International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2017, pp. 233-237
e-ISSN:2278-621X

SPEECH RECOGNITION WITH HIDDEN MARKOV MODEL

Pramitha K[1], Kuris Annie Anton[2] & Manimozhi R[3]

Abstract- Speech is one of the most used way for communication among humans. The interaction between a human and computer is called human interface. Speech has the potential of being an important mode of communication with computer. It's always been fascinating that how a computer may respond to the commands given through speech. Being able to train a computer to recognize human speech will help to reduce so much of time and energy to give commands. This paper gives an overview of using Hidden Markov model algorithm for speech recognition. In this paper various old speech recognition algorithms have been compared with Hidden Markov model. A comparative study of various techniques of speech recognition has been done. Hidden Markov model has been chosen over various speech recognition techniques due to the statistical methods used in it. The reasons why this has occurred are that the models are very rich in mathematical structure and hence can form the theoretical basis for use in a wide range of applications and the models, when applied properly, work very well in practice for several important applications. The paper concludes with the research and development in the speech recognition techniques with the usage of Hidden markov model.
Keywords – Hidden Markov model, Speech recognition techniques, Comparison, Automatic speech recognition (ASR).

1. INTRODUCTION

Speech recognition is also called as Automatic speech recognition or computer speech recognition. It means interpreting voice of the computer and performing given task or the ability to match a voice against a available or given vocabulary. The actual task is to make the computer to understand spoken language. By "understand" we intend to respond fittingly and change over the info discourse into another medium e.g. text. Speech recognition is therefore sometimes called as speech-to-text (STT). A speech recognition framework comprises of a mouthpiece, for the individual to talk into and speech recognition programming; a computer to take and interpret the speech; a good quality soundcard for input and/or output; a proper and good pronunciation.

If we had a computer system which can do half as decent a job of recognizing human speech as other human beings can, and do it economically, speech will eventually replace cards, paper tape and even keyboards for communication with computers.

The technique of automatic speech recognition has improved remarkably in the past decade. With the development in accuracy and scope, there has come, for the time being, a strong juncture on a class of statistical methods based on a structure called a hidden Markov model (HMM).

2. ADVANCES IN SPEECH RECOGNITION

A short introduction to different approaches to the speech recognition is given in this section. Acoustic phonetic based speech recognition (1920-1960s)

Machine recognition came into existence in 1920. The earliest attempt to devise an system based on acoustic-phonetic for speech recognition was created in 1950. At Bell laboratories, Davis et al. (1952) built a system for recognition of isolated digit for a single speaker. The developed system was based on measuring the spectral resonances at the vowel part of each digit. Olson and Belar (1956), at RCA laboratories tried to indentify 10 distinct syllables of a single talker included in 10 monosyllabic words. At MIT Lincoln laboratories, Forgie and Forgie (1959) built a vowel recognizer which can identify 10 vowels embedded in a /b/-vowel-/t/ format in a speaker independent manner. The heading for sub subsections should be in Times New Roman 11-point italic with initial letters capitalized and 6-points of white space above the sub subsection head.

An elaborate filter bank spectrum analyzer was used along with the logic that connects the outputs of each channel of the spectrum analyzer to a vowel decision circuit and majority decisions logic scheme was used to choose the spoken vowel. Another hardware effort in Japan was the work of Sakai and Doshita (1962) of Kyoto University, who built a hardware phoneme recognizer. A digit recognizer was designed by Nagata et al. (1963) at Nippon Electric Corporation (NEC) laboratories. Reddy's research program at Carnegie Mellon UniversityVintsyuk (1968) proposed the utilization of dynamic programming methods for time adjusting a pair of speech expressions (known as Dynamic Time Warping (DTW). [13]

Pattern based speech recognition was he most focused area of research in the 1970s because of the fundamental studies done by Velichko and Zagoruyko (1970) in Russia, Sakoe and Chiba in Japan and Itakura in the United States. Itakura's investigation demonstrated how the possibility of Linear Predictive Coding (LPC), which had just been effectively utilized as a part of low bit rate speech coding, could be stretched out to Speech recognition systems using a specific distance measure in

[1] Department of Computer Application, St Aloysius College, AIMIT, Mangaluru, Karnataka, India
[2] Department of Software Technology, St Aloysius College, AIMIT, Mangaluru, Karnataka, India
[3] Asso. Professor, Department of MCA, St Aloysius College, AIMIT, Mangaluru, Karnataka, India

light of LPC spectral parameters. In 1976, `Sambur and Rabiner described a statistical decision approach which referred to recognition of connected digits for speaker dependent as well as speaker independent system.
Continuous word based speech recognition (1980-1990s) During 1980s, the main focus of research was tuned to continuous word recognition. Robust training methodology that has advantages of both averaging and clustering techniques has been presented by Rabiner and Wilpon (1980). A continuous word recognition system [14] can identify a fluently spoken string of words by matching a concatenated pattern of individual words
Hybrid statistical and connectionist (HMM/ANN) based speech recognition (1990-2000s) W. Ma with his associates Compernolle and Katholieke (Ma et al., 1990) introduced a system that combines the good short-time classification properties of time delay neural networks and the good integration and overall recognition capabilities of HMMs. A novel approach for a hybrid connectionist HMM speech recognition system based on the use of a neural network as a vector quantizer has been proposed in (Rigoll, 1994). Bourlard and Morgan (1998) in their work have described the use of ANN as statistical estimator in automatic speech recognition process.
Variational Bayesian (VB) estimation based speech recognition (2000-2010) Pruthi et al. (2000) have developed a speaker-dependent, real-time, isolated word recognizer for Hindi. Gupta composed a segregated word for Hindi Language(2006). System uses continuous HMM and consists of word based acoustic. The work in (Al-Qatab and Ainon, 2010) discusses the development and implementation of an Arabic speech system. System is developed using HTK. System can recognize both continuous speech and isolated words.

3. HIDEN MARKOV MODEL

HMMs is significant for almost all of the modern speech recognition systems and even though the basic framework hasn' t changed very much in the last couple of years or more, the detailed modelling techniques developed within this framework have evolved to a state of considerable sophistication. The result has been constant and important progress hasa been made.HMM is one of most popular models for sequential data. It is popular because it is easy enough that the prameters can be estimated and inference can be done efficiently but is also capable enough to handle real world application.

3.1 HMM Formulation

Hidden Markov model (HMM) is also known as the hidden state and is called the observation
Markov chain
$P(s_{t+1}|s_1 \ldots s_t) = P(s_{k+1}|s_t)$
That is $P(o_t|s_1 \ldots s_t) = P(o_t|s_t)$
So result is
$P(s_1 \ldots s_t, o_1 \ldots o_t) = P(s_1)P(o_1 \mid s_1) \prod_{i=2}^{t} [P(s_i|s_{i-1})P(o_i \mid s_i)]$

3.2 HMM Elements

An HMM for discrete symbol observation
N : - the number of states in the model the state at time t -> S_t
M : -the number of distinct observation symbols per state
V = {v1,v2,…….vm}
a_{ij} :- the state-transition probability distribution : A
$a_{ij} = p[s_{t+i} = j|s_t = i]$ $1<=i ,j<=N$
$b_j(k)$:- the observation symbol probability distribution : B
$b_j(k) = p[o_t = v_k \mid s_t = j]$ $1<=k <=M$
π :- the initial state distribution, π
$\pi_i = p[s_1 = i]$ $1 <= I <= N$
λ Compact Notation of a HMM Model
$\lambda=(A,B, \pi)$
$p(o_1,o_2,\ldots\ldots o_T \mid \lambda, s_1,s_2.....,s_T)$
$= \pi_1 b_{q1}(o_1) a_{q1q2} b_{q2}(o_2) \ldots . a_{q(T-1)} b_{qt}(o_T)$

4. MEASUREMENTS AND MODELING OF SPEECH

Speech is a constant signal. When we talk, our articulatory apparatus (the lips, jaw, tongue, and velum) regulates the air pressure and flow to produce an audible sequence of sounds. Even if the spectral content of any sound in particular may include frequencies up to several thousand hertz, our articulatory configuration (vocal-tract shape, tongue movement, etc.) often does not withstand dramatic changes that are more than 10 times per second. Modeling of speech thus involves two aspects:
(1) Short-time spectral properties of individual sounds analysis, performed at an interval on the order of 10 milliseconds (msec), and
(2) Long-time development of sound sequences characterization, on the order of 100 msec, due to changes in articulatory configuration.

When speech signals are digitally processed, it requires discrete time sampling and waveform quantization. Typically, an analog speech signal is sampled at a rate of 8-20 kilohertz (kHz), and the amplitude of each waveform sample is usually represented by one of 2^{16} = 65,536 values that is, 16-bit quantization of the discrete time signal.

5. SPEECH RECOGNITION USING HMM'S

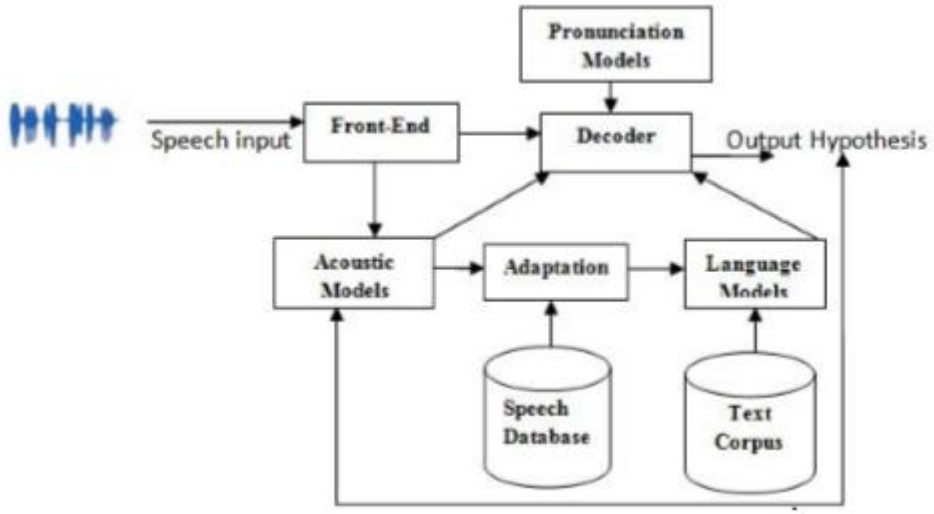

Fig. Speech recognition outline.

5.1 Feature Extraction

Recording of various speech samples of each and every word of the vocabulary is done by separate speakers. Once speech samples are collected, they are converted from analog to digital form by sampling at a frequency of 16 kHz. Recording the speech signals at a regular interval is called sampling. Quantization is done on the collected data if needed to eliminate noise in speech samples. The feature extraction is done on speech samples , and then feature training & feature testing is done. Feature extraction transforms the incoming sound into an internal representation such that it is possible to regain the original signal from it. There are various techniques to extract features like MFCC, PLP, RAST, LPC Costly used is MFCC.

5.2 Mel Frequency Cepstral Coefficients

MFCCs are used because it is created with the help of the knowledge of system of human auditory and is used in every art speech. MFCC is a standard process for extraction of the feature in speech recognition tasks. MFCC consists of certain steps applied on an input speech signal. These computational steps of MFCC consists: - Framing, Windowing, DFTH, Mel filter bank algorithm, computing the inverse of DFT.

5.3 Decoding

It is a very important step among all the steps in the speech recognition process. Decoding is performed to find the best match for the incoming feature vectors using the knowledge base. A decoder takes the actual decision about recognition of a speech pronouncement by merging and optimizing the information transmitted by the acoustic and language models.

5.4 Acoustic Modelling

There are two types of acoustic models i.e. word model and phoneme model. An acoustic model is put into use with the help of different approaches such as HMM, support vector machines (SVM), ANNs, dynamic Bayesian networks (DBN).HMM is used in some form or the other in every state of the art speech and speech recognition system.

5.5 Hidden Markov Model

Acoustic modelling can be done using hidden markov model. There are two stochastic processes which are inter-related which are same as Markov Chain except that the output symbol as well as the transitions are probabilistic. Every HMM state may have a collection of output symbols called as output probabilities and having a limited number of states Q= {q1, q2,q3 ..., qn}. One process is related to the transitions among the states, controlled by a collection of probabilities known as transition probabilities to model the temporal variance of speech. Other process is concerned with the output observations of state O= {o1, o2,o3 ..., on} regulated by Gaussian mixture distributions bj(ot) where 1<=j<=N, to simulate the spectral variance of speech. Any and every sequence of states that has the same length as the symbol sequence is possible, each with a different probability. The sequence of states is considered to be “hidden” from the observer who only sees the sequence of output

symbol, and that is why this model is called Hidden Markov Model. The Markov nature of the HMM i.e. the probability of being in a state is dependent only on the previous state, admits use of the Viterbi algorithm to generate the given sequence symbols, without having to search all possible sequences. At each and every definite instance of time, one process is assumed to be in some state and an observation is produced by the other process which represents the current state. The hidden Markov chain then changes states according to its transition from state i to state j denoted as:
aij= P[Qt+1 = j| Qt = i].

5.6 Language Modelling
Language models are used to get the guidance for the search correct word sequence by predicting the probability of nth word using (n-1) previous words. Language models can be classified into:

Uniform model: Where each word has an equal probability of occurrence.
Stochastic model: Where the probability of occurrence of a word depends on the word preceding it.
Finite state languages: Where the languages use a finite state network to define the allowed word sequences.
Context free grammar: Where it can be used to encode which kind of sentences is allowed.

6. IMPLEMENTATION ISSUES FOR HMMS
There are several practical implementation issues in Hidden Markov Model. The issues are observation sequences, initial parameter estimates, missing data, choice of model size and type.

- *Multiple Observation Sequences*. In the left-right or Bakis model form of HMM, the state proceeds for state 61 at t=1 to state N at t=T in a sequential manner. The main problem with the left-right models is that one cannot use a single observation sequence for re-estimation of the model parameters. Until a transition is made to a successor state, the transient nature of the states within the model only allow a small number of observations for any state. Thus in order to have sufficient data to make reliable estimates of all model parameters, one has to use multiple sequences.
- *Initial estimates of HMM Parameters*. The a choice of the initial estimates of the HMM parameters is an important issue, so that the local maximum is the global maximum of the likelihood function. There is no straightforward way to determine this. Experience has shown that either random or uniform initial estimates of π and A parameters is adequate for giving useful re-estimates of these parameters in almost all cases, and for the B parameters good initial estimates are helpful in the discrete symbol case and are essential in continuous distribution case [16]. The initial estimates of the HMM parameters can be obtained by various ways viz., manual segmentation of the observation sequence(s) into states with averaging of observations within states; maximum likelihood segmentation of observations with averaging; segmental k-means segmentation with clustering, to name a few.
- *Choice of Model*. Another important issue in implementing HMMs is the choice of the type of model ergodic or left-right or some other form, the choice of model size indicating the numbers of states, and the choice of observation symbols – discrete or continuous, single of multi-mixture. There is no simple, theoretically correct way of making such choices. The choices are made depending on the signal being modeled.

7. LIMITATIONS OF HMM IN SPEECH RECOGNITION
There are also some inherent limitations of this statistical model for speech. Some of the major drawback are reviewed here:
T+he assumption that successive observations are independent. The successive observations in reality are rarely independent of each other.
- The Markov assumption itself. Hidden Markov Modeling is based on the Markov property, which states that the probability of being in a given state at time t only depends on the state at time t-1. This is not always the case for speech sounds where dependencies sometimes extend through several states.
- The distribution of individual observation parameters can be well represented as a mixture of Gaussian or auto regressive densities.
- Constant length observation frames. This requirement restricts the possibilities on feature extraction (front end processing). If the frame length be daynmically decided by the front end, better representations could potentially be extracted.
- Trial and error method for choosing a model topology. Findings of various researchers show that the left-to-right architecture performs better that ergodic. But there is no formal method for deciding upon the architecture for solving a problem. Also, there is no method to find out the number of states and transitions required for a model, whether to have alternative paths through a model, whether to use the same topology for all the HMM models in that set.
- The number of parameters needed to set up an HMM is huge. For a simple four-state HMM with five continuous channels, there would be a total of 50 parameters that would need to be evaluated. 40 of the parameters are means and standard deviations, which are themselves aggregate values.

- Amount of data required to train an HMM is very large. As a result of the number of parameters to be estimated in a typical set of HMMs, large training data is hard to be obtained. Sometimes, techniques such as semi continuous HMMs, triphone clustering and interpolation have been successfully used to improve the adverse effects of insufficient training [18].

In spite of these limitations they have been found to work well when applied to certain types of speech recognition problems.

8. STRENGHTS OF HMM

A standout amongst the many difficulties in automatic speech recognition (ASR) that separates the field from traditional classification tasks is the treatment of variable-length input.

A simple example is that, pronunciations of the word hello will be having multiple time durations (therefore multiple feature dimensions) even if the class label is the same. It will be more complicated in continuous ASR, where both the input and output may be of different in length.

In the beginning of ASR, Dynamic Time Warping (DTW) was handling variable-length input problem. This was quickly substituted by Hidden Markov Model (HMM). It turns out HMM is a decent model of how speech is produced. HMM captures the temporal elasticity of speech as well as provides a rigorous framework for modeling the relationship between acoustic features (observation) and a relatively small set of phones (hidden states). With many sophisticated and efficient algorithms for training and decoding developed the properties of HMM are well understood. These are the factors that made HMM more popular in ASR.

HMM has many limitations in the context of speech modeling. The greatest restriction of HMM is the Conditional Independence presumption, which means the perception is indistinguishably and freely dispersed (i.i.d) given a hidden state. This is cannot be true; for example, the segment corresponding to the center state of phone ay is highly correlated. The recent adoption of Deep Neural Network (DNN) for acoustic modeling in place of the traditional Gaussian Mixture Model (GMM) has somewhat alleviated this problem. When the system states are partially observable and the behavior of the system is considered as autonomous HMM acts as a great modeling technique. Although with being partially observable, POMDP could be an effective alternative but the nature of autonomy of the modeled system will determine the ultimate selection of Markov models. Generally system autonomy is dependent on the agent decision making ability with concern of human intervention. HMMs are variant of FSMs where the flexibility of decision process could be perfectly implemented and the selection of output is basically could be described as real system outputs and to reach for efficient performance.

9. CONCLUSION

We have made a comparative study among various speech recognition techniques. Our focus has been to make research and development in speech recognition using hidden markov model. We have shown how hidden markov model is better than older speech recognition techniques, what are the difficulties faced while using the older speech recognition techniques. We have also mentioned the strengths and weakness of hidden markov model which will help future researchers to make further development in speech recognition techniques. The paper is not intended as a survey of all known results in speech recognition and represents only one point of view of important strengths, weaknesses of hidden markov model.

10. REFERENCES

[1] B. H. Juang; L. R. Rabiner, "Hidden Markov Models for Speech Recognition", Technometrics, Vol. 33, No. 3. (Aug., 1991), pp. 251-272. Published in American Statistical Association.

[2] 2X. D. Huang, Y. Ariki, and M. A. Jack (University of Edinburgh) "Hidden Markov Models for Speech Recognition".

[3] CiprianChelba, MichielBacchiani, JohanSchalkwyk (2010-2020) "Challenges in Automatic Speech Recognition" .

[4] Yang Liu, Member, IEEE, Elizabeth Shriberg, Andreas Stolcke, Senior Member, IEEE, Dustin Hillard, Student Member, IEEE, Mari Ostendorf, Fellow, IEEE, Mary Harper, Senior Member, IEEE "Enriching Speech Recognition with Automatic Detection of Sentence Boundaries and Disfluencies".

[5] LAWRENCE R. RABINER, FELLOW, IEEE. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition".

[6] Preeti Saini, Parneet Kaur. CSE Department, Kurukshetra University ACE, Haryana, India.(2013) "Automatic Speech Recognition: A Review" published in International Journal of Engineering Trends and Technology- Volume4.

[7] Markus Forsberg Department of Computing Science Chalmers University of Technology. (2013) "Why is Speech Recognition Difficult?"

[8] Lalit R. Bahi, Peter F. Brown, Peter V. de Souza, and Robert L. Mercer.(1986) "Maximum Mutual Information Estimation of Hidden MarkovModel Parameters for Speech Recognitio" Conference Paper in Proceedings - ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing.

[9] D. R4J REDDY. (1976) "Speech Recognition by Machine: A Review" proceedings of the IEEE.

[10] B.H. Juang & Lawrence R. Rabiner. "Automatic Speech Recognition – A Brief History of the Technology Development".

[11] Raj Redd, Lee D. Erman, R. B. Neely from Carnegie Mellon University (1972). "Working papers in speech recognition."

[12] D.B. Paul. ".Speech Recognition Using Hidden Markov Models".

[13] R K Aggarwal and M. Dave, "Markov Modeling in Hindi Speech Recognition System: A Review", CSI Journal of Computing, vol. 1, no.1,pp. 38-47, 2012.

[14] M.A.Anusuya and S.K.Katti, "Speech Recognition by Machine: A Review", (IJCSIS) International Journal of Computer Science and Information Security, vol. 6, no. 3, pp. 181-205, 2009.