Generalized Hidden Markov Models—Part I: Theoretical Frameworks

Magdi A. Mohamed, *Member, IEEE,* and Paul Gader, *Senior Member, IEEE*

***Abstract*—This is the first paper in a series of two papers describing a novel generalization of classical hidden Markov models using fuzzy measures and fuzzy integrals. In this paper, we present the theoretical framework for the generalization and, in the second paper, we describe an application of the generalized hidden Markov models to handwritten word recognition. The main characteristic of the generalization is the relaxation of the usual additivity constraint of probability measures. Fuzzy integrals are defined with respect to fuzzy measures, whose key property is monotonicity with respect to set inclusion. This property is far weaker than the usual additivity property of probability measures. As a result of the new formulation, the statistical independence assumption of the classical hidden Markov models is relaxed. An attractive property of this generalization is that the generalized hidden Markov model reduces to the classical hidden Markov model if we used the Choquet fuzzy integral and probability measures. Another interesting property of the generalization is the establishment of a relation between the generalized hidden Markov model and the classical nonstationary hidden Markov model in which the transitional parameters vary with time.**

***Index Terms*—Fuzzy integral, fuzzy measures, handwriting recognition, Markov models.**

I. Introduction

HIDDEN Markov model (HMM) is a statistical method that uses probability measures to model sequential data represented by sequence of observation vectors. In this paper, we describe a novel generalization of the classical hidden Markov models that utilizes fuzzy sets, fuzzy measures, and fuzzy integrals. Fuzzy integrals are defined with respect to fuzzy measures whose key property is monotonicity with respect to set inclusion. This property is far weaker than the usual additivity property of probability measures. As a result of the monotonicity of the fuzzy measures, the statistical independence assumption of the classical model is relaxed in the generalized (fuzzy) model. An attractive property of this generalization is that the generalized hidden Markov model (GHMM) reduces to the classical hidden Markov model if we used the Choquet fuzzy integral and probability measures. This property implies that the generalized models include the classical one as a special case.

The classical hidden Markov models have been found to be extremely useful for a wide spectrum of applications in ecology, cryptanalysis, image understanding, speech, and handwriting recognition [1]–[3]. The proposed generalized hidden Markov model shares the ability to model sequential processes with the classical one and, therefore, can be used for similar applications. This proper generalization, which is based on the strong theoretical foundations of fuzzy measures and fuzzy integrals [4], provides more freedom and flexibility to aggregate the sequential information obtained from the observation sequences.

One significant contribution of our research is the establishment of a relation between the generalized hidden Markov model and the classical nonstationary hidden Markov model in which the transitional parameters are allowed to vary with time. The main advantage of our proposed generalized model over the classical nonstationary model is that this nonstationary behavior, which is an extremely desirable property, is achieved naturally and dynamically as a byproduct of the nonlinear aggregation of information using the fuzzy integral. Moreover, the fuzzy model does not require fixing the lengths of the observation sequences and the availability of large training sets in order to learn a large number of transition parameters as for the classical nonstationary model.

Constructing a mathematical framework for generalizing the classical hidden Markov model requires thoughtful application of tools from different disciplines and difficult nonlinear optimization issues. The main factor that makes the classical hidden Markov model a versatile pattern recognition tool is the formulation of the forward and backward variables under statistical independence assumptions to compute the matching scores efficiently. By properly defining fuzzy forward and backward variables, we gain increased flexibility and meaningful matching scores with relaxed assumptions.

The concept of an optimal state sequence is used by many researchers in the field of classical hidden Markov models to design appropriate training and classification techniques. Given a sequence of observation vectors, the difficulty for finding a meaningful optimal state sequence lies with the definition of the optimal state sequence since there are several possible optimality criterion functions. Here, again, we utilize the proper definitions of the fuzzy forward and backward variables to formulate the corresponding nonlinear optimality criterion function and use a fuzzy modification of the classical Viterbi algorithm to determine the fuzzy optimal state sequence. We call this sequence a fuzzy optimal state sequence because its computation involves new parameters that are thought to serve as consistency measures (or robustness factors) that take into account the confidence scores from other states to identify the final optimal state sequence in an appropriate manner.

As for the classical hidden Markov model, there is a computational difficulty for the generalized model. Fortunately, we can overcome this difficulty for certain choices of the fuzzy measure, fuzzy integral, and fuzzy intersection operator by using a scaling procedure similar to that used for the classical model.

Manuscript received June 14, 1996; revised August 10, 1999.

The authors are with the Computer Engineering and Computer Science Department, University of Missouri–Columbia, Columbia, MO 65211 USA.

Publisher Item Identifier S 1063-6706(00)01623-4.

The most difficult aspect of the generalization is the derivation of reestimation formulas used inside the training algorithms as updating rules. We accomplish this difficult task under certain relaxed constraints; the general case remains unsolved. Our generalization opens new future directions for research in modeling sequential processes. There are many aspects that require more study.

The remainder of this paper is devoted to a detailed description of both the existing mathematical frameworks and our generalized hidden Markov model. We provide necessary background material to make the motivation for the generalization clear and to make the paper self-contained. In Section II, we provide a review of the classical hidden Markov models. We also discuss the implementation issues required for the application of interest. Section III provides a description of fuzzy measures and fuzzy integrals. This material is essential for generalizing the classical hidden Markov models. In Section IV, we formulate proper definitions for the fuzzy forward and backward variables using the notions of fuzzy measures and fuzzy integrals. We also describe our approach for solving the optimization problems required for training the fuzzy models and other implementation issues. Finally, Section V is dedicated to a summary of this study and suggestions for future outlooks.

II. Hidden Markov Models

HMM's are statistical methods (stochastic networks) that have been extremely useful for modeling sequentially changing behavior as in speech and handwriting recognition applications. This technique was applied to speech recognition problems with great success [1]. Since the recognition of handwritten words has many similarities with that of speech, researchers tried to apply this technique to handwritten word recognition.

Formally, a hidden Markov model, as defined by Rabiner in [1], "is a doubly embedded stochastic process with an underlying process that is not observable (it is hidden), but can only be observed through another set of stochastic processes that produce the sequence of observations." This means that a probabilistic function of a hidden Markov chain is a stochastic process generated by two interrelated mechanisms, an underlying Markov chain having a finite number of states, and a set of random functions, one of which is associated with each state. At discrete instants of time, the process is assumed to be in some state and an observation is generated by the random function corresponding to the current state. The underlying Markov chain then changes states according to its transition probability matrix. The observer sees only the output of the random functions associated with each state and cannot directly observe the states of the underlying Markov chain; hence, the term hidden Markov model.

In principle, the underlying Markov chain may be of any order and the outputs from its states may be multivariate random processes having some continuous joint probability density function [5]. We will restrict ourselves in this paper to consideration of Markov chains of order one, i.e., those of which the probability of transition to any state depends only upon that state and its predecessor.

A. Elements of a Discrete HMM

Most of the material and notation presented in this section are adapted from Rabiner [1]. There are a finite number, say N, of states in the model. At each time step, a new state is entered based upon a transition probability distribution which depends on the previous state (the Markovian property). If the transition parameters are held constant with respect to time the model is called stationary, otherwise it is called nonstationary. After each transition is made, an observation output symbol is produced according to a probability distribution, which depends on the current state. The probability distribution is held fixed for the state regardless of when and how the state is entered. This means that the properties of the process are held steady, except for minor fluctuations, for a certain period of time and then, at certain instances, a gradual change to another set of properties occurs. We now formally define the following model notation for a first-order discrete observation HMM:

- T Length of observation sequence (total number of time steps).
- N Number of states in the model.
- M Number of observation symbols.
- S $\{S_1, S_2, \cdots, S_N\}$, states.
- Q $\{q_1 q_2 \cdots q_T\}$, state sequence.
- V $\{v_1, v_2, v_3, \cdots, v_M\}$ discrete set of possible observations.
- q_t State visited at time t.
- A $\{a_{ij}\}$, $a_{ij} = P(q_{t+1} = S_j | q_t = S_i)$, state transition probability distribution.
- B $\{b_j(k)\}$, $b_j(k) = P(v_k \text{ at } t | q_t = S_j)$, observation symbol probability distribution in state j.
- π $\{\pi_i\}$, $\pi_i = P(q_1 = S_i)$, initial state distribution.

We use the compact notation $\lambda = (A, B, \pi)$ to indicate the complete parameter set of the model. Given the form of the hidden Markov model $\lambda = (A, B, \pi)$, there are three key problems of interest that must be solved for the model to be useful in real-world applications. These problems are the following.

B. The Classification Problem

The probability of an observation sequence $O = O_1, O_2, \cdots, O_T$ given a model λ, $P(O|\lambda)$ can be used to perform classification. The straightforward way of computing $P(O|\lambda)$ is by enumerating every possible state sequence. Assuming statistical independence of observations, it follows that:

$$P(O|\lambda) = \sum_{\text{all } Q} P(O, Q|\lambda)$$
$$= \sum_{\text{all } Q} \pi_{q_1} b_{q_1}(O_1) \prod_{t=2}^{T} a_{q_{t-1} q_t} b_{q_t}(O_t). \quad (1)$$

This method of computing $P(O|\lambda)$ requires $O(TN^T)$ computations. A method called the forward–backward procedure takes $O(TN^2)$ computations. Consider the forward variable $\alpha_t(i)$ defined as

$$\alpha_t(i) = P(O_1 O_2 \cdots O_t, q_t = S_i | \lambda). \quad (2)$$

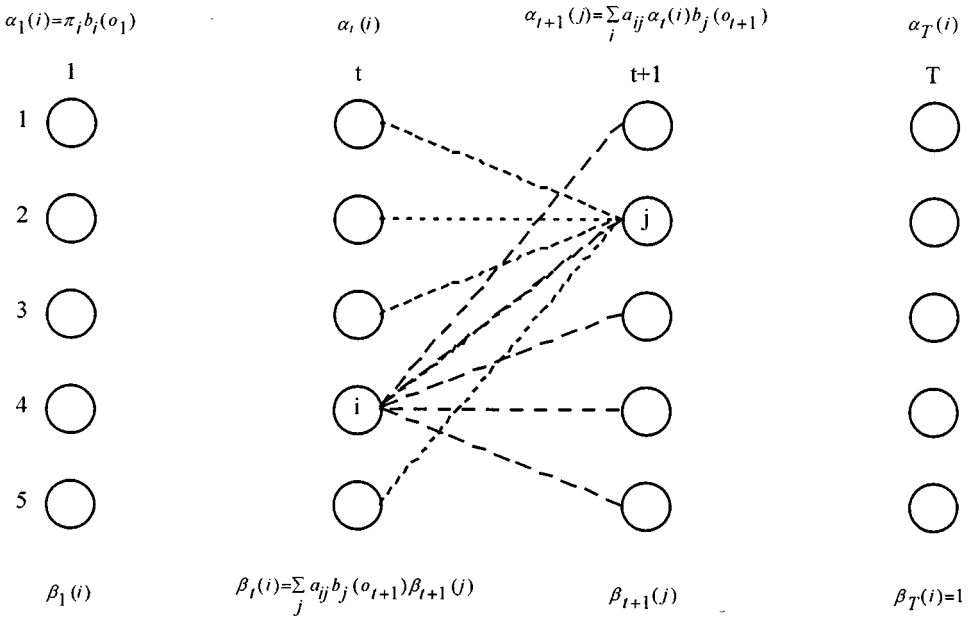


Fig. 1. Forward and backward variables computations.

We can solve for $\alpha_t(i)$ inductively as follows:

Initialization: For all $1 \le i \le N$

$$\alpha_1(i) = \pi_i b_i(O_1) \tag{3}$$

Induction: For all $1 \le t \le T-1$ and $1 \le j \le N$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i)a_{ij}\right] b_j(O_{t+1}) \tag{4}$$

Termination:

$$P(O|\lambda) = \alpha_T(i). \tag{5}$$

Equation (4) relies on assuming statistical independence. To show that, let us denote the left- and right-hand side of (4) by LHS and RHS, respectively. The derivation proceeds as follows:

$$\begin{aligned}\text{RHS} &= \left[\sum_{i=1}^{N} \alpha_t(i)a_{ij}\right] b_j(O_{t+1}) \\ &= \left[\sum_{i=1}^{N} P(O_1, O_2, \cdots, O_t, q_t = S_i) \right. \\ &\quad \left. \cdot\, P(q_{t+1} = S_j | q_t = S_i)\right] P(O_{t+1}|q_{t+1} = S_j) \\ &= \left[\sum_{i=1}^{N} P(O_1, O_2, \cdots, O_t | q_t = S_i) P(q_t = S_i) \right. \\ &\quad \left. \cdot\, P(q_{t+1} = S_j | q_t = S_i)\right] P(O_{t+1}|q_{t+1} = S_j). \end{aligned} \tag{6}$$

Assuming q_{t+1} is independent of $O_1, O_2, \cdots, O_t$, then

$$\begin{aligned}\text{RHS} &= \left[\sum_{i=1}^{N} P(O_1, O_2, \cdots, O_t, q_{t+1} = S_j | q_t = S_i) \right. \\ &\quad \left. \cdot\, P(q_t = S_i)\right] P(O_{t+1}|q_{t+1} = S_j) \\ &= [P(O_1, O_2, \cdots, O_t, q_{t+1} = S_j)] P(O_{t+1}|q_{t+1} = S_j) \\ &= [P(O_1, O_2, \cdots, O_t | q_{t+1} = S_j) P(q_{t+1} = S_j)] \\ &\quad \cdot P(O_{t+1}|q_{t+1} = S_j). \end{aligned} \tag{7}$$

Assuming O_{t+1} is independent of $O_1, O_2, \cdots, O_t$, then

$$\begin{aligned}\text{RHS} &= P(O_1, O_2, \cdots, O_t, O_{t+1} | q_{t+1} = S_j) P(q_{t+1} = S_j) \\ &= P(O_1, O_2, \cdots, O_t, O_{t+1}, q_{t+1} = S_j) \\ &= \alpha_{t+1}(j) \\ &= \text{LHS}. \end{aligned} \tag{8}$$

In a similar manner, we consider a backward variable $\beta_t(i)$ defined as

$$\beta_t(i) = P(O_{t+1}O_{t+2}\cdots O_T | q_t = S_i, \lambda) \tag{9}$$

and again we can solve for $\beta_t(i)$ inductively as follows.

Initialization for all $1 \le i \le N$

$$\beta_T(i) = 1. \tag{10}$$

Induction for all $1 \le t \le T-1$ and $1 \le i \le N$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j). \tag{11}$$

Termination fFor any t such that $1 \le t \le T-1$

$$P(O|\lambda) = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j). \tag{12}$$

Fig. 1 illustrates the operations required to compute the forward and backward variables.

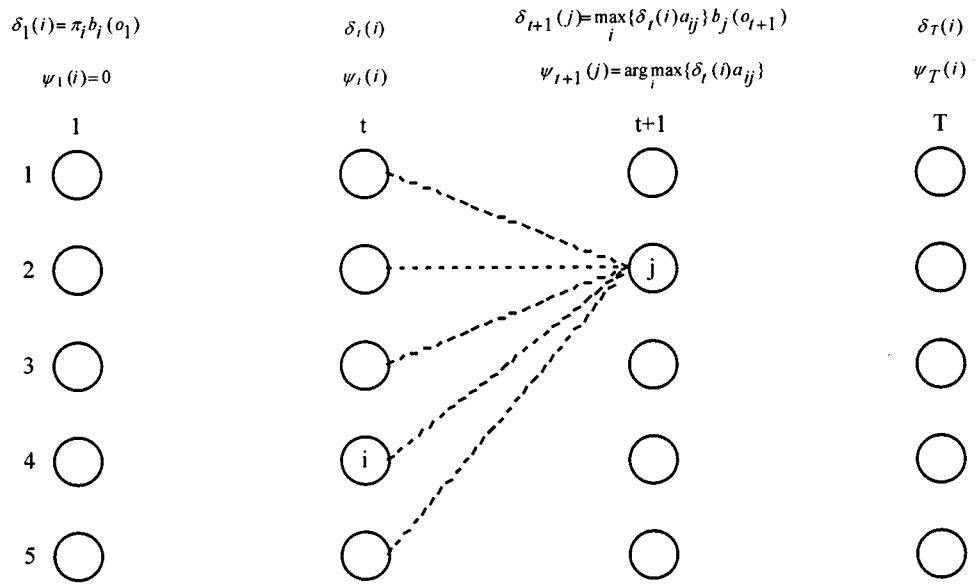


Fig. 2. Viterbi algorithm.

C. The Optimal-State Sequence Problem

There are several possible ways of finding the optimal state sequence associated with the given observation sequence. The difficulty lies with the definition of the optimal state sequence, i.e., there are several possible optimality criteria. One possible optimality criterion is to choose the states which are individually most likely. This optimality criterion maximizes the expected number of correct individual states. To implement the solution we define the variable

$$\gamma_t(i) = P(q_t = S_i|O, \lambda) \tag{13}$$

$\gamma_t(i)$ can be computed as

$$\begin{aligned}\gamma_t(i) &= \alpha_t(i)\beta_t(i)/P(O|\lambda) \\ &= \alpha_t(i)\beta_t(i) \Big/ \sum_{j=1}^{N} \alpha_t(j)\beta_t(j).\end{aligned} \tag{14}$$

Using $\gamma_t(i)$, we can solve for the individually most likely state q_t at time t, $1 \le t \le T$, as

$$q_t = \underset{1\le i\le N}{\operatorname{argmax}}\{\gamma_t(i)\}. \tag{15}$$

The major problem with the above criterion and solution occurs when there are disallowed transitions. In this case, the obtained optimal state sequence may, in fact, be an impossible state sequence. This drawback points to the necessity of global constraints on the derived optimal-state sequence. An optimality criterion of this type is to find the state sequence with the highest probability, i.e., to maximize $P(O, Q|\lambda)$. A formal technique for finding this solution exists and is called the Viterbi algorithm. This algorithm defines a quantity

$$\delta_t(i) = \max_{q_1,q_2,\cdots q_{t-1}} P(q_1, q_2, \cdots, q_t = S_i, \quad O_1O_2\cdots O_t|\lambda). \tag{16}$$

Similarly $\delta_{t+1}(i)$ can be computed inductively and the procedure can be stated as follows:

Initialization for $1 \le i \le N$

$$\delta_1(i) = \pi_i b_i(O_1) \tag{17}$$
$$\psi_1(i) = 0. \tag{18}$$

Recursion for $2 \le t \le T$ and $1 \le j \le N$

$$\delta_t(j) = \max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]b_j(O_t) \tag{19}$$
$$\psi_t(j) = \underset{1\le i\le N}{\operatorname{argmax}}[\delta_{t-1}(i)a_{ij}]. \tag{20}$$

Termination

$$P^* = \max_{1\le i\le N}[\delta_T(i)] \tag{21}$$
$$q_T^* = \underset{1\le i\le N}{\operatorname{argmax}}[\delta_T(i)]. \tag{22}$$

Backtracking for all $1 \le t \le T-1$

$$q_t^* = \psi_{t+1}(q_{t+1}^*). \tag{23}$$

Fig. 2 below illustrates the sequence of operations required for the Viterbi algorithm.

D. The Training Problem

Given any finite observation sequence as training data, we cannot optimally train the model. We can, however, choose A, B, and π such that $P(O|\lambda)$ is locally maximized. The Baum–Welch method is an iterative algorithm that uses the forward and backward probabilities to solve the problem of training by parameter estimation. To implement the solution we first define the variable $\gamma_t(i)$ the probability of being in state S_i at time t and then define $\xi_t(i, j)$ the probability of being in

state S_i at time t and state S_j at time $t+1$, given the model and the observation sequence, i.e.

$$\begin{aligned}\gamma_t(i) &= P(q_t = S_i|O, \lambda) \\ &= \alpha_t(i)\beta_t(i)/P(O|\lambda) \qquad (24)\\ \xi_t(i, j) &= P(q_t = S_i, q_{t+1} = S_j|O, \lambda) \\ &= \alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)/P(O|\lambda). \qquad (25)\end{aligned}$$

Now we have

$$\sum_{t=1}^{T-1} \xi_t(i,j) = \text{expected number of transitions made (from } S_i \text{ to } S_j) \qquad (26)$$

$$\sum_{t=1}^{T-1} \gamma_t(i) = \text{expected number of transitions from } S_i. \qquad (27)$$

The Baum–Welch reestimation formulas for A, B, and π are

$$\overline{\pi}_i = \gamma_1(i) \qquad (28)$$

$$\overline{a}_{ij} = \sum_{t=1}^{T-1} \xi_t(i, j) \Big/ \sum_{t=1}^{T-1} \gamma_t(i) \qquad (29)$$

$$\overline{b}_j(k) = \sum_{t=1,\, O_t=k}^{T} \gamma_t(j) \Big/ \sum_{t=1}^{T} \gamma_t(j). \qquad (30)$$

Iterative application of these formulas will converge to a local maxima of $P(O|\lambda)$.

E. Continuous Observation Densities in HMM

For most applications, the observations are continuous signals (or vectors). Vector quantization of these continuous signals can degrade performance significantly. Moreover, the codebooks generated by the quantization process are constructed using training data from all classes. When a new class of shapes is added, we need to reconstruct the codebook and retrain all system modules. On the other hand, for HMM's with continuous observation densities we do not need to train the system from the beginning since there is no codebook to be constructed. We only need to train the newly added class. Hence, HMM's with continuous observation densities offer some advantages over discrete HMM's.

The class of densities $B = \{b_j(\cdot)\}$ we consider is the class of mixtures of the form

$$b_j(O_t) = \sum_{m=1}^{M} \omega_{jm} f_{jm}(O_t) \qquad (31)$$

where $f_{jm}(O_t) = N(O_t, \mu_{jm}, U_{jm})$ is a multivariate Gaussian density with mean μ_{jm} and covariance matrix U_{jm}. The mixture gains ω_{jm} satisfy the stochastic constraint

$$\sum_{m=1}^{M} \omega_{jm} = 1 \qquad 1 \le j \le N \qquad (32)$$

where $\omega_{jm} \ge 0$, $1 \le j \le N$ and $1 \le m \le M$.

The reestimation formulas [6], [7] for the coefficients

$$\overline{\omega}_{jm} = \frac{\sum_{t=1}^{T-1}\sum_{i=1}^{N} \alpha_t(i)a_{ij}\omega_{jm}f_{jm}(O_{t+1})\beta_{t+1}(j)}{\sum_{m=1}^{M}\sum_{t=1}^{T-1}\sum_{i=1}^{N} \alpha_t(i)a_{ij}\omega_{jm}f_{jm}(O_{t+1})\beta_{t+1}(j)} \qquad (33)$$

$$\overline{\mu}_{jm} = \frac{\sum_{t=1}^{T-1}\sum_{i=1}^{N} [\alpha_t(i)a_{ij}\omega_{jm}f_{jm}(O_{t+1})\beta_{t+1}(j)]O_{t+1}}{\sum_{m=1}^{M}\sum_{t=1}^{T-1}\sum_{i=1}^{N} [\alpha_t(i)a_{ij}\omega_{jm}f_{jm}(O_{t+1})\beta_{t+1}(j)]}. \qquad (34)$$

The reestimation formulas [1] for the coefficients of the mixture densities can also be rewritten as [see, also, (35) at the bottom of the page]

$$\overline{w}_{jm} = \frac{\sum_{t=1}^{T} \gamma_t(j, m)}{\sum_{k=1}^{M}\sum_{t=1}^{T} \gamma_t(j, k)} \qquad (36)$$

$$\overline{\mu}_{jm} = \frac{\sum_{t=1}^{T} \gamma_t(j, m)O_t}{\sum_{t=1}^{T} \gamma_t(j, m)} \qquad (37)$$

$$\overline{U}_{jm} = \frac{\sum_{t=1}^{T} \gamma_t(j, m)(O_t - \mu_{jm})(O_t - \mu_{jm})'}{\sum_{t+1}^{T} \gamma_t(j, m)} \qquad (38)$$

$$\overline{U}_{jm} = \frac{\sum_{t=1}^{T-1}\sum_{i=1}^{N} [\alpha_t(i)a_{ij}\omega_{jm}f_{jm}(O_{t+1})\beta_{t+1}(j)](O_{t+1} - \mu_{jm})(O_{t+1} - \mu_{jm})'}{\sum_{m=1}^{M}\sum_{t=1}^{T-1}\sum_{i=1}^{N} [\alpha_t(i)a_{ij}\omega_{jm}f_{jm}(O_{t+1})\beta_{t+1}(j)]} \qquad (35)$$

$$\gamma_t(j,m) = \left[\frac{\alpha_t(j)\beta_t(j)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)}\right]\left[\frac{w_{jm}f_{jm}(O_t)}{\sum_{k=1}^{M}w_{jk}f_{jk}(O_t)}\right] \quad (39)$$

where $\gamma_t(j, m)$ is the probability of being in state j at time t with the kth mixture component accounting for O_t.

F. Initial Estimates of HMM Parameters

Rabiner described a procedure for providing good initial estimates of the model parameters called the segmental K-means algorithm [1], [8], [9]: An initial model estimate can be chosen randomly, uniformly, or on the basis of any available model that is appropriate to the data. Following model initialization, the set of training observation sequences is segmented into states based on the current model. This segmentation is achieved by finding the optimal state sequence via the Viterbi algorithm. In the case where we are using discrete symbol densities, each of the observation vectors within a state is coded using the M-codeword codebook and $b_j(k)$ is updated as the number of vectors with codebook index k in state j divided by the number of vectors in state j.

In the case that we are using continuous observation densities, the segmental K-means procedure is used to cluster the observation vectors within each state S_j into a set of M clusters, where each cluster represents one of the M mixtures of the $b_j(O_t)$ density. From the clustering, an updated set of model parameters is derived as follows:

- ω_{jm} number of vectors classified in cluster m of state j divided by the number of vectors in state j
- μ_{jm} sample mean of the vectors classified in cluster m of state j
- U_{jm} sample covariance matrix of the vectors classified in cluster m of state j.

Updated estimates of the a_{ij} coefficients can be obtained by counting the number of transitions from state i to j and dividing it by the number of transitions from state i to any state. An updated model is obtained from the new model parameters and the formal reestimation procedure is used to reestimate all model parameters. The resulting model is then compared to the previous model by computing a distance score that reflects the statistical similarity of the two HMM's. If the model distance score exceeds a threshold, then the old model is replaced by the new reestimated model and the overall training loop is repeated, otherwise convergence is assumed and the final model parameters are saved.

G. Implementation Issues for HMM's

For a sufficiently long observation sequence, the dynamic range of $\alpha_t(i)$ computation will exceed the precision range of any computer. There exists a scaling procedure which is used to multiply $\alpha_t(i)$ by a scaling coefficient that is independent of i. A similar scaling is done to the $\beta_t(i)$ coefficients since these also tend to approach zero. At the end of the computation, the scaling coefficients are canceled out [1]. When using the Viterbi algorithm to determine the optimal state sequence, no scaling is required if use logarithms.

Another implementation issue is related to the modification of the reestimation procedure to handle multiple observation sequences. Let the set of the K training observation sequences be $O = \{O^1, O^2, \cdots, O^K\}$, where $O^k = (O_1^k, O_2^k, \cdots, O_{T_k}^k)$ is the kth observation sequence. Assuming that each observation sequence is independent of every other, the goal is to adjust the parameters of the model λ to maximize

$$P(O|\lambda) = \prod_{k=1}^{K} P(O^k|\lambda) = \prod_{k=1}^{K} P^k. \quad (40)$$

The modified reestimation formula for the transition probabilities is

$$\overline{a}_{ij} = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T_k-1}\alpha_t^k(i)a_{ij}b_j(O_{t+1}^k)\beta_t^k(j)}{\sum_{j=1}^{N}\sum_{k=1}^{K}\sum_{t=1}^{T_k-1}\alpha_t^k(i)a_{ij}b_j(O_{t+1}^k)\beta_t^k(j)} \quad (41)$$

using the scaled forward and backward variables. Similar results are obtained for the other parameters.

III. Fuzzy Integrals

Fuzzy integrals are nonlinear functionals that can be used to combine multiple sources of uncertain information. The integrals are evaluated over a set of information sources. The function being integrated supplies a confidence value for a particular hypothesis from the standpoint of each individual source of information. A distinguishing characteristic of fuzzy integrals is that they utilize information concerning not only the worth or importance of the individual sources but also information concerning the worth or importance of subsets of these sources to arrive at a reasonable numeric confidence value for the particular hypothesis or decision under consideration. Recently, fuzzy integrals have been proven to be quite useful in many pattern recognition applications such as automatic target recognition (ATR), handwriting recognition, nonlinear image filtering, and multiple classifier fusion [10]–[14].

Fuzzy integrals are defined with respect to fuzzy measures [4]. The key property of fuzzy measures is monotonicity with respect to set inclusion. This property is far weaker than the usual additivity property of probability measures. A probability measure is a particular case of a fuzzy measure since the additivity property is a special case of the monotonicity property. Other examples of fuzzy measures are the belief and plausibility measures defined in Dempster–Shafer belief theory.

In the following sections, we describe fuzzy measures, the formulation of fuzzy integrals, and the basic components required for our generalization of the classical HMM's.

A. Fuzzy Measures

The additivity hypothesis of the probability measure is not well-suited for modeling systems that manifest a high degree of interdependencies among sources of information. Sugeno [13] introduced the concept of a fuzzy measure as a more flexible model.

Let X be an arbitrary set and Ω a σ algebra of subsets of X. A set function $g: \Omega \rightarrow [0, 1]$ defined on Ω, which has the following properties is called a fuzzy measure.

Boundary Conditions

$$g(\phi) = 0, \qquad g(X) = 1. \tag{42}$$

Monotonicity

$$\text{If } A, B \subset \Omega \text{ and } A \subset B, \text{ then } g(A) \leq g(B). \tag{43}$$

Continuity

$$\text{If } F_n \in \Omega \text{ for } 1 \leq n < \infty \text{ and the sequence } \{F_n\} \text{ is monotone (in the sense of inclusion), then } \lim_{n\rightarrow\infty} g(F_n) = g(\lim_{n\rightarrow\infty} F_n). \tag{44}$$

By the nature of the definition of a fuzzy measure g, the measure of the union of two disjoint subsets cannot be directly computed from the component measures. In light of this, Sugeno introduced the so-called λ-fuzzy measure satisfying the following additional property for all $A, B \subset X$ with $A \cap B = \phi$:

$$g(A \cup B) = g(A) + g(B) + \lambda g(A) g(B), \quad \text{for some } \lambda > -1. \tag{45}$$

Let $X = \{x_1, x_2, \cdots, x_n\}$ be a finite set and let $g^i = g(\{x_i\})$. The values g^i are referred to as the densities of the λ-fuzzy measure g. The value of λ can be found from the equation $g(X) = 1$, which is equivalent to solving

$$\lambda + 1 = \prod_{i=1}^{n} (1 + \lambda g^i). \tag{46}$$

B. Sugeno Integral

The Sugeno fuzzy integral combines objective evidence for a hypothesis with the prior expectation of the importance of that evidence to the hypothesis. Using the notion of fuzzy measures, Sugeno originally defined the concept of fuzzy integrals as follows.

Let (X, Ω) be a measurable space and let $h: X \rightarrow [0, 1]$ be an Ω-measurable function. The Sugeno fuzzy integral over $A \subseteq X$ of the function h with respect to a fuzzy measure g is defined by

$$\begin{aligned} \int_A h(x) \circ g(\cdot) &= \sup_{E \subseteq X} \left[\min \left(\min_{x \in E} h(x), g(A \cap E) \right) \right] \\ &= \sup_{\alpha \in [0, 1]} [\min(\alpha, (g(A \cap F_\alpha))] \end{aligned} \tag{47}$$

where $F_\alpha = \{x | h(x) \geq \alpha\}$. The calculation of the Sugeno fuzzy integral when X is a finite set is easily given [10], [14]. Suppose $h(x_1) \geq h(x_2) \geq \cdots \geq h(x_n)$, (if not, X is rearranged so that this relation holds). Then a Sugeno fuzzy integral e with respect to a fuzzy measure g over X can be computed by

$$e = \max_{i=1}^{n} [\min(h(x_i), g(A_i))] \tag{48}$$

where $A_i = \{x_1, x_2, \cdots, x_i\}$.

When g is the λ-fuzzy measure, the values of $g(A_i)$ can be computed recursively as

$$g(A_1) = g(\{x_1\}) = g^1 \tag{49}$$

$$g(A_i) = g^i + g(A_{i-1}) + \lambda g^i g(A_{i-1}), \quad \text{for } 1 \leq i \leq n. \tag{50}$$

Thus, the calculation of the fuzzy integral with respect to a λ-fuzzy measure only requires the knowledge of the fuzzy densities.

A fuzzy integral over a fuzzy set $\tilde{A}$ is defined by

$$\int_A h(x) \circ g(\cdot) = \int_X [h_A(x) \wedge h(x)] \circ g(\cdot) \tag{51}$$

where $h_A(x)$ is the membership function of the fuzzy set $\tilde{A}$.

C. Choquet Integral

The original definition given by Sugeno [13] for the fuzzy integral is not a proper extension of the usual (Lebesgue) integral, in the sense that the Lebesgue integral is not recovered when the measure is additive. To avoid this drawback, Murofushi and Sugeno [15] proposed the so-called Choquet integral, referring to a functional defined by Choquet in a different context. In addition to this property, Grabisch [16], [17] showed that the Choquet integral shares many important properties with the Sugeno integral.

Let h and g be defined as for the Sugeno integral. The Choquet integral is defined by

$$\int_X h(x) \circ g(\cdot) = \int_0^1 g(A_\alpha)\, d\alpha \tag{52}$$

where $A_\alpha = \{x | h(x) > \alpha\}$.

If X is a discrete set, the Choquet integral can be computed as follows:

$$e = \sum_{i=1}^{n} [h(x_i) - h(x_{i-1})] g_i^n \tag{53}$$

where $h(x_1) \leq h(x_2) \leq \cdots < h(x_n)$

$$h(x_0) = 0$$

and

$$g_i^j = \begin{cases} g(\{x_i, x_{i+1}, \cdots, x_j\}), & i \leq j \\ 0, & \text{otherwise.} \end{cases}$$

Since $g(A_\alpha)$ is a monotonic nonincreasing function of α, it is also possible to redefine Choquet integral as

$$e = \int_0^1 \alpha\, dg(A_\alpha) \tag{54}$$

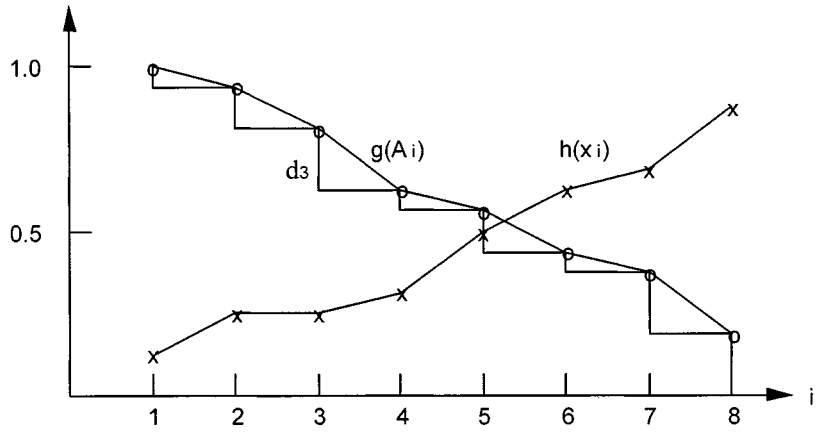


Fig. 3. Computation of Sugeno and Choquet integrals.

with the same assumptions as before. Define $\delta_i = g_i^n - g_{i+1}^n$. Then the computation for the finite set case is given by

$$e = \sum_{i=1}^{n} h(x_i)[g_i^n - g_{i+1}^n]. \tag{55}$$

If g is a probability measure, then $[g_i^n - g_{i+1}^n] = g_i^i$ and the expectation is a weighted sum that is independent of the ordering of the x_i's. In this sense the Lebesgue integral is recovered from the Choquet integral when the measure is additive (probability measure). Fig. 3 illustrates graphically the computation of the Sugeno and the Choquet integrals

$$h(x_1) \le h(x_2) \le \cdots \le h(x_N) \tag{56}$$

$$A_i = \{x_i, x_{i+1}, \cdots, x_N\} \tag{57}$$

$$e_{\text{Sugeno}} = \max_i[\min(h(x_i), g(A_i))] \tag{58}$$

$$e_{\text{Choquet}} = \sum_{i=1}^{N} h(x_i)[g(A_i) - g(A_{i+1})] = \sum_{i=1}^{N} h(x_i)\, d_i. \tag{59}$$

D. Conditional Fuzzy Measures

Conditional fuzzy measures are similar to conditional probabilities [13]. Let X and Y be two universes. A conditional fuzzy measure on Y with respect to X is a fuzzy measure $\sigma_Y(\cdot|x)$ on Y for any fixed $x \in X$. A fuzzy measure g_Y on Y is induced by $\sigma_Y(\cdot|x)$ and a fuzzy measure g_X as follows.

For $B \subset Y$,

$$g_Y(B) = \int_X \sigma_Y(B|x) \circ g_X(\cdot). \tag{60}$$

Now, g_X corresponds to an *a priori* probability and $\sigma_Y(B|x)$ to a conditional probability. For this reason, g_X may be called an *a priori* fuzzy measure. Note that $\sigma_Y(B|x)$ measures the grade of fuzziness of the statement, "One of the elements of B results because of x" [13]. Fig. 4 below illustrates the computation graphically.

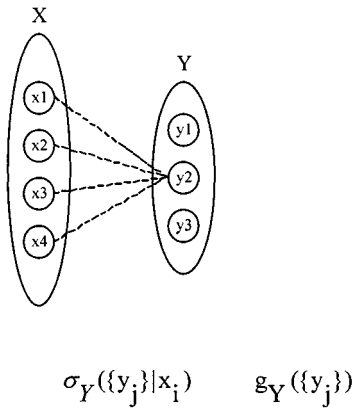


Fig. 4. Conditional fuzzy measures.

E. Other Fuzzy Measures

The basic notion of a fuzzy integral using the Sugeno measure has been demonstrated to be a useful tool. It can be improved by using more general measures or by using different fuzzy aggregation operators in the definition of the fuzzy integral. Recently, Keller and Tahani have extended the fuzzy integral information fusion approach to a large family of measures called S-decomposable measures [10], [14]. Given a triangular conorm S, an S-decomposable measure g has the property: IF $A \cap B = \phi$, THEN

$$g(A \cup B) = S(g(A), g(B)). \tag{61}$$

Possibility measures are simple examples of such S-decomposable measures where S is the maximum operator. An important property of this class is that the measure of an arbitrary set of information sources can be computed if the densities are known, as with the Sugeno measures. Many other S-decomposable measures can actually be constructed by the definition from a set of density values for a given t-conorm S if the boundary conditions hold, i.e., one must guarantee that $g(X) = 1$. This will clearly happen if one of the densities has the value one. This follows simply from the fact that

$$g(X) = g(\{x_1\} \cup \{x_2\} \cup \cdots \cup \{x_n\}) = S(g(\{x_1\}), g(\{x_2\}), \cdots, g(\{x_n\})) \tag{62}$$

and $g(\{x_i\})$ is the density value for $i = 1, \cdots, n$.

IV. Fuzzy Integral Extensions of Hidden Markov Models

Since the definition of the Markov property is a statement about conditional expectations, our generalization relies heavily on the use of the conditional fuzzy measure, which is one of the many things the measure-theoretic framework provides [18], [19]. Let us formally define the model notation for our extension to a fuzzy HMM, $\hat{\lambda} = (\hat{A}, \hat{B}, \hat{\pi})$ as follows:

T, O, N, q_t, and S	Same as for the standard HMM.
Ω	Space of observation vectors.
$X = \{x_1, x_2, \cdots, x_N\}$	States at time t, i.e., nodes at time slot t in the lattice structure resulting from folding the states of the model through time.
$Y = \{y_1, y_2, \cdots, y_N\}$	States at time $t+1$.
$\hat{\pi}_S(\cdot)$	Fuzzy measure on S, we refer to this measure as the initial state fuzzy measure.
$\hat{\pi}_i = \hat{\pi}_S^i = \hat{\pi}_S(\{S_i\})$	Initial state fuzzy density.
$\hat{\pi} = [\hat{\pi}_i]$	Vector of initial state fuzzy densities.
$\hat{b}_j(\,)$	Conditional fuzzy measure on Ω with respect to state S_j, we refer to this as the symbol fuzzy measure for state S_j.
$\hat{b}_j(O_t)$	Symbol fuzzy density.
$\hat{B} = [\hat{b}_i(O_t)]$	Matrix of symbol densities.
$\hat{a}_Y(\cdot\|X)$	Conditional fuzzy measure on Y with respect to $x \in X$, we refer to this measure as the transition fuzzy measure.
$\hat{a}_{ij} = \hat{a}_Y(\{y_j\}\|x_i)$	Transition fuzzy density.
$\hat{A} = [\hat{a}_{ij}]$	Matrix of the transition fuzzy densities.

Our interpretation for the initial state fuzzy density is that $\hat{\pi}_i$ measures the grade of certainty of the statement that the initial state is S_i, i.e., $q_1 = S_i$. An extension of this interpretation could be given for the initial state fuzzy measure as follows: for any $G \subset S$, $\hat{\pi}_S(G)$ measures the grade of certainty of the statement that the initial state is contained in G, i.e., $q_1 \in G$.

The interpretation for the symbol fuzzy density is that $\hat{b}_j(O_t)$ measures the grade of certainty of the statement that we observed O_t given that we are visiting state S_j. An extension of this interpretation could be given for the symbol fuzzy measure as follows: for a set $H \subset \Omega$, $\hat{b}_j(H)$ measures the grade of certainty of the statement that any of the vectors contained in H is observed given that we are visiting state S_j.

For each state S_j, we have a symbol fuzzy measure $\hat{b}_j(\,)$ on the space of observation vectors Ω. At a given time slot t, $1 \leq t \leq T$, the values of the symbol fuzzy densities

$$\hat{b}_1(O_t), \hat{b}_2(O_t), \cdots, \hat{b}_N(O_t)$$

define a fuzzy set over the set of N states. Therefore, we can construct a total of T different fuzzy sets over the set of N states. In this sense, $\hat{b}_j(O_t)$ can also be interpreted as the membership value of observation O_t in state S_j.

The interpretation for the transition fuzzy density is that $\hat{a}_{ij}$ measures the grade of certainty of the statement that visiting y_j (state S_j at time $t+1$) results because of visiting x_i (state S_i at time t). An extension of this interpretation could be given for the transition fuzzy measure as follows: for any $F \subset Y$ and $x \in X$, $\hat{a}_Y(F|x)$, measures the grade of certainty of the statement that visiting one of the elements of F results because of visiting x (one of the states at time t).

A. Fuzzy Formulation of the Forward Variables

Let $\Omega_{1,t}$ denote the space of observation sequence from time slot 1 to time slot t. Let $\Omega_X = \Omega_{1,t} \times X$ denote the Cartesian product of $\Omega_{1,t}$ and (recall that X denotes the states at time t). Let $\hat{\alpha}_{\Omega_X}\colon 2^{\Omega_X} \to [0,1]$ be a fuzzy measure on the space $(\Omega_X, 2^{\Omega_X})$ where for any $E \subset X$, $\hat{\alpha}_{\Omega_X}(\{O_1 \cdots O_t\} \times E)$ measures the grade of certainty of the statement that we observed $O_1 O_2 \cdots O_t$ and we are visiting a state that is contained in E. For a given observation sequence $O_1 O_2 \cdots O_t$ and a state x_i (state S_i at time t) we let $\hat{\alpha}^i_{\Omega_X} = \hat{\alpha}_{\Omega_X}(\{O_1 \cdots O_t\} \times \{x_i\})$ be the fuzzy density for this measure. We will refer to this density as the forward fuzzy variable and denote it by $\hat{\alpha}_t(i)$. The forward fuzzy variable $\hat{\alpha}_t(i)$ measures the grade of certainty of the statement that we observed $O_1 O_2 \cdots O_t$ and we are visiting x_i (state S_i at time t).

Initially, at $t = 1$, the forward fuzzy variables can be computed from the initial state densities and the membership functions by

$$\hat{\alpha}_1(i) = \hat{\pi}_i \wedge \hat{b}_i(O_1) \tag{63}$$

where "$\wedge$" is a fuzzy intersection operator [20].

At any time, a fuzzy measure $\hat{\alpha}_{\Omega_Y}$ on $\Omega_Y = \Omega_{1,t+1} \times Y$ can be constructed from its constituent forward fuzzy variables.

The forward fuzzy variables are computed recursively as

$$\begin{aligned}\hat{\alpha}_{t+1}(j) &= \hat{\alpha}^j_{\Omega_Y} = \hat{\alpha}_{\Omega_Y}(\{O_1 \cdots O_{t+1}\} \times \{y_j\}) \\ &= \int_X \hat{a}_Y(\{y_j\}|x) \circ \hat{\alpha}_{\Omega_X}(\{O_1 \cdots O_t\}, \cdot) \\ &\quad \wedge \hat{b}_j(O_{t+1}).\end{aligned} \tag{64}$$

The above expression offers a more flexible method of computing forward variables than for the standard HMM. Recall that in the derivation of (4) for computing forward variables in the standard HMM there are two assumptions of conditional statistical independence: that the observation at time $t+1$, O_{t+1} is independent of the previous observations $O_1, O_2, \cdots, O_t$ and that the states at time $t+1$ are independent of the same observations $O_1, O_2, \cdots, O_t$. The latter statistical independence assumption is an assumption that the joint measure $P(O_1, O_2, \cdots, O_t, q_{t+1} = S_j)$ can be written as the product $P(O_1, O_2, \cdots, O_t)P(q_{t+1} = S_j)$.

In the fuzzy model, the corresponding assumption is be that the joint measure $\hat{\alpha}_{\Omega_Y}(\{O_1 \cdots O_t\} \times \{y_j\})$ can be written as a combination of two measures defined on $O_1, O_2, \cdots, O_t$ and on the states, respectively. However, we make no assumption that the measures can be decomposed. It is in this sense that the assumption of statistical independence is relaxed.

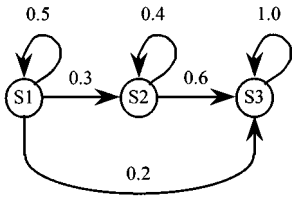


Fig. 5. A sample three-state left-to-right model.

To illustrate the above claimed advantage of our proposed formulation, consider a discrete left-to-right model shown in Fig. 5 and defined as follows:

$$V = \{H, L\}$$
$$S = \{S_1, S_2, S_3\}$$
$$A = \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.0 & 0.4 & 0.6 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}$$

where the training observation sequences are assumed to represent the nonincreasing binary signals given by

$$O^1 = H, H, H$$
$$O^2 = H, H, L$$
$$O^3 = H, L, L$$
$$O^4 = L, L, L.$$

If the symbol probabilities are estimated (initially) according to how many times the symbol appears at time slot 1, 2, and 3, assuming that each time slot may correspond to one of the three states then

$$B = \{b_{jk}\} = \begin{bmatrix} 0.75 & 0.25 \\ 0.50 & 0.50 \\ 0.25 & 0.75 \end{bmatrix}.$$

Given the above set of parameters, we would like to compare the behavior of the HMM and GHMM in terms of $P(O|\lambda)$ and $\hat{P}(O|\lambda)$ when O is:

1) a training sequence such as $O = H, H, H$;
2) a nonmonotonic sequence such as $O = H, L, H$.

If we use HMM (that assumes statistical independence when computing the forward variables)

$$P(H, H, H) = 0.210$$
$$P(H, L, H) = 0.121.$$

The difference between the above two scores is

$$D = P(H, H, H) - P(H, L, H) = 0.089.$$

The value of $P(H, L, H)$ is high compared to the ideal value that is supposed to be close to zero. This result is mainly due to the statistical independence assumed to simplify computations.

Now, if we use the GHMM with the possibility measure (the Choquet integral) and multiplication as the intersection operator

$$\hat{P}(H, H, H) = 0.182$$
$$\hat{P}(H, L, H) = 0.085$$

the difference between the above two scores is

$$\hat{D} = \hat{P}(H, H, H) - \hat{P}(H, L, H) = 0.097.$$

This illustrates that the fuzzy model can have better performance than the classical model. It is worth noting that here we used the same values for the corresponding classical and fuzzy parameters and the most pessimistic fuzzy measure (possibility measure). In the second paper, we compare the classical and fuzzy models using real data.

B. Fuzzy Formulation of the Backward Variable

Let $\hat{\beta}_{\Omega_{t+1,T}}(\cdot|x)$ be a conditional fuzzy measure on $\Omega_{t+1,T}$ with respect to x, where for any subsequence $O_{t+1}O_{t+2}\cdots O_T$, $\hat{\beta}_{\Omega_{t+1,T}}(\cdot|x_i)$ measures the fuzziness of the statement that observing $O_{t+1}O_{t+2}\cdots O_T$ results because of visiting x_i (state S_i at time t).

A conditional fuzzy measure $\hat{\beta}_{\Omega_{t+1,T}}(\cdot|x)$ can be computed from the conditional fuzzy measure $\hat{\beta}_{\Omega_{t+2,T}}(\cdot|y)$ and the transition fuzzy measure $\hat{a}_Y(\cdot|x)$ as follows:

$$\begin{aligned}\hat{\beta}_{\Omega_{t+1,T}}(\{O_{t+1}\cdots O_T\}|x) &= \int_Y [\hat{\beta}_{\Omega_{t+2,T}}(\{O_{t+2}\cdots O_T\}|y) \wedge \hat{b}_j(O_{t+1})] \circ \hat{a}_Y(\cdot|x) \\ &= \int_F \hat{\beta}_{\Omega_{t+2,T}}(\{O_{t+2}\cdots O_T\}|y) \circ \hat{a}_Y(\cdot|x) \qquad (65)\end{aligned}$$

where $\tilde{F}$ is the fuzzy subset of Y given by

$$\tilde{F} = \sum_{j=1}^{N} \hat{b}_j(O_{t+1})/y_j. \qquad (66)$$

We refer to $\hat{\beta}_{\Omega_{t+1,T}}(\cdot|x)$ as the backward fuzzy variable and denote it by $\hat{\beta}_t(i)$. Fig. 6 illustrates the computations of the fuzzy forward and backward variables.

These formulas define a class of generalizations of classical HMM's, one for each type of fuzzy measure, fuzzy integral, and fuzzy intersection operator. If the Choquet integral is chosen with respect to a probability measure and multiplication is used as the intersection operator, then these formulas represent the classical HMM. For any specific choice of measure and integral, there are many implementation issues to consider, both in the training and testing phases. In the next section, we consider the case of the Choquet integral with respect to an arbitrary fuzzy measure and with multiplication as the intersection operator

$$\begin{aligned}\hat{\alpha}_{t+1}(j) &= \hat{\alpha}_{\Omega_Y}(\{O_1\cdots O_{t+1}\} \times \{y_j\}) \\ &= \int_X \hat{a}_Y(\{y_j\}|x) \circ \hat{\alpha}_{\Omega_X}(\{O_1\cdots O_t\}, \cdot) \\ &\quad \wedge \hat{b}_j(O_{t+1}) \\ \hat{\beta}_t(i) &= \hat{\beta}_{\Omega_{t+1,T}}(\{O_{t+1}\cdots O_T\}|x_i) \\ &= \int_Y [\hat{\beta}_{\Omega_{t+2,T}}(\{O_{t+2}\cdots O_T\}|y) \wedge \hat{b}_j(O_{t+1})] \\ &\quad \circ \hat{a}_Y(\cdot|x).\end{aligned}$$

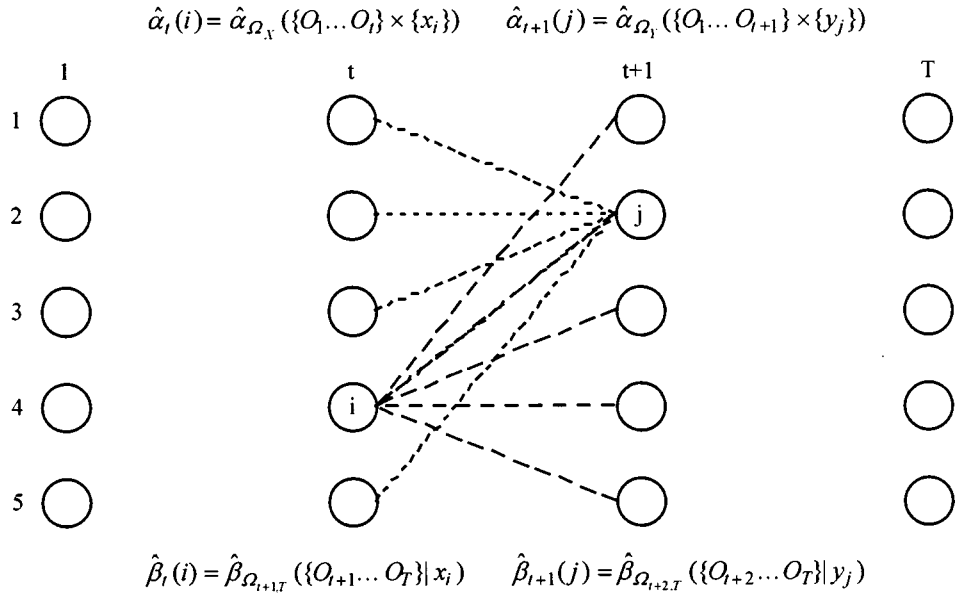


Fig. 6. Fuzzy forward and backward variables computations.

C. GHMM Using the Choquet Integral

As described in Section II, the computation of the Choquet integral is given by

$$e = \sum_{i=1}^{n} h(x_i)\,[g_i^n - g_{i+1}^n]. \tag{67}$$

Define the variable d_i by

$$d_i = [g_i^n - g_{i+1}^n]. \tag{68}$$

Assume that g is a measure satisfying the property that if $g_i^i = 0$ then $d_i = 0$. This condition is satisfied by a wide class of fuzzy measures. We define the variable ρ_i as follows:

$$\rho_i = \begin{cases} d_i/g_i^i, & \text{if } g_i^i \neq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{69}$$

Now, the computation for the Choquet integral is given by

$$e = \sum_{i=1}^{n} h(x_i)\,d_i = \sum_{i=1}^{n} h(x_i)\rho_i g_i^i. \tag{70}$$

This representation of the discrete Choquet integral is very useful for manipulating the GHMM equations and for relating the GHMM to nonstationary HMM's. The forward fuzzy variables are computed as

$$\begin{aligned} \hat{\alpha}_{t+1}(j) &= \hat{\alpha}_{\Omega_Y}^j = \hat{\alpha}_{\Omega_Y}(\{O_1 \cdots O_{t+1}\} \times \{y_j\}) \\ &= \int_X \hat{a}_Y(\{y_j\}|x) \cdot \hat{\alpha}_{\Omega_X}(\{O_1 \cdots O_t\}, \cdot) \wedge \hat{b}_j(O_{t+1}) \\ &= \left[\sum_{i=1}^{N} \hat{a}_{ij} d_t(i,j)\right] \hat{b}_j(O_{t+1}) \end{aligned} \tag{71}$$

where $d_t(i, j)$ represents the difference between the corresponding fuzzy measures and multiplication is used as a fuzzy intersection operator. With this notation, the variable $\rho_t(i, j)$ is given by

$$\rho_t(i, j) = d_i(i, j)/\hat{\alpha}_t(i) \tag{72}$$

then the computation for the fuzzy forward variables reduce to

$$\hat{\alpha}_{t+1}(j) = \left[\sum_{i=1}^{N} \hat{a}_{ij}\rho_t(i, j)\hat{\alpha}_t(i)\right] \hat{b}_j(O_{t+1}) \tag{73}$$

which is similar to the formula for the classical case except for the introduction of the variable $\rho_t(i, j)$ to be computed from the fuzzy measures as described above. Each variable $\rho_t(i, j)$ is a nonlinear function of $\hat{\alpha}_t(k)$ and $\hat{a}_{kj}$, $k = 1, 2, \cdots, N$.

In order to derive reestimation formulas for the GHMM similar to those used for the classical HMM, we redefine the backward fuzzy variable by

$$\hat{\beta}_t(i) = \sum_{j=1}^{N} \hat{a}_{ij}\rho_t(i, j)\hat{\beta}_{t+1}(j)\hat{b}_j(O_{t+1}). \tag{74}$$

It follows that the summation $\sum_{i=1}^{N} \hat{\alpha}_t(i)\hat{\beta}_t(i)$ is independent of t as for the classical case. Let us call the value of this summation the possibility of the observation sequence given the fuzzy model $\hat{\lambda} = (\hat{A}, \hat{B}, \hat{\pi})$ and denote it by $\hat{P}(O|\hat{\lambda})$. It also follows that

$$\begin{aligned} \hat{P}(O|\hat{\lambda}) &= \sum_{i=1}^{N} \hat{\alpha}_t(i)\hat{\beta}_t(i) \\ &= \sum_{\text{all } Q}^{N} \hat{\pi}_{q_1}\hat{b}_{q_1}(O_1) \prod_{t=2}^{T} [\hat{a}_{q_{t-1}, q_t}\rho_t(q_{t-1}, q_t)]\hat{b}_{q_t}(O_t) \\ &= \sum_{\text{all } Q} \hat{P}(O, Q|\hat{\lambda}) \end{aligned} \tag{75}$$

where $\hat{P}(O, Q|\hat{\lambda})$ represents the possibility of the observation $O = \{O_1 O_2 \cdots O_T\}$ and a state sequence $Q = \{q_1 q_2 \cdots q_T\}$ given the fuzzy model $\hat{\lambda} = (\hat{A}, \hat{B}, \hat{\pi})$ that is computed by

$$\hat{P}(O, Q|\hat{\lambda}) = \hat{\pi}_{q_1} \hat{b}_{q_1}(O_1) \prod_{t=2}^{T} [\hat{a}_{q_{t-1}, q_t} \rho_t(q_{t-1}, q_t)] \hat{b}_{q_t}(O_t). \tag{76}$$

This formulation is helpful for developing a constructive and effective training procedure for the fuzzy model by optimizing $\hat{P}(O|\hat{\lambda})$ in a similar way to that of the classical one.

D. Relation Between Generalized HMM and Classical Nonstationary HMM

Recall that a nonstationary HMM is one for which the transition probabilities vary with time. If we define $a'_{ij}(t) = \hat{a}_{ij} \rho_t(i, j)$, then it can be seen that the GHMM can be viewed as a classical nonstationary HMM for which the transition probabilities not only vary with time, but which are dependent upon the observation sequence itself. A major advantage of the GHMM is that this nonstationary behavior is achieved naturally and dynamically as a byproduct of the nonlinear aggregation of information using the fuzzy integral. Moreover, the fuzzy model does not require fixing the lengths of the observation sequences and the availability of more training data in order to learn a large number of transition parameters as for the classical nonstationary model. The additivity constraint of the transition parameters required for all classical HMM's is relaxed for the fuzzy HMM's.

E. Reestimation Formulas for the Choquet Integral GHMM

Let $\hat{P} = \hat{P}(O|\hat{\lambda})$ in a similar manner to that of the classical model; it follows that

$$\begin{aligned} \hat{P} &= \sum_{i=1}^{N} \hat{\alpha}_t(i) \hat{\beta}_t(i) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{N} \hat{\alpha}_t(i) \left(\rho_t(i, j) \hat{a}_{ij}\right) \hat{\beta}_{t+1}(j) \hat{b}_j(O_{t+1}). \end{aligned} \tag{77}$$

Let $\hat{R}$ be the Lagrangian of $\hat{P}$ with respect to the constraint $\sum_{j=1}^{N} \hat{a}_{ij} = 1$

$$\hat{R} = \hat{P} + \sum_{i=1}^{N} \eta_i \left(\sum_{j=1}^{N} \hat{a}_{ij} - 1 \right) = 0 \tag{78}$$

$$\frac{\partial \hat{R}}{\partial \hat{a}_{ij}} = \frac{\partial \hat{P}}{\partial \hat{a}_{ij}} + \eta_i = 0. \tag{79}$$

Multiply by $\hat{a}_{ij}$ and sum over j

$$\sum_{j=1}^{N} \hat{a}_{ij} \frac{\partial \hat{P}}{\partial \hat{a}_{ij}} = -\sum_{j=1}^{N} \hat{a}_{ij} \eta_i = -\eta_i = \frac{\partial P}{\partial \hat{a}_{ij}}. \tag{80}$$

$\hat{P}$ is maximized when

$$\hat{a}_{ij} = \left(\hat{a}_{ij} \frac{\partial \hat{P}}{\partial \hat{a}_{ij}} \right) \Bigg/ \left(\sum_{k=1}^{N} \hat{a}_{ik} \frac{\partial \hat{P}}{\partial \hat{a}_{ik}} \right). \tag{81}$$

The above expression is similar to the expression for the classical case except for the derivatives that can be computed as

$$\begin{aligned} \frac{\partial \hat{P}}{\partial \hat{a}_{ij}} = &\sum_{t=1}^{T-1} \hat{\alpha}_t(i) \hat{\beta}_{t+1}(j) \hat{b}_j(O_{t+1}) \\ &\cdot \left[\rho_t(i, j) + \hat{a}_{ij} \frac{\partial \rho_t(i, j)}{\partial \hat{a}_{ij}} \right]. \end{aligned} \tag{82}$$

If we use a probability measure with the Choquet integral, then $\rho_t(i, j) = 1, \forall t, i, j$, and

$$\frac{\partial \hat{P}}{\partial \hat{a}_{ij}} = \sum_{t=1}^{T-1} \hat{\alpha}_t(i) \hat{\beta}_{t+1}(j) \hat{b}_j(O_{t+1}) \tag{83}$$

which is exactly the expression for the classical case as expected.

The essential question now is, "What is the value of $(\partial \rho_t(i, j)/\partial \hat{a}_{ij})$?" It is very difficult to derive an expression for $\rho_t(i, j)$ as a function of $\hat{a}_{ij}$ because $\rho_t(i, j)$ is computed from the fuzzy forward variables $\hat{\alpha}_t(i)$ that are computed recursively using the Choquet integral. The sorting requirement for computing the fuzzy forward variables makes the derivation of such an expression a complicated task. One approach is to approximate the value by $(\partial \rho_t(i, j)/\partial \hat{a}_{ij}) \cong (\Delta \rho_t(i, j)/\Delta \hat{a}_{ij})$. The problem with this approximation is that we have to store previous values of $\rho_t(i, j)$ in the training procedure, which requires significant memory.

Another approach is to assume a parametric expression that represents $\rho_t(i, j)$ as a monotonic nondecreasing function of $\hat{a}_{ij}$ and $\hat{b}_i(O_t)$ as follows.

Assume that $\rho_t(i, j)$ is the solution of the following differential equations:

$$\hat{a}_{ij} \frac{\partial \rho_t(i, j)}{\partial \hat{a}_{ij}} = \mu_i \rho_t(i, j) \tag{84}$$

$$\hat{b}_i(O_t) \frac{\partial \rho_t(i, j)}{\partial \hat{b}_i(O_t)} = \nu_i \rho_t(i, j) \tag{85}$$

where μ_i and ν_i are positive. These differential equations have solutions. For example the function

$$\rho_t(i, j) = c_{tij} [\hat{a}_{ij}]^{\mu_i} [\hat{b}_i(O_t)]^{\nu_i} \tag{86}$$

(where c_{tij} is positive) is a solution ofthe first equation.

We can also make a similar assumption for deriving a symbol membership estimation formula

$$\frac{\partial \rho_t(i, j)}{\partial \hat{b}_i(O_t)} = \nu_i c_{tij} [\hat{a}_{ij}]^{\mu_i} [\hat{b}_i(O_t)]^{\nu_i - 1} \tag{87}$$

$$\hat{b}_i(O_t) \frac{\partial \rho_t(i, j)}{\partial \hat{b}_i(O_t)} = \nu_i \rho_t(i, j). \tag{88}$$

The second differential equation is needed for the symbol membership reestimation formula. These assumptions allow us to substitute the expression $\mu_i \rho_t(i, j)$ for $\hat{a}_{ij}(\partial \rho_t(i, j)/\partial \hat{a}_{ij})$ inside the updating rule. The term $[1 + \mu_i]$ cancels out of both the numerator and denominator of the equation and it follows

that the reestimation formula (updating rule) for the transition measure becomes

$$\overline{\hat{a}}_{ij} = \frac{\sum_{t=1}^{T-1} \hat{\alpha}_t(i)\rho_t(i,j)\hat{a}_{ij}\hat{\beta}_{t+1}(j)\hat{b}_j(O_{t+1})}{\sum_{k=1}^{N}\sum_{t=1}^{T-1} \hat{\alpha}_t(i)\rho_t(i,k)\hat{a}_{ij}\hat{\beta}_{t+1}(k)\hat{b}_k(O_{t+1})} \tag{89}$$

which is similar to the classical case, except for the presence of ρ. Similarly, we can derive the updating rules for $\hat{\pi}$ and $\hat{B}$ for the discrete case as

$$\overline{\hat{\pi}}_i = \frac{\hat{\alpha}_1(i)\hat{\beta}_1(i)}{\sum_{j=1}^{N} \hat{\alpha}_1(j)\hat{\beta}_1(j)} \tag{90}$$

$$\overline{\hat{b}}_{jk} = \frac{\sum_{t=1,\, O_t=\nu_k}^{T} \hat{\alpha}_t(j)\hat{\beta}_t(j)}{\sum_{t=1}^{T} \hat{\alpha}_t(j)\hat{\beta}_t(j)}. \tag{91}$$

For the continuous case, we model the membership functions $\hat{B} = \{\hat{b}_j(\cdot)\}$ as mixture functions of the form

$$\hat{b}_j(O_t) = \sum_{m=1}^{M} \hat{\omega}_{jm}\hat{f}_{jm}(O_t) \tag{92}$$

where $\hat{f}_{jm}(O_t) = N(O_t, \hat{\mu}_{jm}, \hat{U}_{jm})$ are multivariate Gaussian functions with mean $\hat{\mu}_{jm}$ and covariance matrix $\hat{U}_{jm}$. The reestimation formulae for the coefficients of the mixture functions are (93)–(95), as shown at the bottom of the page.

F. The Fuzzy Viterbi Algorithm

The quantity $\rho_i(i, j)$ provides the basis for defining our modification of the classical Viterbi algorithm. The modification uses the additional information made available by $\rho_t(i, j)$ since it is a nonlinear function of $\hat{\alpha}_t(k)$ and all the transitions $\hat{a}_{kj}$, $k = 1, 2, \cdots, N$. We seek to maximize the function $\hat{P}(O, Q|\hat{\lambda})$ defined by (76). Our modification of the classical Viterbi algorithm is used to perform this maximization. Define a quantity

$$\hat{\delta}_t(i) = \max_{q_1, \cdots, q_{t-1}} \left\{ \hat{\pi}_{q_1}\hat{b}_{q_1}(O_1) \prod_{\tau=2}^{t} \cdot [\hat{a}_{q_{\tau-1},\, q_\tau}\rho_\tau(q_{\tau-1},\, q_\tau)]\hat{b}_{q_\tau}(O_\tau) \right\}. \tag{96}$$

Similarly, $\hat{\delta}_{t+1}(i)$ can be computed inductively using the fuzzy Viterbi algorithmin the following manner.

Initialization for $1 \le i \le N$

$$\hat{\delta}_1(i) = \hat{\pi}_i\hat{b}_i(O_1) \tag{97}$$

$$\hat{\varphi}_1(i) = 0. \tag{98}$$

Recursion for $2 \le t \le T$ and $1 \le j \le N$

$$\hat{\delta}_t(j) = \max_{1\le i\le N} [\hat{\delta}_{t-1}(i)\hat{a}_{ij}\rho_t(i,\, j)]\hat{b}_j(O_t) \tag{99}$$

$$\hat{\varphi}_t(j) = \arg\max_{1\le i\le N} [\hat{\delta}_{t-1}(i)\hat{a}_{ij}\rho_t(i,\, j)]. \tag{100}$$

Termination

$$\hat{P}^* = \max_{1\le i\le N} [\hat{\delta}_T(i)] \tag{101}$$

$$\hat{q}_T^* = \arg\max_{1\le i\le N} [\hat{\delta}_T(i)]. \tag{102}$$

Backtracking for all $1 \le t \le T-1$

$$\hat{q}_t^* = \hat{\varphi}_{t+1}(\hat{q}_{t+1}^*). \tag{103}$$

G. Implementation Issues for the GHMM

As for the classical model, for a sufficiently long observation sequence, the dynamic range of $\hat{\alpha}_t(i)$ computation will exceed

$$\overline{\hat{\omega}}_{jm} = \frac{\sum_{t=1}^{T-1}\sum_{i=1}^{N} \hat{\alpha}_t(i)\hat{a}_{ij}\rho_t(i,\, j)\hat{\omega}_{jm}\hat{f}_{jm}(O_{t+1})\hat{\beta}_{t+1}(j)}{\sum_{m=1}^{M}\sum_{t=1}^{T-1}\sum_{i=1}^{N} \hat{\alpha}_t(i)\hat{a}_{ij}\rho_t(i,\, j)\hat{\omega}_{jm}\hat{f}_{jm}(O_{t+1})\hat{\beta}_{t+1}(j)} \tag{93}$$

$$\overline{\hat{\mu}}_{jm} = \frac{\sum_{t=1}^{T-1}\sum_{i=1}^{N} [\hat{\alpha}_t(i)\hat{a}_{ij}\rho_t(i,\, j)\hat{\omega}_{jm}\hat{f}_{jm}(O_{t+1})\hat{\beta}_{t+1}(j)]O_{t+1}}{\sum_{m=1}^{M}\sum_{t=1}^{T-1}\sum_{i=1}^{N} [\hat{\alpha}_t(i)\hat{a}_{ij}\rho_t(i,\, j)\hat{\omega}_{jm}\hat{f}_{jm}(O_{t+1})\hat{\beta}_{t+1}(j)]} \tag{94}$$

$$\overline{\hat{U}}_{jm} = \frac{\sum_{t=1}^{T-1}\sum_{i=1}^{N} [\hat{\alpha}_t(i)\hat{a}_{ij}\rho_t(i,\, j)\hat{\omega}_{jm}\hat{f}_{jm}(O_{t+1})\hat{\beta}_{t+1}(j)](O_{t+1} - \hat{\mu}_{jm})\,(O_{t+1} - \hat{\mu}_{jm})'}{\sum_{m=1}^{M}\sum_{t=1}^{T-1}\sum_{i=1}^{N} [\hat{\alpha}_t(i)\hat{a}_{ij}\rho_t(i,\, j)\hat{\omega}_{jm}\hat{f}_{jm}(O_{t+1})\hat{\beta}_{t+1}(j)]} \tag{95}$$

the precision range of any computer. This is because we used multiplication as the fuzzy intersection operator and the Choquet integral as the fuzzy integral. Multiplication is used because it is distributive over the summation resulting from using the Choquet integral. We also used the possibility measure as the fuzzy measure. The same scaling procedure for the classical HMM is used for the generalized HMM since we used a possibility measure with the Choquet integral. When using a possibility measure, if we scale the fuzzy forward variables at a certain time the induced forward variables at the next slot will be scaled by the same scaling factor because of the nature of the "max" function. This is a desired property for the scaling technique which is used to multiply $\hat{\alpha}_t(i)$ by a scaling coefficient that is independent of i. A similar scaling is done to the $\hat{\beta}_t(i)$ coefficients by the same scaling factor used for $\hat{\alpha}_t(i)$ since these also tend to approach zero and then at the end of the computation, the scaling coefficients are canceled out.

Another implementation issue, similar to that for the classical case, is related to the modification of the fuzzy reestimation procedure to handle multiple observation sequences. Let the set of the K training observation sequences be $O = \{O^1, O^2, \cdots, O^K\}$ where $O^k = (O_1^k, O_2^k, \cdots, O_{T_k}^k)$ is the kth observation sequence. The goal is to adjust the parameters of the model $\hat{\lambda}$ to maximize

$$\hat{P}(O|\hat{\lambda}) = \prod_{k=1}^{K} \hat{P}(O^k|\hat{\lambda}) = \prod_{k=1}^{K} \hat{P}^k. \tag{104}$$

The modified reestimation formula for the transition fuzzy measures is

$$\overline{\hat{a}}_{ij} = \frac{\sum_{k=1}^{K} \sum_{t=1}^{T_k-1} \hat{\alpha}_t^k(i)\hat{a}_{ij}\rho_t^k(i,j)\hat{b}_j(O_{t+1}^k)\hat{\beta}_t^k(j)}{\sum_{j=1}^{N} \sum_{k=1}^{K} \sum_{t=1}^{T_k-1} \hat{\alpha}_t^k(i)\hat{a}_{ij}\rho_t^k(i,j)\hat{b}_j(O_{t+1}^k)\hat{\beta}_t^k(j)} \tag{105}$$

using the scaled forward and backward variables. Similar results are obtained for the other parameters.

V. Conclusion

We described a generalization of classical HMM's using fuzzy measures and fuzzy integrals resulting in a fuzzy hidden Markov modeling framework. Since the definition of the Markov property is a statement about conditional expectations, our generalization relies heavily on the use of the conditional fuzzy measure. An attractive property of this generalization is the fact that if we used Choquet integral as the general fuzzy integral, multiplication as the fuzzy intersection operator, and a probability measure as the fuzzy measure then we are back to the original probabilistic HMM framework. In this sense, the classical model is one of many models provided by the generalization.

Another property of our generalization is that statistical independence need not be assumed. The expression for computing the fuzzy forward variables inductively does not require decomposition of the joint measure of the previous observation sequence and the current state as is required for the classical case. The fuzzy expression reduces the one required by assuming statistical independence if we use the Choquet integral, probability measure and multiplication as the fuzzy integral, the fuzzy measure and the fuzzy intersection operator respectively.

Another interesting property of our approach for the generalization is the establishment of the relationship between the fuzzy HMM and the classical nonstationary HMM in which the transitional probabilities vary with time. The main advantage of the fuzzy model is that this nonstationary behavior is achieved naturally and dynamically as a byproduct of the nonlinear aggregation of information using the fuzzy integral. Moreover, the fuzzy model does not require fixing the lengths of the observation sequences and the availability of more training data in order to learn a large number of transition parameters as for the classical nonstationary model. The additivity constraint of the transition parameters required for all classical HMM's is not required for the fuzzy HMM's.

Acknowledgment

The authors would like to thank Prof. J. Keller, Prof. X. Zuang, Prof. R. Krishnapuram, and Prof. P. Blackwell of the University of Missouri at Columbia for their valuable discussions and suggestions. The authors would also like to thank the referees for their appreciated detailed review of the manscript.

References

[1] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.

[2] H. Yang and K. Alman, "Two-dimensional shape classification using hidden Markov model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, pp. 1172–1184, Nov. 1991.

[3] M. A. Mohamed and P. D. Gader, "Handwritten word recognition using segmentation-free hidden Markov modeling and segmentation-based dynamic programming techniques," *IEEE Trans. Pattern Anal. Machine Intelligence*, vol. 18, pp. 548–554, May 1996.

[4] Z. Wang and G. Klir, *Fuzzy Measure Theory*, New York: Plenum, 1992.

[5] A. Kundu, "Recognition of handwritten word: First and second order hidden Markov model based approach," *Pattern Recogn.*, vol. 22, no. 3, pp. 457–461, 1988.

[6] B. H. Juang, "Maximum-likelihood estimation for mixer multivariate stochastic observations of Markov chains," *AT&T Tech. J.*, vol. 64, no. 6, July/Aug. 1985.

[7] B. H. Juang and L. R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 1404–1413, Dec. 1985.

[8] ——, "The segmental K-means algorithm for estimating parameters of hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, no. 9, pp. 1639–1641, Sept. 1990.

[9] L. R. Rabiner, J. G. Wilpon, and B. H. Juang, "A segmental K-means training procedure for connected word recognition," *AT&T Tech. J.*, vol. 65, no. 3, pp. 21–31, May/June 1986.

[10] J. Keller, P. Gader, H. Tahani, J. Chiang, and M. Mohamed, "Advances in fuzzy integration for pattern recognition," *Fuzzy Sets Syst.*, vol. 65, pp. 273–283, 1994.

[11] P. D. Gader, M. A. Mohamed, and J. M. Keller, "Dynamic-programming-based handwritten word recognition using the Choquet fuzzy integral as the match function," *J. Electron. Imaging*, vol. 5, no. 1, pp. 15–24, Jan. 1996.

[12] ——, "Fusion of handwritten word classifiers," Pattern Recogn. Lett., to be published.

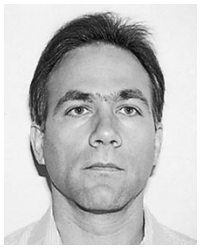
[13] M. Grabisch, "Fuzzy integrals as a generalized class of order filters," in *Proc. Eur Symp Satelite Remote Sensing*, , Rome, Italy, 1994, pp. 128–136.

[14] M. Sugeno, "Fuzzy measures and fuzzy integrals—A survey," in *Fuzzy Automata and Decision Processes*, M. M. Gupta, G. N. Saridis, and B. R. Gaines, Eds, New York: North-Holland, 1977, pp. 89–102.

[15] M. Grabisch and J. Nicolas, "Classification by fuzzy integral: Performance and tests," *Fuzzy Sets Syst.*, vol. 65, pp. 255–273, 1994.
[16] H. Tahani and J. Keller, "Information fusion in computer vision using the fuzzy integral," *IEEE Trans. Syst., Man, Cybern.*, vol. 20, pp. 733–741, May/June 1990.
[17] T. Murofushi and M. Sugeno, "A theory of fuzzy measures: Representations, the Choquet integral, and null sets," *J. Math. Anal. Applicat.*, vol. I59, pp. 532–549, 1991.
[18] M. Grabisch and M. Sugeno, "Multi-attribute classification using fuzzy integral," presented at the 1st Int. Conf. Fuzzy Syst., San Diego, CA, Mar. 1992, pp. 47–55.
[19] M. Grabisch and M. Schmitt, "Mathematical morphology, order filters, and fuzzy logic," in *Proc., 4th Int. Conf. FUZZ-IEEE*. Yokohama, Japan, July 1995, pp. 2103–2108.
[20] M. A. Mohamed and P. D. Gader, "Generalization of hidden Markov models using fuzzy integrals," in *Proc., NAFIPS/IFIS/NASA'94*, San Antonio, TX, Dec. 1994, pp. 3–7.
[21] M. A. Mohamed, "Handwritten word recognition using generalized hidden Markov models," Ph.D. dissertation, Univ. Missouri-Columbia, Columbia, MO, May 1995.
[22] G. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic: Theory and Applications*, New York: Prentice-Hall, 1995.

Magdi A. Mohamed (M'93) received the B.Sc. degree in electrical engineering from the University of Khartoum, Sudan, in September 1983, and the M.S. (computer science) and Ph.D. (electrical and computer engineering) degrees from the University of Missouri-Columbia, in 1991 to 1995, respectively.

From 1985 to 1988, he worked as a Teaching Assistant at the Computer Center, University of Khartoum, Sudan. He also worked as a Computer Engineer for consultation and hardware support at Computer Man Ltd. in Khartoum, Sudan, and as an Electrical and Communication Engineer at Sudan National Broadcasting Corporation in Omdurman, Sudan. From 1991 to 1995 he was a Research and Teaching Assistant in the Department of Electrical and Computer Engineering, University of Missouri-Columbia. He worked as a Visiting Professor in the Computer Engineering and Computer Science Department at the University of Missouri-Columbia from 1995 to 1996. He is currently working as a Research Scientist at Motorola Human Interface Laboratories, Lexicus Division. His research interests include online and offline handwriting recognition, image processing, computer vision, fuzzy set theory, neural networks, pattern recognition, parallel and distributed computing, artificial intelligence, and fractals and chaos theory.



Paul Gader (M'87–SM'99) received the Ph.D. degree for research in image processing from the University of Florida, Gainesville, in 1986.

Since 1986, he has worked as a Senior Research Scientist at Honeywell Systems and Research Center, Minneapolis, MN, as an Assistant Professor of Mathematics at the University of Wisconsin-Oshkosh, and as a Research Engineer and Manager at the Environmental Research Institute of Michigan (ERIM), Ann Arbor, MI. He is currently an Associate Professor of Computer Engineering and Computer Science at the University of Missouri-Columbia. He has been actively involved in handwriting recognition research since 1989. He has performed research in handwritten and machine-printed line, word, and character segmentation, zip code and street number location and recognition, post office box detection and recognition, on handwritten digit and alphabetic character recognition, handwritten word recognition, and on multiple classifier fusion in digit, character, and word recognition. He is also conducting research in land-mine detection, mathematical morphology, medical imaging, automatic target recognition, and neural networks.