HIDDEN MARKOV MODELS: A GUIDED TOUR

Alan B. Poritz

Institute for Defense Analyses, Communications Research Division
Princeton, NJ 08540

ABSTRACT

Hidden Markov modeling is a probabilistic technique for the study of time series. Hidden Markov theory permits modeling with many of the classical probability distributions. The costs of implementation are linear in the length of data. Models can be nested to reflect hierarchical sources of knowledge. These and other desirable features have made hidden Markov methods increasingly attractive for problems in language, speech and signal processing. In this paper, the basic ideas are introduced by elementary examples in the spirit of the Polya urn models. The main tool in hidden Markov modeling is the Baum-Welch (or forward-backward) algorithm for maximum likelihood estimation of the model parameters. This iterative algorithm is discussed both from an intuitive point of view as an exercise in the art of counting and from a formal point of view via the information-theoretic Q-function. Selected examples drawn from the literature illustrate how the Baum-Welch technique places a rich variety of computational models at the disposal of the researcher.

1. Introduction Hidden Markov modeling is a technique for the study of observed items arranged in a discrete-time series. The items in the series can be countably or continuously distributed; they can be scalars or vectors. The technique uses stochastic methods; a time series is generated and analyzed by a parametric probability model. It is parametric in the sense that it is completely described by a finite list of real numbers. A hidden Markov model has two components: a finite-state Markov chain and a finite set of output probability distributions. If the model is looked at generatively, the Markov chain synthesizes a sequence of states, (called a *path*) and the output distributions then turn this path into a time series. If it is looked at analytically, an observed time series gives evidence about the hidden path and the parameters of the generating model.

The work of Markov [48] and Shannon [65],[66] was concerned with Markov chains. The state sequence is observed in a Markov chain; see Billingsley [15]. In a hidden Markov model, the output probabilities impose a veil (Ferguson, [28]) between the state sequence and the observer of the time series. In the effort to lift the veil, a substantial body of theory has been developed over the past twenty-five years. The initial work dealt with finite probability spaces and addressed the problems of tractability of probability computation, the recovery of the hidden states, iterative maximum-likelihood estimation of model parameters from observed time series and the proof of consistency of the estimates; see Baum [10], Baum and Eagon [11], Baum and Petrie [12].

A major development in the theory (1970) was the maximization technique of Baum, Petrie, Soules and Weiss [13] that extended coverage to many of the classical distributions. This work has itself lead to a wide range of theoretical outgrowths. They include a number of generalizations of both the spatial and temporal components of the models, for example: variable-duration hidden Markov models [30], continuous multivariate hidden Markov models [47], hidden-filter hidden Markov models [58], and trainable finite-state (hidden) grammars [8]. A special case of the results in [13] has been designated by Dempster *et al* as the EM algorithm; see [23], especially pp. 28-29 and [62].

In the past few years there has been an explosive growth in the number of papers reporting applications of hidden Markov modeling. The applications are wide-ranging and if we include papers that reference algorithms derived from hidden Markov techniques they constitute a formidable body of literature. Some of the areas of research are: automatic speech recognition [5], [6], [37], [1], [61], [44], [19], [59], language modeling [17], [54], [35], [38], [40], coding theory [18], [2], pattern recognition [67], signal processing [25], [26], financial modeling [20], biological monitoring [70], and biostatistics [34],[56].

A number of survey papers with emphasis on applications to speech and language have helped to popularize the subject: Jelinek [37], Bahl *et al* [4], Levinson *et al* [46], and Levinson [43]. Here, we focus on the models themselves; examples and especially applications are presented mainly to clarify ideas. I thank J.D. Ferguson, L.A. Liporace and A.G. Richter for sharing their insights.

2. Mixtures as degenerate hidden Markov models Suppose we have an urn containing a mixture of black and white balls. Let $b(B)$ be the fraction of black balls and $b(W)$ the fraction of white balls so that $b(B) + b(W) = 1$. We will treat the urn as a probability model (see Feller [27]) and refer to the vector $\lambda = (b(B), b(W))$ as the *parameter vector of the model.* We generate a T-long *observation sequence* of colors $\mathbf{O} = (O_1, \ldots, O_T)$ by sampling the urn T times at random. We use the phrase "at random" to mean "with replacement and according to the uniform distribution". This is the classical situation of Bernoulli trials. Let $\#B$ be the number of black balls drawn and $\#W$ the number of white balls. The probability of the *sequence* of observations is $P_\lambda(\mathbf{O}) = b(B)^{\#B} b(W)^{\#W}$. If λ is unknown, its *maximum-likelihood estimate*, [21], is $\hat{\lambda} = (\#B/T, \#W/T)$. Estimates formed by ratios of observed counts are a persistent theme in what follows.

Consider next a model made from two urns (Urn 1 and Urn 2) and a mug (call it Mug 0). Each of the urns has its own mixture of black balls and white balls. The mug contains a mixture of marked stones: the marking on a stone is either "state 1" or "state 2". See Figure 1. Let a_{01} be the fraction of stones with the state 1 marking, and a_{02} the fraction with the state 2 marking so that $a_{01} + a_{02} = 1$. The parameter vector is now: $\lambda = (a_{01}, a_{02}, b_1(B), b_1(W), b_2(B), b_2(W))$. A T-long *observation sequence* of colors $\mathbf{O} = (O_1, \ldots, O_T)$ is generated as follows. At each time t in the interval $1, 2, \ldots, T$ select a stone from the mug at random. The marking on the stone is, say, "state s_t", (where s_t is either a 1 or a 2). Now select a ball at random from Urn s_t. O_t is the color of that ball.

This simple example already possesses properties associated with a hidden Markov model: it is a generative mechanism for creating observations and the mechanism is a stochastic process with a hidden

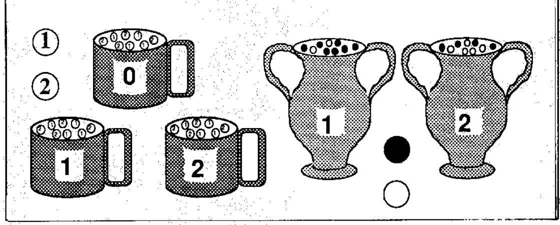

Figure 1. Urns containing colored balls. Mugs containing marked stones.

component. In the process of generating the observed sequence of colors $\mathbf{O}$, a sequence of stones (*i.e.,* states) $\mathbf{s} = (s_1, s_2, \ldots, s_T)$ is also generated. Since $\mathbf{s}$ is not observed, it is referred to it as a ***hidden*** sequence or path. For example, with $T = 6$, the state sequence obtained by sampling stones from the mug might be $\mathbf{s} = (1, 1, 1, 2, 1, 2)$ and the observation sequence obtained by sampling balls from the urns might be $\mathbf{O} = (B, W, B, W, W, B)$. See Figure 2. Ferguson's veil is due to the urns whose sampling obscures the view of the sequence of stones. The observer obtains only probabilistic evidence about the stones.

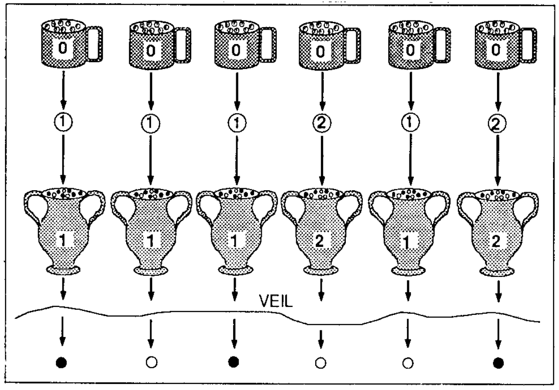

Figure 2. Sampling of the Urns veils the sampling of a single Mug.

The probability of an observation sequence generated by the model involves a sum over all possible configurations of the hidden component. However, since each observation is generated independently, the probability of the entire sequence is easily evaluated:

$$\begin{aligned}
P_\lambda(\mathrm{O}) &= \prod_{t=1}^{T} P_\lambda(O_t) = \prod_{t=1}^{T}(P_\lambda(O_t, s_t = 1) + P_\lambda(O_t, s_t = 2)) \\
&= \prod_{t=1}^{T}(P_\lambda(O_t|s_t = 1)P_\lambda(s_t = 1) + P_\lambda(O_t|s_t = 2)P_\lambda(s_t = 2)) \\
&= \prod_{t=1}^{T}(a_{01}b_1(O_t) + a_{02}b_2(O_t)) = \prod_{k=B,W}(a_{01}b_1(k) + a_{02}b_2(k))^{\#k}
\end{aligned}$$

3. Recovery of a mixture model from observations We have discussed the synthesis of observations from the model. We turn now to the inverse problem: the analysis of the model and the hidden path from these observations. The recovery of such information is a central issue in hidden Markov modeling.

The *prior probability* of state s at time τ is $P_\lambda(s_\tau = s) = a_{0s}$. Given the additional evidence of the observation sequence we define $\gamma_\tau(s)$, the *posterior probability* of state s at time τ to be $\gamma_\tau(s) = P_\lambda(s_\tau = s|\mathbf{O})$. We then have

$$\begin{aligned}
\gamma_\tau(s) =& P_\lambda(\mathrm{O}, s_\tau = s)/P_\lambda(\mathrm{O}) = P_\lambda(O_\tau, s_\tau = s)\prod_{t\neq\tau} P_\lambda(O_t)/P_\lambda(\mathrm{O}) \\
=& a_{0s}b_s(O_\tau)/(a_{01}b_1(O_\tau) + a_{02}b_2(O_\tau))
\end{aligned}$$

and $\gamma_\tau(1) + \gamma_\tau(2) = 1$. Thus, if the parameters of the true model are known, we can make an educated guess about an event (the stone chosen at time τ) that occurred during the generation of the observations but which was not itself observed. In this manner, the entire hidden path can be estimated.

Conversely, if the parameter values of the true model are unknown, but (somehow) the hidden path is made known to us, we can form a maximum likelihood estimate $\hat{\lambda}$, of the true model. For example, $\hat{b}_1(W)$ is the ratio of the number of times a white ball is drawn from Urn 1 to the number of times a ball is drawn from Urn 1.

Usually, however, we are given neither the path nor the true model, but only the observations (which we assume to come from the true model). Closed-form maximum-likelihood parameter estimation is no longer possible. What can be done ? One natural approach is the following. We start with any model whose parameter vector λ contains no zeros. Probability computations are, for the moment, to be based on this model, as if it were the true model. Recalling the definition of $\gamma_t(1)$, we see that $\sum_{t=1,T}\gamma_t(1)$ is the expected number of draws from Urn 1, given the observations and the model. Similarly $\sum_{t:O_t=W}\gamma_t(1)$ is the expected number of draws from Urn 1 that yield a white ball, again conditioned on the observations and the model. It is intuitively appealing to use this evidence to replace $b_1(W)$ by $\bar{b}_1(W) = \sum_{t:O_t=W}\gamma_t(1)/\sum_{t=1,T}\gamma_t(1)$. If a_{01} is then replaced by $\bar{a}_{01} = \sum_{t=1,T}\gamma_t(1)/T$, and so on, a new model $\bar{\lambda}$ is created.

It turns out this intuitive idea is well-founded. Direct calculation (Hartley, 1958, [34]), early computer experiments (Welch, [71]) and the theory to be discussed later all indicate that (except at critical points of the probability as a function of λ) the probability of the observations calculated according to the new model is greater than the probability according to the old model: $P_{\bar{\lambda}}(\mathrm{O}) > P_\lambda(\mathrm{O})$. If $\bar{\lambda}$ is now thought of as an old model, the procedure can be repeated until there is little or no further improvement in the model (as measured by the increase in probability). We make several such starts randomly dispersed across the space of models. The best one, $\hat{\lambda}$, is our estimate of the true model. The hidden state sequence is estimated from this final model.

It is implied by the content of Baum and Eagon [11] and Petrie [57] that we generically recover the true model from a sufficiently long observation sequence. By "generically" we mean to imply some caveats: there are symmetries associated with the naming of states, and there are ambiguities caused by a true model with an equal number of stones of each type, or with identical mixtures of balls in both urns, or with zeros in the parameter vector. Grim, [33], applied this method to independent sampling of mixtures for a number of classical distributions.

4. When time matters: an elementary hidden Markov model The order of the observations played no role in the previous example. It would have been enough to know how many balls of each color were drawn. We go to the trouble of collecting a time series rather than merely a histogram, precisely when we expect that there is information in the order in which the items are dealt out. By enriching our model to include a Markov chain, we model dependencies between adjacent observations by stochastic dependencies between the hidden states.

Consider then a model consisting of two urns (Urn 1 and Urn 2 again) and three mugs (Mug 1 and Mug 2 in addition to Mug 0). Each mug has its own mixture of stones marked "state 1" and "state 2". See Figure 1. The parameter vector is now: $\lambda = (a_{01}, a_{02}, a_{11}, a_{12}, a_{21}, a_{22}, b_1(B), b_1(W), b_2(B), b_2(W))$.

Generate a T-long observation sequence $\mathbf{O}$ as follows. Select a stone at random from Mug 0; its marking is, say, "state s_1". Select a ball at random from Urn s_1; its color is O_1. Now select a stone at random from Mug s_1; its marking is, say, "state s_2". Continue in this way using the current state to obtain both the current observation and the next state until a total of T observations $\mathbf{O} = (O_1, O_2, \ldots, O_T)$

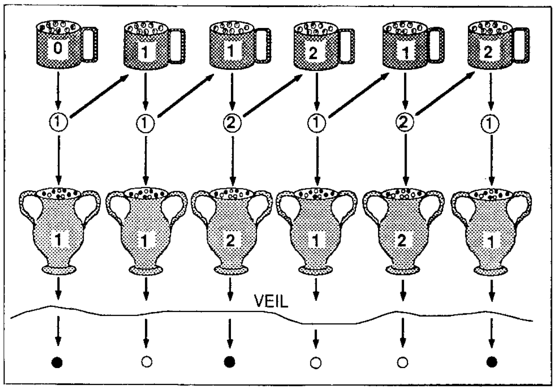

Figure 3. Sampling of the Urns veils the sampling of a sequence of Mugs.

have been generated. See Figure 3. Again denote the hidden state sequence by $\mathbf{s} = (s_1, s_2, \ldots, s_T)$.

Abstractly, what we now have is an *(order one) 2-state hidden Markov model.* It is an order one model because each successor state is selected as a probabilistic function of one predecessor state. Although there are 2 states in this example, any finite number, S, of states is possible. (In the example this would correspond to S urns and $S+1$ mugs each with a mixture of stones bearing S different markings). It is convenient to let S also be the name of the set of states.

The probability vector $a_0 = (a_{01}, a_{02}, \ldots, a_{0S})$ (associated in the example with Mug 0) is the *initial distribution.* The S by S row-stochastic matrix $A = (a_{rs})$ (whose rth row is associated with Mug r) is the *transition matrix.* a_0 and A together constitute a Markov chain of order one. The hidden state sequence $\mathbf{s}$ is produced by this chain. See Figure 4.

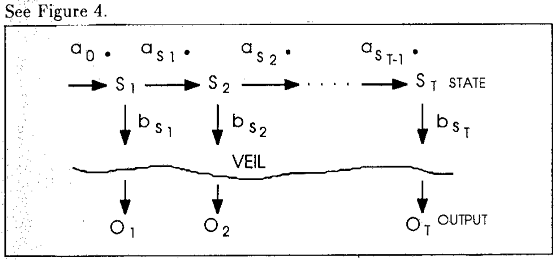

Figure 4. Generating a state sequence and observation sequence from a hidden Markov model.

A hidden Markov model is said to have a ***finite output alphabet*** if the observed items (also called *outputs*) lie in a finite set with K elements. In our example, the outputs are $\{B, W\}$ and $K = 2$. Other examples of alphabets include the first K integers, the English alphabet plus word space, the items in a vector-quantized code book, the set of phonemes in a language, and the set of words in a language. Alternatively, the observations can lie in a continuum or in a countably infinite set. Examples of observations in a continuum include real and complex, scalar and vector signal values (as seen for example in speech, or at the output of biological or industrial monitoring devices).

In the case of a finite alphabet, for each state s, the vector $b_s = (b_s(1), b_s(2), \ldots, b_s(K))$ (associated in the example with Urn s) is called the *output probability vector for state s.* The output probability can be general (an arbitrary distribution) or it can be parameterized (as, say, a binomial distribution). In the case of continuous output, each state s is associated with its own parametric probability density, b_s. The output probabilities map the state sequence $\mathbf{s}$ into the observation sequence $\mathbf{O}$; see Figure 4. again.

A hidden Markov model is summarized by its parameter vector $\lambda = (a_0, A, b_1, b_2, \ldots, b_S)$.

Although many of the early papers referred to the model as a "probabilistic function of a Markov chain", the less cumbersome phrase "hidden Markov model" coined by L. P. Neuwirth is better known today. The description of a hidden Markov model by the sampling of urns also began with Neuwirth [55]. In some of the literature, *e.g.,* [7], [37], [4], [3], output probabilities depend on a pair of states (*i.e.,* $b_{s,r}(k)$). The two variants have parallel theories. More generally, output probabilities can depend on a fixed window back into the recent past on both states and observations. We note that a hidden Markov model is not a Markov process of any finite order.

5. Recovery of a hidden Markov model from observations

Having discussed the generation of observations from a hidden Markov model, we once again turn to the inverse problem: the analysis of the model from observations of it. Following the expository style of J.D.Ferguson [29], we consider three basic problems of hidden Markov analysis. Suppose we are given a T-long sequence $\mathbf{O}$ of observations on an alphabet K, and an integer S. We assume that the observations were generated by an S-state hidden Markov model. Let $\lambda = (a_0, A, b_1, b_2, \ldots, b_S)$ be a model, but not necessarily the true model for the data. Let P_λ stand for probability or probability density according to what is required for the model. The basic problems are:

I. Compute $P_\lambda(\mathbf{O})$, the probability of $\mathbf{O}$ based on λ.
II. Estimate the true model in the maximum likelihood sense, that is, find the model $\hat{\lambda}$ that maximizes $P_{\hat{\lambda}}(\mathbf{O})$.
III. Estimate the hidden state sequence $\mathbf{s}$ from $\mathbf{O}$ and $\hat{\lambda}$.

Let $\mathbf{S}$ be the set of state sequences (or ***path space***), then: $P_\lambda(\mathbf{O}) = \sum_{\mathbf{s}\in\mathbf{S}} P_\lambda(\mathbf{O}, \mathbf{s}) = \sum_{\mathbf{s}\in\mathbf{S}} P_\lambda(\mathbf{O}|\mathbf{s})P_\lambda(\mathbf{s})$. Since $P_\lambda(\mathbf{O}|\mathbf{s}) = \prod_{t=1,T} b_{s_t}(O_t)$ and $P_\lambda(\mathbf{s}) = P_\lambda(s_1)\prod_{t=2,T} P_\lambda(s_t|s_1, \ldots, s_{t-1}) = a_{0s_1}\prod_{t=2,T} a_{s_{t-1}s_t}$, we have:

$$P_\lambda(\mathbf{O}) = \sum_{\mathbf{s}\in\mathbf{S}} a_{0s_1} b_{s_1}(O_1) \prod_{t=2}^{T} a_{s_{t-1}s_t} b_{s_t}(O_t).$$

$\mathbf{S}$ contains S^T members, so that $P_\lambda(\mathbf{O})$ is a sum of S^T terms. This sum becomes intractable as S and T grow. However, there is a better way to make the computation: [69], [18], [13], [10], [2], [31]. For any time t and state s define

$$\begin{aligned}\alpha_t(s) =& P_\lambda(O_1, \ldots, O_t, s_t = s)\\ \beta_t(s) =& P_\lambda(O_{t+1}, \ldots, O_T | s_t = s).\end{aligned}$$

We have $\alpha_1(s) = a_{0s} b_s(O_1)$ and for any $t = 2, \ldots, T$

$$\alpha_t(s) = \sum_{r\in S} \alpha_{t-1}(r) a_{rs} b_s(O_t).$$

Thus $P_\lambda(\mathbf{O}) = \sum_{s\in S} \alpha_T(s)$. This iteration solves Problem I by a calculation that grows linearly in T rather than exponentially, as we might have expected. Let $\beta_T(s) = 1$; for any $t = T-1, \ldots, 1$ we have

$$\beta_t(s) = \sum_{r\in S} a_{sr} b_r(O_{t+1}) \beta_{t+1}(r).$$

The alpha and beta inductions are frequently called the *forward* and *backward* calculations; see [37], [4]. Another exposition, derived from Ferguson's presentation [29], can be found in [60].

We observe that $P_\lambda(\mathbf{O}, s_t = s) = \alpha_t(s)\beta_t(s)$, since the later observations are independent of the earlier ones, in the presence of $s_t = s$. It is now easy to compute certain important posterior probabilities

$$\begin{aligned}\gamma_t(s) &\overset{\text{def}}{=} P_\lambda(s_t = s|\mathbf{O}) = P_\lambda(\mathbf{O}, s_t = s)/P_\lambda(\mathbf{O}) = \alpha_t(s)\beta_t(s)/P_\lambda(\mathbf{O})\\ \gamma_t(r, s) &\overset{\text{def}}{=} P_\lambda(s_t = r, s_{t+1} = s|\mathbf{O}) = \alpha_t(r) a_{r,s} b_s(O_{t+1}) \beta_{t+1}(s)/P_\lambda(\mathbf{O}).\end{aligned}$$

Once again we see that although $P_\lambda(\mathbf{O}, s_t = s)$ is a sum across all state sequences $\mathbf{s} \in \mathbf{S}$ that pass through s at time t, the posterior probability $\gamma_t(s)$ is efficiently calculated from the alphas and betas. See Figure 5., keeping in mind that $\alpha_t(s)$ is a sum across all partial state sequences ending at time t in state s and $\beta_t(s)$ is a sum across all partial state sequences beginning at time t in state s.

Assume for the moment that we believe the current model λ. Problem III can then be solved in several different senses [29]. If we want the state sequence $\mathbf{s} \in \mathbf{S}$ with the highest probability among all state sequences, then a dynamic program (Bellman's algorithm, [14]) determines the sequence for us. We need only replace the summation by maximization in the alpha induction and keep track, with a pointer, $\pi(s, t)$, to the next-to-last state on the highest-probability partial state-sequence ending in state s at time t. Although for many purposes this path is adequate, the sequence whose state at time t is $\arg\max_{s\in S} \gamma_t(s)$ possesses a greater expected number of correct states. Of course it may

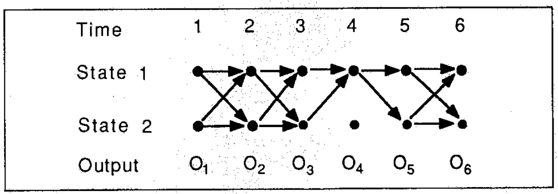

Figure 5. For fixed t, each of the $\gamma_t(s)$ is proportional to the sum of the $P_\lambda(\mathbf{O}, \mathbf{s})$ over all paths $\mathbf{s}$ that pass thru state s at time t.

contain some zero probability transitions. If that is a concern also, then a dynamic program through the $\gamma_t(r, s)$ array will get the legal sequence with the greatest expected number of correct state digraphs.

6. The Baum-Welch Algorithm The original solution of Problem II appeared in [11]; that paper dealt specifically with a finite alphabet and the general output distribution. A yet more fruitful technique based on the Kullback-Leibler number [41] was presented by Baum, Petrie, Soules and Weiss in [13]. For models λ and $\bar{\lambda}$ they defined the *auxiliary* or *Q-function*:

$$Q(\lambda, \bar{\lambda}) = \sum_{\mathbf{s} \in \mathbf{S}} P_\lambda(\mathbf{O}, \mathbf{s}) \log P_{\bar{\lambda}}(\mathbf{O}, \mathbf{s}).$$

Although at first glance Q looks more complicated than P, it is in fact easier to work with . First they showed that $Q(\lambda, \bar{\lambda}) > Q(\lambda, \lambda) \Rightarrow P_{\bar{\lambda}}(\mathbf{O}) > P_\lambda(\mathbf{O})$. (A three line proof, discovered by Liporace, is given in [47].) Next they observed that λ is a critical point of P if and only if it is a critical point of Q as a function of $\bar{\lambda}$ (*i.e.,* with λ held fixed): $\partial P_\lambda / \partial \lambda_i |_\lambda = \partial Q(\lambda, \bar{\lambda}) / \partial \bar{\lambda}_i |_{\bar{\lambda} = \lambda}$ for any coordinate λ_i of λ. Finally they showed that for a broad class of models, Q as a function of $\bar{\lambda}$, has a single critical point and this point is its unique global maximum. We refer to this point (this model) as the *Baum-Welch reestimate.*

The class of models covered in [13] includes the general distribution and the binomial distribution in the finite alphabet case, the Poisson distribution in the case of countable outputs and both the univariate normal and Gamma distributions among the continuous densities. For the general distribution in the finite alphabet case, the Baum-Welch reestimates are:

$$\begin{aligned}
\bar{a}_{0s} &= \gamma_1(s) \\
\bar{a}_{rs} &= \sum_{t=1}^{T-1} \gamma_t(r, s) / \sum_{s' \in S} \sum_{t=1}^{T-1} \gamma_t(r, s') \\
\bar{b}_s(k) &= \sum_{t: O_t = k} \gamma_t(s) / \sum_{t=1}^{T} \gamma_t(s)
\end{aligned}$$

for states $r, s \in S$ and letters $k \in K$.

These formulas have intuitive interpretations based on expected values computed with respect to the old model λ. For example $\bar{b}_s(k)$ is the ratio of the expected number of times that letter k is observed on a visit to state s to the expected number of times state s is visited. If the Q-function is specialized to the case of a mixture (that is, to a hidden Markov model of order zero), the formulas reduce to the intuitive reestimates given in the example in §3. See also Liporace [47].

The idea for solving Problem II is to iteratively improve the model parameters with the aid of Q. Start from some initial model λ and find the Baum-Welch reestimate, $\bar{\lambda}$, by maximizing the Q-function. Now take $\bar{\lambda}$ to be the new initial model and repeat the process. By what has been said, at each step, one of two things must happen: either $P_{\bar{\lambda}}(\mathbf{O}) > P_\lambda(\mathbf{O})$ or λ is a critical point of $P_\lambda(\mathbf{O})$. If $P_\lambda(\mathbf{O})$ has only finitely many critical points, then starting from several scattered initial models and iterating each to convergence, one generally obtains a good estimate of the maximum likelihood model $\hat{\lambda}$. See Figure 6. In practice, iterations are continued until some *ad hoc* criterion of convergence is satisfied. This technique for solving Problems I and II is known both as the *Baum-Welch algorithm* and as the *forward-backward algorithm.*

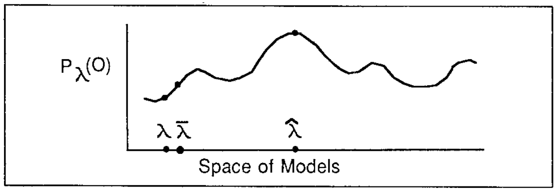

Figure 6. The probability function $P_\lambda(\mathbf{O})$ on the space of models λ. Baum-Welch reestimation $\lambda \rightarrow \bar{\lambda}$ climbs local hills. $\hat{\lambda}$ is the maximum likelihood model.

The technique was developed independently by M.I. Shlesinger in a paper on pattern recognition that appeared in the Russian journal Kibernetika in 1968 [67]. Although that paper dealt with the case of mixtures (*i.e.,* without the Markov chain), the main ideas in the algorithm were exposed. Shlesinger pointed out the important fact that the algorithm applies to models whose output probabilities belong to a particular parametric class whenever the weighted maximum likelihood problem can be solved for an arbitrary single distribution in that class.

In a later paper, [25], Shlesinger and N.A. Esin extended the discussion to hidden Markov models and to hidden Markov models with time-registered transition probabilities and multiple independent observation sequences. We note in passing that if $\mathcal{O} = (\mathbf{O}^1, \mathbf{O}^2, \ldots, \mathbf{O}^\mathbf{M})$ is a set of M independent observation sequences of various lengths generated by a single hidden Markov model, then the auxiliary function Q for the ensemble (appropriately defined in terms of state polysequences) is related to the auxiliary functions $Q^m, m \in M$ for the constituent sequences by the formula: $Q(\lambda, \bar{\lambda}) / P_\lambda(\mathcal{O}) = \sum_{m=1,M} Q^m(\lambda, \bar{\lambda}) / P_\lambda(\mathbf{O}^\mathbf{m})$. The formula is useful in assembling Baum-Welch reestimates.

Since increasing Q increases P the ideas above can be implemented even when there is no single maximizing critical point of Q. In such circumstances, gradient or other numerical methods are relied on to increase Q. Even if a single critical point of Q exists, there may not be a closed form expression for the Baum-Welch reestimate; again numerical methods are needed. An example of this kind, applied to the gamma distribution, is worked out by Levinson in [44].

7. Modeling temporal structure The Markov chain structure is the representation of the flow of information in a hidden Markov model. In complex problems there may exist a hierarchy of levels of information. For example, in speech one has at least semantic, syntactic, and acoustic-phonetic levels. The entire information hierarchy can be assembled into one grand hidden Markov model or *integrated network*, [6], [5], [4], with a sparse transition matrix. At the finest level, a number of elementary hidden Markov models directly produce observed items. Successively higher levels of information are embedded by linking together models formed at lower levels. A commonly used elementary model was introduced by Bakis [9]. In the Bakis model the transition matrix has zeros on all diagonals below the main diagonal and on all diagonals more than two above it. Paths enter only in state 1 and exit only from state S. Bakis "machines" successfully represent events in speech both at the integrated network level [6], [1], [19] and at lower levels [61], [59]. More general basic units can be considered.

It is possible to elaborate the temporal structure in a hidden Markov model for a time series. There is good reason to want to do this. The distribution of the lengths of repeated visits to a given state necessarily falls off geometrically in a hidden Markov model. But it may be that the process that plays out the states is not registered one-for-one with the process that generates the observations; indeed the meter may vary with the state. If observations arrive too infrequently we may want to allow for visits to states without production of visible output (or increase the rate of sampling). If they arrive too often we may want to allow at least some minimum number of observations for each visit to a state. Furthermore, the basic units of the hidden process can be complex: during a visit to the unit one sort of an output distribution may operate in the early portion and another sort later on. These types of concerns are addressed in the following example.

Instead of producing a single observation according to b_s during a visit to state s one could first sample a *duration distribution* $\mathcal{D}_s$ associated with the state and remain in the state for d times (with probability $\mathcal{D}_s(d), d = 1, 2, \ldots, D(s)$.) A total of d observations would be produced according to b_s or some other more elaborate output probability rule associated with state s. The next state would then be chosen according to the transition probabilities and the process repeated for this state.

This idea is called *variable duration* or *variable-length output*; it was developed by Ferguson [30], originally to model pitch in speech. Variable duration increases computational burden, but the inductive calculations remain linear in T. Let $\gamma_t(s, d)$ be the posterior probability that a d long visit to state s begins at time t; it can be calculated from appropriately defined alphas and betas. The Baum-Welch reestimate for the duration distribution for state s is then

$$\bar{\mathcal{D}}_s(d) = \sum_t \gamma_t(s,d) / \sum_t \sum_{d'} \gamma_t(s,d')$$

for $d = 1, 2, \ldots, D(s)$. Poisson distributed durations were developed in [30] and the binomial was mentioned as another tractable alternative. Russell and Moore also examined Poisson distributed durations in [64]. The Gamma distribution is the basis for the duration distributions studied by Levinson [44], [45]. The variable duration idea can be approximated with Bakis machines. The corresponding output probabilities can be forced to be equal (the technical term for this is *tied*, [4]) or allowed to be free.

The time order dependencies inherent in a Markov chain view may not be appropriate to model a particular time series. "Word order in sentences" is a commonly cited example of this problem. A more general class of dependencies is introduced into the model with the replacement, due to Baker [8], of the Markov chain by a context-free grammar, thus creating a *hidden grammar model.* States are elaborated into a finite set of non-terminal symbols $\mathcal{N}$ and a finite set of terminals $\mathcal{K}$. The transitions are replaced by probabilistic production rules of two types. Those of one type, $\{a_{efg}\}$, split a non-terminal e into a pair of non-terminals f and g and those of the other type $\{c_{ek}\}$ send a non-terminal e into a terminal k. The output probabilities are defined on terminal symbols only . The idea is that each non-terminal e sitting at a node in the parsing tree influences the entire interval of observations hanging on the branch of the tree that stems from e.

For hidden grammars, the forward-backward algorithm becomes an *inside-outside* algorithm whose computational burden grows as T^3. See Figure 7. Both P and Q become sums over parses instead of sequences. Let $\mathcal{I}$ be the set of split intervals $I = [t, u, v]$ where $1 \leq t, t \leq u, u \lessdot v, v \leq T$. Also define $\mathcal{J}$ be the set of non-empty intervals $J = [u, v]$ contained in $[1, T]$. Let $\gamma_I(e, f, g)$ be the (posterior) probability that e produces the observations in the interval $[t, v]$, f produces those in $[t, u]$ and g produces those in $[u+1, v]$ conditioned on the observations $\mathbf{O}$. Similarly let $\gamma_J(e, k)$ be the (posterior) probability that e and k both produce the observations in the interval $[u, v]$ again conditioned on $\mathbf{O}$. The Baum-Welch reestimation formulas are once more ratios of expected values:

$$\bar{a}_{efg} = \sum_{I \in \mathcal{I}} \gamma_I(e, f, g) / \sum_{f', g' \in \mathcal{N}} \sum_{I \in \mathcal{I}} \gamma_I(e, f', g')$$
$$\bar{c}_{ek} = \sum_{J \in \mathcal{J}} \gamma_J(e, k) / \sum_{k' \in \mathcal{K}} \sum_{J \in \mathcal{J}} \gamma_J(e, k').$$

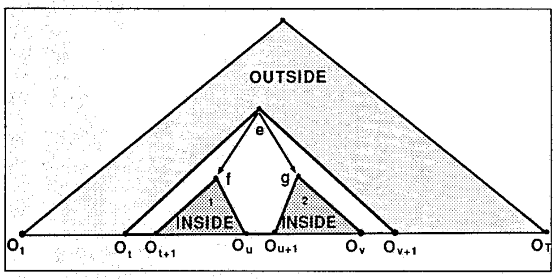

Figure 7. The inside-outside algorithm for hidden grammars leads to the posterior probabilities and the Baum-Welch reestimate.

8. Modeling spatial structure The Q-function view-point has been useful in widening the scope of the spatial (output) component of models for time series. We give some examples below.

Liporace [47] extended the Q-function and Baum-Welch reestimation to the broad class of hidden Markov models with elliptically-symmetric continuous outputs; included in this class are the N dimensional multivariate Gaussian densities. In such a model, for each state s, there is a multivariate Gaussian density with an N-long mean vector μ_s and a positive-definite N by N covariance matrix Σ_s. The Baum-Welch reestimates are:

$$\bar{\mu}_s = \sum_{t=1}^{T} \gamma_t(s) O_t / \sum_{t=1}^{T} \gamma_t(s)$$
$$\bar{\Sigma}_s = \sum_{t=1}^{T} \gamma_t(s)(O_t - \bar{\mu}_s) \otimes (O_t - \bar{\mu}_s) / \sum_{t=1}^{T} \gamma_t(s)$$

where for vectors u and v, $u \otimes v$ is the matrix whose ijth entry is the product $u_i v_j$. The reestimates are valid if, among the T observation vectors, some $N+1$ of them form an affine basis. These densities have had wide applicability for time series of observations of real vectors: [63],[59],[44],[3]. See also Nádas [50].

Given a signal, $Y = (y_{-N+1}, \ldots, y_0, \ldots, y_T)$, we can consider it as a time series generated by a ***hidden-filter*** hidden Markov model; see Poritz [58]. In the all-pole (that is, auto-regressive, or linear predictive) case, we associate with each state s an all-pole filter A_s of degree N, $(A_s = (a_N(s), \ldots, a_1(s))$ and a positive gain-factor σ_s. When state s is active, the next sample of the signal y_t is generated by applying the filter A_s to the most recent N samples of the signal and adding a sample u_t of $\mathcal{N}(0, \sigma_s^2)$ noise: $y_t = \sum_{j=1,N} a_j(s) y_{t-j} + u_t$. The Baum-Welch reestimates are determined as follows: For any t let $V_t = (y_{t-N}, \ldots, y_{t-1})$. Then for any $s \in S$, we have:

$$\bar{A}_s = \left(\sum_{t=1}^{T} \gamma_t(s) y_t V_t \right) \left(\sum_{t=1}^{T} \gamma_t(s) V_t \otimes V_t \right)^{-1},$$
$$\bar{\sigma}_s^2 = \sum_{t=1}^{T} \gamma_t(s) y_t (y_t - \langle V_t, \bar{A}_s \rangle) / \sum_{t=1}^{T} \gamma_t(s).$$

For models with no zero transitions, non-singularity of the N by N signal covariance matrix, $\sum_{t=1,T} V_t \otimes V_t$, assures invertibility of the expected state covariance-matrices above. These models approximate the behavior of a dynamic system (for example, a vocal tract, [58] and Juang and Rabiner [39]) as a time-dependent, noisy articulation of items drawn from a finite set of elementary steady state systems under the control of a Markov chain.

We may want to associate several output distributions with observations generated by what intuitively appears to be a single state (for example, a spoken vowel can be stressed or unstressed, nasalized or not, *etc.*). Instead of sampling a single output distribution per state, during a visit to a state s we could first sample a ***finite mixture distribution*** $\mathcal{M}_s$ associated with the state. Thus we first choose an index i (with probability $\mathcal{M}_s(i), i = 1, 2, \ldots, M(s)$) and then produce output according to a distribution $b_{s,i}(k)$ dependent on both state and index.

This structure is known as a ***mixtures*** hidden Markov model. Let $\gamma_t(s, i)$ be the posterior probability that state s is visited at time t and index i is selected; it can be computed from alphas and betas. The Baum-Welch reestimate for the mixture distribution for state s is then

$$\bar{\mathcal{M}}_s(i) = \sum_t \gamma_t(s, i) / \sum_t \gamma_t(s)$$

for $i = 1, 2, \ldots, M(s)$. This formula is the obvious hidden Markov analog of the formula given by Liporace in [47] for reestimating a single mixture distribution. Richter [63] described a mixture hidden Markov model (as an application of [47]) to handle speech data that is peaked and extended in comparison to a single Gaussian density. Each Richter mixture is a ***homothetic*** set of multivariate Gaussians; that is, densities with common means and covariance matrices that are scalar multiples of one another. Applications of these models are discussed in Bahl *et al* [3]. Mixtures have been employed in a number of hidden Markov studies, for instance in the hidden filter models described in [39].

9. Experimental Data Times-series obtained by data collection are not actually generated by hidden Markov models. Theoretical justification for maximum likelihood estimation (consistency of the estimate, [21], [12]) is therefore removed. Justification for use of the models rests on their success in applications; they are tractable approximations to a true model whose form is unknown. A number of studies have been aimed at finding an alternate criterion to maximum likelihood in selecting the parameters of the model: these include Mercer's maximum mutual-information [3], [53] and the work of Ephraim *et al* on minimum cross-entropy [24] (see also [32], [68]). Ideas in related areas of information geometry and alternating estimation that bear on this problem include [22], [42], and [49]. Nádas [50], [51], [52] discusses a number of practical and theoretical issues on the question of model estimation.

Even assuming the truth of the hidden Markov assumption, many practical considerations arise; we mention some of them here. During the recovery of the hidden state sequence with a dynamic program it is frequently sufficient to save the pointers for only a limited time back from the current time; see Brown *et al* [16]. To handle large state spaces or to restrict the state sequences to respect additional constraints, there are approximate methods; the use of a *stack* algorithm for example, [72], [36], [4]. Scaling of computation is needed to avoid underflow and overflow; see [46]. The quantity of data is often inadequate to produce reliable estimates of the probability of rare events. Several studies have addressed this sparse data question. The deleted interpolation technique of Jelinek and Mercer is summarized in [4]. Katz [40] smoothed probabilities based on Turing's estimate. Levinson *et al* [46] discussed reestimation with probabilities constrained away from zero.

Several useful points can be made regarding data preparation; see Poritz and Richter [59]. Time series are frequently obtained from continuous waveforms. The series is mapped by projection into a sequence of *feature vectors*. These projection operators include digital sampling, PCM coding, vector quantization, band pass filtering, subpopulation filtering and any other operators with less than full rank. In general, such operators destroy information. In particular, experience shows that *a priori* codebook quantization should be used with caution. If enlarging the codebook improves performance, then hidden Markov models based on continuous densities will frequently outperform models based on finite alphabets. Carrying this to the extreme, only modeling of the sampled waveform directly avoids unwarrented data manipulation.

It is important to test the goodness of fit of a converged hidden Markov model. A poor fit may sometimes be overcome. A one-to-one non-linear transformation of continuous data may improve results with Gaussian hidden Markov models (a linear map has no effect). The logarithm applied to positive real observations, for example, often pulls in offending outliers. A mixtures model may cope with multimodal or skewed continuous data. Another choice of parametric distribution may be called for. The number of states can be experimented with; the chi-square theory is discussed in [12]. The time series dynamics may carry Markov structure; improvement in performance has been obtained by concatenating items from nearby times into a single jointly-distributed poly-observation [59],[3].

Prior significance is often attached to individual states or collections of states in a hidden Markov model. Models constructed on labelled or scripted data are then tested on new data obtained from the same or similar sources. By imposing "known" constraints, less is expected from the statistical estimation process in the hope of making it easier to obtain a satisfactory model. This method has many proven advantages; but it does have its drawbacks.

In fact, as was originally observed by Neuwirth, the opposite point of view can be very revealing: a class of hidden Markov models whose states embody no prior assumptions of meaning can be used as a tool for the discovery of structure in a time series. An estimate of the maximum likelihood model in the class is computed by the Baum-Welch algorithm. The significance of the states is based on the parameters of this model. *A priori* labelling is replaced by *probabilistic labelling.* The idea has been used to model text (Cave and Neuwirth [17]), phonetics (Neuburg [54]) and speech (Poritz [58]). A variant discussed by Jelinek in [38] combines deterministic and probabilitic labelling in a language model. Results are frequently intuitively satisfying; that is, they agree with prior conceptions. They also introduce statistically important novel structure that is appreciated only after the fact.

Bibliography

1 A. Averbuch *et al*, "Experiments with the Tangora 20000 word speech recognizer," *Proc. ICASSP, Dallas* (1987), pp. 701-704

2 L. R. Bahl, J. Cocke, F. Jelinek and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Inform. Theory* **IT-20** (1974), pp. 275-320

3 L. A. Bahl, P. F. Brown, P. V. de Souza and R. L. Mercer, " Speech recognition with continuous-parameter hidden Markov models," *IBM Research Report RC13123,* T.J. Watson Research Center, Yorktown Heights, NY (1987)

4 L. R. Bahl, F. Jelinek and R. L. Mercer, "A maximum likelihood approach to continuous speech recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence* **PAMI-5** (1983), pp.179-190

5 J. K. Baker, "Stochastic modeling as a means of automatic speech recognition," *Ph.D. Dissertation,* Carnegie-Mellon Univ. (1975)

6 J. K. Baker, "The Dragon system - an overview," *IEEE Trans. Acoustics, Speech and Signal Process.* **ASSP-23** (1975), pp.24-29

7 J. K. Baker, "Stochastic modeling for automatic speech understanding" in *Speech Recognition* (D. R. Reddy, editor), Academic Press, New York (1975)

8 J. K. Baker, "Trainable grammars for speech recognition," *Speech Communications Paper, 97th Meeting of Acoustical Society of America*, Cambridge, MA (1979)

9 R. Bakis, "Continuous speech recognition via centisecond acoustic states," *J. Acoustical Society Am.* **59** *Supp. 1* (1976)

10 L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of a Markov process," *Inequalities* **III** (1972), pp.1-8

11 L. E. Baum and J. A. Eagon, "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology," *Bull. Amer. Math. Soc.* **73** (1967), pp.360-363

12 L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Ann. Math. Stat.* **37** (1966), pp.1554-1563

13 L. E. Baum, T. Petrie, G. Soules and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Stat.* **41** (1970), pp.164-171

14 R. E. Bellman, *Dynamic Programming,* Princeton Univ. Press, Princeton, NJ (1957)

15 P. Billingsley, *Statistical Inference for Markov Processes,* Univ. of Chicago Press, Chicago (1961)

16 P. F. Brown, J. C. Spohrer, P. H. Hochschild and J. K. Baker, "Partial traceback and dynamic programming," *Proc. ICASSP, Paris* (1982), pp.1629-1632

17 R. L. Cave and L. P. Neuwirth, "Hidden Markov models for English," in *Proceedings of the Symposium on the Applications of Hidden Markov Models to Text and Speech* (J. D. Ferguson, editor), IDA-CRD, Princeton, NJ (1980), pp.16-56

18 R. W. Chang and J. C. Hancock, "On receiver structures for channels having memory," *IEEE Trans. Inform. Theory* **IT-12** (1966), pp. 463-468

19 Y. L. Chow *et al*, "BYBLOS: The BBN Continuous Speech Recognition System," *Proc. ICASSP, Dallas* (1987), pp.89-92

20 T. M. Cover, "An algorithm for maximizing expected log investment return," *IEEE Trans. Inform. Theory* **IT-30** (1984), pp. 369-373

21 H. Cramér, *Mathematical Methods in Statistics*, Princeton Univ. Press, Princeton, NJ (1946)

22 I. Csisźar and G. Tusnady, "Information geometry and alternating minimization procedures," *Tech. Report,* Math. Inst. Hungarian Acad. Sci., Budapest (1982)

23 A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J.R. Stat. Soc.* **B 39** (1977), pp.1-38

24 Y. Ephraim, A. Dembo and L. R. Rabiner, "A minimum discrimination information approach for hidden Markov modeling," *Proc. ICASSP, Dallas* (1987), pp.25-28

25 N. A. Esin and M. I. Shlesinger, "Synthesis of a probabilistic finite-state grammar describing a given set of sequences," *Kibernetika* (1977), pp. 116-120

26 M. Feder, A. V. Oppenheim and E. Weinstein, "Methods for noise cancellation based on the EM algorithm," *Proc. ICASSP, Dallas* (1987), pp.201-204

27 W. Feller, *An Introduction to Probability Theory and Its Applications,* Vol.**I**, John Wiley, New York, 2nd Edition (1958)

28 J. D. Ferguson, *Unpublished lectures,* (1974)

29 J. D. Ferguson, "Hidden Markov analysis: an introduction," in *Proc. of the Symposium on the Applications of Hidden Markov Models to Text and Speech* (J. D. Ferguson, editor), IDA-CRD, Princeton, NJ (1980), pp.8-15

30 J. D. Ferguson, "Variable duration models for speech," in *Proc. of the Symposium on the Applications of Hidden Markov Models to Text and Speech* (J. D. Ferguson, editor), IDA-CRD, Princeton, NJ (1980), pp.143-179

31 G. D. Forney Jr., "The Viterbi Algorithm," *IEEE Proc.* **61** (1973), pp.266-278

32 R. M. Gray, A. H. Gray, G. Rebolledo and J. E. Shore, "Rate distortion speech coding with a minimum discrimination information distortion measure." *IEEE Trans. Inform. Theory* **IT-27** (1981), pp. 708-721

33 J. Grim, "An algorithm for maximizing a finite sum of positive functions and its application to cluster analysis," *Prob. Control and Inform. Theory,* **10** (1981), pp.427-437

34 H. O. Hartley, "Maximum likelihood estimation from incomplete data," *Biometrics* **14** (1958), pp.174-194

35 A. S. House and E. P. Neuburg, "Toward automatic identification of the language of an utterance, I. Preliminary methodological considerations," *J. Acoustical Society Am.* **62** (1977), pp. 708-713

36 F. Jelinek, "Fast sequential decoding algorithm using a stack," *IBM J. Res. and Develop.* **13** (1969), pp.675-685

37 F. Jelinek, "Continuous speech recognition by statistical methods," *IEEE Proc.* **64** (1976), pp.532-556

38 F. Jelinek, "Self-organized language modeling for speech recognition," *IBM Research Report,* T.J. Watson Research Center, Yorktown Heights NY (1985)

39 B-H. Juang and L. R. Rabiner, " Mixture autoregressive hidden Markov models for speech signals," *IEEE Trans. Acoustics, Speech and Signal Process.* **ASSP-33** (1985), pp. 1404-1413

40 S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE Trans. Acoustics, Speech and Signal Proc.* **ASSP-35** (1987), pp.400-401

41 S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Stat.* **22** (1951), pp.79-86

42 H. J. Landau, "Maximum entropy and the moment problem," *Bull. Amer. Math. Soc. New Series* **16** (1987), pp.47-77

43 S. E. Levinson, "Structural methods in automatic speech recognition," *IEEE Proc.* **73** (1985), pp.1625-1650

44 S. E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Computer Speech and Lang.* **1** (1986), pp.29-45

45 S. E. Levinson, "Continuous speech recognition by means of acoustic/phonetic classification obtained from a hidden Markov model," *Proc. ICASSP, Dallas* (1987), pp.93-96

46 S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *Bell Sys. Tech. J.* **62** (1983), pp. 1035-1074

47 L. A. Liporace, "Maximum likelihood estimation for multivariate observations of Markov sources," in *Proc. of the Symposium on the Applications of Hidden Markov Models to Text and Speech* (J. D. Ferguson, editor), IDA-CRD, Princeton, NJ (1980), pp.57-87. and *IEEE Trans. Inform. Theory* **IT-28** (1982), pp. 729-734

48 A. A. Markov, "An example of statistical investigation in the text of "Eugene Onyegin" illustrating coupling of "tests" in chains," *Proc. Acad. Sci. St. Petersburg VI Ser.* **7** (1913), pp. 153-162.

49 B. R. Musicus, "Iterative algorithms for optimal signal reconstruction and parameter identification given noisy and incomplete data," *Ph.D. Dissertation and Tech. Report 496,* Mass. Insititute of Technology Cambridge MA (1982)

50 A. Nádas, "Hidden Markov chains, the forward-backward algorithm, and initial statistics," *IEEE Trans. Acoustics, Speech and Signal Process.* **ASSP-31** (1983), pp.504-506

51 A. Nádas, "A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood," *IEEE Trans. Acoustics, Speech and Signal Proc.* **ASSP-31** (1983), pp.814-817

52 A. Nádas, "Estimation of probabilities in the language model of the IBM speech recognition system," *IEEE Trans. Acoustics, Speech and Signal Process.* **ASSP-32** (1984), pp.859-861

53 A. Nádas and D. Nahamoo, "On robust training in speech recognition," to appear *IEEE Trans. Acoustics, Speech and Signal Proc.*

54 E. P. Neuburg, "Markov models for phonetic text," *J. Acoustical Society. Am.* **50** (1971), p. 116(A)

55 L. P. Neuwirth, *Unpublished lectures* (1970)

56 J. Ott, "Counting methods (EM algorithm) in human pedigree analysis; linkage and segregation analysis," *Ann. Human Genetics* **40** (1977), pp. 443-454

57 T. Petrie, "Probabilistic functions of finite state Markov chains," *Ann. Math. Stat.* **40** (1969), pp. 97-115

58 A. B. Poritz, "Linear predictive hidden Markov models and the speech signal," in *Proc. of the Symposium on the Applications of Hidden Markov Models to Text and Speech* (J. D. Ferguson, editor), IDA-CRD, Princeton, NJ (1980), pp.88-142 and summarized in *Proc. ICASSP, Paris* (1982), pp. 1291-1294

59 A. B. Poritz and A. G. Richter, "On hidden Markov models in isolated word recognition," *Proc. ICASSP, Tokyo* (1986), pp.705-708

60 L. R. Rabiner and B. H. Juang, " An introduction to hidden Markov models," *IEEE ASSP Magazine* (Jan 1986), pp.4-16

61 L. R. Rabiner, S. E. Levinson and M. M. Sondhi, "On the application of vector quantization and hidden Markov models to speaker independent isolated word recognition," *Bell Sys. Tech. J.* **62** (1983), pp. 1075-1105

62 A. R. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *Siam Review* **26** (1984), pp.195-239

63 A. G. Richter, "Maximum likelihood estimation of a mixture of multivariate Gaussian distributions with homothetic covariance matrices," *Unpublished manuscript,* (1980), and presented at *Advances in Speech Processing Conf.,* IBM Europe Institute, Oberlech, Austria (1986)

64 M. J. Russell and R. K. Moore, "Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition," *Proc. ICASSP, Tampa* (1985), pp.5-8

65 C. C. Shannon, "A mathematical theory of communications," *Bell Sys. Tech. J.* **27** (1948), pp. 379-423, 623-656.

66 C. C. Shannon, "Prediction and entropy of printed English," *Bell Sys. Tech. J.* **30** (1951), pp. 50-64.

67 M. I. Shlesinger, "The interaction of learning and self-organization in pattern recognition," *Kibernetika* **4** (1968), pp. 81-88

68 J. E. Shore, "Minimum cross-entropy spectral analysis," *IEEE Trans. Acoustics, Speech and Signal Process.* **ASSP-29** (1981), pp.230-237

69 R. L. Stratonovich, "Conditional Markov processes," *Theory of Probability and its Applications* **v** (1960), pp. 156-178

70 Y. Vardi, L. A. Shepp, L. Kauffman, "A statistical model for positron emission tomography," *J. Am. Stat. Assn.* **80** (1985), pp. 8-37

71 L. R. Welch, *Unpublished work*

72 K. S. Zigangirov, "Some sequential decoding procedures," *Probl. Pered. Inform.* **2** (1966), pp.13-25