A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition

LAWRENCE R. RABINER, FELLOW, IEEE

Although initially introduced and studied in the late 1960s and early 1970s, statistical methods of Markov source or hidden Markov modeling have become increasingly popular in the last several years. There are two strong reasons why this has occurred. First the models are very rich in mathematical structure and hence can form the theoretical basis for use in a wide range of applications. Second the models, when applied properly, work very well in practice for several important applications. In this paper we attempt to carefully and methodically review the theoretical aspects of this type of statistical modeling and show how they have been applied to selected *problems in machine recognition of speech.*

I. INTRODUCTION

Real-world processes generally produce observable outputs which can be characterized as signals. The signals can be discrete in nature (e.g., characters from a finite alphabet, quantized vectors from a codebook, etc.), or continuous in nature (e.g., speech samples, temperature measurements, music, etc.). The signal source can be stationary (i.e., its statistical properties do not vary with time), or nonstationary (i.e., the signal properties vary over time). The signals can be pure (i.e., coming strictly from a single source), or can be corrupted from other signal sources (e.g., noise) or by transmission distortions, reverberation, etc.

A problem of fundamental interest is characterizing such real-world signals in terms of signal models. There are several reasons why one is interested in applying signal models. First of all, a signal model can provide the basis for a theoretical description of a signal processing system which can be used to process the signal so as to provide a desired output. For example if we are interested in enhancing a speech signal corrupted by noise and transmission distortion, we can use the signal model to design a system which will optimally remove the noise and undo the transmission distortion. A second reason why signal models are important is that they are potentially capable of letting us learn a great deal about the signal source (i.e., the real-world process which produced the signal) without having to have the source available. This property is especially important when the cost of getting signals from the actual source is high.

Manuscript received January 15, 1988; revised October 4, 1988. The author is with AT&T Bell Laboratories, Murray **Hill,** NJ 079742070, USA.
IEEE Log Number 8825949. In this case, with a good signal model, we can simulate the source and learn as much as possible via simulations. Finally, the most important reason why signal models are important is that they often work extremely well in practice, and enable us to realize important practical systems—e.g., prediction systems, recognition systems, identification systems, etc., in a very efficient manner.

These are several possible choices for what type of signal model is used for characterizing the properties of a given signal. Broadly one can dichotomize the types of signal models into the class of deterministic models, and the class of statistical models. Deterministic models generally exploit some known specific properties of the signal, e.g., that the signal is a sine wave, or a sum of exponentials, etc. In these cases, specification of the signal model is generally straightforward; all that is required is to determine (estimate) values of the parameters of the signal model (e.g., amplitude, frequency, phase of a sine wave, amplitudes and rates of exponentials, etc.). The second broad class of signal models is the set of statistical models in which one tries to characterize only the statistical properties of the signal. Examples of such statistical models include Gaussian processes, Poisson processes, Markov processes, and hidden Markov processes, among others. The underlying assumption of the statistical model is that the signal can be well characterized as a parametric random process, and that the parameters of the stochastic process can be determined (estimated) in a precise, well-defined manner.

For the applications of interest, namely speech processing, both deterministic and stochastic signal models have had good success. In this paper we will concern ourselves strictly with one type of stochastic signal model, namely the hidden Markov model (HMM). (These models are referred to as Markov sources or probabilistic functions of Markov chains in the communications literature.) We will first review the theory of Markov chains and then extend the ideas to the class of hidden Markov models using several simple examples. We will then focus our attention on the three fundamental problems' for HMM design, namely: the

'The idea of characterizing the theoretical aspects of hidden Markov modeling in terms of solving three fundamental problems is due to Jack Ferguson of IDA (Institute for Defense Analysis) who introduced it in lectures and writing.

0018-9219/89/0200-0257$01.00 © 1989 IEEE

evaluation of the probability (or likelihood) of a sequence of observations given a specific HMM; the determination of a best sequence of model states; and the adjustment of model parameters so as to best account for the observed signal. We will show that once these three fundamental problems are solved, we can apply HMMs to selected problems in speech recognition.

Neither the theory of hidden Markov models nor its applications to speech recognition is new. The basic theory was published in a series of classic papers by Baum and his colleagues [1]-[5] in the late 1960s and early 1970s and was implemented for speech processing applications by Baker [6] at CMU, and by Jelinek and his colleagues at IBM [7]-[13] in the 1970s. However, widespread understanding and application of the theory of HMMs to speech processing has occurred only within the past several years. There are several reasons why this has been the case. First, the basic theory of hidden Markov models was published in mathematical journals which were not generally read by engineers working on problems in speech processing. The second reason was that the original applications of the theory to speech processing did not provide sufficient tutorial material for most readers to understand the theory and to be able to apply it to their own research. As a result, several tutorial papers were written which provided a sufficient level of detail for a number of research labs to begin work using HMMs in individual speech processing applications [14]-[19]. This tutorial is intended to provide an overview of the basic theory of HMMs (as originated by Baum and his colleagues), provide practical details on methods of implementation of the theory, and describe a couple of selected applications of the theory to distinct problems in speech recognition. The paper combines results from a number of original sources and hopefully provides a single source for acquiring the background required to pursue further this fascinating area of research.

The organization of this paper is as follows. In Section II we review the theory of discrete Markov chains and show how the concept of hidden states, where the observation is a probabilistic function of the state, can be used effectively. We illustrate the theory with two simple examples, namely coin-tossing, and the classic balls-in-urns system. In Section III we discuss the three fundamental problems of HMMs, and give several practical techniques for solving these

problems. In Section IV we discuss the various types of HMMs that have been studied including ergodic as well as left-right models. In this section we also discuss the various model features including the form of the observation density function, the state duration density, and the optimization criterion for choosing optimal HMM parameter values. In Section V we discuss the issues that arise in implementing HMMs including the topics of scaling, initial parameter estimates, model size, model form, missing data, and multiple observation sequences. In Section VI we describe an isolated word speech recognizer, implemented with HMM ideas, and show how it performs as compared to alternative implementations. In Section VII we extend the ideas presented in Section VI to the problem of recognizing a string of spoken words based on concatenating individual HMMs of each word in the vocabulary. In Section VIII we briefly outline how the ideas of HMM have been applied to a large vocabulary speech recognizer, and in Sec tion IX we summarize the ideas discussed throughout the paper.

II. DISCRETE MARKOV PROCESSES[2]

Consider a system which may be described at any time as being in one of a set of N distinct states, $.5_1, 5_2, \bullet\bullet\bullet$, SN, as illustrated in Fig. 1 (where $N = 5$ for simplicity). At regularly spaced discrete times, the system undergoes a change of state

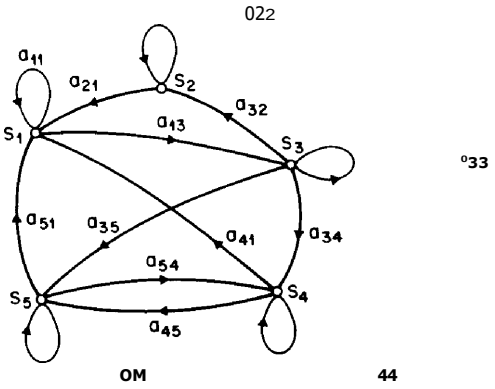


Fig. 1. A Markov chain with 5 states (labeled S_1 to S_5) with selected state transitions.

(possibly back to the same state) according to a set of probabilities associated with the state. We denote the time instants associated with state changes as $t = 1, 2, \bullet\bullet\bullet$, and we denote the actual state at time t as ch. A full probabilistic description of the above system would, in general, require specification of the current state (at time t), as well as all the predecessor states. For the special case of a discrete, first order, Markov chain, this probabilistic description is truncated to just the current and the predecessor state, i.e.,

$$P[qr=Silgr\text{-}1= \mathbf{Si},\ qt\text{-}\ 2 = Skr$$

$$= Ptqt = Siiqt\text{-}i = \qquad (1)$$

Furthermore we only consider those processes in which the right-hand side of (1) is independent

of time, thereby leading to the set of state transition probabilities $a_{,1}$ of the form

$$a_q = P[ch = SiIq\ \__1 = \quad 1 \quad f,1 \quad N \quad (2)$$

with the state transition coefficients having the properties

$$0 \qquad (3a)$$

$$\sum_{N} a_1 = 1 \qquad (3b)$$

since they obey standard stochastic constraints.

The above stochastic process could be called an observable Markov model since the output of the process is the set of states at each instant of time, where each state corresponds to a physical (observable) event. To set ideas, consider a simple 3-state Markov model of the weather. We assume that once a day (e.g., at noon), the weather is

[2] **A** good overview of discrete Markov processes is in (20, ch. 5].

observed as being one of the following:

State 1: rain or (snow)
State 2: cloudy
State 3: sunny.

We postulate that the weather on day *t* is characterized by a single one of the three states above, and that the matrix *A* of state transition probabilities is

$$A = \{a,j = \begin{matrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{matrix}.$$

Given that the weather on day 1 (t = 1) is sunny (state 3), we can ask the question: What is the probability (according to the model) that the weather for the next 7 days will be "sun-sun-rain-rain-sun-cloudy-sun • • • "? Stated more formally, we define the observation sequence o as o = {S3, S3, S3, Si, Si, S3, 52, S3} corresponding to *t* = 1, 2, - • • , 8, and we wish to determine the probability of o, given the model. This probability can be expressed (and evaluated) as

P(OlModel) = P[S_3, S3, S3/ S1/ S1, S3, S2, S_3IModel] = Pisd Fis3ls31 • F[S31s31 • PIS-11.531 • P[SilSi] • Pi5315-11 • P[S21S3] • P[S31.52] = 73 • a33 • a33 • a31 • all • a13 • a32 • a23

= 1 • (0.8)(0.8)(0.1)(0.4)(0.3)(0.1)(0.2) = 1.536 x 10^{-4}

where we use the notation

$$r, = P[q, = Si], \quad 1 Ls. i \quad N \tag{4}$$

to denote the initial state probabilities.

Another interesting question we can ask (and answer using the model) is: Given that the model is in a known state, what is the probability it stays in that state for exactly *ddays?* This probability can be evaluated as the probability of the observation sequence

$$= \{S,, S,, S,, \bullet\bullet, S,, \quad * S_i\},$$
$$1\ 2\ 3 \qquad d\ d+1$$

given the model, which is

$$P(OlModel, = Si) = (aid^{d-1}(1 — au) = p,(d). \tag{5}$$

The quantity *p,(d)* is the (discrete) probability density function of duration *d* in state *i*. This exponential duration density is characteristic of the state duration in a Markov chain. Based on *p,(d),* we can readily calculate the expected number of

observations (duration) in a state, conditioned on starting in that state as

(6a)

$$= > d(add\text{-}lo - au \quad 1) = \quad \_(6b)$$
$$d=1 \qquad\qquad 1 - a_i,$$

Thus the expected number of consecutive days of sunny weather, according to the model, is 1/(0.2) = 5; for cloudy it is 2.5; for rain it is 1.67.

A. Extension to Hidden Markov Models

So far we have considered Markov models in which each state corresponded to an observable (physical) event. This model is too restrictive to be applicable to many problems of interest. In this section we extend the concept of Markov models to include the case where the observation is a probabilistic function of the state—i.e., the resulting model (which is called a hidden Markov model) is a doubly embedded stochastic process with an underlying stochastic 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. To fix ideas, consider the following model

$$= \underset{d=1}{z} dp_i(d)$$

of some simple coin tossing experiments.

Coin Toss Models: Assume the following scenario. You are in a room with a barrier (e.g., a curtain) through which you cannot see what is happening. On the other side of the barrier is another person who is performing a coin (or multiple coin) tossing experiment. The other person will not tell you anything about what he is doing exactly; he will only tell you the result of each coin flip. Thus a sequence of *hidden* coin tossing experiments is performed, with the observation sequence consisting of a series of heads and tails; e.g., a typical observation sequence would be

$$0 = 02\ 03 \cdots \boldsymbol{OT}$$
$$= 3C\ 3C\ 3333C\ 333C \cdots 3C$$

where 3C stands for heads and 3 stands for tails.

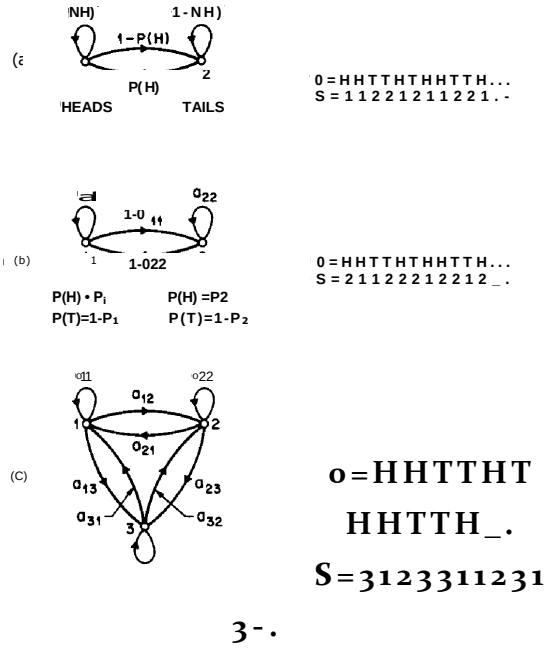
Given the above scenario, the problem of interest is how do we build an HMM to explain (model) the observed sequence of heads and tails. The first problem one faces is deciding what the states in the model correspond to, and then deciding how many states should be in the model. One possible choice would be to assume that only a single biased coin was being tossed. In this case we could model the situation with a 2-state model where each state corresponds to a side of the coin (i.e., heads or tails). This model is depicted in Fig. 2(a).[3] In this case the Markov model is observable, and the only issue for

complete specification of the model would be to decide on the best value for the bias (i.e., the probability of, say, heads). Interestingly, an equivalent HMM to that of Fig. 2(a) would be a degenerate 1-state model, where the state corresponds to the single biased coin, and the unknown parameter is the bias of the coin.

A second form of HMM for explaining the observed sequence of coin toss outcome is given in Fig. 2(b). In this case there are 2 states in the model and each state corresponds to a different, biased, coin being tossed. Each state is characterized by a probability distribution of heads and tails, and transitions between states are characterized by a state transition matrix. The physical mechanism which accounts for how state transitions are selected could itself be a set of independent coin tosses, or some other probabilistic event.

A third form of HMM for explaining the observed sequence of coin toss outcomes is given in Fig. 2(c). This model corresponds to using 3 biased coins, and choosing from among the three, based on some probabilistic event.

'The model of Fig. 2(a) is a memoryless process and thus is a degenerate case of a Markov model.



o_{33}

STATE

1 2 3

NH) P_1 P_2 P_3

PIT) 1-1) 1-P_2 1-P_3

Fig. 2. Three possible Markov models which can account for the results of hidden coin tossing experiments. (a) 1-coin model. (b) 2-coins model. (c) 3-coins model.

Given the choice among the three models shown in Fig. 2 for explaining the observed sequence of heads and tails, a natural question would be which model best matches the actual observations. It should beclear that the simple 1-coin model of Fig. 2(a) has only 1 unknown parameter; the 2-coin model of Fig. 2(b) has 4 unknown parameters; and the 3-coin model of

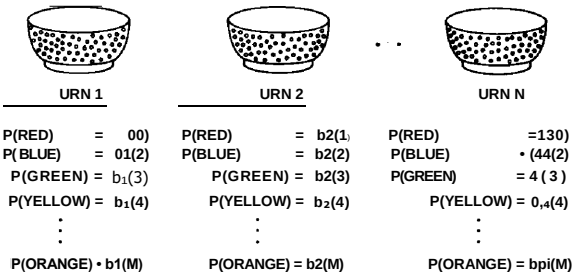


Fig. 3. An N-state urn and ball model which illustrates the general case of a discrete symbol HMM.

Fig. 2(c) has 9 unknown parameters. Thus, with the greater degrees of freedom, the larger HMMs would seem to inherently be more capable of modeling a series of coin tossing experiments than would equivalently smaller models. Although this is theoretically true, we will see later in this paper that practical considerations impose some strong limitations on the size of models that we can consider. Furthermore, it might just be the case that only a single coin is being tossed. Then using the 3-coin model of Fig. 2(c) wou Id be inappropriate, since the actual physical event would not correspond to the model being used—i.e., we would be using an underspecified system.

The Urn and Ball Model': To extend the ideas of the HMM to a somewhat more complicated situation, consider the urn and ball system of Fig. 3. We assume that there are *N* (large) glass urns in a room. Within each urn there are a large number of colored balls. We assume there are M distinct colors

of the balls. The physical process for obtaining observations is as follows. A genie is in the room, and according to some random process, he (or she) chooses an initial urn. From this urn, a ball is chosen at random, and its color is recorded as the observation. The ball is then replaced in the urn from which it was selected. A new urn is then selected

[4]The urn and ball model was introduced by Jack Ferguson, and his colleagues, in lectures on HMM theory. according to the random selection process associated with the current urn, and the ball selection process is repeated. This entire process generates a finite observation sequence of colors, which we would like to model as the observable output of an HMM.

It should be obvious that the simplest HMM that corresponds to the urn and ball process is one in which each state corresponds to a specific urn, and for which a (ball) color probability is defined for each state. The choice of urns is dictated by the state transition matrix of the HMM.

B. Elements of an HMM

The above examples give us a pretty good idea of what an HMM is and how it can be applied to some simple scenarios. We now formally define the elements of an HMM, and explain how the model generates observation sequences.

An HMM is characterized by the following:

1) *N,* the number of states in the model. Although the states are hidden, for many practical applications there is often some physical significance attached to the states or to sets of states of the model. Hence, in the coin tossing experiments, each state corresponded to a distinct biased coin. In the urn and ball model, the states corresponded to the urns. Generally the states are interconnected in such a way that any state can be reached from any other state (e.g., an ergodic model); however, we will see later in

this paper that other possible interconnections of states are often of interest. We denote the individual states as $S = S2, \bullet\bullet\bullet, SO$, and the state at time t as q_t.

2) M, the number of distinct observation symbols per state, i.e., the discrete alphabet size. The observation symbols correspond to the physical output of the system being modeled. For the coin toss experiments the observation symbols were simply heads or tails; for the ball and urn model they were the colors of the balls selected from the urns. We denote the individual symbols as $V = \{v_1, v_2, \bullet\bullet\bullet, v_{A,1}\}$.

3) The state transition probability distribution $A = \{a_{11}\}$ where

$$a, = Pfq_t, = S_t rq_t = S_i],1 \quad _i j \quad N. \quad (7)$$

For the special case where any state can reach any other state in a single step, we have $a_{„} > o$ for all i, j. For other types of HMMs, we would have $a_y = o$ for one or more (i, j) pairs.

4) The observation symbol probability distribution in state *j, B = {b,(k)},* where

$$b_i(k) = P[v_k \text{ at } tlq, = 1 \text{ s } j \quad N \quad k \text{ s M.} \quad (8)$$

5) The initial state distribution 7r

= {7r,} where

$$= P[q1 = Si], \quad 1 \qquad N. \qquad (9)$$

Given appropriate values of *N, M, A, B,* and 7r, the HMM can be used as a generator to give an observation sequence

$$o = \quad o_2 \cdot \cdot O_T \qquad (10)$$

(where each observation o, is one of the symbols from *V,* and *7* is the number of observations in the sequence) as follows:

1) Choose an initial state $q_1 = S,$ according to the initial state distribution W.
2) Set $t = 1$.
3) Choose o, = v_k according to the symbol probability distribution in state *S,,* i.e., *b,(k).*
4) Transit to a new state $q_{t+i} = S_i$ according to the state transition probability distribution for state S_i, i.e.,
5) Set $t = t + 1$; return to step 3)if $t <$ T; otherwise terminate the procedure.

The above procedure can be used as both a generator of observations, and as a model for how a given observation sequence was generated by an appropriate HMM.

It can be seen from the above discussion that a complete specification of an HMM requires specification of two model parameters *(N* and M), specification of observation symbols, and the specification of the three probability measures *A, B,* and Ir. For convenience, we use the compact notation

$$X = (A, B, \mathbf{r})$$

to indicate the complete

parameter set of the model. C. *The*

Three Basic Problems for HMMs[5]

Given the form of HMM of the previous section, there are three basic problems of interest that must be solved for the model to be useful in real-world applications. These problems are the following:

Problem 1: Given the observation sequence o = o_1 o2 . • • *OD* and a model X = *(A, B, 7r),* how do we efficiently compute P(OIX), the probability of the observation sequence, given the model?

Problem 2: Given the observation sequence o = o_1 o_2

• • • *OT,* and the model X, how do we choose a corresponding state sequence Q = q_1 q_2 • • • q_T which is optimal in some meaningful sense (i.e., best "explains" the observations)?

Problem 3: How do we adjust the model parameters X = *(A, B, 21)* to maximize P(01X)?

'The material in this section and in Section III is based on the ideas presented by Jack Ferguson of IDA in lectures at Bell Laboratories.

Problem 1 is the evaluation problem, namely given a model and a sequence of observations, how do we compute the probability that the observed sequence was produced by the model. We can also view the problem as one of scoring how well a given model matches a given observation sequence. The latter viewpoint is extremely useful. For example, if we consider the case in which we are trying to choose among several competing models, the solution to Problem 1 allows us to choose the model which best matches the observations.

Problem 2 is the one in which we attempt to uncover the hidden part of the model, i.e., to find the "correct" state sequence. It should be clear that for all but the case of degenerate models, there is no "correct" state sequence to be found. Hence for practical situations, we usually use an optimality criterion to solve this problem as best as possible. Unfortunately, as we will see, there are several reasonable optimality criteria that can be imposed, and hence the choice of criterion is a strong function of the intended use for the uncovered state sequence. Typical uses might be to learn about the structure of the model, to find optimal state sequences for continuous speech recognition, or to get average statistics of individual states, etc.

Problem 3 is the one in which we attempt to optimize the model parameters so as to best describe how a given observation sequence comes about. The observation sequence used to adjust the model parameters is called a training sequence since it is used to "train" the HMM. The training problem is the crucial one for most applications of HMMs, since it allows us to optimally adapt model parameters to observed training data—i.e., to create best models for real phenomena.

To fix ideas, consider the following simple isolated word speech recognizer. For each word of a W word vocabulary, we want to design a separate N-state HMM. We represent

the speech signal of a given word as a time sequence of coded spectral vectors. We assume that the coding is done using a spectral codebook with M unique spectral vectors; hence each observation is the index of the spectral vector closest (in some spectral sense) to the original speech signal. Thus, for each vocabulary word, we have a training sequence consisting of a number of repetitions of sequences of codebook indices of the word (by one or more talkers). The first task is to build individual word models. This task is done by using the solution to Problem 3 to optimally estimate model parameters for each word model. To develop an understanding of the physical meaning of the model states, we use the solution to Problem 2 to segment each of the word training sequences into states, and then study the properties of the spectral vectors that lead to the observations occurring in each state. The goal here would be to make refinements on the model (e.g., more states, different codebook size, etc.) so as to improve its capability of modeling the spoken word sequences. Finally, once the set of W HMMs has been designed and optimized and thoroughly studied, recognition of an unknown word is performed using the solution to Problem 1 to score each word model based upon the given test observation sequence, and select the word whose model score is highest *(Le.,* the highest likelihood).

In the next section we present formal mathematical solutions to each of the three fundamental problems for HMMs.

We shall see that the three problems are linked together tightly under our probabilistic framework.

Ill. SOLUTIONS TO THE THREE BASIC PROBLEMS OF HMMs

A. Solution to Problem

We wish to calculate the probability of the observation sequence, o = o_1 o2 • • OT, given the model X, i.e., P(OIX). The most straightforward way of doing this is through enumerating every possible state sequence of length T(the number of observations). Consider one such fixed state sequence

$$Q = \text{ch } q2 \bullet \bullet \bullet qT \qquad (12)$$

where q_1 is the initial state. The probability of the observation sequence o for the state sequence of (12) is

$$P(o1\ Q, X) = II\ P(o_t lq_{\prime\prime}\ X) \qquad (13a)$$
$$i=7$$

where we have assumed statistical independence of observations. Thus we get

$$P(o1\ Q, X) = b,(o_1)\ b_{\prime\prime}(o_2) \bullet \bullet \bullet b,(o_7\text{-}). \qquad (13b)$$

The probability of such a state sequence Q can be written as

$$P(QIX) = {}^{7}\text{ciiachq2a92q3} \bullet \bullet \text{aq,}$$
$$\text{iciT} \bullet \qquad (14)$$

The joint probability of o and Q, i.e., the probability that o and Q occur simultaneously, is simply the product of the above two terms, i.e.,

$$P(o,\ QIX) = P(o1Q,\quad P(Q,\ X). \qquad (15)$$

The probability of o (given the model)is obtained by summing this joint probability over all possible state sequences *q* giving

$$P(OIX) = P(o1Q,\ P(QIX)\text{al } Q \qquad (16$$

$$= 71_{\text{-}9_{\prime\prime}}\text{b},(o_1)\ a_{gic2}b,_{\prime\prime}(o_2) \qquad (17$$
$$\text{acI}_{\text{T\_}}\ {}_{i}ch\text{-}bc_r(QT) \bullet$$

The interpretation of the computation in the above equation is the following. Initially (at time t = 1) we are in state cF, with probability 7r,, and generate the symbol o_1 (in this state) with probability b„(o_1). The clock changes from time t to $\boldsymbol{t + 1}$ ($\boldsymbol{t}$ = 2) and we make a transition to state q_2 from state q_1 with probability $a_{g,g2}$, and generate symbol o_2 with probability b,(o_2). This process continues in this manner until we make the list transition (at time *T)* from state qr-7 to state *(T_r* with probability a_{gr}, and generate symbol Or with probability b,(o_7-).

A little thought should convince the reader that the calculation of P(o1 X), according to its direct definition (17) involves on the order of 2T • N^T calculations, since at every t = 1, 2, • • • , T, there are N possible states which can be reached (i.e., there are N^T possible state sequences), and for

each such state sequence about $2T$ calculations are required for each term in the sum of (17). (To be precise, we need $(2T - 1)N^T$ multiplications, and $N^T - 1$ additions.) This calculation is computationally unfeasible, even for small values of N and T; e.g., for $N = 5$ (states), $T = 100$ (observations), there are on the order of $2 \cdot 100 \cdot 5^{100} \approx 10^{72}$ computations! Clearly a more efficient procedure is required to solve Problem 1. Fortunately such a procedure exists and is called the forward-backward procedure.

The Forward-Backward Procedure [2], [3][6]: Consider the forward variable $a_t(i)$ defined as

$$at(i) = P(o_1\ o_2 \cdots O_t,\ q_t = S_i | X) \quad (18)$$

i.e., the probability of the partial observation sequence, $o_1\ o_2 \cdots O_t$, (until time t) and state S_i at time t, given the model X. We can solve for $a_t(i)$ inductively, as follows:

1) Initialization:

$$a_1(i) = r_i b_i(o_1), \quad 1 < i < N. \quad (19)$$

2) Induction:

$$a_{t+1}(j) = \left[\sum_{i=1}^{N} a_t(i) a_{ij} \right] b_j(O_{t+1}), \quad 1 < t < T - 1 \quad (20)$$

Step 1) initializes the forward probabilities as the joint probability of state S_i and initial observation o_1. The induction step, which is the heart of the forward calculation, is illustrated in Fig. 4(a). This figure shows how state S_i can be

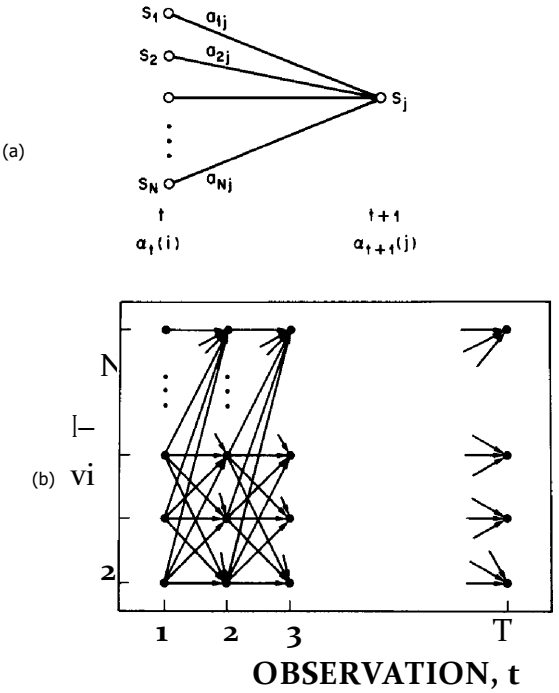


Fig. 4. (a) Illustration of the sequence of operations required for the computation of the forward variable $oi_{r+I(/)}$. (b) Implementation of the computation of $ot_t(i)$ in terms of a lattice of observations t, and states i.

'Strictly speaking, we only need the forward part of the forward-backward procedure to solve Problem 1. We will introduce the backward part of the procedure in this section since it will be used to help solve Problem 3.

reached at time $t + 1$ from the N possible states, S_i, 1 i N, at time t. Since $a_t(i)$ is the probability of the joint event that o_1 o_2 • • O_r are observed, and the state at time t is S_i, the product «,(i)a,, is then the probability of the joint event that o_1 02 • • • Q are observed, and state S_i is reached at time $t + 1$ via state S, at time t. Summing this product over all the N possible states S_i, 1 s i_sNat time t results in the probability of S, at time $t + 1$ with all the accompanying previous partial observations. Once this is done and S_i is known, it is easy to see that «,, j) is obtained by accounting for observation o_{t+1} in state), i.e., by multiplying the summed quantity by the probability $b,(o,,_i)$. The computation of (20) is performed for all states j, 1 5 j Ls. N, for a given t; the computation is then iterated for $t = 1, 2,$ • • • , T - 1. Finally, step 3) gives the desired calculation of P(OIX) as the sum of the terminal forward variables $a_T(i)$. This is the case since, by definition,

$$cir(i) = P(o1\ o2 \bullet\bullet\bullet OD\ qr = S,IX) \quad (22)$$

and hence P(o1X) is just the sum of the «$_T$M's.

If we examine the computation involved in the calculation of $a_t(j)$, **1** t T, 1 5_ j N, we see that it requires on the order of N^2T calculations, rather than $2TN^T$ as required by the direct calculation. (Again, to be precise, we need $N(N + 1)(T - 1) + N$ multiplications and $N(N - 1)(T - 1)$ additions.) For $N = 5$, T = 100, we need about 3000 computations for the forward method, versus 10^{72} computations for the direct calculation, a savings of about 69 orders of magnitude.

The forward probability calculation is, in effect, based upon the lattice (or trellis) structure shown in Fig. 4(b). The key is that since there are only N states (nodes at each time slot in the lattice), all the possible state sequences will re-merge into these N nodes, no matter how long the observation sequence. At time $t = 1$ (the first time slot in the lattice), we need to calculate values of $ot_i(i)$, 1 i N. At times $t = 2, 3,$ • • , T, we only need to calculate values of $a_t(j)$, 1 j Ls N, where each calculation involves only N previous values of $a_{t\_\ 1}(i)$ because each of the Ngrid points is reached from the same N grid points at the previous time slot.

In a similar manner/we can consider a backward variable o,(i) defined as

$$o,(i) = P(o,_{+1}\ o_{1+2} \bullet\bullet\bullet o^{-}rigt = Si, X) \quad (23)$$

i.e., the probability of the partial observation sequence from $t + 1$ to the end, given state S_i at time t and the model X. Again we can solve for o,(i) inductively, as follows:

1) Initialization:

$$o7(i) = 1, \quad 1 < i < N. \qquad (24)$$

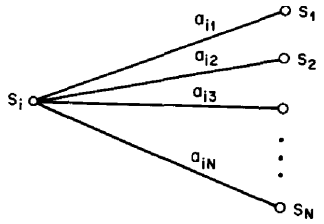
2) Induction:

$$Ot(\blacktriangleright) = aiibi(Ot+1)\ Ot+1(i)1$$

$$t = T - 1, T - 2, \bullet\bullet, 1, 1 \le i \le N. \qquad (25)$$

The initialization step 1) *arbitrarily* defines (3,-(i) to be 1 for all *i.* Step 2), which is illustrated in Fig. 5, shows that in order to have been in state S_i at time *t,* and to account for the

'Again we remind the reader that the backward procedure will be used in the solution to Problem 3, and is not required for the solution of Problem 1.



t +

$_t ti)$ Qt +On

Fig. 5. Illustration of the sequence of operations required for the computation of the backward variable d3,(i).

observation sequence from time $t + 1$ on, you have to consider all possible states S_i at time $t + 1$, accounting for the transition from $S,$ *to* S_i (the $a_{,1}$ term), as well as the observation $o_{,+1}$ in state j (the $b_i(o_{1+1})$ term), and then account for the remaining partial observation sequence from state j (the $o_{,+1}(j)$ term). We will see later how the backward, as well as the forward calculations are used extensively to help solve fundamental Problems 2 and 3 of HMMs.

Again, the computation of $o,(i)$, $1 \le t \le T, 1 \le i \le N,$ requires on the order of N^2T calculations, and can be computed in a lattice structure similar to that of Fig. 4(b).

B. Solution to Problem 2

Unlike Problem 1 for which an exact solution can be given, there are several possible ways of solving Problem 2, namely 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. For example, one possible optimality criterion is to choose the states $q,$ which are *individually* most likely. This optimality criterion maximizes the expected number of correct individual states. To implement this solution to Problem 2, we define the variable

$$'Yr(i) = P(cli = 5, 1^o, X) \qquad (26)$$

i.e., the probability of being in state 5, at time $t,$ given the observation sequence $o,$ and the model X. Equation (26) can be expressed simply in terms of the forward-backward variables, i.e.,

$$7t(i) - \frac{at(i)\,OM)at(i)\,OM)}{P(01X) \quad {}^{N} a_t(i)\, l^{3}\,(i)} \qquad (27)$$

$$i = 1$$

since «,*(i)* accounts for the partial observation sequence $o_1\ o_2 \cdots o$, and state *S,* at *t,* while $,3_t(i)$ accounts for the
remainder of the observation sequence $^{o}t+2 \cdots {}^{\circ}Tr$,
given state *S,* at *t.* The normalization factor P(OIX) =
at(i), $o_1(i)$ makes $-y_t(i)$ a probability measure so that

$$\ell = -y_t(i) = 1. \qquad (28)$$

Using -*MO,* we can solve for the individually most likely state *q,* at time *t,* as

$$q, = \text{argmax}\ Fyt(i)l,1 \qquad t\ \mathbf{5}\_\ T. \qquad (29)$$

Although (29) maximizes the expected number of correct states (by choosing the most likely state for each *t),* there could be some problems with the resulting state sequence. For example, when the HMM has state transitions which have zero probabilty = o for some *i* and *j),* the "optimal" state sequence may, in fact, not even be a valid state sequence. This is due to the fact that the solution of (29) simply determines the most likely state at every instant, without regard to the probability of occurrence of *sequences* of states.

One possible solution to the above problem is to modify the optimality criterion. For example, one could solve for the state sequence that maximizes the expected number of correct pairs of states *(qt, gm),* or triples of states *(qt, qt.", qt+2),* etc. Although these criteria might be reasonable for some applications, the most widely used criterion is to find the *single* best state sequence (path), i.e., to maximize P(Q10, X) which is equivalent to maximizing *P(Q,* 01X). A formal technique for finding this single best state sequence exists, based on dynamic programming methods, and is called the Viterbi algorithm.

Viterbi Algorithm 121], 122]: To find the single best state sequence, Q = {q_1 q2 • • • q_7-}, for the given observation sequence o = {o_1 o_2 • • • O_r}, we need to define the quantity

$$6,(i) = \max P[q, q2 \bullet\bullet\bullet qt = it\ o1\ o2 - \bullet\bullet\bullet\ ^{o}tlxl \quad (30)$$

i.e., $S_t(i)$ is the best score (highest probability) along a single path, at time *t,* which accounts for the first *t* observations and ends in state *S.* By induction we have

$$^{6}t41(j) = [\max o_r(i)a,,l \bullet \mathrm{bi}(O_t{+}1). \quad (31)$$

To actually retrieve the state sequence, we need to keep track of the argument which maximized (31), for each *t* and *j.* We do this via the array Iii_t(j). The complete procedure for finding the best state sequence can now be stated as follows:

1) Initialization: It should be noted that the Viterbi algorithm is similar (except for the backtracking step) in implementation to the forward calculation of (19)-(21). The major difference is the maximization in (33a) over previous states which is used in place of the summing procedure in (20). It also should be clear that a lattice (or trellis) structure efficiently implements the computation of the Viterbi procedure.

C. Solution to Problem 3 [1]-15]

The third, and by far the most difficult, problem of HMMs is to

determine a method to adjust the model parameters *(A, B,* -ir) to maximize the probability of the observation sequence given the model. There is no known way to analytically solve for the model which maximizes the probability of the observation sequence. In fact, given any finite observation sequence as training data, there is no optimal way of estimating the model parameters. We can, however, choose X = *(A, B, 7r)* such that P(o1X) is locally maximized using an iterative procedure such as the Baum-Welch method (or equivalently the EM (expectation-modification) method [23]), or using gradient techniques [14]. In this section we discuss one iterative procedure, based primarily on the classic work of Baum and his colleagues, for choosing model parameters.

In order to describe the procedure for reestimation (iterative update and improvement) of HMM parameters, we first define $E_f(i, j)$, the probability of being in state S, at time t, and state S_i at time $t + 1$, given the model and the observation sequence, i.e.

$$tt(i,D = P(clt = Si, qt\text{-}F1 = Si1o, X). \quad (36)$$

The sequence of events leading to the conditions required by (36) is illustrated in Fig. 6. It should be clear, from the

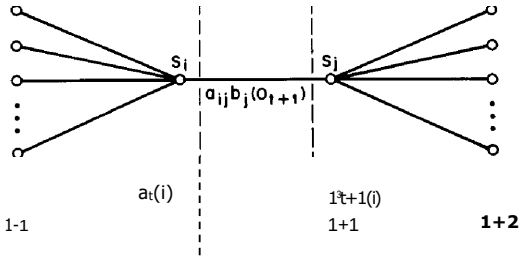


Fig. 6. Illustration of the sequence of operations required for the computation of the joint event that the system is in state S_i at time t and state S_j at time $t + 1$.

definitions of the forward and backward variables, that we can write $E_t(i, j)$ in the form

$$r(i, j) = \frac{ce_r(i)\, aiibi(O_{r+1})\, \underline{Or+ i(j)}}{P(o1X)}$$

$$a_t(i)\ a_{ii}b_i(O_{t+1})\ \underline{Or + l(j)}\ N\ N$$

$$a_t(i)\ aiihi(Or+1)$$

$$j=1$$

where the numerator term is just $P(qt = ch = S_i, OIX)$ and the division by P(o1X) gives the desired probability measure.

$$b_1(i) = 7r_ib_i(0_1), \quad 1 \le i \le N \quad (32a)$$

$$11/1(i) = 0. \quad (32b)$$

2) Recursion:

$$= \max_{1 \le i \le N} [k\_1(i)aijIbi(Ot), \quad 2 \le j \le N \quad (33a)$$

$$= \text{argmax} \quad 2 \le t \le T \quad 1 \le j \le N. \quad (33b)$$

3) Termination: (34a)

$$P^* = \max_{1 \le i \le N} [S_T(i)]$$

$$q; = \text{argmax}_{1 \le i \le N} [4(/)]. \quad (34\ b)$$

4) Path (state sequence) backtracking: $q't^* = 0{,}1(cg4\text{-}1)$, $t=$ T — 1, T — 2, • 1.(35)

(37)

We have previously defined $-y_t(i)$ as the probability of being in state S_i at time t, given the observation sequence and the model; hence we can relate $'y,(i)$ to t_t(i, *1)* by summing over *j,* giving

$$Tt(i) = \sum_{=1} Et(i, \qquad (38)$$

If we sum $-y_t(i)$ over the time index t, we get a quantity which can be interpreted as the expected (over time) number of times that state S_i is visited, or equivalently, the expected number of transitions made from state S_i (if we exclude the time slot t T from the summation). Similarly, summation of $E_t(i, j)$ over t (from $t = 1$ *to* $t = T$ — 1) can be interpreted as the expected number of transitions from state S_i to state S_i. That is

$$\sum_{t=1}^{T-1} -y,(i) = \text{expected number of transitions from } S_i \qquad (39a)$$

$$\sum_{t=-1}^{T-1} E_t(i, j) = \text{expected number of transitions from } S_i \text{ to } S_,. \qquad (39b)$$

Using the above formulas (and the concept of counting event occurrences) we can give a method for reestimation of the parameters of an HMM. A set of reasonable reestimation formulas for 7, A, and B are lihood estimate of the HMM. It should be pointed out that the forward-backward algorithm leads to local maxima only, and that in most problems of interest, the optimization surface is very complex and has many local maxima.

The reestimation formulas of (40a)-(40c) can be derived directly by maximizing (using standard constrained optimization techniques) Baum's auxiliary function

$$Q(X, T\blacksquare) = \text{Q } P(Q10, X) \log [P(0, Qua)] \qquad (41)$$

over X. It has been proven by Baum and his colleagues [6], [3] that maximization of Q(X, X) leads to increased likelihood, i.e.

$$\max [Q(X, X)] \quad P(OIX) \quad P(01 X). \qquad (42)$$

Eventually the likelihood function converges to a critical point.

Notes on the Reestimation Procedure: The reestimation formulas can readily be interpreted as an implementation of the EM algorithm of statistics [23] in which the E (expectation) step is the calculation of the auxiliary function Q(X, X), and the M (modification) step is the maximization over X. Thus the Baum-Welch reestimation equations are essentially identical to the EM steps for this particular problem.

An important aspect of the reestimation procedure is that the

stochastic constraints of the HMM parameters, namely

$$\sum_{i=1}^{N} \pi_i = 1 \tag{43a}$$

$$a_{ij} = \frac{\text{expected number of transitions from state } S_i \text{ to state } S_j}{\text{expected number of transitions from state } S_i} \tag{40b}$$

$$\bar{\pi}_i = \text{expected frequency (number of times) in state } S_i \text{ at time } (t = 1) \tag{40a}$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{40c}$$

$$\bar{b}_j(k) = \frac{\text{expected number of times in state } j \text{ and observing symbol } v_k}{\text{expected number of times in state } j} = \frac{\sum_{t=1}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \quad \text{s.t. } O_t = v_k$$

If we define the current model as $\lambda = (A, B, \pi)$, and use that to compute the right-hand sides of (40a)-(40c), and we define the reestimated model as $\bar{\lambda} = (\bar{A}, \bar{B}, \bar{\pi})$, as determined from the left-hand sides of (40a)-(40c), then it has been proven by Baum and his colleagues [6], [3] that either 1) the initial model λ defines a critical point of the likelihood function, in which case $\bar{\lambda} = \lambda$; or 2) model $\bar{\lambda}$ is more likely than model λ in the sense that $P(O|\bar{\lambda}) > P(O|\lambda)$, i.e., we have found a new model $\bar{\lambda}$ from which the observation sequence is more likely to have been produced.

Based on the above procedure, if we iteratively use $\bar{\lambda}$ in place of λ and repeat the reestimation calculation, we then can improve the probability of O being observed from the model until some limiting point is reached.

$$\sum_{j=1}^{N} a_{ij} = 1, \quad 1 \le i \le N \tag{43b}$$

The final result of this reestimation procedure is called a maximum like-

$$\sum_{k=1}^{M} b_i(k) = 1, \tag{43c}$$

are automatically satisfied at each iteration. By looking at the parameter

estimation problem as a constrained optimization of P(o1 X) (subject to the constraints of (43)), the techniques of Lagrange multipliers can be used to find the values of 7„ a_{id}, and $b_i(k)$*which* maximize *P* (we use the notation *P* = *P(o1X)* as short-hand in this section). Based on setting up a standard Lagrange optimization using Lagrange multipliers, it can readily be shown that *Pis* maximized when

the following conditions are met:

$$\frac{\partial P}{\partial \bar{a}_{ij}} = 0 \tag{44a}$$

$$\frac{\partial P}{\partial \bar{b}_j(k)} = 0 \tag{44b}$$

$$\frac{\partial P}{\partial \bar{\pi}_i} = 0 \tag{44c}$$

where

$$\bar{a}_{ij} = \frac{a_{ij}}{\sum_{k=1}^{N} a_{ik}}$$

$$\bar{b}_j(k) = \frac{b_j(k)}{\sum_{l=1}^{M} b_j(l)}$$

$$\bar{\pi}_i = \frac{\pi_i}{\sum_{k=1}^{N} \pi_k}$$

By appropriate manipulation of (44), the right-hand sides of each equation can be readily converted to be *identical* to the right-hand sides of each part of (40a)-(40c), thereby showing that the reestimation formulas are indeed exactly correct at critical points of P. In fact the form of (44) is essentially that of a reestimation formula in which the left-hand side is the reestimate and the right-hand side is computed using the current values of the variables.

Finally, we note that since the entire problem can be set up as an optimization problem, standard gradient techniques can be used to solve for "optimal" values of the model parameters (14]. Such procedures have been tried and have been shown to yield solutions comparable to those of the standard reestimation procedures.

IV. TYPES OF HMMs

Until now, we have only considered the special case of ergodic or fully connected HMMs in which every state of the model could be reached (in a single step) from every other state of the model. (Strictly speaking, an ergodic model has the property that every state can be reached from every other state in a finite number of steps.) As shown in Fig. 7(a), for an $N = 4$ state model, this type of model has the property that every a_{ij} coefficient is positive. Hence for the example of Fig. 7a we have

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

For some applications, in particular those to be discussed later in this paper, other types of HMMs have been found to account for observed properties of the signal being modeled better than the standard ergodic model. One such model is shown in Fig. 7(b). This model is called a left-right model or a Bakis model [11], [10] because the underlying state sequence associated with the model has the property that as time increases the state index increases (or stays the same), i.e., the states proceed from left to right. Clearly the left-right type of HMM has the desirable property that it can readily model signals whose properties change over time-

e.g., speech. The fundamental property of all left-right

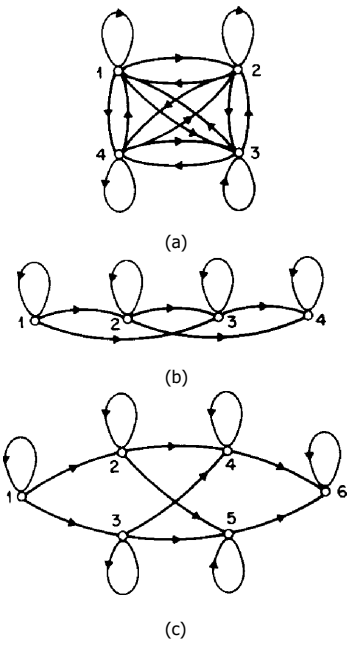


Fig. 7. illustration of 3 distinct types of HMMs. (a) A 4-state ergodic model. (b) A 4-state left-right model. (c) A 6-state parallel path left-right model.

HMMs is that the state transition coefficients have the property

$$= 0, \quad 1 < i \qquad (45)$$

i.e., no transitions are allowed to states whose indices are lower than the current state. Furthermore, the initial state probabilities have the property

$$\{0, i \quad 1$$

$$1, \ i = 1 \qquad (46)$$

since the state sequence must begin in state 1 (and end in state *N).* Often, with left-right models, additional constraints are placed on the state transition coefficients to make sure that large changes in state indices do not occur; hence a constraint of the form

$$a, = 0, \quad j > i + A \qquad (47)$$

is often used. In particular, for the example of Fig. 7(b), the value of is 2, i.e., no jumps of more than 2 states are allowed. The form of the state transition matrix for the example of Fig. 7(b) is thus

$$\Delta - \begin{vmatrix} al & a1 & a1 & 0 \\ 0 & a_{22} & a2 & a_{24} \\ 0 & 0 & a3 & a3 \\ 0 & 0 & 0 & a_{44} \end{vmatrix}$$

It should be clear that, for the last state in a left-right model, that the state transition coefficients are specified as

$$aNN = 1 \qquad (48a)$$

$$am, — 0, \qquad i < N. \qquad (48b)$$

Although we have dichotomized HMMs into ergodic and left-right models, there are many possible variations and combinations possible. By way of example, Fig. 7(c) shows a cross-coupled connection of two parallel left-right HMMs. Strictly speaking, this model is a left-right model (it obeys all the a_q constraints); however, it can be seen that it has certain flexibility not present in a strict left-right model (i.e., one without parallel paths).

It should be clear that the imposition of the constraints of the left-right model, or those of the constrained jump model, essentially have no effect on the reestimation procedure. This is the case because any HMM parameter set to zero initially, will remain at zero throughout the reestimation procedure (see (44)).

A. Continuous Observation Densities in HMMs [24]-[26]

All of our discussion, to this point, has considered only the case when the observations were characterized as discrete symbols chosen from a finite alphabet, and therefore we could use a discrete probability density within each state of this model. The problem with this approach, at least for some applications, is that the observations are continuous signals (or vectors). Although it is possible to quantize such continuous signals via codebooks, etc., there might be serious degradation associated with such quantization. Hence it would be advantageous to be able to use HMMs with continuous observation densities.

In order to use a continuous observation density, some restrictions have to be placed on the form of the model probability density function (pdf) to insure that the parameters of the pdf can be reestimated in a consistent way. The most general representation of the pdf, for which a reestimation procedure has been formulated [24]-[26], is a finite mixture of the form

$$b_i(o) = \sum_{m=1} c_{ij} A[o, p_{in}), 11,,,1, \quad 1 \quad j^{c} N \quad (49)$$

where o is the vector being modeled, $c_{i,,}$ is the mixture coefficient for the mth mixture in state j and DT is any log-concave or elliptically symmetric density [24] (e.g., Gaussian), with mean vector R_{im} and covariance matrix $u,,,$ for the mth mixture component in state j. Usually a Gaussian density is used for N. The mixture gains c_{jm} satisfy the stochastic constraint

$$\text{E } c_{/a)} = 1, \quad 1\,5\_\, j\,5\, N \qquad (50a)$$

$$_{f}E \qquad k)\ O_t \qquad (53$$

$$Iajk —^{1} T \qquad (54$$

$$t=1 \qquad k)$$

$$7; \quad \underline{t=} 7t(j, k)((it — ft,k)(o, — \text{Fi}_{j}\text{k})'$$

g.jjk

where prime denotes vector transpose and where $y_i.(j, k)$ is the probability of being in state j at time t with the kth mixture component accounting for o„ i.e.,

$$\frac{(MD\,k^{at}(j)}{H} \quad \frac{c_{ik}a(00\ pm,\ Um)}{\sum_{m=1}^{M} Cir,^{01}(^{O}tr\ \mu;m\cdot\ U\blacksquare m)\ m}$$

$$at^{(i)}\ at(i)$$

(The term ^$Mb\ k$) generalizes to $y_t(j)$ of (26) in the case of a simple mixture, or a discrete density.) The reestimation formula for $a,_i is$ identical to the one used for discrete observation densities (i.e., (40b)). The interpretation of (52)-(54) is fairly straightforward. The reestimation formula for c_m is the ratio between the expected number of times the system is in state j using the kth mixture component, and the expected number of times the system is in state j. Similarly, the reestimation formula for the mean vector P_{ik} weights each numerator term of (52) by the observation, thereby giving the expected value of the portion of the observation vector accounted for by the kth mixture component. A similar interpretation can be given for the reestimation term for the covariance matrix *Up,*

$$\sum_{t=1}^{T} {}'yt(j,\ k)$$

c.- o, $1 \quad j \quad N, 1\ 5\_m\ M$ (50b)

so that the pdf is properly normalized, i.e.,

$$b_i(x)\ \mathrm{dx} = 1,\ 1 \quad j\ s\ N. \qquad (51)$$

The pdf of (49) can be used to approximate, arbitrarily closely, any finite, continuous density (52)

B.

Autoregressive HMMS [27], [28]

Although the general formulation of continuous density HMMs is applicable to a wide range of problems, there is one other very interesting class of HMMs that is particularly applicable to speech processing. This is the class of auto-regressive HMMs [27], [28]. For this class, the observation vectors are drawn from an autoregression process.

To be more specific, consider the observation vector o with components (x_o, x_1, x_2, • • ,$x,_{<\_-1}$). Since the basis prob-ability density function for the observation vector is Gaussian autoregressive (or order *p),* then the components of o are related by

$$Ok = {}^{-}E\ aiOk\_i\ e_k \qquad (55)$$

function. Hence it can be shown [24]-[26] that the reestimation formulas for the coefficients of the mixture density, i.e., c_{jm}, u. and *CIA,* are of the form

$$(5(0,\ a) = r_a(0)\ \mathrm{r}(0) + 2\sum_{1=1} r_a(i)\ r(i) \qquad k)$$

$$a' = [1,\ a_1,\ a_2,\ \bullet\bullet\bullet,\ a_p]$$

$$\bar{c}_{jk} = \frac{\sum_{t=1}^{T} \sum_{k=1}^{M} y_t(j, k)}{\sum_{t=1}^{T} \sum_{i=1}^{M} y_t(j, k)}$$

where e_k, $k = 0, 1, 2, \cdots, K - 1$ are Gaussian, independent, identically distributed random variables with zero mean and variance σ^{-2}, and a_i, $i = 1, 2, \cdots, p$, are the autoregression or predictor coefficients. It can be shown that for large K, the density function for σ is approximately

$$f(\sigma) = (2\pi a^2)^{-1/2} \exp\left\{-\frac{}{2a^2} \delta(\sigma, a)\right\} \tag{56}$$

where

(57a)

(57b)

$$r_a(i) = \sum_{n=0}^{P} a_n a_{n+i}, \quad (a_0 = 1), \quad 1 \le i \le p \qquad (57c)$$

$$r(i) = \sum_{n=0}^{K-1-i} x_n x_{n+i}, \quad 0 \le i \le p. \qquad (57d)$$

In the above equations it can be recognized that $r(i)$ is the autocorrelation of the observation samples, and $r_a(i)$ is the autoregressive coefficients.

The total (frame) prediction residual a can be written as

$$E[E(e_n)^2] = Ka' \qquad (58)$$

where a^2 is the variance per sample of the error signal. Consider the normalized observation vector

$$o = \frac{o}{fc; ORP} \qquad (59)$$

where each sample x, is divided byNRP, i.e., each sample is normalized by the sample variance. Then $f(o)$ can be written as

$$f(o) = \left(\frac{K}{(2\pi)^{K/2}}\right) \exp\left(-\frac{1}{2} \boldsymbol{b(O, a)}\right). \qquad (60)$$

In practice, the factor K (in front of the exponential of (60)) is replaced by an *effective* frame length Kwhich represents the effective length of each data vector. Thus if consecutive data vectors are overlapped by 3 to 1, then we would use

$= K/3$ in (60), so that the contribution of each sample of signal to the overall density is counted exactly once.

The way in which we use Gaussian autoregressive density in HMMs is straightforward. We assume a mixture density of the form

$$b_t(o) = \sum_{m=1} c_{i,n} b_{i,n}(o) \qquad (61)$$

where each $b_{i,n}(o)$ is the density defined by (60) with auto-regression vector $\boldsymbol{a_{in}}$, (or equivalently by autocorrelation vector $\boldsymbol{r_a}$), *i.e.,*

$$I), (o) = \left(\frac{K}{(2\pi)^{K/2}}\right) \exp\left(-\frac{1}{2} \boldsymbol{b(o, a_{im})}\right). \qquad (62)$$

(a)

(b) A reestimation formula for the sequence autocorrelation, $r(i)$ of (57d), for the jth state, kth mixture, component has been

(c) derived, and is of the form

$$rik = \frac{\sum_{t=1}^{T} y_t(j, k) \cdot \boldsymbol{r_t}}{\sum_{t=1}^{T} 7,(J, k)} \qquad (63a)$$

where $T_t(j, k)$ is defined as the probability of being in state j at time t and using mixture component k, i.e.,

$$7t(l, k) = \frac{a_t(i)\, i3\,(j) c_{ik} b_{ik}(o)}{\sum_{i=1}^{N} \boldsymbol{a_t(j)}\,)3,(j) \sum_{k=1}^{nT} c_{ik} b_{ik}(o_t)} \qquad (63b)$$

It can be seen that r_{ik} is a weighted sum (by probability of occurrence) of the normalized autocorrelations of the frames in the observation sequence. From r_{ik}, one can solve a set of normal equations to obtain the corresponding auto-regressive

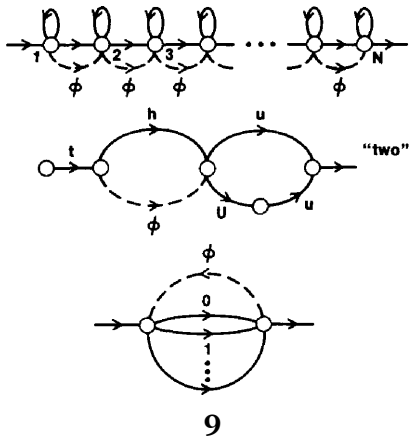
coefficient vector 5_m, for the kth mixture of state

j. The new autocorrection vectors of the autoregression coefficients can then be calculated using (57c), thereby closing the reestimation loop.

C. *Variants on HMM Structures—Null Transitions and Tied States*

Throughout this paper we have considered HMMs in which the observations were associated with states of the model. It is also possible to consider models in which the observations are associated with the arcs of the model. This type of HMM has been used extensively in the IBM continuous speech recognizer [13]. It has been found useful, for this type of model, to allow transitions which produce no output—i.e., jumps from one state to another which produce no observation [13]. Such transitions are called null transitions and are designated by a dashed line with the symbol used to denote the null output.

Fig. 8 illustrates 3 examples (from speech processing tasks) where null arcs have been successfully utilized. The



9

Fig. 8. Examples of networks incorporating null transitions. (a) Left-right model. (b) Finite state network. (**c**) Grammar network.

example of part (a) corresponds to an HMM (a left-right model) with a large number of states in which it is possible to omit transitions between any pair of states. Hence it is possible to generate observation sequences with as few as 1 observation and still account for a path which begins in state 1 and ends in state *N*.

The example of Fig. 8(b) is a finite state network (FSN) representation of award in terms of linguistic unit models (i.e., the sound on each arc is itself an HMM). For this model the null transition gives a compact and efficient way of describing alternate word pronunciations (i.e., symbol delections).

Finally the FSN of Fig. 8(c) shows how the ability to insert a null transition into a grammar network allows a relatively simple network to generate arbitrarily long word (digit)

sequences. In the example shown in Fig. 8(c), the null transition allows the network to generate arbitrary sequences of digits of arbitrary length by returning to the initial state after each individual digit is produced.

Another interesting variation in the HMM structure is the concept of parameter tieing [13]. Basically the idea is to set up an equivalence relation between HMM parameters in

different states. In this manner the number of independent parameters in the model is reduced and the parameter estimation becomes somewhat simpler. Parameter tieing is used in cases where the observation density (for example) is known to be the same in 2 or more states. Such cases occur often in characterizing speech sounds. The technique is especially appropriate in the case where there is insufficient training data to estimate, reliably, a large number of model parameters. For such cases it is appropriate to tie model parameters so as to reduce the number of parameters (i.e., size of the model) thereby making the parameter estimation problem somewhat simpler. We will discuss this method later in this paper.

D. Inclusion of Explicit State Duration Density in HMMs[8] [29], 1301

Perhaps the major weakness of conventional HMMs is the modeling of state duration. Earlier we showed (5) that the inherent duration probability density $p_i(d)$ associated with state S_i, with self transition coefficient a_{ii}, was of the form

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

= probability of d consecutive observations in state S_i.

(64)

For most physical signals, this exponential state duration density is inappropriate. Instead we would prefer to explicitly model duration density in some analytic form. Fig. 9

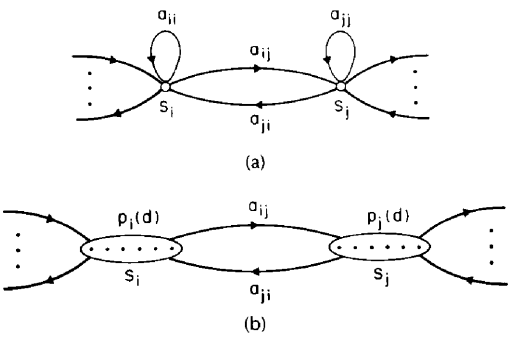


Fig. 9. Illustration of general interstate connections of (a) a normal HMM with exponential state duration density, and (b) a variable duration HMM with specified state densities and no self transitions from a state back to itself.

illustrates, for a pair of model states S_i and S_i, the differences between HMMs without and with explicit duration density. In part (a) the states have exponential duration densities based on self-transition coefficients $a_{„}$ and a_{ll}, respectively. In part (b), the self-transition coefficients are set to zero, and an explicit duration density is specified.[9] For this case, a

[°]In cases where a Bakis type model is used, i.e., left-right models where the number of states is proportional to the average duration, explicit inclusion of

state duration density is neither necessary nor is it useful.

'Again the ideas behind using explicit state duration densities are due to Jack Ferguson of IDA. Most of the material in this section is based on Ferguson's original work. transition is made only after the appropriate number of observations have occurred in the state (as specified by the duration density).

Based on the simple model of Fig. 9(b), the sequence of events of the variable duration HMM is as follows:

1) An initial state, $q_1 = S_i$, is chosen according to the initial state distribution w,.
2) A duration d_1 is chosen according to the state duration density $p_{,i}(d_1)$. (For expedience and ease of implementation the duration density $p_q(d)$ is truncated at a maximum duration value D.)
3) Observations $o_1\ o_2 \cdots O_{tt}$, are chosen according to the joint observation density, $b_{,,}(o_1\ o_2 \cdots O_{th})$- Generallywe assume independent of observations so that $13_{,1}(o_1\ o_2 \cdots ' Odi) = 1^{-}o_{11}\ b_{,,,}(o_t)$.
4) The next state, $q_2 = S_r$ is chosen according to the state transition probabilities, $a_{q1,2}$, with the constraint that $a_{m2} = o$, i.e., no transition back to the same state can occur. (Clearly this is a requirement since we assume that, in state q_t, exactly d_1 observations occur.)

A little thought should convince the reader that the variable duration HMM can be made equivalent to the standard HMM by setting $p,(d)$ to be the exponential density of (64).

Using the above formulation, several changes must be made to the formulas of Section III to allow calculation of P(o1X) and for reestimation of all model parameters. In particular we assume that the first state begins at $t = 1$ and the last state ends at $t = T$, i.e., entire duration intervals are included with the observation sequence. We then define the forward variable $a_t(i)$ as

$$a_t(i) = P(o_1\ o2 \cdots Ot, S, \text{ ends at } tl\ X). \quad (65)$$

We assume that a total of r states have been visited during the first t observations and we denote the states as Q_1, $q2$, $\cdots$, q_r with durations associated with each state of d_1, d_2, $\cdots$, d_r. Thus the constraints of (65) are

$$q, = S_i \quad (66a)$$

$$\sum_{s=1} d, = t. \quad (66b)$$

Equation (65) can then be written as

$$a_t(i) = \sum_{q} p_{a,}(d_i)\ P(o_1\ o2 \cdots Od\text{1}icii) \quad (67)$$

. $^{a}cm2p42^{(d2)\ P}(Od1+1 \cdots$

$°d,+c1_{2}1q2) \cdots$

• a, $_{i}p_{„}(d_{r})$ P(Od_{i} d2 + • • ' $d, _{\_\ 1}$ $+1$ ' " $Orlq1)$

where the sum is over all states q and all possible state durations d. By induction we can write $a_{t}(j)$ as

$$a_{t}(j) = \sum_{r=1}^{N} \sum_{d=1}^{D} a_{r\_d}(i)\, a_{,i}p_{i}(d) \quad II \quad b_{l}(o_{5}) \qquad (68)$$

s--t-d+ 1

where D is the maximum duration within any state. To initialize the computation of $a_{t}(j)$ we use

$$a_{1}(i) = 7r_{i}p,(1) \cdot b,(o_{1}) \qquad (69a)$$

$$a_{2}(i) = 7_{1}\text{-}\ {}_{i}p,(2)b_{1}(o_{5}) + \sum_{s=1}^{2} (x_{i}(j)\, a_{ii}p,(1)\, b1(o_{2}) \qquad (69b)$$

$t=1$

$1^{*}\ i$

$$a_3(\blacktriangleright) = 74),(3)\ II\ b_1(0_5) + \sum_{d=1}^{2N} \sum_{j=1}^{} a_{3-d}(j)\ a.\ p,(d)$$

$$\bullet\ \prod_{s=4-d}^{3} b_i(o_5) \qquad (69c)$$

etc., until $«_D(i)$ is computed; then (68) can be used for all $t > D$. It should be clear that the desired probability of o given the model X can be written in terms of the a's as

$$1^{)}(01X) = a_T(i) \qquad (70)$$

as was previously used for ordinary HMMs.

In order to give reestimation formulas for all the variables of the variable duration HMM, we must define three more forward-backward variables, namely

$$a;'(i) = P(o_1\ o2 \bullet\bullet\bullet O_{,}, S.\ \text{begins at}\ t + 1IX) \qquad (71)$$

$$St(i) = P(Or+1 \bullet\bullet\bullet O_T IS, \text{ends at}\ t, X) \qquad (78)$$

$$137(i) = P(O_{t+}, \bullet\bullet O_T IS, \text{begins at}\ t + 1, X). \qquad (79)$$

The relationships between a, a^*, o, and 13^* are as follows:

$$4(j) = a_t(i)a_{,i} \qquad (74)$$

$$at(i)d=1\ at^*\text{-}d(i)1);(d)\ s = td + 1\ b_i(o_s) \qquad (75)$$

$$o,(i) = \sum_{j=1}^{N} a_{,i}\,o7(j) \qquad (76)$$

$$o;'^{(i)} = E_1\ Ot+\ d(i)\ PO\ )_s\ )\ 11_1\ b,(o_5). \qquad (77)$$

$$d = = t +$$

Based on the above relationships and definitions, the reestimation formulas for the variable duration HMM are

$$\frac{7,\ oO(i)\ 7,}{P(o1X)} = \frac{a_t(i)\ aiig(j)}{\sum_{t=1}^{NT} at(i)\ a1^3(j)\ j=1\ t\text{-}1}$$

$$\sum_{t=1\_r<t}^{T} o17(i) \bullet \text{s.t.o,=k}$$

$$\sum_{k=1}^{M} \sum_{\text{s.t.ot=vk}}^{T} t=1\ _r<t 4(i)\ \ 1^3{:}(i)\ qr(j)\ 1^3,(i)1 \qquad (80)$$

$$= \frac{\sum_{t=1}^{T} E\ 4o1\ Pi(d)\ Ot+d(i)\ _{t+d}\ WOO}{\sum_{s=t+1}^{t+d}} \qquad (81)$$

$$p,(d) = \frac{}{} \qquad (83)$$

(72)

with parameters v, and m and with mean v,m^{-1} and variance

Reestimation formulas for m and v, have been derived and used with good results [19]. Another possibility, which has been used with good success, is to assume a uniform duration distribution (over an appropriate range of durations) and use a path-constrained Viterbi decoding procedure [31].

(73)

$$\sum_{d=1}^{} E\ \sum_{t=1}^{} a7\ (i)\ p1(d)\ lit+d(i)\ IIb(o_s) \qquad 1)$$

The interpretation of the reestimation formulas is the following. The formula for Tr, is the probability that state i was the first state, given o. The formula for 54 *is almost the same as for the usual* ***HMM*** except *it uses the condition that the alpha terms in which a state ends at t,* join with the beta

terms in which a new state begins at $t + 1$. The formula for $b,(k)$ (assuming a discrete density) is the expected

number of times that observation $O_t = v_k$ occurred in state *i,* normalized by the expected number of times that any observation occurred in state i. Finally, the reestimation formula for *p,(d)* is the ratio of the expected number of times state i occurred with duration *d,* tothe expected number of times state *i* occurred with any duration.

The importance of incorporating state duration densities is reflected in the observation that, for some problems, the quality of the modeling is significantly improved when explicit state duration densities are used. However, there are drawbacks to the use of the variable duration model discussed in this section. One is the greatly increased computational load associated with using variable durations. It can be seen from the definition and initialization conditions on the forward variable a,(i), from (68)-(69), that about ***D*** times the storage and *D²12* times the computation is required. For *D*on the order of 25 (as is reasonable for many speech processing problems), computation is increased by a factor of 300. Another problem with the variable duration models is the large number of parameters *(D),* associated with each state, that must be estimated, in addition to the usual HMM parameters. Furthermore, for a fixed number of observations T, in the training set, there are, on average, fewer state transitions and much less data to estimate *p,(d)* than would be used in a standard HMM. Thus the reestimation problem is more difficult for variable duration HMMs than for the standard HMM.

One proposal to alleviate some of these problems is to use a parametric state duration density instead of the non-parametric *p,(d)* used above [29], [30]. In particular, proposals include the Gaussian family with

$$p_i(d) = 1M(d, p,,, \qquad (82)$$

with parameters and *4,* or the Gamma family with

E. Optimization Criterion—ML, MMI, and MDI [32], [33]

The basic philosophy of HMMs is that a signal (or observation sequence) can be well modeled if the parameters of an HMM are carefully and correctly chosen. The problem with this philosophy is that it is sometimes inaccurate—either because the signal does not obey the constraints of the HMM, or because it is too difficult to get reliable estimates of all HMM parameters. To alleviate this type of problem, there has been proposed at least two alternatives to the standard maximum likelihood (ML) optimization procedure for estimating HMM parameters.

The first alternative [32] is based on the idea that several HMMs are to be designed and we wish to design them all at the same time in such a way so as to maximize the discrimination power of each model (i.e., each model's ability

to distinguish between observation sequences generated by the correct model and those generated by alternative models). We denote the different HMMs as $X_{,,}$ v = 1, 2, • • • , V. The standard ML design criterion is to use a separate training sequence of observations o" to derive model parameters for each model X. Thus the standard ML optimization yields

$$= \max P(o1Xp). \quad (84)$$

The proposed alternative design criterion [31] is the maximum mutual information (MMI) criterion in which the average mutual information I between the observation sequence o" and the *complete* set of models X = (X1, X2, • • • , Xv) is maximized. One possible way of implementing this[l°] is

$$= \max [\log P(OPIX,) — \log E_{-1}$$

$$P(o1\ Xo1 \quad (85)$$

i.e., choose X so as to separate the correct model X. from all other models on the training sequence o". By summing (85) over all training sequences, one would hope to attain the most separated set of models possible. Thus a possible implementation would be

$$/* = \max_{I} E [\log P(o"IX_{„}) — \log <._,Y$$

$$P(ORIX,,,,)\}. \quad (86)$$

$$x \qquad w=1$$

There are various theoretical reasons why analytical (or reestimation type) solutions to (86) cannot be realized. Thus the only known way of actually solving (86) is via general optimization procedures like the steepest descent methods [32].

The second alternative philosophy is to assume that the signal to be modeled was not necessarily generated by a Markov source, but does obey certain constraints (e.g., positive definite correlation function) [33]. The goal of the design procedure is therefore to choose HMM parameters which minimize the discrimination information (DI) or the cross entropy between the set of valid (i.e., which satisfy the measurements) signal probability densities (call this set Q), and the set of HMM probability densities (call this set P_h), where the DI between Q and P_5 can generally be written in the form

$$D(QiiP5) = q(y)\ In\ (q(y)Ip(y))\ dy \quad (87)$$

where q and p are the probability density functions corresponding to Q and P_5. Techniques for minimizing (87) (thereby giving an MD1 solution) for the optimum values of

X = (A, *B, 7r)* are highly nontrivial; however, they use a generalized Baum algorithm as the core of each iteration, and thus are efficiently tailored to hidden Markov modeling [33].

It has been shown that the ML, MMI, and MD1 approaches can all be uniformly formulated as MDI approaches.' The three approaches differ in either the probability density attributed to the source being modeled, or in the model wIn (85) and (86) we assume that all words are equiprobable, i.e., p(w) = 1/V.

"Y. Ephraim and L. Rabiner, "On the Relations Between Modeling Approaches for Speech Recognition,"to appear in IEEE TRANSACTIONS ON INFORMATION THEORY. effectively being used. None of the approaches, however, assumes that the source has the probability distribution of the model.

F. Comparison of HMMs [34]

An interesting question associated with HMMs is the following: Given two HMMs, X_1 and X2, what is a reasonable measure of the similarity of the two models? A key point here is the similarity criterion. By way of example, consider the case of two models

$A_1 = (A_1, B_1, 7f_1)$ $X_2 = (A_2, B_2, 71^{-2})$ with

$$A_1 = \begin{vmatrix} p & 1-p \\ 1- & p \end{vmatrix} = \begin{vmatrix} & 1 & ql \\ q & q & \end{vmatrix}$$

$$21_1 = [1/2\ 1/2] \qquad {}^{1-.5]}_{\cdot\_1-s\ s} \qquad {}^{?}r2 = [1/2\ 1/2].$$

and

$$A_2 = \begin{bmatrix} r & 1-rl \\ 1-r & r \end{bmatrix} B_2 =$$

For X_1 to be equivalent to X_2, in the sense of having the same statistical properties for the observation symbols, i.e., $E[O_r = v_k iX_i] = E[O_t = vklX_2]$, for all v_k, we require

$$pq + (1-p)(1-q)rs + (1-r)(1-s)$$

or, by solving for s, we get

$$s = \frac{p + q - 2pq}{1-2r}$$

By choosing (arbitrarily) $p = 0.6$, $q = 0.7$, $r = 0.2$, we get $s = 13/30 = 0.433$. Thus, even when the two models, X_1 and X2, look ostensibly very different (i.e., A_l is very different from A2 and B_1 is very different from 8_2), statistical equivalence of the models can occur.

We can generalize the concept of model distance (dissimilarity) by defining a distance measure $D(X_1, X_2)$, between two Markov models, X_1 and X2, as

$$D(X_1, X_2) = {}^{1}\log P(o^{(2)}1X1) - \log P(o^{12)}1X2)] \qquad (88)$$

where $o^{21} = o_1 o_2\ o_3 \cdots O_T$ is a sequence of observations *generated* by model X2 [34]. Basically (88) is a measure of how well model A_l matches observations generated by model X2, relative to how well model X2 matches observations generated by itself. Several interpretations of (88) exist in terms of cross entropy, or divergence, or discrimination information [34].

One of the problems with the distance measure of (88) isthat it is nonsymmetric. Hence a natural expression of this measure is the symmetrized version, namely

$$D_s(X_,, X_2) = \frac{D(X_1X_2) + D(X_2, X_i)}{2} \qquad (89)$$

V. IMPLEMENTATION ISSUES FOR HMMs

The discussion in the previous two sections has primarily dealt with the theory of HMMs and several variations on the form of the model. In this section we deal with several practical implementation issues including scaling, multiple