International Journal of Engineering Research in Electronic and Communication Engineering (IJERECE) Vol 3, Issue 5, May 2016

A Study of Hybrid ANN-HMM Model for ASR

[1] Ashish Shendre [2] Dr. Sanjay Nalbalwar
[1][2] Electronics & Telecommunication Department
Dr. B.A.T.U. Lonere, Raigad
Maharashtra, India
[1] a.shendre@yahoo.com [2] nalbalwar_sanjayan@yahoo.com

***Abstract*— ASR (Automatic Speech Recognition) is most important for human computer interaction. In spite of tremendous advance in the ASR technique it is still far-away from human level performance. HMM (Hidden Markov model) based ASR system has made advancement but faced lots of limitation in practical problems. ANN (Artificial Neural Network) was used to replace HMM but the attempt was not that successful when trying to model high level dependency. Then attempt was made to combine this two to get the better performance. ANN-HMM hybrid model was introduced to combine the benefit of both and to increase the overall performance of ASR system. Large number of different hybrid architecture has proposed in the literature. Objective of this review paper is to focus on ANN-HMM. Most of the part of the paper is focused on describing the architecture of HMM, Neural Networkand how to use ANN for improvement of speech recognition.**

***Keywords*—Artificial Neural Nerwork, Hidden Morkov Model, Automatic Speech Recongnition**

I. INTRODUCTION

Research in the field of Automatic Speech Recognition (ASR) has been made high advance in human-machine interaction. ASR is having wide area of application like in the area of speech to Text (STT), Text to Speech(TTS), Speech to Speech(STS) conversions. ASR based on Hidden Markov Model(HMM) are widely used as it is having capability to model acoustic as wellas temporal feature of the speech signal. It maximizes the probability of generating the observation sequence for a model. Training algorithm used does not have the better discrimination capability. Research on HMM based ASR was started in early 90's [2], but publication based on HMM was made in the publication that was not generally read by researcher which are working in ASR so the field remain un undiscovered. Basic theory related to HMM was originally published in series of publication [2][11][12].

The processes which are happening in real-world generate output and the output is a signal. Signal is having various types depending on the nature of signal. We are interested in modeling such signal using signal model. Signal models are important because they provide the information about the process which has generated the signal. Signal model is important because it can be used for various purposes. Let us consider that we have to filter the noisy signal then signal model gives the information about source of signal which will be helpful for de-noising. Signal model unable us to realize the important practical system like recognition system, identification system etc.

Signal models are divided into two classes one is deterministic model and second is statistical model. Deterministic model involve in determination of some know properties of signal like signal is combination of another two signal etc. In this, simple parameters are required to estimate like amplitude of signal if signal is sine wave then phase of the sine wave may be one of the parameter is to be estimated. Second class is statistical model involve in determination of statistical properties of signal.

Initial development in this area was concentrated on speaker dependent discrete word recognition for large vocabulary, or speaker independent small size vocabulary, then research has concentrated on continuous speech recognition for large vocabulary speaker independent task [14]. First section of the paper is focused on the introductory part for ASR and different application of HMM. Second section describes the basic structure of HMM. Third section describes application of ANN in HMM. Section four describes ASR system based on HMM.

II. HIDDEN MORKOV MODEL

A probabilistic function of Hidden Markov Chain is a stochastic process generated by two interconnected process [1]. Hidden Markov Chain has a finite number of states and each state is associates with random function. We will assume that the process will be in one of the state of the model for any discrete instant of time and process will produce observation sequence depending on the state. Markov chain will change the state according to state transition probability. Observer only can see the output of random function which is associated with each state and

states are not observable and hence called as Hidden Markov Model.

We can consider the process of generating speech signal from the vocal cord as a Hidden Markov process. We can consider vocal track is in one of the state and for each state it will produce finite number of observation.

To clear the idea of the markov process let us consider the three different state of the day in terms of temperature i.e.low (l), medium(m) and high(h) are three different state of a day as shown in the fig. 1 if we observe every day of the weak and noted the temperature then we will get some sequence of this three different state total seven samples. State of the hmm represent present day temperature and process of transition from one state to another state can be model by construction such model. now consider we constructed such transition for

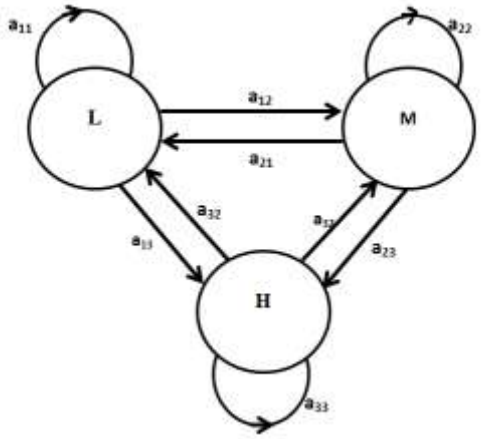

Fig.1 Three State Hmm Model

Three Different Season Of Year And Constructed Hmm Models For Three Different Season We Will Be Having Some Different Pattern Of The Temperature Summer Season May Contain More number Of High (H) Than The Other Season And Hence every Season Can Be Characterized By The Pattern Of Observation now Suppose We Are Having Some Un-Know Pattern And This Un-Know Pattern Is Feed To Three Hmm And One Hmm Model will give the highest probability and hence we can determine to which model the given pattern belong to. How to calculate the probability associate with given hmm model is discussed next. Let us consider the state transition probability matrix for fig. 1 hmm model

$$A_{ij}=\begin{bmatrix}0.3 & 0.2 & 0.5\\ 0.4 & 0.4 & 0.2\\ 0.1 & 0.3 & 0.6\end{bmatrix}$$

Rows of the matrix represent present state of the day as low (l), medium (m), and high (h) respectively and column of matrix represent the state of the next day. For example element of matrix a_{23} represent transition probability from 2nd i.e. from medium state (m) to 3rd state i.e. High state (h).transition matrix is characteristic of the hmm model and we can calculate what is the probability of next seven day i.e. What is probability that for example it will take state of the day as {l, m, m, h, l, m, h} which is called as observation sequence denoted by o so from transition table we can calculate the probability of the state sequence as follow.

$$P(o|m) = p\{l, m, m, h, l, m, h|model\}$$
$$= p(l) * p(m|l) * p(m|m) * p(h|m) * p(l|h) * p(m|l) * p(h|m)$$
$$= \mathbf{\Pi}_l * a21 * a22 * a32 * a13 * a21 * a3$$
$$= 1 * 0.4 * 0.4 * 0.3 * 0.5 * 0.4 * 0.3$$
$$= 0.0028$$

Similarly if we know state transition probability matrix of all the hmm models then we can calculate p(o|m) for each hmm. Matrixes are constructed such that it will give highest probability if the patterns belong to its model and hence we can differentiate between the patterns or we can say that the pattern is classified.

Hmm model is characterized by following elements.

A) N number of state.

Although some state of hmm may be hidden. To design hmm we should know the number of state in the hmm. There is some physical significance between number of state and the physical process to which the HMM going to model.

B) M number of observation symbol

For each state, observation symbol correspond to physical output of the process like in the previous example bag correspond to the number of state and the book peaked from the bag correspond to observation of the process.

C) State transition probability

HMM model can be in of the state and it can go the next state depending on the transition matrix. If Model can jump from one state to any another remaining state then state transition probability $a_{ij} > 0$ for all i and j otherwise for some i and j it will be zero.

D) Observation Symbol probability distribution

The probability distribution of each observation symbol for each state is required to obtain complete the probability distribution.

E) Initial state probability distribution

What is probability of the states to be first state of observation sequence or the first state of the process that is we are modeling is given by initial state probability distribution.

Advantages and Drawbacks of HMM models

Standard HMM procedure for speech recognize has capability to recognize large vocabulary [10000 words], speaker independent and continuous speech recognition. HMM can deal with temporal aspect of speech again there is power full training and decoding algorithm that make HMM more robust for speech recognition task. HMMs can also be used to implement phonological rule, constructing the word model from the phonemes and language model from word model. Training algorithm in HMM maximize likelihood instead of increase in posterior probability this leads to poor discrimination capability, assumption about the state sequence as first order Markov chain lead to decrease the accuracy.

III. ARTIFICIAL NEURAL NETWORK

ANN has widely used for the classification purpose and its application for wide variety has got more attention of researcher. In 1980's number of researcher started to use ANN for speech classification.

As focus of this review paper is to use ANN for statistical ASR. We are going to use ANN for phonetic probability estimation and this phonetic posterior probability are going to use as a parameter of HMM [3][4][5].

ANN is having large number of advantages that will make ANN useful for speech recognition.

- ANN inherently useful for discriminant learning.So when feature vectors are given to ANN as input the output of ANN network will be the class to which this feature vectors belong to.
- When Neural Network is trained on the large database it will less affected by distortion when used for classification.
- They are more robust in discriminating by using fewer amounts of training data.
- Input to the ANN are the feature vector and feature vectors will belong to any one of thephoneme so such kind of Neural Network structure are not useful for continuous speech recognition as there will be loss of inter dependence of the phoneme with previous as this drawback can be overcome by using recurrent neural Network.

A) Posterior probability estimation from ANN

Several authors have showed that the output of ANN can be interpreted as estimate of posterior probability of the classes of the input vectors [2][3][4].

Multilayer Perceptron[MLP]one form of ANN commonly used for speech recognition purpose [6]. MLP is having feed-forward architecture which is able to model the speech unitswhich are depend on some previous output of recognizer. It is proved that with one Hidden layer and enough hidden unit is capable of modeling any input output mapping[7].In training of Neural Network connection weight has to be modified to map input vector to with desired output. Back-Propagation algorithm[8]is generally used for adjusting the weight of the neural network.

Simplest way for speech recognition using ANN is to give complete input vector corresponding to the word or phonemes to the Neural Network and train the Neural Network using various training algorithm. And associate the output to one of the phoneme or word but this approach is useful for isolated speech recognition or small word speech recognition and not useful for large world speech recognition. ANN used for speech recognition should be able to consider the input sequence which are sequential as most of the time speech is depend on some previous state [8]Recurrent Neural Network are useful for such type of pattern recognition system.

Various structure of neural network like MLP, RNN TDNN can be used for speech recognition but this are useful only for small vocabulary speech recognition this network are not useful for large size, continuous speech recognition.

III. ASR BASED ON HMM

Speech signal is transformed into feature vectors the process is called as feature extraction. Most widely used feature extraction technique is based on Mel Frequency Cepstral Coefficients (MFCC's)[9]. One of the recent developments in the MFCC is Delta-Delta MFCC which improves the speech recognition. Some other techniques used are Linear predictive coding (LPC),Linear Predictive Cepstral Coefficients (LPCC); Perceptual Linear Prediction (PLP); And Neural Predictive Coding (NPC). Extracted feature are given to phone model to identify the segment of speech to which this extracted feature belong to. Then word model will transform the local hypothesis (phone model) to global hypothesis means pattern of phonemes are used to form a word, language model will help to add syntactic, semantic constrain on the recognition.

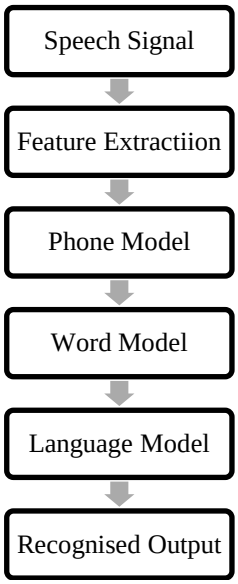

Fig. 2Speech Recognitions System Based on HMM

Extracted feature vectorsare the characteristic of speech signal. Consider the speech signal X is converted into feature vector $\{x_1, x_2, x_3, x_4 \ldots\ldots, x_L\}$ this feature vectors are obtained by overlapping the speech signal and they are of fixed size. Problem is to find out the best sequence of word which will produce the feature vector Y. Decoder will then try to find out the word sequence {w} which best matches with the feature vectorsX and it will select word sequence which is having highest probability this can be written as follow

$$W = \max\{P(w|X)\}$$

It is difficult to find the *P(w|X)* so it is approximated by using the Bayes rule.HMM is used to construct phone model from the feature vector and phone model is used to obtain word model and constrain can be applied on word model to obtain the language model. And hence *P(w|X)* can be approximated to *P(q|x)* where q represent the state of the HMM. Bayes Rule can be used to calculate the probability state*(q),*of HMM given the observation sequence*(X),*

$$\frac{P(Xn|qk)}{P(Xn)} = \frac{P(qk|Xn)}{P(qk)}$$

$$P(qk|Xn) = \frac{P(Xn|qk) * p(qk)}{P(Xn)}$$

Probability of generating sequence X given state qk i.e. P(X|qk) is calculated from acoustic model. Probability of word sequence is calculated from language model. From the above equation we can calculate which HMM model best describing the given feature vector and hence we can recognize the speech.

IV. CONCLUSION

Even if HMM largely used for speech recognition such type of statistical models are having some limitation as we are assuming present state of the HMM is depend only on the previous state. Training data is going to affect the recognition performance so there should be large amount training data available. Hybrid ANN/HMM model required more number of computation than when only HMM is used for the recognition as hybrid model required training of ANN. So to use hybrid model for recognition we require hardware that is capable of performing fast computation.

REFERENCES

[1] Lawrence R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition" *proceedings of IEEE, Vol. 77, No. 2, February 1989.*

[2] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markovchains," Ann. Math. Stat.vol. 37, pp. 1554-1563,1966.

[3] NelsonMorgon and Herve A. Bourlard, "Neural Networks for Statistical Recognition of Continuous Speech," Proceedings of the IEEE, vol. 83, no. 5, May 1995.

[4] Michael D. Richard and Richard P. Lippmann,"Neural Network Classifiers Estimate Bayesian a posteriori Probabilities," Neural Computation Volume 3 issue 4 1991.

[5] Xian Tang, "Hybrid Hidden Markov Model and Artificial NeuralNetwork for Automatic Speech Recognition," Pacific-Asia Conference on Circuits,Communications and System 2009.

[6] T. Robinson, "An application of recurrent nets to phone probability estimation," IEEE Trans. Neural Networks, vol. 5, up.1994.

[7] T. Poggio and F. Girosi, "Networks for approximation and learning," Proc. IEEE vol. 78, pp. 1481-1497, Sept. 1989.

[8] D. E. Rumelhart, G. E. Hinton and R. J. Williams, "Learning internal representations by error

propagation," in Parallel Distributed Processing, Eds. Cambridge MA: MIT Press, 1986, vol. 1, pp. 318-362.

[9] Selina Chu, Shrikanth Narayanan C.-C. Jay Kuo, "Environmental Sound Recognition With Time–Frequency Audio Features," IEEE transaction on audio, speech and language processing, VOL. 17, NO. 6, August 2009.

[10] Mark Gales and Steve Young, "The Application of Hidden Markov Models in Speech Recognition" *foundations and Trendsin Signal Processing Vol. 1, No. 3 (2007) 195–304*

[11] L. E. Baum and J. A. Egon "An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology" *Bull. Amer. Meteorol. Soc. vol. 73, pp. 360-363, 1967.*

[12] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains" *Ann. Math. Stat.vol. 41, no. 1, pp. 164-171, 1970.*

[13] S. E. Levinson, L. R. Rabiner, M. M. Sondhi, "Speaker Independent Isolated Digit Recognition Using Hidden Markov Models*" ICASSP 83, BOSTON.*