ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

The Learning Method of Speech Recognition Based on HMM

Vidwath R Hebse [1], Anitha G [2]

Student, Dept. of CSE, UBDTCE, VTU, India

Associate Professor, Dept. of CSE, UBDTCE, VTU, India

ABSTRACT: This paper helps to improve the student's interest towards speech recognition. It implements Automatic Speech Recognition System based on the following: Preprocessing, Feature Extraction Technique (MFCC- Mel Frequency Cepstrum Coefficients) and Hidden Markov Model (used in recognition phase) .Its purpose is that the human voice can be converted to a computer-readable input, Such as buttons, binary-coded, or sequence of characters. Acoustic phonetic approach, Knowledge based approach and Pattern recognition approach are the three approaches used in speech recognition. In this paper Pattern recognition technique is used. This approach requires no explicit knowledge of speech. This approach has two steps – namely, training of speech patterns based on some generic spectral parameter set and recognition of patterns via pattern comparison.

KEYWORDS: Automatic Speech Recognition (ASR), HMM model, MFCC-Mel Frequency Cepstrum Coefficients

I. INTRODUCTION

Communicating with a machine in a natural mode such as speech brings out not only several technological challenges, but also limitations in our understanding of how people communicate so effortlessly. The key is to understand the distinction between speech processing (as is done in human communication) and speech signal processing (as is done in a machine). When people listen to speech, they apply their accumulated knowledge of speech in relation to a language to capture the message. The language is unique to humans; it is the most basic form of human transmission of information. As the development of computer technology, communication between people and computers become more extensive and in-depth, so that the computer can understand human language, is not only a cherished ideal of mankind since the birth of the computer, but also an important research direction of multiple disciplines.

Real time continuous speech recognition is a computationally demanding task, and one which tends to benefit from increasing the available computing resources. A typical speech recognition system starts with a preprocessing stage, which takes a speech waveform as its input, and extracts from it feature vectors or observations which represent the information required to perform recognition. This stage is efficiently performed by software. The second stage is recognition, or decoding, which is performed using a set of phoneme-level statistical models called hidden Markov models (HMMs). Word-level acoustic models are formed by concatenating phone-level models according to a pronunciation dictionary. These word models are then combined with a language model, which constrains the recognizer to recognize only valid word sequences.

II. LITERATURE SURVEY

In [1] authors proposed Inter-word Co articulation modeling and MMIE training for improved connected digit recognition The authors describe developments by the speech research group at CRIM (Centre de Recherche Informatique de Montreal), in the field of speaker-independent connected digit recognition, using hidden Markov models (HMMs) trained with maximum mutual information estimation (MMIE). The experiments described were all performed on the complete adult portion of the TIDIGITS corpus. Techniques that made it possible to improve greatly the recognition rate are described. New results include a 0.28% word error rate and a 0.84% string error rate with two models per digit (one for male and one for female speakers) using context-dependent discrete HMMs.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

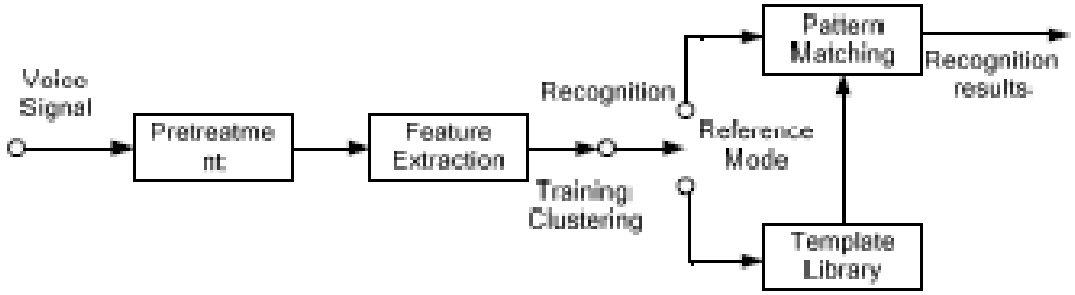
In [2] authors present a research method to directly recognize greeting voice without segmentation to avoid error recognition because of error segmentation. The basic principle of biometric pattern recognition is applied to speaker-independent and continuous speech recognition of greeting. The high-dimension space covering theory is applied to the learning process of speaker-independent and continuous speech recognition of greeting during the construction of voice pattern. And the dynamic search method of high-dimensional space covering is implemented on the dynamic search of continuous speech of greeting to recognize it. Some satisfactory recognition results are obtained by experiment.

In [3] authors proposed Robust Features for Noisy Speech Recognition using MFCC Computation from Magnitude Spectrum of Higher Order Autocorrelation Coefficients. Noise robustness is one of the most challenging problems in automatic speech recognition. The goal of robust feature extraction is to improve the performance of speech recognition in adverse conditions. The mel-scaled frequency cepstral coefficients (MFCCs) derived from Fourier transform and filter bank analysis are perhaps the most widely used front-ends in state-of-the-art speech recognition systems. One of the major issues with the MFCCs is that they are very sensitive to additive noise. To improve the robustness of speech front-ends we introduce, in this paper, a new set of MFCC vector which is estimated through three steps. First, the relative higher order autocorrelation coefficients are extracted. Then magnitude spectrum of the resultant speech signal is estimated through the fast Fourier transform (FFT) and it is differentiated with respect to frequency. Finally, the differentiated magnitude spectrum is transformed into MFCC-like coefficients. These are called MFCCs extracted from Differentiated Relative Higher Order Autocorrelation Sequence Spectrum (DRHOASS). Speech recognition experiments for various tasks indicate that the new feature vector is more robust than traditional mel-scaled frequency cepstral coefficients (MFCCs) in additive noise conditions.

III. HMM-BASED SPEECH RECOGNITION

A. *Principle and System for Speech Recognition (AUTOMATIC SPEECH RECOGNITION)*

Voice recognition technology, also known as automatic speech recognition (ASR), its purpose is that the human voice can be converted to a computer-readable input, Such as buttons, binary-coded, or sequence of characters. Speech recognition, is different from speaker recognition, Speaker Recognition is not to identify words which is in the vocabulary content.[1] Speech recognition systems usually assumed that the voice signal is encoded into one or a plurality of symbol sequence of the information entity. In order to achieve the reverse operation, namely to identify the sequence of symbols of a given speaker's voice, the first continuous speech waveform is converted into a long discrete parameter vector sequence. Assuming that the sequence of this parameter vector is a speech waveform accurately represented in a vector corresponding to the period of time (typically, 10ms and so on), the voice signal can be regarded as smooth. This characteristic can be used to good description of the HMM. The concept of the Markov model is a discrete-time domain finite state automata, hidden Markov model HMM refers to the internal state of this Markov model is not visible to the outside world, the outside world can only see the output value of each moment. The acoustic characteristic of the speech recognition system, the output value is usually calculated from the respective frames. HMM portrayed speech signal the need to make two assumptions, one internal state of the transfer is only related to a previous state, and the other is that the output value is only relevant to the current state (or the current state of the transfer), these two assumptions greatly reduced the model complexity.



The basic configuration of the speech recognition system

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

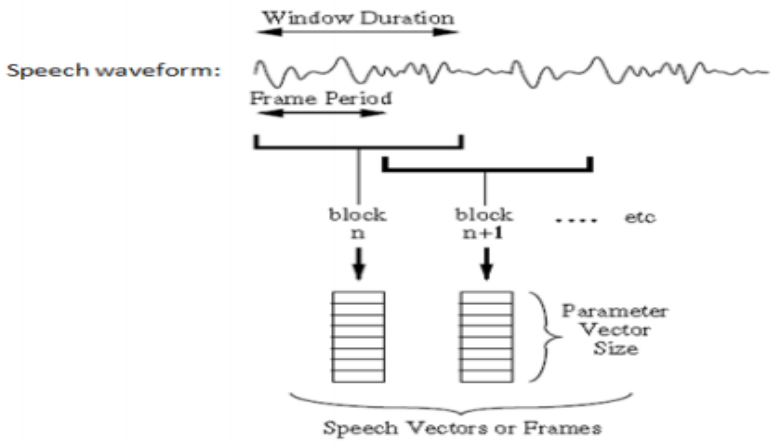
A.1 Pre-emphasis:

In order to flatten speech spectrum, a pre-emphasis filter is used before spectral analysis. Its aim is to compensate the high-frequency part of the speech signal that was suppressed during the human sound production mechanism.

A.2 Frame Blocking and Windowing:

The speech signal is divided into a sequence of frames where each frame can be analyzed independently and represented by a single feature vector. Since each frame is supposed to have stationary behavior, a compromise, in order to make the frame blocking, is to use a 20-25 ms window applied at 10 ms intervals (frame rate of 100 frames/s and overlap between adjacent windows of about 50%), as Holmes & Holmes exposed in 2001. In order to reduce the discontinuities of the speech signal at the edges of each frame, a tapered window is applied to each one. The most common used window is Hamming window[2].

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi(n-1)}{N-1}\right)$$



Speech Endpoint Detection

In the process of speech recognition, When the system receives a signal containing voice, system will detect and locate speech endpoint, removal of excess noise before and after the speech, Complete voice will be submitted to the next level recognition. Voice endpoint detection algorithm is mainly based on the energy of the voice,zero crossing rate, LPC coefficients, information entropy, cepstral, band variance and so on. The endpoint detection effects and the actual environmental noise has a great relationship, Therefore, the endpoint detection of pre-denoising can improve the recognition rate. We introduce a traditional detection methods based on short-term energy, short-time zero-crossing rate.

Characteristic Parameter Extraction

Sub-frame and endpoint detection is complete, the next feature extraction parameters. Feature extraction amount is the effective characteristics of the signal is extracted from the speech signal, but also try to remove the noise information of the speech signal, to improve the accuracy of identification. Since voice having a short-time characteristic, the speech characteristic parameters by frame information extraction, frame feature vector. A voice after feature extraction, into a vector sequence. This vector sequence to train for and then some kind of model for speech recognition voice template. Voice characteristic parameter of extraction is very important, directly affects the accuracy of the speech recognition.
A good speech features to meet the requirements of three:

(1)Can effectively extract the signal characteristics of the speech, including the channel characteristics of the human auditory model;

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

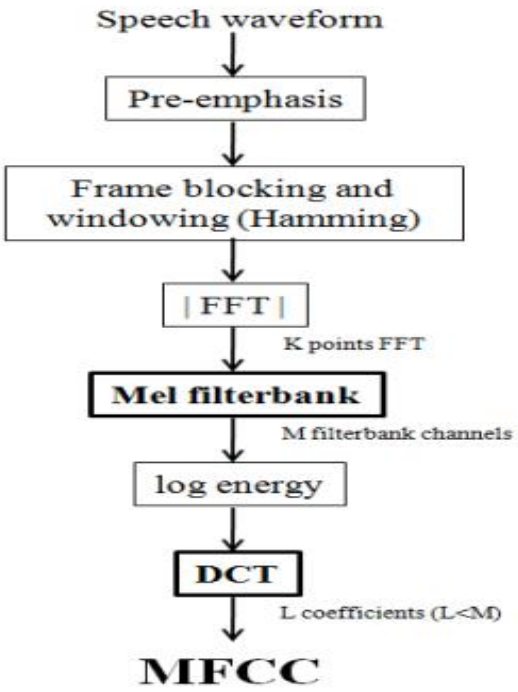
(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

(2)Good independence between the order parameter;
(3)The characteristic parameters have an efficient method of calculating.

A.3 Mel-Cepstrum

Davis & Mermelstein (1980) pointed the Mel Frequency Cepstrum6 Coefficients (MFCC) representation as a beneficial approach for speech recognition (Huang et al., 2001).[3] The MFCC is a representation of the speech signal defined as the real cepstrum of a windowed short-time signal derived from the FFT of that signal (Huang et al, 2001) which, is first subjected to a log-based transform of the frequency axis (mel-frequency scale), and then decorrelated using a modified Discrete Cosine Transform (DCT-II). Figure illustrates the complete process to extract the MFFC vectors from the speech signal. It is to be emphasized that the process of MFCC extraction is applied over each frame of speech signal independently.



MFCC extraction process

After the pre-emphasis and the frame blocking and windowing stage, the MFCC vectors will be obtained from each speech frame. The process of MFFC extraction will be described below considering in any instant that all the stages are being applied over speech frames.
The first step of MFCC extraction process is to compute the Fast Fourier Transform (FFT) of each frame and obtain its magnitude. The FFT is a computationally efficient algorithm of the Discrete Fourier Transform (DFT). If the length of the FFT, is a power of two ($K=2^n$), a faster algorithm can be used, so a zero-padding to the nearest power of two within speech frame length is performed.
The next step will be to adapt the frequency resolution to a perceptual frequency scale which satisfies the properties of the human ears (Molau et al., 2001), such as a perceptually mel-frequency scale. This issue corresponds to Mel filterbank stage.
The last step involved in the extraction process of MFCC is to apply the modified DCT to the log-spectral-energy vector, obtained as input of mel filterbank, resulting in the desired set of coefficients called Mel Frequency Cepstral Coefficients.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

B. HIDDEN MARKOV MODELS

B.1 Introduction

The Hidden Markov model (HMM) is a very powerful mathematical tool for modeling time series. It provides efficient algorithms for state and parameter estimation, and it automatically performs dynamic time warping for signals that are locally squashed and stretched. It can be used for many purposes other than acoustic modeling.

B.2 Markov Chains

Hidden Markov models are based on the well-known Markov chains from probability theory that can be used to model a sequence of events in time. Figure below shows such a graphical network representation of such a model, it has two states a and band some connections indicated by arrows that show how one can get from one state to another

The topology of the network shows an important property of Markov chains, namely that the next state only depends on the current state the model is in, regardless of how it got in the current state; this property is often referred to as the Markov property. By starting in one of the two states and at each time step moving through the model following the arrows out of the current state to the other state or once again to the same state, sequences of a's and b's can be generated[4].

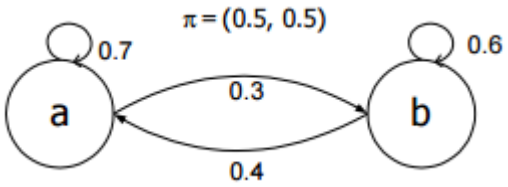


Fig 1: A Markov Chain

The arrows leaving a state are annotated with a probability that indicates how likely it is that this particular transition out of the state will be chosen. As a transition has to be made the probabilities associated with all arrows leaving a state should sum to one. The distribution π indicates how likely each state is to be the start state, in Figure 1 both states are equally likely to be the start state. Using these probabilities a Markov model can be used for recognition. Imagine that we have two processes that produce outputs that can be encoded as sequences of a's and b's and each of these processes can be modeled by a Markov model, the one from Figure 1 and the one from Figure 2.

If we now receive for example a sequence abba we can calculate for each model the probability that it generated this sequence by simply multiplying the probabilities along the path that corresponds to the sequence 2 .

Probability model 1: 0.5 •0.3 •0.6 •0.4 = 0.036

Probability model 2: 0.4 •0.5 •0.7 •0.3 = 0.042

After which we may conclude that the output is most likely to be generated by the second process.

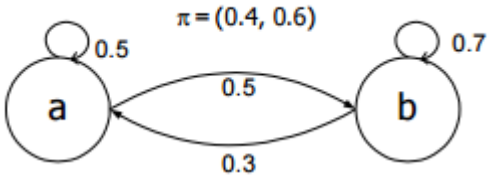


Fig 2: A second Markov model

The concept of the Markov model is a discrete-time domain finite state automata, hidden Markov model HMM refers to the internal state of this Markov model is not visible to the outside world, the outside world can only see the output value of each moment. The acoustic characteristics of the speech recognition system, the output value is usually calculated from the respective frames.HMM portrayed speech signal the need to make two assumptions, one internal state of the transfer is only related to a previous state, and the other is that the output value is only relevant to the current state (or the current state of the transfer), these two assumptions greatly reduced the modelcomplexity.HMM

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

scoring decoding algorithm and corresponding training forward algorithm, Viterbi algorithm and forward-backward algorithms.

C. PROCESSES AND IMPLEMENTATION

HMM is one of the ways to capture the structure in this sequence of symbols. In order to use HMMs in speech recognition, one should have some means to achieve the following:

- Evaluation: Given the observation sequence $O = (o_1, o_2, \ldots, o_t)$ and a HMM $\lambda = (A, B, \pi)$ to choose a corresponding state sequence $Q = q_1, q_2, \ldots, q_t$ which optimal in some meaningful sense, given the HMM.
- Training: To adjust the HMM parameters $\lambda = (A, B, \pi)$ to maximize $P(O \mid \lambda)$.

The following are some of the assumptions in the Hidden Markov Modeling for speech.

- Successive observations (frames of speech) are independent and therefore the probability of sequence of observation $P = (o_1, o_2, \ldots, o_t)$ can be written as a product of probabilities of individual observations, i.e. $O = (o_1, o_2, \ldots o_t) = \prod_i^T P(O_i)$

- Markov assumption: The probability of being in a state at time t, depends only on the state at time $t-1$.

The problems associated with HMM are explained as follows:

(a) Evaluation: Evaluation is to find probability of generation of a given observation sequence by a given model. The recognition result will be the speech unit corresponding to the model that best matches among the different competing models. Now to find $P(O \mid \lambda)$, the probability of observation sequence $O = (o_1, o_2, \ldots, o_t)$ given the model λ i.e. $P(O \mid \lambda)$.

(b) Decoding: Decoding is to find the single best state sequence, $Q = q_1, q_2, \ldots, q_t$, for the given observation sequence $O = (o_1, o_2, \ldots, o_t)$. Consider $\delta_t(i)$ defined as

$$\delta_t(i) = \max_{(q_1, q_2 \ldots q_{T-1})} P[q_1, q_2 \ldots\ldots q_t = i, o_1, o_2 \ldots\ldots o_t \mid \lambda]$$

that is $\delta_t(i)$ is the best score along single path at time t, which accounts for the t observations and ends in state i. by induction,

$$\delta_{i+1}(j) = [\max_i \delta_t(i) a_{ij}] b_j(o_{t+1})$$

(c) Training (Learning): Learning is to adjust the model parameters (A, B, π) to maximize the probability of the observation sequence given the model. It is the most difficult task of the Hidden Markov Modeling, as there is no known analytical method to solve for the parameters in a maximum likelihood model. Instead, an iterative procedure should be used. Baum-Welch algorithm is the extensively used iterative procedure for choosing the model parameters. In this method, start with some initial estimates of the model parameters and modify the model parameters to maximize the training observation sequence in an iterative manner till the model parameters reach a critical value.

(d)Identification

Use the Viterbi algorithm to dynamically find the hidden Markov model state transition sequence (ie identify the results), the time complexity is far less than the total probability formula.

The Viterbi algorithm is widely used in the dynamic programming algorithm in the field of communication, the algorithm in speech recognition applications. The total probability formula, you can calculate the output probability of the system, but were unable to find an optimum state transition path. Using the Viterbi algorithm can be found not only a good enough state transition path and the path corresponding to the output probability can also be obtained. Meanwhile, with the Viterbi algorithm to calculate the output probability of the amount of computation required is much smaller than the amount of calculation of the total probability formula.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

IV. CONCLUSION

The conclusion of this study of recognition and hidden markov model has been carried out to develop a voice based user machine interface system. In various applications we can use this user machine system and can take advantages as real interface, these application can be related with disable persons those are unable to operate computer through keyboard and mouse, these type of persons can use computer with the use of Automatic Speech Recognition system, with this system user can operate computer with their own voice commands (in case of speaker dependent and trained with its own voice samples). Second application for those computer users which are not comfortable with English language and feel good to work with their native language i.e. English, Kannada, Hindi

REFERECES

[1] R.Cardin, Y. Normadin and E. Millen, Inter-word coarticulation modeling and MMIE training digit recognition, ICASSP, p243-246-199.
[2] Hong Ye,Youzheng Zhang and Jianwei Shen Study on Speech Recognition of Greeting Based on Biometric Pattern Recognition.
[3] Amita dev and Bansal poonam , Robust Features for Noisy Speech Recognition using MFCC Computation from Magnitude Spectrum of Higher Order Autocorrelation Coefficients.
[4]Mingjia An,Zhengtao Yu, Jianyi Guo , The Teaching Experiment of Speech Recognition based on HMM.
[5] Ir. P. Wiggers, Dr. drs. L.J.M. Rothkrantz, AUTOMATIC SPEECH RECOGNITION USING HIDDEN MARKOV MODELS.
[6] Noelia Alcaraz Meseguer, Speech Analysis for Automatic Speech Recognition.
[7] Bhupinder Singh, Neha Kapur, Puneet Kaur, Speech Recognition with Hidden Markov Model: A Review.