BIBLIOGRAPHY

[1] Baum, Leonard E., Petrie, Ted, Soules, George, and Weiss, Norman, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, Annals of Mathematical Statistics, Vol. 41, No. 1, 1970, pp. 164-171.

[2] Liporace, Louis A., Maximum likelihood analysis for multi-variate observations of Markov sources, (These Proceedings).

[3] Liporace, Louis A., Statistical analysis of time series, (These Proceedings).

[4] Ferguson, John D., Variable duration models for speech, (These Proceedings).

[5] Markel, J.D., and Gray, A.H. Jr., Linear prediction of speech, Springer-Verlag, New York, (1976).

VARIABLE DURATION MODELS FOR SPEECH

John D. Ferguson

ABSTRACT

In this paper we discuss an extension of the basic theory of hidden Markov models to allow for the fact that speech sounds have varying durations. The extension introduces a distribution on the duration spent in a state, and allows a sequence of observations to be made while remaining in a single state.

VARIABLE DURATION MODELS FOR SPEECH

John D. Ferguson

0. Hidden Markov models

In the case of analyzing printed text using hidden Markov models ala Cave-Neuwirth [4], the observations are discrete and are always of exactly the same length: one character. In trying to analyze speech in a similar way, we have to allow for continuous measurements, and states persisting for various durations. The extension of the basic technique to continuous measurements was discussed in Liporace's paper [6]. Here we discuss the question of variable duration states.

We are interested in a stochastic process which produces observations in the following way. There is a Markov chain with S states, i, $i = 1,2,\ldots,S$, with initial distribution (a_i), transition matrix (a_{ij}), and output matrix $(b_j(k))$ which expresses the probability that the observation will be k, given state j. [We assume, for exposition, a finite output alphabet. We will remark later on the extension of the results obtained to the case of continuously distributed output.]

In the usual case [1], a sequence of observations $\mathcal{O}_1,\mathcal{O}_2,\ldots,\mathcal{O}_T$ is produced by a hidden Markov model with one observation per state. The actual sequence of events is taken to be:

(A) Choose an initial state i_1 according to the initial distribution a_i.

(B) Choose an observation $\mathcal{O}_1$ according to the distribution $b_{i_1}(\mathcal{O})$.

(C) Choose the next state i_2, according to the transition probability distribution $a_{i_1,j}$ $(j = 1,2,\ldots,S)$.

(D) Choose an observation $\mathcal{O}_2$ according to the output distribution $b_{i_2}(k)$, $k = 1,2,\ldots,K$, etc.

In the case of variable duration hidden Markov models, there is an additional ingredient: associated with each state, i, there is a duration distribution $P(d|i)$, $(d = 1,2,\ldots,D)$ which determines the length of time spent in state i, i.e., the number of observations produced while in state i. The sequence of events is now:

(A) Choose an initial state i_1 according to the initial distribution a_i.

(B) Choose a duration d_1 according to the duration distribution $P(d|i_1)$.

(C) Choose observations $\mathcal{O}_1,\mathcal{O}_2,\ldots,\mathcal{O}_{d_1}$ according to $b_{i_1}(\mathcal{O}_1,\ldots,\mathcal{O}_{d_1})$. [This is often, for convenience, taken to be $\Pi_{t=1}^{d_1} b_{i_1}(\mathcal{O}_t)$.]

(D) Choose the next state i_2 according to the transition probabilities $a_{i_1,j}$ $(j = 1,2,\ldots,S)$.

(E) Choose a duration d_2 according to the duration distribution $P(d|i_2)$.

(F) Choose observations $\mathcal{O}_{d_1+1}$, $\mathcal{O}_{d_1+2}$, $\ldots$, $\mathcal{O}_{d_1+d_2}$ according to $b_{i_2}(\mathcal{O}_{d_1+1},\ldots,\mathcal{O}_{d_1+d_2})$, etc.

It is important to note that there is no external demarcation between the observations arising from state i_1 and those arising from i_2.

Let us note that the variable duration model includes the usual model in either of the following two ways:

(i) For each state j, set $P(d=1|j) = 1$. Thus each state lasts exactly one time, and we return to the usual model.

(ii) For each state, j, set $P(d|j) = \rho_j^{d-1}(1-\rho_j)$ for some ρ_j between 0 and 1. Thus the duration distribution is geometric. We also require $a_{ii} = 0$ for each i. Such a model then mimics the usual case.

Conversely, with considerable labor, the variable duration case can be realized in the old framework. Take as states the triples (i,x,d) where i is one of the variable duration states, d a duration, and x a counter, $1 \leq x \leq d$, which indicates the number of observations produced so far while in state i. Transitions are of the form $(i,x,d) \rightarrow (i,x+1,d)$, unless $x = d$, when $(i,d,d) \rightarrow (j,1,d')$ for some j, some d'.

In the Poritz LPC model, described in an earlier paper in this collection, [5], it would be a definite advantage to allow for variable duration. It is clear that phones or phone categories have characteristic duration distributions which are not, in general, geometric. Thus the basic model, without variable duration, should be somewhat ill-fitting.

As usual in the study of hidden Markov models, there are three basic problems:

(i) Given $\mathcal{O}_1,\ldots,\mathcal{O}_T$ and a model, (a_i), (a_{ij}), $(b_j(\mathcal{O}))$, $(P(d|i))$ compute the likelihood of the sequence $\mathcal{O}_1,\ldots,\mathcal{O}_T$.

(ii) Given $\mathcal{O}_1,\ldots,\mathcal{O}_T$ and a state space of size S, find the maximum likelihood model.

(iii) Given $\mathcal{O}_1,\ldots,\mathcal{O}_T$ and a model, find the most likely state sequence either in the sense of the single state sequence of maximum posterior probability, or the state sequence with maximum per state posterior probability. In the case of variable duration hidden Markov models, this problem is complicated by decisions about the durations of the individual states.

We will discuss all three of these problems for our case.

1. Computation of likelihood for variable duration models

Suppose we have observations $\mathcal{O} = \mathcal{O}_1,\ldots,\mathcal{O}_T$ and a variable duration model. We wish to compute the likelihood of the sequence $\mathcal{O}$ given the model. We will make the simplifying assumption that the first state commences at time 1, and the last state ends at time T [no observations are "lost" on either end]. It is messy, but possible, to handle the situation where the first state begins at or before observation $\mathcal{O}_1$, and the last state ends at or after observation $\mathcal{O}_T$. We will not discuss this complication here.

To compute the likelihood, and for later use, we define an analogue of the usual alpha calculation [1].

Let

$$\alpha_t(i) = \Pr(\mathcal{O}_1\ldots\mathcal{O}_t \text{ and state } i \text{ terminates at } t).$$

This probability is a sum over paths $p = (i_1,d_1),\ldots,(i_r,d_r)$ such that $i_r = i$ and $d_1 + \ldots + d_r = t$, of the joint probability of the path and the observations. Thus

$$\alpha_t(i) = \sum_p \text{Prob}(p \;\&\; \mathcal{O}_1\mathcal{O}_2\ldots\mathcal{O}_t)$$

$$= \sum_p a_{i_1} \cdot P(d_1|i_1) \cdot P(\mathcal{O}_1\ldots\mathcal{O}_{d_1}|i_1) \cdot a_{i_1 i_2} \cdot P(d_2|i_2) \cdot P(\mathcal{O}_{d_1+1}\ldots\mathcal{O}_{d_1+d_2}|i_2)\ldots a_{i_{r-1} i_r} \cdot P(d_r|i_r) \cdot P(\mathcal{O}_{d_1+\ldots+d_{r-1}+1}\ldots\mathcal{O}_t|i_r) \cdot$$

We note that, as in the usual case, there is an effectively computable inductive process for calculating the alphas. The likelihood we want is simply $\Sigma_i \alpha_T(i)$.

We have, inductively:

$$\alpha_t(j) = \sum_{i=1}^{S} \sum_{d=1}^{D} \alpha_{t-d}(i) \cdot a_{ij} \cdot P(d|j) \cdot P(\mathcal{O}_{t-d+1} \cdots \mathcal{O}_t | j)$$

ALPHA INDUCTIVE FORMULA

To complete the calculation of likelihood, given a model, we have to show how to initialize the alphas. Since we have assumed that some state must commence at time 1, we can compute:

$$\alpha_1(i) = a_i P(d=1|i) \cdot P(\mathcal{O}_1|i), \quad i = 1,2,\ldots,S,$$

$$\alpha_2(i) = a_i P(d=2|i) \cdot P(\mathcal{O}_1\mathcal{O}_2|i) + \sum_j \alpha_1(j) \cdot a_{ji} \cdot P(d=1|i)P(\mathcal{O}_2|i),$$

$$\alpha_3(i) = a_i P(d=3|i) \cdot P(\mathcal{O}_1\mathcal{O}_2\mathcal{O}_3|i) + \sum_{d<3} \sum_j \alpha_{3-d}(j) a_{ji} P(d|i) P(\mathcal{O}_{4-d} \cdots \mathcal{O}_3|i)$$

etc. until $\alpha_D(i)$ is completed, after which we use the basic inductive formula. [There is an alternative approach involving a dummy array $\alpha_0(i)$, $\alpha_{-1}(i)$, ..., $\alpha_{-D+1}(i)$ and dummy observations $\mathcal{O}_0, \mathcal{O}_{-1}, \ldots, \mathcal{O}_{-D+1}$.]

For speech, the results of this section can be used as follows: given several competing models, say for different talkers, or for different phrases, and given a segment of speech, we can compute the probability of the segment given the model. Together with the prior probability of the model, this yields the posterior probability of the model given the segment.

2. Iterative method for obtaining the maximum likelihood model

In this section we show how the Baum algorithm [1] can be adapted to the case of variable duration hidden Markov models to provide an iterative method for finding the maximum likelihood model.

In the Poritz scheme, [5], extended, we would thus be able to estimate the LPC parameters for each state as well as the distribution on the duration for which these LPC parameters are to be held fixed.

Given any model M, with likelihood $L(M)$, the Baum algorithm produces another model $M' = \mathcal{T}(M)$ such that $L(M') > L(M)$, unless $M = M'$ in which case M must be a critical point of the likelihood function. The algorithm is used iteratively. We start with a

crude model M_0, and compute, successively, $M_1 = \mathcal{T}(M_0)$, $M_2 = \mathcal{T}(M_1)$, etc. Since the likelihood increases at each step, (unless a critical point is reached), then either the likelihood converges, or, in the continuous case, possibly approaches $+\infty$.

3. Computations

We discuss how we can use this algorithm in our case. In the following, the dependence of all probabilities on the current model is suppressed.

We will need three other inductively computable arrays: $\alpha_t^*(i)$, $\beta_t(i)$, $\beta_t^*(i)$ for $t = 0,1,\ldots,T$, $i = 1,\ldots,S$.

We define

$$\alpha_t^*(i) = P(\mathcal{O}_1\ldots\mathcal{O}_t \;\&\; i \text{ commences at } t+1),$$

$$\beta_t(i) = P(\mathcal{O}_{t+1}\ldots\mathcal{O}_T | i \text{ terminates at } t),$$

$$\beta_t^*(i) = P(\mathcal{O}_{t+1}\ldots\mathcal{O}_T |\; i \text{ commences at } t+1).$$

Then we have:

$$\text{(i)}\quad \alpha_t^*(j) = \sum_i \alpha_t(i)a_{ij}\ ,$$

$$\text{(ii)}\quad \alpha_t(i) = \sum_d \alpha_{t-d}^*(i)\cdot P(d|i)\cdot P(\mathcal{O}_{t-d+1}\ldots\mathcal{O}_t|i),$$

$$\text{(iii)}\quad \beta_t(i) = \sum_j a_{ij}\beta_t^*(j),$$

$$\text{(iv)}\quad \beta_t^*(i) = \sum_d \beta_{t+d}(i)P(d|i)\cdot P(\mathcal{O}_{t+1}\ldots\mathcal{O}_{t+d}|i)\ .$$

The betas, as in the Baum algorithm, are computed from right to left, and there is a similar problem in starting up.

We now note some relations (letting $\mathcal{O}$ denote $\mathcal{O}_1\mathcal{O}_2\ldots\mathcal{O}_T$ and $P = Pr(\mathcal{O})$):

(A) $\alpha_t(i)\cdot\beta_t(i)/P = \text{Prob}(\text{state } i \text{ terminates at time } t, \text{ given } \mathcal{O})$,

(B) $\alpha_t^*(i)\cdot\beta_t^*(i)/P = \text{Prob}(\text{state } i \text{ commences at time } t+1, \text{ given } \mathcal{O})$,

(C) $S_t(i) = \sum_{\tau<t} \alpha_\tau^*(i)\cdot\beta_\tau^*(i)/P$

$= \text{Expected number of times state } i \text{ started at time } t \text{ or before, given } \mathcal{O}$,

(D) $E_t(i) = \sum_{\tau<t} \alpha_\tau(i)\beta_\tau(i)/P$

$= \text{Expected number of times state } i \text{ terminated before time } t, \text{ given } \mathcal{O}$,

(E) $\gamma_t(i) = S_t(i) - E_t(i)$

= Prob(state i was used to produce observation $\mathcal{O}_t$, given $\mathcal{O}$),

[Hence $\forall t$, $\Sigma_i \gamma_t(i) = 1$.]

(F) $\sum_t \gamma_t(i)$ = Expected total duration spent in state i, given $\mathcal{O}$,

(G) $S_T(i)$ = Expected total number of times state i commenced, given $\mathcal{O}$

= $E_{T+1}(i)$ = expected number of times state i terminated, given $\mathcal{O}$,

(H) $\sum_t \gamma_t(i)/S_T(i)$ = estimated average duration of state i, given $\mathcal{O}$,

(I) $C(i) = \frac{a_i \beta_0^*(i)}{P}$ = Probability that state i was the first state, given $\mathcal{O}$,

(J) $C(i,j) = \sum_t \frac{\alpha_t(i) a_{ij} \beta_t^*(j)}{P}$ = Expected number of transitions from i to j, given $\mathcal{O}$,

(K) $C(i,d) = \sum_t \alpha_t^*(i) P(d|i) \cdot P(\mathcal{O}_{t+1} \ldots \mathcal{O}_{t+d} | i) \cdot \beta_{t+d}(i)/P$

= Expected number of times state i occurred with duration d, given $\mathcal{O}$,

(L) $C(i,k) = \sum_{t \ni \mathcal{O}_t = k} \gamma_t(i)$

= Expected number of times that state i occurred with observation $\mathcal{O}_t = k$, given $\mathcal{O}$.

Most of these relations are quite simple. Let us discuss only (A) and (E).

For (A), $\alpha_t(i)\beta_t(i)$ is the sum of the probabilities for all paths (i.e., state sequences with durations) which have the property that state i terminates at time t. This holds since $\alpha_t(i)$ accounts for all paths to the left of t, ending in i, and $\beta_t(i)$ accounts for all paths to the right, given i ends at t, and any pairs can be combined.

This is in direct analogy to [1], but note that

$$\sum_i \frac{\alpha_t(i)\beta_t(i)}{P} \leq 1 \quad ,$$

since it is not necessarily true that some state must end at time t.

For (E), we argue as follows: the probability that i is used at time t is

$$\sum_{\tau<t} \sum_{d \ni t \leq \tau+d} \frac{\alpha^*_\tau(i) P(d|i) \cdot P(\mathcal{O}_{\tau+1} \ldots \mathcal{O}_{\tau+d}|i) \beta_{\tau+d}(i)}{P} ,$$

by examining paths, as above.

This can be rewritten:

$$\sum_{\tau<t} \sum_{\text{all } d} \frac{\alpha^*_\tau(i) P(d|i) \cdot P(\mathcal{O}_{\tau+1} \ldots \mathcal{O}_{\tau+d}|i) \beta_{\tau+d}(i)}{P}$$

$$- \sum_{\tau<t} \sum_{\substack{\text{all } d \ni \\ t>\tau+d}} \frac{\alpha^*_\tau(i) P(d|i) \cdot P(\mathcal{O}_{\tau+1} \ldots \mathcal{O}_{\tau+d}|i) \beta_{\tau+d}(i)}{P} .$$

The first quantity is

$$\sum_{\tau<t} \frac{\alpha^*_\tau(i) \cdot \beta^*_\tau(i)}{P} ,$$

using the inductive formula (iv). The second quantity, replacing $\tau + d$ by θ, τ by $\theta - d$, is

$$\sum_{\theta<t} \sum_{d} \frac{\alpha^*_{\theta-d}(i) \cdot P(d|i) \cdot P(\mathcal{O}_{\theta-d+1} \ldots \mathcal{O}_\theta|i) \beta_\theta(i)}{P}$$

$$= \text{(using inductive formula (ii))}$$

$$= \sum_{\theta<t} \frac{\alpha_\theta(i) \beta_\theta(i)}{P} ,$$

as asserted.

Now we show how to use these various computed quantities to reestimate the parameters of the model.

The new initial probabilities are

$$\boxed{\hat{a}_i = C(i)} \qquad \text{(see (I))}.$$

The new transition probabilities are

$$\boxed{\hat{a}_{ij} = C(i,j) \Big/ \sum_j C(i,j) \quad \text{(see (J))} .}$$

The new duration probabilities are

$$\boxed{\hat{P}(d|i) = C(i,d) \Big/ \sum_d C(i,d) \quad \text{(see (K))} .}$$

For the reestimates of observation probabilities, we make the assumption of independence, although other hypotheses can be dealt with.

The new observation probabilities are

$$\boxed{\hat{P}(\mathcal{O}=k|i) = C(i,k) \Big/ \sum_k C(i,k) \quad \text{(see (L))} .}$$

4. Proof of increasing likelihood

We show how the proof of [1] that the likelihood must increase can be adapted to fit our case.

Let (cf. [1]) λ be a generic symbol for the parameters of our model. Thus $\lambda = ((a_i), (a_{ij}), (P(d|i)), (P(k|i)))$. Define

$$Q(\lambda,\lambda') = \sum_{\text{Paths } p} P_\lambda(p\&\mathcal{O}) \log P_{\lambda'}(p\&\mathcal{O}) .$$

The sum is over all paths

$$p : (i_1,d_1),(i_2,d_2),\ldots,(i_N,d_N) \ni \sum_{\tau=1}^{N} d_\tau = T .$$

The first probability, $P_\lambda(p\&\mathcal{O})$, denotes the probability of the path p and the observation sequence $\mathcal{O}$, given the parameter set λ. The second probability $P_{\lambda'}(p\&\mathcal{O})$ denotes the probability of the same path and $\mathcal{O}$, given a second parameter set λ'. This auxiliary function greatly facilitates the proof of increasing likelihood.

Lemma 1. $Q(\lambda,\lambda') - Q(\lambda,\lambda) \leq P_{\lambda'}(\mathcal{O}) - P_\lambda(\mathcal{O})$. The inequality is strict unless, for each path, p, $P_\lambda(p\&\mathcal{O}) = P_{\lambda'}(p\&\mathcal{O})$.

Proof. We use the elementary inequality $\log x \leq x - 1$, which is strict unless $x = 1$.

Thus,

$$Q(\lambda,\lambda') - Q(\lambda,\lambda) = \sum_p P_\lambda(p\&\mathcal{O}) \ln \frac{P_{\lambda'}(p\&\mathcal{O})}{P_\lambda(p\&\mathcal{O})}$$

$$\leq \sum_p P_\lambda(p\&\mathcal{O}) \left(\frac{P_{\lambda'}(p\&\mathcal{O})}{P_\lambda(p\&\mathcal{O})} - 1 \right) = P_{\lambda'}(\mathcal{O}) - P_\lambda(\mathcal{O}) .$$

As shown in [1], a stronger inequality holds:

$$Q(\lambda,\lambda') - Q(\lambda,\lambda) \leq P_\lambda(\mathcal{O}) \ln \frac{P_{\lambda'}(\mathcal{O})}{P_\lambda(\mathcal{O})} ,$$

but Lemma 1 will suffice for our purposes.

This lemma relates the Q function to the likelihood. It indicates that, for fixed λ, if we can find any $\lambda' \ni Q(\lambda,\lambda') > Q(\lambda,\lambda)$ then $P_{\lambda'}(\mathcal{O}) > P_\lambda(\mathcal{O})$. One natural method is therefore to maximize $Q(\lambda,\lambda')$ as a function of λ'.

Lemma 2.

$$\frac{Q(\lambda,\lambda') - Q(\lambda,\lambda)}{P_\lambda(\mathcal{O})}$$

$$= \sum_i C_\lambda(i) \ln \frac{a'_i}{a_i} + \sum_{i,j} C_\lambda(i,j) \ln \frac{a'_{ij}}{a_{ij}}$$

$$+ \sum_{i,d} C_\lambda(i,d) \ln \frac{P'(d|i)}{P(d|i)} + \sum_{i,k} C_\lambda(i,k) \ln \frac{P'(k|i)}{P(k|i)} .$$

[Here the C's refer to the counts made for reestimation, based on parameters λ, described in Section 3, (I), (J), (K), (L).]

Proof. A path can be defined by a state sequence, with associated durations:

$$(i_1,d_1),(i_2,d_2),\ldots,(i_N,d_N) \ni \sum_\tau d_\tau = T \quad .$$

Given a path, p, the observation string is decomposed into segments $\mathcal{O}_p^{(1)} = \mathcal{O}_1\ldots\mathcal{O}_{d_1}$, $\mathcal{O}_p^{(2)} = \mathcal{O}_{d_1+1}\ldots\mathcal{O}_{d_1+d_2},\ldots,$ $\mathcal{O}_p^{(N)} = \mathcal{O}_{d_1+\ldots+d_{N-1}+1}\ldots\mathcal{O}_T$. Using this notation, we have:

$$\begin{aligned}
Q(\lambda,\lambda') &= \sum_p P_\lambda(p\&\mathcal{O}) \ln P_{\lambda'}(p\&\mathcal{O}) \\
&= \sum_p P_\lambda(p\&\mathcal{O}) \ln \Big\{ a'_{i_1} \cdot \prod_{\tau=2}^{N} a'_{i_{\tau-1}i_\tau} \cdot \\
&\qquad \prod_{\tau=1}^{N} P'(d_\tau|i_\tau) \cdot \prod_{\tau=1}^{N} P'(\mathcal{O}_p^{(\tau)}|i_\tau) \Big\} \\
&= \sum_p P_\lambda(p\&\mathcal{O}) \ln a'_{i_1} + \sum_p \sum_{\tau=2}^{N} P_\lambda(p\&\mathcal{O}) \ln a'_{i_{\tau-1}i_\tau} \\
&+ \sum_p \sum_{\tau=1}^{N} P_\lambda(p\&\mathcal{O}) \ln P'(d_\tau|i_\tau) \\
&+ \sum_p \sum_{\tau=1}^{N} P_\lambda(p\&\mathcal{O}) \ln P'(\mathcal{O}_p^{(\tau)}|i_\tau) \quad .
\end{aligned}$$

Hence:

$$\begin{aligned}
&\frac{Q(\lambda,\lambda') - Q(\lambda,\lambda)}{P_\lambda(\mathcal{O})} \\
&= \frac{1}{P_\lambda(\mathcal{O})} \sum_p P_\lambda(p\&\mathcal{O}) \ln (a'_{i_1}/a_{i_1}) \\
&+ \frac{1}{P_\lambda(\mathcal{O})} \sum_p \sum_{\tau=2}^{N} P_\lambda(p\&\mathcal{O}) \ln (a'_{i_{\tau-1}i_\tau}/a_{i_{\tau-1}i_\tau}) \\
&+ \frac{1}{P_\lambda(\mathcal{O})} \sum_p \sum_{\tau=1}^{N} P_\lambda(p\&\mathcal{O}) \ln (P'(d_\tau|i_\tau)/P(d_\tau|i_\tau)) \\
&+ \frac{1}{P_\lambda(\mathcal{O})} \sum_p \sum_{\tau=1}^{N} P_\lambda(p\&\mathcal{O}) \ln (P'(\mathcal{O}_p^{(\tau)}|i_\tau)/P(\mathcal{O}_p^{(\tau)}|i_\tau)) \quad .
\end{aligned}$$

Each of the four summands here will be related to the corresponding summand in the conclusion of the lemma.

First of all,

$$\begin{aligned}
&\frac{1}{P_\lambda(\mathcal{O})} \sum_p P_\lambda(p\&\mathcal{O}) \ln (a'_{i_1}/a_{i_1}) \\
&= \frac{1}{P_\lambda(\mathcal{O})} \sum_{i=1}^{S} \ln (a'_i/a_i) \sum_{p \ni i_1=i} P_\lambda(p\&\mathcal{O}) \\
&= \sum_{i=1}^{S} \ln (a'_i/a_i) \cdot \frac{a_i \cdot \beta_0^*(i)}{P_\lambda(\mathcal{O})} = \sum_{i=1}^{S} \ln (a'_i/a_i) \cdot C_\lambda(i),
\end{aligned}$$

according to (I).

Secondly,

$$\frac{1}{P_\lambda(\mathcal{O})} \sum_p \sum_{\tau=2}^{N} P_\lambda(p\&\mathcal{O}) \ln (a'_{i_{\tau-1}i_\tau}/a_{i_{\tau-1}i_\tau})$$

$$= \frac{1}{P_\lambda(\mathcal{O})} \sum_{i,j} \sum_{N} \sum_{\tau=2}^{N} \sum_{\substack{p \text{ of length } N, \\ \ni i_{\tau-1}=i \\ i_\tau=j}} P_\lambda(p\&\mathcal{O}) \ln \left(\frac{a'_{i_{\tau-1}i_\tau}}{a_{i_{\tau-1}i_\tau}} \right)$$

$$= \sum_{i,j} \ln \left(\frac{a'_{ij}}{a_{ij}} \right) \cdot \frac{1}{P_\lambda(\mathcal{O})} \sum_{t=1}^{T-1} \sum_{\substack{p \ni i \text{ ends at } t \\ j \text{ starts at} \\ t+1}} P_\lambda(p\&\mathcal{O})$$

$$= \sum_{i,j} \ln \left(\frac{a'_{ij}}{a_{ij}} \right) \cdot C_\lambda(i,j), \quad \text{(from (J))} .$$

Thirdly,

$$\frac{1}{P_\lambda(\mathcal{O})} \sum_p \sum_{\tau=1}^{N} P_\lambda(p\&\mathcal{O}) \ln (P'(d_\tau|i_\tau)/P(d_\tau|i_\tau))$$

$$= \frac{1}{P_\lambda(\mathcal{O})} \sum_{i,d} \ln (P'(d|i)/P(d|i)) \sum_{\substack{p, \tau \ni \\ i_\tau=i \\ d_\tau=d}} P_\lambda(p\&\mathcal{O})$$

$$= \sum_{i,d} \ln (P'(d|i)/P(d|i)) \cdot C_\lambda(i,d), \quad \text{(from (K))} .$$

Finally we consider the last sum. To complete the proof we recall the assumption of independence of the observations within a state. We also need some new notation. If p is a path, $p = (i_1,d_1),(i_2,d_2)\ldots,(i_N,d_N)$, then the τ^{th} segment refers to the observations $\mathcal{O}_p^{(\tau)} = \mathcal{O}_{d_1+\ldots+d_{\tau-1}+1} \cdots \mathcal{O}_{d_1+\ldots+d_\tau}$. Let U_τ denote the set of integers from $d_1 + \ldots + d_{\tau-1} + 1$ to $d_1 + \ldots + d_\tau$, inclusive. Then

$$P(\mathcal{O}_p^{(\tau)}|i_\tau) = \prod_{t \in U_\tau} P(\mathcal{O}_t|i_\tau) .$$

Thus

$$\frac{1}{P_\lambda(\mathcal{O})} \cdot \sum_p \sum_{\tau=1}^{N} P_\lambda(p\&\mathcal{O}) \ln \frac{P'(\mathcal{O}_p^{(\tau)}|i_\tau)}{P(\mathcal{O}_p^{(\tau)}|i_\tau)}$$

$$= \frac{1}{P_\lambda(\mathcal{O})} \sum_p \sum_{\tau=1}^{N} P_\lambda(p\&\mathcal{O}) \sum_{t \in U_\tau} \ln \frac{P'(\mathcal{O}_t|i_\tau)}{P(\mathcal{O}_t|i_\tau)}$$

$$= \sum_{i,k} \ln \frac{P'(k|i)}{P(k|i)} \cdot \frac{1}{P_\lambda(\mathcal{O})} \sum_p \sum_{\substack{\tau \ni \\ i_\tau=i}} \sum_{\substack{t \in U_\tau \\ \ni \mathcal{O}_t=k}} P_\lambda(p\&\mathcal{O})$$

$$= \sum_{i,k} \ln \frac{P'(k|i)}{P(k|i)} \cdot C_\lambda(i,k), \quad \text{according to (L)} .$$

Lemma 3. The reestimates $(\hat{a}_i)$, $(\hat{a}_{ij})$, $(\hat{P}(d|i))$, and $(\hat{P}(k|i))$ given in Section 3 respectively maximize the four summands comprising

$$\frac{Q(\lambda,\lambda') - Q(\lambda,\lambda)}{P}$$

as a function of λ'.

Proof. All four parts are similar. We prove part 4.

We are to maximize

$$\sum_{i,k} C(i,k) \ln \frac{P'(k|i)}{P(k|i)}$$

subject to

$$\sum_{k} P'(k|i) = 1, \quad \text{for} \quad i = 1,2,\ldots,S .$$

Equivalently, maximize

$$\sum_{i,k} C(i,k) \ln P'(k|i) \quad \text{with} \quad \sum_{k} P'(k|i) = 1 .$$

The maximum clearly occurs when the P' are proportional to the C, (e.g., by Lagrange multipliers). Thus

$$P'(k|i) = C(i,k) \Big/ \sum_{k} C(i,k) ,$$

as asserted.

Corollary. The likelihood increases if we replace our current parameters (a_i), (a_{ij}), $(P(d|i))$, $(P(k|i))$ by $(\hat{a}_i)$, $(\hat{a}_{ij})$, $(\hat{P}(d|i))$, $(\hat{P}(k|i))$.

More is in fact true: any convex combination of old and new parameters also increases likelihood.

Lemma 4. Let

$$a_i' = \theta a_i + (1-\theta)\hat{a}_i, \quad 0 \le \theta \le 1 ,$$

$$a_{ij}' = \eta_i a_{ij} + (1-\eta_i)\hat{a}_{ij}, \quad 0 \le \eta_i \le 1 ,$$

$$P'(d|i) = \mu_i P(d|i) + (1-\mu_i)\hat{P}(d|i), \quad 0 \le \mu_i \le 1 ,$$

and

$$P'(k|i) = \nu_i P(k|i) + (1-\nu_i)\hat{P}(k|i), \quad 0 \le \nu_i \le 1 .$$

Then $P_{\lambda'}(\mathcal{O}) \ge P_{\lambda}(\mathcal{O})$ for any choice of θ, (η_i), (μ_i), (ν_i).

Proof.

$$\frac{P_{\lambda'}(\mathcal{O}) - P_{\lambda}(\mathcal{O})}{P_{\lambda}(\mathcal{O})} \ge \frac{Q(\lambda,\lambda') - Q(\lambda,\lambda)}{P_{\lambda}(\mathcal{O})}$$

$$= \sum_{i} C(i) \ln \frac{a_i'}{a_i} + \sum_{i,j} C(i,j) \ln \frac{a_{ij}'}{a_{ij}}$$

$$+ \sum_{i,d} C(i,d) \ln \frac{P'(d|i)}{P(d|i)} + \sum_{i,k} C(i,k) \ln \frac{P'(k|i)}{P(k|i)} .$$

The argument is the same for each sum.

For example,

$$\sum_{i,j} C(i,j) \ln \frac{a'_{ij}}{a_{ij}}$$

$$= \sum_i \sum_j C(i,j) \ln \left(\frac{\eta_i a_{ij} + (1-\eta_i)\hat{a}_{ij}}{a_{ij}} \right)$$

$$\geq \sum_i \sum_j C(i,j) \left(\eta_i \ln \frac{a_{ij}}{a_{ij}} + (1-\eta_i) \ln \frac{\hat{a}_{ij}}{a_{ij}} \right)$$

(by concavity of ln),

$$= \sum_i (1-\eta_i) \sum_j C(i,j) \ln \left(\frac{\hat{a}_{ij}}{a_{ij}} \right) \geq 0$$

by maximality of the $\hat{a}_{ij}$.

5. Posterior estimation procedures

One standard method of finding the "best" state sequence in the usual case of hidden Markov models, is dynamic programming ("the Viterbi algorithm" [2]). We consider how that technique can be adapted to our case.

We wish to find the state sequence

$$p_{max} = (i_1,d_1),\ldots,(i_N,d_N)$$

which has maximum posterior probability. Equivalently,

we maximize $P(p\&\mathcal{O})$ over p. We assume a fixed model, and the path found will be the most likely, for this model, given the observations.

Let $\delta_t(i) = \max \text{Prob}(\mathcal{O}_1 \ldots \mathcal{O}_t)$ over all partial paths p for which state i terminates at time t.

$\delta^*_t(i) = \max \text{Prob}(\mathcal{O}_1 \ldots \mathcal{O}_t)$ over all partial paths p for which state i starts at time t + 1.

Then, analogously to the α,α^* formulas,

$$\delta^*_t(j) = \max_i \delta_t(i) a_{ij} ,$$

$$\delta_t(i) = \max_d \delta^*_{t-d}(i) P(d|i) P(\mathcal{O}_{t-d+1} \ldots \mathcal{O}_t | i) .$$

As usual in dynamic programming, we keep a record of best predecessors, so that we can retrace our steps when we have finished the inductive calculation of $\delta_T(i)$, all i.

An alternative procedure gives (in the case where the model is exactly correct) the path with the maximum expected number of correct state-time assignments. This is also described in [2]. The analogous method here is simply to choose, at time t, that state i_t which maximizes $\gamma_t(i)$. The resulting state sequence (which

may have sudden discontinuities and impossible durations) gives the maximum expected number of correct state-time assignments.

Other procedures can be used to produce decisions with a variety of optimality properties. For example, if we define

$$\gamma_t(i,d) = \frac{\alpha_t^*(i)P(d|i)P(\mathcal{O}_{t+1}\ldots\mathcal{O}_{t+d}|i)\beta_{t+d}(i)}{P}$$

then γ represents the posterior probability that state i occupied the time segment from $t+1$ to $t+d$. If we choose triplets t,i,d with the largest values of $\gamma_t(i,d)$ we expect the maximum number of correctly matched states with segments. Of course, there may be overlaps, and near-duplicates. If a complete path is desired, with maximum expected number of correctly matched state-segment pairs, this can be arranged using a dynamic program (variable duration) to find the path

$$p = (i_1,d_1),\ldots,(i_N,d_N) \ni \gamma_0(i_1,d_1) + \gamma_{d_1}(i_2,d_2) + \gamma_{d_1+d_2}(i_3,d_3) + \ldots$$

is maximal.

This path, while complete, may have impossible state transitions $i_\tau \to i_{\tau+1}$ in it. If it is desired to eliminate these, then a dynamic program can again be employed to find the path with all transitions having nonzero probability, and maximal sum of gammas for segments. The ideas of this section will allow us to estimate, in the Poritz scheme [5] extended, which phone class is (probably) operative at each moment, and when transitions occur (probably). This amounts to probabilistic labeling and segmentation.

6. Special duration distributions

There are a few special duration distributions for which reestimation is particularly simple. For example, suppose that, for state i, the duration distribution is Poisson, with mean $\mu_i + 1$, starting at 1. Thus

$$P(\text{Duration} = 1 + d|i) = e^{-\mu_i} \cdot \frac{\mu_i^d}{d!}, \quad d = 0,1,\ldots .$$

This distribution is potentially infinite, but for the problem at hand, we know $d \leq T$. Also, if we impose an arbitrary upper bound $d \leq D_{max}$, there may be very little loss. In any case, we can show that the reestimate for the duration distribution (by maximizing the

Q-function) depends only on the expected average duration spent in state i:

$$\hat{\mu}_i = \frac{\Sigma_t \gamma_t(i)}{S_T(i)} - 1 \quad \text{(see (F) \& (G))} \quad .$$

Lemma 5. If

$$P'(1+d|i) = e^{-\mu_i'} \frac{\mu_i'^d}{d!} \quad ,$$

then $Q(\lambda,\lambda')$ is maximized, as a function of μ_i', at

$$\mu_i' = \frac{\Sigma_t \gamma_t(i)}{S_T(i)} - 1 \quad .$$

Proof.

$$\frac{Q(\lambda,\lambda') - Q(\lambda,\lambda)}{P(\mathcal{O})} = \sum_i C(i) \ln \frac{a_i'}{a_i} + \ldots$$

(according to Lemma 2). We are only interested in maximizing

$$\sum_d C(i,d) \ln \frac{P'(d|i)}{P(d|i)}$$

since the other terms either do not involve state i, or do not refer to duration. Thus we wish to maximize

$$F(\mu_i') = \sum_d C(i,d) \ln \left[e^{-\mu_i'} \cdot \frac{\mu_i'^{d-1}}{(d-1)!} \right]$$

$$= \sum_d C(i,d)[-\mu_i' + (d-1) \ln \mu_i' - \ln (d-1)!] \quad .$$

Now

$$\frac{\partial F}{\partial \mu_i'} = -\sum_d C(i,d) + \sum_d \frac{(d-1)C(i,d)}{\mu_i'} \quad .$$

The maximum occurs at

$$\mu_i' = \frac{\Sigma_d (d-1)C(i,d)}{\Sigma_d C(i,d)}$$

$$= \frac{\Sigma_d d \cdot C(i,d)}{\Sigma_d C(i,d)} - 1 \quad .$$

Here $\Sigma_d d \cdot C(i,d)$ is the expected total duration spent in state $i = \Sigma_t \gamma_t(i)$, and $\Sigma_d C(i,d)$ is the expected number of segments in which i was the state $= S_T(i) = E_{T+1}(i)$.

Of course, similar arguments apply whenever the maximum of the Q-function just depends on the average duration, for example:

$P(d|i)$ can be geometric:

$$P(d|i) = \rho_i^{d-1}(1-\rho_i) \quad \text{for some } \rho_i ,$$

or binomial on the range 1 to $M_i + 1$, (where M_i is fixed and known):

$$P(1+d|i) = \binom{M_i}{d} \rho_i^d (1-\rho_i)^{M_i-d}, \quad d = 0,\ldots,M_i .$$

We observe, following [3], that whenever the maximum likelihood parameters for $P(d|i)$ can be determined, given the number of occurrences $n(i,d)$ for each d, then the maximum of the Q-function can be similarly found, simply by replacing the numbers $n(i,d)$ by the expected numbers $C(i,d)$. This maximization of Q need not, of course, be merely algebraic, as above, but might entail an iterative approximative scheme.

7. Extension to continuously distributed output

As described in [1], and extended in [3], it is not hard to replace $P(k|i)$ by a probability density function on continuous, possibly multivariate observations. For application to the Poritz scheme, [5], or to the ideas proposed in the next paper, by Liporace [7], this extension of the variable duration idea to continuously distributed output is necessary.

We illustrate by allowing the output to be Gaussian distributed (in one-dimension for simplicity) with mean μ_i, variance σ_i^2, in state i. We again assume independence of the observations occurring while in state i. Our conclusion is that the reestimate which maximizes the Q-function is

$$\boxed{\begin{aligned} \hat{\mu}_i &= \frac{\Sigma_t \mathcal{O}_t \gamma_t(i)}{\Sigma_t \gamma_t(i)} , \\ \hat{\sigma}_i^2 &= \frac{\Sigma \mathcal{O}_t^2 \gamma_t(i)}{\Sigma_t \gamma_t(i)} - \hat{\mu}_i^2 . \end{aligned}}$$

For the proof, we need an analogue of Lemma 2 for the case of continuous output.

Lemma 2′.

$$\begin{aligned} \frac{Q(\lambda,\lambda') - Q(\lambda,\lambda)}{P_\lambda(\mathcal{O})} &= \sum_i C(i) \ln \frac{a_i'}{a_i} \\ &+ \sum_{i,j} C(i,j) \ln \frac{a_{ij}'}{a_{ij}} \\ &+ \sum_{i,d} C(i,d) \ln \frac{P'(d|i)}{P(d|i)} \\ &+ \sum_{i,t} \gamma_t(i) \ln \frac{P'(\mathcal{O}_t|i)}{P(\mathcal{O}_t|i)} . \end{aligned}$$

(This formula holds in both the continuous and discrete cases interpreting $P'(\mathcal{O}_t|i)$ and $P(\mathcal{O}_t|i)$ as likelihoods, but the Lemma 2 form is simpler in the discrete case.)

Proof. We can take up the proof of Lemma 2 in the consideration of the last sum:

$$H = \frac{1}{P_\lambda(\mathcal{O})} \cdot \sum_p \sum_{\tau=1}^{N} P_\lambda(p\&\mathcal{O}) \sum_{t\in U_\tau} \ln \frac{P'(\mathcal{O}_t|i_\tau)}{P(\mathcal{O}_t|i_\tau)} \quad .$$

We have:

$$H = \frac{1}{P_\lambda(\mathcal{O})} \sum_i \sum_t \sum_{\substack{p\ni \text{ state} \\ i \text{ is used} \\ \text{at time } t}} P_\lambda(p\&\mathcal{O}) \ln \frac{P'(\mathcal{O}_t|i)}{P(\mathcal{O}_t|i)}$$

$$= \frac{1}{P_\lambda(\mathcal{O})} \sum_{i,t} \ln \frac{P'(\mathcal{O}_t|i)}{P(\mathcal{O}_t|i)} \cdot \sum_{\substack{p\ni \text{state} \\ i \text{ is used} \\ \text{at time } t}} P_\lambda(p\&\mathcal{O})$$

$$= \sum_{i,t} \ln \frac{P'(\mathcal{O}_t|i)}{P(\mathcal{O}_t|i)} \cdot \gamma_t(i),$$

as asserted.

We now employ Lemma 2′ for the Gaussian example at hand. Here

$$P'(\mathcal{O}_t|i) = \frac{1}{\sqrt{2\pi}\,\sigma_i'} \, e^{-1/2 \frac{(\mathcal{O}_t-\mu_i')^2}{\sigma_i'^2}} \quad .$$

Thus

$$H = \sum_{i,t} \gamma_t(i) \ln \frac{P'(\mathcal{O}_t|i)}{P(\mathcal{O}_t|i)}$$

$$= \sum_{i,t} \gamma_t(i) \left[- \ln \sigma_i' - \frac{1}{2} \frac{(\mathcal{O}_t-\mu_i')^2}{\sigma_i'^2} + \ln \sigma_i + \frac{1}{2} \frac{(\mathcal{O}_t-\mu_i)^2}{\sigma_i^2} \right] .$$

Hence

$$\frac{\partial H}{\partial \mu_i'} = \sum_t \gamma_t(i) \frac{(\mathcal{O}_t-\mu_i')}{\sigma_i'^2} \quad ,$$

and

$$\frac{\partial H}{\partial \sigma_i'} = -\sum_t \frac{\gamma_t(i)}{\sigma_i'} + \sum_t \frac{(\mathcal{O}_t-\mu_i')^2}{\sigma_i'^3} \gamma_t(i) \quad .$$

The maximum occurs at

$$\hat{\mu}_i = \sum_t \gamma_t(i) \, \mathcal{O}_t \Big/ \sum_t \gamma_t(i) \quad ,$$

$$\hat{\sigma}_i^2 = \sum_t (\mathcal{O}_t - \hat{\mu}_i)^2 \cdot \gamma_t(i) \Big/ \sum_t \gamma_t(i)$$

$$= \sum_t \mathcal{O}_t^2 \cdot \gamma_t(i) \Big/ \sum_t \gamma_t(i) - \hat{\mu}_i^2 \ ,$$

as asserted.

This analytic approach to maximizing Q is applicable to a variety of simple continuous distributions, including the multi-variate Gaussian, but, as shown in [3], all that is really necessary is the ability to find maximum likelihood estimators, given actual observations, analytically, or otherwise.

8. Duration 0

In the preceding analyses, durations were required to be at least one. In some special cases it is useful to allow a state to have duration 0. In these cases, the inductive formulas for $\alpha, \alpha^*, \beta, \beta^*$ go awry, and special methods must be used. For example, if we have a hidden Markov model for a specific word or sentence, then a definite state order is assumed. But it is quite possible for a speaker to omit, entirely, one or more of the phones in a sequence, especially when excited. For this reason we may wish to allow a state or states to

have 0 duration. In this particular example, it is possible to accommodate 0 duration in a simple way. Since the states are ordered, there is no way to return to an earlier state in the assumed sequence, thus obtaining an infinite loop of 0 duration states. The argument below describes how to handle 0 duration even in the general case, where infinite loops can occur.

Suppose we have a variable duration hidden Markov model with parameters (a_i), (a_{ij}), $(P(d|i),$ $d = 0,1,\ldots,D)$, $(b_j(k))$. Suppose we also have a sequence of observations $\mathcal{O} = \mathcal{O}_1\mathcal{O}_2\ldots\mathcal{O}_T$. We show how to compute the probability of $\mathcal{O}$, given the model. Let $\bar{P}(d|i) = P(d|i)/(1-P(0|i))$ for each i, and $d > 0$. Thus $\bar{P}(d|i) = \text{Prob}(\text{duration is } d, \text{ given that it is not } 0)$. Similarly, let $\bar{a}_{ij} = \text{Prob}(j \text{ is the next state after } i \text{ for which the duration is not } 0)$. Then

$$\bar{a}_{ij} = a_{ij} \cdot (1-P(0|j))$$

$$+ \sum_k a_{ik}P(0|k) \cdot a_{kj}(1-P(0|j))$$

$$+ \sum_{k,\ell} a_{ik}P(0|k) \cdot a_{k\ell}P(0|\ell) \cdot a_{\ell j}(1-P(0|j)) + \cdots$$

This infinite sum can either be approximated, or computed exactly by summing an infinite series of powers of the matrix $M = (a_{k\ell}P(0|\ell))$, obtaining $(I-M)^{-1}$, etc.

Lastly, we define $\bar{a}_i$ = Prob(first state with nonzero duration is i)

$$= a_i(1-P(0|i)) + \sum_j a_j P(0|j) a_{ji} (1-P(0|i))$$

$$+ \sum_{k,j} a_k P(0|k) \cdot a_{kj} P(0|j) \cdot a_{ji} (1-P(0|i))$$

$$+ \ldots ,$$

which is handled similarly.

Then the hidden Markov model with parameters $(\bar{a}_i)$, $(\bar{a}_{ij})$, $(\bar{P}(d|i), d = 1,2,\ldots,D)$, and $(b_j(k))$ can be treated in the standard manner, and allow the computation of the probability of $\mathcal{O}$.

Alternative approaches might involve structuring the Markov chain so that no infinite sequences of 0-duration states can occur, or approximating the model by truncating all such chains to a small finite length.

REFERENCES

[1] Baum, Leonard, E., Petrie, Ted, Soules, George, and Weiss, Norman, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, Annals of Mathematical Statistics, Vol. 41, No. 1, pp. 164-171, 1970.

[2] Forney, G. David, Jr., The Viterbi Algorithm, Proceedings of the IEEE, Vol. 61, No. 3. pp. 268-278, March 1973.

[3] Dempster, A. P., Laird, N. M., Rubin, D. B., Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, Series B (Methodological), Vol. 39, pp. 1-38, 1977.

[4] Cave, Robert L., Neuwirth, Lee P., Hidden Markov models for English, these Proceedings.

[5] Poritz, Alan B., Linear predictive hidden Markov models, these Proceedings.

[6] Liporace, Louis A., Maximum likelihood analysis for multivariate observations of Markov sources, these Proceedings.

[7] Liporace, Louis A., Statistical analysis of time series, these Proceedings.