AN NEW METHOD USED IN HMM FOR MODELING FRAME CORRELATION

Guo Qing, Zheng Fang, Wu Jian and Wu Wenhu

Speech Lab., Dept. of Computer Science and Technology
Tsinghua Univ., Beijing, 100084, P.R.China
qguo@sp.cs.tsinghua.edu.cn

ABSTRACT

In this paper we present a novel method to incorporate temporal correlation into a speech recognition system based on conventional hidden Markov model (HMM). In our new model the probability of the current observation not only depends on the current state but also depends on the previous state and the previous observation. The joint conditional PD is approximated by non-linear estimation method. As a result, we can still use mixture Gaussian density to represent the joint conditional PD for the principle of any PD can be approximated by mixture Gaussian density. The HMM incorporated temporal correlation by non-linear estimation method, which we called it FC HMM does not need any additional parameters and it only brings a little additional computing quantity. The results in the experiment show that the top 1 recognition rate of FC HMM has been raised by 6 percent compared to the conventional HMM method.

1. INTRODUCTION

Hidden Markov modeling (HMM) techniques have been used successfully for speech recognition in the last ten years due to their ease of implementation and modeling flexibility. The success or failure of a HMM system relies on how well the models can characterize the nature of real speech. The underlying assumption in this scheme is that speech is quasi-stationary and these stationary parts can be represented by a single state of a HMM. In the traditional HMM algorithms the probability of duration of a state decreases exponentially with time which is not appropriate for representing the temporal structure of speech. With this in mind, a number of attempts have been made to incorporate some additional knowledge into the traditional HMM scheme [1]-[2]. Typical methods of them are incorporating duration information, the inclusion of higher-order feature sets and the use of correlation among neighboring outputs, etc.

Various approaches have been tried to take account of frame correlation for more realistic modeling. M.Ostendorf et al. [3] propose Stochastic Segment Model, which consists of 1) a time warping of the variable–length segment X into a fixed–length segment Y, and 2) a joint density function of the parameters of the resample segment Y (Gaussian density). They think the segment model represents spectral/temporal structure over the entire phoneme. Similarly, V.Digalakis et al. [4] propose Dynamical System Model. All the two methods tries to directly express speech feature trajectories. While they seem to be successful in extracting dynamic cues for speech recognition under a suitable trajectory assumption, they are not based on widely available HMM technology.

In the case of continuous HMM's, a Gaussian probability density function (PDF) assumption is made between adjacent feature vectors in C.J.Wellekens[5] . In P.Kenny[6], a linear prediction technique is used to parameterize frame correlation.

Paliwal [7] incorporated temporal correlation into discrete HMM's by conditioning the probability of the current observation on the current state as well as on the previous observation. With this approach, an output probability distribution (PD) is constructed for each possible pair of state and observation symbols. In their model, $b_{jX_{t-1}}(X_t) = P(X_t \mid X_{t-1}, q_j^t, \lambda)$ is used to replace $b_j(X_t) = P(X_t \mid q_j^t, \lambda)$ of traditional HMM. The number of parameters in this model to be estimated may increase too excessively to get reliable estimation for the output PD's. S.Takahashi [8][9] propose a bigram-constrained (BC) HMM which has solved this problem. The probability of the current observation in BC HMM depends on the current state as well as on the previous observation too. But a BC HMM is obtained by combining a VQ-code bigram and the traditional HMM. So the number of parameters to be estimated in BC HMM is less than the number of the full parameterization method proposed by Paliwal. A remarkable point of BC HMM is that it has provided a method to combine the joint conditional PD by two separate conditional PD. N.S.Kim [10] propose an algorithm based on Extended Logarithmic Pool which can estimate the joint conditional PD more precisely.

2. MODELING FRAME CORRELATION

In traditional HMM (we only discuss first order left-to-right Markov model), we think the probability of the current observation only depends on the current state, while it doesn't depend on the previous state and the previous observation. In this model the probability of the observation vector Y_t given that the current state is q_t is represented as $P(Y_t \mid q_t, \lambda)$ which is characterized by $b_{q_t}(Y_t)$.

In BC HMM proposed by S.Takahashi think the probability of the current observation not only depends on the current state but also depends on the previous observation. It means that the probability of the observation vector Y_t given that the current state is q_t is represented as $P(Y_t \mid Y_{t-1}, q_t, \lambda)$ which is characterized by $b_{q_t Y_{t-1}}(Y_t)$. To actualize the estimation of parameters of the model and the reliability of parameters estimation, BC HMM only need to characterize $P(Y_t \mid Y_{t-1})$ and

$b_{q_t}(Y_t)$. Then $b_{q_tY_{t-1}}(Y_t)$ is computed by combining them. At last $b_{q_t}(Y_t)$ is replaced by $b_{q_tY_{t-1}}(Y_t)$ which is used for speech recognition. So BC HMM can avoid the problem caused by the full parameterization of Paliwal.

However, the topology shown in figure 1 seems can reflect frame correlation more precisely which means that the probability of the current observation not only depends on the current state but also depends on the previous state and the previous observation. Then the probability of the observation vector Y_t given that the current state is q_t is represented as $P(Y_t \mid Y_{t-1}, q_{t-1}, q_t, \lambda)$ which is characterized by $b_{q_{t-1}Y_{t-1}q_t}(Y_t)$ (q_{t-1} is the state in $t-1$). The same as the model of Paliwal that using limited train data to full parameterize this model is nearly impossible. So we need to find an approximate arithmetic to compute $b_{q_{t-1}Y_{t-1}q_t}(Y_t)$.

Further, we can adopt the first-order forward and backward frame correlation model shown in figure 2. Then the probability of the observation vector Y_t given that the current state is q_t is represented as $P(Y_t \mid Y_{t-1}, Y_{t+1}, q_t, q_{t-1}, q_{t+1}, \lambda)$ which is characterized by $b_{q_tY_{t-1}q_{t-1}Y_{t+1}q_{t+1}}(Y_t)$.

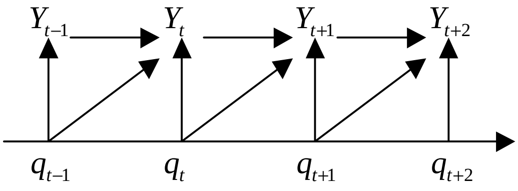

Figure 1

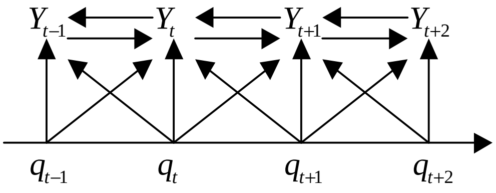

Figure 2

Now we discuss how to estimate $p(Y_t \mid Y_{t-1}, q_t, q_{t-1}, \lambda)$ using non-linear formula.

$$\begin{aligned} p(Y_t \mid Y_{t-1}, q_t, q_{t-1}, \lambda) &= \frac{p(Y_t, Y_{t-1}, q_t, q_{t-1} \mid \lambda)}{p(Y_{t-1}, q_t, q_{t-1} \mid \lambda)} \\ &= \frac{p(Y_t \mid q_t, \lambda)\, p(q_t \mid \lambda)\, p(Y_{t-1}, q_{t-1} \mid Y_t, q_t, \lambda)}{p(Y_{t-1}, q_{t-1} \mid q_t, \lambda)\, p(q_t \mid \lambda)} \\ &= \frac{p(Y_{t-1}, q_{t-1} \mid Y_t, q_t, \lambda)}{p(Y_{t-1}, q_{t-1} \mid q_t, \lambda)}\, p(Y_t \mid q_t, \lambda) \end{aligned} \qquad (1)$$

Let $f(Y_{t-1}, Y_t, q_{t-1}, q_t, \lambda) = \dfrac{p(Y_{t-1}, q_{t-1} \mid Y_t, q_t, \lambda)}{p(Y_{t-1}, q_{t-1} \mid q_t, \lambda)}$ (2)

Then

$$p(Y_t \mid Y_{t-1}, q_t, q_{t-1}, \lambda) = f(Y_{t-1}, Y_t, q_{t-1}, q_t, \lambda)\, p(Y_t \mid q_t, \lambda) \qquad (3)$$

Further, we approximate $p(Y_{t-1}, q_{t-1} \mid q_t, \lambda)$ of right term of (2) by $p(Y_{t-1}, q_{t-1} \mid \lambda)$. We can get:

$$f(Y_{t-1}, Y_t, q_{t-1}, q_t, \lambda) \approx \frac{p(Y_{t-1}, q_{t-1} \mid Y_t, q_t, \lambda)}{p(Y_{t-1}, q_{t-1} \mid \lambda)} \qquad (4)$$

Then, we use a non-linear estimation formula to compute the right term of the above formula, i.e.:

$$\frac{p(Y_{t-1}, q_{t-1} \mid Y_t, q_t, \lambda)}{p(Y_{t-1}, q_{t-1} \mid \lambda)} \approx h(b_{q_{t-1}}(Y_{t-1}), b_{q_t}(Y_t)) \qquad (5)$$

At last, we obtain the non-linear estimation formula of $p(Y_t \mid Y_{t-1}, q_t, q_{t-1}, \lambda)$:

$$\begin{aligned} p^*(Y_t \mid Y_{t-1}, q_t, q_{t-1}, \lambda) &= f(Y_{t-1}, Y_t, q_{t-1}, q_t, \lambda)\, p(Y_t \mid q_t, \lambda) \\ &\approx h(b_{q_{t-1}}(Y_{t-1}), b_{q_t}(Y_t))\, p(Y_t \mid q_t, \lambda) \end{aligned} \qquad (6)$$

As above, we can estimate $P(Y_t \mid Y_{t-1}, Y_{t+1}, q_t, q_{t-1}, q_{t+1}, \lambda)$ using non-linear probability estimation too.

$$\begin{aligned} p(Y_t \mid Y_{t-1}, Y_{t+1}, q_t, q_{t-1}, q_{t+1}, \lambda) &= \frac{p(Y_t, Y_{t-1}, Y_{t+1}, q_t, q_{t-1}, q_{t+1} \mid \lambda)}{p(Y_{t-1}, Y_{t+1}, q_t, q_{t-1}, q_{t+1} \mid \lambda)} \\ &= \frac{p(Y_t \mid q_t, \lambda)\, p(q_t \mid \lambda)\, p(Y_{t-1}, Y_{t+1}, q_{t-1}, q_{t+1} \mid Y_t, q_t, \lambda)}{p(Y_{t-1}, Y_{t+1}, q_{t-1}, q_{t+1} \mid q_t, \lambda)\, p(q_t \mid \lambda)} \\ &= \frac{p(Y_{t-1}, Y_{t+1}, q_{t-1}, q_{t+1} \mid Y_t, q_t, \lambda)}{p(Y_{t-1}, Y_{t+1}, q_{t-1}, q_{t+1} \mid q_t, \lambda)}\, p(Y_t \mid q_t, \lambda) \\ &\approx \frac{p(Y_{t-1}, Y_{t+1}, q_{t-1}, q_{t+1} \mid Y_t, q_t, \lambda)}{p(Y_{t-1}, Y_{t+1}, q_{t-1}, q_{t+1} \mid \lambda)}\, p(Y_t \mid q_t, \lambda) \\ &\approx h(b_{q_{t-1}}(Y_{t-1}), b_{q_t}(Y_t), b_{q_{t+1}}(Y_{t+1}))\, p(Y_t \mid q_t, \lambda) \end{aligned} \qquad (7)$$

3. FRAME CORRELATION (FC) HMM

In this section, we use the concept of non-linear estimation to incorporate the correlation of neighboring frames into the traditional HMM. For simplicity, we only take the case of first-order forward frame correlation, which means that the current observation symbol relates only with the observation and the state on the immediate previous frame as figure 2 shows.

FC HMM $\lambda = (N, \pi, A, B, FC)$ which incorporate frame correlation can be defined as follows:

1) N, the number of states in the model;

2) $\pi = \{\pi_i\}$, where $\pi_i = P[q_1 = i], 1 \le i \le N$ is the initial probability of the model being in state i, and they satisfy the constraint $\sum_{i=1}^{N} \pi_i = 1$.

3) $A = \{a_{ij}\}, 1 \le i, j \le N$, $a_{ij} = P[q_{t+1} = j \mid q_t = i]$

4) $B = b_i(O)$ is the probability density function (pdf) for the observation O given that the state is i.

In our system we adopt a state observation density, $b_i(O)$, of the form, $b_i(O) = \sum_{m=1}^{M} c_{mi} N[O, \mu_{mi}, U_{mi}]$

i.e., a continuous mixture density where O is the observation vector (e.g., cepstral coefficient vector resulting from the LPC analysis), c_{mi} is the mixture weight for the mth component in state i, μ_{mi} is the mean vector for mixture m in state i, and U_{mi} is the covariance matrix for mixture m in state i.

5) The frame correlation PD， $FC = \{ f(Y_{t-1}, Y_t, q_{t-1}, q_t) \}$, in which:

We approximate $f(Y_{t-1}, Y_t, q_{t-1}, q_t)$ by a non-linear estimation formula $h(b_j(Y_{t-1}), b_{q_t}(Y_t))$.

In FC HMM, $P[O_1 O_2 \cdots O_T \mid \lambda]$ can be computed as follows:

$$\begin{aligned} P[O_1 O_2 \cdots O_T \mid \lambda] &= \sum_{\psi=(q_1,q_2,\cdots,q_T)} P[O_1 O_2 \cdots O_T, \psi \mid \lambda \\ &= \sum_{\psi} \left[P(\psi \mid \lambda) P(O_1 O_2 \cdots O_T, \psi \mid \lambda) \right] \\ &= \sum_{\psi} \left[P(\psi \mid \lambda) \prod_{t=1}^{T} p^*(O_t \mid O_{t-1}, q_{t-1}, q_t, \lambda) \right] \\ &= \sum_{\psi} \left[P(\psi \mid \lambda) \prod_{t=1}^{T} f(O_t, O_{t-1}, q_{t-1}, q_t) b_{q_t}(O_t) \right] \\ &\approx \sum_{\psi} \left[P(\psi \mid \lambda) \prod_{t=1}^{T} h(b_{q_{t-1}}(O_{t-1}), b_{q_t}(O_t)) b_{q_t}(O_t) \right] \end{aligned} \tag{8}$$

While using the forward-backward formula to reestimate parameters, $\alpha_t(j)$ is modified as follows:

$$\begin{aligned} \alpha_t(j) &= \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} h(b_i(O_{t-1}), b_j(O_t)) b_j(O_t) \\ &= \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_{iO_{t-1}j}(O_t) \end{aligned} \tag{9}$$

$$\begin{aligned} \beta_t(j) &= \sum_{i=1}^{N} \beta_{t+1}(i) a_{ji} h(b_j(O_t), b_i(O_{t+1})) b_i(O_{t+1}) \\ &= \sum_{i=1}^{N} \beta_{t+1}(i) a_{ji} b_{jO_{t-1}i}(O_{t+1}) \end{aligned} \tag{10}$$

Considering the principle that using mixture Gaussian density can approach any PD, we think that the reestimated matrix B characterize the probability density function incorporated frame correlation, i.e. $p(Y_t \mid Y_{t-1}, q_t, q_{t-1}, \lambda)$.

4. COMPLEXITY ANZLYZING

In this section, we analyze the computing complexity and memory complexity of FC HMM incorporated frame correlation by non-linear estimation. Obviously, the model exploits the principle of that using mixture Gaussian density can approach any PD, so $p^*(Y_t \mid Y_{t-1}, q_{t-1}, q_t, \lambda)$ can be characterized by the weighted sum of M normal distribution. As a result, the model does not bring any memory complexity.

Whenever the training of the model or the recognizing of the model, FC HMM only need to compute $f(Y_{t-1}, Y_t, q_{t-1}, q_t)$ additionally. While the non-linear estimation formula $h(b_j(Y_{t-1}), b_{q_t}(Y_t))$ is used to approximate $f(Y_{t-1}, Y_t, q_{t-1}, q_t)$, only very limited addition and multiplication is added so FC HMM only needs a little more computing complexity than the traditional HMM.

In the model proposed by Paliwal the number of the parameters of B matrix is up to $M^2 N$ which is M times than that of the traditional HMM. BC HMM needs to compute $p(Y_t \mid Y_{t-1})$ in the training of the model so it needs to estimate $M^2 T$ parameters additionally. In addition, when recognizing BC HMM needs to compute $p(Y_t \mid Y_{t-1})$ additionally and especially it needs adjusting the weights of mixture Gaussian density so the computing complexity of speech recognizing is augmented greatly.

5. RECOGNITION RESULTS

5.1 Speech Database and Features

The speech database used in experiment is “863 assessment” male speech database. The database consists of 1560 sentences which is parted to three groups as A, B and C. The number of sentences of each group is nearly equal. 38 male each utters one part of sentences. The entire database is parted to the training set, the testing set 1 and the testing set 2. The Testing-Set 1 is the utterance of untrained-speaker. The Testing-Set 2 is the untrained-utterance of trained-speaker.

In experiment, we adopt a five state first order left-to-right Markov model. The output probability of observations in each state is characterized by 5-mixture Gaussian density. 16-dimension cepstral coefficients derived by LPC analysis are used as features of each frame.

5.2 Comparison of FC HMM and THMM

The recognition rates of FC HMM and THMM are shown in Table 1. From which we can see that whether in Training-Set or Testing-Set recognition effect of FC HMM are both better than that of THMM. To Training-Set, the Top1 recognition rate of FC HMM is 6 percent higher than that of THMM. To Testing-Set 2, i.e. testing utterances of trained-speaker, FC HMM is 4 percent higher than THMM. To Testing-Set 1, i.e. utterances of untrained-speaker, FC HMM is 3 percent higher than THMM.

Model	Recognition Set	Top1	Top2	Top5	Top10
HMM	Train-Set	59.93	76.29	90.12	95.05
	Test-Set 1	32.50	48.29	70.05	82.98
	Test-Set 2	41.22	58.96	80.47	89.88
FC-HMM	Train-Set	66.01	81.10	92.70	96.59
	Test-Set 1	35.78	51.53	72.02	84.42
	Test-Set 2	45.63	63.50	82.80	91.23

Table 1: The recognition rate of FC HMM

From the results we can see that the FC HMM which characterizes frame correlation by using non-linear estimation formula brings a significant improvement than the traditional HMM. While FC HMM does not bring any more memory complexity than THMM and it just brings a little more additional computing quantity than THMM so we can say FC HMM is an efficient method to model frame correlation in HMM.

6. SUMMARY

In this paper we present a novel method to incorporate temporal correlation into a speech recognition system based on conventional hidden Markov models (HMM's). At first, we use the joint conditional probability to represent frame correlation. Then, we use a non-linear probability estimation formula to characterize the correlation of adjacent frames. The methods reported before bring either a large increase of the model parameters or a lot of additional computing quantity. The FC HMM reported in this paper does not bring any more memory complexity and it just brings a little more additional computing quantity so we can say FC HMM is an efficient method to model frame correlation in HMM. Another advantage of the method is that it can be easily incorporated into HMM which we have already had. How to more precisely non-linear estimate the first-order forward frame correlation and how to use the method in high-order forward and backward frame correlation is needed to furthermore researching.

7. REFERENCES

[1] Zhi-Ping Hu and Satoshi Imai, Modeling Improvement of the Continuous Hidden Markov Model for Speech Recognition. Proc. Int. Conf. Acoustics, Speech, and Signal Processing, pp.373-376, 1992,

[2] Padma Ramesh and Jay G. Wilpon, Modeling State Durations in Hidden Markov Models for Automatic Speech Recognition. Proc. Int. Conf. Acoustics, Speech, and Signal Processing, pp.381-384, 1992.

[3] M. Ostendorf and S.Roukos, A stochastic segment model for phoneme-based continuous speech recognition. IEEE Trans. On Acoustics, Speech and Signal Processing, pp. 1857-1869, 1989.

[4] V. Digalakis, J. R. Rohlicek and M.Ostendorf, A dynamical system approch to continuous Speech Recognition. Proc. Int. Conf. Acoustics, Speech, and Signal Processing, pp.289-292, 1991.

[5] C. J. Wellekens, Explicit correlation in hidden Markov model for Speech Recognition. Proc. Int. Conf. Acoustics, Speech, and Signal Processing, pp.383-386, 1987.

[6] P. Kenny, M. Lennig and P. Mermelstein, A linear predictive HMM for vector-valued observations with applications to speech recognition. IEEE Trans. On Acoustics, Speech and Signal Processing, pp. 220-225, 1990.

[7] K. K. Paliwal, Use of temporal correlation between successive frames in hidden Markov model based Speech recognizer. Proc. Int. Conf. Acoustics, Speech, and Signal Processing, pp.215-218, 1993.

[8] S. Takahashi, Phonemic HMM constrained by statistical VQ-code transition. Proc. Int. Conf. Acoustics, Speech, and Signal Processing, pp. 553-556, 1992.

[9] S. Takahashi, Phoneme HMM's constrained by frame correlations. Proc. Int. Conf. Acoustics, Speech, and Signal Processing, pp. 219-222, 1993.

[10] Nam Soo Kim and Chong Kwan Un, Frame-correlated hidden Markov model based on extended logarithmic pool. IEEE Trans. On Acoustics, Speech and Signal Processing, pp. 149-160, 1997