A STUDY ON THE COMBINATION OF HIDDEN MARKOV MODELS AND MULTI-LAYER PERCEPTRON FOR SPEECH RECOGNITION

J. M. Song

Speech Technology Research Group
Department of Electrical Engineering
The University of Sydney

ABSTRACT - This paper presents an enhanced speech recognition algorithm by combining continuous hidden Markov modelling (HMM) with a multi-layer perceptron (MLP). The first stage of speech recognition, carried out by the HMM, selects a small group of candidates and projects incoming speech vectors into state normalized vectors. The second stage of MLP classifies each normalized vector generated by the HMM and determines the best candidate. In this architecture, the HMM plays a role of pre-classification, while the MLP is used for decision refinement. A simple speaker-independent isolated digits telephone speech database was used to test this approach. The result shows that the recognition performance increases from 92.9% to 93.8%.

INTRODUCTION

It is well known that hidden Markov modelling can be used successfully for automatic speech recognition. The power of this statistical pattern matching approach lies in the capability to modelling the spectral characteristics and capture the underlying temporal or dynamic structure of speech signal. However, the standard HMM based on the maximum likelihood criterion (ML) still suffers from some weaknesses caused by several assumptions. For instance, the HMM requires each acoustic feature vector to be independent, and the probability distribution of acoustic vectors in the corresponding space to be approximated by mixtures of Gaussian distributions. Even the model training paradigm remains questionable, i.e. each model is trained independently, rather than competitively. All these problems reduce the discriminative power in classification. Although some of these problems can be tackled within the HMM framework by using different optimum criteria, such as maximum mutual information (MMI) or corrective training, the advantages of these approaches are not totally clear.

There has been a lot of research devoted to the field of artificial neural networks (ANN) and it has been demonstrated that ANNs produce powerful discriminative functions to classifying static patterns. In ANN knowledge or constraints are not encoded in individual units, rules, or procedures, but distributed across many simple computing units. Uncertainty is modeled not as likelihoods or probability density functions in individual model, but by patterns of activity of many units. The training paradigm is competitive, i.e. not only does it increases the chance for correct responses, but it also discourages the mis-classified outputs. Once the training phase is finished, the classification task can be performed very fast due to the massive parallel structure of neural networks. Unfortunately ANNs also have certain weaknesses for use in speech recognition. For instance ANNs require a fixed-size vector applied to input layer and speech pre-segmented into speech units (word or sub-word) due to their inability to deal with the time sequential nature of speech. The research in the incorporation of HMM and ANN has become very attractive (Bourlard et al, 1992). The work presented in this paper combines of hidden Markov models and multi-layer perceptron together within a speech recognition system and demonstrates that a better recognition performance can be achieved by taking advantage of the both components.

THE HMM APPROACH

A probabilistic function of a hidden Markov chain is a stochastic process of two interrelated mechanisms, an underlying Markov chain having the finite number of states, and a set of random functions, each of which is associated with the individual state. At a discrete instant of time, the process is assumed to be in

some state and an observation is generated by the random function corresponding to the current state. The underlying Markov chain then changes state according to the transition probabilities. The observer sees only the output of random functions associated with each state and cannot directly observe the states of the underlying Markov chain. In the use of HMM approach for speech recognition, we have to give a set of HMM specifications, such as the HMM topology and the type of probability function density associated with each state. In the speech recognition context, one of the most elusive aspects in HMM is its interpretation of states shown in the Figure 1. Since they are hidden from observation, it is hard to determine what do they represent exactly. One interpretation of the states in HMM is that they describe the vocal configuration during speaking. When the configuration changes, the underlying states changes accordingly. This viewpoint is the basis for projection of a sequence of acoustic vectors into a state-normalized vector.

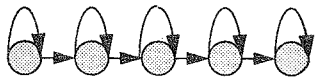

Figure 1. A 5-state left-to-right Markov chain

It is well known that speech signals are non-stationary. However this non-stationary process can be approximated by a concatenation of a sequence of stationary processes, which in turn represents the short term speech signal. Although this point of view is generally accepted, how to merge incoming acoustic feature vectors into a sequence of stationary parts remains an open question. One method is to measure the distance of adjacent acoustic vectors and set up a rather *ad hoc* threshold to decide whether a particular pair of vector should be merged. This method based on simple distance metric is not reliable, particularly if we want to normalized all incoming acoustic vectors to a single fixed number of vectors. The inherent temporal structure might be destroyed by an unreasonable forced merging mechanism. However, the HMM structure could help to deal with this acoustic vector normalization problem. Let's suppose that in a HMM, state self transition represents a stationary phenomenon, while state forward transition represents a non-stationary phenomenon. All acoustic vectors located onto the same state after Viterbi alignment should belong to the same stationary process. Therefore we can merge them together and use a centroid vector to represent this cluster of vectors. In view of this the incoming acoustic vectors from a front end can be normalized to state normalized acoustic feature vectors. This is suitable for further classification by artificial neural network such as multi-layer perceptron.

MULTI-LAYER PERCEPTRON APPROACH

One of the most popular ANNs is probably the feed forward multi-layer perceptron. The capabilities of MLP stem from the nonlinearities used within nodes and the competitive learning paradigm. A three layer MLP used in our experiment is shown in the Figure 2.

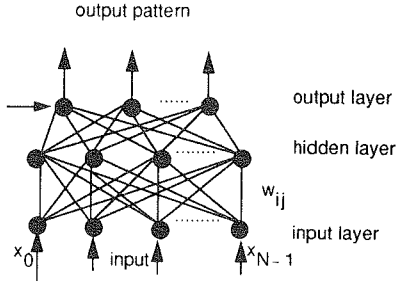

Figure 2. A three layer multi-layer perceptron network

To train this multi-layer perceptron to perform a pattern classification task, we must adjust the weights of each unit in such a way that the error between the desired output and the actual output is reduced. This calculation is carried out by the back propagation algorithm which is a gradient descent method that tries to minimize the mean squared error of the system by moving down the gradient of the error curve.

In Figure 2, the components of the state normalized MFCC vector applied to the input layer are denoted by $x0,...x_{N-1}$, the output of node j at any layer is determined by a nonlinear sigmoid function

$$o_j = \frac{1}{1 + e^{-net_j}}$$ where the value of net_j is the sum of weighted output from previous layer i,

$$net_j = \sum w_{ji} o_i$$

The threshold of each unit is represented by a weight with an input set to a constant value 1. At the output layer, the node associated with the maximum value is chosen as the recognized result.

HMM/MLP TRAINING

Speech Database

In order to test the methodology of combination of HMM/MLP, we used 11 digits (from zero through nine, plus oh) as vocabulary. The speech data was collected through telephone line in noisy environment, the speakers participated into the collection of training and testing database are of different language background, some of them speak English with strong foreign accent. We used 50 speakers for training both HMM and MLP, each training speaker spoke the 11 digits once. The testing database consisted of 40 different speakers. The speech database used in the experiment is limited but varied to a large degree because of the speaking style. However, our intention is to see what kind of additional discriminative ability and robustness can be brought to the HMM by adding the MLP.

Continuous HMM Training

We used 5 states and 2 mixtures of Gaussian pdfs with diagonal covariances, the Markov chain consisted of a left-to-right topology with self transition and transition to next state without skip. We chose this simple HMM configuration deliberately to find out how the succeeding MLP could help to increase the recognition performance. The HMM training procedure used is the segmental k-means algorithm and refined by the Baum-Welch algorithm. The individual word is segmented from embedding noise automatically via forced recognition (Song and Samoulien, 1992). The acoustic feature vector consists of 12 mel frequency cepstrum coefficients. When a set of HMM models has been trained, each training speech is passed through the HMM system and decoded via the Viterbi algorithm in a supervised recognition style. The resulting partitioning of acoustic feature vectors onto 5 states was investigated and a mean vector is used to represent all vectors located in the same state. In doing this, we project a variable length of acoustic vectors into an augmented fixed size vector for the succeeding MLP training.

MLP Training

We used the same training database to train a three layer perceptron. All training word tokens were state-normalized to a fixed length acoustic vector and the values of all parameters were scaled within the range (-1,1). The input layer consists of 60 nodes, and hidden layer consists of 120 nodes. In order to accelerate the MLP training procedure, we use a conjugate-gradient optimization algorithm (Powell, 1977). The conjugate-gradient algorithm is usually able to locate the minimum of a multivariate function much faster than the gradient-descent procedure. It is important to note that the conjugate-gradient technique eliminates the choice of critical parameters (such as the learning-rate and momentum parameters of back propagation), making it much easier to use.

The procedure involved in the HMM/MLP training is shown in Figure 3.

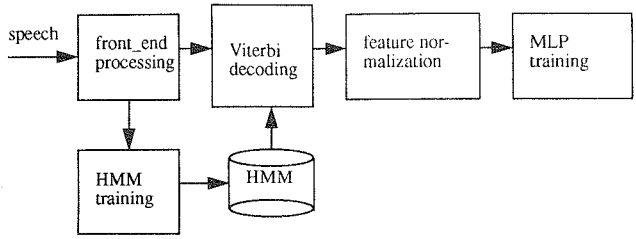

Figure 3. Block diagram for HMM/MLP training

RECOGNITION ALGORITHM AND EXPERIMENTAL RESULTS

To carry out the isolated word speech recognition, we combined the HMM with the MLP together in the way shown in the Figure 4. The front end produces a sequence of 12 dimensional MFCC vector which is matched against each HMM model. The Viterbi algorithm aligns this sequence of MFCC vectors into 5 states of each model and produce a state normalized acoustic pattern along with the accumulated likelihood value. In order to carry out this processing, we have to find out the best state sequence $S = \{s_1, s_2, \ldots s_T\}$ for the sequence of MFCC vectors $O = (O_1, O_2, \ldots O_T)$ where T denotes the frame number corresponding to the input speech. The likelihood that this acoustic sequence matches each HMM model is computed recursively and the best state transition is recorded.

Once the Viterbi alignment is finished, a state-normalized acoustic pattern is generated by each model in the following way:

for each state $1 \le i \le N$

for each MFCC vector O_t located to this state

$$\hat{O}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} O_{t_j}$$

Finally we concatenate the 5 (N=5) centroid vectors to represent the normalized acoustic pattern. Therefore the vector size of this acoustic pattern is 60 (5×12). Note that for one incoming speech signal, we have V normalized acoustic patterns, where V is the number of models. Considering the HMM is generally good at speech recognition, we only use 3 normalized acoustic patterns corresponding to the top 3 word candidates and pass them to the MLP for further refinement. In the MLP the re-classification is forced within these three candidates and ignore the other choices. Finally the normalized acoustic pattern with the largest output value from the MLP is chosen as the recognition output.

Table 1 shows the comparison of the recognition result obtained from the HMM only and that from the combined HMM/MLP.

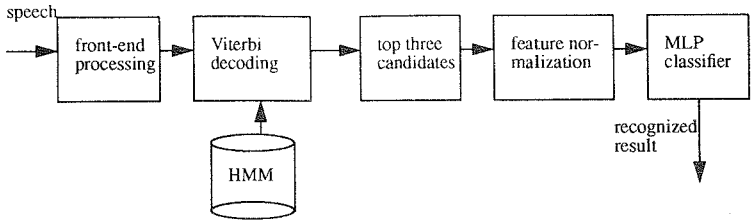

Figure 4. HMM/MLP recognition block diagram

HMM	HMM/MLP
92.9%	93.8%

TABLE 1. Comparison between HMM and HMM/MLP on multi-national speaker independent 11 isolated digit recognition. Training database: 50 speakers/digit, Testing database: 40 speakers/digit

DISCUSSION

One of strengths of the HMM approach is to solve the speech recognition and segmentation simultaneously using Viterbi algorithm. The optimum criterion used in the Viterbi algorithm is to maximize the joint probability of incoming acoustic events and state sequence conditioned on a model, the model associated with the maximum joint probability is chosen as recognized output. i.e.

$$\lambda = argmax_{\lambda_i} \{ max_S P(O, S | \lambda_i) \}$$

where $O = \{O_1, ...O_T\}$ and $S = \{S_1, ...S_T\}$ represent acoustic events and state sequence respectively, λ denotes a HMM model.

In the double maximization presented above, the first maximization on the state sequence S is undertaken by the dynamic programming approach, i.e. whatever the first decision is, the following decisions should be optimal with respect to the previous decisions. Because of this sequential decision making mechanism, the temporal structure is fully explored both from forward searching and backward tracking. By mapping acoustic vectors onto states, a static pattern is assumed to be generalized by concatenation of a number of non-stationary vectors and therefore the ANNs could be used to further classify this static pattern. However, the cost function used in the Viterbi decoding is certainly different from the cost function used in the MLP which produces a least squared error between the corrected response and estimated response. This inconsistency could reduce the effectiveness of the combination of HMM with MLP. One of the methods to deal with this problem is to use a generalized probabilistic decent method (Katagiri et al, 1991). Further research is needed in this area.

CONCLUSION

We have presented an approach of combining HMM with MLP into speech recognition system. In this architecture, incoming acoustic feature vectors can be normalized to a fixed size and further decision on the output from HMM is performed by the MLP. A simple experiment of speaker independent isolated digit recognition is carried out to test the effectiveness of this combination. The performance is improved from

92.9% to 93.8% for a simple HMM structure and this HMM combined with a three layer MLP respectively.

REFERENCES:

H. Bourlard, N. Morgan and S. Renale, (1992), *Neural nets and hidden Markov models: Review and generalizations*, Speech Communication, Vol. 11, pp. 237-246, .

J.M. Song and A. Samoulien, (1992), *A robust speaker independent isolated word recognizer over the telephone network based on a modified HMM approach* , elsewhere in this Proceedings.

S.Katagiri, C.H.Lee and B.H.Juang, (1991), *New discriminative training algorithms based on the generalized probilistic decent method*, Proc. *IEEE-SP Workshop on Neural Network for Signal Processing* 91.

M. J. D. Powell, (1979), *Restart Procedures for the Conjugate Gradient Method*, Mathematical Programming, Vol. 12, pp. 241-254..