OPEN ACCESS

International Journal Of Modern Engineering Research (IJMER)

Marathi Speech Recognition System Using Hidden Markov Model Toolkit

Sangramsing Kayte,

Department of Computer Science and Information Technology
Dr. Babasaheb Ambedkar Marathwada University, Aurangabad

ABSTRACT:- Speech recognition is the process of converting an acoustic waveform into the text similar to the information being conveyed by the speaker. In the present era, mainly Hidden Markov Model based speech recognizers are used. This paper aims to build a speech recognition system for Marathi language. Hidden Markov Model Toolkitis used to develop the system. It recognizes the isolated words using acoustic word model. The system is trained for 30 Marathi words. Training data has been collected from ten speakers. The experimental results show that the overall accuracy of the presented system is 89%.

Keywords:- HMM, HTK,MFCC, ASR,Marathi, ASR.

I. INTRODUCTION

Speech is the most natural way of communication. Everyone knows his tongue language from his childhood. It also provides an efficient means of man-machine communication. Generally, transfer of information between human and machine is accomplished via keyboard, mouse etc. But human can speak more quickly instead of typing. Speech input offers high bandwidth information and relative ease of use. It also permits 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 [1]. Similarly speech output is more impressive 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 [2]. This paper aims to develop and implements speech recognition system for Marathi language.

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 DragonNaturally-Speech as well as research groups are working on this field especially for European languages. However some works for south Asian languages including Marathi have also been donebut no one provides efficient solution for Marathi language [3][4]. The lack of effective Marathi speech recognition system and its local relevance has motivated the authors to develop such small size vocabulary system.

1.2 PAPER CONTRIBUTION

The authors have developed Marathi speech recognition system for isolated word. Hidden Markov Model (HMM) is used to train and recognize the speech that uses MFCC to extract the features from the speech-utterances. To accomplish this, Hidden Markov Model toolkit (HTK) designed for speech recognition is used. HTK is developed in 1989 by Steve Young at the Speech Vision and Robotics Group of the Cambridge University Engineering Department (CUED). Initially, HTK training tools are used to train HMMs using training utterances from a speech corpus. Then, HTK recognition tools are used to transcribe unknown utterances and to evaluate system performance by comparing them to reference transcriptions [5].

II. 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:

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 [6]. 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.

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. It includes the process of measuring some important characteristic of the signal such as energy or frequency response (i.e. signal measurement),

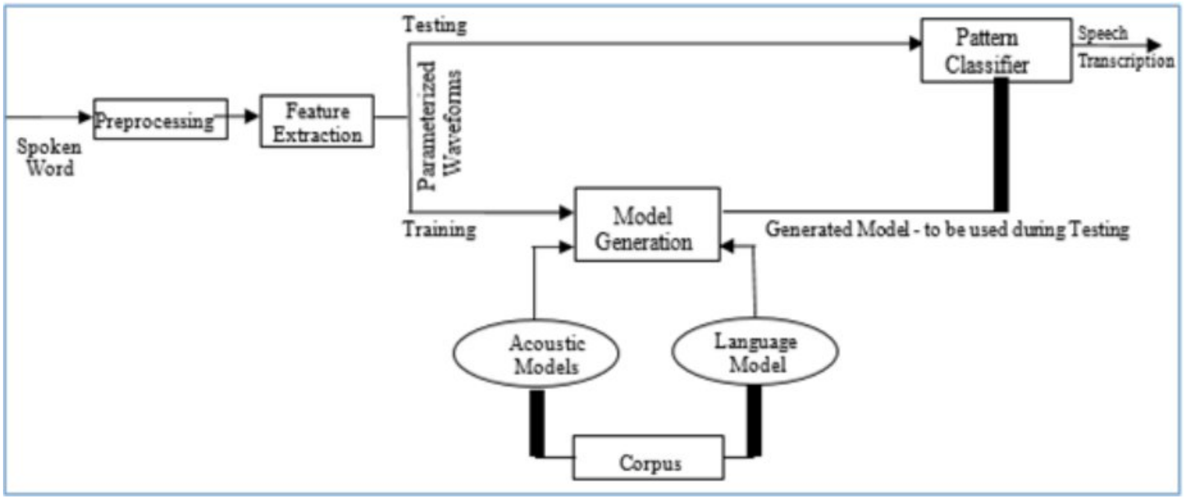


Fig 1: Developed ASR system architecture

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).
Model Generation: The model is generated using various approaches such as Hidden Markov Model (HMM) Artificial Neural Networks, Dynamic Bayesian Networks (DBN), Support Vector Machine (SVM) 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 [7].

III. HIDDEN MARKOV MODEL AND HTK

Hidden Markov Model (HMM) is a doubly 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 [8] [9].

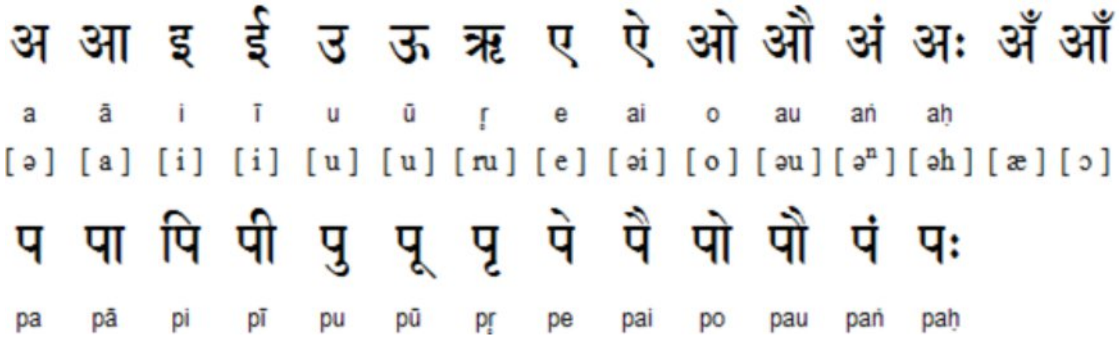
HMM are statistical frameworks, based on the Markov chain with unknown parameters. Hidden Markov Model is a system which consists of nodes representing hidden states. The nodes are interconnected by links which describes the conditional transition probabilities between the states. Each hidden state has 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 deal with speech recognition using the Hidden Markov Model. HTK mainly runs on the Linux platform. However, to run it on Windows, interfacing package Cygwin is used.

IV. MARATHI CHARACTER SET

Marathi is mostly written in a script called Nagari or Devanagari which is phonetic in nature. Marathi sounds are broadly classified as the vowels and consonants [10].
Vowels: In Marathi, there is separate symbol for each vowel. There are 12 vowels in Marathi 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 Matrass are given in table 2 [11][12].

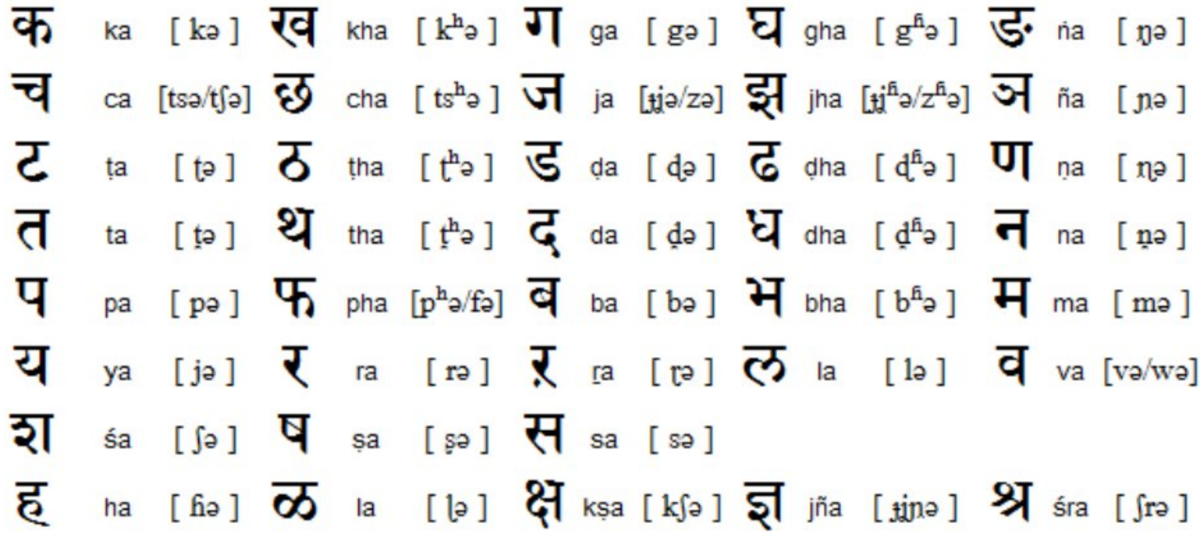
Table 2 Marathi Vowel Set



अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ	अं	अः	अँ	ऑ
a	ā	i	ī	u	ū	ṛ	e	ai	o	au	aṅ	aḥ		
[ə]	[a]	[i]	[i]	[u]	[u]	[ru]	[e]	[əi]	[o]	[əu]	[əⁿ]	[əh]	[æ]	[ɔ]
प	पा	पि	पी	पु	पू	पृ	पे	पै	पो	पौ	पं	पः		
pa	pā	pi	pī	pu	pū	pṛ	pe	pai	po	pau	paṅ	paḥ		

Consonants: The consonant set in Marathi 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. The complete Marathi consonant set with their phonetic property is given in table 3.

Table 3 Marathi Consonant Set



क	ka	[kə]	ख	kha	[kʰə]	ग	ga	[gə]	घ	gha	[gʱə]	ङ	ṅa	[ŋə]
च	ca	[tsə/tʃə]	छ	cha	[tsʰə]	ज	ja	[ʤ̤ə/zə]	झ	jha	[ʤ̤ʱə/zʱə]	ञ	ña	[ɲə]
ट	ṭa	[ʈə]	ठ	ṭha	[ʈʰə]	ड	ḍa	[ɖə]	ढ	ḍha	[ɖʱə]	ण	ṇa	[ɳə]
त	ta	[t̪ə]	थ	tha	[t̪ʰə]	द	da	[d̪ə]	ध	dha	[d̪ʱə]	न	na	[n̪ə]
प	pa	[pə]	फ	pha	[pʰə/fə]	ब	ba	[bə]	भ	bha	[bʱə]	म	ma	[mə]
य	ya	[jə]	र	ra	[rə]	ऱ	ṛa	[ɽə]	ल	la	[lə]	व	va	[və/wə]
श	śa	[ʃə]	ष	ṣa	[ʂə]	स	sa	[sə]						
ह	ha	[ɦə]	ळ	la	[ɭə]	क्ष	kṣa	[kʃə]	ज्ञ	jña	[ʤ̤ɲə]	श्र	śra	[ʃrə]

Other Characters: Apart from consonants and vowels, there are some other characters used in Marathi language are: anuswar (◌◌ं), visarga (◌◌ः), chanderbindu (◌◌ँ), >, ऽ, @, ◌ॉ. 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.

V. IMPLEMENTATION

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

5.1 SYSTEM DESCRIPTION

Marathi 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. Secondly, unknown utterances are transcribed using the HTK recognition tools. System is trained for 30 Marathi words. Word model is used to recognize the speech.

5.2 DATA PREPARATION

Training and testing a speech recognition system needs a collection of utterances. System uses a data-set of 30 words. The data is recorded using unidirectional microphones. 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 eight people (5 male and 3 female) are used to train the system. Each one is asked to utter each word four times. Thus giving a total of 960 ((8*4)*30) speech files. Speech files are stored in .wav format.

5.3 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.

5.4 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.

For prototype models, authors uses 5-11 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 EVALUATION

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 word at least once a time. For testing five speakers are used. The recognition results are shown in table 4. Overall word-accuracy and word-error rate of the system is 94.63% and 5.37% respectively

Table 4 Performance evaluation results

Subject Number	No. of spoken words	No. of Recognized word	% word accuracy	Word error rate
1	30	28	93.34	6.66
2	43	41	95.35	4.65
3	27	27	100.00	0.00
4	36	34	94.44	5.56
5	30	27	90.00	10.00

VI. CONCLUSION

In this paper, the speech recognition system for Marathi language has been developed. The presented system recognizes the isolated words using acoustic word model. The training of the system has been done using 30 Marathi words. During the development of the system, the training data has been collected from the eight 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 94.63% and 5.37%. The future works involves the development of system for more vocabulary size and to improve the accuracy of the system.

REFERENCES

[1]. Sangramsing N.kayte “Marathi Isolated-Word Automatic Speech Recognition System based on Vector Quantization (VQ) approach” 101th Indian Science Congress Jammu University 03th Feb to 07 Feb 2014.

[2]. Monica Mundada, Bharti Gawali, Sangramsing Kayte "Recognition and classification of speech and its related fluency disorders" International Journal of Computer Science and Information Technologies (IJCSIT)

[3]. Monica Mundada, Sangramsing Kayte, Dr. Bharti Gawali "Classification of Fluent and Dysfluent Speech Using KNN Classifier" International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 9, September 2014

[4]. Syama, R (2008) Speech Recognition System for Malayalam. Department of Computer ScienceCochin University of Science & Technology, Cochin.

[5]. 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.

[6]. 6)Becchetti, C and Ricotti, L P (2008) Speech Recognition Theory and C++ Implementation, John Wiley & Sons.

[7]. Huang, X D, Ariki, Y and Jack M A (1990) Hidden Markov Models for Speech Recognition. Edinburg University Press.

[8]. Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "A Marathi Hidden-Markov Model Based Speech Synthesis System" IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. I (Nov - Dec. 2015), PP 34-39e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197

[9]. 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.

[10]. Sangramsing Kayte, Monica Mundada "Study of Marathi Phones for Synthesis of Marathi Speech from Text" International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-4, Issue-10) October 2015

[11]. Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "Di-phone-Based Concatenative Speech Synthesis System for Hindi" International Journal of Advanced Research in Computer Science and Software Engineering - Volume 5, Issue 10, October-2015

[12]. Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte “Di-phone-Based Concatenative Speech Synthesis Systems for Marathi Language” OSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 5, Ver. I (Sep –Oct. 2015), PP 76-81e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197

[13]. Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "A Corpus-Based Concatenative Speech Synthesis System for Marathi" IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. I (Nov - Dec. 2015), PP 20-26e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197