Volume 6, Issue 9, September 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Speech Recognition based on Hidden Markov Model Toolkit (HTK) with BODO Language

[1]Zoha Arfeen, [2]Jyoti Aggarwal
[1] M.Tech Student, [2] Asistant Proferssor
[1, 2] Raj Shree Institute of Management and Technology, Dr. A.P.J.Abdul Kalam Technical University, Lucknow, Uttar Pradesh, India

Abstract: ***This Speech recognition is the process of converting an acoustic waveform into the text similar to the information being conveyed by the speaker. The speech recognition in multilingual is found to be very difficult to improve over separately trained systems. We experiments on different evaluation approach to speech recognition in multilingual, in which the phone sets are entirely distinct. The model has parameters not tied to specific states but that are shared across languages . The text-to-speech synthesis systems require a accurate prosody labels to generate natural-sounding speech. This research paper is to build a speech recognition system for Bodo language using Hidden Markov Model Toolkit (HTK). The system is trained for continuous Bodo speech and the continuous Bodo speech has been taken from male Bodo speakers.***

Keywords: Automatic Speech Recognition (ASR), Bodo, HMM, HTK, ,Isolated word ASR, Mel Frequency Cepstral Coefficient (MFCC),

I. INTRODUCTION

Speech system often requires simple information such as languages of the input, voice-gender i.e. male/female to be used to create the continuous speech and speech recognition sounds uttered by a speaker which are converted to equivalent waveform. Data are present in text ,audio and video format. With broader bandwidths there has been an increase in audio and video content on the Internet. There has been a huge increase in the amount of data generated and stored as computers and Internet but Automatic recognition of affect in speech is no longer for systems that interact with people using spoken language: It is an important ingredient for good service and a good respond to users, understanding, adapting to them. The Speech interfacing involves speech synthesis and speech recognition. The main function of Speech synthesizer is to take the text as input and converts it into the speech output and the Speech recognizer converts the spoken word into text. Our paper is concern to develop and implements a speech recognition system for Bodo language. This is involved to develop a TTS and ASR for Indian languages, and we consider it for Bodo language. In this approach statistics like word frequency, syllable frequency, word length, sentence length Etc is used to compare the corpora of ten Indian languages. In particular the following data are extracted for: [a) Word frequency distribution tables and the percentages of words in the corpus b) Number of distinct words required for coverage of certain percentage of corpus. [c) Syllable frequencies (unisyllables, bisyllables, Trisyllables) and pattern extraction [d] Word length distribution graphs and their analysis [f] Sentence length distribution graphs and their analysis. The Hidden Markov Model (HMM) is used to train and recognize the speech that uses MFCC to extract the features from the male speakers. For this, Hidden Markov Model toolkit (HTK) is designed for speech recognition. Hidden Markov Model toolkit (HTK) is developed in 1989 by Steve Young at the Speech Vision and Robotics Group of the Cambridge University Engineering Department (CUED) . At the begging , HTK training tools are used to train HMMs using Training utterances from a speech corpus. HTK recognition tools are used to transcribe unknown utterances and to evaluate system performance. A method using Gaussian Mixture Model (GMM) for statistical pattern classification is suggested to reduce computational load.

The major application areas of Automatic speech recognition(ASR) systems are dictation, controlling the programs, automatic telephone call processing and query based information system such as travel information system , weather report information system etc. Keeping it mind and its applications into consideration our paper aim to develop a GCC Compiler based speech recognizer for limited vocabulary based on HMM (Hidden Markov Model) using HTK open source toolkit in Linux environment for Bodo language.

II. HIDDEN MARKOV MODEL BASED SYNTHESIS HTS.

The **Hidden Markov Model** (HMM) is a popular statistical tool for modeling a wide range of time series data. In the context of natural language processing (NLP), HMMs have been applied with great success to problems such as part - of-speech tagging and noun-phrase chunking. The Hidden Markov Model (HMM) is a powerful statistical tool for modeling generative sequences that can be characterized by an underlying process generating an observable sequence. HMMs have found application in many areas interested in signal processing, and in particular speech processing, but have also been applied with success to low level NLP tasks such as part-of-speech tagging, phrase chunking, and

extracting target information from documents. A HMM consists of a number of states, each of which is associated with a probability density function. The model parameters are the set of probability density functions, and a transition matrix that contains the probability of transitions between states. HMM-based recognition algorithms are classified into two types, namely, phoneme level model and word-level model. The word-level HMM has excellent performance at isolated word tasks and is capable of representing speech transitions between phonemes. However, each distinct word has to be represented by a separate model which leads to extremely high computation cost (which is proportional to the number of HMM models). The phoneme model on the other hand can help reproduce a word as a sequence of phonemes. Hence new words can be added to the dictionary without necessitating additional models. Hence phoneme model is considered more suitable in applications with large sized vocabularies and where addition of word is a essential possibility. In our approach speech utterances are used to extract spectral (Mel-Cepstral Coeff.), excitation parameters In case of speech synthesis, pipelined RSA algorithm is used to find the most probable path through Hidden Markov Models that can generate speech signal feature vectors like MFCC (Mel Cepstral Coeff.) which are used to generate speech signal.

Mel Cepstral Coefficients and excitation parameters i.e. fundamental frequency *F*0 are extracted from the speech database and use them for Hidden Markov Models training acoustic models. Here HTK is a toolkit for building Hidden Markov Models (HMMs). It is an open source set of modules written in ANSI C, which is related to speech recognition using the Hidden Markov Model. Below there is an example of Hiden Markov Model is shown.

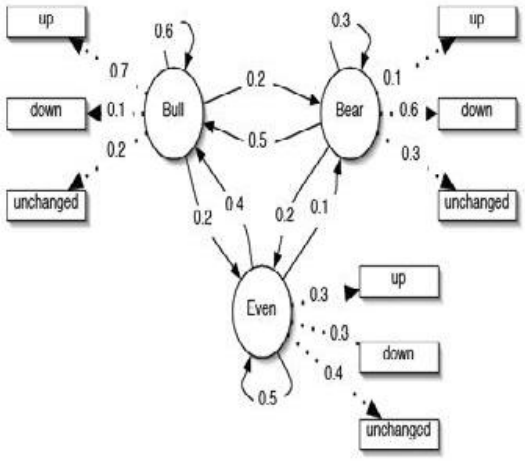

Fig-1 Hidden Markov Model Example

III. THE ACOUSTIC MODEL

In the statistical framework for speech recognition, the main problems are to find the most likely word sequence, which can be described by the equation **Ŵ=argw maxP(W/X).** Using the Bayes equation, we get **Ŵ= argw maxP(W/X)p(w)** The term *P*(*X*/*W*) in the above equation can be realized by the Acoustic model. An acoustic model is a file that contains a statistical representation of each distinct sound that makes up a spoken word. It contains the sounds for each word found in the Language model.

IV. ACOUSTIC MODEL GENERATION

The system is implemented here employ Hidden Markov Model (HMM) for representing speech sounds. HMM consists of a number of states, each of which is associated with a probability density function. The parameters of a HMM comprises of the parameters of the set of probability density functions, and a transition matrix that contains the probability of transition between states. Here the MFCC feature vectors are extracted from speech signals and their associated transcriptions are used to estimate the parameters of HMMs which is called ASR. The HMM Tool Kit, HTK-3.4 is used for training models over 38 context -dependent Bodo phonemes used in the application. The basic acoustic units are context dependent phonemes, that is, tri-phones modeled by left-to-right, 5-state, HMMs.

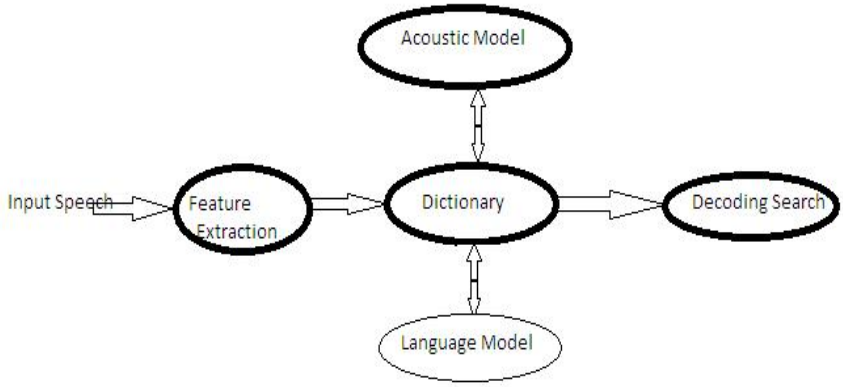

Fig-2 Acoustic Model

In case of speech recognize, the system generally consists of two phases. They are pre-processing and post-processing. In Pre-processing involves feature extraction and the post-processing comprises of building a speech recognition engine. The Speech recognition engine usually consists of knowledge about building an acoustic model, dictionary and grammar.

V. BODO LANGUAGE

Before 1953, the Bodo language had no standard form of writing. It had a history of using Deodhai, Roman and Assamese scripts. At present, Bodos adopted the Devanagari script. But, there is a huge difference in the usage of the letters in Bodo language from the Devanagari script. Bodo language shares some common salient features with other languages belonging to the Bodo group. These features are similar in terms of phonology, morphology, syntax, and vocabulary. Bodo language is closely associated with the Dimasa language of the state of Assam and with the Garo language of the state of Meghalaya, and also with Kokborok language of Tripura. It important to note that ,among the four districts of present Bodo land , namely, Kokrajhar, Chirang, Baksa and Udalguri, the language is heard in pure form only in the district of Udalguri.

VI. PHONOLOGICAL STRUCTURE

The Phonology is the study of speech sound and their functions within the sound system of a particular language. Phonemes are considered as the basic unit of a language. The Bodo phonemes consists of 6(six) vowels and 16(sixteen) consonants. Out of these 16 consonants 2(two) are semi vowel. They are as shown below-

a. Vowels : अ, आ, इ, उ, ए, औ

: ख, ग, ङ, ज, थ, द, न, फ, ब, म, र,

b. Consonants ल, स, ह

c. Semi Vowels : य, व

VII. FEATURE EXTRACTION

The Feature extraction involves conversion of speech waveform to parametric representation. The most important parameters in speech processing are found in the frequency domain. Since our research involves speech recognition, Mel scale cepstral analysis (MEL) is used to generate the MFCC, which characterizes various speech sounds. The MFCC is the evident example of a feature set that is extensively used in speech recognition. As the frequency bands are positioned logarithmically in MFCC it approximates the human system response more closely than any other system. The main technique of evaluating the MFCC is based on the short-term analysis, and hence from each frame a MFCC vector is evaluated. To extract the coefficients the speech sample is taken as the input and hamming window is applied to minimize the discontinuities of a signal. Then DFT is used to generate the Mel filter bank. In case of Mel frequency warping, the width of the triangular filters varies and so the log total energy in a critical band around the center frequency is included and after the warping the numbers of coefficients are computed. Finally for the cepstral coefficients calculation the Inverse Discrete Fourier Transformer is used. It transforms the log of the quefrench domain coefficients to the frequency domain where N is the length of the DFT. Now MFCC can be computed by using the formul

Mel(f)= 2595*log10(1+f/700)

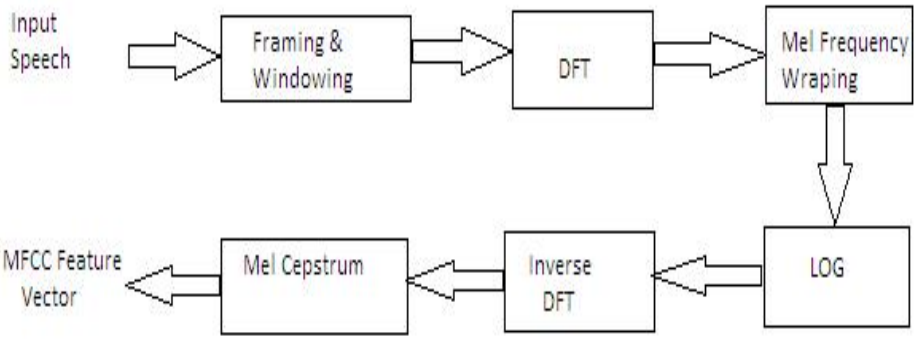

Fig-3 MFCC Block diagram

A. Pre-emphasis:

First the speech signal **s(n)** is sent to a high-pass filter, **s2(n) = s(n) - a*s(n-1)**, where **s2(n)** is the output signal and the value of a is in between 0.9 and 1. In our experiment we have used a = 0.95. The goal of pre-emphasis is to compensate the high frequency part. After pre-emphasis sounds became sharper with a smaller volume.

B. Frame blocking:

The input speech signal is then segmented into frames of 10~20 ms with optional overlap of 1/3~1/2 of the frame size. Generally the frame size is equal the power of two in order to facilitate the use of **Fast Fourier Transform (FFT),** otherwise we have to do zero padding to the nearest length of power of two.

C. Fast Fourier Transform (FFT):

It is an efficient algorithm to compute the Discrete Fourier Transform (DFT) and its inverse. The FFT computes the DFT and produces exactly the same result as evaluating the DFT directly; the main difference is that an FFT is much faster. Let *x*0, ...,*xN*-1 be complex numbers. The DFT is defined by the equation

N-1

XK = $\sum$ Xne $^{-i2\prod Kn/N}$, where K=0,……N-1 n=0

Evaluating the definition which requires *O*(*N*2) operations: there are *N* outputs *Xk*, and each output requires a sum of *N* terms. The FFT is a technique to compute the same results in O(*N* log *N*) operations. In our algorithm, known split radix FFT (RS-FFT) algorithm is used, which is a divide and conquer algorithm that recursively breaks down a DFT of any composite size *N* = *N*1*N*2 into DFTs of sizes *N*1 and *N*2, with O(*N*) multiplications by complex roots of unity called **twiddle** factors, at the time of perform FFT on a frame, we assume that the signal within a frame is periodic, and Continuous.

D. Windowing:

In windowing, each data frame is multiplied with a window function to keep the continuity of the first and the last points in the frame. If the signal in a frame is denoted by **s(n), n= 0,…N-1**, then the signal after windowing is **s(n)*w(n),** where **w(n)** is the window function. In our research, we used different types of window functions, such as Kaiser and Blackman windows , Hamming, Hanning,Rectangular, Welch ,Triangle.

The Welch equation , Welch(i) = 1 – $[(1- N/2)/N/2]^2$ Commonly used as a window for power spectral estimation. Welch window **where 0 <= i < N**

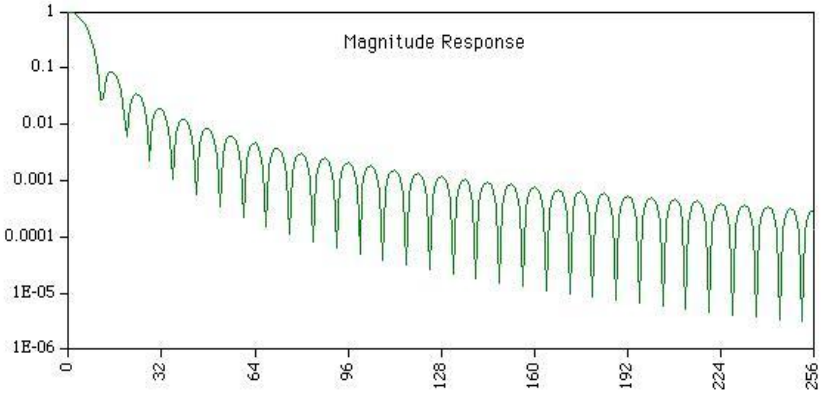

Fig-4 Window for power spectral estimation

E. The Language Model:

The Hamming window that we have used for the experiment is defined by the equation

W(n).54-0.46cos[2n/(N-1)]

where, N = number of samples in each frame. Let **X(n)=**input signal **and Y(n)** = Output signal.

Y(n)=X(n)W(n),

the Fast Fourier transform is used to convert each frame of **N** samples from time domain into frequency domain and then the components of the magnitude spectrum of the analyzed signal are calculated by using the equation

Y(w)=FFT[h(r)*x(r)]=H(w)X(w).

Next the main important point in signal processing is Mel-frequency transformation. The Compensation for non-linear perception of frequency is evaluated by the bank of triangular band filters with the linear distribution of frequencies so called Mel-frequency range which is described by the equation. **fmel=2595** Where **f** represents the frequency in linear range and **fme** the corresponding frequency in nonlinear Mel-frequency range. The **Apriori** probability of a word sequence on semantics, syntax, and pragmatics of the language is recognized. It can be understood by the Language Model that contains a list of words and their probability of occurrence in a sequence, which is independent of the acoustic signal. The probability of a word sequence is given as . p(w1,w2,w3,w4,……..wn)=p(w). By applying Chain rule the probability of n^{th} word is

$P(w1^n)$ = P(w1)P(w2|w1) $P(w2|w1^2)$…….$P(wn|w1^{n-1})$

$P(w_1^n)= \prod^{\prod}_{k=1} P(w_k|w_1^{k-1})$

Language Model or Grammar essentially defines constraints on what the Speech Recognition Engine expect as input can.

Since the speech can be described as an act of producing voice through the use of the vocal cords so a speech signal can be understood as a sequence of phonemes or observations. The important and useful parameters in speech recognitions are found in the frequency domain. The different features of speech can be evaluated by cepstral analysis - MFCCs: They have Speech-related information. _ LPCCs, which contain Speaker's information, and the number of feature vectors that constitute the observation sequence is of variable size These are the best sequences modeled by Hidden Markov Models.

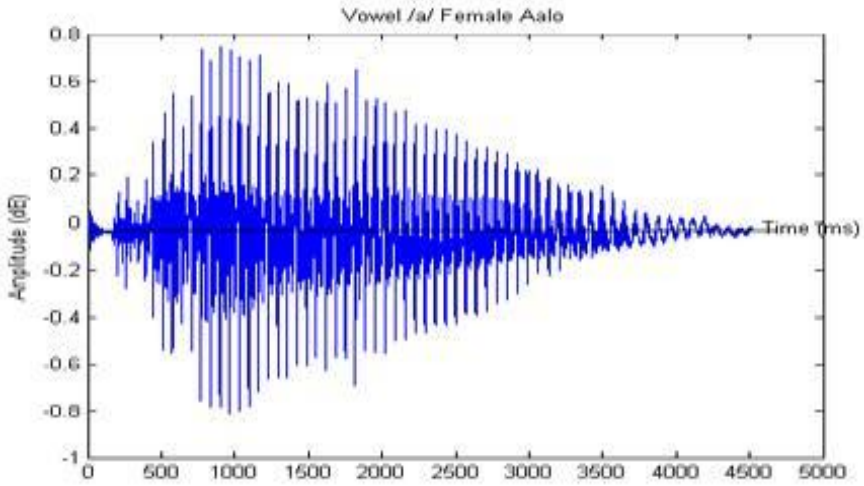

Fig- 5 Speech input information of Bodo speech

VIII. RESULT AND DISCUSSION

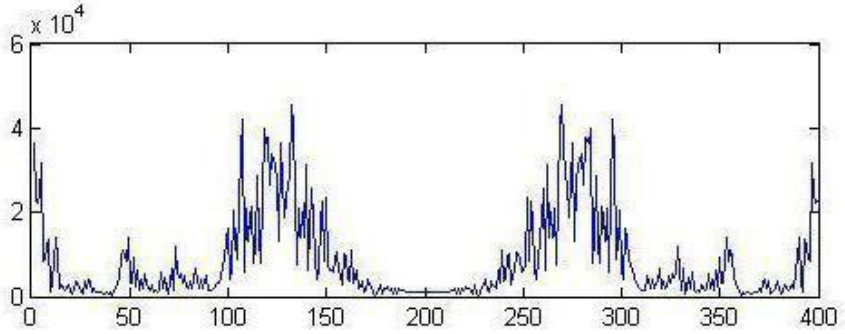

Fig-6 Mel frequency cepstral coefficients speech system

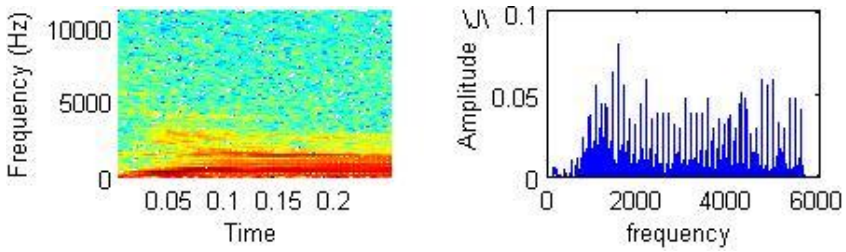

Fig-7 speech patern of Bodo speech

IX. CONCLUSION

Finally**,** an efficient and fast Automatic Speech Recognition system for regional languages say Bodo is needed in the present situation. Our experiment evaluated through this paper is a step towards the development of ASR system in Bodo language. Our work can be extended to emotional continuous speech recognition. From the results, the system is sensitive to changing spoken methods and changing scenarios, so the accuracy of the system is a challenging area to work and hence, various speech enhancements and noise reduction techniques may be applied for making system more fast, efficient and accurate.

REFERENCES

[1] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, The HTK Book, Microsoft Corporation and Cambridge University Engineering Department, 2009.

[2] W.A.Lea, " Trends in Speech Recognition", Prentice Hall, 1980.

[3] http://www-2.cs.cmu. edu/~robust/Tutorial

[4] Modelling Word Duration for Better Speech Recognition by Venkata Ramana Rao Gadde SRI International Menlo Park, USA

[5] R. K. Aggarwal, and M. Dave "Acoustic modeling problem for automatic speech recognition system: conventional methods (Part I)" International journal Speech Technology, Springer, Vol.14, issue 2, 2011.

[6] HTK "Hidden Markov Model Toolkit", available at http://htk.eng.cam.ac.uk,2012.

[7] Anusuya, M. A., & Katti, S. K.. Front end analysis of speech recognition: A review. International Journal of Speech Technology,Springer, Vol.14, pp. 99–145, 2011.

[8] SPHINX,Sphinx, available at http://cmusphinx.sourceforge.net/html/cmusphinx.php, 2011.

[9] K. Kumar and R. K. Aggarwal "Hindi Speech Recognition System using HTK" International journal of Computing and Business Research ISSN Vol. 2 issue 2 May 2011.

[10] R. Kumar "Comparison of HMM and DTW for Isolated Word Recognition of Punjabi Language" In Proceedings of Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Sao Paulo, Brazil. Vol. 6419 of Lecture Notes in Computer Science (LNCS), pp. 244– 252, Springer Verlag, November 8-11,

[11] B. A. Q. Al-Qatab and R. N. Ainon, "Arabic Speech Recognition Using Hidden Markov Model Toolkit (HTK)", Paper presented at International Symposium in Information Technology (ITSim). Kuala Lumpur, June 15-17, 2010.

