AT&T Technical Journal
Vol. 64, No. 2, February 1985
Printed in U.S.A.

A Probabilistic Distance Measure for Hidden Markov Models

By B.-H. JUANG and L. R. RABINER*

(Manuscript received July 31, 1984)

We propose a probabilistic distance measure for measuring the dissimilarity between pairs of hidden Markov models with arbitrary observation densities. The measure is based on the Kullback-Leibler number and is consistent with the reestimation technique for hidden Markov models. Numerical examples that demonstrate the utility of the proposed distance measure are given for hidden Markov models with discrete densities. We also discuss the effects of various parameter deviations in the Markov models on the resulting distance, and study the relationships among parameter estimates (obtained from reestimation), initial guesses of parameter values, and observation duration through the use of the measure.

I. INTRODUCTION

Consider two N-state first-order hidden Markov models specified by the parameter sets $\lambda_i = (\mathbf{u}^{(i)}, \mathbf{A}^{(i)}, \mathbf{B}^{(i)})$, $i = 1, 2$, where $\mathbf{u}^{(i)}$ is the initial state probability vector, $\mathbf{A}^{(i)}$ is the state transition probability matrix, and $\mathbf{B}^{(i)}$ is either an $N \times M$ stochastic matrix (if the observations are discrete) or a set of N continuous density functions (if the observations are continuous).[1] Our interest in this paper is to define a distance for every such pair of hidden Markov models (λ_1, λ_2) so we can measure the dissimilarity between them. Another goal is to study the properties of hidden Markov models, using the distance measure, in order to understand the model sensitivities.

* Authors are employees of AT&T Bell Laboratories.

The need of such a distance measure arises mainly in estimation and classification problems involving Hidden Markov Models (HMMs). For example, in using a reestimation algorithm to iteratively estimate the model parameters,[2] a distance measure is necessary not only to monitor the behavior of the reestimation procedure, but to indicate the expected performance of the resulting HMM. In classification, a good distance measure would greatly facilitate the nearest-neighbor search, defining Voronoi regions[3] or applying the generalized Lloyd algorithm[4] for hidden Markov model clustering.

The only measure for comparing pairs of HMMs that has appeared previously in the literature is the one proposed by Levinson et al. for discrete-observation density hidden Markov models.[1] The distance, which is a Euclidean distance on the state-observation probability matrices, is defined as

$$d(\lambda_1, \lambda_2) \triangleq ||\mathbf{B}^{(1)} - \mathbf{B}^{(2)}|| \triangleq \left\{\frac{1}{MN} \sum_{j=1}^{N} \sum_{k=1}^{M} [b_{jk}^{(1)} - b_{p(j)k}^{(2)}]^2\right\}^{1/2}, \quad (1)$$

where $\mathbf{B}^{(i)} = [b_{jk}^{(i)}]$ is the state-observation probability matrix in model λ_i and $p(j)$ is the state permutation that minimizes the measure of eq. (1). The metric of eq. (1) was called a "measure of estimation error" in Ref. 1 and was used to characterize the estimation error occurring in the reestimation process. Minimum bipartite matching was used to determine the optimum state permutation for aligning the states of the two models. The measure of eq. (1) did not depend at all on estimates of $\mathbf{u}$ or $\mathbf{A}$, since it is generally agreed that the $\mathbf{B}$ matrix is, in most cases, a more sensitive set of parameters related to the closeness of HMMs than the $\mathbf{u}$ vector or the $\mathbf{A}$ matrix.

The distance measure of eq. (1) is inadequate for the following reasons: (1) it does not take into account the deviations in all the parameters of the HMM; (2) its evaluation requires a great deal of computation in the discrete case and probably would become intractable when dealing with continuous-observation hidden Markov models; and (3) it is unreliable when comparing HMMs with highly skewed densities. Hence, our aim is to find a distance measure that truly measures the dissimilarity between pairs of hidden Markov models, can be easily evaluated, is reliable for any pair of Markov models, and is meaningful in the probabilistic framework of the HMM itself.

In this paper, we propose such a distance measure for comparing pairs of HMMs that follows the concept of divergence,[5] cross entropy, or discrimination information.[6] The distance measure, denoted by $D(\lambda_1, \lambda_2)$, has the form

$$\log \Pr(\mathbf{O}_T | \lambda_1) - \log \Pr(\mathbf{O}_T | \lambda_2),$$

where $\mathbf{O}_T$ symbolizes an observation sequence of T observations. Because the distance measure is the difference in log probabilities of the observation sequence conditioned on the models being compared, it will sometimes be referred to as "divergence distance" or "directed divergence measure." In the next section, we formally define the distance measure, and we discuss Petrie's results[7] that further give the distance measure theoretical justification. In Section III, numerical examples related to discrete-observation hidden Markov models are given. We show the effects of individual parameter deviations upon the distance measure and demonstrate several interesting properties of discrete-observation models that are made explicit through the use of the proposed distance measure. A discussion of the use of such a distance measure in continuous-observation models is given in Ref. 8, where hidden Markov models with continuous mixture densities are discussed.

II. DEFINITION OF THE PROPOSED HMM DISTANCE MEASURE

In this section, we define the distance measure for any pair of Markov models, discuss Petrie's Limit Theorem and statistical analysis of probabilistic functions of Markov chains,[7] and then give the proposed distance measure an interpretation from the Kullback-Liebler statistic point of view. The presentation is explicit for discrete-observation models but can easily be extended to continuous-observation cases.

Let $\mathscr{A}_s = \{1, 2, \cdots, N\}$ be a state alphabet, and let $\mathscr{A}_0 = \{y_1, y_2, \cdots, y_M\}$ be an observation alphabet. The Cartesian product $\mathscr{O}_\infty = \prod_{t=1}^{\infty} \mathscr{A}_{0t}$, $\mathscr{A}_{0t} = \mathscr{A}_0$, for all t, forms an observation space in which every point $\mathbf{O}$ has coordinate $\mathbf{o}_t \in \mathscr{A}_{0t} = \mathscr{A}_0$. We are concerned about a class of stochastic processes generated by a hidden Markov source defined by an $N \times N$ ergodic stochastic matrix $\mathbf{A} = [a_{ij}]$ and by an $N \times M$ stochastic matrix $\mathbf{B} = [b_{jk}]$. Matrix $\mathbf{A}$, the state transition probability matrix, generates a stationary Markov process $\mathbf{S} = \cdots \mathbf{s}_{t-1}\mathbf{s}_t\mathbf{s}_{t+1} \cdots$ according to $a_{ij} = \Pr\{\mathbf{s}_{t+1} = j \mid \mathbf{s}_t = i\}$. Based upon $\mathbf{S}$, $\mathbf{B}$ generates $\mathbf{o}_t$ according to $b_{jk} = \Pr\{\mathbf{o}_t = \mathrm{y}_k \mid \mathbf{s}_t = \mathrm{j}\}$. Let $\mathbf{a}^* = [a_1, a_2, \cdots, a_N]$ be the stationary absolute distribution vector for $\mathbf{A}$, i.e., $\mathbf{a}^*\mathbf{A} = \mathbf{a}^*$, where $*$ denotes the transpose. Then, matrices $\mathbf{A}$ and $\mathbf{B}$ define a measure, denoted by $\mu(\cdot \mid \lambda)$, where $\lambda = (\mathbf{A}, \mathbf{B})$, on $\mathscr{O}_\infty$ by

$$\mu(\mathbf{O}_T \mid \lambda) = \sum_{\text{all } \mathbf{S}_T} a_{\mathbf{s}_0} \prod_{t=1}^{T} a_{\mathbf{s}_{t-1}\mathbf{s}_t} b_{\mathbf{s}_t I(\mathbf{o}_t)}, \tag{2}$$

where $\mathbf{O}_T = (\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_T)$ is the observed sequence up to time T (i.e., a truncated $\mathbf{O}$), $\mathbf{S}_T = (\mathbf{s}_0, \mathbf{s}_1, \cdots, \mathbf{s}_T)$ is the corresponding unobserved state sequence, and $I(\cdot)$ is the index function

$$I(\mathbf{o}_t) = k \quad \text{if} \quad \mathbf{o}_t = y_k.$$

Let Λ_a be the space of $N \times N$ *ergodic* stochastic matrices, Λ_b be the space of $N \times M$ stochastic matrices, and $\Lambda = \Lambda_a \times \Lambda_b$. Clearly, $\lambda \in \Lambda$, and for every point in Λ there is a stationary measure $\mu(\cdot \mid \lambda)$ associated with it.

Now consider a probability space $(\mathscr{O}_\infty, \mu(\cdot \mid \lambda_0))$, which will be abbreviated as $(\mathscr{O}_\infty, \lambda_0)$ in the following without ambiguity. Let an observation sequence $\mathbf{O}_T$ be generated according to the distribution $\mu(\cdot \mid \lambda_0)$.

For each T and each $\mathbf{O} \in \mathscr{O}_\infty$, define the function $H_T(\mathbf{O}, \lambda)$ on Λ by

$$H_T(\mathbf{O}, \lambda) = \frac{1}{T} \log \mu(\mathbf{O}_T \mid \lambda). \tag{3}$$

Each $H_T(\cdot, \lambda)$ is thus a random variable on the probability space $(\mathscr{O}_\infty, \lambda_0)$. Also, for a given fixed observation $\mathbf{O}_T$, $H_T(\mathbf{O}, \lambda)$ is a function on Λ. Petrie[7] proved (limit theorem) that for each λ in Λ,

$$\begin{aligned} \lim_{T\to\infty} H_T(\mathbf{O}, \lambda) &= \lim_{T\to\infty} \frac{1}{T} \log \mu(\mathbf{O}_T \mid \lambda) \\ &= H(\lambda_0, \lambda) \end{aligned} \tag{4}$$

exists almost everywhere $\mu(\cdot \mid \lambda_0)$. Furthermore,

$$H(\lambda_0, \lambda_0) \geqslant H(\lambda_0, \lambda) \tag{5}$$

with equality if and only if $\lambda \in G(\lambda_0) = \{\lambda \in \Lambda \mid \mu(\cdot \mid \lambda) = \mu(\cdot \mid \lambda_0)$ as measures on $\mathscr{O}_\infty\}$. Define $\Lambda_T(\mathbf{O}) = \{\lambda' \in \Lambda \mid H_T(\mathbf{O}, \lambda)$ is maximized at $\lambda'\}$. Then, $\Lambda_T(\mathbf{O}) \rightarrow G(\lambda_0)$ almost everywhere $\mu(\cdot \mid \lambda_0)$ (see Ref. 7, Theorem 2.8). The results give further justification to the well-known reestimation procedure[9] for Markov modeling.

With the above background, we define a distance measure $D(\lambda_0, \lambda)$ between two Markov sources λ_0 and λ by

$$\begin{aligned} D(\lambda_0, \lambda) &= H(\lambda_0, \lambda_0) - H(\lambda_0, \lambda) \\ &= \lim_{T\to\infty} \frac{1}{T} [\log \mu(\mathbf{O}_T \mid \lambda_0) - \log \mu(\mathbf{O}_T \mid \lambda)]. \end{aligned} \tag{6}$$

The aforementioned limit theorem guarantees the existence of such a distance measure and eq. (5) ensures that $D(\lambda_0, \lambda)$ is nonnegative. $D(\lambda_0, \lambda) = 0$ if and only if $\lambda \in G(\lambda_0)$, a point that is indistinguishable by the associated probability measure.

By invoking ergodicity,[10] we see that the distance is in fact the Kullback-Leibler number[6] between measures $\mu(\cdot \mid \lambda_0)$ and $\mu(\cdot \mid \lambda)$. If $\mathscr{H}_0$ and $\mathscr{H}_1$ are the hypotheses that $\mathbf{O}_T$ is from the statistical population with measure $\mu(\cdot \mid \lambda_0)$ and $\mu(\cdot \mid \lambda_1)$, respectively, $D(\lambda_0, \lambda)$ is then the average information per observation sample in $\mathbf{O}_T$ for discrimination in favor of $\mathscr{H}_0$ against $\mathscr{H}_1$. Since $\mathbf{O}_T$ is generated according to $\mu(\cdot \mid \lambda_0)$, $\lim_{T\to\infty}(1/T)\log \mu(\mathbf{O}_T \mid \lambda_0)$ should be a maximum over Λ, and $D(\lambda_0, \lambda)$ is a measure of directed divergence, from λ_0 to λ, manifested by the observation $\mathbf{O}_T$.

The distance measure of eq. (6) is clearly nonsymmetric. A natural extension of this measure is the symmetrized version of eq. (6), i.e.,

$$D_s(\lambda_0, \lambda) = \frac{1}{2}[D(\lambda_0, \lambda) + D(\lambda, \lambda_0)], \tag{7}$$

which is the average of the two nonsymmetric distances. $D_s(\lambda_0, \lambda)$ is symmetric with respect to λ_0 and λ and represents a measure of the difficulty (or ease) of discriminating between $\mu(\cdot \mid \lambda_0)$ and $\mu(\cdot \mid \lambda)$, or equivalently, λ_0 and λ. For our purpose, however, there is no particular requirement that the distance be symmetric, and our study will mainly concentrate on the definition of eq. (6).

III. DISCRETE-OBSERVATION HIDDEN MARKOV MODELS

Using the distance measure of eq. (6), we have studied the behavior of several discrete-observation hidden Markov models. In this section, we present some results on the sensitivities of the reestimation procedure to observation sequence length, initial parameter estimates, etc. We begin with a discussion of the evaluation of such a distance measure.

3.1 Evaluation of the distance measure

Evaluation of the distance of eq. (6) is rather straightforward. A standard Monte Carlo simulation procedure based upon a good random number generator is used to generate the required observation sequence $\mathbf{O}_T$ according to the given distribution $\mu(\cdot \mid \lambda_0)$. The probabilities of observing the generated sequence from models λ_0 and λ are then calculated respectively. By way of example, Fig. 1a shows the logarithm of $\mu(\mathbf{O}_T \mid \lambda_0)$ and $\mu(\mathbf{O}_T \mid \lambda)$, respectively, as a function of the observation duration T. The resulting distance $D(\lambda_0, \lambda)$ is then plotted in Fig. 1b. For this example, $\lambda_0 = (\mathbf{A}_0, \mathbf{B}_0)$, $\lambda = (\mathbf{A}, \mathbf{B})$, $N = M = 4$, where

$$\mathbf{A}_0 = \begin{bmatrix} 0.8 & 0.15 & 0.05 & 0 \\ 0.07 & 0.75 & 0.12 & 0.06 \\ 0.05 & 0.14 & 0.8 & 0.01 \\ 0.001 & 0.089 & 0.11 & 0.8 \end{bmatrix} \quad \mathbf{B}_0 = \begin{bmatrix} 0.3 & 0.4 & 0.2 & 0.1 \\ 0.5 & 0.3 & 0.1 & 0.1 \\ 0.1 & 0.2 & 0.4 & 0.3 \\ 0.4 & 0.3 & 0.1 & 0.2 \end{bmatrix}$$

and

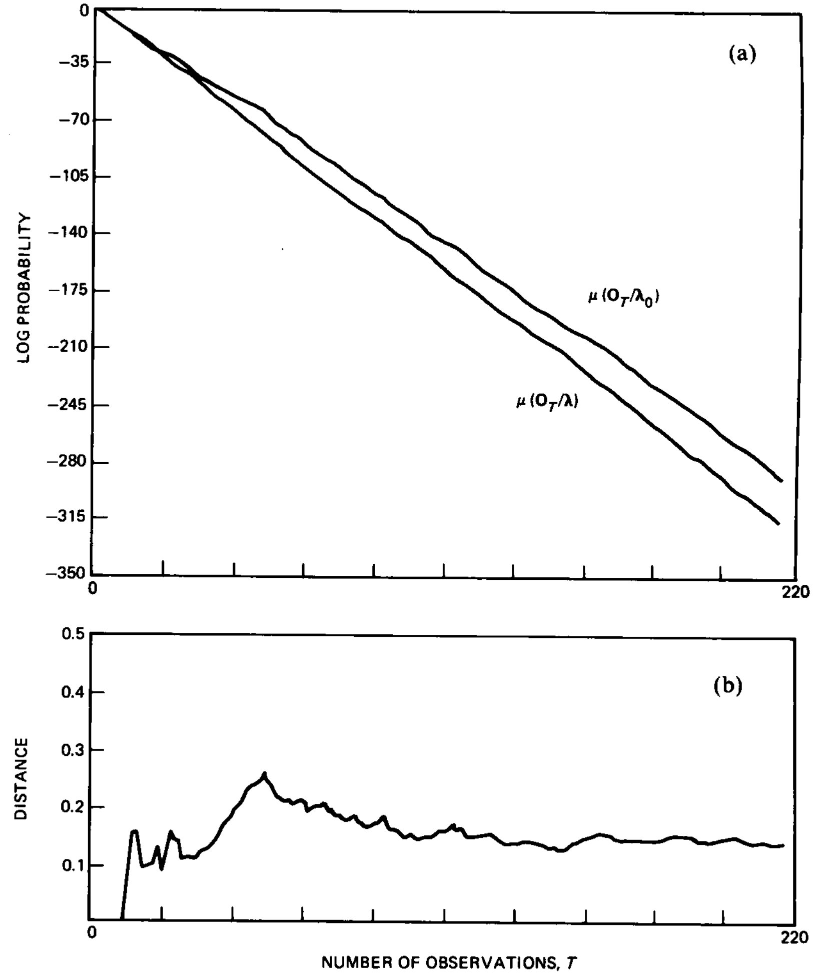

Fig. 1—(a) Log probabilities $\mu(\mathbf{O}_T \mid \lambda_0)$ and $\mu(\mathbf{O}_T \mid \lambda)$ versus the number of observations for a pair of models that are close in distance. (b) Distance $D(\lambda_0, \lambda)$ versus the number of observations for the same pair of models.

$$\mathbf{A} = \begin{bmatrix} 0.4 & 0.25 & 0.15 & 0.2 \\ 0.27 & 0.45 & 0.22 & 0.06 \\ 0.35 & 0.14 & 0.4 & 0.11 \\ 0.111 & 0.119 & 0.23 & 0.54 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 0.1 & 0.15 & 0.65 & 0.1 \\ 0.2 & 0.3 & 0.4 & 0.1 \\ 0.3 & 0.3 & 0.1 & 0.3 \\ 0.15 & 0.25 & 0.4 & 0.2 \end{bmatrix}.$$

We can see from Fig. 1 that, for this example, it takes around 150 observation samples to converge to a distance of 0.14 (to within statistical fluctuations). It is readily shown that the number of observations needed for convergence of the distance to a fixed value is dependent on N and M.

Although the definition of the distance of eq. (6) requires that the pair of models being compared both be ergodic and that there exist a stationary absolute distribution vector $\mathbf{a}$ such that $\mathbf{a}^* \mathbf{A} = \mathbf{a}^*$, practical evaluation of the distance can still be performed for other types of Markov models. We often define the distance measure by replacing the stationary equilibrium distribution vector with the initial state probability vector. In the case of left-to-right models,[1] we use a series of restarted sequences as the generated sequence for distance evaluation, because of the trap state in left-to-right models. In fact, except for some possible minor theoretical discrepancies (which might be traced back to the problem of nonergodic model estimation), the proposed distance measure appears to work quite reliably for any pair of such HMMs. Particularly, in the previous example, the initial state probability vectors associated with models λ_0 and λ were $\mathbf{u}_0^* = [0.75\ 0.15\ 0.05\ 0.05]$ and $\mathbf{u}^* = [0.4\ 0.25\ 0.15\ 0.2]$, respectively.

3.2 Effects of parameter deviations on the distance

We are interested in studying the relationship between parameter deviation and model distance, as well as the relative sensitivity of the distance to different parameter sets that define the HMMs. To illustrate such parameter sensitivities, we have studied HMMs whose parameters are related to the matrices $\mathbf{W}_1$ and $\mathbf{W}_2$,

$$\mathbf{W}_1 = \begin{bmatrix} 0.3 & 0.4 & 0.2 & 0.1 \\ 0.5 & 0.3 & 0.1 & 0.1 \\ 0.1 & 0.2 & 0.4 & 0.3 \\ 0.4 & 0.3 & 0.1 & 0.2 \end{bmatrix}, \quad \mathbf{W}_2 = \begin{bmatrix} 0.05 & 0.1 & 0.65 & 0.25 \\ 0.1 & 0.05 & 0.75 & 0.1 \\ 0.45 & 0.45 & 0.05 & 0.05 \\ 0.05 & 0.1 & 0.65 & 0.2 \end{bmatrix},$$

and to the vector $\mathbf{v}^* = [0.75\ 0.15\ 0.05\ 0.05]$. In particular, model λ_0 is defined by $(\mathbf{u}_0, \mathbf{A}_0, \mathbf{B}_0)$, where $\mathbf{u}_0 = \mathbf{v}$ and $\mathbf{A}_0 = \mathbf{B}_0 = \mathbf{W}_1$. We chose $\mathbf{A}_0 = \mathbf{B}_0$ to avoid a priori numerical difference in different parameter sets. The alternate model, $\lambda = (\mathbf{u}, \mathbf{A}, \mathbf{B})$, is varied from λ_0 by modifying, in turn, either $\mathbf{A}$ or $\mathbf{B}$.

We first study the effect of changes in only the state transition probability matrix on the computed distance. We form a sequence of models $\lambda = (\mathbf{u}, \mathbf{A}, \mathbf{B})$, where $\mathbf{u} = \mathbf{u}_0 = \mathbf{v}$, $\mathbf{B} = \mathbf{B}_0 = \mathbf{W}_1$ and

$$\mathbf{A} = \left(\frac{1}{1+\delta}\right)\mathbf{W}_1 + \left(\frac{\delta}{1+\delta}\right)\mathbf{W}_2, \tag{8}$$

with δ varying from 0.001 to 0.991 in 99 equal steps. For each pair (λ_0, λ), $D(\lambda_0, \lambda)$ is then evaluated. The bottom curve in Fig. 2a shows a plot of $D(\lambda_0, \lambda)$ as a function of the deviation factor δ. Furthermore, for potential geometric interpretations, we calculate the signal-to-noise ratio γ_A, defined by

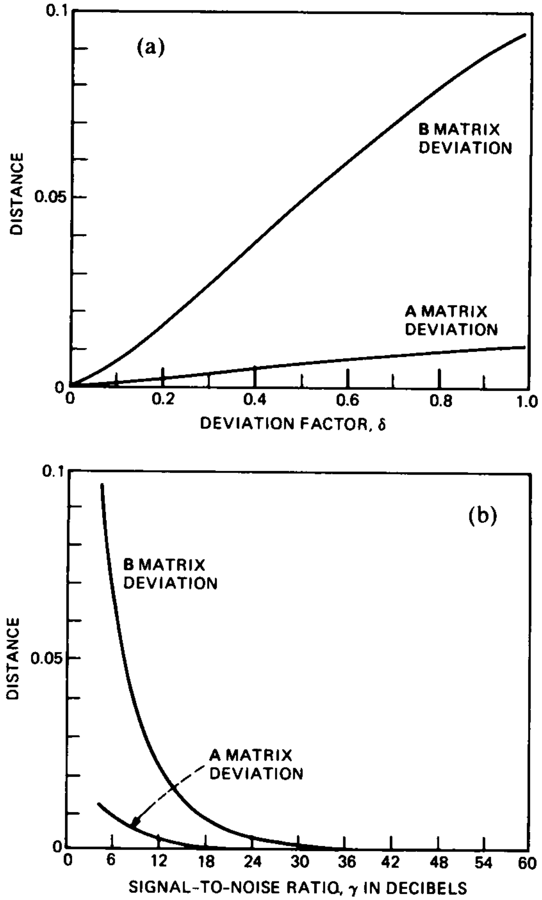

Fig. 2—(a) Relationship between the probabilistic distance and the model deviation factor δ of the **A** and **B** parameters for a pair of HMMs. (b) Relationship between the probability distance and measured parameter deviations of the **A** and **B** parameters for a pair of HMMs.

$$\gamma_{\mathrm{A}} = 10 \log_{10} \frac{||A_0||^2}{||A_0 - A||^2}, \tag{9}$$

where $||\cdot||$ denotes matrix norm ($||\mathbf{A}||^2 = \sum_i \sum_j a_{ij}^2$ for $\mathbf{A} = [a_{ij}]$). For small δ, **A** is very close to $\mathbf{A}_0$ and γ_{A} is large. Accordingly, the distance $D(\lambda_0, \lambda)$ as a function of γ_{A} is plotted in Fig. 2b. (Note that small values of δ in Fig. 2a correspond to large values of γ_{A} in Fig. 2b—i.e., the direction of the curves is reversed.)

Similarly, we study the effect of changes in only the observation probability matrix **B**. The sequence of models $\lambda = (\mathbf{u}, \mathbf{A}, \mathbf{B})$ for comparison is formed by setting $\mathbf{u} = \mathbf{u}_0 = \mathbf{v}$, $\mathbf{A} = \mathbf{A}_0 = \mathbf{W}_1$ and

$$\mathbf{B} = \left(\frac{1}{1+\delta}\right)\mathbf{W}_1 + \left(\frac{\delta}{1+\delta}\right)\mathbf{W}_2, \tag{10}$$

again, with δ varying from 0.001 to 0.991. The relationships, $D(\lambda_0, \lambda)$ versus δ and $D(\lambda_0, \lambda)$ versus $\gamma_{\mathbf{B}}$,

$$\gamma_{\mathbf{B}} = 10 \log_{10} \frac{||\mathbf{B}_0||^2}{||\mathbf{B}_0 - \mathbf{B}||^2}, \tag{11}$$

are shown as the upper curves in Figs. 2a and b, respectively.

Both curves of Fig. 2b show a simple monotonic exponential trend for the example studied. This exponential trend may be intuitively anticipated from eq. (2), which shows that μ is in the form of a product. This monotonic relationship is, in general, true when the signal-to-noise ratio is adequately high, i.e., models are close enough in the Euclidean distance sense. This result is consistent with Theorem 3.19 in Ref. 7, which gives the set $\mathbf{G}(\lambda_0)$ a geometric interpretation. For more complicated models or other types of deviations than those of eqs. (8) and (10), however, the simple monotonic exponential relationship of the type shown in Fig. 2b may not be observed in low signal-to-noise ratio regions.

Another important property of the distance measure, as seen in Fig. 2, is that deviations in the observation probability matrix **B** give, in general, larger distance scores than similar deviations in the state-transition matrix **A**. Thus, the **B** matrix appears to be numerically more important than the **A** matrix in specifying a hidden Markov model. It is our opinion that this may be a desirable inherent property of hidden Markov models for speech recognition applications.

3.3 Examples of the use of the distance in model estimation

3.3.1 Ergodic models

Consider the following models:

1. λ_a: $N = M = 4$, balanced model

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0 & 0.5 \\ 0.5 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0 & 0.5 \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix};$$

2. λ_b: $N = M = 4$, skewed model

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0.25 & 0.75 \\ 0.15 & 0 & 0 & 0.85 \\ 0.2 & 0.8 & 0 & 0 \\ 0 & 0.22 & 0.78 & 0 \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} 0.25 & 0.75 & 0 & 0 \\ 0 & 0.15 & 0.85 & 0 \\ 0 & 0 & 0.1 & 0.9 \\ 0.2 & 0 & 0 & 0.8 \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix};$$

3. λ_c: $N = M = 5$, deterministic observation

$$\mathbf{A} = \begin{bmatrix} 0 & 0.8 & 0.1 & 0.1 & 0 \\ 0 & 0.5 & 0.5 & 0 & 0 \\ 0.4 & 0 & 0.2 & 0 & 0.4 \\ 0.3 & 0.2 & 0.1 & 0.4 & 0 \\ 0.2 & 0.1 & 0.2 & 0.4 & 0.1 \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

It should be pointed out that λ_a is a balanced model in which transitions as well as observations are equiprobable within the structural constraints, while λ_b is a skewed model with the same Markov chain structure as λ_a. Model λ_c has a unique observation probability matrix, namely the identity matrix, which links observations to distinct model states.

A number of observation sequences, $\mathbf{O}_T$, of different duration were generated from these models. Then, for each $\mathbf{O}_T$ sequence, a model estimate, generically denoted as λ_a', λ_b', or λ_c', was obtained using the reestimation algorithm, which, starting from an arbitrary guess, iterated until a certain convergence criterion was met.[1]

Each sequence $\mathbf{O}_T$ of duration T thus corresponds to a model estimate for which the divergence distance can be evaluated from the generating model. Figures 3a, b, and c are plots of $D(\lambda_a, \lambda_a')$, $D(\lambda_b, \lambda_b')$, and $D(\lambda_c, \lambda_c')$ respectively, as a function of the duration T. These figures display typical simulation results of the statistical reestimation technique. Important considerations behind the simulation process include: (1) characteristics of the generating source, such as λ being a balanced or skewed model; (2) effectiveness of the estimation technique; and (3) the number of observations needed for a good estimate. Here we provide qualitative discussions of the plotted results.

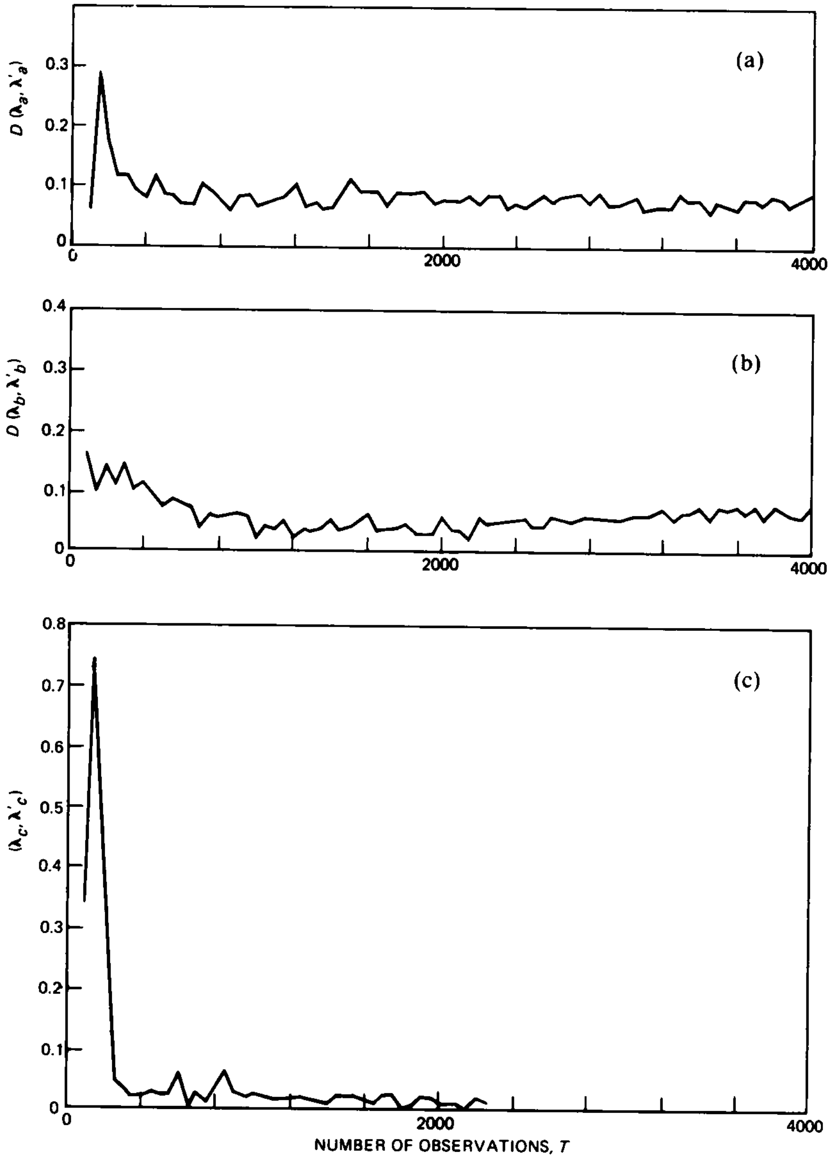

Fig. 3—Distance performance of reestimated ergodic models as a function of the observation duration: (a) λ_a, balanced model; (b) λ_b, skewed model; (c) λ_c, model with deterministic observation.

Figure 3a indicates that the distance between λ_a and λ_a' stabilizes after T grows beyond about five hundred samples. The distance for $T > 500$ is small (about 0.085), with a range of statistical variation between ±0.025. The distance scores of Fig. 3b do not seem to be as well behaved as those of Fig. 3a. Although the estimate λ_b' for λ_b may be as good as λ_a', judging from the distance, λ_b' appears to be more data dependent. A slow drifting from $D \simeq 0.04$ at $T = 1000 \sim 2000$ region to $D \simeq 0.07$ at $T = 3000 \sim 4000$ region is seen. This can be attributed to the fact that λ_b is a skewed model and the associated measure $\mu(\cdot \mid \lambda_b)$ has a slightly wider dynamic range than $\mu(\cdot \mid \lambda_a)$; hence deviations in D are manifested over a broad range of values of T. Those generated sequences, $\mathbf{O}_T$, of high $\mu(\mathbf{O}_T \mid \lambda_b)$ will result in a close estimate λ_b', and the wide dynamic range in $\mu(\cdot \mid \lambda_b)$ will directly translate into the observed variations in $D(\lambda_b, \lambda_b')$ for long observation sequences. This long-term drifting of D is reminiscent of the residual difference between uncorrelated and highly correlated sources in statistical data analysis.

The results of Fig. 3c indicate that when the generating source involves only a Markov chain and does not have variations in the observation density, very good estimates can be obtained with a small amount of data. Also, the **B** matrix, because it is an identity matrix, greatly narrows the range of $\mu(\cdot \mid \lambda_c)$, resulting in negligible variations in $D(\lambda_c, \lambda_c')$ when $\mathbf{O}_T$ is sufficiently long.

3.3.2 Left-to-right models

Another series of simulations dealt with nonergodic models of the types shown in Fig. 4. These models are identical to the three models SRC195, SRC295, and SRC395 studied in Ref. 1. We denote these models by λ_{195}, λ_{295}, and λ_{395} as in Fig. 4. For these models, $N = 5$, $M = 9$, $\mathbf{u}^* = [1\ 0\ 0\ 0\ 0]$, and

$$\mathbf{B} = \begin{bmatrix} 0.7 & 0.3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.8 & 0.2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.2 & 0.8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.3 & 0.7 \end{bmatrix}.$$

An additional model, λ_{595}, which had the same state transition probability and initial state probability as λ_{295} but with the following **B** matrix

$$\mathbf{B} = \begin{bmatrix} 0.8 & 0.1 & 0.1 & 0 & 0 & 0 & 0 \\ 0 & 0.1 & 0.8 & 0.1 & 0 & 0 & 0 \\ 0 & 0 & 0.1 & 0.8 & 0.1 & 0 & 0 \\ 0 & 0 & 0 & 0.1 & 0.8 & 0.1 & 0 \\ 0 & 0 & 0 & 0 & 0.1 & 0.8 & 0.1 \end{bmatrix},$$

was also studied.

The observation matrix **B** for λ_{195} through λ_{395} is *non-overlapping*; observations generated during one state cannot appear during another state. As was the case for model λ_c in the previous section, these

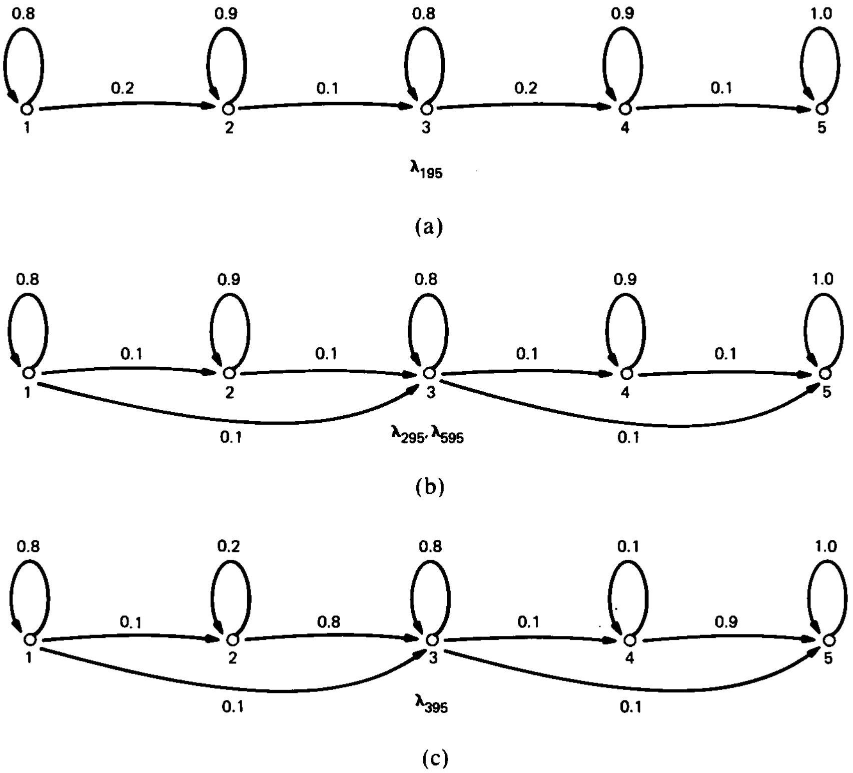

Fig. 4—Left-to-right hidden Markov models: (a) λ_{195}, (b) λ_{295} and λ_{595}, and (c) λ_{395}, used in the study of model parameter estimation sensitivity.

models show a rigid correspondence between states and observations. In this case, we can observe some particular effects of the coupling between matrices **A** and **B** upon model estimates as well as the distance.

The same simulation procedure as above was followed: (1) observation sequences were first generated; (2) model estimates were then obtained using the reestimation algorithm with different initial guesses; and (3) distances between the generating model and the estimated model were calculated and plotted as a function of T, the *total* sequence duration. (Note that because of the trap state in these models, the measurement sequence is a series of restarted sequences and T is the total duration.) Four kinds of initial guesses of model parameters were used. Type 1 is a totally random guess (except for the necessary stochastic constraints). Type 2 is a random guess with known state-transition constraints; that is, elements in **A** corresponding to prohibited transitions are initially set to null, while others are randomly chosen with stochastic constraints. Type 3 is a random guess

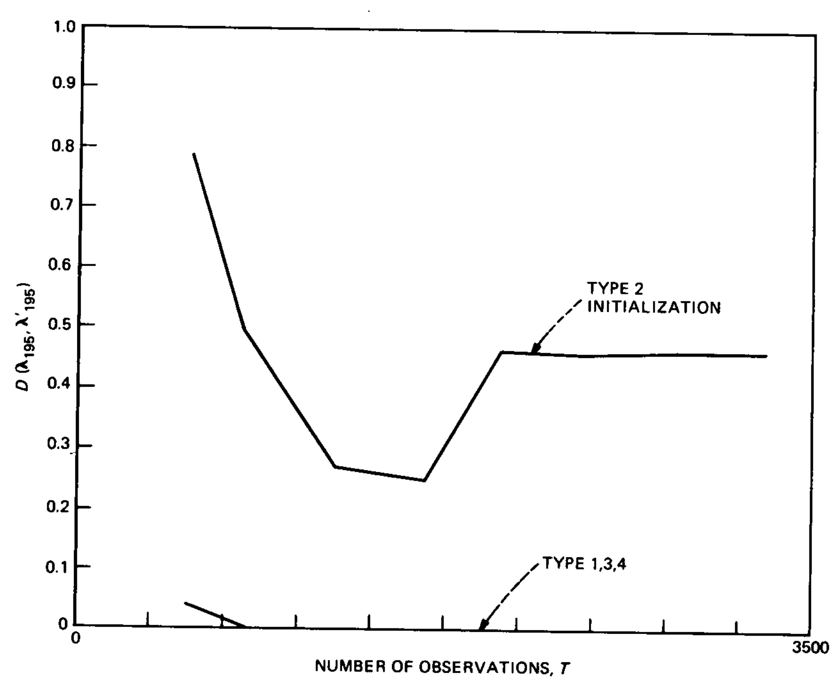

Fig. 5—Computed model distance versus number of observations, for the four types of parameter initial guesses, for model λ_{195}.

with both known state-transition constraints and known state-observation constraints, so in the initial matrices **A** and **B**, those elements corresponding to prohibited transitions and impossible observations are set to null. Type 4 is the generating model itself. Type 4 is useful for studying the convergence properties of the reestimate algorithm itself, since the sequence is unlikely to display complications often observed in sequences that converge to a local optimum.

A set of curves showing the measured distances versus the number of observations, for the four types of initial guess, are plotted in Figs. 5, 6, 7, and 8, corresponding to λ_{195}, λ_{295}, λ_{395}, and λ_{595}, respectively. Figure 5 shows that for model λ_{195}, the model estimates with Type 1, 3, and 4 initial guesses quickly converge to the generating model λ_{195}, i.e., the distances became essentially 0. Note that λ_{195} has a highly constrained structure with high probability of staying in the current state. This, combined with the fact that **B** is non-overlapping, says that this source would most probably produce observation sequences in which the corresponding state sequence is well defined, and the duration of each state (as determined from the **A** matrix) is unlikely to differ from one another dramatically. Type 2 initial guesses maintain the same Markov chain structure, but with random transition

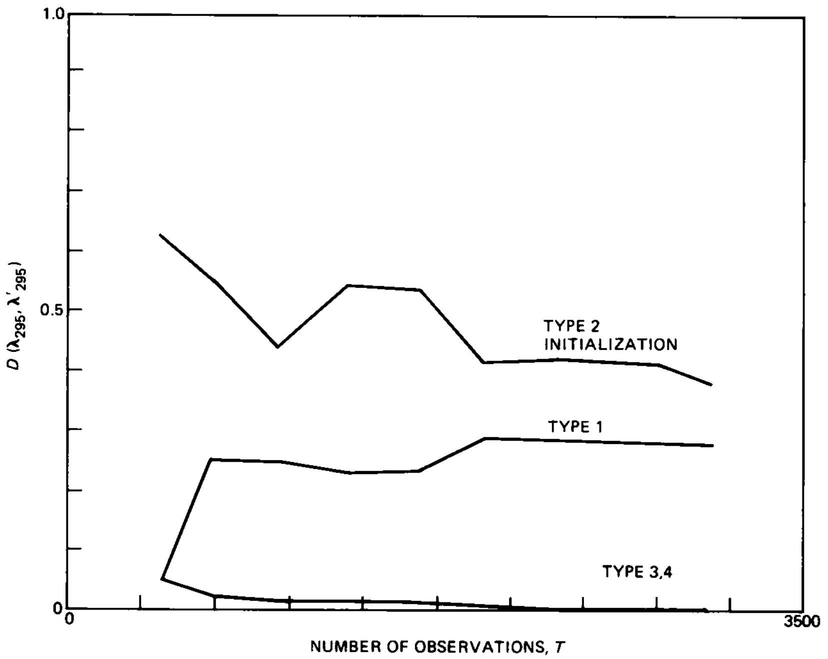

Fig. 6—Computed model distance versus number of observations, for the four types of parameter initial guesses, for model λ_{295}.

probabilities different from the generating source. In fitting the observation sequence to a Type 2 initial estimate, the constrained **A** matrix is changed very little in the reestimation process, which instead mainly extends and modifies the **B** matrix to include, in one state, different observations that originally occurred in different states. Depending on the initial guess values, the resultant **B** matrix may be significantly overlapped, thereby leading to a significant distance from the generating source. Figure 5 shows that this analysis is indeed the case for model λ_{195}. With Type 3 initial guesses, the initial constraints in matrices **A** and **B** are retained through the reestimation process, and optimization of the **A** matrix is independent of that of the **B** matrix. Furthermore, optimization of the **B** matrix, in the current case, is carried out independently for each state, because with the initial constrained **B** matrix, the underlying state sequence is immediately known. Therefore, the results from using Type 4 initial guesses are virtually identical to the results from using Type 3 initial guesses, as shown in Figs. 5 through 8, for all models that were studied.

For λ_{295}, trends similar to those of Fig. 5 are observed and shown in Fig. 6, but some problems due to the allowed state-skipping transitions are observed for Type 1 initial parameter estimates. As explained above, estimated models based upon Type 2 initial guesses are at a

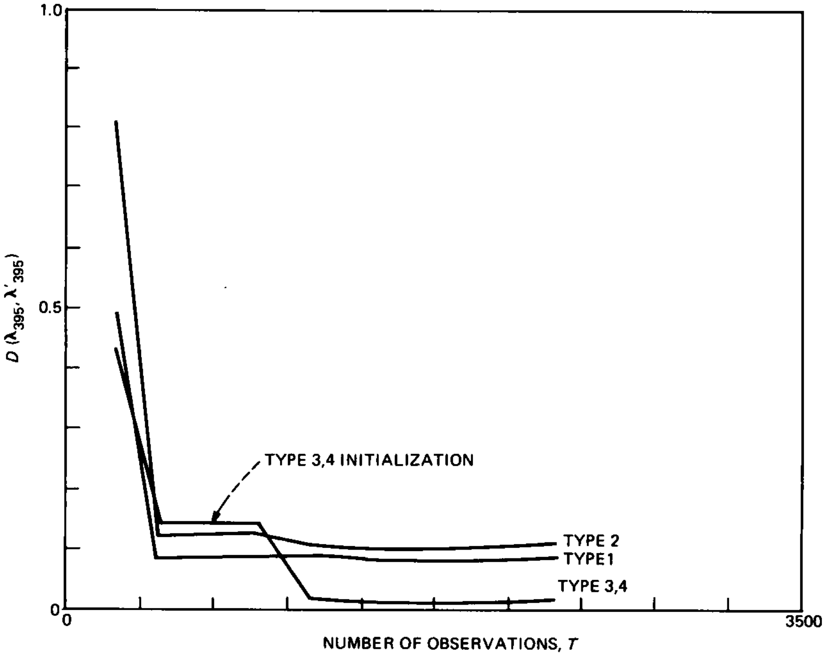

Fig. 7—Computed model distance versus number of observations, for the four types of parameter initial guesses, for model λ_{395}.

significant distance from the generating source. With Type 1 initial guesses, the converged results are only slightly better than those with Type 2 guesses. Type 3 and 4 initial guesses lead to virtually the optimal estimate, i.e., models with essentially zero distance from the generator.

The results using model λ_{395} are shown in Fig. 7. Model λ_{395} has a state-transition structure similar to that of λ_{295}, but with different transition probabilities. The fact that $a_{22} = 0.2$ and $a_{44} = 0.1$ in λ_{395} makes it essentially a three-state model (i.e., two of the states are highly transient). Again, the dependence of model estimates upon the type of initial guess is similar to what is mentioned above, except the distances now are smaller than those obtained using λ_{295}. However, as seen in Fig. 7, when the total duration is small (i.e., 244 samples), the estimated models are at a significant distance from the generating source, regardless of the initial guess. This is because states 2 and 4 are not well represented in the observation sequences. The sudden drop of distance for Type 3 and 4 initial guesses at $T \simeq 1150$ samples indicates that the transient states of the generating model are sufficiently well represented for $T \geq 1150$, and with proper initial guesses, an estimate virtually identical to the generating source can be obtained.

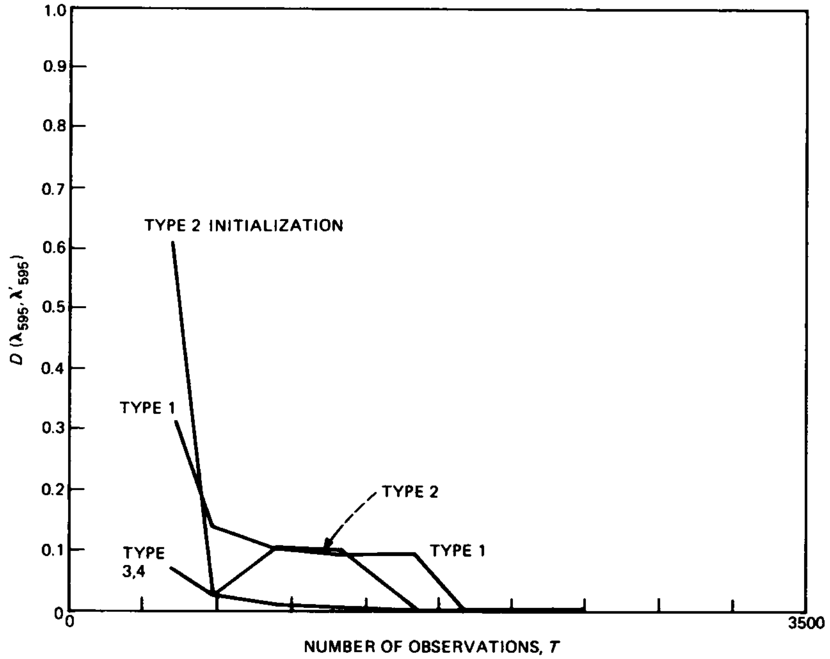

Fig. 8—Computed model distance versus number of observations, for the four types of parameter initial guesses, for model λ_{595}.

The results for model λ_{595} are given in Fig. 8. Since source λ_{595} has an overlapping observation matrix **B**, many of the phenomena that occurred in λ_{195}, λ_{295}, and λ_{395} no longer appear. Indeed, as shown in Fig. 8, estimated models of nearly zero distance from the generating source have been obtained regardless of the initial guess, provided the observation sequences are sufficient in duration. The effects of initial guess are manifested only in the way the estimate converges as T grows.

Figures 5, 6, 7, and 8 not only provide results pertaining to the performance of model estimates and its relationship to model initialization as well as observation length, but also show the effectiveness of the distance measure of eq. (6) in measuring the dissimilarity between any pair of hidden Markov models.

IV. CONCLUSION

We have defined a probabilistic distance measure for hidden Markov models. The measure is consistent with the probabilistic modeling technique and can be efficiently evaluated through Monte Carlo procedures. The distance measure was employed in the study of relative parameter sensitivities as well as the relationship among model estimate, initial guess for the reestimation algorithm, and the observation sequence duration for discrete density hidden Markov models. Much of the behavior of hidden Markov models and the reestimated results have been observed through the use of such a distance measure. The study in turn confirms the effectiveness and reliability of the distance measure. Potential applications of the distance measure may include hidden Markov model selection as well as clustering.

REFERENCES

1. S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition," B.S.T.J., *62*, No. 4 (April 1983), pp. 1035–74.
2. L. E. Baum et al., "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," Ann. Math. Statist., *41* (1970), pp. 164–71.
3. J. H. Conway and N. J. A. Sloane, "Voronoi Regions of Lattices, Second Moments of Polytypes, and Quantization," IEEE Trans. Inform. Theory, *IT-28* (March 1982), pp. 227–32.
4. Y. Linde, A. Buzo, and R. M. Gray, "An Algorithm for Vector Quantizer Design," IEEE Trans. Commun., *COM-28* (January 1980), pp. 84–95.
5. S. Kullback, *Information Theory and Statistics*, New York: Wiley, 1958.
6. R. M. Gray et al., "Rate-Distortion Speech Coding With a Minimum Discrimination Information Distortion Measure," IEEE Trans. Inform. Theory, *IT-27*, No. 6 (November 1981), pp. 708–21.
7. T. Petrie, "Probabilistic Functions of Finite State Markov Chains," Ann. Math. Statist., *40*, No. 1 (1969), pp. 97–115.
8. L. R. Rabiner et al., unpublished work.
9. L. E. Baum and J. A. Eagon, "An Inequality With Applications to Statistical Estimation for Probabilistic Functions of a Markov Process and to a Model for Ecology," Bull. AMS, *73* (1967), pp. 360–3.
10. P. Billingsley, *Ergodic Theory and Information*, New York: Wiley, 1965.

AUTHORS

Biing-Hwang Juang, B.Sc. (Electrical Engineering), 1973, National Taiwan University, Republic of China; M.Sc. and Ph.D. (Electrical and Computer Engineering), University of California, Santa Barbara, 1979 and 1981, respectively; Speech Communications Research Laboratory (SCRL), 1978; Signal Technology, Inc., 1979–1982; AT&T Bell Laboratories, 1982; AT&T Information Systems Laboratories, 1983; AT&T Bell Laboratories, 1983—. Before joining AT&T Bell Laboratories, Mr. Juang worked on vocal tract modeling at Speech Communications Research Laboratory, and on speech coding and interference suppression at Signal Technology, Inc. Presently, he is a member of the Acoustics Research Department, where he is researching speech communications techniques and stochastic modeling of speech signals.

Lawrence R. Rabiner, S.B. and S.M., 1964, Ph.D., 1967 (Electrical Engineering), The Massachusetts Institute of Technology; AT&T Bell Laboratories, 1962—. Presently, Mr. Rabiner is engaged in research on speech communications and digital signal processing techniques. He is coauthor of *Theory and Application of Digital Signal Processing* (Prentice-Hall, 1975), *Digital Processing of Speech Signals* (Prentice-Hall, 1978), and *Multirate Digital Signal Processing* (Prentice-Hall, 1983). Member, National Academy of Engineering, Eta Kappa Nu, Sigma Xi, Tau Beta Pi. Fellow, Acoustical Society of America, IEEE.