Chapter 1
Introduction

1.1 Background

Speech is the predominant mode of communication in human-human interaction scenario. It is quite natural that speech would be the preferred mode for human-machine interaction as well [James Allen, 1994]. Speech based interaction is preferred over other input technologies as it is more natural and provides less effort for interaction. The speech based interaction is going to be the future interface mechanism to interact with intelligent systems to ensure the ease of use. Development of computer hardware with enormous computing power will not have parity with user interface techniques to extract the complete potential of the hardware as we depend on the conventional keyboard and mouse based interaction. The interfaces are still a bottleneck for exploiting the high computational power of the system. The speech driven interface can actually provide the best speed with that individual's capability to ensure speed of interaction with machine. Hence, many researchers are motivated to work for automatic speech recognition by machine. The studies on automatic speech recognition falls into two classes, they are isolated word recognition system and continuous speech recognition system [Mariani, J,1989]. An isolated word recognition system recognizes one word at a time. To use such a system we must pause between each word. A continuous speech recognition system recognizes speech as we normally speak it, with words flowing together in a continuous stream. Most of the systems currently available in the market are isolated word
Continuous speech recognition systems are under active development, however, they are nearing practical use. Some popular examples are Google speech recognition system, CMU’s Sphinx and IBM ViaVoice.

One of the most difficult aspects of performing automatic speech recognition is its multi-disciplinary nature, and tendency of most researchers to apply a monolithic approach to individual problems. To build an efficient speech recognition system, it requires knowledge and expertise in a wide range of disciplines. Various disciplines of speech recognition are as follows [Rabiner and Juang, 1993].

**Signal processing:** The process of extracting the relevant information from the speech signals in an efficient and robust manner. This includes preprocessing of speech signal, enhancement of signals, extracting the most robust spectral properties of the signal, which includes the time varying and frequency varying properties of the signal.

**Physics (Acoustic):** The science of understanding the relationship between the physical speech signal and physiological mechanisms – the human vocal tract mechanism that produce the speech and human speech perception mechanisms – speech hearing.

**Pattern Recognition:** Set of algorithms that is used to cluster data to create one or more prototypical patterns of data ensemble and to match (compare) a pair of patterns on the basis of feature measurement of patterns. In this, our objective is to extract a robust parameter which has most discriminatory power in between the different classes and similarity within the class.
**Communication and information theory:** The procedures for estimating parameters of statistical models; the methods for detecting the presence of a particular speech pattern, the set of modern coding and decoding algorithms (including dynamic programming approach, stack algorithms and Viterbi algorithms) used to search in large finite grid for a best path corresponding to a best recognized sequence of words.

**Linguistics:** The relationships between sounds (phonology) words in a language (syntax), meaning of spoken words (semantics) and sense derived from the meaning (pragmatics). This also addresses the issues of parsing and grammar.

**Physiology:** Understanding the higher order mechanisms within human central nervous systems that account for speech production and perception in human beings. Many modern techniques try embedding this type of knowledge within the framework of artificial neural network.

**Computer Science:** Study of efficient algorithms for implementing the software or hardware and various other practical methods for speech recognition.

**Psychology:** The science of understanding the factors that enables a technology to be used by human beings in practical task.

Measuring the performance or efficiency of Speech recognition system, depends on the conditions under which it is evaluated; under sufficiently narrow conditions almost any system can attain human-like accuracy, but it's much harder to achieve good accuracy under general conditions [Joe Tebelskis,1995].
conditions of evaluation and the efficiency of any system can vary along the following dimensions:

**Vocabulary size and confusability:** As a general rule, it is easy to discriminate among a small set of words, but error rates naturally increase as the vocabulary size grows. On the other hand, even a small vocabulary can be hard to recognize if it contains confusable words, i.e., words having almost similar pronunciation.

**Speaker dependence vs. independence:** By definition, a speaker dependent system is intended for use by a single speaker, but a speaker independent system is intended for use by any speaker. Speaker independence is difficult to achieve because a system's parameters become tuned to the speaker(s) that it was trained on, and these parameters tend to be speaker-specific. Speaker dependency is highly correlated to the type of acoustic signal features extracted from digital speech. The feature vectors like LPC (Linear Predictive Coding) and MFCC (Mel frequency Cepstral coefficients) are reported as highly speaker independent parameters to build automatic speech recognition system [Vergin.R and D. O'Shaughnessy, 1991].

**Isolated, discontinuous, or continuous speech:** Isolated speech means single words; discontinuous speech means full sentences in which words are artificially separated by silence; and continuous speech means naturally spoken sentences. Isolated and discontinuous speech recognition is relatively easy because word boundaries are detectable and the words tend to be clearly pronounced.

**Task and language constraints:** The modern speech recognition systems are built by combining the acoustic features with language models for improving the
recognition accuracy. Language models are nothing but the constraints imposed on speech recognition system to recognize an incoming acoustic signals considering the fact that we speak a valid sentence i.e., both syntactically and semantically aligned to predefined rules. The syntactic and semantic of the language may dismiss the hypothesis hence syntactically or semantically invalid sentence may not be recognized. For example when user says "I like fruits" is a valid case where as if user says "I fruits like " is not. Constraints are often represented by a grammar, which ideally filters out unreasonable sentences so that the speech recognizer evaluates only plausible sentences. The type of grammar depends on the type of task user would like to perform. The amount of grammar or syntactic variability should be defined while designing the speech recognition system [Parneet Kaur et.al., 2012].

**Read vs. spontaneous speech:** Systems can be evaluated on speech that is either read from prepared scripts, or speech that is uttered spontaneously. Spontaneous speech is vastly more difficult, because it tends to be peppered with dis-fluencies like "uh" and "um", false starts, incomplete sentences, stuttering, coughing, and laughter; and moreover, the vocabulary is essentially unlimited, so the system must be able to deal intelligently with unknown words. Detecting and flagging their presence, and adding them to the vocabulary, may require some interaction with the user [Parneet Kaur et al., 2012].

**Adverse conditions:** A system's performance can also be degraded by a range of adverse conditions. These include environmental noise, acoustical distortions (e.g., echoes, room acoustics); different microphones (e.g., close-speaking, Omni-
directional, or telephone); limited frequency bandwidth (in telephone transmission); and altered speaking manner (shouting, whining, speaking quickly, etc.) [Joe Tebelskis, 1995], [Vinay Kumar Gupta, 2011].

Automatic speech recognition is a problem of interest for researchers since 1950's. Over a period of time, different methods and techniques were experimented to build automatic speech recognition system. Initial research was focused to extract the highly discriminant feature vectors from the acoustic signals and that is used for recognizing the speech. Acoustic-Phonetic properties of various phonemes like formant frequencies, pitch, voiced-unvoiced regions, energy, nasality, frication of various phonetic units were studied and modeled for recognition. A set of parallel detectors are implemented to detect the presence of above said acoustic-phonetic properties. Segmentation and labeling process find the most stable region according to how well the features within the region matches with individual phonetic units which were already modeled and kept as reference template. This was the most sophisticated part of the First generation speech recognition system. This method requires extensive knowledge of the acoustic properties of phonetic units. The existence of phonetic units is assumed as apriori knowledge in this method where as the knowledge of acoustic properties of these phonetic units are often established in a posteriori manner. This knowledge base is incomplete, because building a complete knowledge base of the acoustic properties of all phonetic units at different context is not practically feasible. Hence building the reliable system through this method were not attracted much for researchers [Rabiner and Juang, 1993]. Later various signal processing techniques like filter bank analyzers, linear predictive coefficients,
Cepstral Coefficients and Discrete Fourier Transforms (DFT) were employed to extract the acoustic features from the speech signal recognition purpose. These features were used for creating reference pattern through training process. Pattern Classification algorithms are used to measure the similarity of an incoming pattern (test pattern) with a reference pattern to take appropriate decision for recognition [Rabiner and Juang, 1993]. The Dynamic Time Warping (DTW) algorithms and Artificial Neural Networks were popular among these classification methods used for speech recognition purpose. During 1980's Hidden Markov based Statistical models were introduced for speech recognition system which could yield effective results compared to acoustic phonetic method [Rabiner, L., 1989]. During 1990's introduction of language models provided a different direction to speech recognition research by considering the language related features, that was ignored in the earlier discussed techniques [D. Jurafsky, J.H Martin., 2002]. The inclusion of language models could provide highly reliable results in the constraint environment i.e, task / domain specific speech recognition systems in the noiseless or less noisy environment. The system performance was not adequate to deploy in public noisy places. This motivates the researchers to investigate the extraction of acoustic independent parameters from the context of speech. The speech specific lip movements and hand gestures, which are co-expressed with speech were captured and considered for improving the recognition accuracy of automatic speech recognition system in adverse environments [Stéphane Dupont and Juergen Luettin, 2000]. The inclusion of additional inputs for improving the accuracy of speech recognition system thereby improving the Human Computer Interaction (HCI) evolved a new research area.
known as Multimodal Interaction. Multimodal Interface provides an efficient way of interaction with computers employing different types of input modalities. Most of the practical implementation of Multimodal interfaces is designed; speech as major modality and other modalities are implemented as supportive modalities for correcting the ambiguities in recognized speech message [Pui-Yu Hui and Helen M., 2014]. Simple integration of various input modalities will not provide an effective interaction with machine. The logical synchronization of various input messages should be done effectively in order to understand the semantics of the communicated message. For example user can say a command "Copy this file to that folder", and also show the source and destination folders through gestures. The first pointing gesture denotes the source, the file to be copied and the second pointing gesture denotes the destination folder, where the file is to be copied. If the first and second gestures are not synchronized with the recognized speech signal, the whole scenario will get effected and it will go wrong. Human cognitive systems are capable enough to do the logical synchronization of various messages perceived from various input channels in order to derive the semantics of the message. While devising multimodal system, we should also device an algorithm to perform the logical synchronization of various messages in order to ensure the semantic of the message communicated. In this present work we report the multimodal based techniques for improving the speech recognition accuracy, thereby improving the human-computer interaction techniques. The Multimodal based interaction should be seen objectively, i.e., the purpose of the communication need to be met, no matter from which individual recognizer the
message is recognized. More specifically different input modality recognizers will work together to meet the purpose of communication.

1.2 Motivation

Providing human like capabilities to computer systems; which can think, work and act according to context is being a passion and challenging area for researchers who work in the area of artificial intelligence and robotics. The first and foremost capability we all would like to see in any artificial intelligence based system would be speech based interaction. Speech based interaction is something unique to human being, which make human differ from other living beings. Hence building Automatic Speech Recognition (ASR) system is being a hot topic for decades. Years of effort from the research community could provide an efficient and commercial system for practical use in a constraint environment. Quite a good number of commercial and open source systems are available in English and other European languages like German, French, Italian, etc. Dragon Naturally speaking, Google Speech Engine, CMU’s Sphinx, IBM Viavoice, are some of the well known solutions [WR-Wikipedia, Speech 2015]. All these systems uses the acoustic signal features along with language specific knowledge for recognizing the speech. Hence the speech recognition system available in one language cannot be used for recognizing other language unless the language specific knowledge coupled with acoustic signals are given to the system while training and acoustic model building phase [D. Jurafsky, J.H Martin,2002]. Most of the studies in the area of automatic speech recognition system reported in the literature are built for English language [Anusuya and Katti,2009]. Among Asian countries, popular
systems and studies are reported for Chinese, Japanese and Korean languages [Pratik K.K, et al., 2014]. Specific to Indian context there are solution from Nuance Corporations, i.e Dragon Naturally speaking , which is English based ASR for Indian Speakers (Indian accent English). Even if practical solutions are not available, lot of studies are reported in the area of speech recognition system for Indian languages like Hindi [Kumar M, et al., 2004], Tamil, Marathi, Bengali, Malayalam etc. [Sunil kumar., 2002] [Prajith, P.,2008], [Cini Kurian,2014], [Jinal H.T and Dipti B.S, 2015]. In Malayalam Language in the area of continuous and isolated word speech recognition system, not much study is reported.

The extraction of acoustic features from the speech signals are one of the prominent tasks in building ASR systems for isolated word recognition and continuous speech recognition. Researchers have experimented with various types of acoustic features extracted from speech. Most of the studies reported in the literature use Mel frequency Cepstral Coefficients and it's derivatives for building ASR systems [Rabiner and Juang, 1993], [Cong Thanh, et al., 2010], [D. Jurafsky, J.H Martin, 2002]. Non-linear based feature extraction techniques were also studied and used for building automatic speech recognition system. In this method the speech data is considered as non-linear time series data and numerous techniques were introduced for nonlinear based time series analysis. The work done by Prajith P, reports a promising results in the area of usage of nonlinear based features for vowel recognition. In this work he used the conventional MFCC based features joined with time domain based Phase Space Point Distribution (PSPD) features for recognizing the vowels sounds of Malayalam Speech. In his studies he reports 5 to 7 % increase in recognition accuracy against
mere MFCC features [Prajith P, 2008]. This work is further extended by Thasleema T.M and built a consonant recognition system for Malayalam speech using PSPD features [Thasleema T.M, 2012]. No studies are reported in literatures about the usage of hybrid feature vector (Frequency domain feature joined with time domain feature) for developing isolated word recognition or continuous speech recognition system. The present study reports the experimental results on use of PSPD based feature vector with MFCC features for isolated word recognition and continuous speech recognition in Malayalam Language.

Providing error tolerant ASR solution in the acoustic adverse (noisy) environment is a challenging problem for researchers even now. Extraction of non-acoustic features from the context of speech production for improving the accuracy of automatic speech recognition has recently gained wide popularity among researchers. Lip reading and hand and body gesture features are extracted from the context of speech and it is used as an augmenting feature for improving the accuracy of speech recognition. Over a period of time, researchers also experimented with various other input techniques like online handwriting, pen gesture recognition and also inputs from other sensors, like accelerometer, haptic sensors, gyroscope etc. coupled with conventional speech recognizers to improve the man-machine interaction capabilities. The method of using multiple input output modalities for interacting with computer system are known as multimodal interaction. The present work is motivated by the knowledge that, little attempt were reported for building a multimodal based speech recognition system for Indian Languages more specifically for Malayalam Language.
1.3 The Outline of Thesis

The present study, emphasize on the enhancement of isolated word recognition and continuous speech recognition system for Malayalam language, through multimodal techniques. The proposed algorithms and system components are discussed in detail and is validated with experimental results. This thesis also discusses the future directions of multimodal research in order to build efficient human - machine interaction techniques. The overall organization of thesis is detailed as follows.

The chapter 2 discusses the review of previous work in the area of automatic speech recognition system. The Chapter starts with the history of Automatic speech recognition system, there after it reviews the work in the area of acoustic feature extraction. Studies on frequency domain and time domain based features are reviewed in this section. The next section reviews the use of language models for building the continuous speech recognition system. The work reported in literatures in the area of use of visual features for improving the accuracy of speech recognition is reviewed. This section mainly reviews the work in the area of lip feature extraction and hand gesture extraction which are co-expressed with speech and used for improving the accuracy of speech recognition system. The state of art of multimodal based techniques for improving the accuracy of automatic speech recognition and human machine interaction are reviewed in the last section.

The chapter three has two main sections. The first section discusses the characteristics of speech signal (both acoustic and language characteristics) and
the second part describes the architecture of modern speech recognition system. The classifications of phonemes based on different properties specific to Malayalam language are discussed in this section. The architecture of modern speech recognition framework is discussed in section three. Various integral components of modern speech recognition system are covered in this part.

The chapter four deals with acoustic parameters and language models used for building the continuous speech recognition system for Malayalam language. The method of extraction of the time domain based non-linear Phase Space Point Distribution (PSPD) features from the speech signal are discussed in this section. The acoustic model using PSPD feature vector combined with the Mel frequency Cepstral Coefficients (MFCC) are used as acoustic features and this is used for building the acoustic models are discussed in this chapter. Language models which are used for speech recognition system is also discussed in detail. The procedure for N-gram language model building from a given corpora is described with examples. This chapter also explains how continuous speech recognition system works with acoustic models and language models in order to recognize a given input speech signals. The last section of chapter four discusses the experimental setup and simulation results. The performance of system with MFCC based features and Joint features (MFCC+PSPD) are evaluated experimentally with and without language models under different signal to noise ratio. The chapter concludes with a discussion on the importance conducting further studies for developing an automatic speech recognition system in the noisy environment.
Chapter five proposes a method for building of automatic speech recognition system for noisy environment (adverse environment for acoustic signals) using non-acoustic parameters. This chapter starts with various studies reported in the literature about the use of gestures, i.e., visual features for improving the accuracy of speech recognition system. This chapter proposes the algorithms for building isolated word recognition and continuous speech recognition system using the non-acoustic feature vectors, i.e. visual features which are co-expressed with speech.

Chapter six and seven describes the gesture recognition system, which are used as a support system for improving the accuracy of automatic speech recognition. The gestures which are co-expressed with speech are extracted from the context of the speech and this information is used for improving the recognition accuracy in the adverse environment for acoustic signals. Both static and dynamic gestures are studied as part of this thesis. Methods for recognizing static gestures are discussed in chapter six and the procedures for recognizing dynamic gestures are described in chapter seven. Both the studies are experimentally validated and results are tabulated in the respective chapters.

Chapter eight reports the simulation experiment and results of an isolated word recognition system employing audio visual parameters as proposed in chapter five. Both static and dynamic gestures are extracted from the context of speech and used as augmenting features for improving the accuracy of isolated word recognition system. The experimental results are tabulated under different signal to noise ratios for speech signals with static and dynamic gestures
separately. This chapter concludes with a justification for building the continuous speech recognition system using visual features along with language models.

Chapter nine describes the paradigm shift in the area of speech recognition and human computer interaction through Multimodal Interaction techniques. Theory and best practices for building the multimodal system are detailed in this chapter. The Architecture of a proposed multimodal system is discussed in here. The multimodal message generation and Multimodal message understanding are also discussed in the subsequent sections. In this study in addition to speech, both hand gestures and pen-based gestures are implemented as an augmenting modality for improving the interaction. The algorithms proposed as for building a multimodal system for Malayalam language are validated through experiments. Chapter ten concludes the work with the importance of conducting further studies in the area of multimodal system to improve the human computer interaction methods.