MALAYALAM WORD IDENTIFICATION FOR SPEECH RECOGNITION SYSTEM

Maya Moneykumar[1] , Elizabeth Sherly[2] and Win Sam Varghese[3]

[1]*Indian Institute of Information Technology and Management (IIITM-K), Kerala*
[2]*Indian Institute of Information Technology and Management (IIITM-K), Kerala*
[3]*Indian Institute of Information Technology and Management (IIITM-K), Kerala*
E-mail: maya.mphilcs2@iiitmk.ac.in[1] , sherly@iiitmk.ac.in[2], sam.varghese@iiitmk.ac.in[3]

Abstract: Automatic Speech Recognition (ASR) Systems have long been a goal of artificial intelligence researchers. The lack of state-of-the art ASR System has been a major hindrance due to its complexity in reproducing in Computer. Hidden Markov Models (HMMs) are used heavily in most current speech recognition systems for both phoneme and syllable based approach. In this paper, we also propose to use HMM model, but based on energy measure and Mel Frequency Cepstral Coefficient (MFCC) to determine the syllable based segmentation and features of power spectrum of speech signal. The system was trained with utterances of 3 male and 2 female speakers and the database included 40 utterances. The training and testing was done with bi-syllable words and the implementation of the system has been done using Hidden Markov Model Toolkit (HTK).

INTRODUCTION

Speech is the primary means of communication between people, and is the most natural way to exchange information. Speech can provide a convenient interface to control devices. The main goal of speech recognition is to get efficient ways for humans to communicate with computers. Some of the speech recognition applications require speaker dependent/independent isolated word recognition. Automatic Speech Recognition (ASR) is a technology that enables a device to recognize and understand spoken words, by digitization of the signals and matching its pattern against stored patterns.

ASR is a branch of Artificial Intelligence (AI) and is related with number of fields of knowledge such as acoustics, linguistics and pattern recognition. The aim of ASR research is to allow a computer to recognize speech in real-time as human understands. The words that are spoken by any person independent of vocabulary size, noise, speaker characteristics and channel conditions is a major challenge in ASR. The accuracy rate of speech recognition depends on the feature extraction techniques as well as the training methods used. Most current speech recognition systems use hidden Markov models (HMMs) to deal with the temporal variability of speech. Although ASR technology is not yet at the point where machines understand different speech, it is used in a number of applications and services [3,8,16].

The earliest attempts to design ASR by machine were made in 1950s which developed an isolated digit recognition system. At present , a lot of advancements has happened in the field of speech recognition. The windows operating system now provides a well versed speech recognition system. New technology mobile phones are also incorporated with speech recognition applications. The main focus of research groups were in building ASR systems for English. Syllable based speech identification was attempted in English Language also, which has proven that using syllable level information for word identification can very much reduce the error rate. ASR systems developed for Indian languages also have shown better results in Digit Identification and isolated word identification. [7]

Isolated word identification systems for languages like Hindi, Punjabi and Tamil using HTK have shown satisfactory results. A notable work in Tamil - A syllable based isolated word recognizer for Tamil handling OOV (Out Of Vocabulary) words", uses a sub word based continuous speech recognizer for word identification. Noteworthy works happened in Malayalam also, among which the significant one was from CDAC, Trivandrum where they developed a speech recognition system for visually impaired people. MFCC method was used as front-

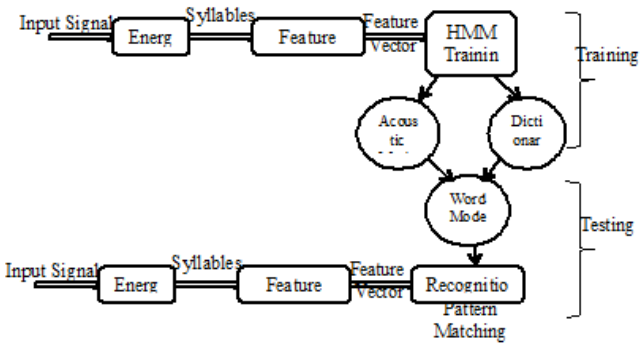
end to extract acoustic features from the input signal and a hybrid model integrating rule based and statistical method was used to handle pronunciation variations in the dictionary. Performance analysis on Tamil and Telugu and English data has shown a satisfactory result for syllable based word identification technique, which in fact will be suitable for Malayalam also.

Among all the methodologies used for speech recognition problem, Hidden Markov Models are best statistical models that are most accurate in modeling the speech parameters. Many researchers proposed different methodologies for HMM based speech recognition system where most of them had used different feature extraction methods to get the speech features like fundamental frequency, energy coefficients etc. Many of the studies show that the use of MFCC as feature to speech recognition yield good result. Other feature extraction methods like LPC are good for speech coding, but not recognition.

Methodology & Implementation

The word identification system designed for Malayalam language uses a syllable based segmentation approach. Instead of training the system with independent words, we use syllables and phoneme combinations for training and identification. This eliminates the problem of maintaining large set of training data, as different words common syllable as well as phone segments. New words can be added to the vocabulary without building new models for the existing syllables and phonemes corresponding to the word.

Fig. 1 Architecture of a Speech Recognition System



Preprocessing

While building an ASR system, the preprocessing stage consists of recording of speech utterances, segmentation of speech signals into syllables, preparation of a pronunciation dictionary and lot more. As the first step 40 different utterances of bi-syllable word in Malayalam were recorded using the tool, Praat. The training and testing of the system is done based on syllables and hence the utterances were segmented into syllable units. For the training purpose, syllables which were segmented manually were used and the same done automatically were used for testing. Manual segmentation was done using Praat, by observing the change in Formant frequency. Automatic segmentation was done based on the energy measure. The speech signal was divided into overlapping frames for which the energy measure was calculated.

Fig2. Speech signal plotted for word / ൽ൦൦

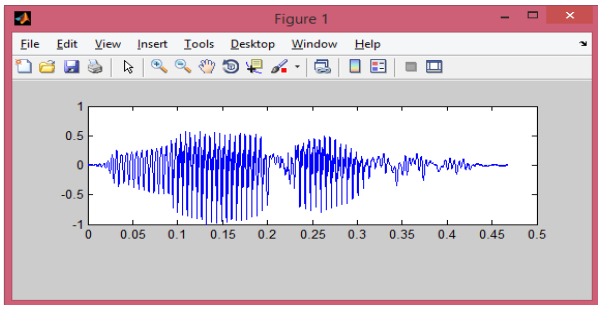


Fig3: Energy plotting for / ൽ൦൦

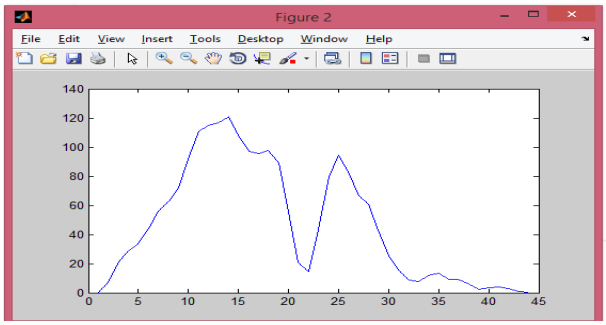
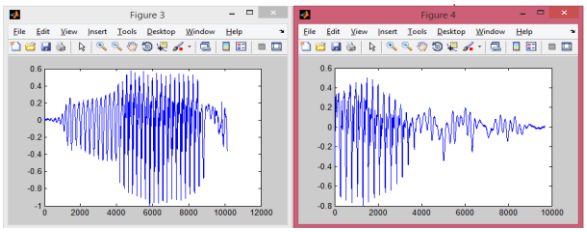
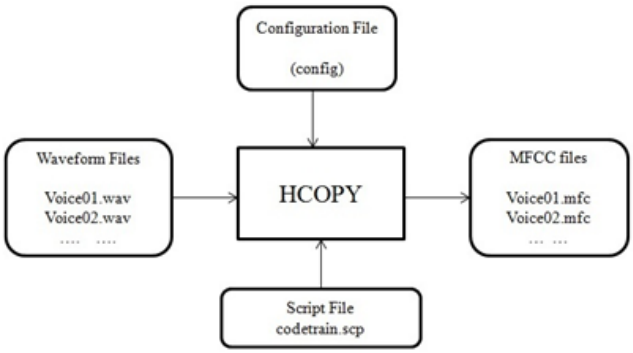


Fig: 4 syllable1- / ൽ syllable2- / ൦൦



Dictionary preparation has an equally important role in speech recognition where a dictionary is built with a sorted list of required wordsTo train a set of HMMs , every file of training data must have an associated syllable level transcription. The transcriptions are provided in the form of Master Label File. [15]. The final stage of data preparation in preprocessing is to parameterize the raw speech waveforms into sequence of feature vectors. For each signal frame, 39 MFCC coefficients are extracted.

Fig 5. Generating mfc files [15]



Training the Acoustic Models

The first step in HMM initialization and training is to choose a priori topology for each HMM.

- Number of states
- Form of the observation functions(associated with each state)
- Disposition of transition between states.

In HTK, a HMM is described in a text description file. Here, we have a total of 5 states in HMM including two non-emitting states, 1 and 5. If the dictionary contains multiple pronunciation of the same word, realigning the training data is done. The recognizer considers all pronunciations for each word and outputs the pronunciation that best matches the acoustic data.

Recognizing Test Data and Evaluation of the Result

Like the training corpus preparation the testing signals were also converted into series of feature vectors (.mfcc). The test data's MFCC coefficients and the final iteration HMM model of the train data set were given as parameters to the HTK tool HVite that performs recognition. For performance analysis, HTK provides a tool called HResults. It computes the recognition statistics such as percentage of correctly recognized words. In the recognition statistics, the first line gives the sentence-level accuracy based on the total number of transcriptions generated by the recognizer which are identical to the reference transcriptions. The second line contains the numbers concerning the word accuracy of the transcriptions generated by the recognizer [15].

$$\text{Correctness \%} = \frac{N-D-S}{N} \times 100$$

$$\text{Accuracy \%} = \frac{N-D-S-I}{N} \times 100$$

$$\text{Word Error Rate (WER)} = \frac{S+D+I}{N} \times 100$$

where N = Total number of utterances, D = Number of deletion error, I = Number of insertion error, S = Number of substitution error and H = Number of utterances recognized.

Results And Discussions

The system was trained with utterances of 3 male and 2 female speakers. The database included 40 utterances. The word utterances were manually segmented into syllables and saved as .wav files. These wave files were given as training data to the system. When an input utterance was given for testing, automatic segmentation of the speech signal was done and the parameters extracted. The pattern matching happened between the trained syllables and the syllabified segments in the test data. By giving multiple pronunciations to a single word, the dictionary was made stronger which improved the accuracy. The test data were categorized into 2 groups A, B consisting of 25 words each. During the testing phase, the system showed an accuracy of 76% and 80% respectively for the sets A and B. The results are good for a single speaker. The accuracy rate can be improved to some extent, if number of utterances for training is increased.

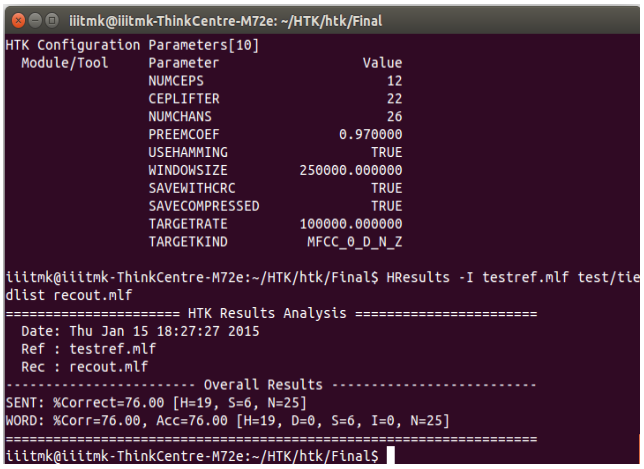


Fig 8: Test result with dataset A

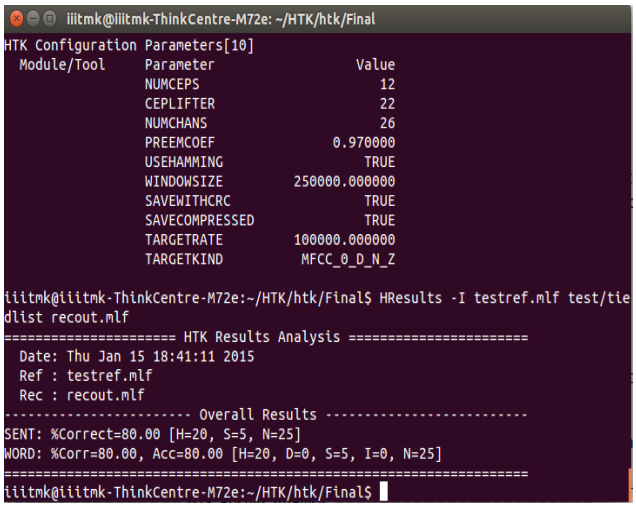


Fig 9: Test result with dataset B

While syllable based word identification is giving a reasonable accuracy, on the other hand when the system was trained using phoneme level classification of a word, the accuracy got reduced to around 40%. Phone level identification has proved successful in the case of languages like English whereas, in Malayalam, since the phone boundaries cannot be classified accurately, identification rate is less.

Expr	Train Set	Test Set	N	D	I	S	H	Accuracy
Exp1	A,B	A,B	50	0	0	23	27	54%
Exp2	A,B	A	25	0	0	6	19	76%
Exp3	A,B	B	25	0	0	5	20	80%

When the system was trained with the entire data set and then tested, in the initial case only 23 utterances were identified correctly. After making changes in the pronunciation dictionary and then training the system, the test results were improved. The syllabification of words also plays an important role in the word recognition accuracy rate. Syllables contain a vowel nucleus and initial and final consonants. The main problem faced during syllabification was that there is no general rule as to where syllable boundaries should be located. This syllabification problem has affected the accuracy rate to a great extent. Accuracy in boundary identification of syllables may improve the efficiency.

It can be concluded from the above results that conducting a number of experiments and changing the word representation in the pronunciation dictionary, apparent changes will happen in the accuracy rate. The utterances for training were recorded in a noisy environment which has affected the performance. A properly pronounced and recorded data (recorded in a noiseless environment) will improve the result. The database should contain maximum number of utterances with which the system should get trained in order to build a much stronger ASR system.

Conclusion and Future Scope

A syllable based word identification system for Malayalam using HTK on the Linux platform was performed using MFCC and HMM. The training was conducted for 40 vocabularies of bi-syllable words. The system was trained to identify the utterances within the training vocabulary, with a moderate accuracy. As a next step, the system can be trained to handle out of vocabulary (OOV) words as well. The work can be extended to handle n-syllable words, with a large vocabulary, that would serve as the first step for speech-to text system for continuous speech recognition.

REFERENCE

[1] Irfan Ahmed , Nasir Ahmad , Hazrat Ali and Gulzar Ahmad “The Development of Isolated Words Pashto Automatic Speech Recognition System” Proceedings of the

18th International Conference on Automation & Computing, Loughborough University, Leicestershire, UK, IEEE, 8 September 2012

[2] Bassam A. Q. Al-Qatab and Raja N. Ainon "Arabic Speech Recognition Using Hidden Markov Model Toolkit(HTK)", IEEE,2010

[3] Cini Kurian and Kannan Balakrishnan "Continuous Speech Recognition System for Malayalam Language" International Journal of Computing and Business Research (IJCBR) ,Volume 3 ,Issue 1, January 2012

[4] A Akila and E. Chandra"Isolated Tamil Word Speech Recognition System Using HTK" International Journal of Computer Science Research and Application, Vol. 03, Issue. 02,pp. 30-38, 2013

[5] A Akila and E. Chandra "Performance enhancement of syllable based Tamil speech recognition system using time normalization and rate of speech" CSIT , 2(2):77–84, June 2014

[6] Gajanan Pandurang Khetri, Satish L. Padme, Dinesh Chnadra Jain,Dr. H. S. Fadewar, Dr. B. R. Sontakke and Dr. Vrushsen P. Pawar " Automatic Speech Recognition for Marathi Isolated Words" International Journal of Application or Innovation in Engineering & Management (IJAIEM) , Volume 1, Issue 3, November 2012

[7] Sukhminder Singh Grewal and Dinesh Kumar "Isolated Word Recognition System for English Language", International Journal of Information Technology and Knowledge Management, vol 2,No. 2,pp.447-450, July-December 2010

[8] Smrithy K Mukundan "Shreshta Bhasha Malayalam Speech Recognition using HTK ", International Journal of Advanced Computing and Communication Systems (IJACCS) , vol.1 Issue 1,March 2014

[9] G Lakshmi Sarada, A Lakshmi ,Hema A.Murthy and T Nagarajan "Automatic transcription of continuous speech into syllable like units for Indian languages" Sadhana vol.34,Part 2,pp 221-233, April 2009

[10] Santosh K. Gaikwad, Bharti W. Gawali and Pravin Yannawar "A Review on Speech Recognition Technique", International Journal of Computer Applications, vol.10,No.3, November 2010.

[11] Manish P. Kesarkar "Feature Extraction for speech recognition",M.Tech Credit Seminar Report, Electronics Systems Group, EE Dept,IIT Bombay, November 2003

[12] Xuedong Huang and Li Deng "An overview of Modern Speech Recognition" , Microsoft Corporation

[13] D.S Shete, Prof.S.B Patil and Prof.S.B Patil "Zero crossing rate and Energy of the speech signal of Devanagari Script" IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 4, Issue 1, Ver. I (Jan. 2014), PP 01-05

[14] Rupali S Chavan And Dr. Ganesh S Sable "An Implementation Of Text Dependent Speaker Independent Isolated Word Speech Recognition Using HMM" , International Journal Of Engineering Sciences & Research Technology, September 2013