ISSN: 2278 – 909X
International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 4, Issue 5, May 2015

Speech Recognition using Hidden Markov Model and Viterbi Algorithm

Mr. Sanjay Bhardwaj, Mr. Sunil Pathania, Mr. Rajesh Akela

Abstract-**Automatic speech recognition, an exhilarating area of research thatcan allow us to interrelate with computers using spoken commands.Speech Recognition is the process in which certain words of a particular speaker will automatically recognized, based on the information included in individual speech waves.Hidden Markov Models (HMMs) are widely used in recognition applications, most notably speech recognition. The sound can be visualized and analyzed in several ways, the recorded signal which is called as test data, is compared with the original signal which is called as the trained data. Hidden Markov Model (HMM) is the most widely used statistical speech recognition technique, as they provide a simple and operative framework for modelling time-varying speech sequences of being one word or a complete sentence. The Viterbi algorithm utilizes the trellis diagram to compute the path metrics and uses Euclidean distance metric to find the "best" partial path.**

***Index Terms*—Hidden Markov Model, Speech recognition,Viterbi algorithm**

I. INTRODUCTION

Speech/Language is one of the most important medium by which a communication can take place. With the advent and widespread use of communication tools, it has really helped in the growth of the analysis and processing of the speech signals. Nowadays, with the growth & advancement in the field of speech analysis and processing, physically challenged peoplesuch as blind and deaf can easily communicate with the machines. With the help language we express our ideas, emotions, thoughts etc. and all these are mixed and combined together which results in the formation of speech in very effective way.

The most basic method of speech recognition is to decipher the speech signal in a chronological manner based on the observed auditory features of the signal and known associations between auditory features and phonetic codes. One of the method or the approach that is used for speech recognition is "Pattern Recognition". In this method, the approach is straight forward is to train it about the speech patterns and then the recognition is carried out on the basis of the comparison of the trained patterned with the patterns to be tested. Training procedure takes out the acoustic properties and characteristics of the pattern [1]. This type of approach and the method of recognition of the speech characteristics via training is called pattern classification because the machine acquirestheauditoryassets of the speech class are dependable and repeatable across all training symbols of the pattern. The main argument is the comparison stage, which does the direct assessment of the unfamiliar speech with each possible pattern learned in the training stage and categorizes the unknown speech according to the goodness of match of the patterns.

The developments in Speech recognition technology for various languages have already taken place. Speaker recognition is the process of spontaneouslyidentifying who is speaking based on unique characteristics contained in speech waves. In this approach and method it is possible to use the speaker's speech to corroborate their individuality and control admittance to facilities such as dialing on the basis of voice, accessing the database, security control for privatedata areas, distantadmittance to computers and other IVR based services [2]. All speaker recognition systems at the highest level can be divided into two supplements namely Feature extraction and Feature matching. In the process of Feature extraction unique information and other characteristics from voice data that can be used to identify the speaker. Whereas in the process Feature matching, the actual procedures of identifying the speaker is carried out by comparing the extracted voice characteristics with a database of known speakers and based on this the outcome is decided whether the speaker has been identified or not.

There are many methods used to characterize a voice signal for speaker recognition tasks. These techniques comprise of Mel- Frequency Cepstrum Coefficients (MFCC), Linear Prediction Coding (LPC)and Auditory Spectrum-Based Speech Feature (ASSF),

the Out of all these techniques of speaker recognition the MFCC technique which an analogy of the human ear. So this type of technique is also depends on the critical bandwidth frequencies, with filters that are spaced linearly at low frequencies and logarithmic at high frequencies to seize the significant characteristics and features of speech[3]. The entire process ofMFCC can be subdivided into five sub-processes. First sub-process is "Frame Blocking" is based on the fact that speech will be stationary for a small interval of time, thus the speech waveform is more or less divided into sub-frames of approximately 20-30 milliseconds. Second is the "Windowing block" whichbrings out the smoothness in the signal by minimizes the discontinuities of the signal by tapering the beginning and end of each frame to zero. The "FFT block"translates each sub-frame from the time domain to the frequency domain. And the last "Mel-frequency wrapping block", the signal is schemed against the Mel-spectrum to simulate human hearing. It has been already studied, proved and verified that human hearing does not track the linear scale but tracks the Mel-spectrum scale, this scale has linear pacing below 100 Hz and logarithmic scaling above 100 Hz. In the final sub-process, the Mel-spectrum plot is transformed back to the time domain.

All this sub-process results in matrices which are called as the Mel-Frequency Cepstrum Coefficients [3]. These coefficient gives out comprehensive information in the spectral form, which delivers a simple and distinctiveillustration of the spectral properties of the voice signal which is the significant for demonstrating and identifying the voice individualities of the speaker. After the speech characteristics has been changed into MFCC matrices, any of the technique can be applied to profileup the speaker recognition representations using the data accomplished in the feature extraction segment and then subsequently recognize any word or sentence articulated out by an unknown speaker. The various techniques used to find out any resemblances and metamorphoses in the MFCC matrices contain, Hidden Markov Modeling (HMM). The identification of the unknown speaker is based on the lowest distance measured from the input pattern. In the HMM, the speaker acknowledgement is based on the smallest distance measured between feature matrix and unknown matrix [4].

A speaker voice patterns for the identical and undistinguishable sentencesarticulated out by the same speaker but at different times, result different categorization of MFCC matrices. Thus the main purpose of creating speaker model/process is that it can over-come these type dissimilarities and generatea result which is able to represent speaker's features. This type is called as "Stochastic modeling"that allows us to model the speaker's features by telling the speech production. This stylefor modeling the speech and the speaker is the Hidden Markov Model technique. A Hidden Markov model has finite set of states, where shift between states is characterized by transition probability matrix, presuming that the probability of being in state "Si" at time t only depends on the state occupied at time t -1. If the state probability vector is known for t = 0, the probability vector for the next observation moments can be computed [5]

In the Hidden Markov model (HMM) as the name itself signifies,the state sequence i.e. how it moves in between the states, what is the sequence of the states, cannot be observed directly as it is hidden and only the observation sequence is known. Sofor the observational sequence, Probability Density Function (PDF) describes the probability "Pi" for that observation vector. The procedure of Hidden Markov speaker model training is well-defined as the determination of the optimum model factors, identification of unidentified speaker, from the given set of training vectors from a specific speaker stored in reference data. The best approach exist is to train an HMM is the Viterbi algorithm which use the Maximum Likelihood (ML) criterion [5]. The Viterbi algorithm only reflects the most credible state sequence. The complication of a speaker model is avital factor which governs the performance of a model, where the optimumdifficulty is reliant on the total training data.

The procedure of speaker recognition using Hidden Markov Model (HMM) is as follows:

1. For each and every word articulated by the speaker be it a single word or a complete sentence, Hidden Markov model must be constructedfor the same in such a way that the characteristics of every word can be completely found out.
2. A calculation of probabilities for all promising reference models against the unknown model must be accomplishedby using the Viterbi algorithm which does so by the selection of the reference with the highest model probability value.

II. METHODOLOGY USED

We consider a speaker identification system with discrete modules for speech signal processing, training, classification, and speaker database (Fig. 1). The system functions in training mode or recognition

mode. The two different chains of arrows starting from the signal processing module describe the data flow (Fig. 1).

The system input in training mode is a collection of speech samples from "N"different speakers. A signal processing model is applied to produce a set of feature vectors for each speaker distinctly. Then a mathematical model is fitted to the feature vector set. We use the vector quantization (VQ) model to signify the statistical distribution of the features of each speaker. Each feature vector set is substituted by a codebook, which is a smaller set of code vectors with fixed size. Codebooks are deposited in the speaker database to represent the speakers. A mutual goal of the codebook design is to diminish the quantization distortion of the training data, i.e., we look for code vectors which minimize the distortion, when training vectors are substituted by their nearest neighbors in the codebook to generate the codebook. In the recognition mode, the input speech sample is processed by the same signal processing methods as in the training. The features are quantized using each codebook in the database. The speaker whose codebook gives the least distortion is acknowledged. If desired, the system lists the smallest distortions and corresponding speakers[5].

Most of today's speech recognition systems are created on the basis of Hidden Markov Models. These are arithmetical models that yieldorder of symbols or quantities. HMMs are used in speech recognition because a speech signal can be viewed as a piecewise immobile and stationary signal [7]. In a short time periods 20-30 milliseconds, speech can be expected as a stationary. Thus speech can be thought of as a Markov model for much stochastic purposes. HMMs are also used for the reason because they can be trained routinely and are simple and computationally reasonable to use. HMM speech recognition method for isolated words is shown in Fig 1 [6].This entire process speech recognition can be divided into three parts:

1) Feature extractions: Among many methods of feature extraction, most commonly used is the Mel Frequency Cepstral Coefficients (MFCC).
2) Vector quantization: In the process of the vector quantizer (VQ), it helps inidentification of the constellations in the set of auditory vectors and thus determines a symbolic vector for each constellation. The coding of this vector, and of all the vectors for other constellation gives the sequence of codebook vectors.
3) HMM probability distribution: Next process is to find out the probability for a given word.

Thus HMM is a double layer stochastic process: one Markov process modeling the temporal structure of speech and the second a set of state output processes modeling the stationary character of speech signal. It is possible to use HMM for any unit of speech. For small vocabulary recognition systems, HMMs can be used to directly model words. For large vocabularies, HMMs are defined on sub-word units like phonemes. Conversely, for a given speech utterance that represents a word, the probability that the utterance has been produced when pronouncing a certain word, can be calculated in the same way. Therefore, different hypotheses of words can be tested and the most probable can be chosen. This is the recognition method based on HMM.

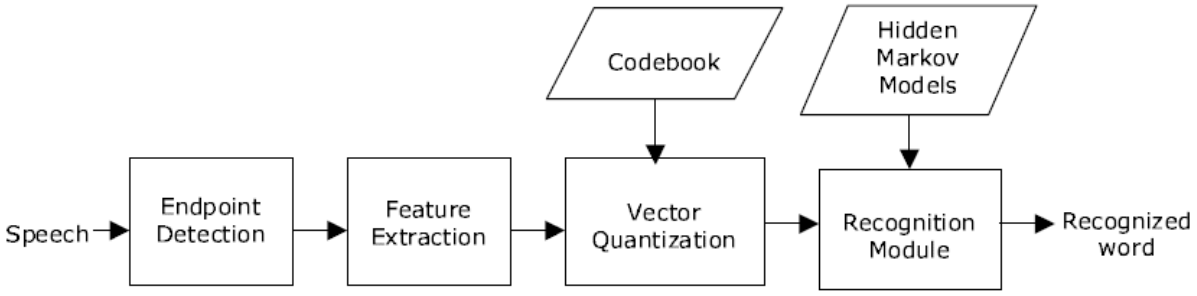

Fig. 1

III. PROGRAM STRUCTURE

Feature Extraction:

After Opening Matlab and setting the speaker recognition folder as the active directory, gives the option for training of all the "eight" pre-recorded signals as well testing, as in Fig. 2(a), Fig 2(b) and Fig 2(c)

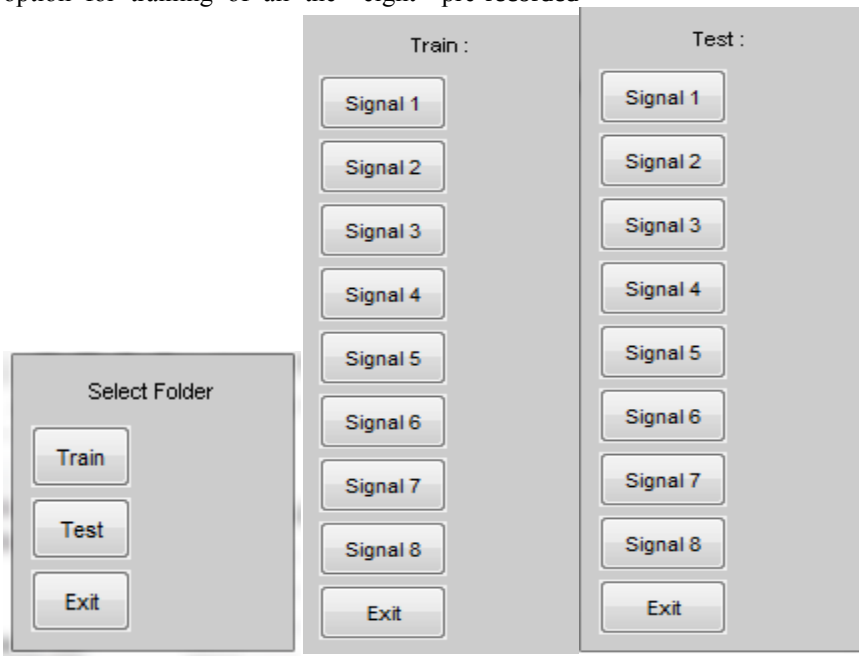

Fig. 2(a) Fig. 2(b) Fig. 2(c)

Feature Matching:

For feature matching, it calculates the distances as well probabilities using the Hidden Markov Model and Viterbi algorithm. With the help of these the words are recognized giving all the details about the "wav" files being compared along with the ID's of the speaker, as in Fig. 3(a) and Fig. 3(b)

Fig. 3(a)

```
Train\S7.wav to be compared
MFCC coefficients computation and VQ codebook training in progress
Completed.

For User #1 Dist :12.4842
For User #2 Dist :12.9247
For User #3 Dist :10.9775
For User #4 Dist :13.7924
For User #5 Dist :12.4336
For User #6 Dist :11.1458
For User #7 Dist :10.8346
For User #8 Dist :12.4062
Matching sound:
File:s7.wav
Location:Test
Recognized speaker ID:7
```

```
Location:Test
Recognized speaker ID:7

Train\S8.wav to be compared
MFCC coefficients computation and VQ codebook training in progress...
Completed.

For User #1 Dist :13.0612
For User #2 Dist :11.1952
For User #3 Dist :11.3335
For User #4 Dist :11.513
For User #5 Dist :12.0602
For User #6 Dist :9.9973
For User #7 Dist :12.125
For User #8 Dist :9.9885
Matching sound:
File:s8.wav
Location:Test
Recognized speaker ID:8
```

Fig. 3(b)

IV. CONCLUSION

In this paper, we have discussed the speech recognition problem in using HMM and Viterbi framework. HMMs have the ability to model temporal variation in speech signal, Viterbi algorithm is able to derive the posterior probability without making any assumption. It was also observed that better the training, better recognition rate.

V. REFERENCES

[1] "Speaker Recognition Using MFCC and Vector Quantization Model".

[2] E. Darren Ellis Department of Computer and Electrical Engineering – University of Tennessee, Knoxville Tennessee 37996 topic on "Design of a Speaker Recognition Code using MATLAB"

[3] Topic on "Noise estimation Algorithms for Speech Enhancement in highly non-stationary environments"ijcsi.org/papers/IJCSI-8-2-39-44.pdf

[4] Topic on "Extraction of Pitch and Formants and its Analysis to identify 3 different emotional states of a person" ijcsi.org/papers/IJCSI-9-4-1-296-299.pdf

[5] Design of Matlab®-Based Automatic Speaker Recognition Systems,Jamel Price and Ali Eydgahi

[6] Speech Recognition Using Hidden Markov Model withNeural Network Probability Estimators.Dharmendra P Kanejiya, IIT Delhi

[7]www.ijarcsse.com/docs/papers/10_October2012/Volume_2_issue_10_October2012/V2I10-0162.pdf.

[8] "ECE341 So You Want to Try and do Speech Recognition." A Simple Speech Recognition Algorithm. 15 April 2003. 1 July 2005.

[9] "Isolated Word, Speech Recognition using Dynamic Time Warping." Dynamic Time Warping. 14 June 2005.

[10] J. Bilmes, "A Gentle Tutorial on the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden MarkovModels. Technical Report, University of Berkeley, ICSI-TR- 97-021. April 1998.

[11] Koolwaaji, Johan, "iSpeak- Consultancy in Speech Technology," Fundamentals of HMM Based Speaker Verification, May 2001

[12] "Speech Recognition by Dynamic Time Warping." Speech Recognition by Dynamic Time Warping. 20 April 1998. 06 July 2005.

Mr. Sanjay Bhardwaj

Sanjay Bhardwaj is pursuing Master's of Technologyin electronics and Communication at Shoolini University, Solan.He is doing his M.Tech thesis in the field of Speech Processing.

Mr. Sunil Pathania

Mr. Sunil Pathania, working as Assistant Professor in School of Electrical & Computer Engineering at Shoolini University, Solan H.P. I have total teaching experience of 3.5 years. Major work area is Embedded System, Neural Network, Control System and Image Processing (MATLAB). I am an engineering graduate from IEET Baddi. I have done Master degree from IIT Roorkee. My native state is Himachal

Mr. Rajesh Akela

Rajesh Kumar Akela received the B.Tech degree in Electronics Engineering from Kurukshetra University and M.Tech degree in Control Instrumentation from Institute of Instrumentation Engineering, department Kurukshetra University Haryana India. Currently he is Assistance Professor in Electronic and Communication Engineering Department at School of Engg. & Tech. Shoolini University Solan HimachalPradesh India. His research interest is in Electronic circuit design, Programmable Logic Controller, Digital electronics, Control Engineering, Microprocessor and Microcontroller.