IOSR Journal of Engineering
May. 2012, Vol. 2(5) pp: 1020-1023

Performance of Hindi Speech Isolated Digits In HTK Environment

Sharmila[1], Dr. Neeta Awasthy[2], Dr .R.K.Singh[3]
RAJKUMARGOEL INSTITUTE OF TECHNOLOGY GHAZIABAD,
RAJKUMARGOEL INSTITUTE OF TECHNOLOGY FOR WOMEN GHAZIABAD

Abstract: *Speech recognition is the way of converting an acoustic signal into the text similar to the meaning full information being conveyed by the source. today's, mainly Hidden Markov Model (HMMs) based speech recognizers are used. This research aims to make a speech recognition system for Hindi language recognition with Hidden Markov Model Toolkit (HTK). HTK recognizes the isolated digits using acoustic digits model. The system is trained for 10 Hindi digits. Training data has been collected from twenty four speakers. The experimental results show that the overall accuracy of the presented system.*

Key words: *Mel Frequency Cepstral Coefficient (MFCC); HTK; HMM; Automatic Speech recognition(ASR), Hindi speech isolated speech recognition*

1. INTRODUCTION

Speech is the transfer of facts from one person to another person . Everyone knows his tongue language from his childhood. It also gives us an efficient means of man-machine communication. Normally, transfer of information between human and machine is accomplished via keyboard, mouse etc. But human can speak more quickly instead of writing. Speech input takes high bandwidth information and relative ease of use. It also provides the user's hands and eyes to be busy with a task, which is particularly valuable when users are in motion or in natural field settings . so speech result is more effective and understandable than the text output. Speech interfacing provides the ways to these issues. Speech interfacing involves speech synthesis and speech recognition. Speech synthesizer takes the text as input and converts it into the speech output i.e. it act as text to speech converter. Speech recognizer converts the spoken word into text. This paper aims to develop and implements speech recognition system for Hindi language. Now a days, due to the versatile applications of speech recognition, it is the most promising field of research. Our daily life activities, depends on mobile applications, weather forecasting, agriculture, healthcare etc. all involves speech recognition 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. Used to transcribe unknown utterances and to evaluate system performance by comparing them to reference transcriptions.

Apart from introduction in section 1, the paper is organized as follows. Some of the related works are presented in section 2. Section 3 presents the architecture and functioning of proposed ASR. Section 4 describes the Hidden Markov Models and HTK. Hindi character set is shown in section 5. Section 6 deals with implementation work. Section 7 concludes the paper.

2. AUTOMATIC SPEECH ECOGNITION (ASR)SYSTEM ARCHITECTURE

The developed speech recognition system architecture is shown in figure A. ASR 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:

Pre-processing: Speech-signal is an analog waveform which cannot be directly processed by digital systems. Hence pre-processing is done to transform the input speech into a form that can be processed by recognizer (Becchetti, 2008). To find this, firstly the speech-input is digitized. The 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.

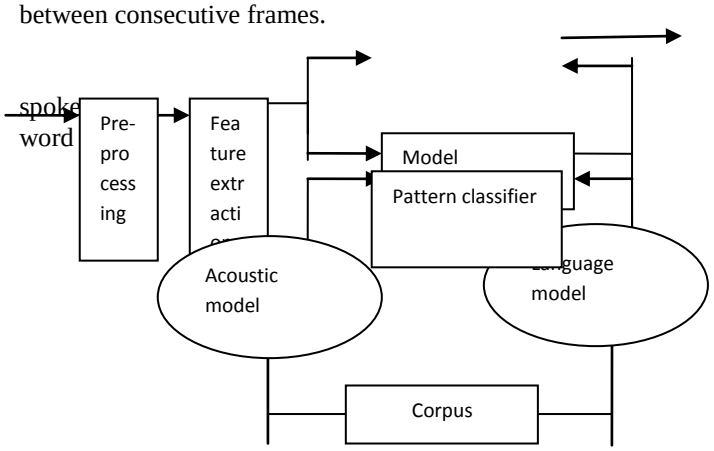


Figure 1.1 Developed ASR system architecture

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 speech 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. It includes the process of measuring some important term and characteristic of the speech signal such as energy or frequency response (i.e. signal measurement 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). 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 digits. The classification problem can be stated as finding the most probable sequence of digits given the acoustic input O (Jurafsky and Martin, 2009), which is computed as:

$$P(w/o)=p(o/w).p(w)/p(o) \quad \text{.....1}$$

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

3. HIDDEN MARKOV MODEL AND HTK

Hidden Markov Model (HMM) is a double stochastic process with one that is not directly observable. This hidden stochastic process can be observed only through another set of stochastic processes that can produce the observation sequence. HMMs are the so far most widely used acoustic models. The reason is just it provides better performance than other methods. HMMs are widely used for both training and recognition of speech system.
HMM are statistical frameworks, based on the Markov chain process with unknown parameters. Hidden Markov Model is a system which consists of nodes representing hidden states. The nodes are interconnected by links which describes the conditional transition probabilities between the states. Each hidden state has an associated set of probabilities of emitting particular visible states.
HTK is a toolkit for building Hidden Markov Models (HMMs). It is an open source set of modules written in ANSI C which always deals with speech recognition using the Hidden Markov Model4.

4. 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).
Vowels: In Hindi language, 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 (Marta) is attached to the consonant. The vowels with equivalent Matras are given in tableI.

Table I Hindi Vowel Set

vowel	अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	औ	ऋ	ॠ
matra	-	◌ ◌ा	◌	◌ ◌ी	◌	◌	◌	◌	◌ ◌ो	◌ ◌ौ	◌	◌

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 (has an additional "h" sound) of the first one. Thus four consonants of each Vargs are [unvoiced], [unvoiced, aspirated], [voiced], [voiced, aspirated] respectively. 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 II.

Phonetic Property / Category	Primary Consonants (unvoiced)		secondary consonants (voiced)		Nasal
Gutturals (कवग)	क	ख	ग	घ	ङ
Palatals (चवग)	च	छ	ज	झ	ञ
Cerebrals (टवग)	ट	ठ	ड	ढ	ण
Dental (तवग)	त	थ	द	ध	न
Labials (पवग)	प	फ	ब	भ	म
Semivowels	य, र, ल, व				
Sibilants	श, ष, स				

Aspirate	ह

Other Characters: Apart from consonants and vowels, there are some other characters used in Hindi language are: anuswar (ं), visarga (ः), chanderbindu (ँ), >, ?, @, A. Anuswar indicates the nasal consonant sounds. Anuswar sound depends upon the character following it. Depending upon the varg of following character, sound wise it represents the nasal consonants of that vargs

5. IMPLEMENTATION

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

5.1 SYSTEM DESCRIPTION

Hindi Speech recognition system is developed using HTK toolkit on the Linux platform. HTK3.4 and ubuntu10.04 are used. Firstly, the HTK training tools are used to estimate the for 10 Hindi Digits model. Parameters of a set of HMMs using training utterances and their associated transcriptions. Secondly, unknown utterances are transcribed using the HTK recognition tools (Hidden Markov Model Toolkit, 2011). System is trained is used to recognize the speech

5.2 DATA PREPARATION

For our purpose we recorded ten repetitions of the Hindi Digits from "shunya" to "nau" of speakers of different age groups and dialect. Then we manually segmented each digit sample. The speech signal is digitized using the A/D (analog to digital) converter. The appropriate sample rate must be chosen in order to insure the quality of the speech. The low pass anti-aliasing filter is implemented before the A/D converter so that the frequencies of the speech signal can be band limited to avoid aliasing between the base bands. The sampling rate is chosen as 16 KHZ. We know that speech signal is a time-varying signal, but over a short period of time (0 to 100 msec) its characteristics are almost stationary. When the speech signal is sampled or digitized, we can analyze the discrete-time representation in a short time interval, such as 10-30 ms.

Speech signal has been recorded using "cool software". The database consists of speech samples from twenty-four speakers. Sampling rate has been taken as 16 KHZ, 16-bit, mono. Speech signal is digitized by the software itself. Segmentation of speech signal is done using "cool" software. The database consists of 10 samples of each Hindi digit of each speaker, i.e., 2400 speech samples, 240 samples of each digit. The microphone was kept at a distance of 3-4 inches from the speaker's mouth. Microphones from "Intex" and "iball" company were used for recording.

5.3 FEATURE EXTRACTION

During this step, the data recorded is parameterized into a sequence of features. For this purpose, HTK tool H Copy 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
13 MFCCs with 12 el cepstrum plus log energy and their first and second order derivatives.

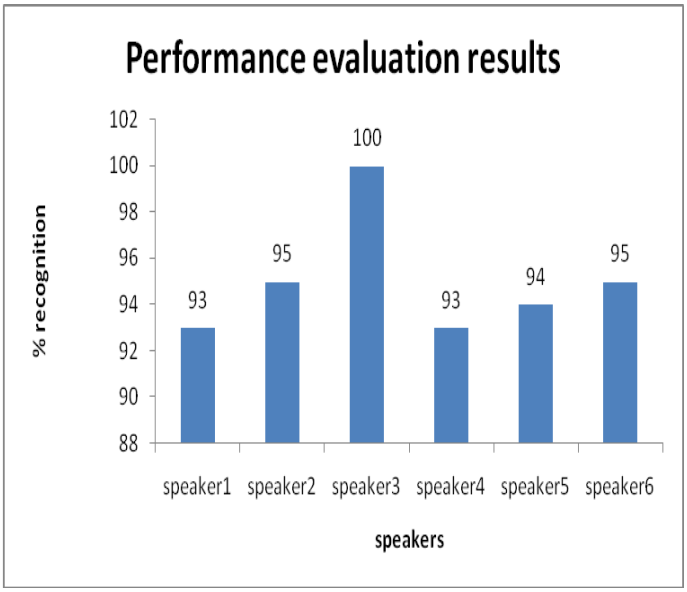
5.4 TRAINING OF THE HMM

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 digits, model for silent (sil) must be included.

For prototype models, authors uses 3-5 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.

5.5 PERFORMANCE RESULT

During evaluation, system is responsible for generating the transcription for an unknown utterance. The model generated during the training phase is responsible for evaluation. In order to evaluate the system performance, speakers are asked to utter each digit at least once a time. For testing ten speakers are used. The recognition results are shown



in graph.

6. CONCLUSION

In this research, the speech recognition system for Hindi language has been developed. The presented system recognizes the isolated digits using acoustic digit model. The training of the system has been done using 10 Hindi digits. During the development of the system, the training data has been collected from the 24 different speakers. The system has also been tested in the room environment. The implementation of the system has been done using Hidden Markov Model Toolkit (HTK). It has been observed from the performed experiments that the accuracy and word error rate of the proposed system is shown in the graph. The future works involves the development of system for more vocabulary size and to improve the accuracy of the system

7. References

1. Rabiner, L R (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE, Vol.77, No.2, pp. 257-286.
2. Huang, X D, Ariki, Y and Jack M A (1990) Hidden Markov Models for Speech Recognition. Edinburg University Press.
3. Wilinski, P, Solaiman, B, Hellion A and Czamecki, W (1998) Towards the Border between Neural and Markovian Paradigms. IEEE Transactions on Systems, Man and Cybernetics. Vol.28, No. 2, pp. 146-159.Indian Script Code for Information Interchange – ISCII (1999) Bureau of Indian Standards. New Delhi. India.
4. Pruthi T, Saksena, S and Das, P K (2000) Swaranjali: Isolated Word Recognition for Hindi Language using VQ and HMM. International Conference on Multimedia Processing and Systems (ICMPS), IIT Madras.
5. Guo, G and Li, S Z (2003) Content Based Audio Classification and Retrieval by SVMs. IEEE Trans. Neural Networks, 14, (January 2003), pp. 209-215.
6. Rai, N (2005) Isolated word speaker Independent Speech recognition for Indian Languages, Department of Computer Science and Engineering, Indian Institute of Technology, Kanpur.
7. Deng, Li (2006) Dynamic Speech Models: Theory, Applications, and Algorithms. Morgan and Claypool.
8. Gupta, R (2006) Speech Recognition for Hindi, M. Tech. Project Report, Department of Computer Science and Engineering, Indian Institute of Technology, Bombay, Mumbai.
8. Rao, R R, Nagesh, A, Prasad, K. and Babu, K E (2007) Text-Dependent Speaker Recognition System for Indian Languages. International Journal of Computer Science and Network Security, Vol. 7, No. 11.
9. Becchetti, C and Ricotti, L P (2008) Speech Recognition Theory and C++ Implementation, JohnWiley & Sons.
10. Deivapalan, P G and Murthy, H A (2008) A syllable-based isolated word recognizer for Tamil handling OOV words, The National Conference on Communications, pp. 267-271.
11. Elshafei, M, Al-Muhtaseb, H. and Al-Ghamdi, M (2008) Speaker-Independent Natural Arabic Speech Recognition System, the International Conference on Intelligent Systems ICIS 2008, Bahrain.
12. Syama, R (2008) Speech Recognition System for Malayalam. Department of Computer Science Cochin University of Science & Technology, Cochin.
13. Jurafsky, D and Martin, J H (2009) Speech and Language Processing, Pearson Education, New Delhi, India.
14. Young, S, Evermann, G, Gales, M, Hain, T, Kershaw D, Liu, X, Moore, G, Odell, J Ollason, D, Povey, D, Valtchev V and Woodland P (2009) The HTK Book, Microsoft Corporation and Cambridge University Engineering Department.
15. Aggarwal, R K and Dave, M (2010) Fitness Evaluation of Gaussian Mixtures in Hindi Speech Recognition System, First International Conference on Integrated Intelligent Computing, SJB Institute of Technology, Bangalore.
16. Al-Qatab, B A Q and Anion, R N (2010) Arabic Speech Recognition Using Hidden Markov Model Toolkit (HTK), International Symposium in Information Technology (ITSim). June15-17, Kuala Lumpur.
17. Jain, A, Aggarwal, R, Garg, A and Kumar, K (2010) Speech Recognition System using MFCC, Proceedings of All India Conference on Recent Emergence and Scope of Electronics Architecture, Haryana, India.
18. Cygwin, Retrieved Jan 15, 2011, from www.cygwin.com.
19. Hidden Markov Model Toolkit (HTK), Retrieved Jan 10, 2011, from http://htk.eng.cam.ac.uk. Velthuis, Retrieved Jan 29, 2011, from