Volume 2, Issue 12, December 2012 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

HTK Based Telugu Speech Recognition

P. Vijai Bhaskar
AVNIET ,Hyderabad
A.P., India

Prof. Dr. S. Rama Mohan Rao
Principal, GCET, Hyderabad
A.P., India

A.Gopi
AVN IET ,Hyderabad
A.P., India

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

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

I. Introduction

Psychological processing involves converting information from one form to another. Speech system often requires simple information such as languages of the input, voice-gender(male/female)to be used to create the continuous speech. In speech recognition sounds uttered by a speaker are converted to equivalent waveform. Data is present in text as well as in audio and video format. Before the arrival of Discrete Selection languages and broad bandwidth connections, people used to limit themselves to text and image data. Now, with broader bandwidths available, there has been an increase in audio and video content on the Internet. Online video and audio sites such as YouTube, Google video, etc are among the most visited sites on the Internet. There has been a huge increase in the amount of data generated and stored as computers and Internet increasingly becoming part of our everyday life. But Automatic recognition of affect in speech is no longer a luxury for systems that interact with people using spoken language: It is a necessary ingredient for showing good service and respect to users, understanding them, adapting to them, and responding appropriately to how they speak. 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 Telugu language.

This actively involved in developing TTS and ASR for Indian languages, and is currently developing unit selection voice for Telugu. 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 were extracted for comparison purpose: [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) Entropy of Indian languages [e) Word length distribution graphs and their analysis [f) Sentence length distribution graphs and their analysis. Also comparison is done both among the languages and within the same language.

The authors have developed Telugu speech recognition system.. Hidden Markov Model (HMM) is used to train and recognize the speech that uses MFCC to extract the features from the male speakers. To accomplish this, Hidden Markov Model toolkit (HTK) (Young etal. 2009; Hidden Markov Model Toolkit, 2011) 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. HTK recognition tools are used to transcribe unknown utterances and to evaluate system performance.

State likelihood evaluation in Hidden Markov model (HMM) using mixture of Gaussians is one problem that needs to be solved. A novel method using Gaussian Mixture Model (GMM) for statistical pattern classification- tion is suggested to reduce computational load. Development of speech interfaces in Hindi for IT based services is a work in progress. Efforts to compensate for different accents in Hindi. Apart from Hindi ASR, speech recognition systems are being developed in other languages like Arabic, Malaya-lam, Tamil, Bengali, Telugu, etc

Some of the major application areas of Automatic speech recognition 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 all the difficulties and its wide applications into consideration the paper aims 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 telugu language.

II. Hidden Markov Model based synthesis

HTS. It was initially implemented for Japanese language but, today, can be implemented for various languages viz. Hindi, English, Tamil, Telugu etc. It is used easily for implementing prosody and various voice characteristics on the basis of probabilities without having large databases. In this approach speech utterances are used to extract spectral (Mel-Cepstral Coeff.), excitation parameters

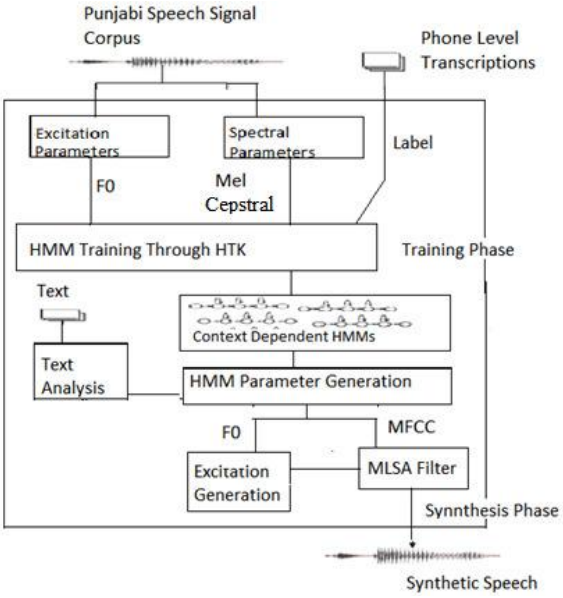


Fig1.Hidden Markov Model

In 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, in turn, to generate speech signal. Mel Cepstral Coefficients and excitation parameters i.e. fundamental frequency *F*0 are extracted from the speech database and concatenated further to use them for Hidden Markov Models training acoustic models.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

III. The Acoustic Model

In a statistical framework for speech recognition, the problem is to find the most likely word sequence, which can be described by the equation

$$\hat{W}=\arg_w \max P(W/X)$$

Applying the Bayes equation, we get

$$\hat{W}= \arg_w \max P(W/X)p(w)$$

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.

The speech recognition system implemented here uses Hidden Markov Models (HMM) for representing speech sounds. A HMM is a stochastic model. 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.

IV. Acoustic Model Generation

The speech recognition system implemented here employs Hidden Markov Model (HMM) for representing speech sounds. A 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.

The MFCC feature vectors extracted from speech signals and their associated transcriptions are used to estimate the parameters of HMMs. This process is called ASR system training. HMM Tool Kit, HTK-3.4 was used for training models over 29 context-dependent Telugu phonemes used in the chosen application. The basic acoustic units are context dependent phonemes, that is, tri-phones modeled by left-to-right, 5-state, HMMs.

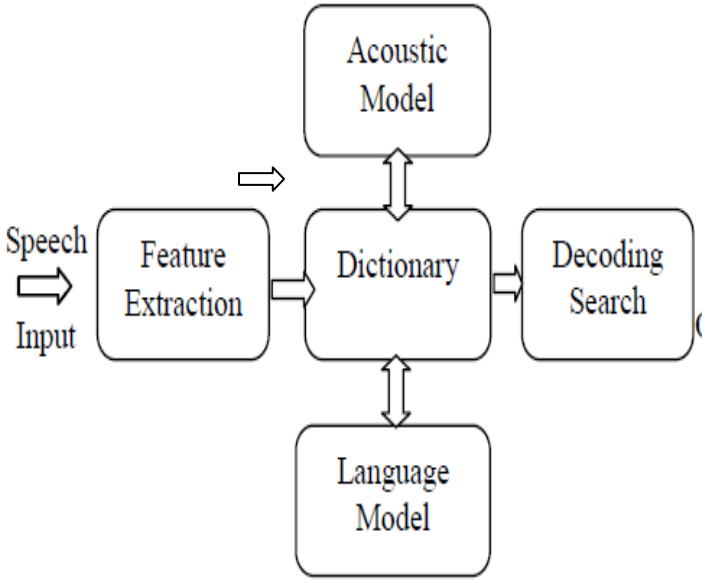


Fig2.Acoustic Model

In order to recognize speech, the system usually consists of two phases. They are called pre-processing and post-processing. Pre-processing involves feature extraction and the post-processing stage comprises of building a speech recognition engine. Speech recognition engine usually consists of knowledge about building an acoustic model, dictionary and grammar. Once all these details are given correctly, the recognition engine identifies the most likely match for the given input, and it returns the recognized word.

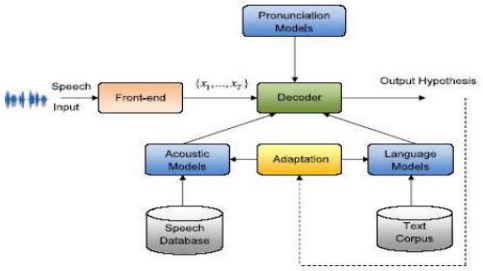


Fig3.Acoustic Model flow

V. Feature Extraction

Feature extraction involves convertion of speech waveform to some type of parametric representation. Most useful parameters in speech processing are found in the frequency domain. Since this project involves speech recognition, Mel scale cepstral analysis (MEL) is used to generate the mfcc co efficient which characterizes various speech sounds. Mel frequency cepstral coefficients result from the Discrete Cosine Transform of the filter bank Spectrum (in dB).The MFCC is the most 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. Technique of computing MFCC is based on the short-term analysis, and thus from each frame a MFCC vector is computed. In order 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 will be used to generate the Mel filter bank. According to 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. After warping the numbers of coefficients are obtained. Finally the Inverse Discrete Fourier Transformer is used for the cepstral coefficients calculation; It transforms the log of the quefrench domain coefficients to the frequency domain where N is the length of the DFT. MFCC can be computed by using the formula

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

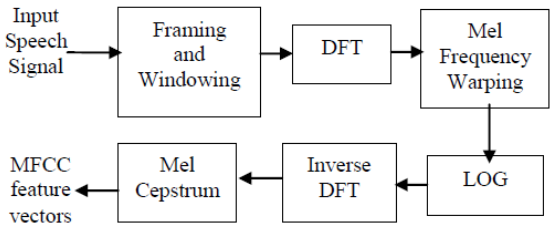


Fig4 .Mel Frequency Cepstral Coefficients Blocks

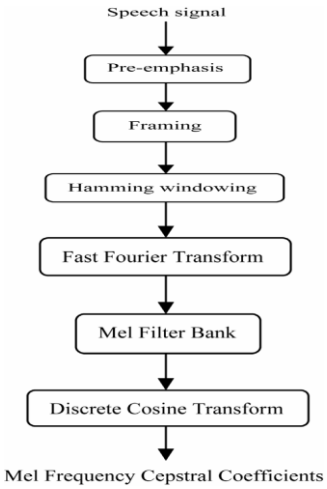


Fig4.Mel Frequency Cepstral Coefficients flow

Pre-emphasis

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 usually between *0.9* and*1.0.* In our research we used *a = 0.98*. The goal of pre-emphasis is to compensate the high frequency part that was suppressed during the sound production mechanism of humans. The speech after pre-emphasis sounds became sharper with a smaller volume.

Frame blocking

The input speech signal is segmented into frames of 20~30 ms with optional overlap of 1/3~1/2 of the frame size. Usually the frame size (in terms of sample points) is equal power of two in order to facilitate the use of FFT. If this is not the case, we need to do zero padding to the nearest length of power of two.

Fast Fourier Transform (FFT)

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

$$Xk=\sum_{n=0}^{N-1} xn\ e^{-i2\pi kn/N} \quad k=0.....N-1$$

Evaluating this definition directly requires *O*(*N*2) operations: there are *N* outputs *Xk*, and each output requires a sum of *N* terms. An FFT is any method to compute the same results in O(*N* log *N*) operations. In our proposed algorithm, a well-known split radix FFT (RS-FFT) algorithm was used. This is a divide and conquer algorithm that recursively breaks down a DFT of any composite size *N* = *N*1*N*2 into many smaller DFTs of sizes *N*1 and *N*2,along with O(*N*) multiplications by complex roots of unity traditionally called twiddle factors, When we perform FFT on a frame, we assume that the signal within a frame is periodic, and Continuous. If this is not the case, we can still perform FFT but the in continuity at the frame's first and last points is likely to introduce undesirable effects in the frequency response which is known as spectral leakage.

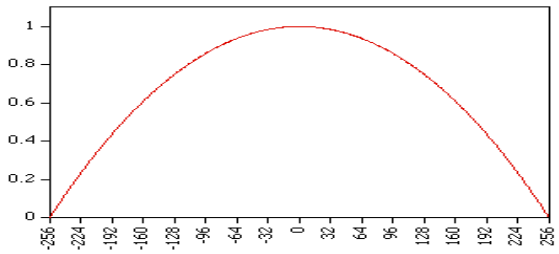
Windowing

In windowing, each data frame has to be multiplied with a window function in order 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 Hamming, Hanning,Rectangular, Bohman, Triangle, Welch, Kaiser and Blackman windows.

Welch window

$$welch(i) = 1 - \left[\frac{i - N/2}{N/2}\right]^2$$

where 0 <= i < N



Commonly used as a window for power spectral estimation. The magnitude response is

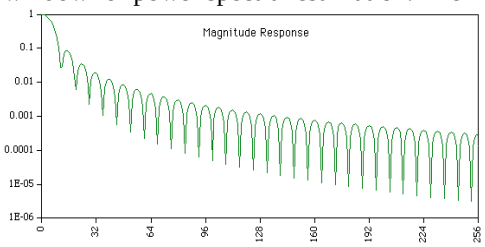


Fig5.window for power spectral estimation

The Language Model

The Hamming window which is used for the purpose is defined by the equation

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

where, *N* = number of samples in each frame.

Let *Y*(*n*) = Output signal and *X*(*n*) = input signal.

$$Y(n)=X(n)W(n)$$

the Fast Fourier transform (FFT) is used to convert each frame of *N* samples from time domain into fre- quency domain. Thus the components of the magnitude spectrum of the analyzed signal are calculated.

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

The most important step in this signal processing is Mel-frequency transformation. Compensation for non- linear perception of frequency is implemented by the bank of triangular band filters with the linear distribution of frequencies along the so called Mel-frequency range. Mel-frequency range is described by the following equation.

$$fmel=2595\log_{10}\left(1+\frac{f}{100}\right)Hz$$

Where *f* is frequency in linear range and *fme t*he corresponding frequency in nonlinear Mel-frequency range.

The *apriori* probability of a word sequence based on syntax, semantics and pragmatics of the language to be recognized. It can be realized by the Language Model which contains a list of words and their probability of occurrence in a given sequence, and is independent of the acoustic signal. The probability of a word sequence is given below.

p(w1,w2,w3,w4,……..wn)=p(w)

By Chain rule the probability of nth word is

$$p\left(w_1^n\right) = p\left(w_1\right) p\left(w_2 \middle| w_1\right) p\left(w_2 \middle| w_1^2\right) \cdots p\left(w_n \middle| w_1^{n-1}\right)$$

$$p\left(w_1^n\right) = \prod_{k=1}^{n} p\left(w_k \middle| w_1^{k-1}\right)$$

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

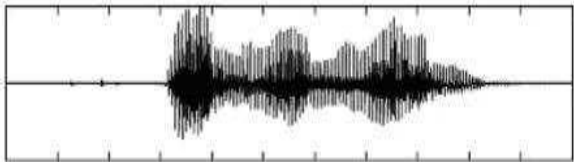


Fig 6.speech input information

Speech can be described as an act of producing voice through the use of the vocal cords.

A speech signal can be understood as a sequence of phonemes or observations.

Most useful parameters in speech processing are found in the frequency domain.

Various defining features of speech can be captured by ceptral analysis

_ MFCCs: They contain Speech-related information.
_ LPCCs: They contain Speaker-related information.

the number of feature vectors that constitute the observation sequence is of variable size These kinds of sequences are best modeled by Hidden Markov Models.

VI.**Results**

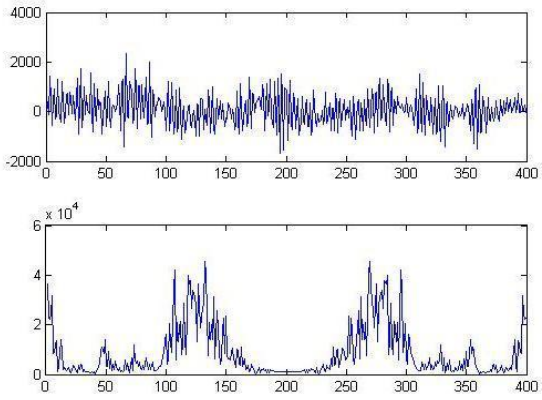


Fig7.Mel frequency cepstral coefficients speech system

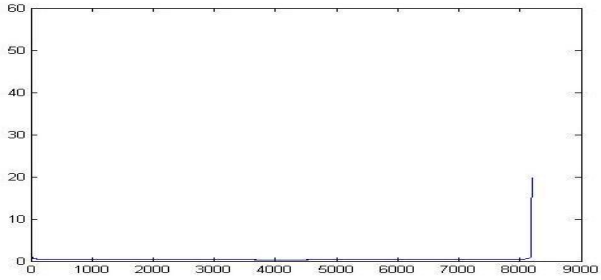


Fig8.Noise free speech frequency system

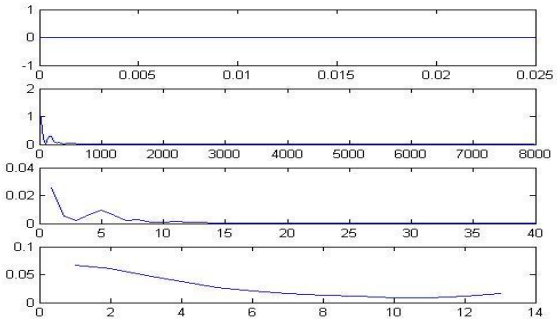


Fig9.speech patterns

VII. CONCLUSION

In conclusion, an efficient, abstract and fast ASR system for regional languages like TELUGU 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 emotional based continuous 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] 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,
2010.

[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..

[12] 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

Author' s Profile

[1]**Mr P. Vijay Bhaskar**, Head of the Department-ECE AVNIET –HYDERABAD obtained his B. Tech., from JNTU, Kakinada and M. Tech., from JNTU, Hyderabad, and both First Class with Distinction and pursuing PhD. from JNTU, Hyderabad, and also having the overall teaching experience of 18 Years and Published 08 papers in various National and International conferences, and guided good number of B. Tech., and M. Tech., Projects

[2] **Prof. Dr. S. Rama Mohana Rao**, who served for 25 years in ISRO Vikram Sarabhai Space Centre Trivandrum from 1971-1996 in various capabilities such as Head, ELP, PCF, EFF and the last being as Deputy Project Director ESP. He has also served for 15 years in various educational institutions as Professor, HOD & Principal. who is an eminent personality known for his versatility, and obtained his Ph D from IISC, Bangalore, and also his credit publications in International & National journals. He has over 15 years of experience as Principal at reputed Engineering colleges.

[3]**Mr A.GOPI** Assistant Professor in Department of-ECE AVNIET –HYDERABAD. having the overall teaching experience of 2 Years and also 2.5 years of experience as project co-ordinator in prowagon technologiy services and Published 05 papers in various National and International conferences, and guided good number of B. Tech., and M. Tech., Projects

