Isolated Speaker Independent Kannada ASR System Using HTK

Shashidhara Nimbargi
PG Scholar: Dept. of CS & E.
S.J.C Institute of Technology
Chickballapur - 562 101, India

Dr.S.N. Chandrashekara
Professor and Head: Dept. of CS & E.
S.J.C Institute of Technology
Chickballapur - 562 101, India

Abstract: Kannada language is one of the oldest languages of the world and has been estimated that it is as old as 2500 years. This is the official language of Karnataka state in India, and there is a sizable Kannada speaking population not only in neighboring states of India, but also in many countries. As per the recent estimate, there are more than around 40 million people who speak this language. With this background of the language, there is a necessity of further research in Kannada speech recognition systems, which will facilitate the development of applications for different commercial as well as social purposes. In this paper, suggested is a method for Kannada speech Recognition system using Hidden Markov model (HMM) and Mel Frequency Cepstral Coefficients (MFCC). This paper briefly talks about the concepts of Speech recognition systems in general, HMM, Hidden Markov Model Toolkit (HTK), Speech data collection, and finally it closes with the outcomes of the work and conclusion.

Keywords: Speech Recognition System, Mel frequency cepstral coefficients (MFCC), Hidden Markov Model (HMM), HTK Tool kit.

1. INTRODUCTION

Speech is one of the fundamental way of communication mode of the human and, man-machine interaction is almost become basic necessity in this era. Developing any kind of speech recognition system which can convert speakers spoken words in to a particular form which a computer can understand and do further processing based on the application requirements can be a very useful tool.

India is a developing nation in the field of Information Systems and not all the population is familiar with English language, especially in rural or under developed regions. Another reason for this is that India has diverse cultural background with number of local or native languages. Kannada is one of such native language, with sizable people using it for daily communication

Fundamentally there are two stages in any speech recognition systems [1] namely, training stage and testing stage. Developing Speech recognition systems particularly for native languages is a complex task because of the reasons such as quasi-stationery characteristic of the speech signal, variation in speech signal based on speaker's age, gender and other such parameters.

The originality of this paper comes from the fact that system is developed for recognizing combination of Kannada words and digits. Even though considerable work is done in many Indian languages [1], [4], [7], not much work is done in Kannada Language.

Additionally whatever work is done for Kannada language was either done only for Kannada digits recognition or speaker dependent implementations. The method proposed here is a speaker independent recognition system. Based on above facts, Kannada speech recognition system calls for further research.

Research in Automatic Speech Recognition (ASR) systems started as early as early 1940's, but the commercial availability of the ASR systems only started in 1980's. Focus of research work in speech recognition systems shifted to statistical modeling framework from pattern recognition systems in the 1980's.

Methods of speech recognition have been developed using HMM based isolated speech recognition for Kannada digits [2] and was shown that the MFCC features combined with first order and second order derivatives given good results. But there is a need for system to be trained with larger data base for improving recognition accuracy. Limitation of this work was that all samples were obtained from a single speaker.

Studies done on speech recognition algorithms [3] shown that MFCC is the most common and relatively faster method of computation and HMM is widely used pattern matching algorithm. Mel-Frequency Cepstral Coefficients is about representation of the short-term power spectrum of a sound, which is based on the linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency.

Earlier studies were done with the aim of choosing the better algorithm based on comparison of various algorithms. According to review of different algorithms for feature extraction techniques, it is found that MFCC has better success rate for speech recognition than Linear Predictive Coding (LPC) and Linear Predictive cepstral coefficients (LPCC) techniques.

Current systems for Kannada Speech Recognition systems still lack the accuracy and robustness. There is still lot of scope for research in Speaker Independent Kannada ASR systems. There is still lack of commercially feasible Kannada ASR solutions which can be used for business as well as for educational purposes.

2. SPEECH REOGNITION

Over a past several decades an enormous experiments and research is conducted in speech recognition technology. Figure 1 shows the basic ASR system using HTK [22], where input to the system is speech file and the output will be recognized text. Input will be a speech file say in wave format which is then preprocessed and feature extraction is done before giving it to recognition module. Speech corpus collection is a very important step in any speech recognition system design. This is because the recognition accuracy of the training module largely depends on the size of the sample data collection and also the various parameters like speaker's age, gender, comfort etc.

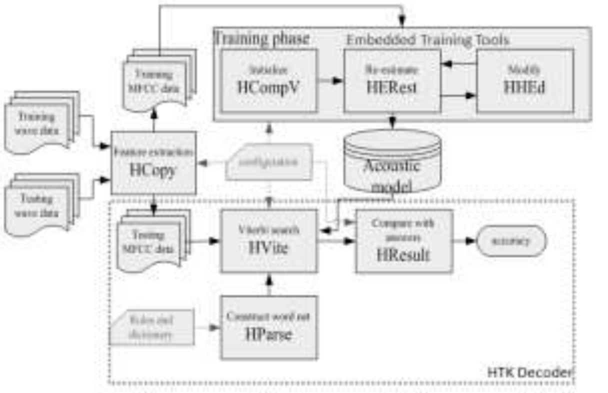


Figure 1: Building ASR using HTK

Speech segmentation, speech enhancement, feature extraction and acoustic pattern classification are major building blocks to develop an ASR engine [5].

2.1 Speech Segmentation

The main pre-processing task in ASR system design and implementation is the endpoint detection. Normally, the input audio speech to the ASR will be a continuous speech signal, wherein there is no boundary information. With such an input, the preprocessing involves, chopping of long periods of speech in to the shorter ones and also elimination of non-speech information. Non-Speech information is nothing but the silent regions of the input signal. The audio input is labeled with the beginning point and end point of the speech. These points are further used to compute the feature vectors. This stage of detection of speech start and end points is normally called as speech boundary detection. This is very important for reduced word error rate and efficient computation of the ASR, which improves the performance recognition of the ASR System [5].

Lot of research work is done in the endpoint detection area. Multiple researchers referred start and endpoint speech detection methods and speech segmentation task by using Short Time Energy (STE) [17], Frame based Teager's Energy, Zero Crossing Rate (ZCR) [18, 19].

2.2 Speech Enhancement

Speech Enhancement which is also a noise cancellation is the preprocessing module for developing a robust ASR system. The main functionality of the module is the suppressing of the noise signals and thus improving the speech quality in real time environmental situations. Enormous amount of work is carried out in the field of noise suppression and speech enhancement. This field is still a research problem, as 100% noise cancellation is still not possible.

2.3 Acoustic Feature Extraction

The feature extraction which is nothing but the parametric representation of a speech signal is an utmost significant task to get better recognition accuracy. Once the start and end points of a speech signal are detected, the next step is the extraction of feature vectors. Enormous amount of research was done in the feature extraction techniques. The acoustic features extraction methods like Mel-Frequency Cepstral Coefficients [20] or Linear Predictive Coefficients [21] are well defined preprocessing techniques in speech recognition.

2.4 Acoustic Pattern Classification

The extracted acoustic feature vectors are used to train a classifier to identify the words spoken by the subject. Hidden Markov Models, Artificial Neural Network (ANN), Dynamic Time Warping (DTW), and Gaussian Mixture Models (GMM) are widely used techniques for speech recognition.

Hidden Markov Model is one of the popular statistical based tools used for speech recognition research [3]. HMM provides simple and yet effective model for modeling time-varying speech signal. In HMM the system under modeling is considered to be a Markov process with hidden states. HMM has been widely used in speech recognition systems.
Hidden Markov model Toolkit (HTK) [6] is a portable toolkit for building and manipulating hidden Markov models. HTK is more suitable for research in speech recognition systems. HTK provides facilities for speech analysis, speech synthesis, Character Recognition, DNA Sequencing and training. HTK consists of set of library modules and tools available in C source form. These tools provide sophisticated facilities for Speech analysis, HMM Training, Testing and Result Analysis.

The software supports to build complex HMM's systems using continuous density mixture Gaussians and discrete distributions. HTK Software architecture is optimised for speech recognition, is very flexible and complete. In addition it has very good documentation and has support for various tools such as data preparation tools, training tools, recognition tools and analysis tools.

In HTK to preprocess Wave files we can use HCopy command. We can setup the configuration of HTK modules using configuration files in which we can mention the parameters like Source or target format, source or target Rate, window size etc.

Speech recognition system can be categorized as training phase or offline processing stage and testing or online processing phase. Figure 2 shows a block diagram of Training Phase. When training a language model, there is a need to include sentence start and end labeling to differentiate between different words.

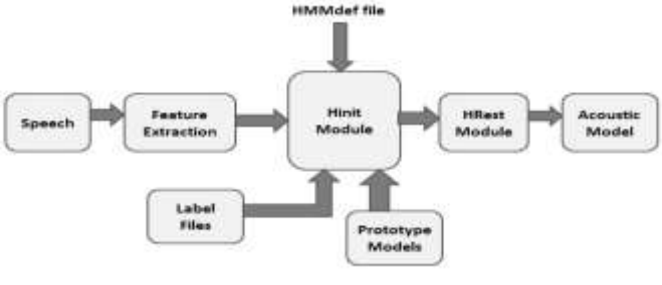


Figure 2: Training phase

Feature extraction is about the technique of preserving the necessary information from the speech signal while removing unwanted information. The process of feature extraction is normally lossy transformation meaning, provided the feature vectors, it is not possible to perfectly reconstruct the original signal. MFCC is one of the prominent feature extraction technique used in speech recognition systems [3]. Human perception of the frequency information does not follow linear scale. In MFCC Mel scale is used which is actually a linear frequency distribution below 1000Hz and logarithmic frequency distribution above 1000Hz [4]. As shown in figure 2, HInit module will have the prototype models as input, which will have the information about number of states, mean value etc.

A typical prototype model file is shown in figure 3. HInit module will iteratively computes an initial set of parameter value of whole word models and HRest is used to further re-estimate the HMM parameters initially computed using set of observation sequences. HVite function is a general-purpose Viterbi word recognizer, which will map a speech file against a network of HMMs and output a transcription for each. HParse is a program that generates word level lattice files from a text file syntax description containing a set of rewrite rules.

Testing phase System will monitor the test inputs and it can detect voice activity by removing noise and silence. Again feature extraction has to be done for the testing speech samples and then it needs to be matched against the already created models to recognize a word.

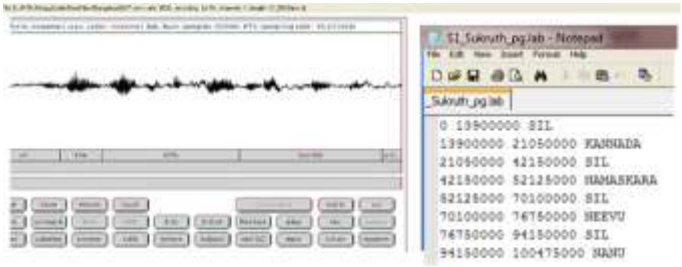
```
~o <VECSIZE> 39 <MFCC>
~h "proto"
<BeginHMM>
<NumStates> 5
<State> 2
<Mean> 39
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 …
<Variance> 39
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 …
<State> 3
<Mean> 13
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 …
<Variance> 13
…
<TransP> 5
0.0 1.0 0.0 0.0 0.0
0.0 0.5 0.5 0.0 0.0
0.0 0.0 0.5 0.5 0.0
0.0 0.0 0.0 0.5 0.5
0.0 0.0 0.0 0.0 0.0
<EndHMM>
```

Figure 3: Prototype Model file

3. KANNADA ASR

The voice samples for training phase used was 31 utterances from 35 for different speakers to check speaker independent performance. Out of 35 training samples 19 were the male speech samples and 16 were samples from female persons.
All the samples are collected in .wav format. Sample from each user was labeled to distinguish between different utterances as shown in figure 4

For testing phase separate utterances of 31 words from 6 different speakers were used. These 6 speakers' data collected was not used in training phase and it was used only for testing purposes. The samples were collected in an environment with controlled noise condition.



```
0 13900000 SIL
13900000 21050000 KANNADA
21050000 42150000 SIL
42150000 52125000 NAMASKARA
52125000 70100000 SIL
70100000 76750000 HEEVU
76750000 94150000 SIL
94150000 100475000 NANU
```

Figure 4: Master Label file

When the test samples were tested against the trained model, accuracy obtained was 100 percent for 17 out of 31 utterances. For the remaining 14 utterances accuracy obtained was up to 83 percent. The results are summarized in Table 1.

Table 1: Results of Kannada ASR

Sl. No	Kannada Word	Accuracy	Sl. No	Kannada Word	Accuracy
1	ಅಜ್ಜಿ	100	17	ಹಬ್ಬ	100
2	ಅರಮನೆ	100	18	ಹತ್ತು	100
3	ಐದು	100	19	ಹಿಂದೆ	83
4	ಬಲಗಡೆ	83	20	ಹೊರಗೆ	83
5	ಬೆಳಗ್ಗೆ	83	21	ಹೃದಯ	83
6	ಬೆಂಗಳೂರು	100	22	ಮೂರು	83
7	ಚೆನ್ನಾಗಿದೆ	83	23	ಮುಂದೆ	83
8	ದಕ್ಷಿಣ	100	24	ನಾಳೆ	100
9	ದೊಡ್ಡದು	100	25	ಊಟ	100
10	ಈದಿನ	83	26	ಪೂರ್ವ	83
11	ಏಳು	100	27	ಸಹಾಯ	100
12	ಎರಡು	83	28	ತರಕಾರಿ	100
13	ಎತ್ತರ	83	29	ಉಪ್ಪು	83
14	ಗಣರಾಜ್ಯ	100	30	ಒಂಬತ್ತು	100
15	ಗಿಡ್ಡ	100	31	ಎಡಗಡೆ	83
16	ಹಾಲು	100			

4. CONCLUSION AND FUTURE WORK

In this study, we have implemented Kannada Speech recognition system using HTK for Kannada words. HTK was found to be very simple and effective tool for research. System was able to identify the speech from different speakers. Accuracy was between 83 to 100 percent for the trained words. The system needs to be trained with larger training samples to increase accuracy further. Also the accuracy of the system for larger databases needs to be researched.

5. REFERENCES

[1] Dalmiya C.P, Dr. Dharun V.S, Rajesh K.P, "An Efficient Method for Tamil Speech Recognition using MFCC and DTW for Mobile Applications", ICT 2013 - Proceedings of 2013 IEEE Conference on Information and Communication Technologies

[2] Muralikrishna H, Ananthakrishna T, Dr. Kumara Sharma, "HMM based Isolated Kannada Digit Recognition System using MFCC, International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2013

[3] Rashmi C R, "Review of Algorithms and Applications in Speech recognition System", 2014 - International Journal of Computer Science and Information Technologies, Vol. 5(4), 5258-5262.

[4] Vibha Tiwari, 2010, MFCC and its applications in speaker recognition, International Journal on Emerging Technologies 1(1): 19-22(2010).

[5] Amaresh P Kandagal, Dr.V. Udayashankara, "A Automatic Bimodal Audiovisual Speech Recognition: A Review, IEEE Conference on Contemporary Computing and Informatics (IC3I), Page(s):940-945, Nov-2014

[6] HTK Site http://htk.eng.cam.ac.uk/

[7] Ankit Kumar, Mohit Dua, Tripti Choudhary, Continuos Hindi Speech Recognition Using Gaussian Mixture HMM, 2014 IEEE Students' Conference on Electrical, Electronics and Computer Science

[8] Loh Mun Yee, Abdul Manan Ahmad,Comparitive study of Speaker Recognition Methods:DTW, GMM and SVM, Faculty of Computer Science & Information system, University of Malaysia

[9] Altangerel Ayush, Bayanduuren Damdinsuren, A design and implementation of HMM based Mongolian Speech Recognition System, School oflnformation and Communication Technology of MUST

[10] Renjith S, Aju Joseph, Anish Babu K.K., Isolated Digit Recognition for Malayalam- An Application Perspective, International Conference on Control Communication and Computing (ICCC), Page(s):190-193, Dec-2013

[11] Chao Wang, Ruifei Zhu, Hongguang Jia, Qun Wei, Huhai Jiang, Tianyi Zhang and Linyao Yu, Design of Speech Recognition System, Third International Conference on Information Science and Technology, Page(s):1042-1044, March 23-25, 2013;

[12] Cini Kurian, A Survey on Speech Recogntion in Indian Languages, International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014, ,6169-6175

[13] Dr E.Chandra, K.Manikandan, M.S.Kalaivani, "A Study on Speaker Recognition System andPattern classification Techniques, International Journal Of Innovative Research In Electrical, Electronics, Instrumentation and Control Engineering", Vol. 2, Issue 2, February 2014

[14] Hemakumar G, Punitha P, Speech Recognition Technology: A Survey on Indian Languages, International Journal of Information Science and Intelligent System, Vol. 2, No.4, 2013

[15] AN.Sigappi, S.Palanivel, Spoken Word Recognition Strategy for Tamil Language, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012

[16] Bassam A. Q. Al-Qatab , Raja N.Ainon,"Arabic Speech Recognition using Hidden Markov Model Toolkit (HTK) ", IEEE, IT Sim, Vol2, Page(s):557-562,June,2010

[17] J. G. Wilpon, L. R. Rabiner, and T. Martin, "An Imporved Word-Detection Algorithm For Telephone-Quality Speech Incorporating Both Syntactic And Semantic Constraints," Bell Labs. Tech. J.,Vol. 63, No.3, pp.479-498.,1984

[18] AN.Sigappi, S.Palanivel, "An Algorithm For Determining The Endpoints of isolated utterances," Bell Syst. Tech. J., Vol. 53, No.2,pp.297-315,1975

[19] R.Tucker,"Voice Activity Detection Using A Periodicity Measure",Proc.Inst. Electr. Eng. I, Vol. 139, No. 4, pp. 377–380, 1992

[20] Seman, N., Bakar, Z.A., Bakar, N.A., "An Evaluation Of Endpoint Detection Measures For Malay Speech Recognition Of An Isolated Words" Information Technology (ITSim), International Symposium, Vol.3, pp.1628-1635, 15-17 June 2010.

[21] Ting HuaNong, Yunus., J., Salleh, S.H.S., "Classification Of MalayMeasures For Malay Speech Sounds Based On Place Of Articulation And Voicing Neural Networks," Electrical and Electronic Technology, International Conference,TENCON.Proceedings of IEEE, vol.1, pp.170-173, 2001.

[22] Steve Young.,et al, The HTK Book (for HTK Version 3.4), December 2006