TECHNOMETRICS, AUGUST 1991, VOL. 33, NO. 3

Hidden Markov Models for Speech Recognition

B. H. Juang and L. R. Rabiner

Speech Research Department
AT&T Bell Laboratories
Murray Hill, NJ 07974

The use of hidden Markov models for speech recognition has become predominant in the last several years, as evidenced by the number of published papers and talks at major speech conferences. The reasons this method has become so popular are the inherent statistical (mathematically precise) framework; the ease and availability of training algorithms for estimating the parameters of the models from finite training sets of speech data; the flexibility of the resulting recognition system in which one can easily change the size, type, or architecture of the models to suit particular words, sounds, and so forth; and the ease of implementation of the overall recognition system. In this expository article, we address the role of statistical methods in this powerful technology as applied to speech recognition and discuss a range of theoretical and practical issues that are as yet unsolved in terms of their importance and their effect on performance for different system implementations.

KEY WORDS: Baum–Welch algorithm; Incomplete data problem; Maximum a posteriori decoding; Maximum likelihood.

Speech recognition by machine has come of age in a practical sense. Numerous speech-recognition systems are currently in operation in applications ranging from a voice dialer for telephone to a voice response system that quotes stock prices on verbal inquiry. What makes these practical benefits happen is the recent technological advances that enable speech-recognition systems to respond reliably to nonspecific talkers with a reasonably sized recognition vocabulary. One such major advance is the use of statistical methods, of which hidden Markov model (HMM) is a particularly interesting one.

The use of HMM's for speech recognition has become popular in the past decade. Although the number of reported recognition systems based on HMM's is too large to discuss in detail here, it is worthwhile to point out some of the most important as well as successful of these systems. These include the early work of the Dragon System at Carnegie Mellon University (Baker 1975), the longstanding effort of IBM on a voice-dictation system (Averbuch et al. 1987; Bahl, Jelinek, and Mercer 1983; Jelinek 1976), the work at AT&T Bell Laboratories, Institute for Defense Analyses, MIT Lincoln Labs, and Philips on whole-word recognition using HMM's (Bourlard, Kamp, Ney, and Wellekens 1985; Lee, Soong, and Juang 1988; Lippman, Martin, and Paul 1987; Poritz and Richter 1986; Rabiner, Juang, Levinson, and Sondhi 1986; Rabiner, Levinson, and Sondhi 1983; Rabiner, Wilpon, and Soong 1989), the DARPA Resource Management task (Chow et al. 1987; Lee 1989), and other related efforts (Derouault 1987; Gupta, Lennig, and Mermelstein 1987). The widespread popularity of the HMM framework can be attributed to its simple algorithmic structure, which is straightforward to implement, and to its clear performance superiority over alternative recognition structures.

Performance, particularly in terms of accuracy, is a critical factor in determining the practical value of a speech-recognition system. A speech-recognition task is often taxonomized according to its requirements in handling specific or nonspecific talkers (speaker-dependent vs. speaker-independent) and in accepting only isolated utterances or fluent speech (isolated word vs. connected word). At present, the state-of-the-art technology can easily achieve almost perfect accuracy in speaker-independent isolated-digit recognition and would commit only 2–3% digit-string errors when the digit sequence is spoken in a naturally connected manner by nonspecific talkers. Furthermore, in speaker-independent continuous speech environments with a 1,000-word vocabulary and certain grammatical constraints, several advanced systems based on HMM have been demonstrated to be able to achieve 96% word accuracy. These results sometimes rival human performance and thus, of course, affirm the potential usefulness of an automatic speech-recognition system in designated applications.

Although hidden Markov modeling has sig-

nificantly improved the performance of current speech-recognition systems, the general problem of completely fluent, speaker-independent speech recognition is still far from being solved. For example, there is no system that is capable of reliably recognizing unconstrained conversational speech, nor does there exist a good way to infer statistically the language structure from a limited corpus of spoken sentences. The purpose of this expository article is, therefore, to provide an overview of the theory of HMM, discuss the role of statistical methods, and point out a range of theoretical and practical issues that deserve attention and are necessary to understand so as to further advance research in the field of speech recognition.

1. MEASUREMENTS AND MODELING OF SPEECH

Speech is a nonstationary signal. When we speak, our articulatory apparatus (the lips, jaw, tongue, and velum, as shown in Fig. 1) modulates the air pressure and flow to produce an audible sequence of sounds. Although the spectral content of any particular sound may include frequencies up to several thousand hertz, our articulatory configuration (vocal-tract shape, tongue movement, etc.) often does not undergo dramatic changes more than 10 times per second. Speech modeling thus involves two aspects: (1) Analysis of the short-time spectral properties of individual sounds, performed at an interval on the order of 10 milliseconds (msec), and (2) characterization of the long-time development of sound sequences, on the order of 100 msec, due to articulatory configuration changes.

To see how speech may be viewed as a nonstationary signal, we show in Figure 2 a short segment (approximately 450 msec long) of the speech waveform corresponding to a recorded utterance of the word "judge." Note that digital processing of a speech signal requires discrete time sampling and quantization of the waveform. Typically, an analog speech signal is sampled at a rate of 8–20 kilohertz (kHz), and the amplitude of each waveform sample is usually represented by one of $2^{16} = 65{,}536$ values—that is, 16-bit quantization of the discrete time signal.

Short-time spectral properties of the digital speech signal are analyzed by successively placing a window over the sampled waveform as illustrated in Figure 2. The window generally has the property that it tapers toward 0 at the ends so as to minimize the discontinuity of the signal outside the window. A short-time spectral window has a typical analysis width of 10–50 msec, and successive windows are normally positioned 10–30 msec apart. A spectral-analysis method is then applied to the windowed signal to produce a parsimonious representation of the spectral properties of the speech waveform within the window. Many spectral-analysis methods have been proposed for speech-signal modeling. These include such standard methods as measurement of the discrete (fast) Fourier transform (FFT), all-pole minimum-phase linear prediction (LPC) methods, and autoregressive/moving average models (Allen and Rabiner 1977; Atal and Hanauer 1971; Cadzow 1982; Makhoul 1975; Markel and Gray 1976; Schafer and Rabiner 1971). Even the more traditional filter-bank method of spectral analysis is still used in some systems (Dautrich, Rabiner, and Martin 1983), particularly in hardware implementations. To emphasize spectral properties that are known to be important to a human listener, auditory models can be incorporated in the overall spectral representation (Cohen 1985; Ghitza 1986). In speech modeling, we often call this short-time spectral vector an observation vector or simply an observation. In Figure 2, where the analysis mechanism is illustrated, we use a 30-msec Hamming window (frame) with successive spectral frames spaced 15 msec apart. FFT spectra of the first four frames of the signal are plotted in the figure, each being fitted with a 10th order LPC all-pole smoothed model spectrum.

To see the development of sound sequences on a relatively long-time basis, we show in Figure 3 a speech waveform corresponding to a sentence, "My cap is off for the judge" (approximately two seconds long), together with a spectrogram plot of the signal. A spectrogram is a plot of successive spectra in which the horizontal and vertical coordinates are time and frequency, respectively, and the darkness at each time-frequency point represents the corresponding spectral magnitude.

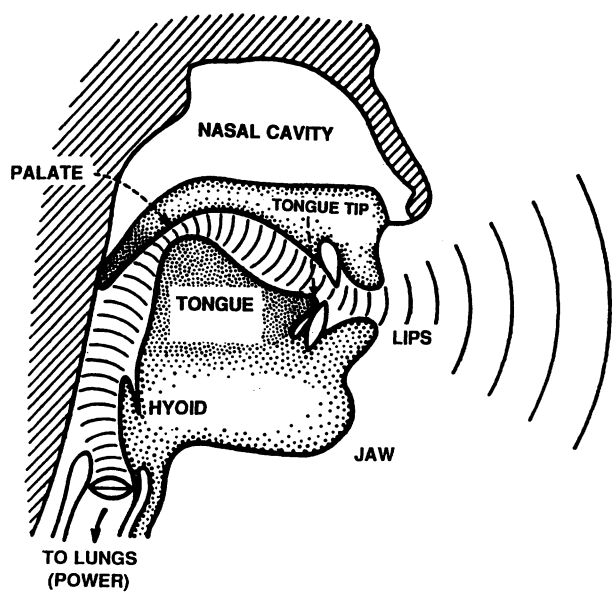

Figure 1. Schematic Description of the Human Vocal System.

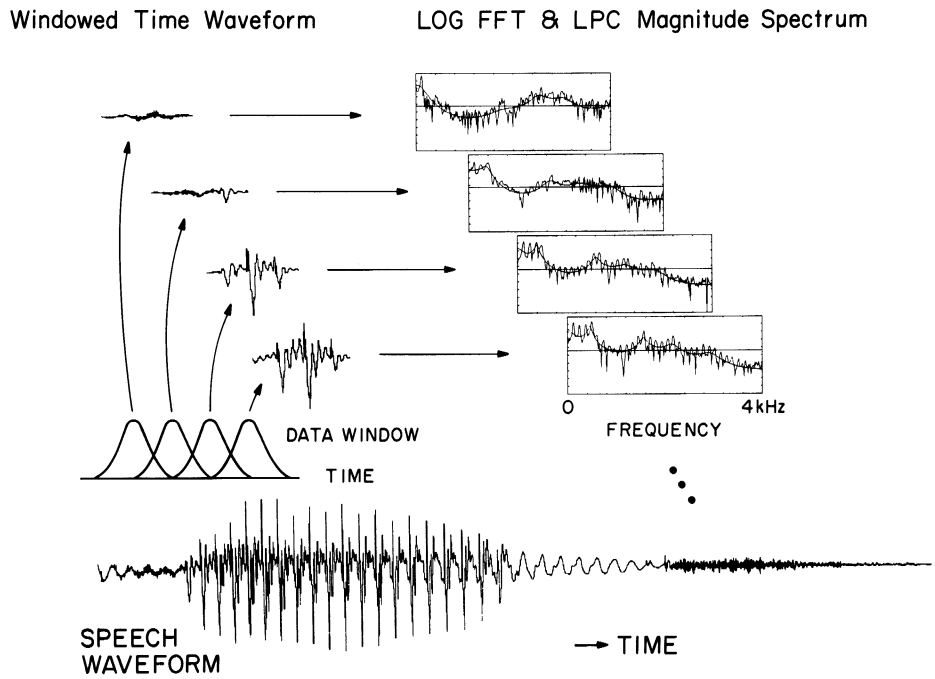

Figure 2. A Segment of Speech Waveform Corresponding to the Word "Judge" and the Resulting Short-Time Spectral Analysis of the First Four Frames.

It has been found (Juang, Rabiner, and Wilpon 1987) that spectral vectors, as represented by the so-called cepstrum, which is defined as the Fourier transform of the log magnitude spectrum—particularly the log magnitude LPC-model spectrum—have several advantages in statistical modeling for speech recognition. Computationally, the cepstrum of a stable all-pole system can be found recursively. Let the polynomial $\mathrm{P}(z) = 1 + p_1 z^{-1} + p_2 z^{-2} + \cdots + p_K z^{-K}$ have all of its roots inside the unit circle. The LPC (smoothed) spectrum of a frame of speech has the form $\sigma/\mathrm{P}(z)$, in terms of the z transform, where σ is the gain term and K is typically on the order of 10–16. Since $\ln \mathrm{P}(z^{-1})$ is analytic inside the unit circle, it can be represented in a Taylor series, leading to the Laurent expansion

$$\ln[\sigma/\mathrm{P}(z)] = \ln \sigma + \sum_{k=1}^{\infty} c(k) z^{-k}.$$

The coefficients $c(k)$ defined previously are often called the LPC cepstrum. Figure 4 shows histograms of the first 12 LPC-cepstral coefficients (derived from 10th-order all-pole models) obtained from a speech data base of 70 sentences spoken by seven people. (The speech material in the data base includes many different sentences and spans a wide range of speech sounds. The bandwidth of the speech signal was limited to 4 kHz and a sampling rate of 8 kHz was used.) For speech recognition, a weighting is generally ap-

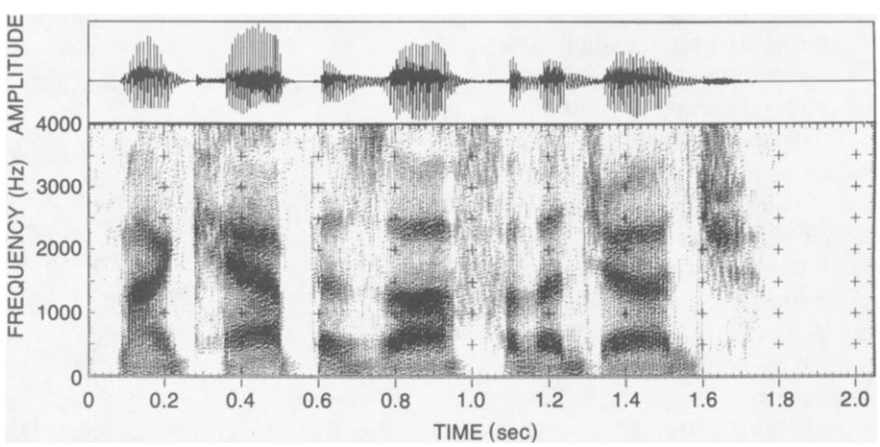

Figure 3. Speech Amplitude and Resulting Spectrogram for the Sentence, "My Cap Is Off for the Judge."

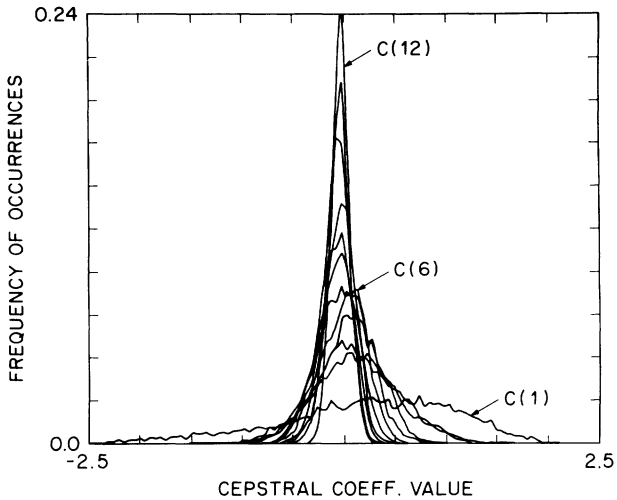

Figure 4. Histograms of the First 12 LPC Cepstral Coefficients From a Seven-Speaker 70-Sentence Speech Data Base.

plied to the LPC cepstrum before further processing (Juang et al. 1987). A vector, such as the cepstrum, that represents a short time speech spectrum is considered an observation of speech.

Another type of speech observation that is often used is a discrete-symbol representation of the spectral vector of each frame that results from a classification procedure called spectral labeling. The discrete symbol is obtained by choosing one out of a finite collection of several hundred spectral prototypes. The chosen spectral prototype is the one that is closest (in some well-defined spectral sense) to the input speech spectrum. Statistical modeling is performed on the index sequence of the closest spectral prototypes. The concept of observation distribution is very different in this case of discrete symbols from that of the continuous distribution of parameters that define a spectrum.

On a longer time basis, there are many ways to characterize the sequence of sounds—that is, running speech—as represented by a sequence of spectral observations. The most direct way is to register the spectral sequence directly without further modeling. If we denote the spectral vector at time t by $\mathbf{O}_t$ and the observed spectral sequence corresponding to the sequence of speech sounds lasts from $t = 1$ to $t = T$, a direct spectral sequence representation is then simply $\{\mathbf{O}_t\}_{t=1}^{T} = (\mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_T)$. Alternatively, one can model the sequence of spectra in terms of a Markov chain that describes the way one sound changes to another. Some perspectives as to how these two seemingly different methodologies relate to each other were given by Juang (1984) and Bridle (1984). In this article, we only discuss the latter case in which an explicit probabilistic structure is imposed on the sound-sequence representation.

HMM Formulation

Consider a first-order N-state Markov chain as illustrated for $N = 3$ in Figure 5. The system can be described as being in one of the N distinct states 1, 2, . . . , N at any discrete time instant t. We use the state variable q_t as the state of the system at discrete time t. The Markov chain is then described by a state transition probability matrix $A = [a_{ij}]$, where

$$a_{ij} = \Pr(q_t = j \mid q_{t-1} = i), \qquad 1 \le i, j \le N \tag{1}$$

with the following axiomatic constraints:

$$a_{ij} \ge 0 \tag{2}$$

and

$$\sum_{j=1}^{N} a_{ij} = 1 \quad \text{for all } i. \tag{3}$$

Note that in (1) we have assumed homogeneity of the Markov chain so that the transition probabilities do not depend on time. Assume that at $t = 0$ the state of the system q_0 is specified by an initial state probability $\pi_i = \Pr(q_0 = i)$. Then, for any state sequence $\mathbf{q} = (q_0, q_1, q_2, \ldots, q_T)$, the probability of $\mathbf{q}$ being generated by the Markov chain is

$$\Pr(\mathbf{q} \mid A, \pi) = \pi_{q_0} a_{q_0 q_1} a_{q_1 q_2} \cdots a_{q_{T-1} q_T}. \tag{4}$$

Suppose now that the state sequence $\mathbf{q}$ cannot be readily observed. Instead, we envision each observation $\mathbf{O}_t$, say a cepstral vector as mentioned previously, as being produced with the system in state q_t, $q_t \in \{1, 2, \ldots, N\}$. We assume that the production of $\mathbf{O}_t$ in each possible state i ($i = 1, 2, \ldots, N$) is stochastic and is characterized by a set of observation probability measures $B = \{b_i(\mathbf{O}_t)\}_{i=1}^{N}$,

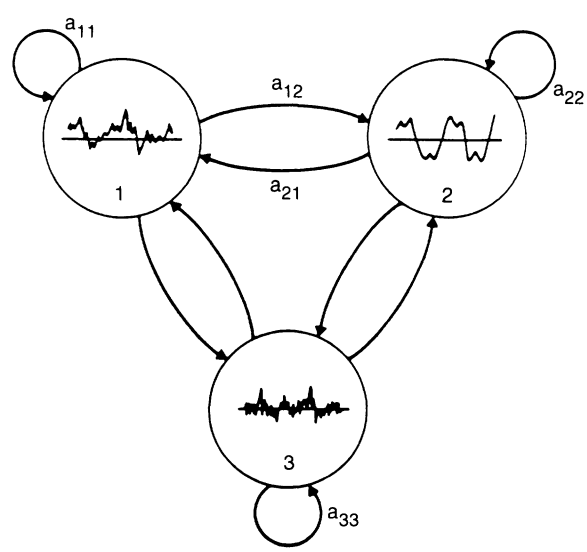

Figure 5. A First-Order Three-State Markov Chain With Associated Processes.

where

$$b_i(\mathbf{O}_t) = \Pr(\mathbf{O}_t \mid q_t = i). \quad (5)$$

If the state sequence $\mathbf{q}$ that led to the observation sequence $\mathbf{O} = (\mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_T)$ is known, the probability of $\mathbf{O}$ being generated by the system is assumed to be

$$\Pr(\mathbf{O} \mid \mathbf{q}, B) = b_{q_1}(\mathbf{O}_1) b_{q_2}(\mathbf{O}_2) \ldots b_{q_T}(\mathbf{O}_T). \quad (6)$$

The joint probability of $\mathbf{O}$ and $\mathbf{q}$ being produced by the system is simply the product of (4) and (6), written as

$$\Pr(\mathbf{O}, \mathbf{q} \mid \pi, A, B) = \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(\mathbf{O}_t). \quad (7)$$

It then follows that the stochastic process, represented by the observation sequence $\mathbf{O}$, is characterized by

$$\Pr(\mathbf{O} \mid \pi, A, B) = \sum_{\mathbf{q}} P(\mathbf{O}, \mathbf{q} \mid \pi, A, B)$$

$$= \sum_{\mathbf{q}} \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(\mathbf{O}_t), \quad (8)$$

which describes the probability of $\mathbf{O}$ being produced by the system without assuming the knowledge of the state sequence in which it was generated. The triple $\lambda = (\pi, A, B)$ thus defines an HMM (8). In the following, we shall refer to λ as the model and the model parameter set interchangeably without ambiguity.

The particular formulation of (8) is quite similar to that of the incomplete data problem in statistics (Dempster, Laird, and Rubin 1977). In terms of the physical process of a speech signal, one interpretation that may be helpful for initial understanding of the problem is that a state represents an abstract speech code (such as a phoneme) embedded in a sequence of spectral observations, and because speech is normally produced in a continuous manner, it is often difficult and sometimes unnecessary to determine how and when a state transition (from one abstract speech code to another) is made. Therefore, in (8) we do not assume explicit, definitive observation of the state sequence $\mathbf{q}$, although the Markovian structure of the state sequence is strictly implied. This is why it is called a "hidden" Markov model.

2. THE STATISTICAL METHOD OF THE HIDDEN MARKOV MODEL

In the development of the HMM methodology, the following problems are of particular interest. First, given the observation sequence $\mathbf{O}$ and a model λ, how do we efficiently evaluate the probability of $\mathbf{O}$ being produced by the source model λ—that is, $\Pr(\mathbf{O} \mid \lambda)$? Second, given the observation $\mathbf{O}$, how do we solve the inverse problem of estimating the parameters in λ? Although the probability measure of (8) does not depend explicitly on q, the knowledge of the most likely state sequence $\mathbf{q}$ that led to the observation $\mathbf{O}$ is desirable in many applications. The third problem then is how to deduce from $\mathbf{O}$ the most likely state sequence $\mathbf{q}$ in a meaningful manner. According to convention (Ferguson 1980) we call these three problems (1) the evaluation problem, (2) the estimation problem, and (3) the decoding problem. In the following sections, we describe several conventional solutions to these three standard problems.

2.1 The Evaluation Problem

The main concern in the evaluation problem is computational efficiency. Without complexity constraints, one can simply evaluate $\Pr(\mathbf{O} \mid \lambda)$ directly from the definition of (8). Since the summation in (8) involves N^{T+1} possible $\mathbf{q}$ sequences, the total computational requirements are on the order of $2T \cdot N^{T+1}$ operations. The need to compute (8) without the exponential growth of computation, as a function of the sequence length T, is the first challenge for implementation of the HMM technique. Fortunately, using the well-known forward-backward procedure (Baum 1972), this exorbitant computational requirement of the direct summation can be easily alleviated.

A forward induction procedure allows evaluation of the probability $\Pr(\mathbf{O} \mid \lambda)$ to be carried out with only a computational requirement linear in the sequence length T and quadratic in the number of states N. To see how this is done, let us define the forward variable $\alpha_t(i)$ as $\alpha_t(i) = \Pr(\mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_t, q_t = i \mid \lambda)$—that is, the probability of the partial observation sequence up to time t and state $q_t = i$ at time t. With reference to Figure 6, which shows a trellis structure implementation of the computation of $\alpha_t(i)$, we see that the forward variable can be calculated inductively by

$$\alpha_t(j) = \left[\sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} \right] b_j(\mathbf{O}_t).$$

The desired result is simply $\Pr(\mathbf{O} \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i)$.

This tremendous reduction in computation makes the HMM method attractive and viable for speech-recognizer designs because the evaluation problem can be viewed as one of scoring how well an unknown observation sequence (corresponding to the speech to be recognized) matches a given model (or sequence of models) source, thus providing an efficient mechanism for classification.

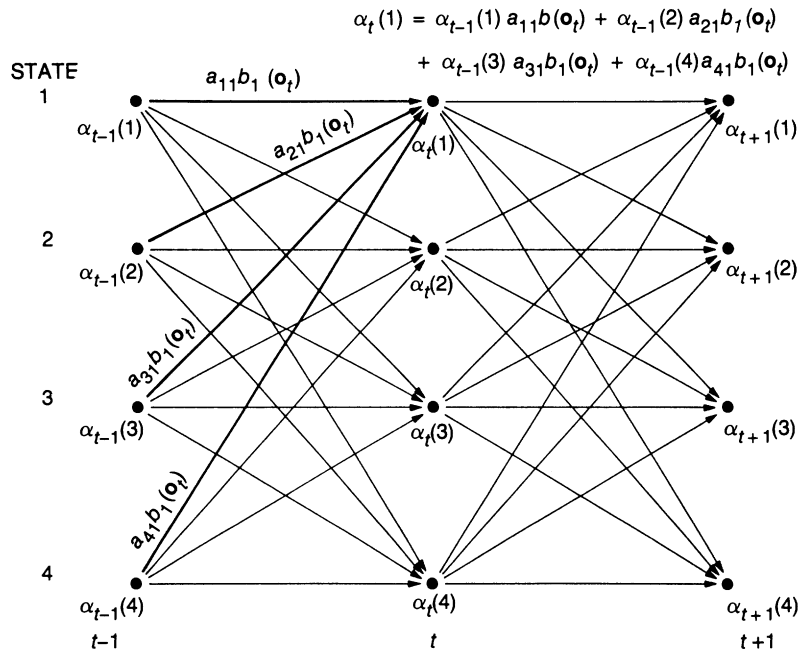

Figure 6. A Trellis Structure for the Calculation of the Forward Partial Probabilities $\alpha_t(i)$.

2.2 The Estimation Problem

Given an observation sequence (or a set of sequences) $\mathbf{O}$, the estimation problem involves finding the "right" model parameter values that specify a model most likely to produce the given sequence. In speech recognition, this is often called "training," and the given sequence, on the basis of which we obtain the model parameters, is called the training sequence, even though the formulation here is statistical.

In solving the estimation problem, we often follow the method of maximum likelihood (ML); that is, we choose λ such that $\Pr(\mathbf{O} \mid \lambda)$, as defined by (8), is maximized for the given training sequence $\mathbf{O}$. The Baum–Welch algorithm (Baum and Egon 1967; Baum and Petrie 1966; Baum, Petrie, Soules, and Weiss 1970; Baum and Sell 1968) (often blended with the forward-backward algorithm because of its interpretation as an extension of the forward induction procedure to the evaluation problem) cleverly accomplishes this maximization objective in a two-step procedure. Based on an existing model λ' (possibly obtained randomly), the first step transforms the objective function $\Pr(\mathbf{O} \mid \lambda)$ into a new function $Q(\lambda', \lambda)$ that essentially measures a divergence between the initial model λ' and an updated model λ. The Q function is defined, for the simplest case, as

$$Q(\lambda', \lambda) = \sum_{\mathbf{q}} \Pr(\mathbf{O}, \mathbf{q} \mid \lambda') \log \Pr(\mathbf{O}, \mathbf{q} \mid \lambda), \quad (9)$$

where $\Pr(\mathbf{O}, \mathbf{q} \mid \lambda)$ is given in (7). Because $Q(\lambda', \lambda) \geq Q(\lambda', \lambda')$ implies $\Pr(\mathbf{O} \mid \lambda) \geq \Pr(\mathbf{O} \mid \lambda')$, we can then simply maximize the function $Q(\lambda', \lambda)$ over λ to improve λ' in the sense of increasing the likelihood $\Pr(\mathbf{O} \mid \lambda)$. The maximization of the Q function over λ is the second step of the algorithm. The algorithm continues by replacing λ' with λ and repeating the two steps until some stopping criterion is met. The algorithm is of a general hill-climbing type and is only guaranteed to produce fixed-point solutions, although in practice the lack of global optimality does not seem to cause serious problems in recognition performance (Paul 1985). Note that the classical EM algorithm of Dempster et al. (1977) parallels closely the Baum–Welch algorithm. As noted by Baum in the discussion section of Dempster et al. (1977), the incomplete data formulation of Dempster et al. is essentially identical to the HMM formulation without the Markov-chain constraints. The Q function of (9) is clearly an expectation operation, so the two-step algorithm is identical to the E(xpectation)-M(aximization) algorithm.

The ML method is, however, not the only possible choice for solving the estimation problem. As will be discussed later, other alternatives are attractive and offer different modeling perspectives.

2.3 The Decoding Problem

As noted previously, we often are interested in uncovering the most likely state sequence that led to the observation sequence $\mathbf{O}$. Although the probability measure of an HMM, by definition, does not explicitly involve the state sequence, it is important in many applications to have the knowledge of the most likely state sequence for several reasons. As an example, if we use the states of a word model to

represent the distinct sounds in the word, it may be desirable to know the correspondence between the speech segments and the sounds of the word, because the duration of the individual speech segments provides useful information for speech recognition.

As with the second problem, there are several ways to define the decoding objective. The most trivial choice is, following the Bayesian framework, to maximize the (instantaneous) a posteriori probability

$$\gamma_t(i) = \Pr(q_t = i \mid \mathbf{O}, \lambda); \tag{10}$$

that is, we decode the state at time t by choosing $\bar{q}_t$ to be

$$\bar{q}_t = \arg \max_{1 \le i \le N} \gamma_t(i). \tag{11}$$

It is also possible to extend the definition of (10) to the cases of pairs of states or triples of states and so on. For example, the rule

$$(\bar{q}_t, \bar{q}_{t+1}) = \arg \max_{1 \le i,j \le N} \Pr(q_t = i, q_{t+1} = j \mid \mathbf{O}, \lambda) \tag{12}$$

will produce the maximum a posteriori (MAP) result of the minimum number of incorrectly decoded state pairs, given the observation sequence $\mathbf{O}$.

Although the preceding discussion shows the flexibility of possible localized decoding, we often choose to work on the entire state sequence $\mathbf{q}$ by maximizing $\Pr(\mathbf{q} \mid \mathbf{O}, \lambda)$ for three reasons: (1) It is optimal for the unknown observation $\mathbf{O}$ in the MAP sense, (2) speech utterances are usually not prohibitively long so as to require locally (rather than globally) optimal decoding, and (3) it is possible to formulate the maximization of $\Pr(\mathbf{q} \mid \mathbf{O}, \lambda)$ in a sequential manner to be solved by dynamic programming methods such as the Viterbi algorithm (Forney 1973).

Maximization of $\Pr(\mathbf{q} \mid \mathbf{O}, \lambda)$ is equivalent to maximization of $\Pr(\mathbf{q}, \mathbf{O} \mid \lambda)$ because $\Pr(\mathbf{O} \mid \lambda)$ is not involved in the optimization process. From (7), we see that

$$\begin{aligned}\Pr(q_1, q_2, \ldots, q_t, \mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_t \mid \lambda) \\ = \Pr(q_1, q_2, \ldots, q_{t-1}, \mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_{t-1} \mid \lambda) \\ \cdot a_{q_{t-1}q_t} b_{q_t}(\mathbf{O}_t).\end{aligned} \tag{13}$$

If we define

$$\begin{aligned}\delta_t(i) \triangleq \max_{q_1, q_2, \ldots, q_{t-1}} \Pr(q_1, q_2, \ldots, q_t = i, \\ \mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_t \mid \lambda),\end{aligned} \tag{14}$$

then the following recursion is true:

$$\delta_{t+1}(j) = [\max_i \delta_t(i) a_{ij}] b_j(\mathbf{O}_{t+1}). \tag{15}$$

The optimal state sequence is thus the one that leads to $\delta_T(\bar{q}_T) = \max_i \delta_T(i)$. This recursion is in a form suitable for the application of the Viterbi algorithm.

2.4 Speech Recognition Using HMM's

The typical use of HMM's in speech recognition is not very different from the traditional pattern-matching paradigm (Duda and Hart 1973). Successful application of HMM methods usually involves the following steps:

1. Define a set of L sound classes for modeling, such as phonemes or words; call the sound classes $V = \{v_1, v_2, \ldots, v_L\}$.
2. For each class, collect a sizable set (the training set) of labeled utterances that are known to be in the class.
3. Based on each training set, solve the estimation problem to obtain a "best" model λ_i for each class v_i $(i = 1, 2, \ldots, L)$.
4. During recognition, evaluate $\Pr(\mathbf{O} \mid \lambda_i)$ $(i = 1, 2, \ldots, L)$ for the unknown utterance $\mathbf{O}$ and identify the speech that produced $\mathbf{O}$ as class v_j if

$$\Pr(\mathbf{O} \mid \lambda_j) = \max_{1 \le i \le L} \Pr(\mathbf{O} \mid \lambda_i). \tag{16}$$

Since the detailed characteristics of how to implement an HMM recognizer are not essential to this article, we will omit them here. Interested readers should consult Jelinek, Bahl, and Mercer (1975) and Levinson, Rabiner, and Sondhi (1983) for more specifics related to individual applications.

3. STRENGTHS OF THE METHOD OF HIDDEN MARKOV MODELS AS APPLIED TO SPEECH RECOGNITION

The strengths of the HMM method lie in two broad areas: (1) Its mathematical framework and (2) its implementational structure. In terms of the mathematical framework, we discuss the method's consistent statistical methodology and the way it provides straightforward solutions to related problems. In terms of the implementational structure, we discuss the inherent flexibility the method provides in dealing with various sophisticated speech-recognition tasks and the ease of implementation, which is one of the crucial considerations in many practical engineering systems.

3.1 The Consistent Statistical Framework of the HMM Methodology

The foundation of the HMM methodology is built on the well-established field of statistics and probability theory. That is to say, the development of the methodology follows a tractable mathematical structure that can be examined and studied analytically.

The basic theoretical strength of the HMM is that it combines modeling of stationary stochastic processes (for the *short-time* spectra) and the temporal relationship among the processes (via a Markov chain) together in a well-defined probability space. The measure of such a probability space is defined by (8).This combination allows us to study these two separate aspects of modeling a dynamic process (like speech) using one consistent framework.

In addition, this combination of short-time static characterization of the spectrum within a state and the dynamics of change across states is rather elegant because the measure of (8) can be decomposed simply into a summation of the joint probability of $\mathbf{O}$, the observation, and $\mathbf{q}$, the state sequence, as defined by (7). The decomposition permits independent study and analysis of the behavior of the short-time processes and the long-term characteristic transitions. Since decoding and recognition are our main concerns, this also provides an intermediate level of decision that can be used to choose among alternate configurations of the models for the recognition task. This kind of flexibility with consistency is particularly useful for converting a time-varying signal such as speech, without clear anchor points that mark each sound change, into a sequence of (sound) codes.

3.2 The Training Algorithm for HMM's

Another attractive feature of HMM's comes from the fact that it is relatively easy and straightforward to train a model from a given set of labeled training data (one or more sequences of observations).

When the ML criterion is chosen as the estimation objective—that is, maximization of $\Pr(\mathbf{O} \mid \lambda)$ over λ—the well-known Baum–Welch algorithm is an iterative hill-climbing procedure that leads to, at least, a fixed-point solution as explained in Section 2.2. If we choose the state-optimized (or decoded) likelihood defined by

$$L_\lambda(\bar{\mathbf{q}}) = \max_{\mathbf{q}} \Pr(\mathbf{O}, \mathbf{q} \mid \lambda), \tag{17}$$

where

$$\bar{\mathbf{q}} = \arg\max_{\mathbf{q}} \Pr(\mathbf{O}, \mathbf{q} \mid \lambda) \tag{18}$$

as the optimization criterion, the segmental k-means algorithm (Juang and Rabiner 1990; Rabiner, Wilpon, and Juang 1986), which is an extended version of the Viterbi training/segmentation algorithm (Jelinek 1976), can be conveniently used to accomplish the parameter training task.

The segmental k-means algorithm, as can be seen from the objective function of (17), involves two optimization steps—namely, the segmentation step and the optimization step. In the segmentation step, we find a state sequence $\bar{\mathbf{q}}$ such that (17) is obtained for a given model λ and an observation sequence $\mathbf{O}$. Then, given a state sequence $\bar{\mathbf{q}}$ and the observation $\mathbf{O}$, the optimization step finds a new set of model parameters $\bar{\lambda}$ so as to maximize (17); that is,

$$\bar{\lambda} = \arg\max_{\lambda}\{\max_{\mathbf{q}} \Pr(\mathbf{O}, \mathbf{q} \mid \lambda)\}. \tag{19}$$

Equation (19) can be rewritten as

$$\bar{\lambda} = \arg\max_{\lambda}\{\max_{\mathbf{q}} [\log \Pr(\mathbf{O} \mid \mathbf{q}, \lambda) + \log \Pr(\mathbf{q} \mid \lambda)]\}. \tag{20}$$

Note that $\max_\lambda[\log\Pr(\mathbf{O} \mid \bar{\mathbf{q}}, \lambda) + \log \Pr(\bar{\mathbf{q}} \mid \lambda)]$ consists of two terms that can be separately optimized since $\log \Pr(\bar{\mathbf{q}} \mid \lambda)$ is a function of only A, the state transition probability matrix, and $\log \Pr(\mathbf{O} \mid \bar{\mathbf{q}}, \lambda)$ is a function of only B, the family of (intrastate) observation distributions. (We neglect the initial state probability for simplicity in presentation.) This separate optimization is the main distinction between the Baum–Welch algorithm and the segmental k-means algorithm.

These two training algorithms (Baum–Welch and segmental k means) both result in well-formulated and well-behaved solutions. [For a theoretical comparison of the two methods in terms of likelihood differences and state posteriori probability deviations, interested readers should consult Merhav and Ephraim (in press).] The segmental k-means algorithm, however, due to the separate optimization of the components of the model parameter set, leads to a more straightforward (simpler with less computation and numerical difficulties) implementation.

The ease of HMM training also extends to the choice of observation distributions. It is known (Juang 1985; Juang and Rabiner 1985; Liporace 1982) that these algorithms can accommodate observation densities that are (a) strictly log-concave densities, (b) elliptically symmetric densities, (c) mixtures of distributions of the preceding two categories, and (d) discrete distributions. These choices of observation distribution in each state of the model allow accurate modeling of virtually unlimited types of data.

3.3 Modeling Flexibility

The flexibility of the basic HMM is manifested in three aspects of the model, namely: model topology, observation distributions, and decoding hierarchy.

Many topological structures for HMM's have been studied for speech modeling. For modeling isolated utterances (i.e., whole words or phrases), we often

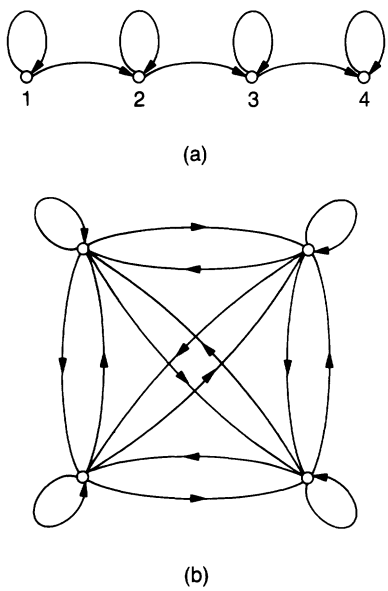

Figure 7. (a) Left-to-Right Hidden Markov Model, (b) Ergodic Hidden Markov Model.

use left-to-right models (Bakis 1976; Rabiner et al. 1983) of the type shown in Figure 7a, since the utterance begins and ends at well-identified time instants (except in the case of very noisy or corrupted speech) and the sequential behavior of the speech is well represented by a sequential HMM. For other speech-modeling tasks, the use of ergodic models (Levinson 1987) of the type shown in Figure 7b is often more appropriate. The choice of topological configuration and the number of states in the model is generally a reflection of the a priori knowledge of the particular speech source to be modeled and is not in any way related to the mathematical tractability or implementational considerations.

In Section 3.2, we pointed out that the range of observation distributions that can be accommodated by well-developed training algorithms is rather large. There are no real analytical problems that make the use of any of this rather rich class of distributions impractical. Since speech has been shown to display quite irregular probability distributions (Jayant and Noll 1984; Juang, Rabiner, Levinson, and Sondhi 1985), both in waveform and spectral parameters, one indeed needs the freedom to choose an appropriate distribution model that fits the observation well and yet is easy to obtain.

In modeling spectral observations, we have found the use of mixture densities (Juang 1985; Juang and Rabiner 1985) beneficial. With $f_i(\cdot)$ denoting the kernal density function, the mixture density assumes the form

$$b(\mathbf{O}) = \sum_{i=1}^{M} c_i f_i(\mathbf{O}), \tag{21}$$

where c_i is the mixture component weight, $\sum_{i=1}^{M} c_i = 1$, and M is the number of mixture components. This mixture distribution function is used to characterize the distribution of the observations in each state. By varying the number of mixture components, M, it can be shown that it is possible to approximate densities of virtually any shape (unimodal, multimodal, heavy-tail, etc.).

With specific constraints, the basic form of the mixture distribution of (21) can be modified to accommodate several other types of distributions, giving rise to the so-called vector quantizer HMM (Rabiner et al. 1983), semicontinuous HMM (Huang and Jack 1989), or continuous HMM (Bahl, Brown, de Souza, and Mercer 1988b; Poritz and Richter 1986; Rabiner et al. 1986).

The choice of observation distributions also extends to the case of the HMM itself. One can form an HMM with each state characterized by another HMM; that is, $b_i(\mathbf{O})$ of each state ($i = 1, 2, \ldots, N$) can assume the form of an HMM probability measure as defined by (8). This principle is the basis of many of the subword unit-based speech-recognition algorithms (Lee, Juang, Soong, and Rabiner 1989; Lee et al. 1988).

3.4 Ease of Implementation

Two areas of concern in the implementation of any algorithm are the potential for numerical difficulties and the computational complexity. The HMM is no exception.

The potential numerical difficulties in implementing HMM systems come from the fact that the terms in the HMM probability measure of (7) and (8) are multiplicative. A direct outcome of the multiplicative chain is the need for excessive dynamic range in numerical values to prevent overflow or overflow problems in digital implementations.

Numerical scaling and interpolation are two reasonable ways of avoiding such numerical problems. The scaling algorithm, well documented by Levinson et al. (1983) and Juang and Rabiner (1985), alleviates the dynamic-range problem by normalizing the par-

tial probabilities, such as the forward variable defined in Section 2.1, at each time instance before they cause overflow or underflow. The scaling algorithm is naturally blended in the forward-backward procedure. Normalization alone, however, does not entirely solve the numeric problems that result from insufficient data support. Insufficient data support can cause spurious singularities in the model parameter space. One may resort to parameter smoothing and interpolation to alleviate such numerical singularity problems. A particularly interesting method to deal with sparse data problems is the scheme of deleted interpolation proposed by Jelinek and Mercer (1980). For HMM speech recognition, some trivial measures such as setting a numeric floor to prevent singularity are often found beneficial and are straightforward to implement (Lee, Lin, and Juang 1991; Rabiner et al. 1986).

With the understanding of the relationship between the trellis structure (Juang 1984) and the decoding structure of the HMM, as discussed in the evaluation problem, we are able to apply the HMM to many complicated problems without much concern as to computational complexity. A simple calculation could verify that a typical off-the-shelf digital signal processor of 10 million floating point operations per second would be able to support a 100-word recognition vocabulary for a real-time performance. Even as recognition vocabularies increase to size 1,000 or more, the required processing often remains pretty much the same because of grammatical constraints that limit the average number of words following a given word to somewhere on the order of 100 words. [This effect is called *word-average branching factor* or *perplexity* and has been shown to be on the order of 100 for several large vocabulary-recognition tasks (Bahl et al. 1980; Chow et al. 1987).]

4. HIDDEN MARKOV MODEL ISSUES FOR FURTHER CONSIDERATION

The basic theory of hidden Markov modeling has been developed over the last two decades. When applied to speech recognition, however, there are still some remaining issues to be resolved. We begin with a discussion of parameter-estimation criteria as applied to optimal decoding of the observation sequence.

4.1 Parameter-Estimation Criteria

The original HMM parameter estimation was formulated as an inverse problem: Given an observation sequence $\mathbf{O}$ and an assumed source model, estimate the (source) parameter set λ, which maximizes the probability that $\mathbf{O}$ was produced by the source. The ML method, which seeks to maximize $\Pr(\mathbf{O} \mid \lambda)$, is optimal according to this criterion. The Baum–Welch reestimation algorithm, as described previously, is a convenient, straightforwardly implementable solution to the ML HMM estimation problem.

The ML method, however, need not be optimal in terms of minimizing classification error rate in recognition tasks in which the observation $\mathbf{O}$ is said to be produced by one of the many (say L) source classes, $\{C_i\}_{i=1}^{L}$. This is the classic problem in isolated and connected word-recognition tasks. To achieve the minimum classification error rate, the classical Bayes rule (Duda and Hart 1973) requires that

$$C^*(\mathbf{O}) = C_i \quad \text{if} \quad C_i = \arg\max_j \Pr(C_j \mid \mathbf{O}), \tag{22}$$

where $\mathbf{O}$ is the unknown observation to be classified into (recognized as) one of the L classes, $C^*(\cdot)$ denotes the decoded class of $\mathbf{O}$, and $\Pr(C_j \mid \mathbf{O})$ is the (true) a posteriori probability of C_j given the observation $\mathbf{O}$. The decision rule of (22) is the well-known MAP decoder. The decision rule of (22) is often written as

$$C^*(\mathbf{O}) = C_i \quad \text{if} \quad C_i = \arg\max_j \Pr(\mathbf{O} \mid C_j) \Pr(C_j) \tag{23}$$

in terms of the class prior $\Pr(C_j)$ and the conditional probability $\Pr(\mathbf{O} \mid C_j)$. It is clear that the difficulty in minimizing misclassification rate stems from the fact that both the prior distribution $\Pr(C_i)$ and the conditional distribution $\Pr(\mathbf{O} \mid C_i)$ are generally unknown and have to be estimated from a given, finite training set. There are practical reasons why this difficulty is hard to overcome.

For instance, to obtain reliable estimates of $\Pr(C_i)$ and $\Pr(\mathbf{O} \mid C_i)$, we generally need a sufficient size training set (i.e., large enough to adequately sample all relevant class interactions). When the vocabulary is large, this is difficult if not impossible to achieve. For example, suppose the vocabulary has 10,000 words, each of which represents a class. If we assume that 10 occurrences each, on average, are needed for reliable estimates of both $\Pr(C_i)$ and $\Pr(\mathbf{O} \mid C_i)$, this amounts to a total of $10{,}000 \times 10 = 10^5$ word utterances. (This is equivalent to ~14 hours of speech, assuming two words per second.) Therefore, some other strategy is required to estimate $\Pr(C_j)$ and $\Pr(\mathbf{O} \mid C_j)$ reliably from a smaller size training set. One possibility is to choose a set of classes to represent, instead of words, a reduced set of subword units—that is, phonemes. This greatly reduces the requirements on the amount of training data since the number of subword classes essentially does not grow with the size of vocabulary. A consequence of using subword unit classes to represent the basic set of speech sounds is that an estimate of class probability, $\Pr(C_i)$,

can be obtained directly from a lexical description of the words, independent of the *spoken* training set. Furthermore, we can estimate $\Pr(\mathbf{O} \mid C_i)$ from each of the subword units in the lexical entry for the words. This type of decomposition—namely, breaking large sound classes like words and phrases into smaller ones like subword units—leads to one particular problem in HMM speech recognition; that is, given an independent (and possibly incorrect) estimate of word probability, $\Pr^*(C_i)$, how do we estimate $\Pr^*(\mathbf{O} \mid C_i)$ such that the Bayes minimum error rate is attained? (Although we have used the example of decomposing words into subword classes, the same concept applies to high levels—that is, decomposing phrases into words.) Note that in this discussion the association between the training data $\mathbf{O}$ and the (subword) class is assumed to be known a priori (often as a result of hand labeling). We call this case the complete label case. Typical examples of complete label systems include most isolated-word and connected-word tasks and the case of hand-segmented and hand-labeled continuous speech recognition. This case is illustrated in Figure 8a, which shows the speech waveform and energy contour for a sequence of isolated digits spoken in a stationary background. It is relatively easy to define (roughly) the point in time at which each spoken digit begins and ends.

The decomposition described previously (namely, from text to words or subword units), although circumventing some of the training problems, leads to other problems—for example, the need for a detailed segmentation and exact labeling of the speech. For large-vocabulary continuous speech recognition in which the training data is extensive, this labor-intensive task cannot realistically be accomplished. Instead, we often have to rely on only partial knowledge of the data. For example, we generally know or assume we know the phoneme sequence of the words in the string, but not the direct correspondence between each phoneme and the segment of speech. We call this case the incomplete label case, and typical examples are the problems of estimating models of subword speech units from continuous speech and those of words in connected word tasks without prior word segmentation. An illustration of this case is given in Figure 8b, which shows the speech waveform and an energy contour for the sentence, "*How many ships are there*"; the boundaries for either individual words or the phonemes making up the words are not shown, nor are they known precisely.

Another problem with the estimation procedure arises when the distribution, particularly the conditional probability $\Pr(\mathbf{O} \mid C_i)$, is postulated to be the same as $\Pr(\mathbf{O} \mid \lambda_i)$, the HMM to be estimated. Since we generally choose the *form* of the observation distribution before we have any solid knowledge of the characteristics of the source in each HMM state, there is the risk of a serious mismatch between the chosen observation distribution model and the actual data source. A similar mismatch potential exists in the

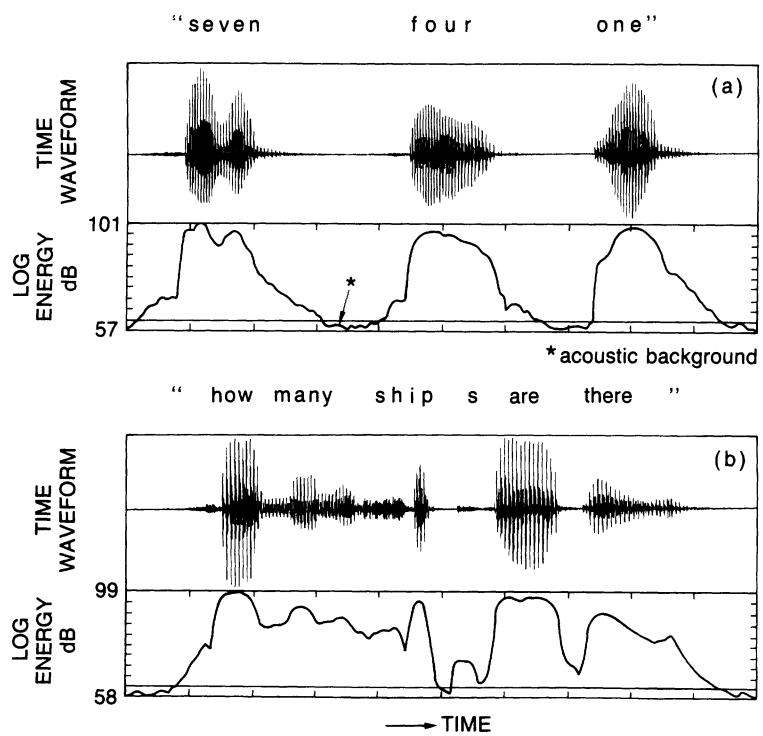

Figure 8. (a) The Waveform and Energy Contour of Three Digits Spoken in Isolation; (b) The Waveform and Energy Contour of a Naturally Spoken Sentence, "How Many Ships Are There?"

Markovian structure of the model; that is, speech signals need not be Markovian. We therefore need to provide a mechanism for compensating such potential errors in modeling. We call this problem the model-mismatch case.

Note that if the chosen class model (the conditional as well as the *a priori* probability) is indeed the correct one the ML method will lead to the asymptotically best recognition performance (in terms of highest correct classification rate) while allowing classes to be added or removed from the system specification without the need for complete retraining of all class-reference models. These assumptions, however, are rarely true in practice. An alternative to the preceding distribution estimate ideas, called *corrective training*, tries to minimize the recognition error rate directly by identifying, during training, sources that lead to recognition errors (or near errors) and applying some type of correction rule to the parameter estimation to reduce the probability of these errors or near-misses. Unlike the preceding three cases, corrective training uses the HMM as a form of *discriminant function* and is not concerned about modeling accuracy per se but more with minimizing the number of recognition errors that occur during training. A more complete discussion of these cases follows.

The Complete-Label Case. In the complete-label case, the association between the training data $\mathbf{O}$ and the class C_i is precisely known a priori during training (e.g., as shown in Fig. 8a). This is a typical situation in classical pattern-classification theory, and all the concerns about supervised learning (Duda and Hart 1973) apply. What is unique in the current speech-recognition problem, however, is the use of a prescribed class prior model $\Pr(C_i)$.

A detailed account of the issues involved in the complete-label case was given by Nadas, Nahamoo, and Picheny (1988). Consider the set of L HMM class models, $\Lambda = \{\lambda_i\}_{i=1}^{L}$, which are used to model the classes $\{C_i\}_{i=1}^{L}$, respectively. The complete set of models, Λ, defines a probability measure

$$\Pr_{\Lambda}(\mathbf{O}) = \sum_i \Pr_{\Lambda}(\mathbf{O} \mid C_i)\Pr_{\Lambda}(C_i). \tag{24}$$

The notation of (24) is slightly different from that used earlier because of the fact that we are explicitly using both classes and associated models simultaneously for classification purposes. Assume that the prior $\Pr^*_{\Lambda}(C_i) = \Pr_0(C_i)$ is given or obtained independent of the *spoken* training set $\{\mathbf{O}^{(i)}\}$, where $\mathbf{O}^{(i)}$ are the portion of the training data that are labeled as C_i. For this case, Nadas et al. (1988) proposed the use of a slightly different training measure—namely, the conditional maximum likelihood estimator, (CMLE) obtained by

$$\Lambda^*_{\text{CMLE}} = \arg\max_{\Lambda} \prod_i \Pr_{\Lambda}(C_i \mid \mathbf{O}^{(i)}). \tag{25}$$

The motivation for choosing this estimator is that the assumed model, $\Pr_{\Lambda}(\mathbf{O}^{(i)} \mid C_i)$ *or* $\Pr(\mathbf{O}^{(i)} \mid \lambda_i)$, may be inappropriate for $\Pr(\mathbf{O}^{(i)} \mid C_i)$ so that the asymptotic optimality of the standard ML estimator is no longer guaranteed. Furthermore, the prior model $\Pr_0(C_i)$ could be incorrect (poor estimate, etc.), so that the true MAP decoder of (22) is virtually impossible to implement.

When the class prior $\Pr_0(C_i)$ is obtained independent of $\{\mathbf{O}\}$ and is not part of the model Λ to be optimized, then

$$\begin{aligned}\Lambda^*_{\text{CMLE}} &= \arg\max_{\Lambda} \prod_i \Pr_{\Lambda}(C_i \mid \mathbf{O}^{(i)}) \\ &= \arg\max_{\Lambda} \sum_i \log \frac{\Pr_{\Lambda}(\mathbf{O}^{(i)}, C_i)}{\Pr_{\Lambda}(\mathbf{O}^{(i)})\Pr_0(C_i)} \\ &= \Lambda^*_{\text{MMI}},\end{aligned} \tag{26}$$

which is the well-known maximum mutual information (MMI) estimator. [Note that the second equality comes from the fact that the logarithm is monotonic and the additive term $\log \Pr_0(C_i)$ does not affect the maximization result.]

The effect of conditional ML estimation in terms of class prior robustness—that is, uncertain or incorrect $\Pr_{\Lambda}(C_i)$—can best be illustrated by the following example from Nadas et al. (1988). Suppose in the training data that there are N_i occurrences of the class C_i and, among these occurrences, the (discrete) observation $\mathbf{O}$ occurs jointly with class C_i $N_{i,\mathbf{O}}$ times. Then, the ML estimates of the prior and the conditional probabilities are, respectively,

$$\Pr^*_{\text{ML}}(C_i) = N_i \Big/ \sum_{j=1}^{L} N_j \tag{27}$$

and

$$\Pr^*_{\text{ML}}(\mathbf{O} \mid C_i) = N_{i,\mathbf{O}}/N_i. \tag{28}$$

The decoder that uses these estimates then decides that an observation $\mathbf{O}$ belongs to class C_i if

$$\frac{N_{i,\mathbf{O}}}{N_i}\frac{N_i}{\sum_{j=1}^{L} N_j} = \frac{N_{i,\mathbf{O}}}{\sum_{j=1}^{L} N_j} = \max_k \frac{N_{k,\mathbf{O}}}{\sum_{j=1}^{L} N_j}, \tag{29}$$

which is simply

$$N_{i,\mathbf{O}} = \max_k N_{k,\mathbf{O}}. \tag{30}$$

This is optimal when the total number of observations approaches infinity. When the prior $\Pr(C_i)$ is prescribed as $\Pr_0(C_i)$ rather than $\Pr^*_{\text{ML}}(C_i)$, however,

the dependence of the decision rule on the a priori probability becomes obvious if we continue to use $\Pr^*_{\mathrm{ML}}(\mathbf{O} \mid C_i)$; that is, (29) becomes

$$\frac{N_{i,\mathbf{O}}}{N_i} \cdot \Pr_0(C_i) = \max_j \frac{N_{j,\mathbf{O}}}{N_j} \Pr_0(C_j). \tag{31}$$

The CMLE criterion of (25), on the other hand, leads to a set of equations

$$\Pr^*_{\mathrm{CMLE}}(\mathbf{O} \mid C_i) = \frac{N_{i,\mathbf{O}}}{N_{\mathbf{O}}} \frac{\sum_{k=1}^{L} \Pr^*_{\mathrm{CMLE}}(\mathbf{O} \mid C_k)\Pr_0(C_k)}{\Pr_0(C_i)}, \tag{32}$$

where $N_{\mathbf{O}} = \sum_{j=1}^{L} N_{j,\mathbf{O}}$. This is obtained with the variational method by defining the Lagrangian for maximization as

$$\sum_{i=1}^{L} \sum_{\mathbf{O}} N_{i,\mathbf{O}} \log \frac{\Pr(\mathbf{O} \mid C_i)\Pr_0(C_i)}{\sum_{k=1}^{L} \Pr(\mathbf{O} \mid C_k)\Pr_0(C_k)} + \sum_{i=1}^{L} \theta_i \left[1 - \sum_{\mathbf{O}} \Pr(\mathbf{O} \mid C_i)\right],$$

where θ_i $(i = 1, 2, \ldots, L)$ are the Lagrange multipliers. Plugging (32) into the MAP rule of (22), we decide an unknown $\mathbf{O}$ to be from class C_i if

$$\frac{N_{i,\mathbf{O}}}{N_{\mathbf{O}}} \frac{\sum_{k=1}^{L} \Pr^*_{\Lambda}(\mathbf{O} \mid C_k)\Pr_0(C_k)}{\Pr_0(C_i)} \cdot \Pr_0(C_i) = \max_j \frac{N_{j,\mathbf{O}}}{N_{\mathbf{O}}} \frac{\sum_{k=1}^{L} \Pr^*_{\Lambda}(\mathbf{O} \mid C_k)\Pr_0(C_k)}{\Pr_0(C_j)} \cdot \Pr_0(C_j), \tag{33}$$

which can be reduced to

$$N_{i,\mathbf{O}} = \max_j N_{j,\mathbf{O}}, \tag{34}$$

which is independent of the prior $\Pr_0(C_i)$. In this sense, it was concluded that if $\Pr_0(C_i)$ is not the *true* prior (because of bad assumptions or estimation errors), the MLE will implement a suboptimal decoder (31), while the CMLE of (25) will lead to the correct MAP decoding result (asymptotically) because of the compensation built into the estimate of $\Pr(\mathbf{O} \mid C_i)$.

Although the criterion of CML or MMI (for training) is attractive in terms of compensation problems associated with MAP decoding, some important concerns remain unanswered. The most immediate concern in using this criterion is the lack of a convenient and robust algorithm to obtain the estimate $\Pr^*_{\mathrm{CMLE}}(\mathbf{O} \mid C_i)$. In many practical situations, the procedure for obtaining the solution may be far more complicated than (32) would imply, particularly when HMM's are involved and the observation distribution is not of a discrete type. Previous attempts (Bahl et al. 1986; Brown 1987) at using the MMI criterion have not produced an estimation procedure that is guaranteed to converge to an optimal solution either. Moreover, even though the preceding example demonstrates the robustness of CMLE against errors in the class prior, it is still not clear if CMLE is more robust than MLE when the form of the model (i.e., HMM) is incorrect for the speech source. Another problem is that the CMLE has a larger variance than the MLE. This therefore undermines the potential gain in offsetting the sensitivity due to inaccurate $\Pr_0(C_i)$ when the decoder based on finite training data is used on test data not included in the training set. In the case of insufficient training data (as is almost always the situation), there are other problems with practical implementations of the procedure.

The Incomplete-Label Case. The case of incomplete labeling arises because of (a) practical difficulties in labeling and segmenting any large continuous-speech data base and (b) the inherent ambiguity among different sound classes in terms of both class definition (i.e., inherent similarities in sound classes) and time uncertainty as realized in speech signals (i.e., it is not clear that exact segmentation boundaries between adjacent sounds universally exist; see Fig. 8b as an example). For the case of decomposing an isolated word into a prescribed phoneme sequence, we usually have a lexical description of the word in terms of the phoneme sequence (as described in a dictionary) and the spoken version of the word without any explicit time segmentation into corresponding phonemes. Under these conditions (which are typical for speech recognition), training of the prescribed subword unit models is rather difficult due to the lack of a definitive labeling relating subword classes to specific intervals of speech. After all, if we do not know for sure that a training token $\mathbf{O}$ is in sound class C_i, the likelihood function $\Pr(\mathbf{O} \mid C_i)$ cannot be defined, not to mention optimized.

There are several ways to handle the problem of incomplete labeling based on the idea of embedded decoding. One way is to retain the constraints of the known class sequence (in the previous example, the phoneme sequence) and solve for the "optimal" set of class models sequentially. Another alternative is to solve for the models of the sound classes simultaneously with the class decoding.

Consider first the case in which we have partial knowledge; that is, a given training token $\mathbf{O}$ is known to correspond to a sequence of u class labels $h = (h_1, h_2, \ldots, h_u)$ (as determined from dictionary lookup of the words realized by $\mathbf{O}$), where $h_j \in \{C_i\}_{i=1}^{L}$. The goal is to obtain the L models $\Lambda = \{\lambda_i\}_{i=1}^{L}$ corresponding to the L sound classes $C = \{C_i\}_{i=1}^{L}$, using the number of segments u and the class labels h as

hard constraints in the decoding. Figure 9 illustrates the decomposition of a word into $u = 4$, 5, or 6 segments. We see the varying segmentations associated with different numbers of base units within the word. The procedure begins by assuming a uniform segmentation $\mathbf{X} = (x_1, x_2, \ldots, x_u)$ of $\mathbf{O}$ into u speech intervals with the ith interval, x_i, corresponding to sound class h_i. (Note that each interval is a sequence of spectral vectors.) Based on this initial segmentation, the likelihood functions are defined and individual class models are obtained by ML (via the forward-backward procedure). For example, if $h_i = C_j$, then the segment x_i is used to define a likelihood function $\Pr(x_i \mid C_j)$ for maximization. As a result, a set of sound unit models are created. With the new set of unit models, we further refine the segmentation of $\mathbf{O}$ into $\mathbf{X}$ (again assuming exactly u segments) by optimally decoding $\mathbf{O}$ using the Viterbi algorithm. This leads to an improved segmentation of $\mathbf{O}$ that can then be used to give a refined set of sound models. This process is iterated until a reasonably stationary segmentation of $\mathbf{O}$ into intervals $\mathbf{X}$ is obtained.

This constrained decoding approach to the incomplete label case has been used in explicit acoustic modeling of phonemic classes (Lee et al. 1989; Lee, Rabiner, Pieraccini, and Wilpon 1990) with good success. There are, however, some theoretical shortcomings to the method. One problem is that the segmentation/decoding results will be different for different numbers of segments u in the given string (see Fig. 9). Thus even the simple expedient of having multiple dictionary definitions for a word can lead to inconsistent segmentation in terms of sound classes. Although the procedure is practical, there appears to be room for theoretical improvements, which in turn may prove beneficial in practical implementations.

An alternative and probably more thorough way of handling the incomplete-label problem is to combine the ideas of segmentation, decoding, and modeling together and try to solve a large network for both the class models and segmentations simultaneously, without any prescribed label-sequence constraint. In this approach, the preceding iterative procedure—that is, recursively and interleavingly improving data segmentation and model estimation via the sequential k-means algorithm—is again used. The key difference, in contrast to the constrained decoding approach discussed previously, is that the label-sequence constraint is no longer retained in each iteration. This allows globally optimal decoding of $\mathbf{O}$, given a set of sound unit models. This advantage, however, comes at the price of giving up the convenient and readily available lexical representation h from the dictionary.

As previously pointed out, the goal of signal modeling is to come up with a parsimonious, consistent representation of the source that displays a certain kind of variation in the observed output. Speech signals are known to have inherent time uncertainty due to speaking style and speaking rate variations, as well as spectral uncertainty because of coarticulation effects, individual speaker characteristics, and so forth. This justifies the use of HMM for an individual sound class. The reason is that the incomplete data problem formulation (Dempster et al. 1977; Rabiner and Juang 1986), based on which the fundamental HMM probability measure of (8) is defined, is particularly appropriate when explicit knowledge of the exact position (labeling) of the particular sound in the middle of an utterance is lacking. On the other hand, speech is a linear code (Chao 1968) in the sense that decoded symbols come out one after another and no two symbols can appear at the same time. Therefore, the objective of optimizing the quantity defined by (7) will have to be accomplished during the recognition process, and alternate

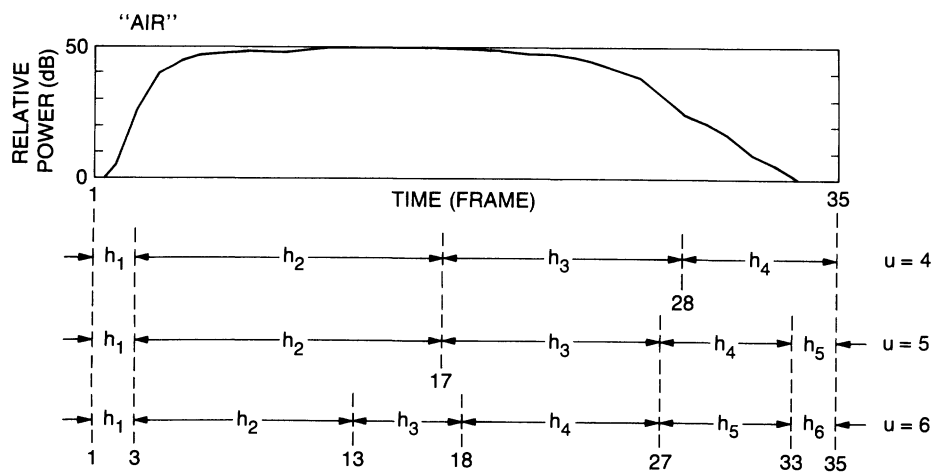

Figure 9. Decomposition of the Word "Air" Into Different Numbers of Segment Units.

use of the Baum–Welch algorithm and the segmental k-means algorithm as laid out by Lee et al. (1989) appears to be a reasonable approach to the problems associated with incomplete-label cases.

Model-Mismatch Issues. In speech modeling, we often choose (assume) the form of model before we actually know enough about the characteristics of the source. When we choose $\Pr(\mathbf{O} \mid \lambda)$ to model $\Pr(\mathbf{O} \mid C)$ for class C and perform ML estimation, the optimality in decoding is meaningful only when $\mathbf{O}$ is indeed generated by the source λ. When the actual source is inconsistent with λ, we need to either improve $\Pr(\mathbf{O} \mid \lambda)$ based on $\mathbf{O}$ or revise the decoding rule in some way. Here, we discuss the first possibility.

Consider two statistical populations with probability densities f_1 and f_2, respectively. The Kullback–Leibler number, cross entropy, I divergence, or *discrimination information* (Good 1963; Hobson and Cheng 1973; Johnson 1979; Kullback 1958) defined by

$$I(f_1:f_2) = I(1:2) = \int f_1(\mathbf{O}) \log \frac{f_1(\mathbf{O})}{f_2(\mathbf{O})}\, d\zeta(\mathbf{O}) \tag{35}$$

is the mean information for discriminating f_1 against f_2. This discrimination information can be used in HMM (Ephraim, Dembo, and Rabiner 1989).

Let $R = (R_1, R_2, \ldots, R_T)$ be a sequence of constraints associated with the observation sequence $\mathbf{O} = (\mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_T)$, which we attempt to model. For example, R could be considered as the autocorrelation of the actual speech data. Let $\Omega(R)$ be the set of all models (probability distributions) that satisfy the constraints R. The minimum discrimination information (MDI) approach to HMM modeling is to find the HMM parameter set λ which minimizes the MDI defined by

$$\begin{aligned} v(R, P) &\triangleq \inf_{Q \in \Omega(R)} I(Q:P) \\ &= \inf_{Q \in \Omega(R)} \int q(\mathbf{O}) \log \frac{q(\mathbf{O})}{p(\mathbf{O} \mid \lambda)}\, d\mathbf{O}. \end{aligned} \tag{36}$$

The estimation criterion is thus the minimum over λ of the MDI of (36). Note that in (36) we have assumed that both distributions P and Q have generalized pdf's, denoted by $p(\cdot \mid \lambda)$ and $q(\cdot)$, respectively. Clearly, the idea of (minimum) MDI model design is to allow a search into a general set of models (distributions) Ω with the HMM measure $\Pr(\mathbf{O} \mid \lambda)$ as a contrasting reference so that the form of the chosen model $P(\cdot \mid \lambda)$ and what is revealed in the given data $\mathbf{O}$ can be better matched under the MDI measure.

MDI modeling involves two steps. First, given an HMM $p(\cdot \mid \lambda)$, we find the solution, q, to the MDI problem of (36). Then, with q given, we optimize the parameter values λ such that $v(R, P)$ is minimized.

The MDI modeling criterion carries some information-theoretic justification and can be related to the MMI approach (Ephraim and Rabiner 1988). There remain several unresolved issues with this method, however. The objective to improve the match between the observations $\mathbf{O}$ and the model λ is embedded in $v(R, P)$, which is computed for all $Q \in \Omega(R)$ with respect to the *chosen* model $P(\cdot \mid \lambda)$. Since $P(\cdot \mid \lambda)$ is often not a good measure of how well the model matches the data because of model mismatch, the goodness of the MDI pdf, q, which minimizes $v(R, P)$, is difficult to measure in recognition tasks, not to mention the effects on classification errors. Furthermore, the nice computational structure of the HMM that makes possible the modeling of long training sequences is not preserved in the MDI pdf, q, resulting in an exceptionally high computational complexity. In practice, some approximations are made to reduce the complexity problem (Ephraim et al. 1989). To date, it has not been demonstrated that the MDI model-design procedure can bring about significant improvements in recognition performance (Ephraim et al. 1989). It is also not clear whether MDI (for model correctness) and CMLE/MMIE (for robust decoding) can be easily jointly formulated for an improved recognizer design.

Corrective Training. As explained earlier, the minimum Bayes risk or error rate is the theoretical recognizer performance bound conditioned on the exact knowledge of both the prior and the conditional distributions. When both distributions are not known exactly and the classifier needs to be designed based on a finite training set, there are several ways to try to reduce the error rate. One method is based on the theoretical link between discriminant analysis and distribution estimation (Duda and Hart 1973). The idea here is to design a classifier (discriminant function) such that the minimum classification-error rate is attained on the training set. In particular, we wish to design a classifier that uses estimates of $\Pr(C_i)$ and $\Pr(\mathbf{O} \mid C_i)$ and that achieves a minimum error rate for the training set in the same way a discriminant function is designed. The reason for using the HMM $\Pr(\mathbf{O} \mid \lambda_i)$ as opposed to other discriminant functions, is to exploit the strengths of the HMM—namely, consistency, flexibility and computational ease, as well as its ability to generalize classifier performance to independent (open) data sets. The generalization capability of HMM's, as discriminant

functions, is somewhat beyond the scope of this article and will not be discussed here. In the following, we focus on the issue of treating the estimation of the distributions $\Pr(C_i)$ and $\Pr(\mathbf{O} \mid \lambda_i)$ as a discriminant-function design to attain the minimum error rate.

Bahl, Brown, de Souza, and Mercer (1988a) were the first to propose an error-correction strategy, which they named corrective training, to specifically deal with the misclassification problem. Their training algorithm was motivated by analogy with an error-correction training procedure for linear classifiers (Nilsson 1965). In their proposed method, the observation distribution is of a discrete type, $B = [b_{ij}]$, where b_{ij} is the probability of observing a vector quantization code index (acoustic label) j when the HMM source is in state i. Each b_{ij} is obtained via the forward-backward algorithm as the weighted frequency of occurrence of the code index (Rabiner et al. 1983). The corrective training algorithm of Bahl et al. (1988a) works as follows. First, use a labeled training set to estimate the parameters of the HMM's $\Lambda = \{\lambda_i\}$ with the forward-backward algorithm. For each utterance $\mathbf{O}$, labeled as C_k, for example, evaluate $\Pr(\mathbf{O} \mid \lambda_k)$ for the correct class C_k and $\Pr(\mathbf{O} \mid \lambda_l)$ for each incorrect class C_l. (The evaluation of likelihood for the incorrect classes need not be exhaustive.) For every utterance in which $\log \Pr(\mathbf{O} \mid \lambda_l) > \log \Pr(\mathbf{O} \mid \lambda_k) - \delta$, where δ is a prescribed threshold, modify λ_k and λ_l according to the following mechanism: (a) Apply the forward-backward algorithm to obtain estimates b'_{ij} and b''_{ij}, using the labeled utterance $\mathbf{O}$ only, for the correct class C_k and incorrect class C_l, respectively, and (b) modify the original b_{ij} in λ_k to $b_{ij} + \gamma b'_{ij}$ and the b_{ij} in λ_l to $b_{ij} - \gamma b''_{ij}$. When the state labels are tied for certain models, the preceding procedure is equivalent to replacing the original b_{ij} with $b_{ij} + \gamma(b'_{ij} - b''_{ij})$. The prescribed adaptation parameter, γ, controls the "rate of convergence," and the threshold, δ, defines the "near-miss" cases. This corrective training algorithm therefore focuses on those parts of the model that are most important for word discrimination, a clear difference from the ML principle.

Although Bahl et al. (1988a) reported that the corrective training procedure worked better (in isolated-word recognition tasks) than models obtained using the MMI or the CMLE criteria, the lack of a rigorous analytical foundation for the algorithm is one problem. Without a better theoretical understanding of the algorithm, the appeal of the method is primarily experimental. Other attempts to design HMM's for minimum error rate or some form of class separation include the work by Ljolje, Ephraim, and Rabiner (1990) and by Sondhi and Roe (1983).

Several other forms of discriminative training were also proposed by Katagiri, Lee, and Juang (1990), who combined the adaptive learning concept in learning vector quantizer design (Kohonen 1986) and the corrective training method described previously, leading to a framework for the analysis of related training/learning ideas. Although this work is still in progress, the key result is that these adaptive learning methods can be formulated as a general-risk minimization procedure based on a probabilistic descent algorithm, ensuring a stochastic convergence result. The generalized risks considered by Katagiri et al. (1990) include the regular classification error, the mean squared error, nonlinear functions (e.g., sigmoid) of the likelihood, and several other general measures. The corrective training algorithm of Bahl et al. (1988a) is just one possible choice for the minimization of a prescribed risk function.

4.2 Integration of Nonspectral and Spectral Features

The use of HMM's in speech recognition has been mostly limited to the modeling of short-time speech spectral information; that is, the observation $\mathbf{O}$ typically represents a smoothed representation of the speech spectrum at a given time. The spectral feature vector has proved extremely useful and has led to a wide variety of successful recognizer designs. This success can be attributed both to the range of spectral-analysis techniques developed in the past three decades, as well as to our understanding of the perceptual importance of the speech spectrum to the recognition of sounds. The success of spectral parameters for characterizing speech was further augmented by the introduction of the so-called delta-cepstrum (Furui 1986), which attempts to model the differential speech spectrum.

Besides spectral parameters, there are other speech features that are believed to contribute to the recognition and understanding of speech by humans. One such category of nonspectral speech features is prosody as it is manifested on both the segmental and the supra-segmental level (Lea 1980). Physical manifestations of prosody in the speech signal include signal energy (suitably normalized), differential energy, and signal pitch (fundamental frequency). There are at least two issues of concern in integrating nonspectral with spectral features in a statistical model of speech: Do such features contribute to the performance of the statistical model for actual recognition tasks, and how should the features be integrated so as to be consistent with their physical manifestations in the production and perception of speech? The first issue is relatively easy to resolve based on experimental results. Several HMM-based recognition systems have incorporated

log energy (and differential log energy) either directly into the feature vector or as an additional feature whose probability (or likelihood) is factored into the likelihood calculation (Bahl et al. 1983; Rabiner 1984; Shikano 1985) with moderate success (i.e., higher recognition performance). The level of performance improvement, however, is considerably smaller than one might anticipate based on the importance of prosody in speech.

The problem with combining spectral and nonspectral features in a statistical framework is one of temporal rate of change. To attain adequate time resolution for the spectral parameters that characterize the varying vocal tract, we need to sample the spectral observations at a rate on the order of 100 times per second (10-msec frame update). The prosodic features of speech characterize stress and intonation, and these occur at a syllabic rate of about 10 times per second. [Of course, we can always oversample the features associated with the prosodic parameters to keep the rate the same as that of the spectral parameters; this in fact is what is currently done in most systems—e.g., Rabiner (1984) and Shikano (1985)]. Furthermore, the sequential characteristic change of different features may be too different to warrant a single, unified Markov chain. Thus one key question is: How do we combine two feature sets with fundamentally different time scales and possibly different sequential characteristics so as to be able to perform optimum modeling and decoding?

A second problem in combining fundamentally different (in nature) features concerns their statistical characterization. To be technically correct, we need to know the joint distribution of the two feature sets. For statistically independent (or often just uncorrelated) feature sets, we can represent the joint density as a product of the individual densities. In practice, however, there is usually some correlation between any pair of speech feature sets; hence some correction for the correlation is usually required. One proposed method of handling this problem is to perform a principal-component analysis on the joint feature set before hidden Markov modeling is performed (Bocchieri and Doddington 1986). Although this alleviates the difficulties somewhat, it is not a totally satisfying solution because the resulting feature set usually has no straightforward physical significance. Furthermore, the set of principal components is a function of the data and hence need not be optimal for unseen data (open-set problem).

4.3 Duration Modeling in HMM's

One of the inherent limitations of the HMM approach is its treatment of temporal duration. Inherently, within a state of an HMM, the probability distribution of state duration is exponential; that is, the probability of staying in state i for exactly d observations is $\Pr_i(d) = (a_{ii})^{d-1}(1 - a_{ii})$. This exponential duration model is inappropriate for almost any speech event.

Several alternatives for implementing different state duration models have been proposed. The most straightforward approach is the concept of a semi-Markov chain (Ferguson 1980; Russell and Moore 1985) in which state transitions do not occur at regular time intervals. More formally, we assume that for a given state sequence $\mathbf{q}$ in which there are $r - 1$ state transitions such that the states visited are $q_1, q_2, \ldots, q_r$ with associated state durations of $d_1, d_2, \ldots, d_r$ (frames), the joint probability of the observation sequence $\mathbf{O}$ and the state sequence $\mathbf{q}$, given the model λ, becomes

$$\begin{aligned}\Pr(\mathbf{O}, \mathbf{q} \mid \lambda) &= \pi_{q_1}\Pr_{q_1}(d_1)\Pr(\mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_{d_1} \mid q_1) \\ &\quad a_{q_1q_2}\Pr_{q_2}(d_2)\Pr(\mathbf{O}_{d_1+1} \cdots \mathbf{O}_{d_1+d_2} \mid q_2) \\ &\quad \vdots \\ &\quad a_{q_{r-1}q_r}\Pr_{q_r}(d_r)\Pr(\mathbf{O}_{d_1+d_2+\cdots+d_{r-1}+1} \cdots \mathbf{O}_T \mid q_r).\end{aligned} \tag{37}$$

The probability measure for such a Markov model is, accordingly, $\Pr(\mathbf{O} \mid \lambda) = \sum_q \Pr(\mathbf{O}, \mathbf{q} \mid \lambda)$. Based on these definitions, modeling of the source, including the duration model $\Pr_i(d_i)$, can be implemented using a hill-climbing reestimation procedure of the type used previously (Levinson 1986). Typically, $\Pr_i(d_i)$ is treated as a discrete distribution over the range $1 \le d_i \le D_{\max}$, where $D_{\max}$ represents the longest possible dwell in any state.

Although preceding formulation handles the duration model *simultaneously* with the Markovian transition and the local (state) distribution models and can lead to analytical solutions, there are drawbacks with such a duration model for speech recognition. One such drawback is the greatly increased computational complexity due to the loss of regularity in transition timing. In the traditional HMM, transitions are allowed to occur at every time frame. In the semi-Markov model, transitions do not occur at regular time [the return transition is part of the duration model $\Pr_i(d_i)$], and this leads to a significantly more complicated lattice for decoding an input string. It was estimated by Rabiner (1989) that the semi-Markov model incurs a factor of 300 increase in computational complexity for a value of $D_{\max} = 25$. The much increased complication in decoding lattice often renders many search algorithms such as the beam search (Lowerre and Reddy 1980) and the stack algorithm (Jelinek 1969) for handling large

problems extremely difficult to implement. Another problem with the semi-Markov model is the large number of parameters ($D_{\max}$) associated with each duration model that must be estimated in addition to the usual HMM parameters. Finally, it is not clear how accurate or how robust the estimated $\Pr_i(d_i)$ needs to be in order to be beneficial in speech recognition.

One proposal (Levinson 1986) to alleviate some of these problems is to use a parametric state-duration distribution model instead of the nonparametric ones used previously. Several parametric distributions have been considered, including the Gaussian family (with constraints to avoid negative durations) and the gamma family. Other less involved attempts include modeling the state duration with a uniform distribution, requiring only the estimate of the duration range, with minimum and maximum duration allowed in a particular state. This simple duration model has been applied with good success (Lowerre and Reddy 1980) for some tasks.

The major difficulty in modeling the durational information is that it is much more sparse than spectral information; that is, there is only one duration per state. Hence we either alter the structure of the HMM (e.g., to a semi-Markov model), thereby losing much of the regularity of the original HMM formulation, or we seek alternative implementation structures, as in the case of prosodic information. For durational information a simple-minded approach is to treat the spectral modeling and the duration modeling as two separate, loosely connected problems. Hence the regular HMM estimation (spectral) is performed on the given observation sequence $\mathbf{O}$. Then the best state sequence $\bar{\mathbf{q}}$, which maximizes $\Pr(\mathbf{O}, \mathbf{q} \mid \lambda)$, is found using the Viterbi algorithm. Finally, esimates of $\Pr_i(d_i)$ are obtained based on the optimal state sequence $\bar{\mathbf{q}}$ by either the ML method or from simple frequency of occurrence counts (Rabiner et al. 1986). Often the duration d_i for state i is normalized by the overall duration T to account for the inherent variation in speaking rate. This approach is usually called the postprocessor duration model because the standard decoding is performed first and the duration information is only available after the initial processing is finished. Although the postprocessor duration model has had some success (Rabiner et al. 1986), the questions of optimality of the estimate, robustness of the solution, and other criteria for successful use of duration information, especially as applied to speech recognition, remain unanswered.

4.4 Model Clustering and Splitting

One of the basic assumptions in statistical modeling is that the variability in the observations from an information source can be modeled by statistical distributions. For speech recognition, the source could be a single word, a subword unit like a phoneme, or a word sequence. Because of variability in the source production (e.g., accents, speed of talking) or the source processing (e.g., transmission distortion, noise), it is often expedient to consider using more than a single HMM to characterize the source. There are two motivations behind this multiple HMM approach. First, lumping all of the variability together from inhomogeneous data sources leads to unnecessarily complex models, often yielding lower modeling accuracy. Second, some of the variability, or rather the inhomogeneity in the source data, may be known a priori, thus warranting separate modeling of the source data sets. Here, our main concern is the first case—that is, automatic modeling of an inhomogeneous source with multiple HMM's, because the latter (manual) case is basically straightforward.

Several generalized clustering algorithms exist, such as the k-means clustering algorithm, the generalized Lloyd algorithm widely used in vector quantizer designs (Linde, Buzo, and Gray 1980) or the greedy growing algorithm found in set-partition or decision-tree designs (Breiman 1984), all of which are suitable for the purpose of separating inconsistent training data so that each divided subgroup becomes more homogeneous and therefore is better modeled by a single HMM. The nearest neighbor rule required in these clustering algorithms is simply to assign an observation sequence $\mathbf{O}$ to cluster i if $\Pr(\mathbf{O} \mid \lambda_i) = \max_j \Pr(\mathbf{O} \mid \lambda_j)$, where λ_j's denote the models of the clusters. Successful application of the model-clustering algorithms to the speech-recognition problem, using the straightforward ML criterion, has been reported (Rabiner, Lee, Juang, and Wilpon 1989). When other estimation criteria are used, however, the interaction between multiple HMM modeling and the Bayes minimum-error-classifier design remains an open question in need of further study.

An alternative to model clustering is to arbitrarily subdivide a given speech source into a large number of subclasses with specialized characteristics and then consider a generalized procedure for model merging based on source likelihood considerations. By way of example, for large-vocabulary speech recognition we often try to build specialized units (context sensitive) for recognition. We could consider building units that are a function of the sound immediately preceding the unit (left-context) and the sound immediately following the unit (right-context). There are about 10,000 such units in English. Many of the units are functionally almost identical. The problem is how to determine which pairs of units should be merged (so that the number of model units is manageable and the variance of the parameter estimate

is reduced). To set ideas, consider two unit models, λ_a and λ_b, corresponding to training observation sets $\mathbf{O}_a$ and $\mathbf{O}_b$, and the merged model λ_{a+b}, corresponding to the merged observation sets $\{\mathbf{O}_a, \mathbf{O}_b\}$. We can then compute the change in entropy (i.e., loss of information) resulting from the merged models as

$$\begin{aligned}\Delta H_{ab} &= H_a + H_b - H_{a+b} \\ &= -\Pr(\mathbf{O}_a \mid \lambda_a)\log \Pr(\mathbf{O}_a \mid \lambda_a) \\ &\quad - \Pr(\mathbf{O}_b \mid \lambda_b)\log \Pr(\mathbf{O}_b \mid \lambda_b) \\ &\quad + \Pr(\{\mathbf{O}_a, \mathbf{O}_b\} \mid \lambda_{a+b})\log \Pr(\{\mathbf{O}_a, \mathbf{O}_b\} \mid \lambda_{a+b})\end{aligned}$$

(Lee 1989). Whenever ΔH_{ab} is small enough, it means that the change in entropy resulting from merging the models will not affect system performance (at least on the training set) and the models can be merged. The question of how small is acceptable is dependent on specific applications. Other practical questions of a similar nature also remain.

4.5 Parameter Significance and Statistical Independence

Although in theory the importance of the state transition coefficients a_{ij} is the same as that of the state observation density $b_j(\mathbf{O})$ [in training, a_{ij} affects the parameter estimate of $b_j(\cdot)$ and vice versa], in normal practice of speech recognition, this is not usually the case. This paradox is due to the discrimination capability of the a_{ij}'s relative to that of the $b_j(\mathbf{O})$'s. Since a_{ij} is the relative frequency of state transition from state i to state j, its dynamic range (especially in a left-to-right model) is severely constrained. The $b_j(\mathbf{O})$ densities, nevertheless, often have almost unlimited dynamic range, particularly when continuous density functions are used, as each of the state densities is highly localized in the acoustic parameter space. When we combine the a_{ij}'s and the $b_j(\mathbf{O})$'s to give the probability distribution of the HMM, we find in practice that, for a left-to-right model, the a_{ij}'s can be neglected entirely (i.e., set all $a_{ii} = a_{i,i+1} = .5$) with no effect on recognition performance.

The preceding analysis points out the existence of the unbalanced numerical significance of the a's and the b's in the likelihood calculation of the HMM as applied to speech recognition. This, however, should not be taken as to totally discredit the usefulness of the Markov-chain contribution in terms of signal modeling. In fact, the transition probability still plays an important role in parameter estimation.

The introduction of semi-Markov models as discussed in Section 3.7 is one way to enhance the significance of the Markov-chain contribution. The probability of a Markovian sequence of (4) is revised to a general form of product $\Pr_{q_1}(d_1)\Pr_{q_2}(d_2)\ldots$ as in (37). This more general form of Markov-chain dependence allows introduction of a singularity in $\Pr_q(d)$, thereby increasing its numerical significance under certain conditions. [For example, if we specify $\Pr_q(d) = 0$ for $d \leq d_0$, the HMM system has to undergo at least d_0 frames in state q before it moves out of that state, regardless of the value of $b_q(\mathbf{O}_t)$ during that period of time.] Such a provision for introducing singularities could have a major impact on the recognition performance (Lowerre and Reddy 1980).

A separate issue associated with the observation densities is the statistical independence assumption. If the state sequence that led to the production of $\mathbf{O}$ is known, the conditional probability of $\mathbf{O}$ as defined in (6) involves a product form that implies statistical independence (Sondhi, Levinson, and de la Noue 1988; Wellekens 1987). The partial HMM measure of (7) shows that the contribution of the underlying Markov chain [as expressed in (4)] is to multiply the result of (6), a result equivalent to assuming that the observations within a state are independent. This greatly simplified assumption may be an area for improvement. The argument is that within a state (especially one representing a stationary sound like a vowel) the observations are highly correlated. Thus assuming independence gives far too much emphasis to the match within such stationary states. The problem is one of form as well as substance. It is difficult to choose an appropriate form for $\Pr(\cdot \mid \mathbf{q}, \lambda)$, and it is even more difficult to estimate the parameters of the chosen form from any reasonably sized training set. The problems are similar to those of any reduced-rate measurement—namely, one measurement of $\Pr(\cdot \mid \mathbf{q}, \lambda)$ for each state rather than several measurements per state. Hence we are trying to estimate parameters of a much more complicated representation from far less data than in the simple, uncorrelated observations case. As such, this problem is one for which there may never be a good solution.

4.6 Higher Order HMM

Until now, almost all HMM formulations for speech recognition have been based on a simple first-order Markov chain. At the acoustic signal-processing level, this low-order modeling may be acceptable because the time scale of processing for each frame of signal is kept on the order of 10 msec. When an HMM is used in higher levels of a recognition system, such as syntactic or semantic processing, the first-order formulation often turns out to be inadequate. This is because the grammatical structure of the English language (and perhaps any other language) obviously

cannot be properly modeled by a first-order Markov chain. Although the structural simplicity of a first-order model makes the computation simple and straightforward, there may be a need to complete the analytical framework of higher order models. Moreover, for such higher order models to be practically useful, many of the implementational advantages of the first-order case may have to be formulated in an appropriate manner.

5. SUMMARY

In this article, we have reviewed the statistical method of HMM's. We showed that the strengths of the method lie in the consistent statistical framework that is flexible and versatile, particularly for speech applications, and the ease of implementation that makes the method practically attractive. We also pointed out some areas of the general HMM method that deserve more attention with the hope that increased understanding will lead to performance improvements for many applications. These areas include the modeling criteria, particularly the problem of minimum classification error, incorporation of new (nonspectral) features as well as prior linguistic knowledge, and the modeling of state durations and its use in speech recognition. With our current understanding, HMM systems have been shown capable of achieving recognition rates of more than 95% word accuracy in certain speaker-independent tasks with vocabularies on the order of 1,000 words. With further progress, it is not difficult to foresee HMM-based recognition systems that perform well enough to make the technology usable in everyday life.

ACKNOWLEDGMENTS

We acknowledge the excellent comments and criticisms provided by Yariv Ephraim and C. H. Lee on earlier drafts of this article.

[*Received October 1990. Revised March 1991.*]

REFERENCES

Allen, J. B., and Rabiner, L. R. (1977), "A Unified Theory of Short-time Spectrum Analysis and Synthesis," *Proceedings of IEEE*, 65, 11, 1558–1564.

Atal, B. S., and Hanauer, S. L. (1971), "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *Journal of the Acoustical Society of America*, 50, 2, 637–655.

Averbuch, A., et al. (1987), "Experiments With the TANGORA 20,000 Word Speech Recognizer," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, New York: IEEE, pp. 701–704.

Bahl, L. R., Bakis, R., Cohen, P. S., Cole, A. G., Jelinek, F., Lewis, B. L., and Mercer, R. L. (1980), "Further Results on the Recognition of a Continuously Read Natural Corpus," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, New York: IEEE, pp. 872–875.

Bahl, L. R., Brown, P. F., de Souza, P. V., and Mercer, R. L. (1986), "Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, New York: IEEE, pp. 49–52.

——— (1988a), "A New Algorithm for the Estimation of Hidden Markov Model Parameters," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, New York: IEEE, pp. 493–496.

——— (1988b), "Speech Recognition With Continuous Parameter Hidden Markov Models," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, New York: IEEE, pp. 40–43.

Bahl, L. R., Jelinek, F., and Mercer, R. L. (1983), "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5, 179–190.

Baker, J. K. (1975), "The DRAGON System—An Overview," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 23, 24–29.

Bakis, R. (1976), "Continuous Speech Word Recognition Via Centisecond Acoustic States," unpublished paper presented at the meeting of the Acoustics Society of America, Washington, DC, April.

Baum, L. E. (1972), "An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes," *Inequalities*, 3, 1–8.

Baum, L. E., and Egon, J. A. (1967), "An Inequality With Applications to Statistical Estimation for Probabilistic Functions of a Markov Process and to a Model for Ecology," *Bulletin of the American Meteorological Society*, 73, 360–363.

Baum, L. E., and Petrie, T. (1966), "Statistical Inference for Probabilistic Functions of Finite State Markov Chains," *The Annals of Mathematical Statistics*, 37, 1554–1563.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970), "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *The Annals of Mathematical Statistics*, 41, 164–171.

Baum, L. E., and Sell, G. R. (1968), "Growth Functions for Transformations on Manifolds," *Pacific Journal of Mathematics*, 27, 211–227.

Bocchieri, E. L., and Doddington, G. R. (1986), "Frame-Specific Statistical Features for Speaker Independent Speech Recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34, 755–764.

Bourlard, H., Kamp, Y., Ney, H., and Wellekens, C. J. (1985), "Speaker-Dependent Connected Speech Recognition Via Dynamic Programming and Statistical Methods," in *Speech and Speaker Recognition*, ed. M. R. Schroeder, Basel, Switzerland: Karger, pp. 115–148.

Breiman, L. (1984), *Classification and Regression Trees*, Monterey, CA: Wadsworth.

Bridle, J. S. (1984), "Stochastic Models and Template Matching: Some Important Relationships Between Two Apparently Different Techniques for Automatic Speech Recognition," in *Proceedings of the Institute of Acoustics, Autumn Conference*, pp. 1–8.

Brown, P. F. (1987), "The Acoustic-Modelling Problem in Automatic Speech Recognition," unpublished Ph.D. thesis, Carnegie Mellon University, Dept. of Computer Science.

Cadzow, J. A. (1982), "ARMA Modeling of Time Series," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4, 124–128.

Chao, Y. R. (1968), *Language and Symbolic Systems*, Cambridge, U.K.: Cambridge University Press.

Chow, Y. L., et al. (1987), "BYBLOS: The BBN Continuous Speech Recognition System," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, New York: IEEE, pp. 89–92.

Cohen, J. (1985), "Application of an Adaptive Auditory Model to Speech Recognition," unpublished paper presented at the 110th Meeting of Acoustical Society of America, Nashville, Tennessee, November 4–8.

Dautrich, B. A., Rabiner, L. R., and Martin, T. B. (1983), "On the Effects of Varying Filter Bank Parameters on Isolated Word Recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 31, 793–806.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 1, 1–38.

Derouault, A. M. (1987), "Context Dependent Phonetic Markov Models for Large Vocabulary Speech Recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, New York: IEEE, pp. 360–363.

Duda, R. O., and Hart, P. E. (1973), *Pattern Classification and Scene Analysis*, New York: John Wiley.

Ephraim, Y., Dembo, A., and Rabiner, L. R. (1989), "A Minimum Discrimination Information Approach for Hidden Markov Modeling," *IEEE Transactions on Information Theory*, 35, 1001–1013.

Ephraim, Y., and Rabiner, L. R. (1988), "On the Relations Between Modeling Approaches for Information Sources," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, New York, IEEE, pp. 24–27.

Ferguson, J. D. (1980), "Hidden Markov Analysis: An Introduction," in *Hidden Markov Models for Speech*, ed. J. D. Ferguson, Princeton, NJ: Institute for Defense Analyses, pp. 8–15.

Forney, G. D. (1973), "The Viterbi Algorithm," *Proceedings of the IEEE*, 61, 268–278.

Furui, S. (1986), "Speaker Independent Isolated Word Recognition Based on Dynamics Emphasized Cepstrum," *Transactions of the Institute of Electronics and Communication Engineers*, 69, 1310–1317.

Ghitza, O. (1986), "Auditory Nerve Representation as a Front-end for Speech Recognition in a Noisy Environment," *Computer Speech and Language*, 1, 109.

Good, I. J. (1963), "Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables," *The Annals of Mathematical Statistics*, 34, 911–934.

Gupta, V. N., Lennig, M., and Mermelstein, P. (1987), "Integration of Acoustic Information in a Large Vocabulary Word Recognizer," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, New York: IEEE, pp. 697–700.

Hobson, A., and Cheng, B. (1973), "A Comparison of the Shannon and Kullback Information Measures," *Journal of Statistical Physics*, 7, 301–310.

Huang, X., and Jack, M. A. (1989), "Unified Techniques for Vector Quantization and Hidden Markov Models Using Semi-continuous Models," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, New York: IEEE, pp. 639–642.

Jayant, N. S., and Noll, P. (1984), *Digital Coding of Waveforms*, Englewood Cliffs, NJ: Prentice-Hall.

Jelinek, F. (1969), "A Fast Sequential Decoding Algorithm Using a Stack," *IBM Journal of Research and Development*, 13, 675–685.

——— (1976), "Continuous Speech Recognition by Statistical Methods," *Proceedings of the IEEE*, 64, 532–536.

Jelinek, F., Bahl, L. R., and Mercer, R. L. (1975), "Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech," *IEEE Transactions on Information Theory*, 21, 250–256.

Jelinek, F., and Mercer, R. L. (1980), "Interpolated Estimation of Markov Source Parameters From Sparse Data," in *Pattern Recognition in Practice*, eds. E. S. Gelsema and L. N. Kanal, Amsterdam: North-Holland, pp. 381–397.

Johnson, R. W. (1979), "Determining Probability Distributions by Maximum Entropy and Minimum Cross-entropy," in *Proceedings of APL79* (ACM 0-89791-005), pp. 24–29.

Juang, B. H. (1984), "On the Hidden Markov Model and Dynamic Time Warping for Speech Recognition—A Unified View," *AT&T Technical Journal*, 63, 1213–1243.

——— (1985), "Maximum Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains," *AT&T Technical Journal*, 64, 1235–1249.

Juang, B. H., and Rabiner, L. R. (1985), "Mixture Autoregressive Hidden Markov Models for Speech Signals," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33, 1404–1413.

——— (1990), "The Segmental k-Means Algorithm for Estimating Parameters of Hidden Markov Models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38, 1639–1641.

Juang, B. H., Rabiner, L. R., Levinson, S. E., and Sondhi, M. M. (1985), "Recent Developments in the Application of Hidden Markov Models to Speaker-Independent Isolated Word Recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, New York: IEEE, pp. 9–12.

Juang, B. H., Rabiner, L. R., and Wilpon, J. G. (1987), "On the Use of Bandpass Liftering in Speech Recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35, 947–954.

Katagiri, S., Lee, C. H., and Juang, B. H. (1990), "A Theory on Adaptive Learning for Pattern Classification," unpublished manuscript.

Kohonen, T. (1986), "Learning Vector Quantization for Pattern Recognition," Report TKK-F-A601, Helsinki University of Technology.

Kullback, S. (1958), *Information Theory and Statistics*, New York: John Wiley.

Lea, W. A. (1980), "Prosodic Aids to Speech Recognition," in *Trends in Speech Recognition*, ed. W. A. Lea, Englewood Cliffs, NJ: Prentice-Hall, pp. 116–205.

Lee, C. H., Juang, B. H., Soong, F. K., and Rabiner, L. R. (1989), "Word Recognition Using Whole Word and Subword Models," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, New York: IEEE, pp. 683–686.

Lee, C. H., Lin, C. H., and Juang, B. H. (1991), "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 39, 4, 806–814.

Lee, C. H., Rabiner, L. R., Pieraccini, R., and Wilpon, J. G. (1990), "Acoustic Modeling for Large Vocabulary Speech Recognition," *Computer Speech and Language*, 4, 127–165.

Lee, C. H., Soong, F. K., and Juang, B. H. (1988), "A Segment Model Based Approach to Speech Recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, New York: IEEE, pp. 501–504.

Lee, K. F. (1989), "Automatic Speech Recognition, the Development of the SPHINX System," Boston: Kluwer.

Levinson, S. E. (1986), "Continuously Variable Duration Hidden Markov Models for Automatic Speech Recognition," *Computer, Speech and Language*, 1, 29–45.

——— (1987), "Continuous Speech Recognition by Means of Acoustic-Phonetic Classification Obtained From a Hidden Markov Model," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, New York: IEEE, pp. 93–96.

Levinson, S. E., Rabiner, L. R., and Sondhi, M. M. (1983), "An

Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition," *Bell System Technical Journal*, 62, 1035–1074.

Linde, Y., Buzo, A., and Gray, R. M. (1980), "An Algorithm for Vector Quantizer Design," *IEEE Transactions on Communications*, 28, 84–95.

Liporace, L. A. (1982), "Maximum Likelihood Estimation for Multivariate Observations of Markov Sources," *IEEE Transactions on Information Theory*, 28, 729–734.

Lippmann, R. P., Martin, E. A., and Paul, D. B. (1987), "Multistyle Training for Robust Isolated Word Speech Recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, New York: IEEE, pp. 705–708.

Ljolje, A., Ephraim, Y., and Rabiner, L. R. (1990), "Estimation of Hidden Markov Model Parameters by Minimizing Empirical Error Rate," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, New York: IEEE, pp. 709–712.

Lowerre, B., and Reddy, R. (1980), "The HARPY Speech Understanding System," in *Trends in Speech Recognition*, ed. W. Lea, Englewood Cliffs, NJ: Prentice-Hall, pp. 340–360.

Makhoul, J. (1975), "Linear Prediction: A Tutorial Review," *Proceedings of the IEEE*, 63, 561–580.

Markel, J. D., and Gray, A. H., Jr. (1976), *Linear Prediction of Speech*, New York: Springer-Verlag.

Merhav, N., and Ephraim, Y. (in press), "Maximum Likelihood Hidden Markov Modeling Using a Dominant Sequence of States," *IEEE Transactions on Acoustics, Speech and Signal Processing*.

Nadas, A., Nahamoo, D., and Picheny, M. A. (1988), "On a Model-Robust Training Method for Speech Recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36, 1432–1436.

Nilsson, N. J. (1965), *Learning Machines*, New York: McGraw-Hill.

Paul, D. B. (1985), "Training of HMM Recognizers by Simulated Annealing," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, New York: IEEE, pp. 13–16.

Poritz, A. B., and Richter, A. G. (1986), "On Hidden Markov Models in Isolated Word Recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, New York: IEEE, pp. 705–708.

Rabiner, L. R. (1984), "On the Application of Energy Contours to the Recognition of Connected Word Sequences," *AT&T Bell Labs Technical Journal*, 63, 1981–1995.

——— (1989), "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, 77, 257–286.

Rabiner, L. R., and Juang, B. H. (1986), "An Introduction to Hidden Markov Models," *IEEE Acoustics, Speech & Signal Processing Magazine*, 3, 4–16.

Rabiner, L. R., Juang, B. H., Levinson, S. E., and Sondhi, M. M. (1986), "Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixture Densities," *AT&T Technical Journal*, 64, 1211–1222.

Rabiner, L. R., Lee, C. H., Juang, B. H., and Wilpon, J. G. (1989), "HMM Clustering for Connected Word Recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pp. 405–408.

Rabiner, L. R., Levinson, S. E., and Sondhi, M. M. (1983), "On the Application of Vector Quantization and Hidden Markov Models to Speaker-Independent Isolated Word Recognition," *Bell System Technical Journal*, 62, 1075–1105.

Rabiner, L. R., Wilpon, J. G., and Juang, B. H. (1986), "A Segmental *k*-Means Training Procedure for Connected Word Recognition," *AT&T Technical Journal*, 65, 21–31.

Rabiner, L. R., Wilpon, J. G., and Soong, F. K. (1989), "High Performance Connected Digit Recognition Using Hidden Markov Models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37, 1214–1225.

Russell, M. J., and Moore, R. K. (1985), "Explicit Modeling of State Occupancy in Hidden Markov Models for Automatic Speech Recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, New York: IEEE, pp. 5–8.

Schafer, R. W., and Rabiner, L. R. (1971), "Design of Digital Filter Banks for Speech Analysis," *Bell System Technical Journal*, 50, 3097–3115.

Shikano, K. (1985), *Evaluation of LPC Spectral Matching Measures for Phonetic Unit Recognition*, technical report, Carnegie Mellon University, Computer Science Dept.

Sondhi, M. M., Levinson, S., and De la Noue, P. (1988), unpublished report, AT&T Bell Labs.

Sondhi, M. M., and Roe, D. (1983), unpublished report, AT&T Bell Labs.

Wellekens, C. (1987), "Explicit Time Correlation in Hidden Markov Models for Speech Recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, New York: IEEE, pp. 384–386.