Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 6, June 2013, pg.233 – 238

RESEARCH ARTICLE

An Overview of Speech Recognition Using HMM

Ms. Rupali S Chavan[1], Dr. Ganesh. S Sable[2]
[1]Department of E&TC, Savitribai Phule Women's Engineering College, Aurangabad, Maharashtra, India
[2]Department of E&TC, Savitribai Phule Women's Engineering College, Aurangabad, Maharashtra, India

[1] chavanrupali452@gmail.com; [2] sable.eesa@gmail.com

Abstract— The Speech is most prominent & one of the natural forms of communication among of human being. The speech is a signal of infinite information. There are different aspects related to speech like speech recognition, speech verification, speech synthesis, speaker recognition, speaker identification etc. The purpose of this project is to study a speech recognition system using HMM. The goal of speech recognition is to determine which speech is present based on spoken information. The system uses MFCC for feature extraction and HMM for pattern training. The success of MFCC combined with their robust and cost-effective computation, turned them into a standard choice in speech recognition applications. And HMM provides a highly reliable way of recognizing speech.

Key Terms: - Discrete Cosine Transform; Fast Fourier Transform; Hidden Markov Model; Mel Frequency Cepstral coefficients; Speech recognition

I. Introduction

Speech recognition is a powerful tool of the information exchange using the acoustic signal. Therefore, not surprisingly, the speech signal is for several centuries the subject of research. Speech recognition is a technology that able a computer to capture the words spoken by a human with a help of microphone. These words are later on recognized by speech recognizer, and in the end, system outputs the recognized words. Speech recognition is basically the science of talking with the computer, and having it correctly recognized. Speech recognition is getting the meaning of an utterance such that one can respond properly whether or not one has correctly recognized all of the words. Data input to a machine is of generic use, but in what circumstances is speech recognition preferred ?An eyes-and-hands-busy user such as a quality control inspector, inventory taker, cartographer, radiologist (medical X-ray reader), mail sorter, or aircraft pilot-is one example. Another use is transcription in the business environment where it may be faster to remove the distraction of typing for the non-typist. The technology is also helpful to handicapped persons who might otherwise require helpers to control their environments. Automatic speech recognition has a long history of being a difficult problem-the first papers date from about 1950. During this period, a number of techniques, such as linear-time-scaled word-template matching, dynamic-time-warped word-template matching, linguistically motivated approaches (find the phonemes, assemble into words, assemble into sentences), and hidden Markov models (HMM), were used. Of all of the available techniques, HMMs are currently yielding the best performance [1].

In speech recognition, database creation (training) and recognition processes are involved. Database creation describes the collection of speaker's voice samples and extraction of features for selected words. And recognition is a process to identify the spoken word by comparing current voice features to pre stored features of voice. In real time, the recognition first it finds the likelihood of the unknown spoken word to the pre stored database of known words and then it make decision of word with the selection of maximum likelihood word. Speech recognition has two categories text dependent and text independent. Text dependent speech recognition identifies the spoken word against the words that were given to him at the time of database collection. In this case the text in recognition phase is same as in training phase. Text independent speech recognition identifies the spoken word irrespective of the words. Speech recognition is also classified as speaker dependent & speaker independent. In speaker dependent type speech of the speakers is recognized only if their speech samples are taken during training. Speaker independent speech recognition identifies the spoken word irrespective of the speakers [2].

In early research Lawrence Rabiner, Biing Hwang Juang in their book "Fundamentals of speech recognition" explained the different techniques like Hidden Markov Models, DTW, LPC, VQ, and MFCC in detail. The HMM systems generally use large acoustic models composed of several thousands of parameters. Dynamic Time Warping (DTW) and Hidden Markov Model (HMM) are two well-studied non-linear sequence alignment (or, pattern matching) algorithm. The research trend transited from DTW to HMM in approximately1988-1990, since DTW is deterministic and lack of the power to model stochastic signals[3].In another research review M.A.Anusuya, S.K.Katti reported on "Speech Recognition by Machine: A Review". They presented a brief survey on Automatic Speech Recognition. They deeply explained the Automatic Speech Recognition system classification, relevant issues of ASR design, Approaches to speech recognition [2].In another research Mahdi Shaneh and Azizollah Taheri suggested the "Voice command recognition system based on MFCC and VQ algorithms". They designed a system to recognition voice commands. They used MFCC algorithm for feature extraction and VQ (vector quantization) method for reduction of amount of data to decrease computation time. In the feature matching stage Euclidean distance was applied as similarity criterion. Because of high accuracy of used algorithms, they got the high accuracy voice command system. They trained initially with one repetition for each command and once in each in testing sessions and got 15% error rate. Secondly they increased the training samples then got zero error rates [4].

In their research H.P. Combrinck and E.C. Botha reported "On The Mel-scaled Cepstrum''. They reported on superior performance of MFCC especially under adverse conditions. Also concluded that it represents a good trade-off between computational efficiency and perceptual considerations [5].

Another research was done by Ahmad A. M. Abushariah, Teddy S. Gunawan, and Othman O. Khalifa in their paper **"**English Digits Speech Recognition System Based on Hidden Markov Models". Two modules were developed, namely the isolated words speech recognition and the continuous speech recognition. Both modules were tested in both clean and noisy environments and showed a successful recognition rates. These recognition rates are relatively successful if compared to similar systems. The recognition rates of multi-speaker mode performed better than the speaker-independent mode in both environments[6].Then Ibrahim Patel and DrY.Shrinivasa Rao in their research paper "Speech recognition using Hidden Markov Model with MFCC Subband technique" concluded that with these methods quality metrics of speech recognition with respect to computational time,learning accuracy get improved[8].In another research of voice recognition using HMM with MFCC for secure ATM by Shumaila Iqbal,Tahira Mahboob and Malik Sikandar recognition accuracy was found to be 86.67% [9].

Here this paper takes an overview of speech recognition system using MFCC and HMM. The Mel Frequency Cepstral Coefficient (MFCC) method is studied here for extracting the features of speech signal. The pre-processing and feature extraction stages of a pattern recognition system serves as an interface between the real world and a classifier operating on an idealised model of reality. Then HMM is used to train these features into the HMM parameters and used to find the log likelihood of entire speech samples. In recognition this likelihood is used to recognize the spoken word.

II. SPEECH RECOGNITION SYSTEM

Speech signal primarily conveys the words or message being spoken. Area of speech recognition is concerned with determining the underlying meaning in the utterance. Success in speech recognition depends on extracting and modelling the speech dependent characteristics which can effectively distinguish one word from another. The speech recognition system may be viewed as working in a four stages as shown in Fig. 1

i. Feature extraction
ii. Pattern training
iii. Pattern Matching.
iv. Decision logic

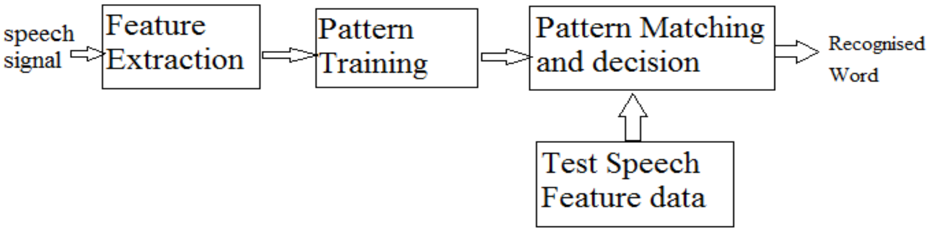

Fig. 1 Speech Recognition System

The feature extraction process is implemented using Mel Frequency Cepstral Coefficients (MFCC) in which speech features are extracted for all the speech samples. Then all these features are given to pattern trainer for training and are trained by HMM to create HMM model for each word. Then viterbi decoding will be used to select the one with maximum likelihood which is nothing but recognized word.

III. MFCC APPROACH

The purpose of this module is to convert the speech waveform to some type of parametric representation. MFCC is used to extract the unique features of speech samples. It represents the short term power spectrum of human speech. The MFCC technique makes use of two types of filters, namely, linearly spaced filters and logarithmically spaced filters. To capture the phonetically important characteristics of speech, signal is expressed in the Mel frequency scale. The Mel scale is mainly based on the study of observing the pitch or frequency perceived by the human. The scale is divided into the units mel. The Mel scale is normally a linear mapping below 1000 Hz and logarithmically spaced above 1000 Hz. Equation (1) is used to convert the normal frequency to the Mel scale the formula used is

$$\text{Mel}=2595 \log_{10}(1+f/700) \quad (1)$$

As shown in Fig 1, MFCC consists of six computational steps. Each step has its own function and mathematical approaches as discussed briefly in the following:

Step 1: Pre–emphasis

This step processes the passing of signal through a filter which emphasizes higher frequency in the band of frequencies the magnitude of some higher frequencies with respect to magnitude of other lower frequencies in order to improve the overall SNR. It increases with This process will increase the energy of signal at higher frequency. [7]

Step 2: Framing

The process of segmenting the sampled speech samples into a small frames. The speech signal is divided into frames of N samples. Adjacent frames are being separated by M (M<N). Typical values used are M = 100 and N= 256(which is equivalent to ~ 30 m sec windowing)

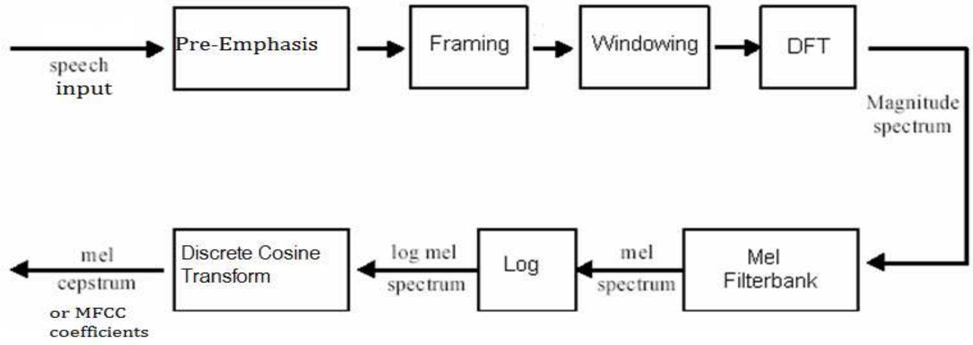

Fig.2 Computational Steps of MFCC

Step 3: Hamming windowing

Each individual frame is windowed so as to minimize the signal discontinuities at the beginning and end of each frame. Hamming window is used as window and it integrates all the closest frequency lines. The Hamming window equation is given as: If the window is defined as

W (n), $0 \leq n \leq N-1$ where

N = number of samples in each frame

Y[n] = Output signal

X (n) = input signal

W (n) = Hamming window, then the result of windowing signal is shown below:

$$Y(n) = X(n) * W(n) \quad (2)$$

$$W(n) = 0.54–0.46 \cos(2\prod n / N-1);\ 0 < n < N-1 \quad (3)$$

Step 4: Fast Fourier Transform

To convert each frame of N samples from time domain into frequency domain FFT is applied.

Step 5: Mel Filter Bank Processing

The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. The bank of filters according to Mel scale as shown in Fig 6 is then performed.

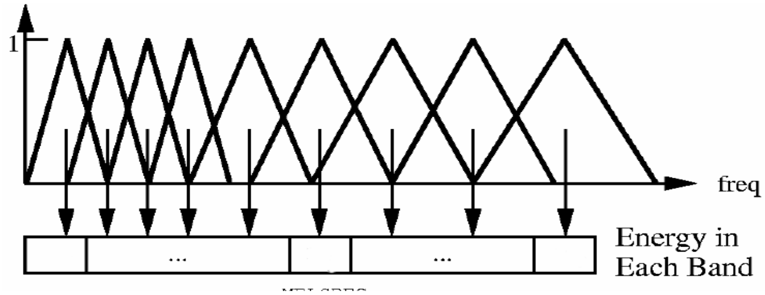

Fig. 3 Mel Filter Bank

This figure shows a set of triangular filters that are used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the centre frequency and decrease linearly to zero at centre frequency of two adjacent filters. Then, each filter output is the sum of its filtered spectral components. The output is mel spectrum consists of output powers of these filters. Then its logarithm is taken and output is log mel spectrum.

Step 6: Discrete Cosine Transform

This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT). The result of the conversion is called Mel Frequency Cepstrum Coefficients. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector.[7][8].

IV. HIDDEN MARKOV MODELLING APPROACH

A hidden Markov model (HMM) is a statistical model in which the system being modeled is assumed to be a Markov process with unknown parameters; the challenge is to determine the hidden parameters from the observable data. In a hidden Markov model, the state is not directly visible, but variables influenced by the state are visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states. A hidden Markov model can be considered a generalization of a mixture model where the hidden variables which control the mixture component to be selected for each observation, are related through a Markov process rather than independent of each other.

HMM creates stochastic models from known utterances and compares the probability that the unknown utterance was generated by each model. This uses theory from statistics in order to (sort of) arrange our feature vectors into a Markov matrix (chains) that stores probabilities of state transitions. That is, if each of our code words were to represent some state, the HMM would follow the sequence of state changes and build a model that includes the probabilities of each state progressing to another state.

HMMs are more popular because they can be trained automatically and are simple and computationally feasible to use HMM considers the speech signal as quasi- static for short durations and models these frames for recognition. It breaks the feature vector of the signal into a number of states and finds the probability of a signal to transit from one state to another. HMMs are simple networks that can generate speech (sequences of cepstral vectors) using a number of states for each model and modeling the short-term spectra associated with each state with, usually, mixtures of multivariate Gaussian distributions (the state output distributions). The parameters of the model are the state transition probabilities and the means, variances and mixture weights that characterize the state output distributions [10]. This uses theory from statistics in order to (sort of) arrange our feature vectors into a Markov matrix (chains) that stores probabilities of state transitions. That is, if each of our code words were to represent some state, the HMM would follow the sequence of state changes and build a model that includes the probabilities of each state progressing to another state.

HMM can be characterized by following when its observations are discrete:

i. N is number of states in given model, these states are hidden in model.
ii. M is the number of distinct observation symbols correspond to the physical output of the certain model.
iii. A is a state transition probability distribution defined by NxN matrix as shown in equation (4).

$$A= \{a_{ij}\}$$

$$a_{ij} = p\{ q_{t+1} = j/q_t = i\},\ 1\leq i, j\leq N_n \qquad (4)$$

$$\sum a_{ij} = 1,\ 1\leq i, j\leq N_n \qquad (5)$$

Where q_t occupies the current state. Transition probabilities should meet the stochastic limitations

B is observational symbol probability distribution matrix (3) defined by NxM matrix equation comprises

$$b_j(k) = p\{o_t=v_k|q_t=j\},\ 1<=j<=N\ ,\ 1<=k<=M \qquad (6)$$

$$\sum b_j(k) = 1,\ 1<=k<=M \qquad (7)$$

Where V_k represents the K^{th} observation symbol in the alphabet, and O_t the current parameter vector. It must follow the stochastic limitations

Π is an initial state distribution matrix (4) defined by Nx1.

$\pi= \{\pi_\iota\}$

$$\pi_i = p\{q_1 = i\}, \quad 1 \leq i \leq N \tag{8}$$

By defining the N, M, A, B, and π, HMM can give the observation sequence for entire model as λ= (A, B, π) which specify the complete parameter set of model [11].

V. Forward Backward Algorithm

The forward backward estimation algorithm is used to train its parameters and to find log likelihood of voice sample. It is used to estimate the unidentified parameters of HMM. It is used to compute the maximum likelihoods and posterior mode estimate for the parameters for HMM in training process. Here we want to find P(O|λ), given the observation sequence O = O1,O2,O3, · · · ,OT .

Forward Algorithm

The forward variable αt(i) is defined as αt(i) = P(o1,o2,… ,ot,qT = i|λ) i.e. the probability of the partial observation sequence (until time t) and state i at time t, given the model λ. αt(i) is inductively computed by following steps:

- Initialization:

$\alpha1(\iota) = \pi i\ Bi\ (o1\ ),1 \leq \iota \leq N$ (9)

Induction:

$\alpha_{t+1}\ (j)\ = \left[\Sigma\alpha_t\ (i)a_{ij}\right]b_j(o_{t+1})$ $,1 \leq t \leq T$ -1 (10)

- Termination:

$P(O|\lambda\ ) = \Sigma\alpha t\ (i)$ (11)

Finally the required P(O| λ) is sum of the terminal forward variables αT (i), this is true because

$\alpha T\ (i) = P(O1,O2, \cdot \cdot \cdot ,OT\ , qT = Si|\lambda)$ (12)

Si is the state at time t. There are N possible states Si (1 ≤ i≤ N), at time t.

Backward Algorithm

The backward variable βt(i) is defined as βt(i) = P(ot+1,ot+2,…,oT,qT = i|λ) i.e. the probability of the partial observation sequence from t+1 to the end, given the state i at time t and the model λ. βt(i) is inductively solved as follows:

- Initialization:

$\beta t\ (i) = 1,1 \leq i \leq N$ (13)

- Induction:

$\beta t\ (i)= \Sigma\ aij\ bj\ bj\ (Ot+1)\ \beta t\ +1(\ j)$ where $t = T -1, T - 2\ldots1,1 \leq i \leq N$ (14)

Combining Forward and Backward variables, we get:

$P(O|\lambda\ ) = \Sigma\alpha t\ (i)\beta t\ (i)\ ,1 \leq t \leq T$ (15)

VI. K-means Algorithm

Segmental k mean algorithm is used to generate the code book of entire features of voice sample. It is used for clustering the observations into the k partitions. K-mean algorithm is used to first partition the input vector into k initial sets by random selection or by using heuristic data. It defines two steps to precede k-mean algorithm. Each observation is assigned to the cluster with the closest mean. And then calculate the new means to be centroids of observation in each cluster by associating each observation with the closest centroids it construct the new partition , the centroids are recalculated for new cluster until it convergence or observations are no longer remains to clustering . It converges extremely fast in practice and return the best clustering found by executing several iterations. [9]

Given a set of observations (**x**1, **x**2, …, **x***n*), where each observation is a *d*-dimensional real vector, *k*-means clustering aims to partition the *n* observations into *k* sets (*k* ≤ *n*) **S** = {*S*1, *S*2, …, *Sk*} so as to minimize within-cluster sum of squares (WCSS):

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

Where $\boldsymbol{\mu}i$ *is* the mean of points in *Si*

VII. Viterbi Algorithm

Using the final re-estimated A, B and π; the value of log likelihood HMM is calculated with respect to all the word models available with the recognition engine by using Viterbi algorithm. The Viterbi algorithm takes model parameters and the observational vectors of the word as input and returns the value of matching with all particular word models. This is the likelihood values of the word (LIHMM) passed to hybrid training model [9].

It says that to find single best state sequence, Q = q1, q2.q3, · · · , qt, (which produces given observation sequence) for a given observation sequence O = o1, o2, o3,⋯,ot, we define a quantity δt(i) = $\max_{q1,q2,\cdots,qt-1} P[q1q2 \cdots qt = i, O1O2 \cdots Ot | \lambda$]

i.e., δt(i) is the best score along a single path, at time t, which account for the first t observations and ends in state Si, by induction

δt+1(j) =$\max_i \delta_t (i)aij]$bj(Ot+1)

In order to find the state sequence we need to keep track of state which maximizes the above equation. We do this via array ψt(j) for each t and stat j. Once the final state is reached corresponding state sequence can be found out using backtracking [9].

VIII. Conclusion

In this over review, we have discussed the speech recognition system using HMM. Here the techniques used in each stage of speech recognition system are discussed. Through this over review it is found that MFCC is used widely for feature extraction of speech because it is noise robust and HMM is best among all modeling techniques as it increases recognition accuracy and speed.

References

[1] SD.B.Paul,"Speech Recognition using Hidden Markov Model."

[2] M.A.Anusuya , S.K.Katti "Speech Recognition by Machine: A Review" International journal of computer science and Information Security 2009

[3] L.R.Rabiner and B.H.jaung ," Fundamentles of Speech Recognition Prentice-Hall, Englewood Cliff, New Jersy,1993.

[4] Mahdi Shaneh, and Azizollah Taheri,"Voice Command Recognition System Based on MFCC and VQ Algorithms", World Academy of Science, Engineering and Technology 57 2009

[5] H. Combrinck and E. Botha, "On the mel-scaled cepstrum," department of Electrical and

[6] Electronic Engineering, University of Pretoria.,Journal of Computer Science 3 (8): 608-616, 2007 ISSN 1549-3636.

[7] Ahmad A. M. Abushariah,Teddy S. Gunawan, Othman O. Khalifa"English Digits Speech Recognition System Based on Hidden Markov Models", International Islamic University Malaysia, International Conference on Computer and Communication Engineering (ICCCE 2010), 11-13 May 2010, Kuala Lumpur, Malaysia

[8] Anjali Bala,ABHIJEET Kumar,Nidhika Birla,"Voice command recognition System Based on MFCC and DTW",International Journal of Engineering Science and Technology,Vol.2(12),2010

[9] Ibrahim Patel,Dr.Y.Srinivasa Rao, , "Speech recognition using Hidden Markov Model With MFCC-Subband Technique." 2010 International Conference on Recent Trends in Information,Telecommunication and Computing.

[10] Shumaila Iqbal,Tahira Mehboob,Malik,"Voice Recognition using HMM with MFCC for secure ATM",IJCS Vol.8,Issue 6

[11] Nov 2011

[12] Vimala C, Dr.V.Radha, "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

[13] Lawrence R. Rabiner, Fellow, IEEE 'A Tutorial On Hidden Markov Model And Selected Applications In Speech Recognition, Proceedings Of The IEEE, Vol. 77, No. 2, February 1989.