International Journal of Engineering Trends and Technology (IJETT) - Volume4 Issue6- June 2013

Hindi Automatic Speech Recognition Using HTK

Preeti Saini[1], Parneet Kaur[2], Mohit Dua[3]

[1] *Preeti Saini CSE, ACE Kurukshetra University, India*

[2] *Parneet Kaur CSE, ACE Kurukshetra University, India*

[3] *Mohit Dua, CSE, NIT, Kurukshetra University, India*

Abstract: Automated Speech Recognition (ASR) is the ability of a machine or program to recognize the voice commands or take dictation which involves the ability to match a voice pattern against a provided or acquired vocabulary. At present, mainly Hidden Markov Model (HMMs) based speech recognizers are used. This paper aims to build a speech recognition system for Hindi language. Hidden Markov Model Toolkit (HTK) is used to develop the system. It recognizes the isolated words using acoustic word model. The system is trained for 113 Hindi words. Training data has been collected from nine speakers. The experimental results show that the overall accuracy of the presented system with 10 states in HMM topology is 96.61 and 95.49%.

Keywords: HMM; HTK; Mel Frequency Cepstral Coefficient (MFCC); Automatic Speech Recognition (ASR); Hindi; Isolated word ASR.

I. INTRODUCTION

Many times key board acts as a barrier between computer and the user. This is true especially for rural areas. This work is an attempt towards reducing the gap between the computer and the people of rural India, by allowing them to use Hindi language, the most common language being used by the people in rural areas. Speech recognition will, indeed, play a very significant role in promoting the technology in the rural areas. Speech is a useful and effective communication medium with machines, especially in the environment where keyboard input is awkward or impossible. The Speech Processing technology (Automatic Speech Recognition and Speech Synthesis) has made great progress for European languages. In India, almost three-fourth of the population lives in rural areas and most of the population is unfamiliar with computers and English. It would be a great boon for Indian society if communication with machines, mainly with computers, in native languages can be made possible. It will enable people to interact with computers in their own language and without the use of keyboard. Speech interfacing involves two distinct areas, speech synthesis and automatic speech recognition (ASR). Speech synthesis is the process of converting the text input into the corresponding speech output, i.e., it acts as a text to speech converter. Conversely, speech recognition is the way of converting the spoken sounds into the text similar to the information being conveyed by these sounds. Among these two tasks, speech recognition is more difficult but it has variety of applications such as interactive voice response system, applications for physically challenged persons and others (Aggarwal and Dave, 2011). There are many public domain software tools available for the research work in the field of speech recognition such as Sphinx from Carnegie Mellon University (SPHINX, 2011), hidden Markov model toolkit (HTK, 2011) and large vocabulary continuous speech recognition (LVCSR) engine Julius from Japan (Julius, 2011). This paper aims to develop and implement speech recognition system for Hindi language using the HTK open source toolkit.

1.1 Motivation

At present, due to its versatile applications, speech recognition is the most promising field of research. Our daily life activities, like mobile applications, weather forecasting, agriculture, healthcare etc. involves speech recognition. Communicating vocally to get information regarding weather, agriculture etc. on internet or on mobile is much easier than communicating via keyboard or mouse. Many international organizations like Microsoft, SAPI and Dragon-Naturally-Speech as well as research groups are working on this field especially for European languages. However some works for south Asian languages including Hindi have also been done (Pruthi et al., 2000; Gupta, 2006; Rao et al., 2007; Deivapalan and Murthy, 2008; Elshafei et al., 2008; Syama, 2008; Al-Qatab et al., 2010) but no one provides efficient solution for Hindi language. The lack of effective Hindi speech recognition system and its local relevance has motivated the authors to develop such small size vocabulary system [1].

II. Related works

This section presents some of the reported works available in the literature that are similar to the presented work.

Among others, some of the works providing ASR system for South-Asian languages are (Al-Qatab and Ainon, 2010; Gupta, 2006; Pruthi et al., 2000). Pruthi et al. (2000) have developed a speaker-dependent, real-time, isolated word recogniser for Hindi. Linear predictive cepstral coefficients (LPCCs) were used for feature extraction and recognition was carried out using discrete HMM. System was designed for two male speakers. The recognition vocabulary consists of Hindi digits. An isolated word speech recognition system for Hindi language has been designed by Gupta (2006). System uses continuous density hidden Markov model (CDHMM) and consists of word-based acoustic units. Again the system vocabulary contains Hindi digits. Recogniser gives good results when tested for sounds used for training the model. For other sounds too, the results are satisfactory. The work in Al-Qatab and Ainon (2010) discusses the development and implementation of an Arabic speech system using HTK. System can recognise both continuous speech as well as isolated words. System uses an Arabic dictionary built manually by the speech-sounds of 13 speakers. MFCCs were used to extract the speech feature vectors. The vocabulary consists of 33 words. This paper shows the design and implementation of Hindi speech system. The system uses a vocabulary of 113 words. The developed system is giving good recognition results for both speaker dependent and speaker independent environments [6].

III. Statistical Framework of an ASR

ASR as shown in Fig. 1 mainly comprises of five parts: Acoustic Analysis for feature extraction, Acoustic model based on statistical HMM approach, Language model, Pronunciation dictionary and the decoder for recognition.

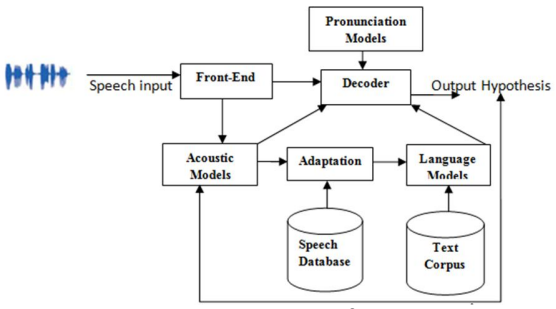

Fig. 1: Block diagram of ASR [4]

The sound waves captured by a microphone at the front end are fed to the acoustic analysis module. In this module the input speech is first converted into series of feature vectors which are then forwarded to the decoder.

This decoding module with the help of acoustic, language and pronunciation models comes up with the results. Mainly, the speech recognition problem can be divided into the following four step i.e. signal parameterization using a feature extraction technique such as MFCC or PLP, acoustic scoring with Gaussian mixture models (GMMs), sequence modeling with hidden Markov models (HMMs) and generating the competitive hypotheses using the score of knowledge sources (acoustic, language and pronunciation models) and selecting the best as final output with the help of a decoder [4].

IV. Automatic Speech Recognition System architecture

The developed speech recognition system architecture is shown in figure 1. It consists of two modules, training module and testing module. Training module generates the system model which is to be used during testing. The various phases used during ASR are:

4.1 Preprocessing: Speech-signal is an analog waveform which cannot be directly processed by digital systems. Hence preprocessing is done to transform the input speech into a form that can be processed by recognizer (Becchetti, 2008). To achieve this, firstly the speech-input is digitized. The digitized (sampled) speech-signal is then processed through the first-order filters to spectrally flatten the signal. This process, known as pre-emphasis, increases the magnitude of higher frequencies with respect to the magnitude of lower frequencies. The next step is to block the speech-signal into the frames with frame size ranging from 10 to 25 milliseconds and an overlap of 50%–70% between consecutive frames [3].

4.2 Feature Extraction: The goal of feature extraction is to find a set of properties of an utterance that have acoustic correlations to the speech-signal, that is parameters that can somehow be computed or estimated through processing of the signal waveform. Such parameters are termed as features. The feature extraction process is expected to discard irrelevant information to the task while keeping the useful one [1].

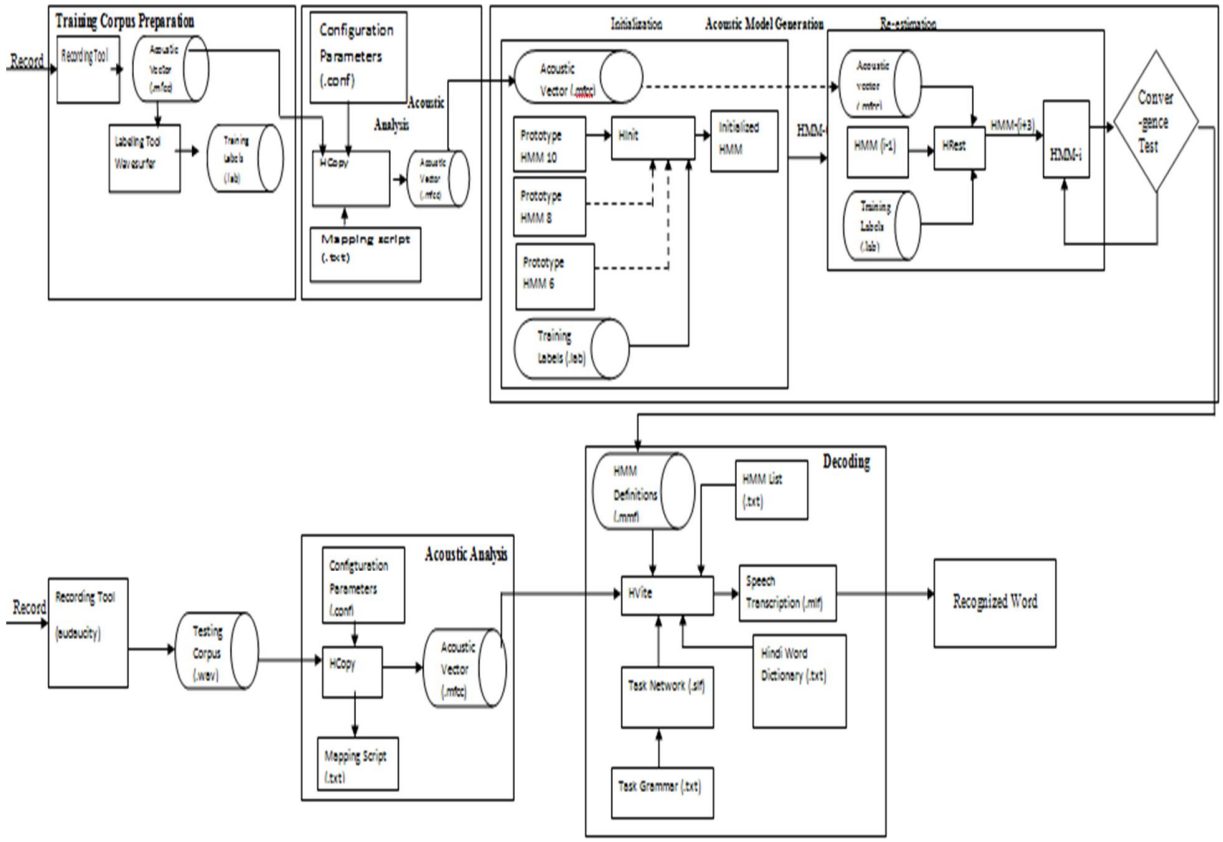

Fig. 1: Developed ASR system architecture

augmenting these measurements with some perceptually meaningful derived measurements (i.e. signal parameterization), and statically conditioning these numbers to form observation vectors (Jain et al, 2010)

4.3 Model Generation: The model is generated using various approaches such as Hidden Markov Model (HMM) (Huang et al., 1990), Artificial Neural Networks (ANN) (Wilinski et al., 1998), Dynamic Bayesian Networks (DBN) (Deng, 2006), Support Vector Machine (SVM) (Guo and Li, 2003) and hybrid methods (i.e. combination of two or more approaches). Hidden Markov model has been used in some form or another in virtually every state-of-the-art speech and speaker recognition system (Aggarwal and Dave, 2010).

Pattern Classifier: Pattern classifier component recognizes the test samples based on the acoustic properties of word. The classification problem can be stated as finding the most probable sequence of words W given the acoustic input O (Jurafsky and Martin, 2009), which is computed as:

$$W = \text{argmax}_W P(W|O). P(W)/P(O) \quad \ldots (1)$$

Given an acoustic observation sequence O, classifier finds the sequence W of words which maximizes the probability P(O |W).P(W) . The quantity P(W) , is the prior probability of the word which is estimated by the language model. P(O |W) is the observation likelihood, called as acoustic model.

IV. Hidden Markov Model and HTK

Hidden Markov model is a doubly stochastic process, generated by two interrelated mechanisms, an underlying Markov chain having a finite number of states and a set of random functions, one of which is associated with each state. At discrete instances of time, one process is assumed to be in some state representing the temporal variability and an observation is generated by another process corresponding to the current state representing the spectral variability. These two stochastic processes have been successfully used to model speech variability and flexible enough for building practical ASR systems. For ASR, only the observed sequence of events is known and underlying transition process

is unobservable. That is why it is called a "hidden" Markov model [1].

HTK is the "Hidden Markov Model Toolkit" developed by the Cambridge University Engineering Department (CUED). This toolkit aims at building and manipulating Hidden Markov Models (HMMs). HTK is primarily used for speech recognition research HTK consists of a set of library modules and tools available in C source form.

V. Hindi Character Set

Hindi is mostly written in a script called Nagari or Devanagari which is phonetic in nature. Hindi sounds are broadly classified as the vowels and consonants (Velthuis, 2011).

5.1 Vowels: In Hindi, there is separate symbol for each vowel. There are 12 vowels in Hindi language. The consonants themselves have an implicit vowel + (अ). To indicate a vowel sound other than the implicit one (i.e. अ), a vowel-sign (Matra) is attached to the consonant. The vowels with equivalent Matras are given in table 5.1 [1].

5.2 Consonants: The consonant set in Hindi is divided into different categories according to the place and manner of articulation. There are divided into 5 Vargs (Groups) and 9 non-Varg consonants. Each Varg contains 5 consonants, the last of which is a nasal one. The first four consonants of each Varg, constitute the primary and secondary pair. The primary consonants are unvoiced whereas secondary consonants are voiced sounds. The second consonant of each pair is the aspirated counterpart of the first one. Remaining 9 non Varg consonants are divided as 5 semivowels, 3 sibilants and 1 aspirate (Rai, 2005). The complete Hindi consonant set with their phonetic property is given in table 5.2 [1].

Vowel	अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	औ	ऋ	ऋ
Sibilants	-	ा	f ि	ी	ु	ू	े	ै	ो	ौ	ृ	ँ?

Table 5.1: Hindi Vowel Set [1]

Phonetic Property	Primary Consonants (unvoiced)		Secondary Consonants (voiced)		Nasal
Category	Un-aspirated	aspirated	un-aspirated	aspirated	
Gutturals (कवगर्)	क	ख	ग	घ	ङ
Patatals (चवगर्)	च	छ	ज	झ	ञ
Cerebrals (टवगर्)	ट	ठ	ड	ढ	ण
Dental (तवगर्)	त	थ	द	ध	न
Labials (पवगर्)	प	फ	ब	भ	म
Semivowels	य, र, ल, व				
Sibilants	श, ष, स				
Aspirate	ह				

Table 5.2: Hindi Consonant Set [1]

VI. IMPLEMENTATION

In this section, implementation of the speech system based upon the developed system architecture has been presented.

6.1 System Description

Hindi Speech recognition system is developed using HTK toolkit on the Linux platform. HTK v3.4 and ubuntu10.04 are used. Firstly, the HTK training tools are used to estimate the parameters of a set of HMMs

using training utterances and their associated transcriptions [2].
Secondly, unknown utterances are transcribed using the HTK recognition tools (Hidden Markov Model Toolkit, 2011). System is trained for 113 Hindi words. Word model is used to recognize the speech.

6.2 Data Interpretation
Training and testing a speech recognition system needs a collection of utterances. System uses a data-set of 113 words. The data is recorded using unidirectional microphone Sony F-V120. Distance of approximately 5-10 cm is used between mouth of the speaker and microphone. Recording is carried out at room environment. Sounds are recorded at a sampling rate of 16000 Hz. Voices of nine people (5 males and 4 females) are used to train the system. Each one is asked to utter each word three times. Thus giving a total of 3051(113*3*9) speech files. Speech files are store in .wav format. Velthuis (Velthuis, 2011) transliteration developed in 1996 by Frans Velthuis is used for transcription

6.3 Feature Extraction
During this step, the data recorded is parameterized into a sequence of features. For this purpose, HTK tool HCopy is used. The technique used for parameterization of the data is Mel Frequency Cepstral Coefficient (MFCC). The input speech is sampled at 16 kHz, and then processed at 10 ms frame rate with a Hamming window of 25 ms. The acoustic parameters are 39 MFCCs with 12 mel cepstrum plus log energy and their first and second order derivatives [1].

6.4 Training the HMM
For training the HMM, a prototype HMM model is created, which are then re-estimated using the data from the speech files. Apart from the models of vocabulary words, model for silent (sil) must be included [1]. For prototype models, authors uses 6, 8 and 10 state HMM in which the first and last are non-emitting states. The prototype models are initialized using the HTK tool HInit which initializes the HMM model based on one of the speech recordings. Then HRest is used to re-estimate the parameters of the HMM model based on the other speech recordings in the training set [1]. **HVite** to generate the output in a transcription file (.mlf). The HVite tool processes the signal using Viterbi Algorithm, based on token passing algorithm, which matches it against the recognizer's Markov models [3].

6.5 Performance Evaluation
The performance of the system is tested against speaker independent parameter by using two types of speakers: one who are involved in training and testing both and the other who are involved in only testing. The second parameter for checking system performance by varying no. of states in HMM topology. A total of 6 distinct speakers are used for this and each one is asked to utter 30-46 words.

6.6 Results
The table 6.6.1 to 6.6.6 shows the evaluation results of the H-SRS. The results shown reveal that when the word length is less and number of states is less, then the performance of the system is better. But as the word length increases and the number of states decrease then the system performance degrades. Since the word length and the number of states is increased, then implemented system performs well. The performance of the system (with 10 states in HMM topology) lies in the range of 96% and 95% with word error rate 6% and 8%.

Recognition by speakers involved both in training and testing: -

Speaker Number	No. of spoken words	Length of word (in characters)	No. of states in HMM topology	No. of Recognized words	% Word accuracy	Word error rate
S1	46	2	6	43	93.47	6.53
S2	37	2	6	35	94.59	5.41
S3	33	2	6	31	93.3	6.7
Total	116			109	93.96	6.04

Table 6.6.1 Recognition by speakers involved in training and testing with 6 states in HMM topology

Speaker Number	No. of spoken words	Length of word (in characters)	No. of states in HMM topology	No. of Recognized words	% Word accuracy	Word error rate
S4	30	3	8	27	90	10
S5	45	3	8	41	91	9
S 6	33	3	8	31	93.93	6.07
Total	108			99	91.66	8.34

Table 6.6.2 Recognition by speakers involved in training and testing with 8 states in HMM topology

Speaker Number	No. of spoken words	Length of word (in characters)	No. of states in HMM topology	No. of Recognized words	% Word accuracy	Word error rate
S1	46	3	10	44	95.65	4.35
S2	32	3	10	32	100	0
S3	40	3	10	38	95	5
Total	118			114	96.61	3.39

Table 6.6.3 Recognition by speakers involved in training and testing with 10 states in HMM topology

Recognition by speakers involved only in testing: -

Speaker Number	No. of spoken words	Length of word (in characters)	No. of states in HMM topology	No. of Recognized words	% Word accuracy	Word error rate
S4	40	2	6	36	90	10
S5	37	2	6	35	94.59	5.41
S6	46	2	6	43	93.47	6.53
Total	123			114	92.68	7.32

Table 6.6.4 Recognition by speakers involved only in testing with 6 states in HMM topology

Speaker Number	No. of spoken words	Length of word (in characters)	No. of states in HMM topology	No. of Recognized words	% Word accuracy	Word error rate
S1	36	3	8	33	91.66	8.34
S2	31	3	8	28	90.32	9.68
S3	35	3	8	32	91.42	8.58
Total	102			93	91.17	8.83

Table 6.6.6 Recognition by speakers involved only in testing with 8 states in HMM topology

Speaker Number	No. of spoken words	Length of word (in characters)	No. of states in HMM topology	No. of Recognized words	% Word accuracy	Word error rate
S4	39	3	10	38	97.43	2.57
S5	34	3	10	32	94.11	5.89
S6	38	3	10	36	94.73	5.27
Total	111			106	95.49	4.51

Table 6.6.5 Recognition by speakers involved only in testing with 10 states in HMM topology

VII. CONCLUSION

An efficient, abstract and fast ASR system for regional languages like Hindi is need of the hour. The work implemented in the paper is a step towards the development of such type of systems. The work may further be extended to large vocabulary size and to spontaneous speech recognition. As shown in results, the system is sensitive to changing spoken methods and changing scenarios, so the accuracy of the system is a challenging area to work upon. Hence, various speech enhancements and noise reduction techniques may be applied for making system more efficient, accurate and fast.

REFERENCES

[1] Kuldeep Kumar R. K. Aggarwal, "Hindi speech recognition system using HTK", International Journal of Computing and Business Research, vol. 2, issue 2, May 2011.

[2] Kuldeep Kumar, Ankita Jain and R.K. Aggarwal, "A Hindi speech recognition system for connected words using HTK", Int. J. Computational Systems Engineering, vol. 1, no. 1, pp. 25-32, 2012.

[3] Mohit Dua, R.K.Aggarwal, Virender Kadyan and Shelza Dua, "Punjabi Automatic Speech Recognition Using HTK", IJCSI International Journal of Computer Science Issues, vol. 9, issue 4, no. 1, July 2012.

[4] M.A.Anusuya and S.K.Katti, "Speech Recognition by Machine: A Review", (IJCSIS) International Journal of Computer Science and Information Security, vol. 6, no. 3, pp. 181-205, 2009.

[5] Rajesh Kumar Aggarwal and Mayank Dave, "Acoustic modeling problem for automatic speech recognition system: conventional methods (Part I)," Int J Speech Technol, pp. 297–308, 2011.

[6] Kuldeep Kumar, Ankita Jain and R.K. Aggarwal, "A Hindi speech recognition system for connected words using HTK", Int. J. Computational Systems Engineering, vol. 1, no. 1, pp. 25-32, 2012.

[7] R.K. Aggarwal and M. Dave, "Performance evaluation of sequentially combined heterogeneous feature streams for Hindi speech recognition system", 01 September 2011.

[8] Rajesh Kumar Aggarwal and M. Dave, "Acoustic modeling problem for automatic speech recognition system: advances and refinements (Part II)", Int J Speech Technol, pp. 309–320, 2011.

[9] J. L. Flanagan, Speech Analysis, Synthesis and Perception, Second Edition, Springer-Verlag, 1972.

[10] Wiqas Ghai and Navdeep Singh,"Literature Review on Automatic Speech Recognition", International Journal of Computer Applications vol. 41– no.8, pp. 42-50, March 2012.

[11] R K Aggarwal and M. Dave, "Markov Modeling in Hindi Speech Recognition System: A Review", CSI Journal of Computing, **v**ol. 1, no.1, pp. 38-47, 2012.

[12] Preeti Saini and Parneet Kaur, "Automatic Speech Recognition- A Review", International journal of Engineering Trends & Technology, pp. 132-136, vol-4, issue-2, 2013.