AT&T Bell Laboratories Technical Journal
Vol. 63, No. 7, September 1984
Printed in U.S.A.

On the Hidden Markov Model and Dynamic Time Warping for Speech Recognition—A Unified View

By B.-H. JUANG*

(Manuscript received January 17, 1984)

This paper gives a unified theoretical view of the Dynamic Time Warping (DTW) and the Hidden Markov Model (HMM) techniques for speech recognition problems. The application of hidden Markov models in speech recognition is discussed. We show that the conventional dynamic time-warping algorithm with Linear Predictive (LP) signal modeling and distortion measurements can be formulated in a strictly statistical framework. It is further shown that the DTW/LP method is implicitly associated with a specific class of Markov models and is equivalent to the probability maximization procedures for Gaussian autoregressive multivariate probabilistic functions of the underlying Markov model. This unified view offers insights into the effectiveness of the probabilistic models in speech recognition applications.

I. INTRODUCTION

Research in speech recognition has produced numerous algorithms and commercially available speech recognizers that all work to some extent.[1] Among these, temporal alignment techniques such as Dynamic Time-Warping (DTW) algorithms[2–4] and Markov modeling[5–7] are two prevailing approaches that are practical and theoretically sound. Both techniques emphatically address nonstationarity in speech signals. The two techniques, however, operate in different manners, as we will discuss briefly.

* AT&T Bell Laboratories.

In a recognition system employing dynamic time warping, a warping procedure based upon a prechosen, well-defined distortion measure aligns the unknown test speech sequence in turn to each reference sequence. The distortion measure must be a meaningful metric of dissimilarity between sound representations, usually the short-time spectra. The objective is to find a reference sequence of a known category or word that has the least dissimilarity to the test sequence after being optimally time aligned. Time alignment involves (time-) warping functions that are dynamic but deterministic representations of the possible variation of sound durations that are evident between the reference and the test sequences. There is, hence, one warping function that best matches the reference and test sequences, resulting in the smallest dissimilarity. The smallest dissimilarity measurement among all categories determines the recognition (recognition by minimum distortion).

Markov modeling techniques, while they may still perform sound pattern comparisons, do not require explicit time alignment. Instead, a probabilistic transition and observation structure is defined for each reference category or word. Such a structure, called a Markov model, includes (1) a state transition probability matrix, (2) an initial probability vector, and (3) an observation probability matrix for discrete probability densities or a set of continuous densities defined by parameter sets, or a mixture of the two when different types of densities are used. During recognition, one computes for each given reference model the probability of observing the test sequence. The model that produces the maximum observation probability is the classification result (recognition by maximum probability).

These two techniques are similar in theory, despite their vastly different operations and results. Confusion from this similarity often renders comparative studies of the two techniques difficult and futile. The purpose of this paper is, then, to give a unified tutorial view of the two techniques and to establish a theoretical link between them such that more fruitful and meaningful comparison can be made and each technique will improve the other technique.

The paper is organized as follows. We first present statistical characteristics of Gaussian autoregressive sources in Section II, which serves as a foundation for the later developments. This topic is well studied in multivariate analysis, and an excellent treatment of it in the context of speech processing can be found in Ref. 8. In Section III we discuss maximum likelihood estimation of Gaussian autoregressive source parameters, and we explicitly demonstrate the relationship between some well-known probability density functions and distortion measures related to linear prediction (LP). In Section IV we discuss some fundamentals of probabilistic functions of Markov chains and their applications in speech recognition. We again show that distortion measures can be cast in the framework of probabilistic functions of Markov chains. We finally discuss dynamic time warping in Section V and show that dynamic time warping employing LP-related measures is equivalent to the recognition by maximum probability procedure, with some specific constraints. Theoretical similarities and differences between the two techniques are then discussed in detail to complete the attempted unified view. We start with Gaussian autoregressive source because it is one of the best known sources and is useful in speech research. More general measure-theoretic steps could have been taken to establish a formal theoretical link between the two methods. It is, nevertheless, our opinion that the present framework of Gaussian autoregressive sources adequately gives a meaningful unified view of the two methods.

II. GAUSSIAN AUTOREGRESSIVE SOURCE

Consider a stationary, zero-mean, Gaussian signal source. The output of the source, subject to observation, is an N-sampled sequence $\{\mathbf{s}_1, \mathbf{s}_2, \ldots \mathbf{s}_N\}$, where each $\mathbf{s}_i$ is a real random variable. The vector notation $\mathbf{s}^t = [\mathbf{s}_1\ \mathbf{s}_2\ \ldots\ \mathbf{s}_N] \in R^N$ denotes the observation. The probability density function of the random vector $\mathbf{s}$ for *known* autocorrelation matrix $\mathbf{C}_N$ is thus

$$f(s \mid \mathbf{C}_N) = \lim_{\substack{\Delta si \to 0 \\ i=1,2,\ldots,N}} \frac{\begin{array}{c}\mathrm{P_r}\{s_1 \leqslant \mathbf{s}_1 \leqslant s_1 + \Delta s_1,\ s_2 \leqslant \mathbf{s}_2 \leqslant s_2 \\ + \Delta s_2, \ldots, s_N \leqslant \mathbf{s}_N \leqslant s_N + \Delta s_N \mid \mathbf{C}_N\}\end{array}}{\Delta s_1 \Delta s_2 \ldots \Delta s_N}$$

$$= (2\pi)^{-N/2} |\mathbf{C}_N|^{-1/2} \exp\left\{-\frac{1}{2} s^t \mathbf{C}_N^{-1} s\right\}, \tag{1}$$

where $s^t = [s_1\ s_2 \ldots s_N]$ is a realization of $\mathbf{s}^t$, $\mathbf{C}_N = [r_{ij}]_{i,j=1}^N$, and $r_{ij} = E\{\mathbf{s}_i\mathbf{s}_j\} = r_{|i-j|}$ due to stationarity.

The source is assumed to be Mth-order autoregressive with coefficients $\mathbf{a}^t = [a_0\ a_1 \ldots a_M]$, where a_0 is always unity. Hence, as shown in Fig. 1, the source can be equivalently viewed as a white Gaussian noise source with unity variance, followed by an all-pole filter $1/A(z)$, where

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \ldots a_M z^{-M},$$

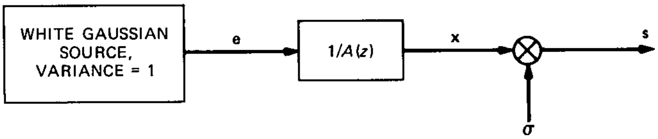


Fig. 1—Gaussian autoregressive source.

and an amplifier with multiplication factor σ. If we denote the output of such a white Gaussian noise source as $\mathbf{e}_i$, $i = \ldots, -1, 0, 1, \ldots$, then the output of the filter and the overall signal source is

$$\mathbf{x}_i = -\sum_{j=1}^{M} a_j \mathbf{x}_{i-j} + \mathbf{e}_i \quad \text{for any} \quad i \tag{2a}$$

and

$$\mathbf{s}_i = \sigma \mathbf{x}_i, \tag{2b}$$

respectively.

For our interests in source identification, we wish to express the density function of (1) in terms of the source parameters σ^2 and $\mathbf{a}$, that is, $f(s \mid \mathbf{a}, \sigma^2)$, or equally importantly, we would like to obtain the gain-independent probability density $f(x \mid \mathbf{a})$. The difficulty here is that the relationship of (2) is not defined for the first M samples since we have only finite sample observation starting from $\mathbf{s}_1$. However, using a classical (Gram-Schmidt) orthogonalization procedure for the first M samples, we may rewrite (2) as

$$\begin{aligned}
&h_{11}\mathbf{s}_1 = \sigma\epsilon_1 \\
&h_{21}\mathbf{s}_1 + h_{22}\mathbf{s}_2 = \sigma\epsilon_2 \\
&\vdots \\
&h_{M1}\mathbf{s}_1 + h_{M2}\mathbf{s}_2 + \ldots h_{MM}\mathbf{s}_M = \sigma\epsilon_M \\
&a_M\mathbf{s}_1 + a_{M-1}\mathbf{s}_2 + \ldots a_1\mathbf{s}_M + \mathbf{s}_{M+1} = \sigma\mathbf{e}_{M+1} \\
&\quad\ddots \\
&\quad a_M\mathbf{s}_{N-M} + a_{M-1}\mathbf{s}_{N-M+1} + \ldots a_1\mathbf{s}_{N-1} + \mathbf{s}_N = \sigma\mathbf{e}_N.
\end{aligned}$$

Denoting $\mathbf{e}^t = [\epsilon_1, \epsilon_2 \ldots, \epsilon_M, \mathbf{e}_{M+1}, \ldots, \mathbf{e}_N]$, we have a system equation for the N observation samples:

$$\sigma\mathbf{e} = \mathbf{H}\mathbf{s} \tag{3a}$$

$$\mathbf{e} = \mathbf{H}\mathbf{x}, \tag{3b}$$

where

$$\mathbf{H} = \left[\begin{array}{cccc|ccc}
h_{11} & & & & & & \\
h_{21} & h_{22} & & & & 0 & \\
\vdots & & & & & & \\
h_{M1} & h_{M2} \ldots & & h_{MM} & & & \\
\hline
a_M & a_{M-1} & \ldots & a_1 & 1 & & \\
 & & & & & & \\
 & & & \cdots & & & \\
0 & a_M & & a_{M-1} & \ldots & a_1 & 1
\end{array}\right]$$

$$= \left[\begin{array}{c|c} \mathbf{H}_1 & 0 \\ \hline \underbrace{\mathbf{H}_2}_{M} & \underbrace{\mathbf{H}_3}_{N-M} \end{array}\right] \begin{array}{l} \}M \\ \}N-M. \end{array}$$

The elements of $\mathbf{e}$ are uncorrelated, and h_{ij} are properly scaled so that $E[\mathbf{e}_i\mathbf{e}_j] = E[\epsilon_i\mathbf{e}_j] = E[\epsilon_i\epsilon_j] = \delta_{ij}$ for any appropriate i and j, thereby giving

$$\begin{aligned} \mathbf{I}_N &= E\{\mathbf{e}\mathbf{e}^t\} \\ &= (\sigma^2)^{-1}\mathbf{H}E\{\mathbf{s}\mathbf{s}^t\}\mathbf{H}^t \\ &= (\sigma^2)^{-1}\mathbf{H}\mathbf{C}_N\mathbf{H}^t, \end{aligned} \tag{4}$$

where $\mathbf{I}_N$ is the $N \times N$ identity matrix. Equation (4) leads to

$$\mathbf{C}_N^{-1} = (\sigma^2)^{-1}\ \mathbf{H}^t\mathbf{H} \tag{5}$$

and

$$|\mathbf{C}_N| = |\mathbf{C}_N^{-1}|^{-1} = (\sigma^2)^N\ |\mathbf{H}|^{-2}. \tag{6}$$

But, since $|\mathbf{H}_3| = 1$ and $|\mathbf{H}| = |\mathbf{H}_1|\cdot|\mathbf{H}_3|$,

$$|\mathbf{C}_N| = (\sigma^2)^N|\mathbf{H}_1|^{-2}.$$

Note that matrix $\mathbf{H}_1$ also corresponds to the diagonalization of the $M \times M$ autocorrelation matrix $\mathbf{C}_M = [r_{ij}]$ for $i, j = 1, 2, \ldots, M$; i.e.,

$$\mathbf{I}_M = (\sigma^2)^{-1}\ \mathbf{H}_1\mathbf{C}_M\mathbf{H}_1^t,$$

where $\mathbf{I}_M$ is the $M \times M$ identity matrix. Therefore,

$$|\mathbf{C}_M| = (\sigma^2)^M\ |\mathbf{H}_1|^{-2}$$

and

$$|\mathbf{C}_N| = (\sigma^2)^{N-M}\ |\mathbf{C}_M|. \tag{7}$$

Given an autocorrelation matrix $\mathbf{C}_M$, $|\mathbf{C}_M|$ can be easily obtained by first diagonalizing $\mathbf{C}_M$ using Cholesky decomposition or, more efficiently, Levinson's recursion algorithm,[9]

$$\mathbf{B}^t\mathbf{C}_M\mathbf{B} = \beta = \begin{bmatrix} \beta_0 & & & 0 \\ & \beta_1 & & \\ & & \ddots & \\ 0 & & & \beta_{M-1} \end{bmatrix},$$

where $\mathbf{B}$ is an upper triangular matrix, the diagonal elements of which are all unity. Therefore,

$$|\mathbf{C}_M| = \prod_{i=0}^{M-1} \beta_i,$$

and as a result,

$$|\mathbf{C}_N| = (\sigma^2)^{N-M}\left(\prod_{i=0}^{M-1} \beta_i\right). \qquad (8)$$

Note that β_i is equivalent to the minimum mean-square error resulting from an ith-order linear prediction of the signal. The probability density function of $\mathbf{s}$ becomes

$$f(s\,|\,\mathbf{C}_N) = (2\pi)^{-N/2}(\sigma^2)^{-(N-M)/2}\left(\prod_{i=0}^{M-1} \beta_i\right)^{-1/2} \exp\{-s^t\mathbf{H}^t\mathbf{H}s/2\sigma^2\}. \qquad (9)$$

For the gain-independent expression,

$$f(x\,|\,\mathbf{C}_N) = (2\pi)^{-N/2}\left(\prod_{i=0}^{M-1} \frac{\beta_i}{\sigma^2}\right)^{-1/2} \exp\left\{-\frac{1}{2}x^t\mathbf{H}^t\mathbf{H}x\right\}. \qquad (10)$$

We further write $x^t\mathbf{H}^t\mathbf{H}x$ explicitly as

$$x^t\mathbf{H}^t\mathbf{H}x = \left(\sum_{j=0}^{M} a_j^2\right)\left(\sum_{i=1}^{N} x_i^2\right) + 2\left(\sum_{j=0}^{M-1} a_j a_{j+1}\right)\left(\sum_{i=1}^{N-1} x_i x_{i+1}\right) + \dots$$
$$+ 2(a_0 a_M)\left(\sum_{i=1}^{N-M} x_i x_{i+M}\right) - Q,$$

where Q represents negligible terms compared to others for $N \gg M$. Letting

$$r_a(i) \triangleq \sum_{j=1}^{M-i} a_j a_{j+i}, \qquad (11)$$

$$r_x(i) \triangleq \sum_{j=1}^{N-i} x_j x_{j+i} = \frac{1}{\sigma^2}\sum_{j=1}^{N-i} s_j s_{j+i} = \frac{r_s(i)}{\sigma^2}, \qquad (12)$$

and

$$\alpha(x;\,\mathbf{a}) = r_a(0)r_x(0) + 2\sum_{i=1}^{M} r_a(i)r_x(i), \qquad (13)$$

we then have an approximation for the density function,

$$f(s\,|\,\mathbf{C}_N) \cong (2\pi)^{-N/2}(\sigma^2)^{-(N-M)/2}\left(\prod_{i=0}^{M-1} \beta_i\right)^{-1/2} \exp\left\{-\frac{1}{2}\alpha(\sigma^{-1}s;\,\mathbf{a})\right\} \qquad (14)$$

or

$$f(x\,|\,\mathbf{C}_N) \cong (2\pi)^{-N/2}\left(\prod_{i=0}^{M-1} \frac{\beta_i}{\sigma^2}\right)^{-1/2} \exp\left\{-\frac{1}{2}\alpha(x;\,\mathbf{a})\right\}. \qquad (15)$$

This function can be evaluated easily if the source parameters are known.

III. MAXIMUM LIKELIHOOD

In many realistic situations, such as dealing with speech signals, the a priori information about the source is usually not available. What is involved in parameterization of speech signals for coding and recognition is mainly estimating and identifying the source parameters from *finite observations*. More specifically, the following two dominant problems arise almost ubiquitously in speech analysis research: (1) Estimation—Given an observation s, what is the best or the most probable set of source parameters that led to the observation? (2) Identification—Given two observations s_1 and s_2, how close are the two observations? Are they close enough to be considered identical? Or, what is the probability that s_2 has been produced by the same source as s_1? The first problem is certainly very much studied in statistical estimation theory as well as in (deterministic) least-squares time-series analysis. Research in identification, particularly in the field of speech processing, resulted in some distance or distortion measures.[10] Thus our main goal here is to integrate the formulation of the two problems in a probabilistic framework to better understand the probabilistic modeling techniques. In the following presentation, we shall focus on the maximum likelihood estimate of Gaussian autoregressive source parameters, as initiated in the previous section.

3.1 Estimation—Autocorrelation method

We discuss here only the autocorrelation method. For other varieties, we suggest that readers consult Refs. 9 and 11.

The observation sequence $s = \{s_1, s_2, \ldots, s_N\}$ is assumed to be very long, i.e., $N \gg M$. It is further argued that observation of the source output, which is infinitely long, is made through some "smooth window," so that the edge problem at the beginning of the observed sequence is avoided, maintaining that

$$\mathbf{e}_i = \sum_{j=0}^{M} a_j \mathbf{x}_{i-j} \quad \text{and} \quad \mathbf{s}_i = \sigma \mathbf{x}_i \quad \text{for all} \quad i, \tag{16}$$

with $a_0 = 1$ and $\mathbf{x}_i = 0$ for $i \leq 0$ and for $i > N$. Hence, the diagonal elements in $\mathbf{H}$ matrix are assumed to be all unity. Equation (6) then becomes

$$|\mathbf{C}_N| = (\sigma^2)^N,$$

and the probability density function, as in eq. (14), is now expressed as

$$f(s \mid \mathbf{C}_N) \simeq (2\pi)^{-N/2}(\sigma^2)^{-N/2}\exp\left\{-\frac{1}{2}\,\alpha(\sigma^{-1}s;\, \mathbf{a})\right\}. \tag{17}$$

Since (17) is a function of $\mathbf{a}$ and σ^2 and it closely approximates $f(s \mid \mathbf{C}_N)$, we shall, in the following, define $f(s \mid \mathbf{a}, \sigma^2)$ as (17). Furthermore, the gain-independent density function is thus

$$f(x \mid \mathbf{a}) = (2\pi)^{-N/2}\exp\left\{-\frac{1}{2}\,\alpha(x;\, \mathbf{a})\right\}. \tag{18}$$

It defines the probability density function of observing a vector x at the output of an all-pole filter $1/A(z)$ driven by a *unity variance* Gaussian i.i.d. sequence.

It is clear that, given an observation $s^{(0)}$, the maximum likelihood estimate of $\mathbf{a}$ is the one that maximizes $f(s^{(0)} \mid \mathbf{a}, \sigma^2)$ or, equivalently, minimizes $\alpha(s^{(0)};\, \mathbf{a})$ because σ here is only a scaling factor. In linear prediction terminology, the optimal $\mathbf{a}^{(0)}$ is obtained by minimizing the prediction-error energy $\alpha(s^{(0)};\, \mathbf{a})$, defined by (13), and $\alpha(s^{(0)};\, \mathbf{a}^{(0)}) = \min_{\mathbf{a}} \alpha(s^{(0)};\, \mathbf{a})$ is called the minimum residual energy.[9] Furthermore, to maximize $f(s^{(0)} \mid \mathbf{a}^{(0)}, \sigma^2)$ with respect to σ^2, the optimal estimate $\sigma^2_{(0)}$, is easily found to be

$$\sigma^2_{(0)} = \alpha(s^{(0)};\, \mathbf{a}^{(0)})/N, \tag{19}$$

which leads to

$$\alpha(\sigma^{-1}_{(0)}s^{(0)};\, \mathbf{a}^{(0)}) = N. \tag{20}$$

This can be verified intuitively by recognizing that there are N valid uncorrelated error samples, $\sigma_{(0)}\mathbf{e}_i = \sum_{j=0}^{M} a_j^{(0)}\mathbf{s}_{i-j}^{(0)}$, $i = 1, 2, \ldots, N$, each having variance $\sigma^2_{(0)}$, resulting in an energy of $\alpha(s^{(0)};\, \mathbf{a}^{(0)}) = N\,\sigma^2_{(0)}$.

3.2 Identification

In this section we establish the relationship between probability and distortion measures for identification purposes. In particular, for the present consideration of Gaussian autoregressive sources, we discuss the role of Itakura-Saito measure and the likelihood ratio measure[12] in probability density functions.

3.2.1 Itakura-Saito measure

In maximum likelihood estimation, we often maximize the log likelihood, $\log\{f(s^{(0)} \mid \mathbf{a}, \sigma^2)\}$, instead of $f(s^{(0)} \mid \mathbf{a}, \sigma^2)$ for convenience, particularly when the probability density function is jointly Gaussian, as in the present case. The log likelihood takes the form

$$\log\{f(s \mid \mathbf{a}, \sigma^2)\} = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2}\,\alpha(\sigma^{-1}s;\, \mathbf{a}). \tag{21}$$

Since $\mathbf{a}^{(0)}$ and $\sigma_{(0)}^2$ are the maximum likelihood estimate based upon $s^{(0)}$,

$$\begin{aligned}\log\{f(s^{(0)} \mid \mathbf{a}^{(0)}, \sigma_{(0)}^2)\} &= -\frac{N}{2}\log(2\pi\sigma_{(0)}^2) - \frac{1}{2}\,\alpha(\sigma_{(0)}^{-1}s^{(0)};\, \mathbf{a}^{(0)}) \\ &= -\frac{N}{2}\log(2\pi\sigma_{(0)}^2) - \frac{N}{2} \\ &= \{\log f(s^{(0)} \mid \mathbf{a}, \sigma^2)\}_{\max}. \end{aligned} \tag{22}$$

The log likelihood difference between the maximum and other arbitrary values is thus

$$\begin{aligned} L_d &= \{\log f(s^{(0)} \mid \mathbf{a}, \sigma^2)\}_{\max} - \log f(s^{(0)} \mid \mathbf{a}, \sigma^2) \\ &= \log f(s^{(0)} \mid \mathbf{a}^{(0)}, \sigma_{(0)}^2) - \log f(s^{(0)} \mid \mathbf{a}, \sigma^2) \\ &= -\frac{N}{2}\log(2\pi\sigma_{(0)}^2) - \frac{N}{2} + \frac{N}{2}\log(2\pi\sigma^2) + \frac{1}{2}\,\alpha(\sigma^{-1}s^{(0)};\, \mathbf{a}) \\ &= \frac{N}{2}\left[\frac{1}{N}\,\alpha(\sigma^{-1}s^{(0)};\, \mathbf{a}) + \log\sigma^2 - \log\sigma_{(0)}^2 - 1\right] \end{aligned} \tag{23}$$

because

$$f(s^{(0)} \mid \mathbf{a}, \sigma^2) = (2\pi\sigma^2)^{-N/2}\exp\left\{-\frac{1}{2}\,\alpha(\sigma^{-1}s^{(0)};\, \mathbf{a})\right\}. \tag{24}$$

The bracketed term in (23) is, in fact, the well-known Itakura-Saito distortion measure[10] between $\{\mathbf{a}^{(0)}, \sigma_{(0)}^2\}$, representing $s^{(0)}$, and $\{\mathbf{a}, \sigma^2\}$, representing another observation s; i.e.,

$$\begin{aligned} d_{IS}(s^{(0)};\, s) = d_{IS}(s^{(0)};\, \{\mathbf{a}, \sigma^2\}) &= \frac{1}{N}\,\alpha(\sigma^{-1}s^{(0)};\, \mathbf{a}) \\ &\quad + \log\sigma^2 - \log\sigma_{(0)}^2 - 1 \\ &= 2L_d/N. \end{aligned} \tag{25}$$

Therefore, the probability of observing $s^{(0)}$ at the output of a source with parameters $\{\mathbf{a}, \sigma^2\}$ is, in terms of the distortion measure,

$$\begin{aligned} f(s^{(0)} \mid \mathbf{a}, \sigma^2) &= (2\pi\sigma^2)^{-N/2}\exp\left\{-\frac{N}{2}\,[d_{IS}(s^{(0)};\, \{\mathbf{a}, \sigma^2\}) \right. \\ &\qquad \left. + \log\sigma_{(0)}^2 - \log\sigma^2 + 1]\right\} \\ &= G(\sigma^2, \sigma_{(0)}^2)\exp\left\{-\frac{N}{2}\,d_{IS}(s^{(0)};\, \{\mathbf{a}, \sigma^2\})\right\}, \end{aligned} \tag{26}$$

where

$$G(\sigma^2, \sigma^2_{(0)}) = (2\pi\sigma^2)^{-N/2}\exp\left\{-\frac{N}{2}[\log \sigma^2_{(0)} - \log \sigma^2 + 1]\right\}. \quad (27)$$

3.2.2 Likelihood ratio measure

In many situations, the desired identification is only based upon the autoregressive parameters $\mathbf{a}$. This is equivalent to comparing two gain-normalized observation sequences. Let $\{\mathbf{a}^{(0)}, \sigma^2_{(0)}\}$ and $\{\mathbf{a}, \sigma^2\}$ be the maximum likelihood estimates corresponding to observations $s^{(0)}$ and s, respectively. The gain-normalized observations are then $x^{(0)} = s^{(0)}/\sigma_{(0)}$ and $x = s/\sigma$. From (18) and (20), the maximum log likelihood for $x^{(0)}$ is

$$\begin{aligned}\{\log f(x^{(0)} \mid \mathbf{a})\}_{\max} &= \log f(x^{(0)} \mid \mathbf{a}^{(0)}) \\ &= -\frac{N}{2}\log(2\pi) - \frac{1}{2}\alpha(x^{(0)}; \mathbf{a}^{(0)}) \\ &= -\frac{N}{2}\log(2\pi) - \frac{N}{2}.\end{aligned}$$

But, since

$$f(x^{(0)} \mid \mathbf{a}) = (2\pi)^{-N/2}\exp\left\{-\frac{1}{2}\alpha(x^{(0)}; \mathbf{a})\right\}, \quad (28)$$

the log likelihood difference between the maximum and another arbitrary value is then

$$\begin{aligned}L_d &= \{\log f(x^{(0)} \mid \mathbf{a})\}_{\max} - \log f(x^{(0)} \mid \mathbf{a}) \\ &= -\frac{N}{2}\log(2\pi) - \frac{N}{2} + \frac{N}{2}\log(2\pi) + \frac{1}{2}\alpha(x^{(0)}; \mathbf{a}) \\ &= \frac{N}{2}\left[\frac{1}{N}\alpha(x^{(0)}; \mathbf{a}) - 1\right]. \quad (29)\end{aligned}$$

The above-bracketed term is the likelihood ratio measure widely employed in vector quantization vocoder designs; that is,

$$\begin{aligned}d_{LR}(s^{(0)}; s) &\triangleq d_{LR}(x^{(0)}; x) \\ &= d_{LR}(x^{(0)}; \mathbf{a}) \\ &= \frac{1}{N}\alpha(x^{(0)}; \mathbf{a}) - 1. \quad (30)\end{aligned}$$

The likelihood ratio measure and the Itakura-Saito measure are closely related,

$$\begin{aligned} d_{LR}(s^{(0)}; s) &= d_{LR}(x^{(0)}; x) \\ &= d_{LR}(x^{(0)}; \mathbf{a}) \\ &= d_{IS}(x^{(0)}; x) \\ &= d_{IS}(x^{(0)}; \mathbf{a}). \end{aligned}$$

As a result, the probability of observing $x^{(0)}$ at the output of an all-pole filter $1/A(z)$ driven by a unity variance Gaussian i.i.d. sequence is

$$\begin{aligned} f(x^{(0)} \mid \mathbf{a}) &= (2\pi)^{-N/2} \exp\left\{-\frac{N}{2}[d_{LR}(x^{(0)}; \mathbf{a}) + 1]\right\} \\ &= P \exp\left\{-\frac{N}{2} d_{LR}(x^{(0)}; \mathbf{a})\right\} \end{aligned} \tag{31}$$

where

$$P = (2\pi)^{-N/2} \exp\left\{-\frac{N}{2}\right\}. \tag{32}$$

Equations (26) and (31) are thus the fundamental link between probability and distortion measures.

3.3 Estimation and identification based upon multiple observations

We have presented parametric estimation of the probability density function based upon a single observation in the above. When several observations are available and known to be from the same source, the estimation turns out to be quite similar to the single-observation case.

Let $s^{(i)}$, $i = 1, 2, \ldots, L$, denote the available observations. These observations are considered to be i.i.d. with probability density

$$f(s^{(i)} \mid \mathbf{a}, \sigma^2) = (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2}\alpha(\sigma^{-1}s^{(i)}; \mathbf{a})\right\}.$$

The joint probability density of observations $s^{(1)}, s^{(2)}, \ldots, s^{(L)}$ is thus

$$\begin{aligned} &f(s^{(1)}, s^{(2)}, \ldots, s^{(L)} \mid \mathbf{a}, \sigma^2) \\ &= \prod_{i=1}^{L} f(s^{(i)} \mid \mathbf{a}, \sigma^2) \\ &= [(2\pi\sigma^2)^{-N/2}]^L \exp\left\{-\frac{1}{2}\sum_{i=1}^{L} \alpha(\sigma^{-1}s^{(i)}; \mathbf{a})\right\}. \end{aligned} \tag{33}$$

As with the single-observation case, the maximum likelihood estimate requires minimization of $\sum_{i=1}^{L} \alpha(\sigma^{-1}s^{(i)}; \mathbf{a})$. But maximizing

$f(s^{(1)}, s^{(2)}, \dots, s^{(2)} \,|\, \mathbf{a}, \sigma^2)$ is, in our interest here, equivalent to maximizing $[f(s^{(1)}, s^{(2)}, \dots, s^{(L)} \,|\, \mathbf{a}, \sigma^2)]^{1/L}$. Since

$$\begin{aligned}[f(s^{(1)}, s^{(2)}, \dots, s^{(L)} \,|\, \mathbf{a}, \sigma^2)]^{1/L} \\ = (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2}\left[\frac{1}{L}\sum_{i=1}^{L} \alpha(\sigma^{-1}s^{(i)}; \mathbf{a})\right]\right\},\end{aligned} \tag{34}$$

the similarity between multiple- and single-observation estimation can easily be seen by comparing (17) and (34). From (13),

$$\begin{aligned}&\frac{1}{L}\sum_{i=1}^{L} \alpha(\sigma^{-1}s^{(i)}; \mathbf{a}) \\ &= \frac{1}{\sigma^2}\frac{1}{L}\sum_{i=1}^{L}\left\{r_a(0)r_s^{(i)}(0) + 2\sum_{j=1}^{M} r_a(j)r_s^{(i)}(j)\right\} \\ &= \frac{1}{\sigma^2}\left\{r_a(0)\left[\frac{1}{L}\sum_{i=1}^{L} r_s^{(i)}(0)\right] + 2\sum_{j=1}^{M} r_a(j)\left[\frac{1}{L}\sum_{i=1}^{L} r_s^{(i)}(j)\right]\right\},\end{aligned} \tag{35}$$

where

$$r_s^{(i)}(j) = \sum_{n=1}^{N-j} s_n^{(i)} s_{n+j}^{(i)},$$

and $s_n^{(i)}$ is simply the nth sample in $s^{(i)}$. Equations (34) and (35) clearly demonstrate that the maximum likelihood estimate of the source parameters, given multiple observations, can be obtained by the same minimization procedure as in the single-observation situation, using the autocorrelation coefficients averaged over all available observations.

The same result holds for the gain-independent case where the joint density for observations $x^{(1)} = s^{(1)}/\sigma_{(1)}$, $x^{(2)} = s^{(2)}/\sigma_{(2)}, \dots, x^{(L)} = s^{(L)}/\sigma_{(L)}$ is, after taking the Lth root,

$$\begin{aligned}[f(x^{(1)}, x^{(2)}, \dots, x^{(L)} \,|\, \mathbf{a})]^{1/L} \\ = (2\pi)^{-N/2} \exp\left\{-\frac{1}{2}\left[\frac{1}{L}\sum_{i=1}^{L} \alpha(x^{(i)}; \mathbf{a})\right]\right\}.\end{aligned} \tag{36}$$

Note that gain independence is maintained by normalizing each observation $s^{(i)}$ with its own estimate $\sigma_{(i)}$. Equation (35) thus becomes

$$\begin{aligned}&\frac{1}{L}\sum_{i=1}^{L} \alpha(x^{(i)}; \mathbf{a}) \\ &= r_a(0)\left[\frac{1}{L}\sum_{i=1}^{L} r_x^{(i)}(0)\right] + 2\sum_{j=1}^{M} r_a(j)\left[\frac{1}{L}\sum_{i=1}^{L} r_x^{(i)}(j)\right]\end{aligned}$$

$$= r_a(0)\left[\frac{1}{L}\sum_{i=1}^{L}\frac{r_s^{(i)}(0)}{\sigma_{(i)}^2}\right] + 2\sum_{j=1}^{M} r_a(j)\left[\frac{1}{L}\sum_{i=1}^{L}\frac{r_s^{(i)}(j)}{\sigma_{(i)}^2}\right]. \qquad (37)$$

Equations (35) and (37) lead to the same optimization procedure as the centroid computation in code-book design for vector quantization.[13] In centroid computation, however, the objective is to minimize the average distortion, while in the current probabilistic framework, the probability is to be maximized. The equivalence between probability and distortion measures is again witnessed.

Once the source parameters are estimated, the probability density is defined just as in the single-observation case, and expressed in (26) or (31) in terms of distortion measures.

IV. PROBABILISTIC FUNCTIONS OF MARKOV CHAINS

Consider a first-order K-state Markov chain governed by a transition probability matrix $\mathbf{V} = [v_{ij}]$, $i, j = 1, 2, \ldots, K$, and an initial probability vector $\mathbf{u}^t = [u_1, u_2, \ldots, u_K]$. Obviously,

$$\sum_{j=1}^{K} u_j = 1, \qquad u_j \geqslant 0 \quad \text{for all} \quad j \qquad (38)$$

and

$$\sum_{j=1}^{K} v_{ij} = 1 \quad \text{for any} \quad i, \qquad (39)$$

because v_{ij} is the probability of making a transition from state i to state j given that the current state is i. For any integer state sequence $\Theta = \theta_0\theta_1 \ldots \theta_T$, where $\theta_i \in \{1, 2, \ldots, K\}$, the probability of Θ being generated by the Markov chain can be easily calculated by

$$\Pr(\Theta \mid \mathbf{V}, \mathbf{u}) = u_{\theta_0} v_{\theta_0\theta_1} v_{\theta_1\theta_2} \cdots v_{\theta_{T-1}\theta_T}. \qquad (40)$$

Now suppose $\Theta = \theta_0\theta_1 \ldots \theta_T$ cannot be observed directly. Instead, we observe a stochastic process $\mathbf{S} = \mathbf{s}_1\mathbf{s}_2 \ldots \mathbf{s}_T$, produced by an underlying state sequence $\theta_1\theta_2 \ldots \theta_T$. Each state, say i, manifests itself through a probability density function $f_i(s)$. We use $\mathbf{F} = \{f_i(\cdot)\}$ to denote such a set of density functions. The probability density of observing $\mathbf{S} = S \triangleq s_1s_2 \ldots s_T$ given a specific state sequence Θ generated by the Markov chain with transition probability matrix $\mathbf{V}$ and initial probability $\mathbf{u}$ is thus

$$f(S \mid \Theta, \mathbf{V}, \mathbf{u}, \mathbf{F}) = f_{\theta_1}(s_1) f_{\theta_2}(s_2) \ldots f_{\theta_T}(s_T). \qquad (41)$$

Each s_i here is a vector without ambiguity. It follows that the probability density of observing S given $\mathbf{V}$ and $\mathbf{u}$ is

$$\begin{aligned} f(S \mid \mathbf{V}, \mathbf{u}, \mathbf{F}) &= \sum_{\text{all } \Theta} f(S, \Theta \mid \mathbf{V}, \mathbf{u}, \mathbf{F}) \\ &= \sum_{\text{all } \Theta} f(S \mid \Theta, \mathbf{V}, \mathbf{u}, \mathbf{F}) \Pr(\Theta \mid \mathbf{V}, \mathbf{u}, \mathbf{F}) \\ &= \sum_{\text{all } \Theta} f(S \mid \Theta, \mathbf{V}, \mathbf{u}, \mathbf{F}) \Pr(\Theta \mid \mathbf{V}, \mathbf{u}) \\ &= \sum_{\theta_0, \theta_1, \ldots \theta_T = 1}^{K} u_{\theta_0} v_{\theta_0 \theta_1} f_{\theta_1}(s_1) v_{\theta_1 \theta_2} f_{\theta_2}(s_2) \ldots v_{\theta_{T-1} \theta_T} f_{\theta_T}(s_T). \end{aligned} \tag{42}$$

The stochastic process $\mathbf{S}$ is characterized by the density $f(S \mid \mathbf{V}, \mathbf{u}, \mathbf{F})$ and the set of probability density functions $\mathbf{F}$, which is assumed to be known and independent of the Markov chain in the above. The triple $(\mathbf{V}, \mathbf{u}, \mathbf{F}) \triangleq \mathbf{M}$ is then called a (hidden) Markov model,[14] and the conditional density for the stochastic process $\mathbf{S}$ may be written as $f(S \mid \mathbf{M})$.

The application of hidden Markov models in speech recognition can now be formulated. It is treated as a classification problem. We wish to recognize utterances known to have been selected from some vocabulary $\mathbf{W}$ of B words $W^{(1)}, W^{(2)}, \ldots, W^{(B)}$. (We use "words" here for convenience. They may not be words in the traditional sense, but merely some lengths of speech utterances.) Every word $W^{(i)}$ is represented by a model $\mathbf{M}_i$. An observation sequence $S = s_1 s_2 \ldots s_T$ of an unknown word is given. We then apply the maximum likelihood rule and classify S as word $W^{(i)}$ iff

$$f(S \mid \mathbf{M}_i) \geqslant f(S \mid \mathbf{M}_j) \quad \text{for any} \quad j = 1, 2, \ldots, B.$$

Such an application presents two problems: evaluating $f(S \mid \mathbf{M}_i)$ and estimating model $\mathbf{M}$ that maximizes the likelihood of a given observation S. Sections II and III showed similar problems.

The computational load in evaluating $f(S \mid \mathbf{M})$ appears to be exponential in T, as we see from (42), which is a sum over all possible state sequences of length T. With the so-called forward-backward algorithm by Baum,[15] however, it is only linear in T. The estimation of model parameters $\mathbf{V}$, $\mathbf{u}$, and $\mathbf{F}$, on the other hand, is less straightforward, and no closed form solution has been found so far. An iterative reestimation algorithm by Baum[15] and Baum et al.[16] is usually employed to attack this estimation problem for a certain class of hidden Markov models, including those with Gaussian autoregressive densities of (1), (17), or (18). We shall briefly discuss the forward-backward algorithm and the reestimation formula for Markov models with Gaussian autoregressive density. Similar developments with applications in speaker identification can be found in the work of Poritz.[17] More rigorous developments of these techniques, as well as their theoretical verifications, can be found in Refs. 15 and 16.

4.1 Forward-backward recursion and trellis structure

Define the forward probabilities $\xi_0(i) = u_i$, $i = 1, 2, \ldots, K$, and

$$\xi_t(i) = \sum_{j=1}^{K} \xi_{t-1}(j) v_{ji} f_i(s_t) \tag{43}$$

for $i = 1, 2, \ldots, K$ and $t = 1, 2, \ldots, T$. Clearly,

$$\xi_t(i) = f(s_1, s_2, \ldots, s_t, \theta_t = i \mid \mathbf{M}).$$

Similarly, define backward probabilities

$$\begin{aligned} \eta_t(i) &= f(s_{t+1}, s_{t+2}, \ldots, s_T \mid \theta_t = i, \mathbf{M}) \\ &= \sum_{j=1}^{K} \eta_{t+1}(j) v_{ij} f_j(s_{t+1}) \end{aligned} \tag{44}$$

and $\eta_T(i) = 1$, for $i = 1, 2, \ldots, K$, and $t = T-1, T-2, \ldots, 0$. $\xi_t(i)$ and $\eta_t(i)$ satisfy

$$\xi_t(i)\eta_t(i) = f(S, \theta_t = i \mid \mathbf{M}). \tag{45}$$

Therefore, we have

$$\begin{aligned} f(S \mid \mathbf{M}) &= \sum_{i=1}^{K} f(S, \theta_t = i \mid \mathbf{M}) \\ &= \sum_{i=1}^{K} \xi_t(i)\eta_t(i) \end{aligned} \tag{46}$$

for any t. In particular, by letting $t = T$,

$$f(S \mid \mathbf{M}) = \sum_{i=1}^{K} \xi_T(i) \tag{47}$$

so that $f(S \mid \mathbf{M})$ can be evaluated from forward probabilities alone and the computation load is thus linear in T.

This forward-backward evaluation technique takes advantage of a trellis structure when reducing the computational burden. The complexity of a tree structure, as Fig. 2 illustrates, grows exponentially in T. It treats distinctive paths differently, as if at instance t, the number of available state indices were K^t, as we noted by the parenthesized index in Fig. 2. The original evaluation formula of (42) displays a tree structure since the summation is directly over all θ_i, $i = 0, 1, 2, \ldots, T$, in the range of 1 through K. This tree structure can be transformed easily into a trellis structure, as Fig. 3 depicts for $K = 4$, by recognizing the fact that at any instance t and any state θ_t there are always only K possible next states regardless of the past transition history. The branches in the tree structure merge into K nodes (states) at every instance. The definition of the forward probabilities can be stated as

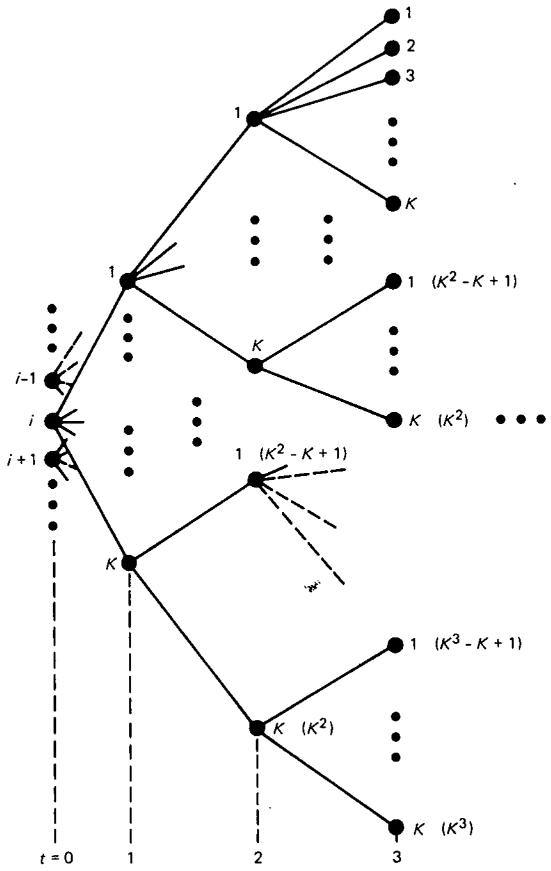


Fig. 2—Tree structure showing exponential growth of complexity.

follows: the density of the event that $\theta_t = i$ and $s_1, s_2, \ldots, s_t$ are observed is obtained by summing, over all j, the densities of the event that $\theta_{t-1} = j$ and $s_1, s_2, \ldots, s_{t-1}$ are observed, multiplied by the transition probability from state j to i and by the density of $\mathbf{s}_t = s_t$ at state i. Thus, as Fig. 3 shows for $K = 4$, $\xi_t(1) = \xi_{t-1}(1)v_{11}f_1(s_t) + \xi_{t-1}(2)v_{21}f_1(s_t) + \xi_{t-1}(3)v_{31}f_1(s_t) + \xi_{t-1}(4)v_{41}f_1(s_t)$. Every $\xi_t(i)$ can be computed recursively from $\xi_{t-1}(i)$, $i = 1, 2, \ldots, K$. Each progression from $t - 1$ to t requires the same amount of computation and, therefore, total computation is linear in T.

The tree/trellis structure also shows that dynamic programming techniques such as the Viterbi algorithm or the (M, L) algorithm, etc., can be employed efficiently to find a particular path Θ_{op} such that $f(S, \Theta_{op} \mid \mathbf{M}) = \max_{\text{all } \theta} f(S, \Theta \mid \mathbf{M})$. Later we will show that dynamic time

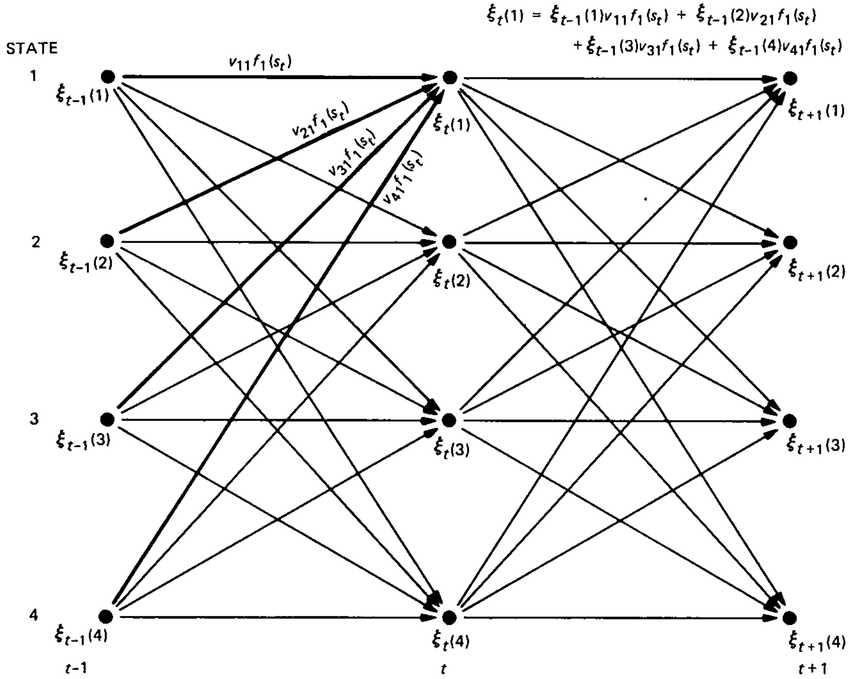


Fig. 3—Trellis structure for evaluating probability in hidden Markov models.

warping displays similar requirements of finding an optimal path (the warping function) to minimize the accumulative distortion.

4.2 Reestimation—Baum-Welch algorithm

We discuss here only the estimation of Markov models with Gaussian autoregressive densities; that is, in $\mathbf{F} = \{f_i\}_{i=1}^{K}$, every f_i takes the form of (17)* and is characterized by parameters $(\mathbf{a}_i, \sigma_i^2)$. Equivalently, we write $\mathbf{F} = \{(\mathbf{a}_i, \sigma_i^2)\}_{i=1}^{K}$. The objective of model estimation here is to find a model $\mathbf{M} = (\mathbf{V}, \mathbf{u}, \mathbf{F})$ that maximizes the likelihood of a given sequence S, for fixed number of states K and order of autoregression M.

Given a sequence S and an arbitrary model $\mathbf{M}$, the Baum-Welch reestimation algorithm iteratively finds another model $\mathbf{M}' = (\mathbf{V}', \mathbf{u}', (\mathbf{a}_i', \sigma_i'^2))$ that leads to $f(S \mid \mathbf{M}') \geq f(S \mid \mathbf{M})$. The algorithm continually improves the estimate and converges to a local optimum. Let

$$\begin{aligned}\gamma_t(i) &\triangleq f(S, \theta_t = i \mid \mathbf{M}) \\ &= \xi_t(i) \cdot \eta_t(i) \qquad (48)\end{aligned}$$

* The gain-independent case can be easily developed in the same way that the previous sections showed.

for $t = 0, 1, 2, \ldots, T$, and $i = 1, 2, \ldots, K$. Further, define

$$\gamma_t(i, j) \triangleq f(S, \theta_{t-1} = i, \theta_t = j \mid \mathbf{M}) \\ = \xi_{t-1}(i) v_{ij} f_j(s_t) \eta_t(j) \tag{49}$$

for $i, j = 1, 2, \ldots, K$, and $t = 1, 2, \ldots, T$. A new estimate of transition probability v_{ij}' for $\mathbf{V}'$ is obtained by

$$v_{ij}' = \frac{\sum_{t=1}^{T} \gamma_t(i, j)}{\sum_{t=1}^{T} \gamma_t(i)}. \tag{50}$$

And by applying Theorem 3.1 of Ref. 16 one chooses $(\mathbf{a}_i', \sigma_i'^2)$, for each $i = 1, 2, \ldots, K$, such that (see Appendix A)

$$\alpha(\bar{s}_i; \mathbf{a}_i') = \min_{\mathbf{a}} \alpha(\bar{s}_i; \mathbf{a}) \tag{51}$$

and

$$\sigma_i'^2 = \frac{\alpha(\bar{s}_i; \mathbf{a}_i')}{N \sum_{t=1}^{T} \gamma_t(i)}, \tag{52}$$

where $\bar{s}_i$ represents a composite observation whose autocorrelation coefficients are $\bar{r}_i(j)$,

$$\bar{r}_i(j) = \sum_{t=1}^{T} \gamma_t(i) r_t(j) \\ = \sum_{t=1}^{T} \gamma_t(i) \sum_{n=1}^{N-j} s_{t,n} s_{t,n+j}. \tag{53}$$

Note that $s_{t,n}$ is the nth sample in observation s_t. The composite autocorrelation $\bar{r}_i(j)$ is, as seen from (53), a weighted average autocorrelation. The weight is the density or relative frequency of the observation S being at state i at instance t. The concept of relative frequency may be helpful in relating (53) to (35) where a uniform average is involved. After $\bar{r}_i(j)$ for $i = 1, 2, \ldots, K$, and $j = 0, 1, 2, \ldots, M$ have been calculated, each $(\mathbf{a}_i', \sigma_i'^2)$ is found by using the same maximum likelihood estimation procedure as in Section II.

In estimating the Markov model parameters based upon multiple observations $S^{(1)}, S^{(2)}, \ldots, S^{(L)}$, we try to maximize the joint density given the same model $f(S^{(1)}, S^{(2)}, \ldots, S^{(L)} \mid \mathbf{M})$. Since $S^{(i)}$ are independently observed,

$$f(S^{(1)}, S^{(2)}, \ldots, S^{(L)} \,|\, \mathbf{M}) = \prod_{i=1}^{L} f(S^{(i)} \,|\, \mathbf{M}). \tag{54}$$

The Baum-Welch algorithm can be equally applied and the key equation of (50) for new estimate of transition probability becomes (see Appendix B)

$$v'_{ij} = \frac{\sum_{t=1}^{T} \left[\sum_{l=1}^{L} \gamma_t^{(l)}(i, j) \right]}{\sum_{t=1}^{T} \left[\sum_{l=1}^{L} \gamma_t^{(l)}(i) \right]}, \tag{55}$$

where

$$\gamma_t^{(l)}(i) = \frac{f(S^{(l)}, \theta_t = i \,|\, \mathbf{M})}{f(S^{(l)} \,|\, \mathbf{M})} \tag{56}$$

and

$$\gamma_t^{(l)}(i, j) = \frac{f(S^{(l)}, \theta_{t-1} = i, \theta_t = j \,|\, \mathbf{M})}{f(S^{(l)} \,|\, \mathbf{M})}. \tag{57}$$

The new estimate of $(\mathbf{a}'_i, \sigma_i'^2)$ of (51) and (52) then follows the same procedure as previously discussed for the maximum likelihood estimate based upon multiple observations. In particular, one can show (see Appendix B) that the new improved estimate $(\mathbf{a}'_i, \sigma_i'^2)$, according to the reestimation algorithm, satisfies

$$\alpha(\bar{s}_i; \mathbf{a}'_i) = \min_{\mathbf{a}} \alpha(\bar{s}_i, \mathbf{a}) \tag{58}$$

and

$$\sigma_i'^2 = \frac{\alpha(\bar{s}_i; \mathbf{a}'_i)}{N \sum_{t=1}^{T} \sum_{l=1}^{L} \gamma_t^{(l)}(i)}, \tag{59}$$

where $\bar{s}_i$ now represents a composite observation whose autocorrelation coefficients are $\bar{r}_i(j)$,

$$\begin{aligned} \bar{r}_i(j) &= \sum_{t=1}^{T} \sum_{l=1}^{L} \gamma_t^{(l)}(i) r_t^{(l)}(j) \\ &= \sum_{t=1}^{T} \sum_{l=1}^{L} \gamma_t^{(l)}(i) \sum_{n=1}^{N-j} s_{t,n}^{(l)} s_{t,n+j}^{(l)}, \end{aligned} \tag{60}$$

with $s_{t,n}^{(l)}$ being the nth sample in the observation vector $\acute{s}_t$ of sequence $S^{(l)}$.

V. DYNAMIC TIME WARPING AND HIDDEN MARKOV MODELS

In this section, we establish the relationship between dynamic time warping using linear predictive coding (LPC) distance measures and hidden Markov models with Gaussian autoregressive densities. We give a unified view on these two techniques such that comparative discussions on the two techniques can be made easily.

Consider two speech sequences, $W = w_1 w_2 \ldots w_{T_w}$ and $Y = y_1 y_2 \ldots y_{T_y}$, called the reference and the test sequence, respectively. Through some warping function $\phi(\cdot)$, or $\zeta(\cdot)$,

$$t_y = \phi(t_w), \qquad t_w = 1, 2, \ldots, T_w \tag{61}$$

$$t_w = \zeta(t_y), \qquad t_y = 1, 2, \ldots, T_y, \tag{62}$$

a correspondence between W and Y can be established. Let $d[\cdot;\cdot]$ be a distortion measure. The conventional dynamic time warping uses dynamic programming techniques to determine ϕ or ζ such that

$$D_\phi = \sum_{t_w=1}^{T_w} d[w_{t_w}; y_{\phi(t_w)}] \tag{63}$$

or

$$D_\zeta = \sum_{t_y=1}^{T_y} d[y_{t_y}; w_{\zeta(t_y)}] \tag{64}$$

is minimized. In recognition, we classify an utterance Y as word $W^{(i)}$ in a vocabulary $\{W^{(1)}, W^{(2)}, \ldots, W^{(B)}\}$ if

$$\begin{aligned} [D_\phi^{(i)}]_{\min} &= \min_{\phi^{(i)}} \left\{ \sum_{t_w=1}^{T_w^{(i)}} d[w_{t_w}^{(i)}, y_{\phi^i(t_w)}] \right\} \\ &\leq \min_{\phi^{(j)}} \left\{ \sum_{t_w=1}^{T_w^{(j)}} d[w_{t_w}^{(j)}, y_{\phi^j(t_w)}] \right\} \quad \text{for} \quad j = 1, 2, \ldots, B \end{aligned} \tag{65}$$

or

$$\begin{aligned} [D_\zeta^{(i)}]_{\min} &= \min_{\zeta^{(i)}} \left\{ \sum_{t_y=1}^{T_y^{(i)}} d[y_{t_y}, w_{\zeta^{(i)}(t_y)}^{(i)}] \right\} \\ &\leq \min_{\zeta^{(j)}} \left\{ \sum_{t_y=1}^{T_y^{(j)}} d[y_{t_y}, w_{\zeta^{(j)}(t_y)}^{(j)}] \right\} \quad \text{for} \quad j = 1, 2, \ldots, B. \end{aligned} \tag{66}$$

The choice of warping directions, i.e., using ϕ or ζ, is rather arbitrary and often is taken into consideration together with some continuity conditions.[2,3]

Now, consider a T_w-state hidden Markov model $\mathbf{M}_w = (\mathbf{V}_w, \mathbf{u}_w, \mathbf{F}_w)$, where $\mathbf{V}_w$ is a $T_w \times T_w$ matrix $\mathbf{V}_w = [v_{ij}]$, $v_{ij} = 1/T_w$, for any $i, j = 1, 2, \ldots, T_w$, $\mathbf{u}_w$ is a T_w-dimensional vector with $u_i = 1/T_w$ for any $i = 1, 2, \ldots, T_w$, and $\mathbf{F}_w$ is the set of Gaussian autoregressive densities $\{f_i\}_{i=1}^{T_w}$. Each f_i is defined by a parameter pair $(\mathbf{a}_{w,i}, \sigma_{w,i}^2)$, which is the maximum likelihood estimate based upon observation w_i. We further consider a particular state sequence, called progressive sequence $\Theta_w = \theta_0\theta_1 \ldots \theta_{T_w}$, where $\theta_i = i$ and where $\theta_0 \epsilon \{1, 2, \ldots, T_w\}$ is arbitrary. Then,

$$\begin{aligned} &f(W, \Theta_w \,|\, \mathbf{M}_w) \\ &= (1/T_w)^{T_w+1} \prod_{i=1}^{T_w} f_i(w_i \,|\, \mathbf{a}_{w,i}, \sigma_{w,i}^2) \end{aligned}$$

and from (22),

$$\begin{aligned} &\log f(W, \Theta_w \,|\, \mathbf{M}_w) \\ &= -\frac{N}{2}\left[\sum_{i=1}^{T_w} \log(2\pi\sigma_{w,i}^2) + T_w\right] - (T_w + 1)\log T_w \\ &= \max_{\Theta\epsilon\{\Theta\}_{(T_w)}} \{\log f(W, \Theta \,|\, \mathbf{M}_w)\}, \end{aligned} \tag{67}$$

where $\{\Theta\}_{(T_w)}$ denotes the set of all state sequences with length T_w. We define a similar model $\mathbf{M}_y = (\mathbf{V}_y, \mathbf{u}_y, \mathbf{F}_y)$ for Y, in which $v_{ij} = 1/T_y$, $u_i = 1/T_y$ for $i, j = 1, 2, \ldots, T_y$, and $\mathbf{F}_y = \{(\mathbf{a}_{y,i}, \sigma_{y,i}^2)\}_{i=1}^{T_y}$, where each $(\mathbf{a}_{y,i}, \sigma_{y,i}^2)$ is estimated based on y_i. We thus also have

$$\begin{aligned} &f(Y, \Theta_y \,|\, \mathbf{M}_y) \\ &= (1/T_y)^{T_y+1} \prod_{i=1}^{T_y} f_i(y_i \,|\, \mathbf{a}_{y,i}, \sigma_{y,i}^2) \end{aligned}$$

and

$$\begin{aligned} &\log f(Y, \Theta_y \,|\, \mathbf{M}_y) \\ &= -\frac{N}{2}\left[\sum_{i=1}^{T_y} \log(2\pi\sigma_{y,i}^2) + T_y\right] - (T_y + 1)\log T_y \\ &= \max_{\Theta\epsilon\{\Theta\}_{(T_y)}} \{\log f(Y, \Theta \,|\, \mathbf{M}_y)\}. \end{aligned} \tag{68}$$

A correspondence between progressive state sequences Θ_w and Θ_y is made through warping functions ϕ or ζ, as defined in (61) or (62). Within such a framework,

$$f(Y,\ \zeta(\Theta_y)\,|\,\mathbf{M}_w) = \left(\frac{1}{T_w}\right)^{T_y+1} \prod_{i=1}^{T_y} f_{\zeta(i)}(y_i\,|\,\mathbf{a}_{w,\zeta(i)},\ \sigma^2_{w,\zeta(i)}) \tag{69}$$

and

$$\log f(Y,\ \zeta(\Theta_y)\,|\,\mathbf{M}_w) = -\frac{N}{2}\sum_{i=1}^{T_y} \log(2\pi\sigma^2_{w,\zeta(i)}) - \frac{1}{2}\sum_{i=1}^{T_y} \alpha(\sigma^{-1}_{w,\zeta(i)} y_i;\ \mathbf{a}_{w,\zeta(i)}) - (T_y + 1)\log T_w. \tag{70}$$

In the above we have used $\zeta(\Theta_y)$ to denote the ζ-warped progressive sequence Θ_y. Note that (68) is also a maximum over all models with fixed $\mathbf{V}_y$ and $\mathbf{u}_y$. The difference between (68) and (70),

$$\log f(Y,\ \Theta_y\,|\,\mathbf{M}_y) - \log f(Y,\ \zeta(\Theta_y)\,|\,\mathbf{M}_w) = \frac{N}{2}\sum_{i=1}^{T_y} \left\{\frac{1}{N}\alpha(\sigma^{-1}_{w,\zeta(i)} y_i;\ \mathbf{a}_{w,\zeta(i)}) + \log \sigma^2_{w,\zeta(i)} - \log \sigma^2_{y,i} - 1\right\} + (T_y + 1)(\log T_w - \log T_y), \tag{71}$$

is thus nonnegative if $T_w \geqslant T_y$ (sufficient). In realistic situations, $\log T_w \simeq \log T_y$, so that

$$\begin{aligned}\log f(Y,\ \Theta_y\,|\,\mathbf{M}_y) - \log f(Y,\ \zeta(\Theta_y)\,|\,\mathbf{M}_w) &\simeq \frac{N}{2}\sum_{i=1}^{T_y} \left\{\frac{1}{N}\alpha(\sigma^{-1}_{w,\zeta(i)} y_i;\ \mathbf{a}_{w,\zeta(i)}) + \log \sigma^2_{w,\zeta(i)} - \log \sigma^2_{y,i} - 1\right\} \\ &= \frac{N}{2}\sum_{i=1}^{T_y} d_{IS}[y_i;\ w_{\zeta(i)}] \\ &= \frac{N}{2} D_\zeta. \end{aligned} \tag{72}$$

Therefore, the accumulative distortion D_ζ in dynamic time warping is directly related to the likelihood difference between the two models in generating Y sequence. To express the density $f(Y\,|\,\mathbf{M}_w)$ in terms of D_ζ we further have

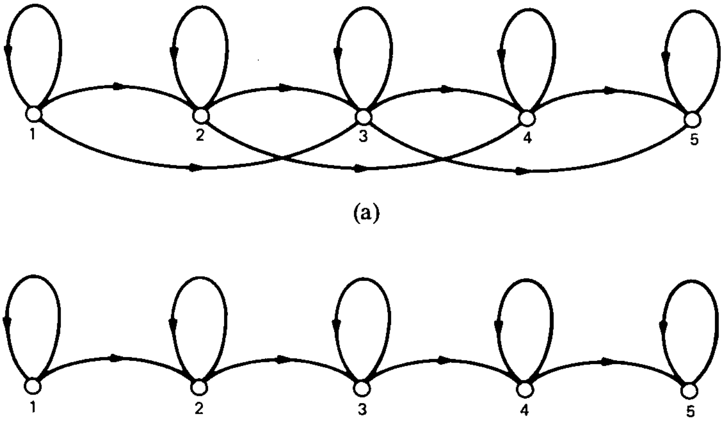
$$\begin{aligned} f(Y\,|\,\mathbf{M}_w) &= \sum_{\text{all } \zeta} f(Y,\ \zeta(\Theta_y)\,|\,\mathbf{M}_w) \\ &\simeq f(Y,\ \Theta_y\,|\,\mathbf{M}_y)\left[\sum_{\text{all } \zeta} \exp\left(-\frac{N}{2}D_\zeta\right)\right]. \end{aligned} \tag{73}$$

The same results can be extended to $f(W\,|\,\mathbf{M}_y)$,

$$
\begin{aligned}
f(W\,|\,\mathbf{M}_y) &= \sum_{\text{all } \phi} f(W,\, \phi(\Theta_y)\,|\,\mathbf{M}_y) \\
&\simeq f(W,\, \Theta_w\,|\,\mathbf{M}_w) \left[\sum_{\text{all } \phi} \exp\left(-\frac{N}{2} D_\phi \right) \right]. \qquad (74)
\end{aligned}
$$

Equation (72) demonstrates that determining a warping function ζ in dynamic time warping is equivalent to finding only the best state sequence that maximizes the density $f(Y,\, \zeta(\Theta_y)\,|\,\mathbf{M}_w)$. On the other hand, the density $f(Y\,|\,\mathbf{M}_w)$ can be calculated by summing $\exp\{-(N/2)D_\zeta\}$ over all possible warping paths ζ and then multiplying this sum of exponential terms by a constant $f(Y,\, \Theta_y\,|\,\mathbf{M}_y)$. Since $f(Y,\, \Theta_y\,|\,\mathbf{M}_y)$ is defined by the unknown sequence Y only, it does not affect the classification problem for recognition formulated above. The accumulative distortions D_ζ are then the key determining factor.

Although we have defined the transition probabilities v_{ij} to be all equal for any i and j, it is not absolutely necessary for the results of (72). It has, however, a simple but important interpretation in the calculation of $f(Y\,|\,\mathbf{M}_w)$. Equal v_{ij} for $i, j = 1, 2, \ldots, T_w$ allows the input sequence Θ_y to be warped in every possible way and in every possible permutation. One thus may not expect a good recognition performance based upon the density $f(Y\,|\,\mathbf{M}_w) = \sum_\zeta f(Y,\, \zeta(\Theta_y)\,|\,\mathbf{M}_w)$, as the time order of the observed speech sequence is crucially important. (A reversed "we" may sound very close to "you"!) For word recognition, some constraints on the transitions are therefore desirable. Markov models for two types of serial constraints, for example, are shown in Fig. 4. In Fig. 4a, single and double transitions are allowed. Similar results linking dynamic time warping and Markov modeling can be obtained in this case if



(a)

(b)

Fig. 4—Markov chains for two types of serial constraints: (a) single and double transitions permitted, and (b) only single transitions permitted.

$$\mathbf{V} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & \cdots & & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & \cdots & \\ 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & \\ \vdots & & & \ddots & & & \\ & & & & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ & & & & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & & & & 0 & 0 & 1 \end{bmatrix}.$$

For the single transition case of Fig. 4b,

$$\mathbf{V} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & & & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & & \\ \vdots & & \ddots & & & \\ & & & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & & & 0 & 0 & 1 \end{bmatrix}.$$

These kinds of constraints all have their counterparts in dynamic time warping, appearing as the continuity conditions for determining the warping function. It is expected that these constant transition probabilities lead to a similar approximation, as appeared in (72). Unlike (71), however, the lack of exactness in (72) is now caused by the increased transition probabilities at the end of the sequence. The idea of unconstrained endpoint algorithm in dynamic time warping to correct for the abrupt change in transition “possibilities” at word boundaries is, hence, noted.

VI. SUMMARY

We have given a unified theoretical view on the two dominant speech recognition techniques, namely dynamic time warping and Markov modeling. We described the role of some well-known distortion measures in the context of probability densities. After the relationship between probability density and distortion measures is made explicit, the similarities between the two techniques can be seen. We have shown that if the underlying transition structure is equiprobable, dynamic time warping is equivalent to the probabilistic modeling technique except that it searches for the best transition path to minimize the accumulative distortion, while the probabilistic technique sums the density along every possible path. The results show that the two techniques may not be mutually exclusive, particularly when the density functions are exponential and LP-related distortion measures are used. The discussion may be helpful in bringing about a better understanding of each technique and possible future improvements.

REFERENCES

1. W. A. Lea, ed., "Speech Recognition: Past, Present, and Future," in *Trends in Speech Recognition,* Englewood Cliffs, NJ: Prentice-Hall, 1980, pp. 39–98.
2. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-23,* No. 1 (February 1975), pp. 67–72.
3. L. R. Rabiner and S. E. Levinson, "Isolated Connected Word Recognition—Theory and Selected Applications," IEEE Trans. Commun., *COM-29,* No. 5 (May 1981), pp. 621–59.
4. H. Sakoe and S. Chiba, "A Dynamic Programming Approach to Continuous Speech Recognition," Proc. Int. Congress on Acoustics, Budapest, Hungary, Paper 20 C-13, 1971.
5. F. Jelinek, "Continuous Speech Recognition by Statistical Methods," Proc. IEEE, *64* (April 1976), pp. 532–56.
6. J. K. Baker, "Stochastic Modeling for Automatic Speech Understanding," in *Speech Recognition,* R. Reddy, ed., New York: Academic Press, 1975.
7. L. R. Rabiner, S. E. Levinson, and M. M. Sondhi, "On the Application of Vector Quantization and Hidden Markov Models to Speaker-Independent Isolated Word Recognition," B.S.T.J., *62,* No. 4 (February 1983), pp. 429–56.
8. F. Itakura, *Speech Analysis and Synthesis Systems Based on Statistical Method.* Doctor of Engineering Dissertation (Dept. of Engineering, Nagoya University, Japan, 1972) (in Japanese).
9. J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech,* New York: Springer-Verlag, 1976.
10. R. M. Gray et al., "Distortion Measures for Speech Processing," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-28* (August 1980), pp. 367–76.
11. J. Makhoul, "Linear Prediction: A Tutorial Review," Proc. IEEE, *63* (April 1975), pp. 561–80.
12. B. H. Juang, D. Y. Wong, and A. H. Gray, Jr., "Distortion Performance of Vector Quantization for LPC Voice Coding," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-30,* No. 2 (April 1982), pp. 294–304.
13. A. Buzo et al., "Speech Coding Based Upon Vector Quantization," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-28* (October 1980), pp. 562–74.
14. 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.
15. L. E. Baum, "An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of a Markov Process," *Inequalities, 3* (1972), pp. 1-8.
16. 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.
17. A. B. Poritz, "Linear Predictive Hidden Markov Models and the Speech Signal," ICASSP-'82 Proc., May 1982, pp. 1291–4.

APPENDIX A

Reestimation—Single Observation

We first show that the new, improved estimate ($\mathbf{a}_i'$, $\sigma_i'^2$) satisfies (51) and (52). We follow Ref. 16 and define the Q-function as

$$Q(\mathbf{M}, \mathbf{M}') = \sum_{\text{all } \Theta} f(S, \Theta \,|\, \mathbf{M}) \log f(S, \Theta \,|\, \mathbf{M}'), \tag{75}$$

where

$$\begin{aligned} \log f(S, \Theta \,|\, \mathbf{M}') \\ &= \sum_{j=1}^{K} (\log u_j')\delta(\theta_0 - j) \\ &\quad + \sum_{k=1}^{K} \sum_{i=1}^{K} \sum_{t=1}^{T} (\log v_{kl}')\delta(\theta_{t-1} - k)\delta(\theta_t - l) \\ &\quad + \sum_{i=1}^{K} \sum_{t=1}^{T} [\log f_i'(s_t)]\delta(\theta_t - i), \end{aligned} \tag{76}$$

$\delta(\cdot)$ is the Kronecker delta, and

$$\begin{aligned} f_i'(s) &= f(s \,|\, \mathbf{a}_i', \sigma_i'^2) \\ &= (2\pi)^{-N/2}(\sigma_i')^{-N} \exp\left\{-\frac{1}{2}\,\alpha(\sigma_i'^{-1}s;\, \mathbf{a}_i)\right\} \end{aligned} \tag{77}$$

according to (17). It has been shown that[16]

$$\begin{aligned} f(S \,|\, \mathbf{M}') \\ &= \sum_{\text{all } \Theta} u_{\theta_0}' \prod_{t=1}^{T} v_{\theta_{t-1}\theta_t}' f_{\theta_t}'(s_t) \\ &\geqslant \sum_{\text{all } \Theta} u_{\theta_0} \prod_{t=1}^{T} v_{\theta_{t-1}\theta_t} f_{\theta_t}(s_t) \\ &= f(S \,|\, \mathbf{M}) \end{aligned}$$

if $Q(\mathbf{M}, \mathbf{M}') \geqslant Q(\mathbf{M}, \mathbf{M})$. Therefore, we may obtain a new, improved estimate by maximizing $Q(\mathbf{M}, \mathbf{M}')$ with respect to $\mathbf{M}'$. Equation (76) shows that the contributions due to $\mathbf{u}'$, $\mathbf{V}'$, and $\mathbf{F}'$ are separated and maximization of $Q(\mathbf{M}, \mathbf{M}')$ can thus be carried out independently with respect to each parameter set. In particular,

$$\begin{aligned} Q(\mathbf{M}, \mathbf{M}') &= Q_u(\mathbf{M}, \mathbf{u}') + Q_{\mathrm{V}}(\mathbf{M}, \mathbf{V}') + Q_{\mathrm{F}}(\mathbf{M}, \mathbf{F}') \\ &= Q_u(\mathbf{M}, \mathbf{u}') + Q_{\mathrm{V}}(\mathbf{M}, \mathbf{V}') + \sum_{i=1}^{K} Q_{fi}(\mathbf{M}, f_i'), \end{aligned}$$

where

$$Q_u(\mathbf{M}, \mathbf{u}') = \sum_{\text{all}\,\Theta} f(S, \Theta\,|\,\mathbf{M}) \sum_{j=1}^{K} (\log u_j')\delta(\theta_0 - j),$$

$$Q_{\mathbf{V}}(\mathbf{M}, \mathbf{V}') = \sum_{\text{all}\,\Theta} f(S, \Theta\,|\,\mathbf{M}) \left[\sum_{k=1}^{K} \sum_{l=1}^{K} \sum_{t=1}^{T} (\log v_{kl}')\delta(\theta_{t-1} - k)\delta(\theta_t - l)\right],$$

and

$$Q_{fi}(\mathbf{M}, f_i') = \sum_{\text{all}\,\Theta} f(S, \Theta\,|\,\mathbf{M}) \sum_{t=1}^{T} [\log f_i'(s_t)]\delta(\theta_t - i).$$

We shall not repeat the treatment of $\mathbf{u}'$ and $\mathbf{V}'$ here, but only discuss the maximization of $Q_{fi}(\mathbf{M}, f_i')$. We rewrite $Q_{fi}(\mathbf{M}, f_i')$ as

$$\begin{aligned} Q_{fi}(\mathbf{M}, f_i') &= \sum_{t=1}^{T} \left[\sum_{\text{all}\,\Theta} f(S, \Theta\,|\,\mathbf{M})\delta(\theta_t - i)\right] \log f_i'(s_t) \\ &= \sum_{t=1}^{T} f(S, \theta_t = i\,|\,\mathbf{M}) \log f_i'(s_t) \\ &= \sum_{t=1}^{T} f(S, \theta_t = i\,|\,\mathbf{M}) \\ &\quad \cdot \left\{-\frac{N}{2}\log 2\pi - \frac{N}{2}\log \sigma_i'^2 - \frac{1}{2}\,\alpha(\sigma_i'^{-1}s_t;\, \mathbf{a}_i')\right\}. \end{aligned} \tag{78}$$

Maximizing $Q_{fi}(\mathbf{M}, f_i')$ with respect to $\mathbf{a}_i'$ is equivalent to minimizing

$$\begin{aligned} &\sum_{t=1}^{T} f(S, \theta_t = i\,|\,\mathbf{M})\alpha(s_t;\, \mathbf{a}_i') \\ &\quad = \sum_{t=1}^{T} f(S, \theta_t = i\,|\,\mathbf{M}) \\ &\qquad \cdot \left\{r_{a'}(0)r_t(0) + 2\sum_{j=1}^{M} r_{a'}(j)r_t(j)\right\} \\ &\quad = r_{a'}(0)\left[\sum_{t=1}^{T} f(S, \theta_t = i\,|\,\mathbf{M})r_t(0)\right] \\ &\qquad + 2\sum_{j=1}^{M} r_{a'}(j)\left[\sum_{t=1}^{T} f(S, \theta_t = i\,|\,\mathbf{M})r_t(j)\right], \end{aligned} \tag{79}$$

where $r_{a'}(j)$ and $r_t(j)$ are the j-lag autocorrelation coefficient of $\mathbf{a}_i'$ and s_t, respectively. Since (79) is simply $\alpha(\bar{s}_i;\, \mathbf{a}_i')$, (51) is thus proved. Equation (52) then follows from maximizing (78) with respect to $\sigma_i'^2$ given $\mathbf{a}_i'$.

APPENDIX B

Reestimation—Multiple Observations

We apply the Baum-Welch algorithm to maximize

$$f(S^{(1)}, S^{(2)}, \ldots, S^{(L)} \,|\, \mathbf{M}) = \prod_{i=1}^{L} f(S^{(i)} \,|\, \mathbf{M}), \tag{80}$$

given multiple observations $S^{(1)}, S^{(2)}, \ldots, S^{(L)}$. For brevity, we use $\{S\}$ to denote the set of observations $S^{(i)}$, $i = 1, 2, \ldots, L$. In addition, for each observation sequence $S^{(i)}$, there is a corresponding probable state sequence $\Theta^{(i)}$, and

$$f(S^{(i)} \,|\, \mathbf{M}) = \sum_{\text{all } \Theta^{(i)}} f(S^{(i)}, \Theta^{(i)} \,|\, \mathbf{M}).$$

We further use $\{\Theta\}$ to denote the set of probable state sequences behind the set of observations $\{S\}$. Then (80) becomes

$$f(\{S\} \,|\, \mathbf{M}) = \sum_{\text{all } \{\Theta\}} f(\{S\}, \{\Theta\} \,|\, \mathbf{M}). \tag{81}$$

Accordingly, we define the Q-function as follows:

$$Q(\mathbf{M}, \mathbf{M}') = \sum_{\{\Theta\}} f(\{S\}, \{\Theta\} \,|\, \mathbf{M}) \log f(\{S\}, \{\Theta\} \,|\, \mathbf{M}'),$$

where

$$\begin{aligned} \log f(\{S\}, \{\Theta\} \,|\, \mathbf{M}') &= \sum_{j=1}^{K} \sum_{i=1}^{L} (\log u_j') \delta(\theta_0^{(i)} - j) \\ &+ \sum_{k=1}^{K} \sum_{l=1}^{K} \sum_{i=1}^{L} \sum_{t=1}^{T} (\log v_{kl}') \delta(\theta_{t-1}^{(i)} - k) \delta(\theta_t^{(i)} - l) \\ &+ \sum_{m=1}^{K} \sum_{i=1}^{L} \sum_{t=1}^{T} [\log f_m'(s_t^{(i)})] \delta(\theta_t^{(i)} - m). \end{aligned} \tag{82}$$

and $f_m'(s)$ is defined as in (77). We shall address the maximization of $Q(\mathbf{M}, \mathbf{M}')$ with respect to $\mathbf{V}' = [v_{kl}']$ and $\mathbf{F}' = \{f_i'\}$. Extension of previous results on the initial probabilities to the current case of multiple observations is straightforward. Again, (82) shows separate contributions from different parameter sets and, hence, we write

$$Q(\mathbf{M}, \mathbf{M}') = Q_u(\mathbf{M}, \mathbf{u}') + Q_{\mathbf{V}}(\mathbf{M}, \mathbf{V}') + Q_{\mathbf{F}}(\mathbf{M}, \mathbf{F}'),$$

where

$$Q_u(\mathbf{M}, \mathbf{u}') = \sum_{\{\Theta\}} f(\{S\}, \{\Theta\} \,|\, \mathbf{M}) \sum_{j=1}^{K} \sum_{i=1}^{L} (\log u_j')\delta(\theta_0^{(i)} - j)$$

$$\begin{aligned} Q_{\mathrm{V}}(\mathbf{M}, \mathbf{V}') &= \sum_{\{\Theta\}} f(\{S\}, \{\Theta\} \,|\, \mathbf{M}) \sum_{k}^{K} \sum_{l}^{K} \sum_{i}^{L} \sum_{t}^{T} (\log v_{kl}') \\ &\quad \cdot \delta(\theta_{t-1}^{(i)} - k)\delta(\theta_t^{(i)} - l) \\ &= \sum_{k=1}^{K} \left\{ \sum_{\{\Theta\}} f(\{S\}, \{\Theta\} \,|\, \mathbf{M}) \sum_{l}^{K} \sum_{i}^{L} \sum_{t}^{T} (\log v_{kl}') \right. \\ &\quad \left. \cdot \delta(\theta_{t-1}^{(i)} - k)\delta(\theta_t^{(i)} - l) \right\} \\ &= \sum_{k=1}^{K} Q_{vk}(\mathbf{M}, \mathbf{v}_k'), \end{aligned}$$

and

$$\begin{aligned} Q_{\mathrm{F}}(\mathbf{M}, \mathbf{F}') &= \sum_{\{\Theta\}} f(\{S\}, \{\Theta\} \,|\, \mathbf{M}) \sum_{m=1}^{K} \sum_{i=1}^{L} \sum_{t=1}^{T} [\log f_m'(s_t^{(i)})]\delta(\theta_t^{(i)} - m) \\ &= \sum_{m=1}^{K} \left\{ \sum_{\{\Theta\}} f(\{S\}, \{\Theta\} \,|\, \mathbf{M}) \sum_{i}^{L} \sum_{t}^{T} [\log f_m'(s_t^{(i)})\delta(\theta_t^{(i)} - m)] \right\} \\ &= \sum_{m=1}^{K} Q_{f_m}(\mathbf{M}, f_m'). \end{aligned}$$

In the above, $\mathbf{v}_k' = [v_{k1}' \; v_{k2}' \ldots v_{kK}']^t$ and

$$\sum_{i=1}^{K} v_{ki} = 1 \quad \text{for} \quad k = 1, 2, \ldots, K. \tag{83}$$

Let Ω be the Langrangian of $Q_{vk}(\mathbf{M}, \mathbf{v}_k')$ with respect to the constraint (83),

$$\Omega = Q_{vk}(\mathbf{M}, \mathbf{v}_k') + \lambda \left[\sum_{i=1}^{K} v_{ki}' - 1 \right].$$

We need to solve the equation

$$\frac{\partial \Omega}{\partial v_{kj}'} = \frac{\partial Q_{vk}}{\partial v_{kj}'} + \lambda = 0. \tag{84}$$

Then, we have

$$\frac{\partial Q_{vk}}{\partial v_{kj}'} = -\lambda = -\sum_{j=1}^{K} v_{kj}' \lambda = \sum_{j=1}^{K} v_{kj}' \frac{\partial Q_{vk}}{\partial v_{kj}'}. \tag{85}$$

Since,

$$\begin{aligned}\frac{\partial Q_{vk}}{\partial v'_{kj}} &= \sum_{\{\Theta\}} f(\{S\}, \{\Theta\} \,|\, \mathbf{M}) \sum_{i=1}^{L} \sum_{t=1}^{T} \frac{1}{v'_{kj}} \delta(\theta_{t-1}^{(i)} - k)\delta(\theta_t^{(i)} - j)\\ &= \frac{1}{v'_{kj}} \sum_{i=1}^{L} \sum_{t=1}^{T} \sum_{\{\Theta\}} f(\{S\}, \{\Theta\} \,|\, \mathbf{M})\delta(\theta_{t-1}^{(i)} - k)\delta(\theta_t^{(i)} - j)\\ &= \frac{1}{v'_{kj}} \sum_{i=1}^{L} \sum_{t=1}^{T} f(S^{(i)}, \theta_{t-1}^{(i)} = k, \theta_t^{(i)} = j \,|\, \mathbf{M})\\ &\quad \cdot f(\{S\} \,|\, \mathbf{M})/f(S^{(i)} \,|\, \mathbf{M}), \qquad (86)\end{aligned}$$

and

$$\begin{aligned}&\sum_{j=1}^{K} v'_{kj} \frac{\partial Q_{vk}}{\partial v'_{kj}}\\ &\quad = \sum_{j=1}^{K} \sum_{i=1}^{L} \sum_{t=1}^{T} f(S^{(i)}, \theta_{t-1}^{(i)} = k, \theta_t^{(i)} = j \,|\, \mathbf{M})\\ &\qquad \cdot f(\{S\} \,|\, \mathbf{M})/f(S^{(i)} \,|\, \mathbf{M}) \qquad (87)\\ &\quad = \sum_{i=1}^{L} \sum_{t=1}^{T} f(S^{(i)}, \theta_{t-1}^{(i)} = k \,|\, \mathbf{M}) f(\{S\} \,|\, \mathbf{M})/f(S^{(i)} \,|\, \mathbf{M}),\end{aligned}$$

we arrive at the solution of (55). The term $f(\{S\} \,|\, \mathbf{M})$ in the above has no effect as it appears in both (86) and (87).

Next, we deal with the maximization of $Q_{fj}(\mathbf{M}, f'_j)$. Note from above that

$$\begin{aligned}Q_{fj}(\mathbf{M}, f'_j) &= \sum_{\{\Theta\}} f(\{S\}, \{\Theta\} \,|\, \mathbf{M}) \sum_{i=1}^{L} \sum_{t=1}^{T} [\log f'_j(s_t^{(i)})]\delta(\theta_t^{(i)} - j)\\ &= \sum_{t=1}^{T} \sum_{i=1}^{L} \left\{ \sum_{\{\Theta\}} f(\{S\}, \{\Theta\} \,|\, \mathbf{M})\delta(\theta_t^{(i)} - j) \right\} \log f'_j(s_t^{(i)})\\ &= \sum_{t=1}^{T} \sum_{i=1}^{L} f(S^{(i)}, \theta_t^{(i)} = j \,|\, \mathbf{M})\\ &\qquad \cdot \left\{ -\frac{N}{2} \log 2\pi - \frac{N}{2} \log \sigma_j'^2 - \frac{1}{2} \alpha(\sigma_j'^{-1} s_t^{(i)}; \mathbf{a}'_j) \right\}\\ &\qquad \cdot f(\{S\} \,|\, \mathbf{M})/f(S^{(i)} \,|\, \mathbf{M}), \qquad (88)\end{aligned}$$

an expression similar to (78). Maximizing $Q_{fj}(\mathbf{M}, f'_j)$ with respect to $\mathbf{a}'_j$ is equivalent to minimizing

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{i=1}^{L} f(S^{(i)}, \theta_t^{(i)} = j \mid \mathbf{M})[f(\{S\} \mid \mathbf{M})/f(S^{(i)} \mid \mathbf{M})]\alpha(s_t^{(i)}; \mathbf{a}_j')\\
&\quad = r_{a'}(0)\cdot\left[\sum_{t=1}^{T}\sum_{i=1}^{L} f(S^{(i)}, \theta_t^{(i)} = j \mid \mathbf{M})[f(\{S\} \mid \mathbf{M})/f(S^{(i)} \mid \mathbf{M})]r_t^{(i)}(0)\right]\\
&\qquad + 2\sum_{m=1}^{M} r_{a'}(m)\cdot\left[\sum_{t=1}^{T}\sum_{i=1}^{L} f(S^{(i)}, \theta_t^{(i)} = j \mid \mathbf{M})\right.\\
&\qquad \left.\cdot[f(\{S\} \mid \mathbf{M})/f(S^{(i)} \mid \mathbf{M})]r_t^{(i)}(m)\right],
\end{aligned}\tag{89}$$

where $r_{a'}(m)$ and $r_t^{(i)}(m)$ are the m-lag autocorrelation coefficient of $\mathbf{a}_j'$ and $s_t^{(i)}$ sequences, respectively. Similar to the development in Appendix A, (89) is simply $\alpha(\bar{s}_j; \mathbf{a}_j')$ and (58) must be satisfied in order to maximize $Q_{fj}(\mathbf{M}, f_j')$ with respect to $\mathbf{a}_j'$. Equation (59) then follows from maximizing (88) with respect to $\sigma_j'^2$ given $\mathbf{a}_j'$.

AUTHOR

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.