



An SVM Based Speaker Independent Isolated Malayalam Word Recognition System

Thushara P V¹, Gopakumar C²

¹M.Tech Scholar, Department of Electronics, College of Engineering, Chengannur
Chengannur, Alappuzha Dist., Kerala-689121, INDIA

²Associate Professor, Department of Electronics, College of Engineering, Karunagapally
Karunagapally, Kollam Dist., Kerala-690523, INDIA

Abstract: As technology advances man-machine interaction is becoming an unavoidable activity. So an effective method of communication with machines enhances the quality of life. If it is able to operate a system by simply commanding, then it will be a great blessing to the users. Speech is the most effective mode of communication used by humans. So by introducing voice user interfaces the interaction with the machines can be made more user friendly. This paper implements a speaker independent speech recognition system for limited vocabulary Malayalam Words in Raspberry Pi. Mel Frequency Cepstral Coefficients (MFCC) are the features for classification and this paper proposes Radial Basis Function (RBF) kernel in Support Vector Machine (SVM) classifier gives better accuracy in speech recognition than linear kernel. An overall accuracy of 91.8% is obtained with this work.

Keywords: Speech Recognition; Feature Extraction; MFCC; Support Vector Machine; RBF kernel; Linear kernel; Raspberry Pi

I. Introduction

Automatic speech recognition can be defined as a technology which enables a system to recognize the input speech signals and interpret the meaning, after which the system should be able to generate some control signals [1]. In this paper, the speech recognition system recognizes Malayalam words with limited vocabulary. There are so many challenges in speech processing like acoustic and phonetic variability. Isolated speech input is used since continuous speech is difficult to process. In continuous speech, it is difficult to find the start and end points of words. Within all these considerations, system recognizes words with better accuracy and greater speed. Malayalam is recently being considered among one of the classic languages in India. Speech recognition system for Malayalam language helps people who are not conversant with English and unaware of using computer. Malayalam is one among the 22 languages spoken in India with classical status [2]. Malayalam belongs to the Dravidian family of languages and most of the Malayalam speakers live in the Kerala, one of the southern states of India. There are 37 consonants and 16 vowels in the language. For the proper functioning of the system, there should be distinct pauses between the words i.e. isolated words. Due to the memory constraints in the handheld device, the vocabulary supported by the system is limited i.e. it is a limited vocabulary speaker independent isolated word recognition system.

II. Literature Survey

Nowadays, innovation in scientific research is focused much more on the interactions between humans and technology and automatic speech recognition is a driving force in this process. Speech recognition technology is changing the way information is accessed, tasks are accomplished and business is done. There are two related speech tasks: speech understanding and speech recognition. Speech understanding is getting the meaning of an utterance such that one can respond properly whether or not one has correctly recognized all of the words. Speech recognition is simply transcribing the speech without necessarily knowing the meaning of the utterance. The two can be combined, but the task described here is purely recognition. Automatic speech recognition (ASR) is the ability of a machine to convert the words that is spoken in to the microphone to recognized words.

A. Speech Production and Perception

Five different elements are associated with speech production and perception. They are speech formulation, human vocal mechanism, acoustic air, perception of the ear, speech comprehension etc.

B. Types of ASR Systems

ASR can be classified in several ways: speaker dependent or independent, discrete or continuous, and small or large vocabulary.

1. Speaker Independent (SI) systems or Speaker Dependent (SD) system: System can recognize a variety of speakers, without any training. Such systems limit the number of words in a vocabulary. But speaker dependent system can only recognize the speech of users it is trained to understand [3].
2. Discrete ASR or Continuous ASR: Discrete ASR recognizes isolated utterances. Here the user must speak unnaturally, leaving distinct pauses between each word. In Continuous ASR, the user can speak naturally, with normal conversational pauses, but it is more difficult for the system to detect the word boundaries.
3. Small vocabulary or large vocabulary system: In small vocabulary ASR, all the words in the vocabulary are trained at least once, whereas large vocabulary systems recognize sounds rather than whole words and are able of recognizing words that have never been in the training set.

C. Factors Affecting Speech Recognition Performance

Different speech recognition systems have different parameters and design methodology according to the application. But in most cases some factors are similar. For instance, Vocabulary size and variability of factors [3]. These factors have significant roles for the accuracy of the system. According to how much vocabulary can be recognized, speech recognition can be divided into three different scales vocabulary speech recognition: Small vocabulary speech recognition, Medium vocabulary speech recognition, large vocabulary speech recognition. Small-scale can identify less than 100 vocabulary while medium-scale can identify more than 100 vocabulary and large-scale can identify more than 1000 vocabulary. Other factors that affect performance of speech recognition system, are variability in speakers, environments, transmission channels and microphones. The variability in speakers involves gender, speed of speaking, regional changes in language. Variability in environment means whether it is noisy or clean. The bandwidth of transmission channel also plays an important role in determining the accuracy of the system as number of samples transmitted changes when bandwidth changes. According to the type of microphone and distance of microphone from mouth, reliability of system may changes. And finally the performance of the system will be in the hands of experts behind the work.

III. Methodology

A speech recognition system is basically a pattern recognition system dedicated to recognize the words spoken into microphone. Proposed automatic speech recognition system (ASR) starts with a feature extraction module where, the input speech waveform is processed to extract the required acoustic feature vectors that are used to characterize the spectral properties of the time varying speech signal. Feature extraction is processing the input speech to extract compact and efficient set of parameters that uniquely represents the speech input. The second stage is the classifier. This stage evaluates the similarities between the input feature vector sequence and trained database to determine which words were most likely spoken. By feature extraction, we calculate Mel Frequency Cepstral Coefficients (MFCCs) [4]. We use MFCCs as the feature for classification because it allows better processing of data, high accuracy for clean speech and it approximates the human auditory system's response. The classifier used for training and testing is Support Vector Machine (SVM). The block diagram of the proposed speech recognition system is shown in figure 1:

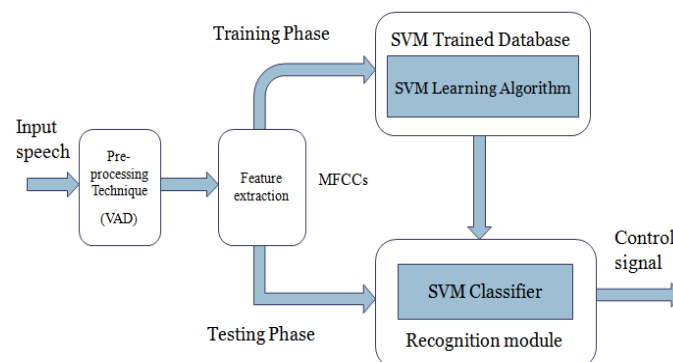


Figure 1. Block Diagram of Proposed System

A. Voice Activity Detection

Voice Activity Detection is a preprocessing technique which detects the start and end point of a word. In the proposed method we use three different features per frame. The first feature is the widely used short-term

energy (E) [5]. Energy is the most common feature for speech/silence detection. Another feature selected for separating speech and silence is zero crossing rate. It is observed that silence parts have high zero crossing rate whereas in speech, zero crossing is low. Besides these two features, it was observed that the most dominant frequency component of the speech frame spectrum can be very useful in discriminating between speech and silence frames.

B. Mel Frequency Cepstral Coefficients (MFCCs)

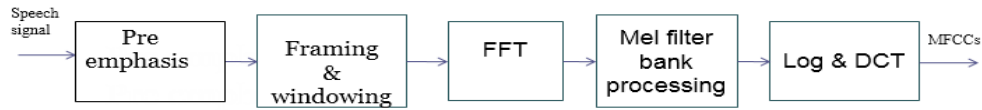


Figure 2. Calculating MFCCs

C. Support Vector Machine (SVM) Classifier

Pattern recognition algorithm used for the proposed system is Support Vector Machine [6]. MFCCs of each of the words is given to the classifier stage for training using SVM learning algorithm. Here 100 samples of each word from different speakers is used to train the classifier and testing is conducted with another different speakers in order to make the system speaker independent. Here we optimize the parameters of SVM by making a comparison between the kernel functions. In this work, we use simple linear kernel and radial basis function kernel for comparison.

IV. Results and Analysis

A. Database Used

For conducting this experiment we choose six Malayalam words. The samples stored in the database are recorded by using a high quality studio-recording microphone at a sampling rate of 16 KHz. Malayalam numerals from one to six is chosen to create the database. Twelve speakers are selected to record the words. Each speaker utters six words with thirty samples each. We have used six male speakers and six female speakers for creating the database. Thus the database consists of a total of 2160 utterances of the spoken words. Speech database is shown in Table 1.

Table 1. Speech Database

NO	WORDS
1	ഒന്ന് (onnu)
2	രണ്ട് (randu)
3	മൂന്ന് (moonnu)
4	നാല് (naalu)
5	അഞ്ച് (anj)
6	ആറു (aaru)

B. Voice Activity Detection

This important front end processing detects the start and end point of words. The result obtained by voice activity detection of word 'onnu' is plotted in Figure 3.

C. Simulation Parameters of Feature Extraction Block

- Sampling frequency = 16000Hz
- Frame duration = 10ms
- Frame overlapping = 5ms
- Number of DFT points = 256
- Number of Mel filters = 24
- Number of MFCC coefficients = 19

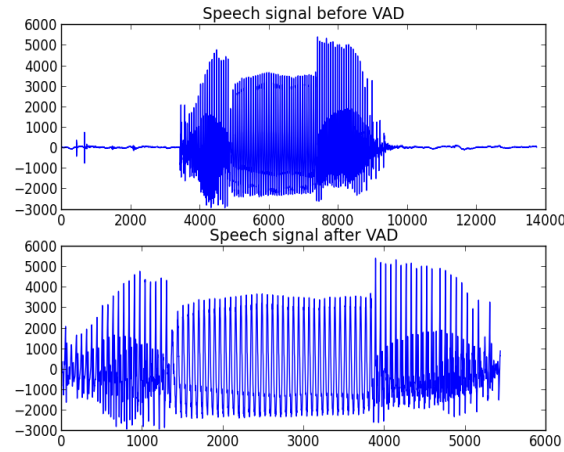


Figure 3. Voice Activity Detection of word 'ONNU'

D. Recognition Results

There are 100 input samples for training and testing is conducted with samples of different speakers for each word. Training is performed with maximum iterations of 100. In SVM, it is possible to use RBF kernel and linear kernel. In this paper, a comparison of accuracy in test results is made for both the cases when RBF and Linear kernel are used. The accuracy obtained in both the experiments are shown in the table II:

Table 2. Accuracy Test Results

Words	Accuracy (%)	
	Linear kernel	RBF kernel
onnu	75	91.6
randu	91.6	91.6
moonnu	84	100
naalu	50	100
anj	68	84
aaru	84	84
Average	75.5	91.8

E. Comparison of Results

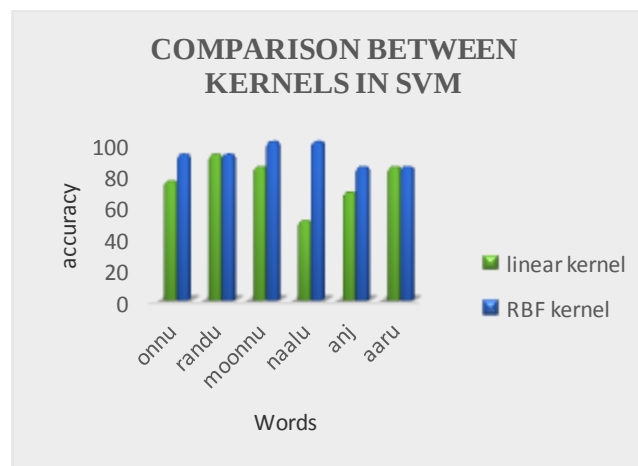


Figure 4. Comparison of accuracies between linear and RBF kernel

The experimental results prove that training with RBF kernel gives better accuracy in recognition than with linear kernel. System trained with linear kernel has got an average accuracy of 75.5% whereas for RBF kernel it is 91.8%.

