NEW FEEDBACK METHOD OF HYBRID HMM/ANN METHODS FOR CONTINUOUS SPEECH RECOGNITION

Tranzai Lee, Daowen Chen

National Lab of Pattern Recognition,Institute of Automation,
Chinese Academy of Sciences, Beijing, P.R. of China
e-mail : lqz , dchen@prldec3.ia.ac.cn

ABSTRACT

In the continuous speech recognition, the co-pronunciation between two successive phonemes seriously disturb recognition effect. It is difficult for pure hidden Markov model(HMM) methods to cope with the co-pronunciation, because HMM methods consider that two successive frames of speech are independant. The hybrid HMM and artificial neural networks(ANN) methods with feedback MLP[1,3] provide the ability to cope with the co-pronunciation by means of the feedback input. In this paper, we propose a new feedback method for feedback hybrid HMM/ANN methods on the basis of the original methods[1,3]. New feedback method provides the more information of co-pronunciation to feedback ANN. As a result, new feedback method falls the error rate 20.4%. Additionally, By means of our previous work, the hybrid mthods HMM/ANN with the feedback double MLP structure, we discuss the method that reduces the computation of the feedback MLP during the recognition.

1. INTRODUCTION

In theory, the hybrid ANN/HMM methods[1,2,3,4,5] have many advantages by comparison with pure HMM methods for the continuous speech recognition. Especially for the hybrid HMM/MLP methods with feedback MLP[1,3], they have the second order Markov model. Therefore, unlike pure HMM methods, the hybrid HMM/MLP mehtods with feedback MLP have made use of the transition information between two sucessive states. If pure HMM also do so, the parameters in HMM will increase enormously and HMM will become intractable.

Pure HMM methods consider that the two successive frames of speech are independent each other. But, it is well known that the co-pronunciation can go on more than 50ms between two successive phonemes. In fact, the transition information means the information of the co-pronunciation. Therefore, the hybrid methods with feedback MLP have utilized the co-pronunciation information. Then, a problem How can the more co-pronunciation information be used can be proposed now. In this paper, we proposed a new feedback method for the hybrid methods with feedback MLP in part 4 on the basis of the original feedback method[1,3]. In order to proof the effect of the new feedback method, our previous work, the hybrid methods with feedback double MLP structure are used.

After we give a brief review to the original hybrid methods in the second part, our new hybrid methods with feedback double MLP structrue and the results of recent experiments are introduced in the third part . In the fourth part, we give our new feedback method for feedback ANN. In the fifth part, by means of new double ANN structure, we give a method and some experiments that reduces the computation of feedback ANN.

2. THE ORIGINAL HYBRID METHOD

In this part, we give a brief review to the original hybrid methods. At the first, we introduce the following symbols that are used in this paper :

$Q=\{s_1,\cdots,s_N\}$: the set of states in HMM;

x_t : the observation at time t , namely, the feature vector of the t -th frame of speech, $1\le t\le T$ and T is the number of the speech's frames .

X_i^j : the observation sequence $\{x_i,\cdots,x_j\}$, $(1\le i<j\le T)$.

Q_1^T : the state sequence $\{s_{i_1},\cdots,s_{i_T}\}$, $s_{i_t}\in Q$, $i=1,\cdots,T$.

MLP that are used in the hybrid methods fall into two types : feedback MLP and feedforward MLP(see [1,3]). They estimate probabilities $p(s_i|\tilde{s}_j, X_{t-c}^{t+c})$ and $p(s_i|X_{t-c}^{t+c})$ individually at time t . Where, X_{t-c}^{t+c} is the contextual input at time t that includes $2c+1$ frames of speech features, with $i,j=1,\cdots,N$. c is a positive interger; $\tilde{s}_j$ means that the state is s_j at time $t-1$ and it is fedback to the input layer[1, 3]. As a result, the hybrid methods also fall into two types : the hybrid methods with feedback MLP and the hybrid methods with feedforward MLP. The probability $p(Q_1^T|X_1^T)$ had to be estimated in the any hybrid methods and different methods vary only in the methods that estimate this probability. Therefore, for the sake of shortness, only methods that estimate this probability are discussed for any hybrid methods in this paper. For the more details, To see [1] and [3]. The methods that the original hybrid methods estimate The probability $p(Q_1^T|X_1^T)$ are the following.

2.1 The Hybrid Methods With Feedback MLP

$$p(Q_1^T \mid X_1^T) = \prod_{t=1}^{T} p(s_{i_t} \mid \tilde{s}_{i_{t-1}}, X_{t-c}^{t+c}) \quad (2.1)$$

$p(s_{i_t} \mid \tilde{s}_{i_{t-1}}, X_{t-c}^{t+c})$, in the right hand side of (2.1), can be gotten from the output of i_t-th unit in the feedback MLP's output layer when the state is $s_{i_{t-1}}$ at time $t-1$ and the feedback MLP's contextual input is X_{t-c}^{t+c} at time t.

2.2 The Hybrid Methods With Feedforward MLP

Because what the feedforward MLP can provides are $p(s_{i_t} \mid X_{t-c}^{t+c})$ and aren't $p(s_{i_t} \mid \tilde{s}_{i_{t-1}}, X_{t-c}^{t+c})$, probability $p(Q_1^T \mid X_1^T)$ is estimated as the following :

$$p(Q_1^T \mid X_1^T) = \prod_{t=1}^{T} p(s_{i_t} \mid X_{t-c}^{t+c}) \quad (2.2)$$

3. NEW HYBRID METHODS

In our previous works, we proposed new hybrid methods that have feedback double MLP structure. In this paper, our new works are based on our previous works. For the sake of clearness, we give the brief review of our previous work in this part.

At time t, MLP's contextual input X_{t-c}^{t+c} can be considered as the observation. It is denoted with y_t. $Y_1^T = y_1, \cdots, y_T$ is the observation sequence. $p(Q_1^T \mid Y_1^T)$ is as much as $p(Q_1^T \mid X_1^T)$ provides discriminant information among the word models. Because maximizing $p(Q_1^T \mid Y_1^T)$ is equivalent to maximizing $p(Q_1^T Y_1^T)$ in the space of Q_1^T, we estimate $p(Q_1^T Y_1^T)$ instead of $p(Q_1^T \mid Y_1^T)$ during recognition. For the second order Markov model, We can get

$$p(Q_1^T Y_1^T) = \prod_{t=1}^{T} p(y_t, s_{i_t} \mid \tilde{s}_{i_{t-1}}) = \prod_{t=1}^{T} p(s_{i_t} \mid \tilde{s}_{i_{t-1}}, y_t) p(y_t \mid \tilde{s}_{i_{t-1}}) \quad (3.1)$$

To compare (2.1) and (3.1), the original hybrid methods with feedback MLP omit $p(y_t \mid \tilde{s}_{i_{t-1}})$. Actually, the original feedback MLP structures cann't provide it.

3.1 Feedback Double MLP Structure

In order to get $p(y_t \mid \tilde{s}_{i_{t-1}})$ that appears in the (3.1), we increase a feedforward MLP in the original feedback MLP structure. This feedforward MLP is denoted with MLP2. Correspondingly, the original feedback MLP is denoted with MLP1. MLP2 has the contextual input in common with MLP1 and the same number of the units as MLP1 in the output layer (Figure 1). The units, in the output layer of MLP2, associate with the HMM states $s_1, \cdots, s_N$, like the units in the output layer of MLP1. For the observation squence Y_1^T and corresponding state squence $Q_1^T = \{s_{i_1}, s_{i_2} \cdots s_{i_T}\}$, the i_{t-1}-th unit's ideal output is 1 and the other are 0 in MLP2's output layer at time t, with $t = 1, \cdots, T$. As training MLP1, MLP2 is trained to minimize the Mean Square Error between its ideal output and its real output. Then, after MLP2 is trained, $g_i^f(y_t)$, the output of the i-th unit in MLP2's output layer when MLP2 is inputed with y_t at time t, is $p(\tilde{s}_i \mid y_t)$, the posterior probabilities that the state is s_i at time $t-1$ under the condition of input y_t at time t, i.e. :

$$g_i^f(y_t) = p(\tilde{s}_i \mid y_t), \text{ with } i = 1, \cdots, N. \quad (3.2)$$

(3.2)'s proof is similar to the proof about MLP2's output(see [1,3]) . By means of Bayes' rule, (3.1) is rewritten as :

$$p(Q_1^T Y_1^T) = \prod_{t=1}^{T} \frac{p(s_{i_t} \mid \tilde{s}_{i_{t-1}}, y_t) p(\tilde{s}_{i_{t-1}} \mid y_t) p(y_t)}{p(s_{i_{t-1}})} \quad (3.3)$$

$p(y_t)$, in the right hand side of (3.3), is the same to any state sequences, therefore, can be omitted. $p(\tilde{s}_{i_{t-1}} \mid y_t)$ can be gotten from the output of i_{t-1}-th unit in the MLP2's output layer, i.e. $g_{i_{t-1}}^f(y_t)$. So far, every item, in the right hand side of (3.1), can be gotten with new MLP structure in Figure 1. We call this method *the hybrid method with feedback double MLP structure.* Obviously, for estimating probabilities $p(Q_1^T Y_1^T)$, (3.3) makes use of more information than (2.1). Therefore, it should performs better then (2.1).

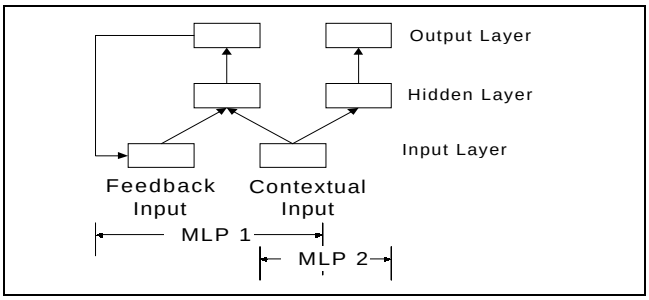

Figure 1. The Feedback Double MLP structure. MLP1 and MLP2 use the common Contextual Input.

3.2 Feedforward Double MLP Structure

Indeed, the major drawback of the hybrid methods with feedback MLP is the high CPU cost and real time is very difficult to be realized during the recognition. To remove the feedback input from the feedback double MLP structure, we get the feedforward double MLP structure and new hybrid methods are derived in sucsession, namely *the hybrid methods with feedforward double MLP structure.* $p(s_{i_t} | \tilde{s}_{i_{t-1}}, y_t)$, in the right hand side of (3.3), is approximated with $p(s_{i_t} | y_t)$, the output of MLP1 that has become a feedforward MLP. Then, according to (3.3), $p(Q_1^T Y_1^T)$ is approximately estimated as following :

$$p(Q_1^T Y_1^T) = \prod_{t=1}^{T} \frac{p(s_{i_t} | y_t) p(\tilde{s}_{i_{t-1}} | y_t)}{p(s_{i_{t-1}})} \qquad (3.4)$$

$p(y_t)$, in the right hand side of (3.3), is omitted.

By the experiments we conducted , we give the comparisons among the following four hybrid methods :

1) The original feedback hybrid method(derived from (2.1));
2) The hybrid method with feedback double MLP structure(derived from (3.3));
3) The original feedforward hybrid method (derived from (2.2));
4) The hybrid method with feedforward double MLP structure(derived from (3.4)).

The details of the experiments :

- The structure of MLP1 and MLP2 : 7 frames of speech features in the contextual input(i.e. c =3). 200 units in the hidden layer. 89 units associated with Mandarin phonemes and background silence in the output layer. Addtional 89 units for the feedback input in the input layer of the feedback MLP1.
- Data : 6 times pronunciation of 1264 Mandarin syllables(including intonation), 4 times for training , 1 time for cross-validation and 1 time for recogntion.
- Recognition : for every pronounciation in the 1264 Mandarin syllables, to judge which syllable's pronounciation of 407 Mandarin syllables(do not including intonation) it is.

The experiment results are listed in Table (1).

Hybrid Method	error rate
1. Original feedback hybrid(2.1)	26.7%
2. New feedback hybrid(3.3)	5.4%
3. Original feedforward hybrid(2.2)	6.3%
4. New feedforward hybrid(3.4)	5.6%

Table (1). *The comparison of four hybrid mehtods*

According to the experiment results, we can reach the following conclusions :

- ♦ To compare to the original hybrid method with feedback MLP, new hybrid method with feedback double MLP structure falls error rate 79.8%.
- ♦ To compare to the original hybrid method with feedforward MLP, new hybrid method with feedforward double MLP structure falls error rate 11.1%. It means that approximation from (3.3) to (3.4) is significent.
- ♦ During the training, new hybrid methods can perform Viterbi alignment better for the segmentation of speeches used in the training. As a result, new hybrid methods can segment speeches more precisely and the training can converges farther.
- ◊ *A problem* that we don't understand : We don't predicted that The error rate of the original feedback hybrid method is higher than that of the original feedforward hybrid method. we wonder whether it means that some problems exist in the original feedback hybrid method.

4. NEW FEEDBACK METHOD IN THE FEEDBACK HYBRID MLP/HMM METHOD

In the previous feedback hybrid methods[1,3], the preceding state, at time $t-1$, is fedback to input layer at time t . But, it is common knowledge that co-pronunciation can goes on several frames between two successive phonemes. Therefore, the original feedback method is improper. We illustrate this problem with following example and give our new feedback method. Suppose 14 frames of speech is the pronunciation from phoneme α to phoneme β, namely, $\alpha\ \alpha\ \alpha\ \tilde{\alpha}\ \tilde{\alpha}\ \tilde{\alpha}\ \tilde{\beta}\ \tilde{\beta}\ \tilde{\beta}\ \tilde{\beta}\ \beta\ \beta\ \beta\ \beta$. where $\tilde{\alpha}$ and $\tilde{\beta}$ is the co-pronunciation of α and β. But, $\tilde{\alpha}$ is more similar to α and $\tilde{\beta}$ is more similar to β. Therefore, $\tilde{\alpha}$ is incorporated into α and $\tilde{\beta}$ into β. According to the original feedback method, at the eighth frame, the precedng state β is fedback to input layer. In fact, from the eighth frame to the ninth frame, pronunciation is greatly influenced by the phoneme α . Thus, the information of co-pronunciation does not tell MLP by feedback input.

Our new feedback method is illustrated as follows. At the first, a threshold λ is gived, for example, λ =3. At current time t (namely, at the t -th frame), if there are the different phonemes(or states) to the current state(state at the time t) within λ frames in front of the t -th frame, the state that is different to the current state and the frame that this state lies at is nearest to current frame is fedback to input layer. Else, current state is fedback to input layer. According to new feedback method, in the previous example, at the eighth frame, because there is a phoneme α within previous λ (=3) frames, α is fedback to input layer. But, at the tenth frame, there is no longer other phoneme except for β within previous λ frames. Therefore, β is fedback to input layer. The new feedback

method can tell MLP the more information of co-pronunciation by the feedback input. Table (2) give the experiment's results in new feedback method. The experiment's details are the same as previous experiments in this paper.

The results in the Table (2) show that new feedback method takes effect to both the original feedback hybrid method and the hybrid method with feedback double MLP for both λ =3 and λ =5 in comparison with the results in the Table (1). But, when λ =3, error rates are lower and fall 34.5% and 20.4%, individually, for two feedback hybrid methods. for two hybrid methods. The optimum λ that makes error rate lowest varies with frame's length and maybe also with the languages. Therefore, we do not attempt to give the optimum λ in this paper.

	error rate(%)	
hybrid method with new feedback method	λ =3	λ =5
1.Original feedback hybrid method	17.5	23.5
2.hybrid method with feedback double MLP	4.3	4.9

Table (2) *the error rate(%) of the hybrid methods with new feedback method for* λ *=3 and* λ =5.

5. THE REDUCTION OF COMPUTATION FOR FEEDBACK HYBRID METHOD

The high CPU cost is the major drawback of the feedback hybrid methods and comes from recurrent computation for every state as a feedback input at every frame. It is fortunate that our new hybrid MLP/HMM method with feedback double MLP structure, given in the part 3, provides a method that reduces the computation of the feedback hybrid method. In fact, of all outputs in the MLP2's output layer, the output of only several units are enough big and the other are too small to take effect. Then, at the first, we select a integer r. Of the r states that have biggest output in the MLP2's output layer, every is fedback to MLP1's input layer and MLP1 is computed individually. For the other states as feedback input, MLP1 are not computed. Thus, the computation is reduced greatly.

By the experiments, we can find optimum r that is smallest and does not increase error rate in comparison with error rate in the fourth row of Table(2)(λ =3). The results of experiments are showed in Table(3). The experiment's details are the same as previous experiments in this paper. According to the results in Table(3), the optimum r is 31. Indeed, the amount of the computation still is too big for r =31. But, when r =10, error rate only slightly increases and still is lower than the error rate of the other methods in the table(1)and table(2).

r	10	11	13	14	15	16	17	31
error rate(%)	5.1	4.9	4.8	4.7	4.6	4.5	4.4	4.3

Table (3) *The relation between* r *and error rate. The optimum* r *is 31 (* λ =3).

By our works mentioned above, the performance of the hybrid methods is enhanced greatly.

6. REFERENCES

[1] Bourlard,H. and Wellekens,C.J.(1990). "Links between Markov models and multilayer perceptrons", IEEE Trans. on PAMI, vol.12,no. 12, pp.1167-1178.

[2] Bourlard,H. "Towards Increasing Speech Recognition Error Rates".

[3] Bourlard,H., Morgan,N.(1994) CONNECTIONIST SPEECH RECOGNITION-- *A Hybrid approach*, Kluwer Academic Publishers

[4] Morgen,N. and Bourlard, H. "Continuous Speech Recognition" ,IEEE Signal Processing Magazine,MAY 1995, pp25-42.

[5] Vincent Fontaine, Christophe Ris, Henri Leich, Johan Vantieghem, Sari Accaino,Dirk Van Compernolle, "Comparison Between Two Hybrid HMM/MLP Approaches in Speech Recognition",