Hidden Markov Model with Nonstationary states

Mohamed Ali MAHJOU, Noureddine ELLOUZE
LSTS-ENIT, le belvedère, BP 37, 1002 Tunis
Tél. : (216 1) 87 47 00   Fax : (216 1) 87 27 29

ABSTRACT

In this paper we explore the nonstationarity of Markov model and we propose a nonstationary Hidden Markov Model (NSHMM) which is defined with a set of dynamic transitions probability parameters $A(t)=\{a_{ij}(t)\}$ that depend on the time t already spent in state i. When compared to traditional models, this NSHMM is defined as generalization of the DHMM. The model was applied to on line recognition of handwritten arabic characters. The characters are represented by a radial sequence which is independent of translation, orientation and homothetie. The complete symbol-generation procedure includes sampling, size normalization and quantization phases. In the training, the model parameters are estimated with Baum-welch algorithm from a set of characters. Experiments have been conducted, and a good classification score has been obtained. The discrimination between characters models has been improved. Experiments showed that this approach can better capture the dynamic nature of arabic script.

Keywords : Hidden Markov Model, state duration, Nonstationary HMM, on line arabic character recognition

1. INTRODUCTION

Hidden Markov model is a double layer-stochastic process, composed of hidden layer that controls the time evolution of an observable layer. An HMM has N distincts states. Each state is uniquely defined by an output probability density, that provides a likehood for a given vector generated by the state. The transition from a state is governed by the state transition probabilities and influenced by the current observation vector. The state observation and transition probabilities provide a probablistic mechanism for association of a time sequence of vectors with a given HMM model denoted $\lambda = (A, \Pi, B)$. However, the standard HMM is useful only for stationary signal sources. The real world signals such as arabic scripts, are highly dynamical. Thus, approximating non stationary signals by a stationary models isn't adequate.

Experiments demonstrate that modelling variable duration can improve the recognition rate.

In a traditionnal HMM, with no explicit duration modelling, the probability of state duration for d consecutive observations $p_i(d)$ associated with state S_i, with self transition coefficient a_{ii}, was exponential expression :

$$p_i(d) = a_{ii}^{d-1}(1-a_{ii})$$

which means : $p_i(d+1) = a_{ij} * p_i(d)$. For most physical signals, this exponential density of duration is not always appropriate. The standard HMM model can be improved by the introduction of duration probabilities. Furguson [Furguson 80] first proposed a model with explicit time duration of states in HMM. The main idea is to replace self-transition probability a_{ii} by an explicit duration density $p_i(d)$. In this case, transition occurs after an appropriate number of observations in the state, where the number is controlled by the duration density.

Gu and Tseng [Tseng 91] proposed a bounded state duration HMM. The state durations are simply bounded by two parameters in the recognition phase. These parameters are estimated from training data.

The majority of efforts, however, are based on the formulation of an autoregressif process for observation vectors as in the autoregressif HMM [Rabiner 85] and its several variants generalized by Deng [Deng 93].

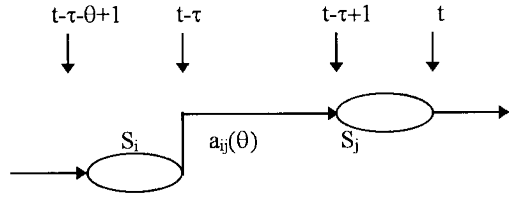
Recently Githza and Sondhi [Ghitza 93] proposed the use of dynamically time-warped templates in each nonstationary state corresponding to a diphone in speech recognition. This model can be regarded as a combination of the HMM and the stochastic segment model where dynamic time-warping replaces linear warping.

Vaseghi proposed a non stationary model for speech recognition [Vaseghi 95]. The transition probabilities are derived from the cumulative density function of state duration. The training of HMMs with duration dependent transitions are based on maximum likehood segmentation of training data using the Viterbi algorithm.

The purpose of this paper is to describe a Hidden Markov Model with nonstationary states for arabic character recognition. The description of the model is presented in section 2, the proposed nonstationary Hidden Markov Model without self transitions extends the description of state duration by replacing the stationary matrix A with a set of dynamic transition probability parameters $A(t)=\{a_{ij}(t)\}$ and matrix of state duration. The properties and the evaluation of the model are detailed in sections 3 and 4, including the initial parameters estimation. The last section concerns the application of the NSHMM to the on line recognition of handwritten arabic characters. Experimental results are analysed to evaluate the performance of character recognition.

2. DESCRIPTION OF THE MODEL

In both standard HMM and DHMM, the transition probabilities $\{a_{ij}\}$ are constant over time. Therefore the Markov chain is discribed as stationary in the strict sense. In contrast, in a NSHMM, the transition process takes account of time duration in a state, and the probability transition to next state is determined based on the duration of current state (figure 2.1). Considering the following example :



the generation of the symbols is as follows :

1- We enter state S_i between $t\text{-}\tau\text{-}\theta$ and $t\text{-}\tau\text{-}\theta+1$.

2- A duration θ is choosen according to the state duration density $p_q(\theta)$. The observations $O_{t-\tau-\theta+1},\ldots,O_{t-\tau}$ are choosen according the joint probability $b_{q1}(O_{t-\tau-\theta+1},\ldots,O_{t-\tau})$. Meanwhile, the observations $O_{t-\tau-\theta+1},\ldots,O_{t-\tau}$ are assumed to be independent, so that we have $b_i(O_{t-\tau-\theta+1},\dots,O_{t-\tau}) = \prod_{d=t-\tau-\theta+1}^{t-\tau} b_i(O_d)$

3- the next state S_j is choosen according to the state transition probabilities $a_{ij}(\theta)$, (same as DHMM no transition back to the same state can occur)

5- similary to step 2, a duration τ is choosen according to the state duration density $p_q(\tau\ )$. The observations $O_{t-\tau+1},\dots,O_t$ are choosen according the joint probability $b_{q1}(O_{t-\tau+1},\dots,O_t)$.

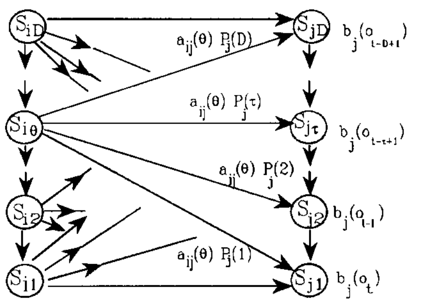


Figure 2.1 : State transition

3. EVALUATION OF THE MODEL

Based on the previous desription of NSHMM process, for a give a sequence of observations, we extend the standard definition of the forward variable α and the backward variable β as follows :

$\alpha_{t:d}(j) = P(o_1,\dots,o_t,\ q_l = S_j\ ,\ t-d+1 \le l \le t \mid \lambda)$

$\beta_{t:d}(i) = P(o_{t+1},\dots,o_T \mid q_l = S_i\ ,\ t-d+1 \le l \le t,\ \lambda)$

then by induction

$\alpha_{t:d}(j)$ =

$$\sum_{\substack{i=1\\ i\neq j}}^{N} \sum_{\delta=1}^{D} \alpha_{t-d:\delta}(i) a_{ij}(\delta) P_j(d) \prod_{s=t-d+1}^{t} b_j(O_s) \qquad (1)$$

Where D is the maximum duration within any state.

To initiate the computation of $\alpha_{t:d}(j)$ we use :

$\alpha_{1:1}(j) = \pi_i\ P_i(1) b_j(o_1)$

thus :

$$\alpha_{2:1}(j) = \sum_{i\neq j}^{N} \alpha_{1:1}(i)\, a_{ij}(1)\, P_j(1)\, b_j(o_2)$$

$$\alpha_{3:1}(j) = \sum_{i\neq j}^{N} \sum_{\delta=1}^{2} \alpha_{2:\delta}(i)\, a_{ij}(\delta)\, P_j(1)\, b_j(o_3)$$

$$\alpha_{3:2}(j) = \sum_{i\neq j}^{N} \sum_{\delta=1}^{1} \alpha_{1:\delta}(i)\, a_{ij}(\delta)\, P_j(2) \prod_{l=1}^{3} b_j(o_{4-d+1})$$

The total computation in eq. (1) which is performed for t=1,...,T, d=1,...,D and j=1,...,N requires about $N^2.T.D^3$ calculations.

Backward variable is similary written inductively:

$\beta_{t:d}(i)$ =

$$\sum_{\substack{j=1\\ j\neq i}}^{N} \sum_{d=1}^{D} a_{ij}(\delta)\beta_{t+d:d}(j) P_j(d) \prod_{s=t+1}^{t+d} b_j(O_s) \qquad (2)$$

To initiate the computation of $\beta_{t:d}(j)$ we use :

$\beta_{T-1:1}\ (i) = 1$

$\beta_{T-2:2}\ (i) = 1$

.....

$\beta_{T-D:D}\ (i) = 1$

Like eq. (1), the total computation in eq. (2) which is performed for t=1,...,T, d=1,...,D and j=1,...,N requires about $N^2.T.D^3$ calculations.

Thus, the number of parameters in NSHMM increases. The storage for the state transition parameters of the NSHMM is N^2D , D times the storage for the DHMM.

It should be clear that the desired probability of O given the model can be written in terms of α 's as :

$$P(O|\lambda) = \sum_{i=1}^{N} \sum_{\tau=1}^{D} \alpha_{T:\tau}(i) = \sum_{i=1}^{N} \alpha_T(i)$$

4. PARAMETERS ESTIMATION

Maximization of $P(O|\lambda)$ over the paramter set is a problem of constrained optimization, where :

$$P(O|\lambda) = \sum_{t=1}^{T} \sum_{j}^{N} \sum_{d}^{D} \sum_{\substack{i=1\\ i\neq j}}^{N} \sum_{\delta=1}^{D} \alpha_{t-d:\delta}(i) a_{ij}(\delta) P_j(d) \prod_{s=t-d+1}^{t} b_j(O_s) \beta_{t:d}(j)$$

With respect to the constraints :

$$\sum_{j=1}^{N} a_{ij}(\tau) = 1 \quad i = 1,\dots,N \ \ and\ \ \tau = 1,\dots,D$$

$$\sum_{k=1}^{M} b_j(k) = 1$$

$$\sum_{j=1}^{N} \pi_i =$$

$$\sum_{d=1}^{D} p_i(d) = 1$$

To solve this problem, we use the iterative procedure of Baum-Welch. In order to describe this procedure, we first define the probability of being in state S_i, at time t- τ for a δ duration, and state S_j at time t for a τ duration given the observation sequence and the model:

$$\xi_t(i,j,\delta,\tau) = P(q_{t-\delta-\tau+1} = s_i, q_{t-\delta-\tau+2} = s_i, \dots, q_{t-\tau} = s_i, q_{t-\tau+1} = s_j, \dots, q_t = s_j,\ q_{t+1} \neq s_j | O,\lambda)$$

$$= \frac{\alpha_{t-\tau:\delta}(i) * a_{ij}(\delta) * P_j(\tau) * \prod_{l=1}^{\tau} b_j(o_{t-\tau+l}) * \beta_{t:\tau}(j)}{P(O|\lambda)}$$

Where the numerator term is just :

$$P(q_{t-\delta-\tau+1} = s_i, q_{t-\delta-\tau+2} = s_i, \dots, q_{t-\tau} = s_i, q_{t-\tau+1} = s_j, \dots, q_t = s_j,\ q_{t+1} \neq s_j, O|\lambda)$$

and the division by $P(O|\lambda)$ gives the desired probability.

We define also the probability of being in state S_j for τ duration:

$$\gamma_t(j,\tau) = \sum_{\delta=1}^{D} \sum_{i=1, i\neq j}^{N} \xi_\tau(i,j,\delta,\tau)$$

Thus, the different parameters of the model can be written as follows :

- First state probability

$$\overline{\Pi}_j = \frac{\Pi_j * \sum_{\tau} P_j(\tau) * \prod_{l=1}^{\tau} b_j(l) * \beta_{\tau:\tau}(j)}{P(O|\lambda)}$$

- Transition matrix probabilities

$$\overline{a}_{ij}(\delta) = \frac{\sum_{t}\sum_{\tau} \xi_t(i,j,\delta,\tau)}{\sum_{t}\sum_{\tau}\sum_{k} \xi_t(i,k,\delta,\tau,)}$$

$$\overline{a}_{ij}(\delta) = \frac{\sum_{t}\sum_{\tau} \alpha_{t-\tau:\delta}(i) * a_{ij}(\delta) * P_j(\tau) * \prod_{l=1}^{\tau} b_j(o_{t-\tau+1}) * \beta_{t:\tau}(j)}{\sum_{t}\sum_{\tau}\sum_{k} \alpha_{t-\tau:\delta}(i) * a_{ik}(\delta) * P_k(\tau) * \prod_{l=1}^{\tau} b_k(o_{t-\tau+1}) * \beta_{t:\tau}(k)}$$

- Matrix of state duration

$$\overline{P}_j(\tau) = \frac{\sum_{t} \gamma_t(j,\tau)}{\sum_{t}\sum_{\tau} \gamma_t(j,\tau)}$$

$$\overline{P}_j(\tau) = \frac{\sum_{t}\sum_{i \neq j}\sum_{\delta} \alpha_{t-\tau:\delta}(i) * a_{ij}(\delta) * P_j(\tau) * \prod_{l=1}^{\tau} b_j(o_{t-\tau+1}) * \beta_{t:\tau}(j)}{\sum_{t}\sum_{i \neq j}\sum_{\delta}\sum_{\tau} \alpha_{t-\tau:\delta}(i) * a_{ij}(\delta) * P_j(\tau) * \prod_{l=1}^{\tau} b_j(o_{t-\tau+1}) * \beta_{t:\tau}(j)}$$

- Matrix of symbol generation

$$\overline{b}_j(k) = \frac{\sum_{t}\sum_{i \neq j}\sum_{\delta}\sum_{\tau} \sum_{\substack{\theta \\ o_{t-\tau+\theta} = o_k}} \xi_t(i,j,\delta,\tau)}{\sum_{t}\sum_{i \neq j}\sum_{\delta}\sum_{\tau} \sum_{\theta} \xi_t(i,j,\delta,\tau)}$$

$$\overline{b}_j(k) = \frac{\sum_{t}\sum_{i \neq j}\sum_{\delta}\sum_{\tau} \sum_{\substack{\theta=1 \\ o_{t-\tau+\theta} = o_k}}^{D} \alpha_{t-\tau:\delta}(i) * a_{ij}(\delta) * P_j(\tau) * \prod_{l=1}^{\tau} b_j(o_{t-\tau+1}) * \beta_{t:\tau}(j)}{\sum_{t}\sum_{i \neq j}\sum_{\delta}\sum_{\tau} \sum_{\theta=1}^{D} \alpha_{t-\tau:\delta}(i) * a_{ij}(\delta) * P_j(\tau) * \prod_{l=1}^{\tau} b_j(o_{t-\tau+1}) * \beta_{t:\tau}(j)}$$

5. PROPERTIES OF THE MODEL

Like in the DHMM, the duration P in state s_i is estimated by:

$$p_i(d) = \frac{\textit{expected number of d long sejour in state } s_i}{\textit{expected number of transitions from state } s_i}$$

The difference between the DHMM and the NSHMM is made by the way the states transition parameters are estimated. In the DHMM the transition probabilities $\{a_{ij}\}$, are constant over time. Therefore the Markov model is described as stationary. In the proposed model, however, state transition probabilities are expressed conditional on how long their current state has been occupied. When matrix A is stationary, the model is reduced to the DHMM.
On the other hand, it can be seen from the definition and the initialization conditions on the forward-backward variables that there are much time calculation and more parameters number associated with state. Thus the reestimation problem is more difficult than both standard HMM and DHMM. To alleviate some of these problem, we can use, like in the DHMM, a parametric state duration density instead of the non-parametric $p_i(d)$. In particular, gaussian family and gamma family can be used [Mahjoub 97].

6. APPLICATION TO THE ON-LINE RECOGNITION OF ARABIC CHARACTERS

Our research to explore the nonstationarity of Markov model is motivated by the desire to develop an on line recognition characters.
There are many uncertainties in handwritten character recognition. Stochastic modeling is a flexible and general method for modelling such problems. It entails the use of probabilistic models to deal with uncertain or incomplete forms. In on-line handwritten character recognition, uncertainty and incompleteness are very frequent. Thus, HMMs have been applied with some success to handwritten script recognition.
We adress in this part the problem of the automatic recognition of arabic handwritten characters. Pen-based computer devices capture handwriting through a special pen attached to an electronic tablet. The digitized data recorded on the tablet are sent to a recognizer which outputs the corresponding typeset sequence of characters. The data capturated include the temporal information in the handwriting, and recognition is done on-line.

6.1 Symbol generation

representation of a 2-D shape with one feature vector includes two aspects of data manipulation. First, 2-D shape is transformed into a 1-D signal sequence. Then, the feature vector is extracted from the 1-D sequence.
The used method to represent a 2-D shape (character) with a 1-D signal is a radial sequence from the center of gravity to the contour of the shape [Yang 91]. We denote this radial sequence r(k), k=1,...,L, where L is the length of the radii sequence(Figure 6.1).

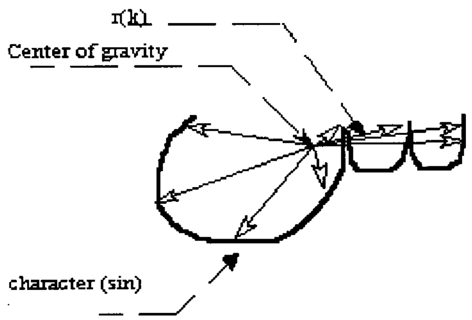


Figure 6.1. : Radial sequence

A direct choice of sampling frequency is Shannon frequency ($f_s \geq 2F$). After sampling, the radial sequence is divided into a T equidistant steps. Considering a size normalization, r(t) can be written as : $r(t) = r(t)/r_{max}$. Since only the discrete HMM is considered, the r(t) must be quantized. The quantization step is choosen experimentally. Its value varies from 0.1 to 0.01.
Each observation $O=\{o_t\}$ is a vector of T dimension. In this case o_t can be calculated as follows :
$\mathbf{o_t = c * r(t)/r_{max}}$, where **c** is quantization constant, $10 \leq c \leq 100$. The following figure (Figure 6.2) shows the complete symbol generation procedure.

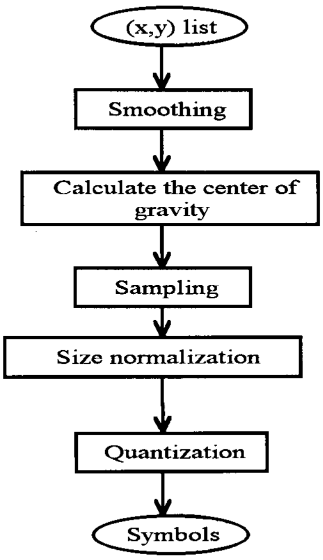


Figure 6.2: Symbol generation procedure

6.3. Recognition results

In order to verify the effectiveness of the proposed nonstationary model, we used 800 forms of arabic characters written by 2 scripters. We have tried different number of states, N=3 to 7, and both stationary and nonstationary transition. We have also compared the results, from M=8 to 20. With the intention of maximazing the recognition rate we choose to study the necessary number of iterations (NI) to obtain the good parameter estimation. Results showed that in training NI is a linear function of LS, where LS is the length of the most long string of vector observation (Figure 6.3).

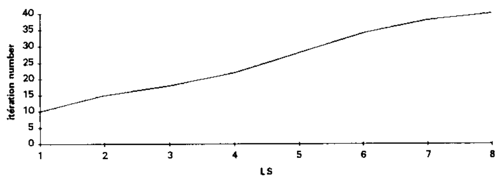


Fig. 6.3. : Iteration number of character « ح » model

Vector observation = « 6765442223344456666 », the most long string = « 6666 » where **LS**=4

With the big complexity of the proposed model, an adequate choice of the value of D can reduce the calculations. With a fixed value of N=5, experiences showed that the good value of D is M-2 . Experimental results indicate that the proposed model offers a great potential for solving complex and dynamic arabic handwritten characters. the recognition rate is affected by the choice of N, M and D. Table 1 shows the recognition rates with different values of N,M and D. It's clear that the best recognition rates correspond to the nonstationary models.

(N,M,D)	HMM(D=1)	DHMM	NSHMM
(5,8,6)	92.01%	93.05%	95.5%
(5,10,8)	92.3%	94.2%	96%
(7,12,10)	94.1%	96%	98.2%
(7,20,16)	94.8%	96.5%	98.6%

Table 1 : Recognition rate

7. CONCLUSION

This paper exposes a nonstationarity of the states in a Hidden Markov model called NSHMM. The proposed model is more general in descriptive capability than that of standard HMM and DHMM. Its parameter estimation has been presented based on the Baum Welch algorithm. In this NSHMM, the state duration is modeled similar to that of the DHMM, with a little increase of storage and computation. The transition probablities are expressed conditional on state duration. The increase number of parameters can be reduced by functional modelling of transition probabilities.
The model has been applied for on-line recognition of handwritten arabic characters. Each character is modeled by a hidden Markov model with output distributions defined by discret parameters.
The proposed NSHMM in this paper increases significantly the flexibility of hidden Markov models for character recognition.
Although, it is a very promising approach for solving the problem of on-line recognition of different set characters.

8. REFERENCES

[Deng 93] L. Deng , » A stochastic models of speech incorporating hierarchical nonstationarity », IEEE, Trans. speech audio proces. , vol. 1, N° 1, Oct. 1993, 99 471-474.
[Furguson 80] Furguson J. D. « variable duration models for speech » , proc. symp. on the application of hidden markov models to text and speech, Princeton, NJ, 1980, pp 143-179.
[Ghitza 93] Ghitza Oded , M. M. Sondhi « Hidden Markov models with templates as non-stationary states : an application to speech recognition », Computer speech and language, pp 101-119
[Krishna 93] Krishna S. Nathan, Jerome, Nahamoo, Bellegarda“ On line handwriting recognition using continuous parameter Hidden Markov Model ”IEEE, pp. 121-124, 1993
[Rabiner 89] Lawrence R. Rabiner « A tutorial on Hidden Markov Models and selected applications in speech recognition » IEEE, Proceedings, vol. 77, N° 2, pp. 257-286, February 1989
[Tseng 91] H. Gu , C. Tseng and L. Lee, « Isolated ulterance speech recognition using hidden Markov models with bounded state durations », IEEE trans. signal process. , vol. 39, N° 8, August 91.
[Vaseghi 95] Vaseghi S. V. « State duration modelling in hidden Markov models », Signal proc. , 41, pp 31-41, 1995
[Veltman 94] Stephan R. Veltman and Ramjee Prasad, Senior Member IEEE « Hidden Markov Models applied to on-line handwritten isolated character recognition » IEEE transactions on image processing, vol. 3, N°3, May 1994.
[Yang 91] Yang He Amlan Kundu “ 2-D shape classification using Hidden Markov Model ” IEEE, PAMI vol. 13, N° 11, November 1991.