defined as a fields concerned with machine recognition of meaningful regularities in noisy or complex environments. A pattern can be a speech signal, a handwritten cursive word, a human face or a finger print image. The design of a pattern recognition system involves the following process such as (1) Data acquisition and preprocessing, (2) Data representation, (3) Decision making.

The problem domain decides the choice of sensors, preprocessing techniques, representation scheme and the decision making model. Generally a well defined and adequately complete pattern recognition problem will lead to a solid pattern representation and a simple decision making scheme. Learning from the set of example (training set) is an important and desired property of most pattern recognition systems. The best known approaches for pattern recognition systems are, 1. Template matching, 2. Statistical classification, 3. Syntactic or structural matching, and 4. Neural networks. In this chapter we present a study on widely used approach for pattern recognition problem using connectionist approach namely Artificial Neural Networks (ANN). Classification studies investigates the recognition capabilities of the ANN using Multi Layer Feed Forward Neural Network (MLFFNN) with Back Propagation Learning algorithm (BPLA). The present al-
The traditional speech feature extraction approach using LPC and MFCA techniques, multi resolution (time-frequency) analysis using wavelets transform based NWHF parameter and time domain approach using RSS based SSPD, are used as feature parameter for recognition studies. The recognition experiments are conducted using ANN in order to identify the credibility of these feature parameters. This chapter is organized in two sections. First section presents the recognition experiments conducted using connectionist approach ANN and the second section describes the simulation experiment conducted for the recognition of the Malayalam V.CV speech unit along with the performance comparison of proposed feature parameters.

The chapter 8 in this thesis discusses on statistical learning pattern classification approaches using Support Vector Machine (SVM) and k-Nearest Neighbor (k-NN) classifier for Malayalam V.CV speech unit recognition. In the first section of this chapter the pattern classification approach using k-NN classifier is discussed. In the second section the pattern recognition approach SVM introduced by Vapnik is discussed. SVM has the inbuilt ability to solve pattern classification problem in a manner close to the optimum for the problem of interest. Furthermore, SVM has the ability to achieve remarkable performance without prior knowledge built into the design of the system. SVM is mainly used for binary classifications. To extend this for multiclass classification a relatively new learning architecture namely Decision Directed Acyclic Graph Support Vector Machine (DDAGSVM) is studied. For N class problem, the DDAG contains, one for each pair of classes.
DDAGSVM works in a kernel induced feature space and uses two class maximal margin hyper plane at each decision node of the DDAG. The DDAGSVM is considerably faster to train and evaluate comparable to other algorithms. In this chapter an effort is made to build a recognition system using SVM based DDAGSVM technique for speech recognition. The classification experiments are carried out using LPCC (LPC, RC, LAR), MFCC, NWHF and SSPD based feature parameters as discussed previously for Malayalam V/CV speech unit classification.

The recognition accuracies obtained using previously discussed feature parameters and classifiers for Malayalam V/CV speech database in which each speech sequences are divided into 256 sample blocks and its multiples are as follows. Using k-NN algorithm, an average recognition accuracy obtained are 36.73 % for LPCC method, 42.92% for MFCC, 49.75 % for NWHF and 54.25 % for SSPD feature methods. Using ANN 45.19% for LPCC, 49.72% for MFCC, 53.65% for NWHF and 59.03% for SSPD are obtained. Finally using SVM the average recognition accuracy obtained are 51.59% for LPCC, 55.6% for MFCC, 65.32% for NWHF and 73.82% for SSPD feature methods.

Another experimental results by grouping the V/CV speech database into six different phonetic classes are as follows. The average recognition accuracies using k-NN are 51.88% for LPCC method, 55.65% for MFCC, 63.53% for NWHF and 66.73% for SSPD feature methods. Using ANN the recognition accuracy obtained are 56.32% for LPCC method, 61.21% for MFCC, 65.2% for NWHF and 74.08 for SSPD feature methods. The average recognition accuracy obtained using SVM are 59.65% for LPCC, 63.15% for MFCC, 81.08% for NWHF and
90.31% for SSPD feature methods. The recognition results using Malayalam V/CV speech database indicate that the proposed SVM based SSPD method is more efficient and can be adopted for developing an efficient speech recognition system for Malayalam language.

Finally, Chapter 9 concludes this work with directions for future research work.