ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 12, December 2014

An open source approach for continuous speech recognition using hidden Markov models

Gurusiddappa Hugar[1], Vishwanath Hiregoudar[2]

Assistant Professor, Computer Science, KLEIT, Hubli, India [1]

Assistant Professor, Computer Science, AGMRCET, Hubli, India [2]

Abstract: Speech is a basic mode of communication between us and most natural efficient form of exchanging information. Speech Recognition is a conversion of an acoustic waveform to text. Speech can be isolated, connected and continuous type. The goal of this work is to recognize a Continuous Speech using Mel Frequency Cepstrum Coefficients (MFCC) to extract the features of Speech signal, Hidden Markov Models (HMM) for pattern recognition and Viterbi Decoder for decoding of speech signal. Continuous Speech files of the TIMIT standard database are used for the work. The recognition success rate is calculated for the entire database, separate Training and Testing files are found in the database and we also prepared a small set of database used in our work. For the complete process we used Hidden Markov Model Tool Kit (HTK) which is an Open source tool developed by Cambridge University Engineering Department (CUED), which contains a set of standard C Programs for feature extraction, model building and for decoding purposes, for the entire work Linux Operating System fedora is used, The objective of the work is to develop an open source HTK based Continuous Speech Recognition & to obtain better recognition accuracy for large vocabulary size.

Keywords: Speech recognition, feature extraction, pattern recognition, HTK, TIMIT, HMM

I. INTRODUCTION

Speech is a common mode of communication for people throughout lives. When humans speak, air passes from the lungs through the mouth and nasal cavity, and this air stream is restricted and changed depending on the position of tongue, teeth and lips. This produces contractions and expansions of the air, an acoustic wave, a sound. The sounds so forms are usually called *phonemes*. The phonemes are combined together to form words [1].

The speech recognition means transforming human speech to a text or to an order to the computer. Basically it is divided in to three types isolated word, connected word and continues type; Continuous speech recognition is all most a natural sentences or group of connected words, this recognition is difficult because they must utilize special methods to determine utterance boundaries. As vocabulary grows larger, confusability between different word sequences grows [2].

HTK is a toolkit developed by Speech Vision and Robotics Group of the Cambridge University Engineering Department (CUED) in 1989 by Steve Young. The Hidden Markov Model Toolkit (HTK) is a portable, open source toolkit for building and manipulating hidden Markov models. TIMIT is an acronym for the Texas Instruments Massachusetts Institute of Technology. TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. TIMIT Database has TIMIT Database has 2342 train wave files and 1680 test wave files.

The fig. 1 illustrates the simple steps we followed for continuous speech recognition.

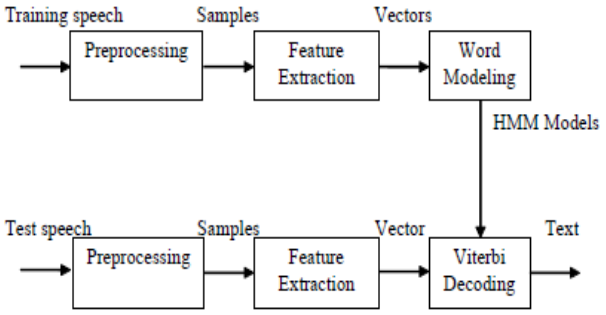

Fig1. Proposed System Block Diagram

The first stage of any recognizer development work is data preparation. MFCC Features are extracted from the training and testing speech files; HMM models are developed only for training files for each phoneme using MFCC features and the transcription (text information about content in speech file) data called word modelling. Each HMM model is represented by 3 to 5 states were in each state is represented by 8 Gaussian Mixture Model (GMM) mixtures for more accuracy they are trained n times. During the testing stage, the Viterbi search algorithm is used for the best state sequence to match the given observation sequence of the test data and represents the text of a speech file on the command prompt. The overall recognition performance is calculated based on word substitution, deletion and insertion errors found during recognition. Number of error counts will be displayed upon recognition [3&4].

Below Sections describes the detailed methodology of a work includes, Feature extraction technique, i.e. MFCC, Pattern Recognition Technique i.e. Building Hidden

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 12, December 2014

Markov Models, Decoding method using Viterbi decoder, complete HTK Process, obtained results from the work, conclusion and references used for the work

II. MFCC FEATURE EXTRACTION

MFCC is based on a perceptual scaled frequency axis. The Mel-scale provides higher frequency resolution on the lower frequencies and lower frequency resolutions on higher frequencies. This scaling is based on the hearing system of the human ear. MFCC algorithm follows different techniques of signal processing used to extract speech features as shown in Fig 2. This algorithm gives a vector of n coefficients; it is a feature, feature extraction method.

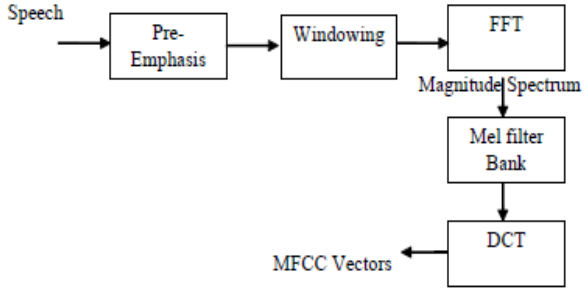

Fig2. MFCC based Feature Extraction

Mel Frequency Cepstrum analysis is done based on the human ear's critical bandwidth, filters spaced linearly at low frequency and logarithmically at high frequencies. The relationship between the linear and Mel scale is given by the equation 1,

$$F(Mel) = [2595*\log10\,[1+f]/700] \quad (1)$$

Filters are placed according to this scale to capture the important characteristics of the speech signal which represents the log spectrum. The conversion of the log spectrum into time domain is done using Discrete Cosine Transform which results in the Mel Frequency Cepstrum Coefficients. DCT is represented by the equation 2.

$$c_n = \sum_{k=1}^{K} (\log S_k)\cos[n(k-1/2)(\pi/k)],\ n=1,2,3\ldots..K \quad (2)$$

Where S_k denotes Mel spectrum coefficients, k=1, 2, 3….K.

The steps to obtain MFCC coefficients are:

Pre–emphasis: The speech signal is put through a low order digital system to spectrally flatten the signal. The widely used pre-emphasis equation is given by the equation 3

$$Z(n) = x(n) - a*x(n-1) \quad (3)$$

With a=0.97.

Framing: The pre-emphasized speech signal is blocked into frames of 25 msec, with adjacent frames being separated by 10 msec at sampling rate 16 kHz.

Hamming window: The next step in processing is to window each individual frame so as to minimize signal discontinuities at the beginning and end of each frame. Typical window used is the Hamming window, given by

$$W(n) = 0.54-0.46\cos(2\pi n/N-1)\ 0\le n\le N-1 \quad (4)$$

The result of windowing is the signal given by

$$y(n) = z(n)*w(n) \quad (5)$$

Fast Fourier Transform: Spectrum of each windowed frame is computed using FFT to obtain

$$Y(w) = FFT[h(n)*y(n)] \quad (6)$$

Mel Filter Bank: The FFT coefficients pass through a set of triangular band pass filters arranged on the Mel scale to obtain filtered spectrum. Figure 3 shows the triangular band pass filters on Mel scale.

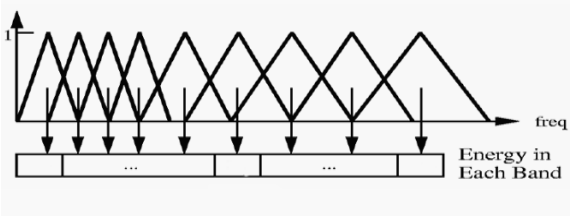

Fig.3: Mel scaled triangular filter bank

Discrete Cosine Transform: Mel Frequency Cepstrum Coefficients are computed by applying Discrete Cosine Transform on log magnitude of triangular band pass filtered output. This MFCC coefficient forms an observation vector.

Thus, after feature extraction each speech frame is represented by MFCC vector having 13 coefficients. Delta features also called the velocity features which depict the relative difference between adjacent frames and Double Delta features also called acceleration coefficients are added to make the total vector length equal to 39 which increases the recognition accuracy by extracting each feature of a speech signal thus these features are used to build efficient word model.

III. HIDDEN MARKOV WORD MODELS

The Hidden Markov Model (HMM) is a very powerful statistical tool for acoustic modelling in speech recognition. It incorporates parametric models, such as GMMs, and provides a unified pattern classification of time varying data sequences via dynamic programming. The HMM has become one of the most powerful statistical methods for modelling speech signals.

An HMM is a Markov chain where the output observation is a random variable generated according to an output probabilistic function associated with each state. Formally, an HMM is defined by:

- $A = \{a_{ij}\}$, the state transition probability matrix, where is the probability of taking a transition from state i to state j.
- $B = \{b_i(o_t)\}$, the set of state output probability distribution; where $b_i(o_t)$ is the probability of emitting o_t when state i is entered.
- $\pi = \pi_i$, the initial state distribution.

Since $\{a_{ij}\}$, $\{b_i(o_t)\}$, and π_i are all probabilities, they must satisfy the following properties:

$$a_{ij}\ge 0, b_i(o_t)\ge 0, \pi_i\ge 0 \text{ for all } i,j \quad (7)$$

$$\sum_{j=1}^{N} a_{ij}=1 \quad (8)$$

$$\sum b_i(o_t)=1 \quad (9)$$

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 12, December 2014

$$\sum_{i=1}^{N} \pi_i = 1 \quad (10)$$

Where N is the total number of states. In the discrete state observation case $b_i(o_t)$, is a discrete probability mass function (PMF). It can be extended to the continuous case with a continuous parametric probability density function (PDF). Conversely, a continuous vector variable can be mapped to a discrete set using vector quantization. A complete HMM can now be defined as:

$$\lambda = (A, B, \pi) \quad (11)$$

The first issue is how to choose the initial estimates of the HMM parameters. The re-estimation algorithm of the HMM finds a local maximum of the likelihood function. Choosing the initial parameters is important so that the local maximum will be or near the global maximum. Setting the initial estimates of the HMM means and variances to global means and variances is usually a good choice. The second issue is how to train the model parameters. The Gaussian mixture training for observation distribution usually starts with a single Gaussian model. The parameters are computed from the training data. Then the Gaussian density function is split to double the number of mixtures and parameters re-trained. After each splitting, several iterations are needed to refine the model. It is shown in practice that this procedure yields fairly good results.

IV. HTK OPERATIONS

All operations carried out using HTK Tool kit are illustrated in fig5.

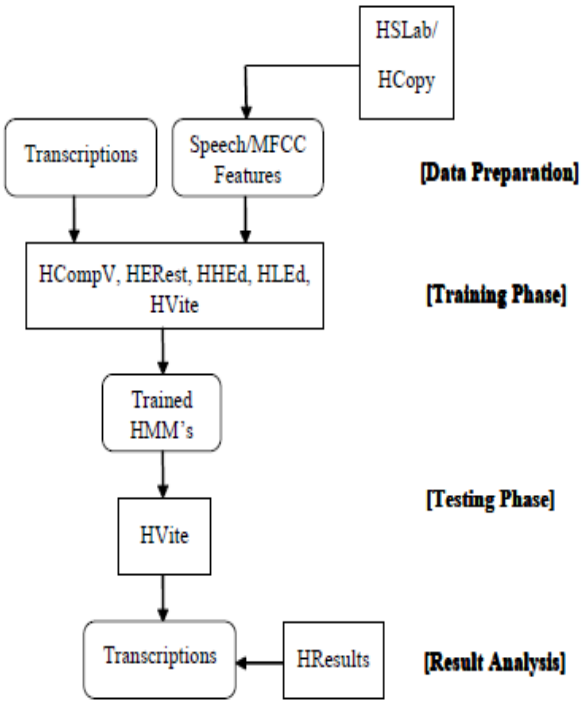

Fig5. HTK Process

The overall process is divided in to four phases: Data Preparation, Training Phase, Testing Phase and Result Analysis, Data Preparation includes speech files information and MFCC features, Training Phase includes HMM models building using features and makes a efficient word models, Testing Phase includes testing of test speech files decoding them by viterbi decoder using trained HMM's and Result Analysis includes representation of recognized text and overall accuracy for whole database which contains all test speech files.

V. RESULTS

The result part is divided in to two parts, result is categorized based on standard database and our database, for our database we just considered 10 speech samples recorded with different persons, in the result analysis Ref is the word model file contains all word details of all speech files, Rec is output file contains the obtained result from the test files, file results includes each file name with a trained words and recognized text, with its accuracy where N is the total number of words in the speech file, H is the number of words recognised among all words present test speech files, D is the number of words missed and S is the number of unnecessary words or extra words recognized due to difficulty in recognition.

For complete TIMIT Database which is complex database, obtained 95.56% of correctness and in our database, which is very small and simple, obtained 100% of correctness. The complete Result is directly copied from the command prompt as shown below.

For TIMIT Database:

HTK Results Analysis
Ref : words.mlf
Rec : recout30.mlf
----------File Results ----------
SX319.rec: 87.50(62.50) [H=7, D=0, S=1, I= 2,N=8]
Aligned transcription: ./TEST/DR1/MDAB0/SX319.lab vs ./TEST/DR1/MDAB0/SX319.rec
LAB: a big goat idly ambled through the farmyard
REC: !ENTER the big goat idly ambled through the farmyard !EXIT
SX139.rec: 100.00(75.00) [H=8, D=0, S=0,I=2, N=8]
Aligned transcription: ./TEST/DR1/MDAB0/SX139.lab vs ./TEST/DR1/MDAB0/SX139.rec
LAB: the bungalow was pleasantly situated near the shore
REC: !ENTER the bungalow was pleasantly situated near the shore !EXIT
Overall Results
SENT: %Correct=0.00 [H=0, S=1680, N=1680]
WORD: %Corr=95.56, Acc=71.06 [H=13845, D=34, S=640, I=3673, N=14519]

For Our Database:

HTK Results Analysis
Ref : words.mlf
Rec : recout.mlf
---------- File Results ----------
MA2.rec: 100.00(50.00) [H= 4, D=0, S=0, I= 2, N= 4]
Aligned transcription: test/MA2.lab vs test/MA2.rec
LAB: i am sdmcet student
REC: !ENTER i am sdmcet student !EXIT
MA2.rec: 100.00(33.33) [H=3, D=0, S=0, I=2, N=3]

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 12, December 2014

Aligned transcription: test/MA2.lab vs test/MA2.rec
LAB: pursued my m.tech
REC: !ENTER pursued my m.tech !EXIT
MA1.rec: 100.00(50.00) [H= 4, D=0,S=0,I=2, N=4]
Aligned transcription: test/MA1.lab vs test/MA1.rec
LAB: i am gurusiddappa hugar
REC: !ENTER i am gurusiddappa hugar !EXIT
Overall Results
SENT: %Correct=0.00 [H=0, S=11, N=11]
WORD:%Corr=100.00,Acc=45.45 [H=11,D=0,S=0,I=6, N=11]

VI. CONCLUSION

Continuous speech recognition is a challenging task performed using HTK tool for TIMIT database and also for our own database. HTK is an open source tool uses Linux OS, provides very good recognition accuracy for almost all words with different accents, it has built in MFCC, HMM and Viterbi programs. HMM models are used to model all phonemes, and Viterbi decoding algorithm used to recognize test sentence. The tool will not recognize the words which are not trained means if the word is not present in the dictionary with its pronunciation, each word should be trained in order to recognize sentences. HTK also provides facility for Adapting HMMs so that Recognition accuracy can be increased and make a system more robust by making speakers independent and dependent. As the future work point of view, The HTK tool can be used to recognize real time speech signal and this work can be extended to recognize regional languages.

REFERENCES

[1]. Lawrence Rabiner and Biing-Hwang Juang, "Fundamental of Speech Recognition", Prentice Hall Processing Series, 1993.
[2]. Lawrence Rabiner and Ronald Schafer, "Digital Processing of Speech Signals", Prentice Hall.
[3]. Cambridge University Engineering Department (CUED), "htk book".
[4]. "Speech Recognition using Hidden Markov Models" by Sharada C. Sajjan and Vijaya C, (WJST) World Journal of Science and Technology 2011, 1(12): 75-78, ISSN: 2231 – 2587.
[5]. "A Review on Speech Recognition Challenges and Approaches", World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 2, No. 1, 1-7, 2012.
[6]. "MFCC and its applications in speaker recognition", International Journal on Emerging Technologies, ISSN: 0975-8364, 1(1): 19-22(2010).
[7]. "HMM Clustering for Connected Word Recognition" by L.R.Rabiner, C.H.Lee, B.H.Juang, and J.G.wilpon AT&T Laboratories.
[8]. "Audio-Visual Speech Modelling for Continuous Speech Recognition", Stéphane Dupont and Juergen Luettin IEEE Transactions on MULTIMEDIA, VOL. 2, NO. 3, September 2000.
[9]. http://www.ijera.com/papers/Vol2_issue3/MH2320712087.pdf
[10]. http://www.ijarcsse.com/docs/papers/Volume_3/5_May2013/V3I4-0389.pdf
[11]. http://www.ijsr.net/archive/v2i3/IJSRON2013583.pdf
[12]. http://en.wikipedia.org/wiki?curid=34420482
[13]. http://www.ee.columbia.edu/ln/labrosa/doc/HTKBook21/node8.html
[14]. http://htk.eng.cam.ac.uk/
[15]. http://www.intechopen.com/books/speech-technologies/phoneme-recognition-on-the-timit-database
[16]. http://pubs.sciepub.com/jcn/1/2/3/index.html

BIOGRAPHIES

Mr. Gurusiddappa Hugar is an Asst.Professor in K.L.E. Institute of Technology Hubli; He received B.E. (CSE) & M.Tech (DE) from VTU Belgaum. His area of interest includes Image Processing, Speech Processing, Web Technologies, all current trends and techniques in Computer Science.

Mr. Vishwanath Hiregoudar is an Asst.Professor in A.G.M.R.C.E.T Varur, Hubli; He received B.E. (CSE) & M.Tech (DE) from VTU Belgaum. His area of interest includes Digital Logic, Microprocessors, Data Compressions, all current trends and techniques in Computer Science.