Speech Recognition Using Hidden Markov Model With Neural Network Probability Estimators

Dharmendra P Kanejiya, IIT Delhi

Abstract

Automatic speech recognition, an exciting area of research, can allow us to interact with computers using 'spoken commands'. Hidden Markov Model (HMM) is the most widely used statistical speech recognition technique. In this paper, I describe HMM based speech recognition system and point to its limitations, some of which can be alleviated by use of Neural Networks. I describe the use of Multilayer Perceptron (MLP) as probability estimators in HMM framework. The practical implementation of this hybrid HMM/MLP based speech recognition system is also illustrated. Experimental results on isolated word, limited vocabulary task are also reported. Particularly, the effect of learning factor variation in training shows that by tuning it properly, the recognition performance can be improved significantly.

I. Introduction

Today when we interact with computers, we normally have to resort to more artificial methods like keyboard, mouse driven menu system or touch-screen. But speech recognition can improve the user interface by allowing spoken commands as input. Since speech is the most natural form of human communication, it has been a long standing challenge to bring human-computer dialogue as close as possible to human-human dialogue. Speech recognition by machine can be considered as a problem of converting a speech waveform into words. A speech recognizer is designed to recognize one of a set of words or phrases specified in a vocabulary.

The principal difficulty for speech recognition lies in the very nature of the speech signal. The speech signal is highly non-stationary and much information is contained in the transient parts What we as humans identify as the same speech component e.g. a certain phoneme, has a large variety of different pronunciations. They vary in time an they depend a lot on the context. Therefore, a few simple operations on the speech signal are not sufficient for recognition and a complicated and sophisticated processing chain has to be designed in order to come close to an acceptable recognition rate.

The state-of-art speech recognition systems are based on Hidden Markov Model (HMM) [1]. These models provide a reasonable statistical superstructure for both the estimation of system parameters and for effective characterization of temporal variation in speech signal. In this respect, Artificial Neural Networks(ANN) still remain problematic. But HMMs have certain limitations like a priori assumption of probability density, maximum likelihood criteria for parameter estimation which is not optimum from classification point of view. On the other end, ANNs don't assume any knowledge of distribution of input and can also be trained in discriminant way.

Finally some efforts have been made to combine the temporal modeling of HMM and discriminant learning property of ANN to get a robust speech recognition system performance [2]. This is called the hybrid HMM/ANN approach which will be the focus of this paper.

In section II, an overview of HMM based speech recognition system is given. Then section III discusses the use of Neural Networks in probability estimation. The hybrid approach for speech recognition is treated in section IV. Then section V provides implementation details and results obtained followed by section VI which concludes.

II. Hidden Markov Model Based Speech Recognition

The procedure [3] for isolated word, discrete density HMM speech recognition is shown in fig 1. There are mainly three procedures: feature extraction, vector quantization and HMM probability scoring. For continuous density HMM case, the vector quantization is absent and HMM includes parameters of mixtures of Gaussians.

After silence removal, feature extraction produces usually a vector every 10 ms, an *acoustic vector*, which represents the salient feature of a window of about 30 ms. A popular choice of features are the Linear Prediction Coefficients (LPC), Mel Frequency Cepstral Coefficients (MFCC).

The vector quantizer detects clusters in the set of acoustic vectors and determines a representative vector for each cluster. This vector is coded, and the string of codebook vectors is fed to the recognizer. With respect to incoming speech signal the data flow is considerably reduced.

The most widely used speech recognizers are based on HMMs. Hidden Markov modeling of speech assumes that speech is a piecewise stationary process, that is , and utterance is modeled as a succession of discrete stationary states with instantaneous transitions between these states. These states are not directly observable but only through observation vectors representing speech features. A simple HMM is shown in fig.2. 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 subword units like phonemes. For a given Markov model of, say, a word, the probability that the pronunciation of the word produces a certain speech signal can be determined. 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.

All automatic speech recognition systems, just as humans, acquire their ability through learning. Speech utterances with known meaning are fed to the system from a database. The system then adapts its parameters such that it reacts similarly to all utterances with the same meaning. HMMs are trained using Baum-Welch algorithm to maximize likelihood.. HMMs inherently incorporate the sequential and statistical character of the speech signal and they have proved their efficiency in speech recognition. However, standard HMMs still suffer from several weaknesses, namely [4]:

- a priori choice of a model topology, e.g. number of states.
- a priori choice of statistical distribution in continuous density HMM for emission probability associated with each state : mixtures of multivariate Gaussians.
- poor discrimination due to the training algorithm which maximizes likelihood instead of a posteriori probabilities.

III. ANN As Probability Estimator

Multilayer Perceptrons (MLP) are probably the best studied class of neural networks. They have layered feedforward architecture with an input layer, zero or more hidden layers and an output layer. Fig 3 shows a two layer perceptron which is used in our experiments described in section V. Each layer is connected to the previous via a weight matrix and operates according to the relation

$$y_i^L = f\left(\sum_j w_{ij}^{L,L-1} y_j^{L-1} \right) \qquad (1)$$

where,

y_i^L output of unit *i* in layer *L*

$w_{ij}^{L,L-1}$ element of the weight matrix between layers *L-1* and *L*

f transfer function of a unit, typically a sigmoid :

$$f(x) = \frac{1}{1+\exp(-x)} \qquad (2)$$

equation (1) can incorporate a bias for y_i^L by assuming a unit in layer *L-1* with a fixed output of 1. MLPs are trained to associate an input vector with a desired output vector. Both classification and regression may be performed in the same framework. In the case of N-class classification, a network with N outputs would be used : one for each class. A "1-from-N" training scheme would thus be used, where the desired output vector would contain a one for the correct class and zero for all other classes.

Training is accomplished via the back-propagation algorithm [5], which is a steepest descent procedure. For large problems stochastic approximation is usually adopted (per sample update rather than batch update) as follows:

$$\mathbf{W}(k+1) = \mathbf{W}(k) + \eta\left[-\frac{\partial E}{\partial \mathbf{W}}\right] + \beta\left[\Delta\mathbf{W}(k-1)\right] \tag{3}$$

where η is the learning factor, β is momentum factor and *E* is the error function defined in [5] which is minimized.

MLPs may be used to estimate probabilities. Bourlard and Wellekens [6] showed that after "1-from-N" training , a MLP output value, given input ***x***, will be an estimate of the posteriori probability $P(c_i/\boldsymbol{x})$ of the corresponding class c_i. For this the output units should be constrained to be non-negative, less than one and sum (over all the classes) to one. One way of achieving this is by adopting a normalized output function such as 'softmax' [7],

$$f(x_i^L) = \frac{\exp(x_i^L)}{\sum_{j\in L} \exp(x_j^L)} \tag{4}$$

where x_i^L is the activation (pretransfer function output) of unit *i* in layer *L*.

By estimating posterior probability using MLP, we are able to perform discriminative training of input features, which is optimum from classification point of view.

IV. Hybrid HMM/MLP Speech Recognition System

As explained in the previous section, while HMMs are effective in modeling temporal variation in speech, they lack discrimination capability because of maximum-likelihood parameter estimation criteria. On the other hand, strength of MLP is in the fact that they don't need to assume about statistical distribution of input as well as they can be trained to exhibit discriminant properties. But they are weak in dealing with the time sequential nature of speech. So recent research [2,4,7] tried to develop hybrid HMM/MLP systems in which MLPs are used to compute the emission probability associated with each state of HMM.

In [7] it is proved that if each output unit of a Multi Layer Perceptron (MLP) is associated with a particular state q_t of the set of states Q={1,2,…N} on which the Markov Models are defined, it is possible to train the MLP to generate *a posteriori* probabilities like $P(q_t = i \mid \boldsymbol{o}_t)$ when $\boldsymbol{o}_t$, a particular acoustic feature vector, is provided to its input. The training data set of the MLP consists of a labeled sequence of *T* acoustic vectors $\{\boldsymbol{o}_1,\boldsymbol{o}_2,\ldots,\boldsymbol{o}_T\}$. At time *t*, the input pattern to MLP is the acoustic vector $\boldsymbol{o}_t$ associated with a particular state. The training of the MLP parameters (weights) is based on the minimization of the following Mean Square Error criterion,

$$E = \frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{N}\left[f_i(\mathbf{o}_t) - d_i(\mathbf{o}_t)\right]^2 \tag{5}$$

where $f_i(\boldsymbol{o}_t)$ represents the *i*th output value of the MLP given $\boldsymbol{o}_t$ at its input and $d_i(\boldsymbol{o}_t)$ is the associated target value and is equal to δ_{il} if the input is known to belong to class *l* of Q ("1-from-N" training).

It is also shown that if the MLP contains enough parameters and if the global minimum of E (5) is reached, the output values of the MLP are the estimates of the *a posteriori* probability density functions which are optimal for classification:

$$f_i^{opt}(\mathbf{o}_t) = P(q_t = i \mid \mathbf{o}_t) \tag{6}$$

Dividing each MLP output by its relevant frequency (*a priori* probability density), results in scaled *likelihood* estimate that is suitable as emission probabilities for HMMs:

$$\frac{p(\boldsymbol{o}_t \mid q_t = i)}{p(\boldsymbol{o}_t)} = \frac{P(q_t = i \mid \boldsymbol{o}_t)}{P(q_t = i)} \tag{7}$$

The training procedure of the hybrid system is quite similar to the one of the standard HMMs. Given a parameter set for the HMM (transition probabilities a_{ij} and initial state probabilities π) and the MLP (weights), the best state sequence can be found with a Viterbi algorithm, leading to a labeling of the feature sequence. This labeling allows the MLP to be trained with a standard back-propagation algorithm updating the weights and to compute new values for transition probabilities. This procedure is repeated iteratively until an optimum is reached.

Thus the implementation schemes for training and testing of hybrid speech recognition system are as shown in block diagrams in fig. 4.

This training procedure has two major advantages in comparison with maximum likelihood estimation. First, it is based on a posteriori probability approximation, which is discriminant by nature. Second the MLP gives a model for the emission probability function without making any assumptions about the distribution of features among the states.

Another advantage of the hybrid system is that these results hold if the MLP is fed with a larger feature window taking more context into account. This procedure allows the time correlation between the successive acoustic vector in the recognition process to be taken into account.

V. Implementation and Results

The theoretical procedure discussed in this paper was implemented for the task of isolated word, limited vocabulary, speaker dependent speech recognition. The database consisted of 50 utterances of the ten digit words (‘zero’ to ‘nine’), out of which 20 utterances were used for training and the rest were used for testing the performance.

A feature vector of 12 dimensional Mel Frequency Cepstral Coefficients (MFCC) was extracted from 16 ms window of speech for every 8 ms. The MLP had 13 input neurons, 19 hidden neurons and 6 output neurons representing 6 states of HMM. An *epoch* period is when all the training patterns are applied once to the neural network. For better training, the number of epochs is more than one. The learning factor η and momentum factor β in weight update equation (3) and the number of epochs were varied and their effect was observed by repeating the training and testing phases for each variation. The recognition performance of these experiments are shown below:

learning factor η	momentum factor β	training epochs	Recognition Rate (%)
1.0	0.07	2	50.00
0.2	0.07	5	83.00
0.19	0.07	10	85.00

In other experiments, the learning factor η was updated in the following manner after each feature vector being processed for training:

$$\eta(k) = \eta(k-1)\left(1 - \frac{\theta}{k}\right) \tag{8}$$

The recognition results in these cases are shown below

learning factor η(0)	momentum factor β	update parameter θ	Recognition Rate (%)
0.20	0.07	0.01	89.00
0.25	0.07	0.08	90.33
0.25	0.07	0.15	83.00
0.25	0.10	0.08	86.00

VI. Conclusion

In this paper, I have discussed the speech recognition problem in combined HMM/ANN framework. HMMs have the ability to model temporal variation in speech signal, but they lack in discrimination property for classification. Neural networks are able to derive the posterior probability without making any assumption as well as are discriminant but are not effective in modeling time-variability of speech. The hybrid approach, which combines best from both, was implemented and the performance was encouraging. More importantly, the effect of variation in learning factor η in MLP weight update equation for training was significant. It was also observed that iteratively decreasing the value of η leads to better training and thus better recognition rate.

References

[1] Rabiner, L. R., "A tutorial on Hidden Markov Models and selected applications in speech recognition", *Proceedings of the IEEE,* Vol. 77, No. 2, pp. 257-285, Feb. 1989.

[2] Bourland, H., Morgan, N., "Continuous speech recognition by connectionist statistical methods", *IEEE Tran. Neural Networks,* Vol. 4, No. 6, Nov. 1993.

[3] Kanejiya, D. P., "Speech Recognition In Noisy Environment", *M.Tech. Thesis*, Department of Electrical Engineering, IIT Delhi, Dec. 1999.

[4] Hennebert, J., Hasler, M., Dedieu, H., "Neural networks in speech recognition", *Proc. 6th Microcomputer School,* Prague, Czech Republic, Sept. 1994.

[5] Bose, N. K., Liang, P., *Neural Networks Fundamentals with Graphs, Algorithms and Applications*, Tata McGraw-Hill, New Delhi, 1998.

[6] Bourlard, H., Wellekens, C. J., "Links between Markov models and Multi-layer perceptrons", *IEEE Tran. Patt. Anal. Machine Intell.*, Vol. 12, pp. 1167-1178, 1990.

[7] Renals, S., Morgan, N., Bourland, H., Cohen, M., and Franco, H., "Connectionist probability estimators in HMM speech recognition", *IEEE Tran. Speech & Audio Proc.*, Vol. 2, No. 1, Part II, Jan. 1994.

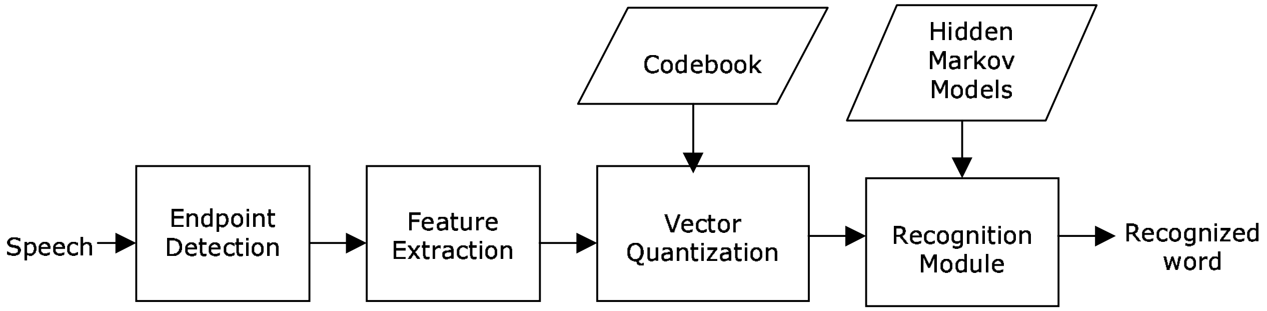

Figure 1 : Isolated word, Discrete density HMM based speech recognition system

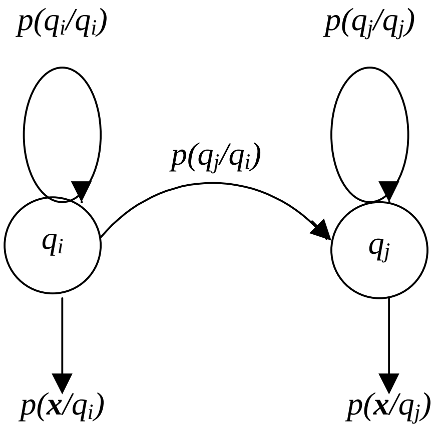

Figure 2 : A two-state left-to-right HMM

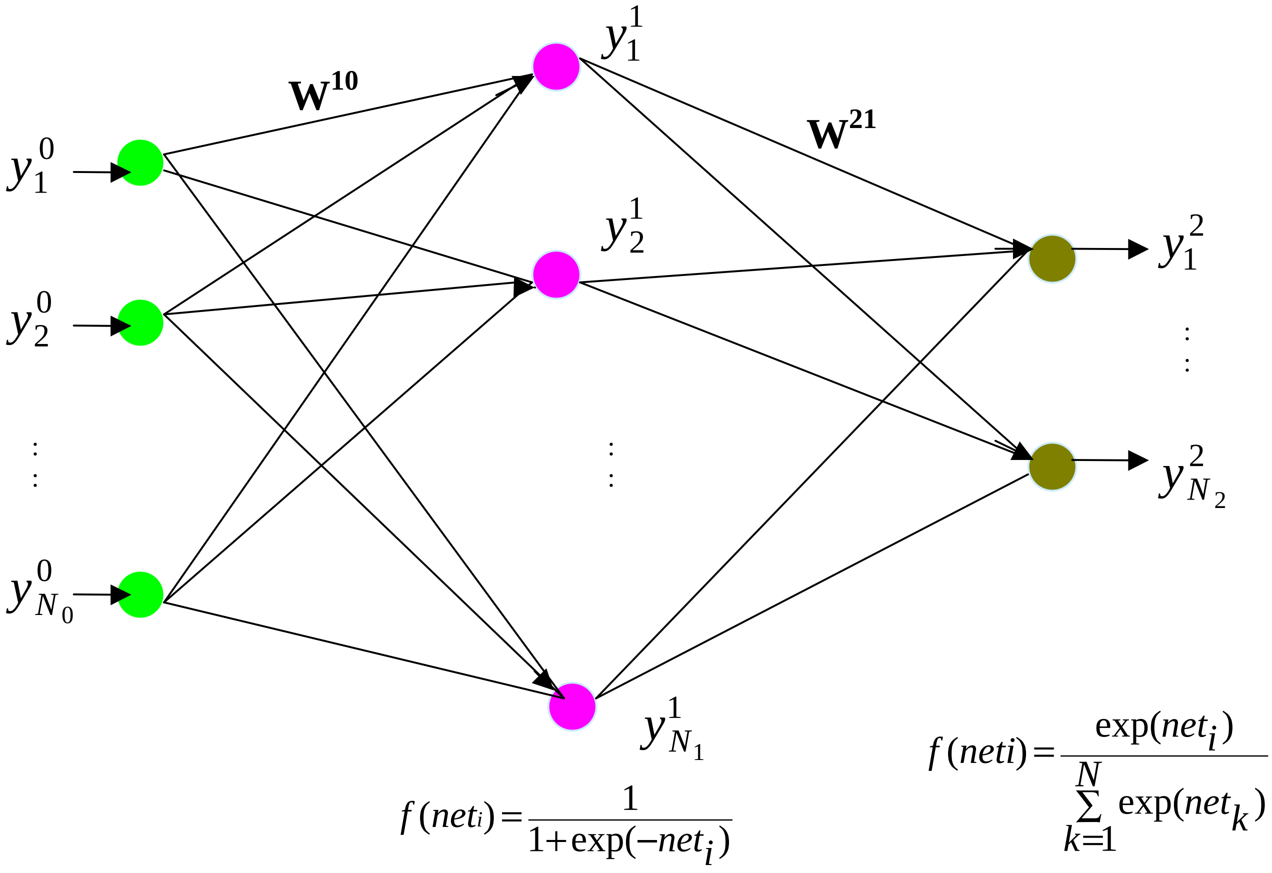

Figure 3 : A two-layer perceptron with activation functions

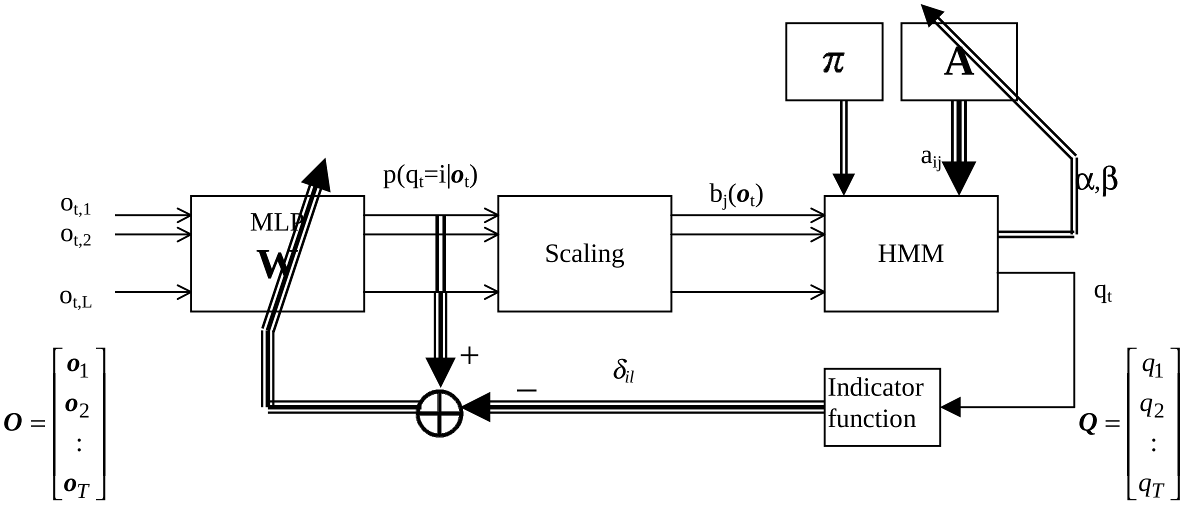
π
A
a_{ij}
α,β
$o_{t,1}$
$o_{t,2}$
$o_{t,L}$
MLP
W
$p(q_t=i|\boldsymbol{o}_t)$
Scaling
$b_j(\boldsymbol{o}_t)$
HMM
q_t
+
−
δ_{il}
Indicator function
$\boldsymbol{O}=\begin{bmatrix}\boldsymbol{o}_1\\ \boldsymbol{o}_2\\ \vdots\\ \boldsymbol{o}_T\end{bmatrix}$
$\boldsymbol{Q}=\begin{bmatrix}q_1\\ q_2\\ \vdots\\ q_T\end{bmatrix}$

Figure 4(a) : Hybrid HMM/MLP Training Procedure

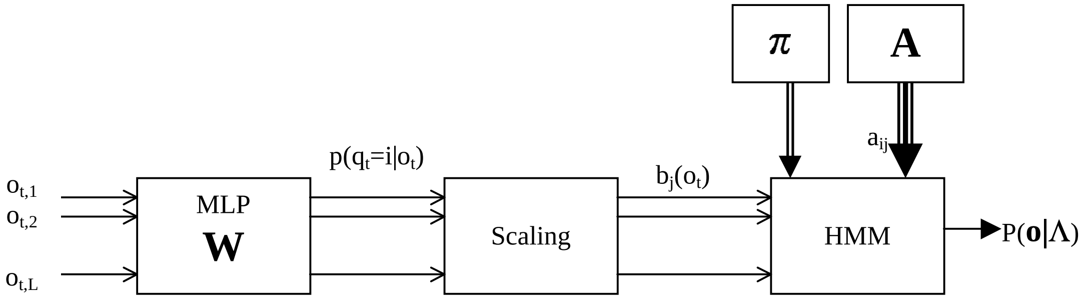

Figure 4(b) : Hybrid HMM/MLP Recognition Procedure