RESEARCH ARTICLE OPEN ACCESS

Automatic Speech Recognition for Ahirani Language Using Hidden Markov Model Toolkit (HTK)

Ajay S. Patil
School of Computer Sciences,
North Maharashtra University
Jalgaon (MS) – India

ABSTRACT

This paper describes the implementation of HMM (Hidden Markov Model) based speaker independent isolated word speech for Ahirani which is a commonly spoken language in the Khandesh region of Maharashtra State of India. The system is developed using Hidden Markov Model ToolKit (HTK). The system is trained on 20 Ahirani words by collecting data from 10 speakers and is tested using data collected from another 10 speakers in room environment. This paper details the experiment by discussing the implementation using the HTK Toolkit. The experimental results show that the performance of the system is 94% and is speaker independent.

Keywords:- Ahirani, Automatic Speech Recognition (ASR), HMM, HTK, Isolated Word ASR, Mel Frequency Cepstral Coefficient (MFCC), Speaker Independent.

I. INTRODUCTION

Speech Recognition is a technology that allows a computer to identify the words that a person speaks into a microphone or telephone. Speech recognition can be defined as the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words [1][2]. Automatic speech recognition (ASR) is one of the fastest growing areas of engineering and technology. Automatic speech recognition systems are developed for English and other major languages spoken in developed countries. Automatic speech recognition systems are under development for Indian languages such as Hindi, Tamil, Telugu, Bengali, Assamese and Marathi. Spoken languages like Ahirani are not explored till now. This work is an attempt to initiate the work on designing and developing a speech recognition system for Ahirani. It is one of the most common language spoken in Khandesh. Khandesh region mainly constitutes Dhulia, Jalgaon and Nandurbar districts. Automatic speech recognition systems have been implemented using various toolkits and software. Most commonly used amongst them are the Hidden Markov Model ToolKit, Sphinx Toolkit, ISIP Production System, Julius Open-Source Large Vocabulary CSR Engine, HMM Toolbox for Matlab etc. Among all these tools the HTK toolkit is the most popularly used tool to design ASR systems. Since it is used in building and manipulating hidden Markov Models it has applications in other research areas as well. HTK is well documented and provides guided tutorials for its use. The ASR for Ahirani discussed in this paper is based on the work onMuhirwe Jackson [3] who designed automatic digit speech recognizer for Kinyarwanda language and Nicolas Moreau [4] who has illustrated design of a basic Yes/No (English) speech recognition system using HTK.

II. RELATED WORK

In recent years may researchers have used HTK to design automatic speech recognition systems using HTK toolkit. Kumar and Aggarwal (2011) built a speech recognition system for Hindi using HTK to recognize the isolated words using acoustic word model. The system is trained for 30 Hindi words collected from eight speakers. Overall accuracy of their system is 94.63%[5]. Dua et al. implemented an isolated word Automatic Speech Recognition system (ASR) for Punjabi using HTK. The system is trained for 115 Punjabi words collected from eight speakers and is tested using samples from six speakers. The overall system performance is 95.63% and 94.08%[6]. Saini et al. (2013) also built an ASR for Hindi using HTK that recognizes isolated words and the system is trained for 113 Hindi words collected from nine speakers. The systems overall accuracy is

96.61% [7]. Agrawal and Dave (2008) implemented a speech recognition system for Hindi. They used a dataset of 100 words and used different windowing functions; they obtained accuracy between 55%-76% for various windowing techniques [8]. Gupta R. (2006) designed a speech recognition system for Hindi Digits [9]. Gawali (2010) et al. developed isolated word recognition system using MFCC and DTW features for Marathi [10]. Work has not been reported for Ahirani language so far. This has been the principle motivation behind undertaking this work. The ASR system designed for Ahirani discussed in this paper, apart from Muhirwe Jackson [3] and Moreau [4], also makes use of concepts studied from the related work stated in this section.

III. HIDDEN MARKOV MODEL TOOLKIT (HTK)

Hidden Markov Model Toolkit i.e., HTK is a portable toolkit developed by the Cambridge University Engineering Department (CUED) freely accessible for download after registration at the URL http://htk.eng.cam.ac.uk. It consists of several library module and C program code and with good documentation (HTK Book[11]). Precompiled binary versions are also available for download (for Unix/Linux and Windows operating systems). Its current 3.4.1 release is stable and has been used by researchers worldwide. Apart from speech recognition it has been applied to character recognition, speech synthesis, DNA sequencing etc. The toolkit provides tools for data preparation, training, testing and analysis (table 1).

Table 1: HTK Tools

Task	**Tools available in HTK**
Data Preparation	HSLab, HCopy, HList, HQuant, HLed
Training	HCompV, HInit, HRest, HERest, HSmooth, HHed, HEAdapt
Testing	HVite, HBuild, HParse, HDMan
Analysis	HResults

IV. ISOLATED AHIRANI WORDS RECOGNITION SYSTEM

As mentioned earlier the design of isolated Ahirani word recognition system is based on Jackson (2005)[3] and Moreau (2002) [4]. As no previous speech corpus is available for Ahirani language, it was necessary to design a speech corpus. It is important that the speech database should be concrete, diverse and should sufficiently represent the language under study. The text chosen to develop speech database must be grammatically correct so that it can be used to record speech samples from various speakers. Training and testing a speech recognition system needs a collection of utterances of identified words. It was experienced that recording several hundred words with multiple utterances from each speaker and labeling it using HSLab took considerable amount of time. So based on statistics of training and testing data in previous works the training data was restricted to 1000 voice samples for 20 Ahirani words. A list of 20 Ahirani words (table 2) spoken in day to life has been selected for this experiment.

Table 2. Twenty Ahirani Words for the Speech Corpus

कारे	वख्खर	लगीन	ननिंद	रांधनी	उखल्डा	न्हयारी	तुन्हं	चुल्हा	चलींतर	खुडा
कोन्ही	फपुटा	जेठीनी	उब्या	रुम्हनं	जपीजाय	घट्या	बैतन	दुल्ड्ली	कंडोलीन	उलतनी

The subsections to follow discuss creation of the speech corpus, acoustical analysis, training the HMMs and creating task definitions and finally testing the performance of the developed system against test data.

4.1 Creation of Speech Corpus

Speech corpus data for training and testing purpose is collected from native speakers of Ahirani considering four speakers each (2 males and 2 females) from the five villages Akulkheda, Ghumawal, Janve, Ghodgaon and Gidhade from Jalgaon and Dhule districts. Voice samples from ten speakers (5 males and 5 females) are used to train the system whereas voice samples from the another ten (5 males and 5 females) were used for testing purpose. For the training corpus each speaker was asked to utter each of the twenty Ahirani words (table 2) five times. The total number of recorded signals in the training corpus is 1000 utterances. In case of the testing data the words were uttered only once by the other set of 10 speakers. The detail of training and testing data is given in table 3. The total number of recorded signal in the testing data set is 200 utterances. The speech signals are recorded using HSLab and Sennheiser PC 350 special edition microphone in room environment at a sampling rate of 16000 Hz. Since, Sennheiser

PC 350 is a head set microphone, the distance between the mouth of the speaker and microphone is nearly similar for all speakers. Signal (speech) files are stored in HTK specific (.sig) format. For each recorded word the process of labeling the signal is manually carried out. There are three successive regions in each of the words recorded viz., start silence (labeled sil), the recorded word (e.g., tiphan) and end silence (again labeled as sil). HTK stores the labeled speech information in a label file with extension .lab. This label file contains start and end sample time for each label for all recorded words.

Table 3. Details of Speech Corpus

Description	Training	Testing
No. of words	20	20
Male speakers	5	5
Female speakers	5	5
No. of utterances	5	1
Recorded words	1000	200

4.2 Acoustical Analysis

For training and recognition, HTK tool requires the processing of the raw signal files created using HSLab. It provides the HCopy tool to convert the original signal to a series of acoustical signals. Although HCopy supports several acoustical analysis coefficients, Mel-scale frequency cepstral coefficient (MFCC) were selected as it takes into consideration the human perception sensitivity with respect to frequencies. The parameters necessary for acoustic analysis such as format of input speech files, features to be extracted, window size, window function, number of cepstral coefficients, pre-emphasis coefficients, number of filter bank channels and length of cepstral filtering is provided to the HCopy in a configuration analysis.conf. The values to these parameters used for this experiment are given in table 4.

Table 4. HCopy Configuration Parameters

Parameter	Value	Description
SOURCEFORMAT	HTK	The format of input speech signal
TARGETKIND	MFCC_0_D_A	12 MFCC coefficients (c1,c2,..,c12), null coefficient (c0, total energy in frame), delta and acceleration coefficients (first & second order derivatives of c0,c1,…,c12)
WINDOWSIZE	250000.0	25000 µs/25ms size of window frame
TARGETRATE	100000.0	10000 µs/10ms
NUMCEPS	12	Number of the MFCC coefficients (c1, c2,…,c12)
USEHAMMING	T	Hamming function is used
PREEMCOEF	0.97	Pre-emphasis coefficient
NUMCHANS	26	Number of filterbank channels
CEPLIFTER	22	Cepstral liftering length

HCopy is also given the list of speech files to process along with the name and location of target file names. HCopy segments each signal file into successive frames of 25 ms, overlapping each other. These segments are then multiplied by the Hamming windowing function. Then from each frame 39 coefficients are extracted. Each acoustical observation (target .mfcc coefficient file) consists of sequence of vectors (containing the following 39 values) stored in the following format.

Fig. 1. Parameter Vector layout in HTK Format File (MFCC_0_D_A)

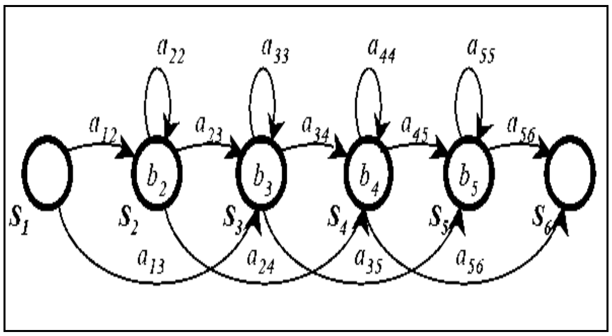
<-------12 MFCC------>				null	<-----------13 Delta----------->					<-----13 Acceleration------->				
C_1	C_2	…	C_N	E	dC_1	dC_2	…	dC_N	dE	DC_1	DC_2	…	DC_N	DE
1	2	…	12	13	1	2	…	12	13	1	2	…	12	13

C_i: Basic coefficients **E**: Log Energy **dC_i, dE**: Delta coefficients **DC_i, DE**: Acceleration coefficients

4.3 HMM Training and Task Definition

Since this work deals with isolated single word speech recognition, we have modeled 21 acoustical events (20 Ahirani words + sil) with Hidden Markov Model. In other words we have designed 21 Hidden Markov Models one for each event. The basic topology (Fig. 2) as for HMM machines as suggested by Nicolaus (2002) [4] is used for all 21 events modeled. Hence, each of the model (an HMM machine) to be built has six states out of which four (S_2 and S_5) are active and two (S_1 and S_6) are non-emitting.

Fig. 2. Basic topology (all HMMs)



The prototype of each of the 21 HMMs are stored in separate files called as HMM description files. The HInit tool is used to initialize each HMM by using time alignment of training data using the Viterbi algorithm. HInit creates the initialized version of each HMM based on the prototype (created earlier one for each HMM), coefficient file (.mfcc) and label files (.lab). The observation function b_i is represented by diagonal matrices with Gaussian distribution. 39 values each for mean matrix (all 0's) and variance matrix (all 1's) is given as initial data. The initial transition probability a_{ij} is as given below.

Fig. 3. Initial Transition Matrix a_{ij}

$$\begin{bmatrix} 0.0 & 0.5 & 0.5 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.4 & 0.3 & 0.3 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.4 & 0.3 & 0.3 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.4 & 0.3 & 0.3 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.5 & 0.5 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix}$$

Each of these initialized HMM was then subjected to HRest, which was repeatedly executed until convergence. The HRest tool estimates the optimal values for the designed Hidden Markov Models. The task grammar (gram.txt), task dictionary (dict.txt) were created as mentioned in the HTK documentation. The task grammar is compiled with HParse tool to obtain the task network. The developed task grammar, task dictionary and 21 HMMs together makes the ASR system for Ahirani and can be used for testing.

4.4 Performance Testing

That test data to be recognized is subjected to acoustic analysis (.mfcc). The HVite tool is used to match the test data with all the 21 HMMs. The result obtained are as given under.

```
====================== HTK Results Analysis =======================
Date: Mon Apr 14 17:35:11 2014
Ref : .\vcrms_project\test\ref.mlf
Rec : .\vcrms_project\test\rec.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=94.00 [H=188, S=12, N=200]
WORD: %Corr=94.00, Acc=94.00 [H=188, D=0, S=12, I=0, N=200]
===================================================================
```

Fig. 3 HTK Result Analysis for Test Data

Here H is the number of correct labels, S is number of substitutions, N is the total number of words given for testing, D is number of deletions, and I is number of insertions. It is observed from the results generated by the toolkit that the sentence recognition (SENT) rate is 94.00 %. Out of N=200 sentences provided as input to HTK, H=188 is number of test data correctly recognized whereas S=12 is number of substitution errors. The statistics given on the second line (WORD) not relevant for this taks and it is meaningful only with more sophisticated types of recognition systems like connected words recognition tasks etc. The system was trained and tested with two different sets of speakers. These resulting recognition rate of 94% indicates that the developed system is speaker independent.

V. FUTURE WORK AND CONCLUSIONS

The systemhas given encouraging results for selected twenty Ahirani words. Since the system is in place, the work can be extended to several hundred speech samples collected from several individuals belonging to different age groups. In future work can be extended for continuous speech recognition of Ahirani. An agriculture services related interactive voice recognition system can also be developed for Ahirani speaking farmers.

ACKNOWLEDGEMENT

This research work is supported by grants received under Vice Chancellor Research Motivation Scheme (VCRMS) of North Maharashtra University, Jalgaon (MS), India.

REFERENCES

[1] Zue, V., Cole, R., Ward, W. (1996). Speech Recognition Survey of the State of the Art in Human Language Technology. Kauii, Hawaii, USA

[2] Mengjie, Z., (2001) Overview of speech recognition and related machine learning techniques, Technical report. retrieved April 14, 2014 from http://www.mcs.vuw.ac.nz/comp/Publications/archive/CS-TR-01/CS-TR-01-15.pdf

[3] Muhirwe Jackson, Automatic Speech Recognition: Human Computer Interface for Kinyarwanda Language, Thesis, Makerere University, Kampala, Uganda, August 2005 retrieved April 14, 2014 from http://www.fon.hum.uva.nl/david/ba_shs/kinyarwanda_rcognizer_final_report_1.pdf

[4] Nicolas Moreau, HTK (v.3.1): Basic Tutorial, retrieved April 14, 2014 from http://www.labunix.uqam.ca /~boukadoum_m/DIC9315/Notes/Markov/HTK_basic_tutorial.pdf (2002).

[5] Kuldeep Kumar, R. K. Aggarwal, Hindi Speech Recognition System using HTK, International Journal of Computing and Business Research, Volume 2 Issue 2 May 2011

[6] Mohit Dua, R.K.Aggarwal, Virender Kadyan, Shelza Dua, Punjabi Automatic Speech Recognition Using HTK, I International Journal of Computer Science Issues, Vol. 9, Issue 4, No 1, July 2012 p. 359-364

[7] Preeti Saini, Parneet Kaur, Mohit Dua, Hindi Automatic Speech Recognition Using HTK, International Journal of Engineering Trends and Technology (IJETT) – Volume 4 Issue 6, June 2013, pp. 2223-2229

[8] Aggarwal R.K., Dave Mayank, Implementing a Speech Recognition System Interface for Indian Languages, in proceedings of the IJCNLP-08 workshop on NLP for less priviledged Languages, pp. 105-112, Hyderabad, India, January 2008.

[9] Gupta, R, and Sivakumar G. “Speech Recognition for Hindi Language”, 2006, M.Tech Project, IIT, Bombay,

[10] Gawali, Bharti W., Gaikwad, S., Yannawar, P., Mehrotra Suresh C., Marathi Isolated Word Recognition System using MFCC and DTW Features (2010), Int. Conf. on Advances in Computer Science 2010, DOI: 02.ACS.2010. 01.73

[11] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev and P. Woodland, The HTK Book (for HTK Version 3.4), Cambridge University, Cambridge, England, 2006