J. of Electronics (China), 17(3): 242-247, July 2000

THE HIDDEN MARKOV MODEL OF CO-ARTICULATION AND ITS APPLICATION TO THE CONTINUOUS SPEECH RECOGNITION

Tranzai Lee, Fang Zheng, Wenhu Wu, Daowen Chen[*]

(Speech Lab., Dept. Of Computer Sciences and technology, Tsinghua University, Beijing 100084)
[*] (National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080)

Abstract the co-articulation is one of the main reasons that makes the speech recognition difficult. However, the traditional Hidden Markov Models(HMM) can not model the co-articulation, because they depend on the first-order assumption. In this paper, for modeling the co-articulation, we propose a more perfect HMM than traditional first order HMM on the basis of our previous works and they give a method in that this HMM is used in continuous speech recognition by means of multilayer perceptrons (MLP), i.e. the hybrid HMM/MLP method with triple MLP structure. The experiment we conduct shows that this new hybrid HMM/MLP method decreases error rate in comparison with our previous works.

Key words Speech recognition; High-order HMM; Hybrid HMM/MLP.

I. Introduction

There are two variable sequences X_1^T and Q_1^T in the Hidden Markov Model(HMM) of any speech that includes T frames. Where $X_1^T = x_1, x_2, \cdots, x_T$ is the observation sequence and x_i is the feature of i -th frame speech with $i = 1,2,\cdots,T$; $Q_1^T = s_{i_1}, s_{i_2}, \cdots, s_{i_T}$ is the state sequence and $i_j \in \{1,2,\cdots,N\}$ with $j = 1,2,\cdots,T$; $Q = \{s_1,\cdots,s_N\}$ is the set of states in HMM. Traditional HMM used in the speech recognition has defined the relation between X_1^T and Q_1^T by means of the first order assumption[1].Therefore, it is called first-order HMM and its Directed Probabilistic Independence Networks(DPIN[1]) is showed in Fig.1.

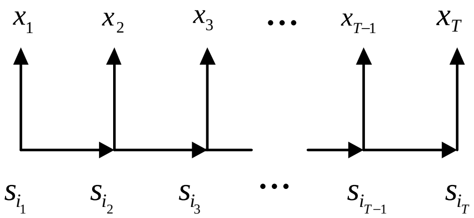

Fig.1 The directed probabilistic independence networks of the first-order HMM.

Because of the first order assumption, HMM and its DPIN are too simple to describe speech. As a result, there are some differences between speech and HMM. For example, first-order HMM assumes that every observation is independent of the other observation and the observation x_t is independent of any states except for state s_{i_t} after $s_{i_{t-1}}$ is given. This is not true because of the existence of co-articulation in speech. In other words, first-order HMM can not model the co-articulation.

In this paper, in order to model the co-articulation, we try to explore new HMMs, i.e., high-order HMMs, that are more perfect than first-order HMM and their application to speech recognition based on our previous works[2,3]. In Section II, some high-order HMMs are given and their traits are analyzed. In Section III, we discuss the application of a high-order HMM in the speech recognition by means of the hybrid HMM/MLP method[4-7]. In Section IV, we conduct a experiment to compare works given in this paper with our previous works[2,3].

II. High-Order HMMs

One of the main reasons that make speech recognition difficult is co-articulation. Therefore, modeling co-articulation becomes very important for speech recognition. It is improper to avoid considering the co-articulation as in the traditional(first-order) HMM. In this part, we try to give a more perfect HMM that can model the co-articulation in some way, i.e., high-order HMM. The co-articulation happens when the articulation vary from a phoneme to next phoneme. It means that a observation is dependent on more states than one in such HMM. In this section, we give a high-order HMM. Our new works are based on our previous works[2,3]. Therefore, at first we briefly review our previous works.

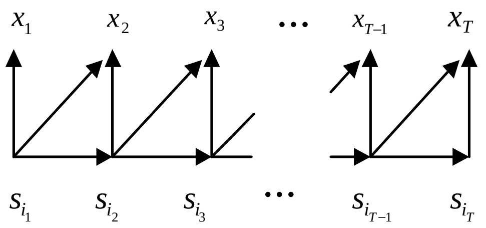

Fig.2 The DPIN of HMM(a) that models the co-articulation of two phonemes. An observation is dependent on two states.

1. Our previous work

In our previous works[2,3], in order to model the co-articulation of two phonemes, we constructed an HMM and its DPIN is showed in Fig.2. For the sake of convenience, this HMM is symbolized into HMM(a). In HMM(a), an observation is dependent on two states. It means that this HMM models the co-articulation of two phonemes.

During recognition, the posterior probabilities $p(Q_1^T|X_1^T)$ must be estimated. However, maximizing $p(Q_1^T|X_1^T)$ is equivalent to maximizing $p(Q_1^T X_1^T)$ in the space of Q_1^T . Therefore, $p(Q_1^T X_1^T)$ have the same capability of discrimination between models as $p(Q_1^T|X_1^T)$. We estimate $p(Q_1^T X_1^T)$ instead of $p(Q_1^T|X_1^T)$ during recognition and we still regard that $p(Q_1^T|X_1^T)$ is estimated. For DPIN, the joint probability distribution $p(u_1,\cdots,u_n)$ can be factorized as follows[1]:

$$p(u_1,\cdots,u_n)=\prod_{i=1}^{n} p(u_i|pa(u_i)) \tag{1}$$

Where $pa(u_i)$ is the set of u_i 's parents with $i=1,2,\cdots,n$. According to Eq(1), using HMM(a), probabilities $p(Q_1^T X_1^T)$ is estimated as follows :

$$p(Q_1^T X_1^T)=\prod_{t=1}^{T} p(s_{i_t}|s_{i_{t-1}})p(x_t|s_{i_{t-1}},s_{i_t}) \tag{2}$$

By means of Bayes rule, Eq(2) is rewritten as :

$$p(Q_1^T X_1^T)=\prod_{t=1}^{T} \frac{p(s_{i_t}|\tilde{s}_{i_{t-1}},x_t)p(\tilde{s}_{i_{t-1}}|x_t)p(x_t)}{p(s_{i_{t-1}})} \tag{3}$$

Where, $\tilde{s}_i$ means that the state is s_i at time $t-1$ when current time is t . In Ref.[2], we use two MLPs to estimate the probabilities that appear in Eq.(3). Thus, the double MLP structure is constructed, i.e., MLP-1, a feedback MLP, estimates $p(s_{i_t}|\tilde{s}_{i_{t-1}},x_t)$ and MLP-2, a feed-forward MLP, estimates $p(\tilde{s}_{i_{t-1}}|x_t)$. Thus, we have gotten the feedback hybrid HMM/MLP method with double MLP structure.

In addition, the information of the contextual dependence plays important role in speech recognition and MLP can easily utilize this information by means of the contextual input that includes $2c+1$ frames[5,6] if it is trained properly. At time t , MLP's contextual input is X_{t-c}^{t+c} , i.e., $x_{t-c}\cdots x_t\cdots x_{t+c}$. As in Ref.[2], X_{t-c}^{t+c} is regarded as the observation instead of x_t at time t . Let $y_t = X_{t-c}^{t+c}$, then $Y_1^T = y_1, y_2,\cdots, y_T$ is the observation sequence. $p(Q_1^T|Y_1^T)$ is as much as $p(Q_1^T|X_1^T)$ provides discriminant information among the word models. Therefore, we will estimate $p(Q_1^T|Y_1^T)$ instead of $p(Q_1^T|X_1^T)$ in this paper, i.e., x_t is replaced with y_t and previous discussions are still tenable.

2 Modeling the co-articulation of 3 phonemes

In HMM(a), the co-articulation of only two phonemes is considered. HMM(a) is still very simple. The HMM that can model the co-articulation of three phonemes will be more perfect than HMM(a) and it means that every observation is dependent on three states. This HMM is symbolized into HMM(b) and its DPIN is showed in Fig.3.

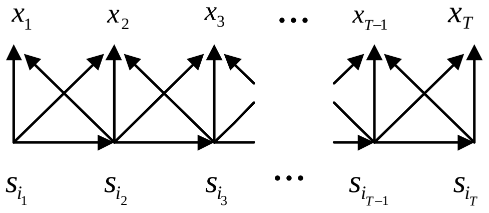

Fig.3 The DPIN of HMM(b) that models the co-articulation of three phonemes. An observation is dependent on three states.

According to Eq(1), using HMM(b), $p(Q_1^T Y_1^T)$ is estimated as follows :

$$p(Q_1^T Y_1^T) = \prod_{t=1}^{T} p(s_{i_{t+1}} | s_{i_t}) p(y_t | \tilde{s}_{i_{t-1}}, s_{i_t}, \bar{s}_{i_{t+1}}) \tag{4}$$

Where $\overline{S}_i$ means that the state is s_i at time $t+1$when current time is$t$. Indeed, Eq.(4) utilizes more information than Eq.(2) for estimating $p(Q_1^T Y_1^T)$. But the cost to estimate $p(y_t | \tilde{s}_{i_{t-1}}, s_{i_t}, \bar{s}_{i_{t+1}})$ appeared in the right hand side of Eq.(4) will be so large that it is beyond the practicability.

To resolve this problem, a feasible method is that HMM(c) (Fig.4) is used instead of HMM(b). HMM(c) utilizes the relation between y_{t-1} and S_{i_t} in different way from HMM(b). During the recognition, the current state of search path (state sequence) must be decided at any time. In previous HMMs, S_{i_t} is dependent only on $S_{i_{t-1}}$ at time t after $S_{i_{t-1}}$ is given. For deciding current state, HMM(c) utilizes the probabilities of S_{i_t} after y_{t-1} and $S_{i_{t-1}}$ are given, i.e., the relation among S_{i_t} , y_{t-1} and $S_{i_{t-1}}$ gotten from speech database statistically.

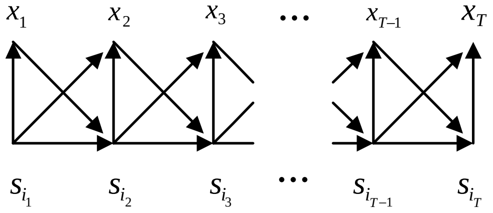

Fig.4 The DPIN of HMM(c) that utilizes the relation between state at time t and observation at time t-1.

According to Eq.(1), using HMM(c), $p(Q_1^T Y_1^T)$ is estimated as follows :

$$p(Q_1^T Y_1^T) = \prod_{t=1}^{T} p(s_{i_t} | \tilde{s}_{i_{t-1}}, y_{t-1}) p(y_t | \tilde{s}_{i_{t-1}}, s_{i_t}) \tag{5}$$

III. The Hybrid HMM/MLP with Triple MLP Structures

As in Ref.[2], by means of Bayes rule, we can get :

$$p(y_t|\tilde{s}_{i_{t-1}}, s_{i_t}) = \frac{p(s_{i_t}|\tilde{s}_{i_{t-1}}, y_t)\,p(y_t|\tilde{s}_{i_{t-1}})}{p(s_{i_t}|s_{i_{t-1}})} \tag{6}$$

$$p(y_t|\tilde{s}_{i_{t-1}}) = \frac{p(s_{i_{t-1}}|y_t)\,p(y_t)}{p(s_{i_{t-1}})} \tag{7}$$

Thus, Eq.(5) can be rewritten as :

$$p(Q_1^T Y_1^T) = \prod_{t=1}^{T} p(s_{i_t}|\tilde{s}_{i_{t-1}}, y_{t-1}) \cdot \frac{p(s_{i_t}|y_t, \tilde{s}_{i_{t-1}})\,p(\tilde{s}_{i_{t-1}}|y_t)\,p(y_t)}{p(s_{i_t}|\tilde{s}_{i_{t-1}})\,p(\tilde{s}_{i_{t-1}})} \tag{8}$$

$p(y_t)$, in the right hand side of Eq.(8), is the same for any state sequence Q_1^T and can be omitted. As in Ref.[2], $p(s_{i_t}|y_t, \tilde{s}_{i_{t-1}})$ and $p(\tilde{s}_{i_{t-1}}|y_t)$ can be estimated using double MLP structure. In order to get $p(s_{i_t}|\tilde{s}_{i_{t-1}}, y_{t-1})$, another MLP is needed. The inputs of this MLP are $\tilde{s}_{i_{t-1}}$ and y_{t-1}. Thus, we construct a network structure that includes three MLPs (Fig.5) for estimating $p(Q_1^T Y_1^T)$ using Eq.(8). We call this network structure *triple MLP structure.*

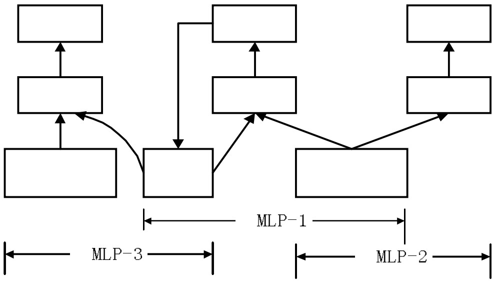

Fig.5 Triple MLP structure. The first layer(the upper layer) is the output layer, the second layer is the hidden layer and the third layer is the input layer. In the input layer, y_t is the contextual input at time t ; $\tilde{s}_{i_{t-1}}$ is the state at time $t-1$ when current time is t .

In Fig.5, MLP-1, MLP-2 and MLP-3 are used to estimate the probabilities $p(s_{i_t}|y_t, \tilde{s}_{i_{t-1}})$, $p(\tilde{s}_{i_{t-1}}|y_t)$ and $p(s_{i_t}|\tilde{s}_{i_{t-1}}, y_{t-1})$, individually. Thus, using triple MLP structure to estimate the probabilities used in Eq.(8), a new hybrid HMM/MLP method is gotten. We call this method *the hybrid HMM/MLP with triple MLP structure.* When $\tilde{s}_{i_{t-1}}$ is the input to MLP-1, $\tilde{s}_{i_{t-1}}$ is the feedback input[5,6]. Now, we call $\tilde{s}_{i_{t-1}}$ the state input when $\tilde{s}_{i_{t-1}}$ is the input to MLP-3. The method of state input is the same as the feedback input.

During recognition, it is possible that any state may appear in any time. Therefore, MLP-2 and MLP-3 must be computed for every possible state as $\tilde{s}_{i_{t-1}}$ when current time is t with $t = 1,2,\cdots,T$. Therefore, the amount of computation is increased about twice in comparison with the hybrid method with double MLP structure and is too large to run in real-time during the recognition. Fortunately, the reduction method of computation proposed in Ref.[3] for the double MLP structure still be effective to triple MLP structure and so is the new feedback method proposed in Ref.[3].

IV. The Experiment Result and Conclusion

In this section, we conduct the experiment for the comparison between the hybrid HMM/MLP method with double MLP structure proposed in Refs.[2,3] and the hybrid HMM/MLP method with triple MLP structure proposed in this paper.

The details of the experiments :

(1) The MLP’s structure: 7 frames of speech features in the contextual input(i.e., c =3), 12 order mel-scale cepstrum and a energy extracted from every frame of speech. Thus, there are 91 units in the contextual input. 200 units in the hidden layer. 89 units associated with Mandarin phonemes and background silence in the output layer. Additional 89 units for the feedback input and the state input in the input layer of the feedback MLP-1 and MLP-3.

(2) Data : 6 times pronunciation of 407 Mandarin syllables (including all Mandarin syllables if intonation is not considered. Note that Mandarin is a toned and single syllable language), 4 times for training, 1 time for cross-validation[5] and 1 time for recognition test.

(3) Training and recognition : MLP-1, MLP-2 and MLP-3 are trained simultaneously. For the hybrid HMM/MLP method with double MLP structure, MLP-3 is not used during the recognition.

The experiment results are showed in Tab.1.

The hybrid methods	syllable error rate
With double MLP structure	9.4%
With triple MLP structure	8.4%

Tab.1 The error rate comparison of the hybrid HMM/MLP method with double MLP structure and the hybrid HMM/MLP method with triple MLP structure.

According to the experiment results in Tab.1, the hybrid method with triple MLP structure reduces error 10.6% in comparison with the hybrid method with double MLP structure. We can come to conclusion that HMM(c) and hybrid HMM/MLP method with triple MLP structure take effect.

Unfortunately, one of the hybrid’s main drawbacks is that training takes too long time, and it is more obvious in the hybrid HMM/MLP method with triple MLP structure.

V. References

[1] P. Smyth, D. Heckerman, M. I. Jordan, “Probabilistic independence networks for hidden Markov probability models,” Neural Computation 9, pp227-267(1997).

[2] Tranzai Lee, Daowen Chen, “The hybrid ANN/HMM method with double MLP structure for continuous speech recognition”, The Fourth International Conference on Neural Information Processing (ICONIP’97), Dunedin, New Zealand, September,1997

[3] Tranzai Lee, Daowen Chen, “New feedback method of hybrid HMM/MLP methods for continuous speech recognition.” ICASSP’98, Seattle, USA, MAY 1998.

[4] H., Bourlard, “Forwards increasing speech recognition error rates”. Speech Communication, 15(2), 205-231, 1996.

[5] H. Bourlard, N. Morgan, “Connectionist speech recognition -- *A Hybrid approach*,” Boston / Dordrecht / London, Kluwer Academic Publishers , 1994.

[6] H. Bourlard, C.J. Wellekens,. “Links between Markov models and multilayer perceptrons”, IEEE Trans. on PAMI, 12(1990) 12, 1167-1178.

[7] N. Morgen, H. Bourlard, “Continuous Speech Recognition”, IEEE Signal Processing Magazine, MAY 1995, 25-42.