AUTOMATIC SPEECH PROCESSING USING HTK FOR TELUGU LANGUAGE

SaiSujith Reddy Mankala[1], Sainath Reddy Bojja[2],
V. Subba Ramaiah[3] & R. Rajeswara Rao[4]
[1,2]Department of CSE, MGIT, JNTU, Hyderabad, India
[3]Assisant Professor, Department of CSE, MGIT, JNTU, Hyderabad, India
[4]Associate Professor, Department of CSE, JNTU Kakinada, 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 an already stored vocabulary. At present, mainly speech recognizers based on Hidden Markov Model (HMMs) are used. This paper's aim is to provide a speech recognition system for Telugu language using Hidden Markov Model Toolkit (HTK).It recognizes the isolated words using acoustic word model. The system is trained for 113 Telugu 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 96.64 and 95.46%.

KEYWORDS: *HMM, Wireless Network, Mobile Network, Virus, Worms & Trojon*

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 Telugu language, the most common language being used by the people in rural areas. In rural areas where people cannot operate keyboards and other input devices, Speech recognition will, indeed, play a very prominent role in promoting the usage of modern technology like internet by giving oral input. 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 help for Indian society if inputs to machines, mainly 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 [3][7](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. 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 Telugu language using the HTK [1] open source tool.

II. MOTIVATION

Due to its versatile applications, speech recognition [5] will be one the important fields of research. Many of our daily life activities, like getting information about weather, health care etc.would be enhanced using speech recognition. Commanding a device vocally to get information about weather current affairs etc. on internet or on mobile is much easier than giving inputs via keyboard or mouse. International organizations like Microsoft, Dragon-Naturally-Speech are working on this field especially for many languages. Some works for south Asian languages including Telugu 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 none of the works provided efficient solution for Telugu.

III. STATISTICAL FRAMEWORK OF AN ASR

ASR [8] implementation as shown in Fig. 1 mainly comprises of five functions: Acoustic Analysis for extracting features, Acoustic model based on statistical HMM technique, Language model, Pronunciation dictionary and the decoder used for recognition.

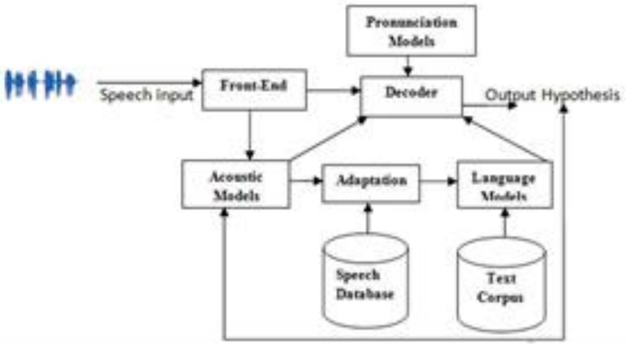


Figure 1. Block Diagram of an ASR.

The sound waves captured by a microphone at the front end are fed to the acoustic module. In this module the input speech is first converted into series of feature vectors which are then forwarded to the decoders. This decoding module with the help of acoustic, language and pronunciation models comes up with the results. Mainly, the speech recognition process can be divided into four logical steps i.e. Signal Parameterization using a feature extraction technique such as MFCC or PLP, acoustic scoring with Gaussian mixture models (GMMs), Sequence modelling using Hidden Markov models (HMMs) and generating the competitive hypothesis using the score of knowledge sources and selecting the best as final output with the help of a decoder.

IV. AUTOMATIC SPEECH RECOGNITION SYSTEM ARCHITECTURE

The architecture for developed speech recognition [6] system is shown in figure 1. It consists of two parts, training and testing. The Training module generates the system model which is used during testing. Various phases used during ASR are:

4.1 Pre-processing

Speech signal is an analog waveform which cannot be directly processed by digital systems. To convert analog signal to digital system accessible one, pre-processing is done. The resulting digitized speech signal (sampled) is then processed through the first-order filters to spectrally flatten the signal. This method, known as pre-emphasis, increases the difference between the magnitudes of higher frequencies with respect to the magnitude of lower frequencies. Then 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.

4.2 Feature Extraction

Feature extraction [2] is the process of extracting relevant information from the speech signal. The goal of feature extraction is to find a set of properties of an utterance that have acoustic correlations to the speech-signal. In other words parameters that can somehow be computed or estimated through processing of the signal waveform are found. Such parameters are called features.

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). This Hidden Markov model has been used in some form or another in virtually almost every state-of-the-art speech and speaker recognition system.

Pattern classifier component recognizes the test samples based on the acoustic properties of word. A Classification problem finds the most probable sequence of words W given the acoustic input O (Jurafsky and Martin, 2009), as follows:

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

For an acoustic sequence O, the classifier finds the sequence of words W 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. The quantity P (O |W) is the likelihood, known as acoustic model.

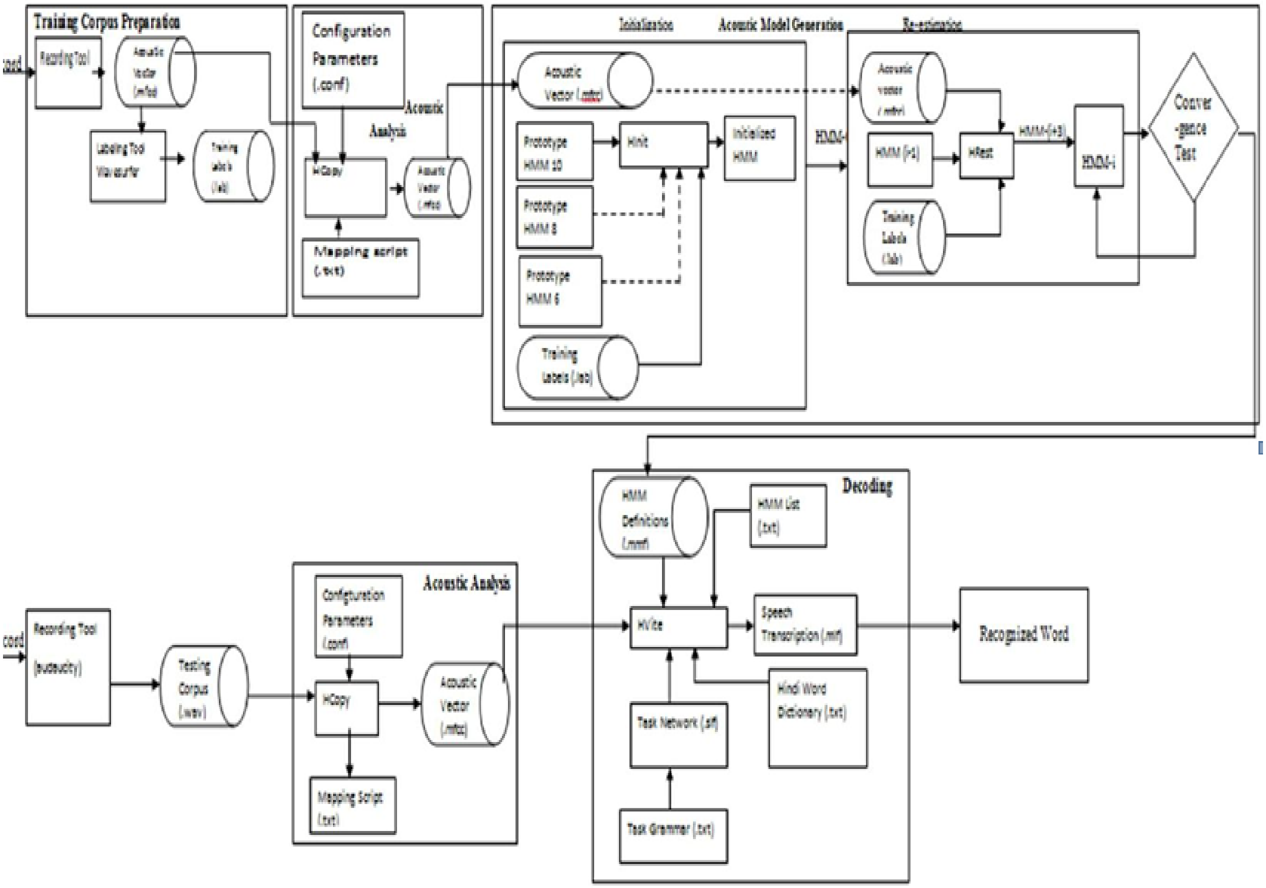
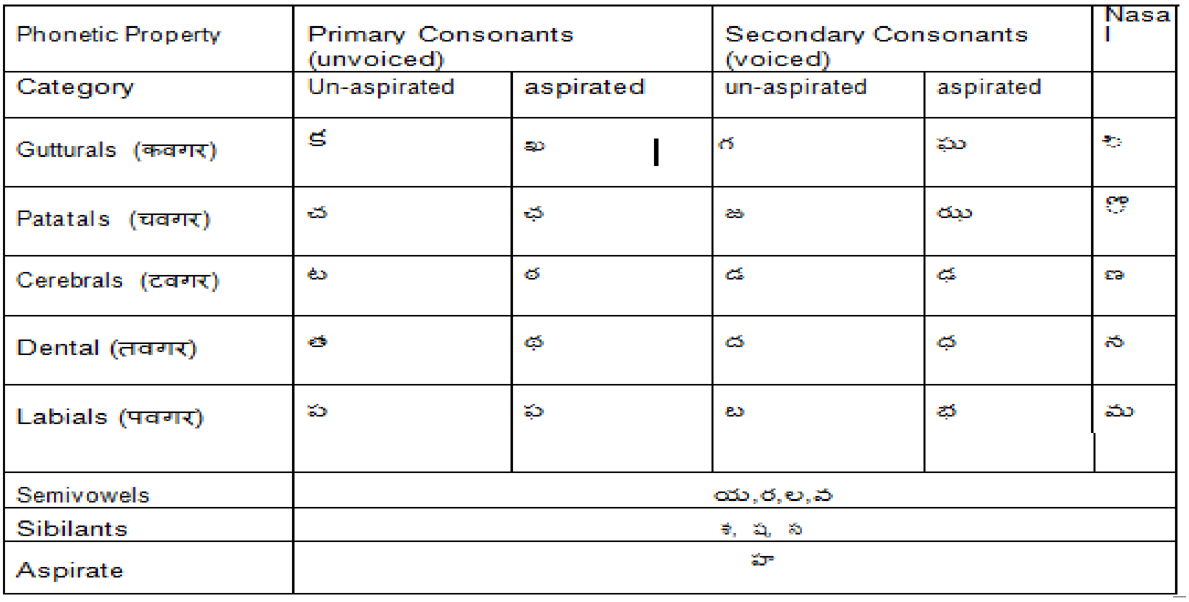


Figure 2. Architecture of ASR system

Table 1. Telugu Vowel Set.

Vowel	అ	ఆ	ఇ	ఈ	ఉ	ఊ	ఎ	ఏ	ఒ	ఓ	ై	ౌ

Table 2. Telugu Consonant Set



Phonetic Property	Primary Consonants (unvoiced)		Secondary Consonants (voiced)		Nasal
Category	Un-aspirated	aspirated	un-aspirated	aspirated	
Gutturals (कवगर)	క	ఖ	గ	ఘ	ఙ
Patatals (चवगर)	చ	ఛ	జ	ఝ	ఞ
Cerebrals (टवगर)	ట	ఠ	డ	ఢ	ణ
Dental (तवगर)	త	థ	ద	ధ	న
Labials (पवगर)	ప	ఫ	బ	భ	మ
Semivowels	య,ర,ల,వ				
Sibilants	శ, ష, స				
Aspirate	హ				

V. Hidden Markov Model And HTK

Hidden Markov Model is a doubly stochastic process, based on two related methods. First is an underlying Markov chain having a finite number of states and the second is using a set of random functions, and one of them 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. 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.

VI. Telugu Character Set

Telugu is mostly written in a script called Telugu script which is phonetic in nature. Telugu is broadly classified into vowels and consonants.

6.1 Vowels

In Telugu, there is separate symbol for each vowel. There are 12 vowels [9] in Telugu language. The consonants themselves have an implicit vowel. The vowels with equivalent matras are given in the table 1.

6.2 Consonants

The consonant set in Telugu is divided into different categories depending on place and manner of articulation. There are divided into 5 Vargs (Groups) and 9 non-Vargs. Each Varg contains 5 consonants, and the last one is a nasal one. The first four consonants of each group (Varg), constitute two pairs namely primary and secondary. Primary consonants are not voiced whereas secondary consonants are voiced sounds. In each pair, the second consonant is the aspirated counterpart of the first one. Remaining 9 non Varg consonants are divided as 5 semivowels, 1 aspirate and 3 sibilants. The complete Telugu consonant set with their phonetic property is given in table 2.

VII. Implementation

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

7.1 System Description

Telugu Speech recognition system is developed using HTK v3.4 toolkit on the Linux platform. Firstly, the HTK training tools are used to estimate the parameters of a set of HMMs using training utterances and their associated transcriptions.

Secondly, the unknown utterances are transcribed using the HTK [4] recognition tools. System is trained for 113 Telugu words. Word model is used to recognise the speech.

7.2 Data Interpretation

Training and testing a speech recognition system requires 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. A sampling rate of 16000 Hz is used to carry out recording at room environment. 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 transliteration developed in 1996 by Frans Velthuis is used for transcription.

7.3 Feature Extraction

During this step, the data recorded is parameterized into a feature sequence.HTK tool HCopy is used for this purpose. The technique used for the parameterization of data is Mel Frequency Cepstral Coefficient (MFCC). Sample rate of 16 kHz is used for the input speech, and then processed at 10 ms frame rate with a hamming window of 25 ms. The acoustic parameters are 35 MFCCs with 12 mel cepstrum plus log energy and their first and second order derivatives.

7.4 Training the HMM

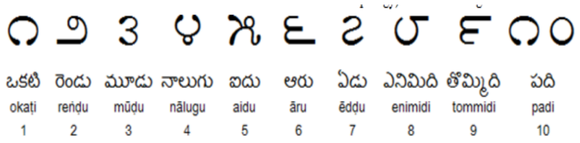
For training the HMM, a prototype HMM model is created, and are then re-estimated using the data from the speech files. Apart from the models of vocabulary words, model for silent (sil) must be included. For prototype models, authors use 6, 8 and 10 state HMM and among them and last are non-emitting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

7.5 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 number 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.

7.6 Results

The tables show 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%.



౧	౨	౩	౪	౫	౬	౭	౮	౯	౧౦
ఒకటి	రెండు	మూడు	నాలుగు	ఐదు	ఆరు	ఏడు	ఎనిమిది	తొమ్మిది	పది
okaṭi	reṅḍu	mūḍu	nālugu	aidu	āru	ēḍḍu	enimidi	tommidi	padi
1	2	3	4	5	6	7	8	9	10

Figure 3. ASR numbering system

7.7 Recognition by Speakers involved in Training and Testing

Table 3. Recognition by speakers involved

Speaker No.	No of words spoken	Length of word	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 4. Recognition by speakers involved in training and testing with 8 states in HMM topology

Speaker No.	No.of words spoken	Length of word	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
S6	33	3	8	31	93.93	6.07
Total	108			99	91.66	8.34

Table 5. Recognition by speakers involved in Training and testing with 10 states in HMM topology

Speaker Number	No. of spoken words	Length of word	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. Recognition by speakers involved in Training and testing with 10 states in HMM topology

Speaker Number	No. of spoken words	Length of word	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	03.47	6.53
Total	123			114	92.68	7.32

VIII. CONCLUSION

An efficient, abstract and fast ASR system for regional languages like Telugu is very much required in today’s world. 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. The system is sensitive to changing spoken methods and changing scenarios as shown in the results. 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 more accurate, effective efficient and fast systems.

REFERENCES

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

[2]. H.S. Jayanna, and S.R.M. Prasanna, “Analysis, feature extraction, modeling and testing techniques for speaker recognition”, IETE Technical Review, 2009, Vol. 26, issue 3, pp. 181-190, 2009.

[3]. Rajesh Kumar Aggarwal & Mayank Dave, "Acoustic modeling problem for automatic speech recognition system: conventional methods (Part I)," Int. J Speech Technology, pp. 297–308, 2011.
[4]. Kuldeep Kumar, Ankita Jain & 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.
[5]. Rajesh Kumar Aggarwal & M. Dave, "Acoustic modeling problem for automatic speech recognition system: advances and refinements (Part II)", Int. J Speech Tech., no. l, pp. 309– 320, 2011.
[6]. J.L.Flanagan, *Speech Analysis, Synthesis and Perception*, Second Edition, Springer-Verlag, 1972.
[7]. Wiqas Ghai & Navdeep Singh, "Literature Review on Automatic Speech Recognition", International Journal of Computer Applications vol. 41– no.8, pp. 42-50, March 2012.
[8]. Preeti Saini & Parneet Kaur, "Automatic Speech Recognition- A Review", International journal of Engineering Trends & Technology, pp. 132-136, vol-4, issue-2, 2013.
[9]. R K Aggarwal and M. Dave, "Markov Modeling in Hindi Speech Recognition System: A Review", CSI Journal of Computing, vol. 1, no.1, pp. 38-47, 2012.

AUTHORS

Sai Sujith Reddy Mankala is pursuing his under graduation (Bachelor of Technology) in CSE at MGIT, Hyderabad. His area of interests includes Speech processing, Computer Networks and Data Base Management Systems.

Sainath Reddy Bojja is pursuing his undergraduation (bachelor of Technology) in CSE at MGIT, Hyderabad. His area of interests includes Speech processing, Computer Networks and Network Security.

V. Subba Ramaiah received his B.Tech. degree in Computer Science and Engineering from SITAMS, JNT University, Chittoor, India, in 2002 and the M.Tech. degree in Computer Science from SIT, JNT University, Hyderabad, India, in 2007. He has been working as Senior Assistant Professor in the department of Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad. His research interests are Speech and Pattern recognition.

R.Rajeswara Rao was born in India in 1976. He received his B.Tech. degree in Computer Science and engineering from Siddhartha Engineering College, Vijayawada, India, in 1999 and the M.Tech. degree in Computer Science and Engineering from College of Engineering, JNT University, Hyderabad, India, in 2003. He has completed his Ph.D degree in computer science and engineering from JNT University, Hyderabad, India, in 2010. His research interests are speech processing and pattern recognition.