Training Hidden Markov Models with Multiple Observations—A Combinatorial Method

Xiaolin Li, *Member*, *IEEE Computer Society*, Marc Parizeau, *Member*, *IEEE Computer Society*, and Réjean Plamondon, *Fellow*, *IEEE*

Abstract—Hidden Markov models (HMMs) are stochastic models capable of statistical learning and classification. They have been applied in speech recognition and handwriting recognition because of their great adaptability and versatility in handling sequential signals. On the other hand, as these models have a complex structure and also because the involved data sets usually contain uncertainty, it is difficult to analyze the multiple observation training problem without certain assumptions. For many years researchers have used Levinson's training equations in speech and handwriting applications, simply assuming that all observations are independent of each other. This paper presents a formal treatment of HMM multiple observation training without imposing the above assumption. In this treatment, the multiple observation probability is expressed as a combination of individual observation probabilities without losing generality. This combinatorial method gives one more freedom in making different dependence-independence assumptions. By generalizing Baum's auxiliary function into this framework and building up an associated objective function using the Lagrange multiplier method, it is proven that the derived training equations guarantee the maximization of the objective function. Furthermore, we show that Levinson's training equations can be easily derived as a special case in this treatment.

Index Terms—Hidden Markov model, forward-backward procedure, Baum-Welch algorithm, multiple observation training.

1 Introduction

Hidden Markov models (HMMs) are stochastic models which were introduced and studied in the late 1960s and early 1970s [1], [2], [3], [4], [5]. As the parameter space of these models is usually superdimensional, the model training problem seems very difficult at first glance. In 1970, Baum et al. published their maximization method which gave a solution to the model training problem with a single observation [4]. In 1977, Dempster et al. introduced the Expectation-Maximization (EM) method for maximum likelihood estimates from incomplete data and, later, Wu proved some convergence properties of the EM algorithm [6], which made the EM algorithm a solid framework in statistical analysis. In 1983, Levinson et al. presented a maximum likelihood estimation method for HMM multiple observation training, assuming that all observations are independent of each other [7]. Since then, HMMs have been widely used in speech recognition [7], [8], [9], [10], [11], [12]. More recently, they have also been applied to handwriting recognition [18], [19], [20], [21], [22] as they are adaptive to random sequential signals and capable of statistical learning and classification.

Although the independence assumption of observations is helpful for problem simplification, it may not hold in some cases. For example, the observations of a syllable pronounced by a person are possibly highly correlated. Similar examples can also be found in handwriting: Given a set of samples of a letter written by a person, it is difficult to assume or deny their independence properties when viewed from different perspectives. Based on these phenomena, it is better not to just rely on the independence assumption.

This paper presents a formal treatment for HMM multiple observation training without imposing the independence assumption. In this treatment, the multiple observation probability is expressed as a combination of individual observation probabilities rather than their product. The dependence-independence property of the observations is characterized by combinatorial weights. These weights give us more freedom in making different assumptions and, hence, in deriving corresponding training equations. By generalizing Baum's auxiliary function into this framework and building up an associated objective function using the Lagrange multiplier method, it is proven that the derived training equations guarantee the maximization of the objective function and, hence, the convergence of the training process. Furthermore, as two special cases in this treatment, we show that Levinson's training equations can be easily derived with an independence assumption and some other training equations can also be derived with a uniform dependence assumption.

The remainder of this paper is organized as follows: Section 2 summarizes the first order HMM. Section 3 describes the combinatorial method for HMM multiple observation training. Section 4 shows two special cases: an independence assumption versus a uniform dependence assumption. Finally, Section 5 concludes this paper.

- *X. Li is with CADlink Technology Corporation, 2440 Don Reid Drive, Suite 100, Ottawa, Ontario, Canada K1H 1E1. E-mail: xli@cadlink.com.*
- *M. Parizeau is with the Département de Génie Electrique et de Génie Informatique, Université Laval, Ste-Foy, Québec, Canada G1K 7P4. E-mail: parizeau@gel.ulaval.ca.*
- *R. Plamondon is with the École Polytechnique de Montréal, Montréal, Québec, Canada H3C 3A7. E-mail: rejean.plamondon@polymtl.ca.*

Manuscript received 5 Dec. 1997; accepted 4 Aug. 1999.
Recommended for acceptance by R. Chellappa.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 107606.

2 First Order Hidden Markov Model

2.1 Elements of HMM

A hidden Markov process is a doubly stochastic process: an underlying process which is hidden from observation and an observable process which is determined by the underlying process. With respect to first order hidden Markov process, the model is characterized by the following elements [10]:

- *set of hidden states*:

$$S = \{S_1, S_2, \cdots, S_N\}, \tag{1}$$

  where N is the number of states in the model,
- *state transition probability distribution*:[1]

$$A = \{a_{ij}\}, \tag{2}$$

  where, for $1 \le i, j \le N$,

$$a_{ij} = P[q_{t+1} = S_j | q_t = S_i] \tag{3}$$

$$\begin{cases} 0 \le a_{ij} \\ \sum_{j=1}^{N} a_{ij} = 1, \end{cases} \tag{4}$$

- *set of observation symbols*:

$$V = \{v_1, v_2, \cdots, v_M\}, \tag{5}$$

  where M is the number of observation symbols per state,
- *observation symbol probability distribution*:[2]

$$B = \{b_j(k)\}, \tag{6}$$

  where, for $1 \le j \le N, \ 1 \le k \le M$,

$$b_j(k) = P[v_k \text{ at } t | q_t = S_j] \tag{7}$$

$$\begin{cases} 0 \le b_j(k) \\ \sum_{k=1}^{M} b_j(k) = 1, \end{cases} \tag{8}$$

  and
- *initial state probability distribution*:

$$\pi = \{\pi_i\}, \tag{9}$$

  where, for $1 \le i \le N$,

$$\pi_i = P[q_1 = S_i] \tag{10}$$

$$\begin{cases} 0 \le \pi_i \\ \sum_{i=1}^{N} \pi_i = 1. \end{cases} \tag{11}$$

  For convenience, we denote an HMM as a triplet in all subsequent discussion:

$$\lambda = (A, B, \pi). \tag{12}$$

1. A is also called transition matrix.
2. B is also called emission matrix.

2.2 Ergodic Model and Left-Right Model

An HMM can be classified into one of the following types in the light of its state transition:

- *ergodic model*: An ergodic model has full state transition.
- *left-right model*:[3] A left-right model has only partial state transition such that $a_{ij} = 0, \forall j < i$.

2.3 Observation Evaluation: Forward-Backward Procedure

Let $O = o_1 o_2 \cdots o_T$ be an observation sequence where $o_t \in V$ is the observation symbol at time t and let $Q = q_1 q_2 \cdots q_T$ be a state sequence where $q_t \in S$ is the state at time t. Given a model λ and an observation sequence O, the observation evaluation problem $P(O|\lambda)$ can be solved using forward-backward procedure in terms of forward and backward variables (Fig. 1):

- *forward variable*:[4]

$$\alpha_t(i) = P(o_1 o_2 \cdots o_t, q_t = S_i | \lambda). \tag{13}$$

  $\alpha_t(i)$ can be solved inductively:

  1. initialization:

$$\alpha_1(i) = \pi_i b_i(o_1), \quad 1 \le i \le N \tag{14}$$

     and
  2. induction:

$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] b_j(o_{t+1}), \quad 1 \le t \le T-1, 1 \le j \le N. \tag{15}$$

- *backward variable*:[5]

$$\beta_t(i) = P(o_{t+1} o_{t+2} \cdots o_T | q_t = S_i, \lambda). \tag{16}$$

  $\beta_t(i)$ can be solved inductively:

  1. initialization:

$$\beta_T(i) = 1, \quad 1 \le i \le N \tag{17}$$

     and
  2. induction:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j), \quad 1 \le t \le T-1, 1 \le i \le N. \tag{18}$$

- *observation evaluation*:

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_t(i) \beta_t(i), \forall t, \tag{19}$$

  especially,

3. This type of model is widely used in modeling sequential signals.
4. That is, the probability of the partial observation sequence $o_1 o_2 \cdots o_t$ with state $q_t = S_i$ given model λ.
5. That is, the probability of the partial observation sequence $o_{t+1} o_{t+2} \cdots o_T$ given state $q_t = S_i$ and model λ.

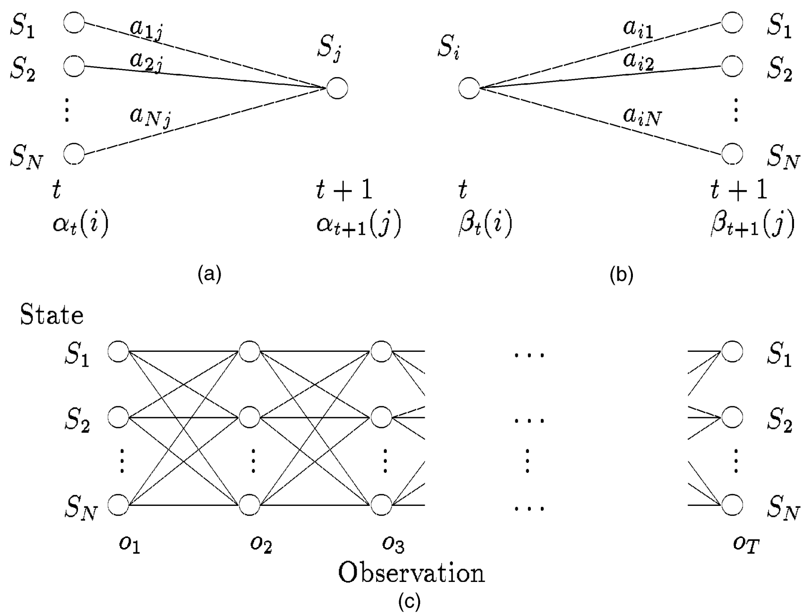

Fig. 1. Illustration of forward-backward procedure. (a) Forward variable. (b) Backward variable. (c) Computation lattice.

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i). \tag{20}$$

It is easy to see that the computational complexity of the forward-backward procedure is $O(TN^2)$.

2.4 Model Training: Baum-Welch Algorithm

Now, let us consider the model training problem: Given an observation sequence O, how do we find the optimum model parameter vector $\lambda \in \Lambda$ that maximizes $P(O|\lambda)$. To solve this problem, Baum et al. defined an auxiliary function and proved the two propositions below [4]:

- *auxiliary function*:

$$Q(\lambda, \bar{\lambda}) = \sum_{Q} P(O, Q|\lambda) \log P(O, Q|\bar{\lambda}), \tag{21}$$

where $\bar{\lambda}$ is the auxiliary variable that corresponds to λ.

Proposition 1. *If the value of $Q(\lambda, \bar{\lambda})$ increases, then the value of $P(O|\bar{\lambda})$ also increases, i.e.,*

$$Q(\lambda, \bar{\lambda}) \geq Q(\lambda, \lambda) \longrightarrow P(O|\bar{\lambda}) \geq P(O|\lambda). \tag{22}$$

Proposition 2. *λ is a critical point of $P(O|\lambda)$ if and only if it is a critical point of $Q(\lambda, \bar{\lambda})$ as a function of $\bar{\lambda}$, i.e.,*

$$\frac{\partial P(O|\lambda)}{\partial \lambda_i} = \left.\frac{\partial Q(\lambda, \bar{\lambda})}{\partial \bar{\lambda}_i}\right|_{\bar{\lambda}=\lambda}, 1 \leq i \leq D, \tag{23}$$

where D is the dimension of λ and λ_i, $1 \leq i \leq D$, are individual elements of λ.

In light of the above propositions, the model training problem can be solved by the Baum-Welch algorithm in terms of joint events and state variables (Fig. 2):

- *joint event*:[6]

$$\begin{aligned}\xi_t(i,j) &= P(q_t = S_i, q_{t+1} = S_j|O, \lambda) \\ &= \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{P(O|\lambda)},\end{aligned} \tag{24}$$

- *state variable*:[7]

$$\begin{aligned}\gamma_t(i) &= P(q_t = S_i|O, \lambda) \\ &= \sum_{j=1}^{N} \xi_t(i,j),\end{aligned} \tag{25}$$

- *parameter updating equations*:

1. state transition probability:

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \quad 1 \leq i \leq N, 1 \leq j \leq N, \tag{26}$$

2. symbol emission probability:

$$\bar{b}_j(k) = \frac{\sum_{t=1, o_t=v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}, \quad 1 \leq j \leq N, 1 \leq k \leq M, \tag{27}$$

3. initial state probability:

$$\bar{\pi}_i = \gamma_1(i), \quad 1 \leq i \leq N. \tag{28}$$

6. That is, the probability of being in state S_i at time t and state S_j at time $t+1$ given the observation sequence O and model λ.

7. That is, the probability of being in state S_i at time t given the observation sequence O and the model λ.

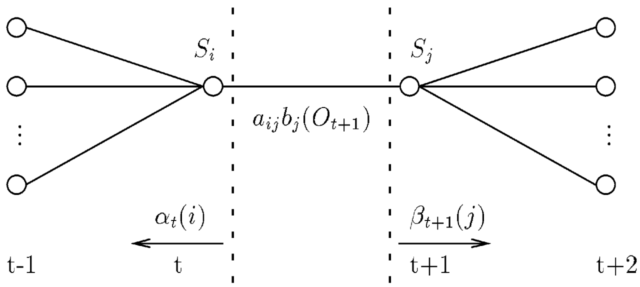

Fig. 2. Illustration of the joint event.

3 Multiple Observation Training

3.1 Combinatorial Method

Now, let us consider a set of observation sequences from a pattern class:

$$O = \{O^{(1)}, O^{(2)}, \cdots, O^{(K)}\}, \tag{29}$$

where

$$O^{(k)} = o_1^{(k)} o_2^{(k)} \cdots o_{T_k}^{(k)}, \quad 1 \le k \le K \tag{30}$$

are individual observation sequences. Usually, one does not know if these observation sequences are independent of each other or not. And, a contravercy can arise if one assumes the independence property while these observation sequences are statistically correlated. In either case, we have the following expressions without losing generality:

$$\begin{cases} P(\mathbf{O}|\lambda) = & P(O^{(1)}|\lambda)P(O^{(2)}|O^{(1)},\lambda)\cdots \\ & P(O^{(K)}|O^{(K-1)}\cdots O^{(1)},\lambda) \\ P(\mathbf{O}|\lambda) = & P(O^{(2)}|\lambda)P(O^{(3)}|O^{(2)},\lambda)\cdots \\ & P(O^{(1)}|O^{(K)}\cdots O^{(2)},\lambda) \\ \vdots & \\ P(\mathbf{O}|\lambda) = & P(O^{(K)}|\lambda)P(O^{(1)}|O^{(K)},\lambda)\cdots \\ & P(O^{(K-1)}|O^{(K)}O^{(K-2)}\cdots O^{(1)},\lambda). \end{cases} \tag{31}$$

Based on the above equations, the multiple observation probability given the model can be expressed as a summation:

$$P(\mathbf{O}|\lambda) = \sum_{k=1}^{K} w_k P(O^{(k)}|\lambda), \tag{32}$$

where

$$\begin{cases} w_1 = & \frac{1}{K} P(O^{(2)}|O^{(1)},\lambda)\cdots P(O^{(K)}|O^{(K-1)}\cdots O^{(1)},\lambda) \\ w_2 = & \frac{1}{K} P(O^{(3)}|O^{(2)},\lambda)\cdots P(O^{(1)}|O^{(K)}\cdots O^{(2)},\lambda) \\ \vdots & \\ w_K = & \frac{1}{K} P(O^{(1)}|O^{(K)},\lambda)\cdots P(O^{(K-1)}|O^{(K)}O^{(K-2)}\cdots \\ & O^{(1)},\lambda) \end{cases} \tag{33}$$

are weights. These weights are conditional probabilities and, hence, they can characterize the dependence-independence property.

Based on the above expression, we can construct an auxiliary function below for model training:

$$Q(\lambda, \bar{\lambda}) = \sum_{k=1}^{K} w_k Q_k(\lambda, \bar{\lambda}), \tag{34}$$

where $\bar{\lambda}$ is the auxiliary variable corresponding to λ and

$$Q_k(\lambda, \bar{\lambda}) = \sum_{Q} P(O^{(k)}, Q|\lambda) \log P(O^{(k)}, Q|\bar{\lambda}), \quad 1 \le k \le K \tag{35}$$

are Baum's auxiliary functions related to individual observations. Since w_k, $1 \le k \le K$, are not functions of $\bar{\lambda}$, we have the following theorems related to the maximization of $P(\mathbf{O}|\lambda)$ [23]:

Theorem 1. *If the value of $Q(\lambda, \bar{\lambda})$ increases, then the value of $P(\mathbf{O}|\bar{\lambda})$ also increases, i.e.,*

$$Q(\lambda, \bar{\lambda}) \ge Q(\lambda, \lambda) \longrightarrow P(\mathbf{O}|\bar{\lambda}) \ge P(\mathbf{O}|\lambda). \tag{36}$$

Furthermore, as w_k, $1 \le k \le K$ are weights that characterize the dependence-independence property of the observations, if one assumes that these weights are constants, one has the following theorem [23]:

Theorem 2. *For fixed w_k, $1 \le k \le K$, λ is a critical point of $P(O|\lambda)$ if and only if it is a critical point of $Q(\lambda, \bar{\lambda})$ as a function of $\bar{\lambda}$, i.e.,*

$$\frac{\partial P(\mathbf{O}|\lambda)}{\partial \lambda_i} = \left. \frac{\partial Q(\lambda, \bar{\lambda})}{\partial \bar{\lambda}_i} \right|_{\bar{\lambda}=\lambda.} \tag{37}$$

In such a case, the maximization of $Q(\lambda, \bar{\lambda})$ is equivalent to the maximization of $P(\mathbf{O}|\lambda)$.

3.2 Maximization: Lagrange Multiplier Method

Based on Theorem 1, one can always maximize $Q(\lambda, \bar{\lambda})$ to increase the value of $P(\mathbf{O}|\bar{\lambda})$, regardless of 1) if the individual observations are independent of one another or not and 2) whether the combinatorial weights are constants or not. Let us consider the auxiliary function with boundary conditions:

$$\begin{aligned} & Q(\lambda, \bar{\lambda}) = \textstyle\sum_{k=1}^{K} w_k Q_k(\lambda, \bar{\lambda}) & \\ & 1 - \textstyle\sum_{j=1}^{N} \bar{a}_{ij} = 0, & 1 \le i \le N \\ & 1 - \textstyle\sum_{k=1}^{M} \bar{b}_j(k) = 0, & 1 \le j \le N \\ & 1 - \textstyle\sum_{i=1}^{N} \bar{\pi}_i = 0, & \end{aligned} \tag{38}$$

we can construct an objective function using Lagrange multiplier method:

$$\begin{aligned} F(\bar{\lambda}) = & Q(\lambda, \bar{\lambda}) + \sum_{i=1}^{N} c_{ai} \left[1 - \sum_{j=1}^{N} \bar{a}_{ij}\right] \\ & + \sum_{j=1}^{N} c_{bj} \left[1 - \sum_{k=1}^{M} \bar{b}_j(k)\right] + c_\pi \left[1 - \sum_{i=1}^{N} \bar{\pi}_i\right], \end{aligned} \tag{39}$$

where c_{ai}, c_{bj}, and c_π are Lagrange multipliers. Differentiating the objective function with respect to individual parameters and finding solutions to corresponding Lagrange multipliers, we obtain the following training equations that guarantee the maximization of the objective function:

1. state transition probability:

$$\bar{a}_{mn} = \frac{\sum_{k=1}^{K} w_k P(O^{(k)}|\lambda) \sum_{t=1}^{T_k-1} \xi_t^{(k)}(m,n)}{\sum_{k=1}^{K} w_k P(O^{(k)}|\lambda) \sum_{t=1}^{T_k-1} \gamma_t^{(k)}(m)}, \quad 1 \le m \le N, 1 \le n \le N, \tag{40}$$

2. symbol emission probability:

$$\bar{b}_n(m) = \frac{\sum_{k=1}^{K} w_k P(O^{(k)}|\lambda) \sum_{t=1, o_t^{(k)}=v_m}^{T_k} \gamma_t^{(k)}(n)}{\sum_{k=1}^{K} w_k P(O^{(k)}|\lambda) \sum_{t=1}^{T_k} \gamma_t^{(k)}(n)}, \quad 1 \le n \le N, 1 \le m \le M, \tag{41}$$

3. initial state probability:

$$\bar{\pi}_n = \frac{\sum_{k=1}^{K} w_k P(O^{(k)}|\lambda) \gamma_1^{(k)}(n)}{\sum_{k=1}^{K} w_k P(O^{(k)}|\lambda)}, \quad 1 \le n \le N. \tag{42}$$

3.3 Convergence Property

The training equations derived by the Lagrange multiplier method guarantee the convergence of the training process. First, these training equations give the zero points of the first order Jacobi differential matrix

$$\frac{\partial F(\bar{\lambda})}{\partial \bar{\lambda}}.$$

Second, the second order Jacobi differential matrix

$$\frac{\partial^2 F(\bar{\lambda})}{\partial \bar{\lambda}^2}$$

is diagonal and all its diagonal elements are negative. Thus, the algorithm guarantees local maxima and hence, the convergence of the training process (see [23] for detailed proofs).

The above training equations are adaptive to both the ergodic model and the left-right model since we do not put any constraints on the model type during the derivation.

4 Two Special Cases: Independence versus Uniform Dependence

4.1 Independence Assumption

Now, let us assume that the individual observations are independent of each other, i.e.,

$$P(\mathbf{O}|\lambda) = \prod_{k=1}^{K} P(O^{(k)}|\lambda). \tag{43}$$

In this case, the combinatorial weights become:

$$w_k = \frac{1}{K} P(\mathbf{O}|\lambda)/P(O^{(k)}|\lambda), \quad 1 \le k \le K. \tag{44}$$

Substituting the above weights into (40) to (42), we obtain Levinson's training equations:

1. state transition probability:

$$\bar{a}_{mn} = \frac{\sum_{k=1}^{K} \sum_{t=1}^{T_k-1} \xi_t^{(k)}(m,n)}{\sum_{k=1}^{K} \sum_{t=1}^{T_k-1} \gamma_t^{(k)}(m)}, \quad 1 \le m \le N, 1 \le n \le N, \tag{45}$$

2. symbol emission probability:

$$\bar{b}_n(m) = \frac{\sum_{k=1}^{K} \sum_{t=1, o_t^{(k)}=v_m}^{T_k} \gamma_t^{(k)}(n)}{\sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_t^{(k)}(n)}, \quad 1 \le n \le N, 1 \le m \le M, \tag{46}$$

3. initial state probability:

$$\bar{\pi}_n = \frac{1}{K} \sum_{k=1}^{K} \gamma_1^{(k)}(n), \quad 1 \le n \le N. \tag{47}$$

4.2 Uniform Dependence Assumption

If we assume that the individual observations are uniformly dependent on one another, i.e.,

$$w_k = const, \quad 1 \le k \le K. \tag{48}$$

Substituting the above weights into (40) to (42), it readily follows that

1. state transition probability:

$$\bar{a}_{mn} = \frac{\sum_{k=1}^{K} P(O^{(k)}|\lambda) \sum_{t=1}^{T_k-1} \xi_t^{(k)}(m,n)}{\sum_{k=1}^{K} P(O^{(k)}|\lambda) \sum_{t=1}^{T_k-1} \gamma_t^{(k)}(m)}, \quad 1 \le m \le N, 1 \le n \le N, \tag{49}$$

2. symbol emission probability:

$$\bar{b}_n(m) = \frac{\sum_{k=1}^{K} P(O^{(k)}|\lambda) \sum_{t=1, o_t^{(k)}=v_m}^{T_k} \gamma_t^{(k)}(n)}{\sum_{k=1}^{K} P(O^{(k)}|\lambda) \sum_{t=1}^{T_k} \gamma_t^{(k)}(n)}, \quad 1 \le n \le N, 1 \le m \le M, \tag{50}$$

3. initial state probability:

$$\bar{\pi}_n = \frac{\sum_{k=1}^{K} P(O^{(k)}|\lambda) \gamma_1^{(k)}(n)}{\sum_{k=1}^{K} P(O^{(k)}|\lambda)}, \quad 1 \le n \le N. \tag{51}$$

5 Conclusions

A formal treatment for HMM multiple observation training has been presented in this paper. In this treatment, the multiple observation probability is expressed as a combination of individual observation probabilities without losing generality. The independence-dependence property of the observations are characterized by the combinatorial weights and, hence, it gives us more freedom in making different assumptions and also in deriving corresponding training equations.

The well-known Baum's auxiliary function has been generalized into the case of multiple observation training and two theorems related to the maximization have been presented in this paper. Based on the auxiliary function and its boundary conditions, an objective function has been constructed using Lagrange multiplier method and a set of training equations have been derived by maximizing the objective function. Similar to the EM algorithm, this algorithm guarantees the local maxima and, hence, the convergence of the training process.

We have also shown, through two special cases, that the above training equations are general enough to include different situations. Once the independence assumption is made, one can readily obtain Levinson's training equations. On the other hand, if the uniform dependence is assumed, one can also have the corresponding training equations.

ACKNOWLEDGMENTS

This research work was supported by NSERC grant OGP0155389 to M. Parizeau and NSERC grant OGP00915 and FCAR Grant ER-1220 to R. Plamondon.

REFERENCES

[1] L.E. Baum and T. Petrie, "Statistical Inference for Probabilistic Functions of Finite State Markov Chains," *Annals of Math. Statistics,* vol. 37, pp. 1,554-1,563, 1966.

[2] L.E. Baum and J.A. Egon, "An Inequality with Applications to Statistical Estimation for Probabilistic Functions of a Markov Process and to a Model for Ecology," *Bull. Amer. Meteorology Soc.,* vol. 73, pp. 360-363, 1967.

[3] L.E. Baum and G.R. Sell, "Growth Functions for Transformations on Manifolds," *Pacific J. Math.,* vol. 27, no. 2, pp. 211-227, 1968.

[4] L.E. Baum, T. Petrie, G. Soules, and N. Weiss, "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *Annals of Math. Statistics,* vol. 41, no. 1, pp. 164-171, 1970.

[5] L.E. Baum, "An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes," *Inequalities,* vol. 3, pp. 1-8, 1970.

[6] C.F.J. Wu, "On the Convergence Properties of the EM Algorithm," *Annals of Statistics,* vol. 11, no. 1, pp. 95-103, 1983.

[7] S.E. Levinson, L.R. Rabiner, and M.M. Sondhi, "An Introduction to the Application of the Theory of Probabilistic Functions of Markov Process to Automatic Speech Recognition," *Bell System Technical J.,* vol. 62, no. 4, pp. 1,035-1,074, 1983.

[8] L.R. Bahl, F. Jelinek, and R.L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 5, pp. 179-190, 1983.

[9] L.R. Rabiner and S.E. Levinson, "A Speaker-Independent, Syntax-Directed, Connected Word Recognition System Based on Hidden Markov Models and Level Building," *IEEE Trans. Acoustic, Speech, and Signal Processing,* vol. 33, no. 3, pp. 561-573, 1985.

[10] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE,* vol. 77, no. 2, pp. 257-286, 1989.

[11] K.-F. Lee, *Automatic Speech Recognition—The Development of SPHINX System.* Kluwer Academic, 1989.

[12] L. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition.* Englewood Cliffs, N.J.: Prentice Hall, 1993,

[13] M. Iwayama, N. Indurkhya, and H. Motoda, "A New Algorithm for Automatic Configuration of Hidden Markov Models," *Proc. Fourth Int'l Workshop Algorithmic Learning Theory (ALT '93),* pp. 237-250, 1993.

[14] A. Kaltenmeier, T. Caesar, J.M. Gloger, and E. Mandler, "Sophisticated Topology of Hidden Markov Models for Cursive Script Recognition," *Proc. Second Int'l Conf. Document Analysis and Recognition,* pp. 139-142, 1993.

[15] P. Baldi and Y. Chauvim, "Smooth On-Line Learning Algorithms for Hidden Markov Models," *Neural Computation,* vol. 6, pp. 307-318, 1994.

[16] S.B. Cho and J.H. Kim, "An HMM/MLP Architecture for Sequence Recognition," *Neural Computation,* vol. 7, pp. 358-369, 1995.

[17] J. Dai, "Robust Estimation of HMM Parameters Using Fuzzy Vector Quantization and Parzen's Window," *Pattern Recognition,* vol. 28, no. 1, pp. 53-57, 1995.

[18] A. Kundu, Y. He, and P. Bahl, "Recognition of Handwritten Word: First and Second Order Hidden Markov Model Based Approach," *Pattern Recognition,* vol. 22, no. 3, pp. 283-297, Mar. 1989.

[19] S.R. Veltman and R. Prasad, "Hidden Markov Models Applied to On-Line Handwritten Isolated Character Recognition," *IEEE Trans. Image Processing,* vol. 3, no. 3, pp. 314-318, 1994.

[20] E.J. Bellegarda, J.R. Bellegarda, D. Nahamoo, and K.S. Nathan, "A Fast Statistical Mixture Algorithm for On-Line Handwriting Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 16, no. 12, pp. 1,227-1,233, Dec. 1994.

[21] G. Rigoll, A. Kosmala, J. Rottland, and C.H. Neukirchen, "A Comparison between Continuous and Discrete Density Hidden Markov Models for Cursive Handwriting Recognition," *Proc. Int'l Conf. Pattern Recognition (ICPR '96),* pp. 205-209, 1996.

[22] J. Hu, M.K. Brown, and W. Turin, "HMM Based On-Line Handwriting Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 10, pp. 1,039-1,045, Nov. 1996.

[23] X. Li, M. Parizeau, and R. Plamondon, "Hidden Markov Model Multiple Observation Training," Technical Report EPM/RT-99/16, Nov. 1999.

Xiaolin Li received the BSc degree in electrical engineering from Guangxi University in 1978, the MSc degree in electrical engineering from the East China Institute of Technology in 1987, and the PhD degree in computer science from the University of Birmingham, England, in 1994. In the 1980s, he worked as an assistant lecturer at Guangxi University and then as a lecturer at the Guilin Institute of Electronics. In 1990, he was a visiting researcher at the University of Manchester, England. After receiving his PhD degree, he worked with the Hong Kong University of Science and Technology, Laval University, and, then, the École Polytechnique de Montréal. Currently, he is a mathematician working with CADlink Technology Corporation in Ottawa, Canada. His research interests include pattern recognition, image/signal processing, linear systems, and computational geometry.

Marc Parizeau received the PhD degree from the École Polytechnique de Montréal in 1992. He is now an assistant professor of computer engineering at the Université Laval, Québec, Canada. His main research interests are in pattern recognition, neural networks, and genetic programming.

Réjean Plamondon received the BSc degree in physics and the MScA and PhD degrees in electrical engineering from the Université Laval, Québec, Canada, in 1973, 1975, and 1978, respectively. In 1978, he became a member of the faculty of the École Polytechnique, Montréal, Canada, where he is currently a full professor. He was the head of the Department of Electrical and Computer Engineering from 1996 to 1998 and he is now the chief executive officer of the École Polytechnique, one of the largest engineering schools in Canada.

Over the past 20 years, Dr. Plamondon has proposed many original solutions to problems in the field of on-line and off-line handwriting analysis and processing. He has based these solutions on his exhaustive studies of human movement generation and perception, particularly as it relates to problems associated with the design of automatic systems for signature verification and handwriting recognition, and has also applied this knowledge to the development of interactive electronic penpads to help children learn handwriting skills and of powerful methods for analyzing and interpreting neuromuscular signals. His major contribution has been the theoretical development of a kinematic theory of rapid human movements which can take into account, with the help of a single equation called a delta-lognormal function, many psychophysical phenomena reported in studies dealing with rapid movements over the past century. The theory has been found to be successful in describing the basic kinematic properties of velocity profiles as observed in finger, hand, arm, head, and eye movements. He then studied and analyzed these biosignals extensively in order to develop creative and powerful methods and systems in various domains of engineering.

His research interests focus on the automatic processing of handwriting: neuromotor models of movement generation and image perception, script recognition, signature verification, signal analysis and processing, electronic penpads, man-computer interfaces via handwriting, forensic sciences, education, and artifical intelligence. He is the founder and director of Laboratoire Scribens, a research group dedicated exclusively to the study of these topics.

Dr. Plamondon is an active member of several professional societies; is the president of the International Graphonomics Society; and is the Canadian representative on the Board of Governors of the International Association for Pattern Recognition (IAPR). From 1988 to 1994, he was thechair of the IEEE Computer Society Technical Committee TC-11. In 1989-1990, he was a fellow at the Netherlands Institute for Advanced Study in Wassenaar. From 1990 to 1997, he was the president of the Canadian Image Processing and Pattern Recognition Society. In 1994, he was named a fellow of the IAPR and, from 1994 to 1998, he was the chair of the IAPR conferences and meetings committee. He is the author or coauthor of numerous publications and technical reports. He has coedited four books and has also published a children's book, a novel, and two collections of poems. He is a fellow of the IEEE.