```
     07/18/2023 08:53    RECAP LIBRARY ANNEX PRODUCTION SYSTEM          PG   18
     RPT# 02376011 ah9692                                               cpyrept.

                                  EDD RETRIEVAL RECEIPT


          Order: 652593
            For: EDD
         Copied: 07/18/2023
        Shipped: 07/18/2023
     Deliver To: 198572131
  Patron E-Mail: rdhawkins@Princeton.EDU
Oth Patron Info:
  Def PickUp Loc: EDD-ReCAP EDD
  Delivery Meth: EDD


   Item BarCode: 32101105432817
     Item Title: Pattern recognition in practice : proceedings of an internat
    Item Author:
Item Call Number: Q327 .P38
  Item Vol/Part:


  Article Title: Interpolated estimation of Markov source parameter
 Article Author: Jelinek, Frederick and Mercer, Robert L.
   Art Vol/Part: ,
       Beg Page: 381      End Page: 397           Total Pages: 0
     Other Info:
          Notes: I've looked for this paper everywhere! Thanks!









      TOTAL COUNT: 1
```

Pattern Recognition in Practice
E.S. Gelsema and L.N. Kanal (eds.)
© North-Holland Publishing Company, 1980

INTERPOLATED ESTIMATION OF MARKOV SOURCE PARAMETERS FROM SPARSE DATA

F. Jelinek and R. L. Mercer

Continuous Speech Recognition Group
Computer Sciences Department
IBM T. J. Watson Research Center
Yorktown Heights, New York 10598

In this paper we introduce a new method of estimating transition probabilities of Markov source models from given sparse data. We first review the forward-backward algorithm which derives parameter estimates having maximum likelihood properties relative to the data. We consider the problem of predicting yet unseen data. We give two heuristic methods based on the interpolated estimator concept that address the problem and present results of simulations confirming the viability of our approach. The algorithms were successfully applied in continuous speech recognition to estimation of parameters of speech processes.

1. Introduction

A Markov source is a collection of states connected to one another by transitions which produce symbols from a finite alphabet. Each transition t from a state s has associated with a probability $q_s^{(t)}$ that t will be taken after s is reached. The Markov source assigns probabilities to all strings of transitions from any chosen state s to any state s'.

We define a Markov source more formally as follows. Let $\mathscr{S}$ be a finite set of states, $\mathscr{T}$ a finite set of transitions, and $\mathscr{A}$ a finite alphabet. The structure of a Markov source is a $1-1$ mapping M: $\mathscr{T}\rightarrow\mathscr{S}\times\mathscr{A}\times\mathscr{S}$. If $M(t)=(\ell,a,r)$ then we refer to ℓ as the predecessor state of t, a as the output symbol associated with t, and r as the successor state of t; we write $\ell=L(t)$, $a=A(t)$, and $r=R(t)$.

The parameters of a Markov source are transition probabilities $q_s(t), s\in\mathscr{S}, t\in\mathscr{T}$, such that

$$q_s(t) = 0 \qquad \text{if } s\neq L(t)$$

and

$$\sum_t q_s(t) = 1, \qquad s\in\mathscr{S} \tag{1}$$

and initial probabilities Q(s), $s\in\mathscr{S}$ such that

$$\sum_{s \in \mathscr{S}} Q(s) = 1. \tag{2}$$

Frequently it is convenient to designate an initial state s_I such that $Q(s_I) = 1$.

Figure 1 is an example of a Markov source given in the usual diagrammatic form indicating output symbols associated with transitions.

In general, the transition probabilities associated with one state are different from those associated with another. This need not always be the case, however. We say that state s_1 is tied to state s_2 if there exists a 1−1 correspondence $T_{s_1 s_2}: \mathscr{T} \to \mathscr{T}$ such that $q_{s_1}(t) = q_{s_2}(T_{s_1 s_2}(t))$ for all transitions t. It is easily verified that the relationship of being tied in an equivalence relation and hence induces a partition of $\mathscr{S}$ into sets of states which are mutually tied together.

A string of n transitions t_1^n(*) for which $L(t_{i+1}) = R(t_i)$, $i = 1, \ldots, n-1$ is called a path. The probability of a path t_1^n is given by

$$P(t_1^n) = Q(L(t_1)) q_{L(t_1)}(t_1) \prod_{i=2}^{n} q_{R(t_{i-1})}(t_i) \tag{3}$$

Associated with path t_1^n is an output sumbol string $a_1^n = A(t_1^n)$. A particular output string a_1^n, may in general arise from more than one path. Thus the probability $P(a_1^n)$ is given by

$$P(a_1^n) = \sum_{t_1^n} P(t_1^n) \delta(A(t_1^n), a_1^n) \tag{4}$$

where

$$\delta(a,b) = \begin{cases} 1 \text{ if } a = b \\ 0 \text{ otherwise.} \end{cases} \tag{5}$$

In practice it is useful to allow transitions which produce no output. These null transitions are represented diagrammatically by interrupted lines (see Figure 2). Rather than deal with null transitions directly, we have found it convenient to associate with them the distinguished letter ϕ. We then add to the Markov source a filter (see Figure 3) which removes ϕ, transforming the output sequences a_1^n into an observed sequence b_1^m, where $b_i \in \mathscr{B} = \mathscr{A} - \{\phi\}$. Although more general sources can be handled, we shall restrict our attention to sources which do not have closed circuits of null transitions.

(*) t_1^n is a short-hand notation for the concatenation of the symbols $t_1, t_2, \ldots, t_n$. Strings are indicated in bold face throughout.

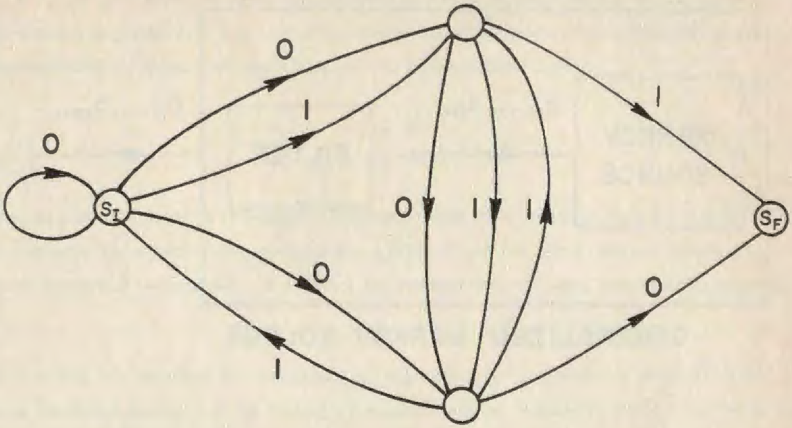

Figure 1
A binary Markov source with output-transition associations indicated.

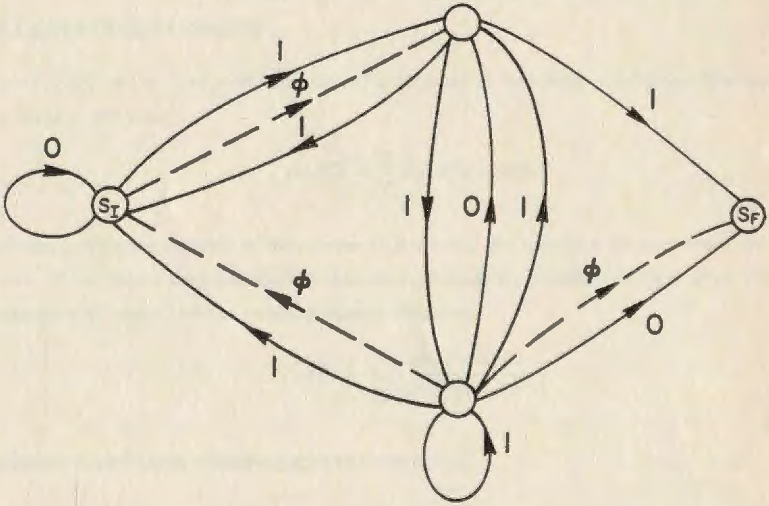

Figure 2
A binary Markov source with null transitions.

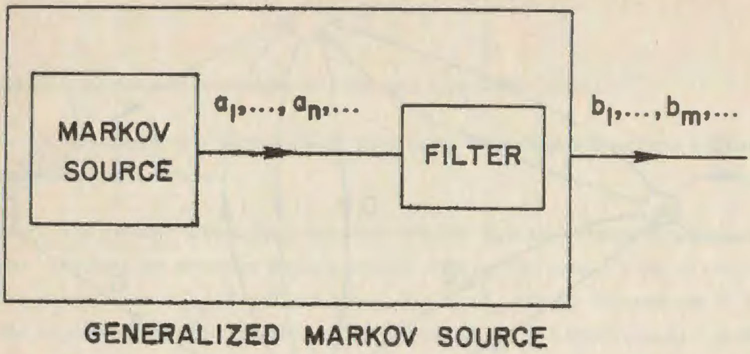
MARKOV SOURCE
$a_1, \ldots, a_n, \ldots$
FILTER
$b_1, \ldots, b_m, \ldots$
GENERALIZED MARKOV SOURCE

Figure 3

A generalized Markov source.
Filter deletes ϕ symbols from the a-sequence.

If t_1^n is a path which produces the observed sequences b_1^m then we say that b_i spans t_j if t_j is the transition which produced b_i or if t_j is a null transition immediately preceding a transition spanned by b_i. In the diagram below, b_1 spans t_1;b_2 spans t_2,t_3, and t_4; and b_3 spans t_5 and t_6.

$$\begin{array}{cccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \\ b_1 & \phi & \phi & b_2 & \phi & b_3 \end{array}$$

The purposee of this paper is to suggest algorithms which from observed data b_1^n attempt to model its actual generator by adjusting the probabilities $q_s(t)$,$s \in \mathscr{S}$, of the given Markov source (i.e., having a specified structure $\mathscr{T} \rightarrow \mathscr{S} \times \mathscr{A} \times \mathscr{S}$. It will not be assumed that the data was actually generated by the source.

In the next two sections we recapitulate an algorithm [1] that seeks to account for the data in a maximum likelihood manner, i.e., by finding $q_s(t)$ maximizing the probability $P(b_1^n)$ given by an expression analogous to (3) that takes into account the existence of null transitions. The rest of the paper is devoted to the sparse data situation where maximum likelihood estimation of b_1^m does not provide a satisfactory model of the underlying generator.

2. The Forward-Backward Algorithm

Let $P_i(t,b_1^m)$ be the joint probability that b_1^m is observed at the output of a filtered Markov source and that b_i spans t. The *count*

$$c(t,b_1^m) \triangleq \sum_{i=1}^{m} P_i(t,b_1^m)/P(b_1^m) \tag{6}$$

is the Bayes a *posteriori* estimate of the number of times that the transition t is used when the string b_1^m is produced. If the counts are normalized so that the total count for transitions from a given state is 1, then it is reasonable to expect that the resulting relative frequency

$$f_s(t,b_1^m) \triangleq \frac{c(t,b_1^m)\delta(s,L(t))}{\sum_{t'} c(t',b_1^m)\delta(s,L(t'))} \tag{7}$$

will approach the transition probability $q_s(t)$ as m increases.

This suggests the following iterative procedure for obtaining estimates of $q_s(t)$.

1. Make initial guesses $q_s^o(t)$ and $Q^o(s)$.

2. Set $j=0$.

3. Compute $P_i(t,b_1^m)$ for all i and t using $q_s^j(t)$ and $Q^j(s)$.

4. Compute $f_s(t,b_1^m)$ and obtain new estimates $q_s^{j+1}(t) = f_s(t,b_1^m)$ and $Q^{j+1}(s) = \sum_{L(t)=s} P_1(t,b_1^m) \mid \sum_{t'} P_1(t',b_1^m)$.

5. Set $j=j+1$.

6. Repeat from 3.

To apply this procedure, we need a simple method for computing $P_i(t,b_1^m)$. Now $P_i(t,b_1^m)$ is given by three terms: the probability $P\{B_1^{i-1} = b_1^{i-1}, S_{i-1} = L(t)\}$ that a string of transitions ending in $L(t)$ will produce the observed sequence b_1^{i-1}, times the probability that t will be taken once $L(t)$ is reached, times the probability that a string of transitions starting with $R(t)$ will produce the remainder of the observed sequence. If $A(t)=\phi$, then the remainder of the observed sequence is b_i^m, if $A(t)\neq\phi$ then, of course $A(t) = b_i$ and the remainder of the observed sequence is b_{i+1}^m. Thus if $\alpha_i(s)$ denotes the probability of producing the observed sequence b_1^i by a sequence of transitions ending in the state s, and $\beta_i(s)$ denotes the probability of producing the observed sequence $b_i^m$by a string of transitions starting from the state s, then

$$P_i(t,b_1^m) = \begin{cases} \alpha_{i-1}(L(t)) \; q_{L(t)}(t) \; \beta_i(R(t)) & \text{if } A(t) = \phi \\ \alpha_{i-1}(L(t)) \; q_{L(t)}(t) \; \beta_{i+1}(R(t)) & \text{if } A(t) = b_i \end{cases} \tag{8}$$

where we defined

$$\alpha_i(s) \triangleq P\{B_1^i = b_1^i, S_i = s\} \tag{9}$$

and

$$\beta_i(s) \triangleq P\{B_i^n = b_i^n \mid S_{i-1} = s\} \tag{10}$$

In (9) and (10) we use the usual notation associated with random variables which are denoted by capitals.

It follows from the Markovian nature of the source that for i≥1

$$P\{B_1^i = b_1^i, S_i = s\} = \sum_{s'} P\{B_1^{i-1} = b_1^{i-1}, S_{i-1} = s'\} \; P\{T{:}A(T) = b_i, R(T) = s \mid S_{i-1} = s'\}$$

$$+ \sum_{s''} P\{B_1^i = b_1^i, S_i = s''\} P\{T{:}A(T) = \phi, R(T) = s \mid S_i = s''\} \tag{11a}$$

and for $i=1$,

$$P\{S_0 = s\} = Q(s) + \sum P\{T{:}A(T) = \phi, R(T) = s \mid S_0 = s'\}P\{S_0 = s'\} \tag{11b}$$

Let us denote the probability that a transition t *ending* in state s and producing the output a is taken by

$$\gamma(t,s,a) \overset{\Delta}{=} q_{L(t)}(t) \;\; \delta(R(t),s) \;\; \delta(A(t),a) \tag{12}$$

Then from (11) and (12) we get the equations

$$\alpha_0(s) = Q(s) + \sum_t \alpha_0(L(t))\gamma(t,s,\phi) \tag{13a}$$

$$\alpha_i(s) = \sum_t \alpha_{i-1}(L(t))\gamma(t,s,b_i) + \sum_t \alpha_i(L(t))\gamma(s,t,\phi) \quad i \geq 1 \tag{13b}$$

Similarly, denoting the probability that a transition t *starting* in state s and producing output a is taken by

$$\xi(t,s,a) = q_{L(t)}(t) \;\; \delta(L(t),s) \;\; \delta(A(t),a). \tag{14}$$

we get the equations

$$\beta_m(s) = 1 \tag{15a}$$

$$\beta_i(s) = \sum_t \beta_i(R(t)) \;\; \xi(t,s,\phi) + \sum_t \beta_{i+1}(R(t)) \;\; \xi(t,s,b_i) \quad i \leq m, \tag{15b}$$

Since α_i appears on both sides of (13), for the latter to be a recursion one must find a sequence of states s in which (13) can be unambiguously evaluated. The requirement that there be no null transition loops makes this possible.

Let $\mathscr{S}_1$ be those states that cannot be reached by a null transition. Let $\mathscr{S}_i, 1=2,3,...,\tau$ be all the states that can be reached by a single null transition from states in $\mathscr{S}_1,...,\mathscr{S}_{i-1}$ and from no other states. Then by (12) $\gamma(t,s,\phi) = 0$ for all $L(t)\epsilon\mathscr{S}_i, s\epsilon\mathscr{S}_j, j \geq i$. It is then possible to evaluate (13) first for all $s\epsilon\mathscr{S}_1$, then for $s\epsilon\mathscr{S}_2$, etc., and finally for $s\epsilon\mathscr{S}_\tau$. For the same reason (15) is a recursion since it can be evaluated first for $s\epsilon\mathscr{S}_\tau$, then for $s\epsilon\mathscr{S}_{\tau-1}$, etc., and lost for $s\epsilon\mathscr{S}_1$.

We now know the correct way of carrying out step 3 of the above iterative procedure. α_i is computed in a forward pass over the data, β_i in a backward pass over the data, and finally $P_i(t,b_1^m)$ from equation (8). We refer to the iterative procedure together with the method described for computing $P_i(t,b_1^m)$ as the Forward-Backward (F-B) algorithm.

3. Maximum Likelihood Properties of the F-B Algorithm

The probability, $P(b_1^m)$, of the observed sequence b_1^m is a function of the probabilities $q_s(t)$. To display this dependence explicitly, we write $P(b_1^m,q_s(t))$. L. E. Baum [1] has proven that $P(b_1^m,q_s^{j+1}(t))\geq P(b_1^m,q_s^j(t))$ with equality only if $q_s^j(t)$ is a stationary point of $P(b_1^m,q_s^j(t))$. This result also holds if the transition distributions of some of the states are held fixed or if some of the states are tied to one another thereby reducing the number of independent transition distributions.

Since tying of states implies equality of certain transition probabilities (see Section 1), it is necessary to modify appropriately the Forward-Backward algorithm. Let the state space $\mathcal{S}$ be partitioned into subsets $\mathcal{S}^*_i, i = 1,2,\ldots,r$ of tied states. This partition plus the tying functions T_{s,s_2} induce a partition of the transition set $\mathcal{T}$ into tied subsets $\mathcal{T}^*_j, j = 1,2,\ldots,k$ as follows: t and t' belong to the same subset $\mathcal{T}^*_j$ if $L(t)\epsilon\mathcal{S}^*_i$ and $L(t')\epsilon\mathcal{S}^*_i$ for some i, and $t' = T_{L(t),L(t')}(t)$. Since a value of i exists such that for all $t^*\epsilon\mathcal{T}^*_j, L(t^*)\epsilon\mathcal{S}^*_i$, it is possible to define a predecessor function L^* such that $L^*(\mathcal{T}^*_j) = \mathcal{S}^*_i$. Then counters $c^*(\mathcal{T}^*_j,b_1^m)$ can be established containing the contributions for transtions $t\epsilon\mathcal{T}^*_j$:

$$c^*(\mathcal{T}^*_j,b_1^m) = \sum_{t\epsilon\mathcal{T}^*_j} c(t,b_1^m) \tag{16}$$

The new transition probability values $q_s^j(t)$ are then obtained from the relative frequencies

$$f_s(t,b_1^m) = \frac{c^*(\mathcal{T}^*_j,b_1^m)\delta(\mathcal{S}^*_i,L^*(\mathcal{T}^*_j))}{\sum_{h=1}^{k} c^*(\mathcal{T}^*_h,b_1^m)\delta(\mathcal{S}^*_i,L^*(\mathcal{T}^*_h))} \tag{17}$$

where $t\epsilon\mathcal{T}^*_j$.

It is also possible to show that the F-B algorithm converges for Gaussian Markov sources, whose outputs are real vectors x. Here to each transition of t there corresponds a mean vector m_t and a covariance matrix Σ_t and when the source undergoes transition t the output vector x is chosen at random, with normal density of mean m_t and covariance Σ_t. Baum [1, Theorem 4] proved that the above maximum likelihood property holds when an "obvious" generalization of the F-B algorithm is used to estimate the parameters m_t,Σ_t (for more detail, see [1]). Again, states may be tied by inducing constraints $m_t = m_{t'}$, and $\Sigma_t = \Sigma_{t'}$.

The question naturally arises whether the distributions $q_s(t)$ can be said to approach the "true" distributions underlying a Markov source that actually produced the observed data b_1^m. Baum and Petri [2] show that under certian quite restricted conditions on the source, as $m\rightarrow\infty$, the critical point achieving the maximum $P(b_1^m)$ would correspond to a distribution assignment $q_s(t)$ equivalent ot the underlying distribution (equivalent in the sense that for all b_1^k the value of $P(b_1^k)$ is the same for the equivalent and underlying source). This desirable property of the F-B algorithm is unfortunately quite illusory from the practical standpoint since (i) it presupposes knowledge of the exact structure of the "true" Markov source, and (ii) it presupposes either unimodality of the parameter space $q_s(t)$ or the existence of a test that the critical point reached by the F-B algorithm is one of those that maximize $P(b_1^m)$ over all critical points. In most real situations, of course, even if the data are generated by a Markov source in the first place, the structure of the latter is unknown and its state space is too large.

The practical effectiveness of the F-B algorithm has been proven in many applications. Some experimental results in the area of speech process modeling can be found in [3].

4. The Sparse Data Problem

In most real situations, Markov source modeling is not done to account best for the observed data b_1^m, but to predict some as yet unseen (and much larger!), output of the unknown generator. Maximum likelihood modeling is very vulnerable to inadequate data. Suppose that the data b_1^m is generated by an unknown Markov source having the same structure as the model, and that for some letter $z \epsilon$)$\mathscr{A}$ there is a unique t leaving a state $s \epsilon \mathscr{S}$ $(L(t) = s)$ such that $A(t) = z$. If $q_s(t)$ is positive but small, it is possible that the source neer took transition t when generating b_1^m. Then regardless of the initial value $q_s^0(t)$, the first iteration of the F-B algorithm may result in counter value $c(t, b_1^m) = 0$. As a consequence, the final estimate of $q_s(t)$ will be zero which for many applications may be fatal to the model. The case of $c(t, b_1^m) = 0$ for many t is quite frequent in practice. Of course, there is a wide variety of less drastic convergence problems which may also arise.

If the obvious step of increasing the training size cannot be taken, then one possibility is to reduce the complexity of the model. This can be done either by restructuring it so as to make the state space $\mathscr{S}$ smaller, or by partitioning $\mathscr{S}$ into subsets $\mathscr{S}^*_i$ of tied states and thereby reducing the number of parameters to be estimated. Such tying would, of course, take place on the basis of some "physical" information regarding the underlying data generator that would allow the imposition of the equivalence structure implied by the tying. Indeed, it is such information that leads to the experimenter's original structural design.

We are interested in how far the "tying-of-states" approach can be taken. In general, one wishes to base one's predictions on as specific a set of circumstances as can be described, but such predictions will be reliable only if these circumstances occurred sufficiently frequently (hence the tying). The following approach to solving this dilemma seems reasonable.

Let $\hat{q}_s(t)$ be Forward-Backward estimates of the transition probabilities based on b_1^m and let $^*\hat{q}_s(t)$ be the corresponding estimates obtained when certain of the states are assumed to be tied. Where the estimates $\hat{q}_s(t)$ are unreliable, we would like to fall back on the more reliably estimated $^*\hat{q}_s(t)$, but where

$\hat{q}_s(t)$ is reliable we would like to use it directly.

A convenient way to achieve this is to choose as final estimates of $q_s(t)$ a linear combination of $\hat{q}_s(t)$ and ${}^*\hat{q}_s(t)$. Thus we let $q_s(t)$ be given by

$$q_s(t) = \lambda_s \hat{q}_s(t) + (1-\lambda_s){}^*\hat{q}_s(t) \tag{18}$$

with λ_s chosen close to 1 when $\hat{q}_s(t)$ is reliable and close to 0 when it is not.

Figure 4 shows the part of the transition structure of the Markov source related to the state s. Equation (18) can be interpreted in terms of the interpolated Markov source shown in Figure 4b in which each state is replaced by three states. In Figure 4, $\hat{s}$ corresponds directly to s in Figure 4. The null transitions from $\hat{s}$ to s and s^* have transition probabilities equal to λ_s and $1-\lambda_s$ respectively. The

The transitions out of s have probabilities $\hat{q}_s(t)$ while those out of s^* have probabilities ${}^*\hat{q}_s(t)$. The structure of the interpolated Markov source is completely determined by the structure of the original Markov source and by the tyings assumed for obtaining more reliable parameter estimates.

The interpretation of (18) as an interpolated Markov source immediately suggestes that the parameters λ_s be determined by the Forward-Backward (F-B) algorithm. However, since the λ parameters were introduced to predict as yet unseen data, rather than to account for the training data b_1^m, the F-B algorithm must be modified. We wish to extract the λ values from data that was not used to determine the distributions $\hat{q}_s(t)$ and ${}^*\hat{q}_s(t)$, otherwise we will again only account for the observed sparse data b_1^m and will not be able to predict well the future data.

An immediate heuristic remedy is that of the *held out* interpolator. Namely, we divide the available data b_1^m into two pairs, b_1^ℓ and $b_{\ell+1}^m$, and then determine $\hat{q}_s$ and $\hat{q}^*{}_s$ from b_1^ℓ, and the interpolation transition probabilities λ_s from the remaining data $b_{\ell+1}^m$. The immediate question then is how large ℓ should be for a given value of m, and what $\hat{q}_s$ and $\hat{q}^*{}_s$ values should be used in the final model: those determined from b_1^ℓ or from b_1^m? Another remedy is that of a *deleted* interpolator which in some sense uses all of the data all of the time. The idea is to compute for every $i=1,\dots,m$ the probabilites $P_i(t,b_1^m)$ used in formula (6) on the basis of $\hat{q}_s$ and $\hat{q}^*{}_s$ values that were obtained by the F-B algorithm from data b_1^m from which a substring including b_i had been eliminated. The precise way to accomplish this is intimately tied to the F-B algorithm and will be described in the next section.

The heuristic justification of the deleted interpolator approach is that $P_i(t,b_1^m)$ concerns the transition t that might have "caused" b_i, and if it is computed as suggested above, then b_i has the character of future,

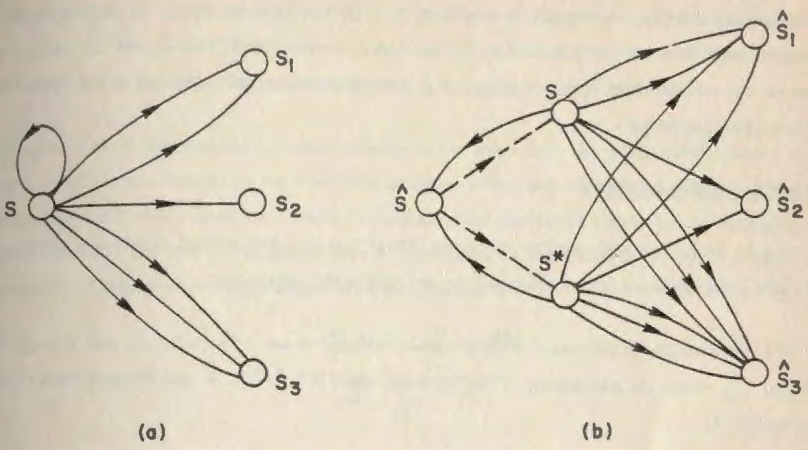

Figure 4

Part (a) represents a segment of a Markov source. Part (b) represents the corresponding segment of an associated interpolated Markov source. $\hat{s}$ denotes the new state corresponding to s with null transitions leaving it. s^* represents the new state corresponding to s whose output transitions have the tied distribution ${}^*\hat{q}_s$.

as yet unobserved data. Thus, the ith contribution in (6) to the value of λ_s will depend on the relative quality of predictions of the deleted data b_i (analogous to the "future") by $\hat{q}_s$ and $^*\hat{q}_s$ whose values are based on the retained data $b_1,\ldots,b_{i-\ell},b_{i+k},\ldots,b_m$ (analogous to the data observed in the "past"), where ℓ and k are positive integers.

5. Fixed Deleted Interpolation

In this section we will describe the *fixed* deleted interpolation method of Markov source modeling. The more general *relaxed* interpolation will be sketched in the next section.

We partition the state space $\mathcal{S}$ into subsets of tied states $\mathcal{S}^*_i, i = 1,\ldots,r$, and through the tying function $T_{s,s'}$ obtain the partitioning of the transition space $\mathcal{T}^*_j, j = 1,\ldots k$, and the predecessor function L^* (see Section 3).

We divide the data b_1^m into n blocks of length $\ell(m{=}\ell n)$. We run the F-B algorithm in the ordinary way, but on the last iteration we establish separate counters.

$$c_j(t,b_1^m) = \sum_{i=(j-1)\ell+1}^{j\ell} P_i(t,b_1^m) \quad j = 1,\ldots,n \tag{19}$$

for each block of data. The counters will give rise to *detailed* distributions

$$\hat{q}_s(t,j) = \frac{\delta(s,L(t)) \sum_{\nu\neq j} c_\nu(t,b_1^m)}{\sum_{t'} \delta(s,L(t')) \sum_{\nu\neq j} c_\nu(t',b_1^m)} \tag{20}$$

and to tied distributions

$$^*\hat{q}_s(t,j) = \frac{\delta(\mathcal{S}^*_i,L^*(\mathcal{T}_g)) \sum_{\nu\neq j} c^*_\nu(\mathcal{T}^*_g,b_1^m)}{\sum_{h=1}^{k} \delta(\mathcal{S}^*_i,L^*(\mathcal{T}_h)) \sum_{\nu\neq j} c^*_\nu(\mathcal{T}^*_h,b_1^m)} \qquad t\in\mathcal{T}^*_g,\ s\in\mathcal{S}^*_i \tag{21}$$

where as before

$$c^*_j(\mathcal{T}^*_g,b_1^m) = \sum_{t\in\mathcal{T}^*_g} c_j(t,b_1^m) \tag{22}$$

Note that $\hat{q}_s(t,j)$ and $^*\hat{q}_s(t,j)$ *do not* depend explicitly (*) on the the output data belonging to the j^{th} block. Thus the data in the j^{th} block can be interpreted as *new* in relation to these probabilities.

To determine the λ_s values, we then run the F-B algorithm on the interpolated Markov source (see Figure 4b) using all of the data b_1^m, but when carrying out the recursions (13) and (15) for α_i and β_i we use probabilities $\hat{q}_s(t,j)$ and ${}^*\hat{q}_s(t,j)$ (in (12) and (14)) if b_i belonged to the j^{th} block.

As presented here, the obtained λ_s values depend on the states $\hat{s}$ of the interpolated Markov source. However, we wish them to depend on the relative reliabilities of the detailed and tied estimates $\hat{q}_s$ and ${}^*\hat{q}_s$. One way to accomplish this it to tie the $\hat{s}$ states according to the number of times the corresponding state s has been reached by the source. In such a case we establish $\nu-1$ thresholds $0 \leq \tau_1 \leq \ldots \leq \tau_\nu - 1 \leq 1 = \tau_\nu$ and in addition to (20) and (21) compute state occupancy estimates

$$\hat{q}(s,j) = \frac{\sum_{t} \delta(s,L(t)) \sum_{\mu \neq j} c_\mu(t,b_1^m)}{\sum_{t'} \sum_{\mu \neq j} c_\mu(t',b_1^m)} \tag{22}$$

In the j^{th} block we then let λ_s have the i^{th} value $\lambda(i)$ if $\tau_{i-1} \leq \hat{q}(s,j) \leq \tau_i$. In this case the count contributions for transitions leaving $\hat{s}$ estimated from the j^{th} block are added to the contents of the *ith* counter pair.

The final interpolated Markov source used to predict new test data is then based on probabilities

$$\hat{q}_s(t) = \frac{\delta(s,L(t)) \sum_{j=1}^{n} c_j(t,b_1^m)}{\sum_{t'} \delta(s,L(t')) \sum_{j=1}^{n} c_j(t',b_1^m)} \tag{24}$$

and

$$ {}^*\hat{q}_s(t) = \frac{\delta(\mathscr{S}^*{}_i, L^*(\mathscr{T}^*{}_g)) \sum_{j=1}^{n} c^*{}_j(\mathscr{T}_g, b_1^m)}{\sum_{h=1}^{k} \delta(\mathscr{S}^*{}_i, L^*(\mathscr{T}^*{}_h)) \sum_{j=1}^{n} c^*{}_j(\mathscr{T}^*{}_h, b_1^m)} \qquad t \epsilon \mathscr{T}^*{}_g, \;\; s \epsilon \mathscr{S}^*{}_i \tag{25}$$

and λ_s values are chosen from the computed set $\lambda(1),\ldots,\lambda(k)$, depending on the range within which the state occupancy estimate

(*)Of course, $\hat{q}_s$ and ${}^*\hat{q}_s$ do depend implicitly on all of the data, since the probabilities $P_i(t,b_1^m)$ used in (19) were obtained by ordinary iteration on all of b_1^m. It is in principle possible not to take this shortcut, a nd run n separate F-B algorithms, one for each deleted block. However, for a sufficiently long ℓ this would produce only a small difference.

$$\hat{q}(s) = \frac{\sum_{t} \delta(s,L(t)) \sum_{j=1}^{n} c_j(t,b_1^m)}{\sum_{t} \sum_{j=1}^{n} c_j(t',b_1^m)} \tag{26}$$

falls.

Unfortunately, not much can be said about desirable block lengths ℓ or threshold sets. $\{\tau_i\}$. In each case, these should be determined experimentally. Clearly, a larger block length ℓ is more conservative since it makes for less accurate $\hat{q}_s(t,j)$ estimates and thus leads to decreasing values of λ_s.

6. Relaxed Deleted Interpolation

In the fixed deleted interpolation method the distributions $\hat{q}_s(t,j)$ and $*\hat{q}_s(t,j)$, are computed first and then used to determine the interpolation parameters λ_s. Finally, the model of the source is built up from $\lambda_s,\hat{q}_s$, and $*\hat{q}_s$. However, inspection of the structure of the model (Figure 4b) leads immediately to the suggestion that all of its statistical parameters should be computed from the data at the same time. The hoped for advantage of such a *relaxed* model is that the parameters might complement each other better.

Since all distributions $\lambda_s,\hat{q}_s(t)$, and $*\hat{q}_s(t)$ are to be determined simultaneously, there is no point in distinguishing between them. We simple include all the states $\hat{s},s$, and $s*$ (see Figure 4b) in one *interpolated state space* $\mathcal{S}^I$ and consider the latter partitioned into tied stes $\mathcal{S}^*_1,\ldots,\mathcal{S}^*_r$. Some $\mathcal{S}^*_i$ will consist of single states only (those that belonged to the original state space $\mathcal{S}$), other $\mathcal{S}^*_i$ sets will be determined by the original tying of the $\mathcal{S}$ space, and the remaining $\mathcal{S}^*_g$ sets consisting of $\hat{s}$ states will force common λ_s within-class values (this classification will be done in accordance with the expected reliability of the original $q_s(t)$ estimates). As before, let the state partition $\mathcal{S}^*_i, i = 1,\ldots,r$, induce a transition partition $\mathcal{T}^*_j, j = 1,\ldots,k$ and a predecessor function L^* over $\mathcal{T}^*_j$. Segment the data string b_1^m into n blocks of length $\ell(m = n)$. Let $\hat{P}_i(t,b_1^m)$ denote the probability $P_i(t,b_1^m)$ defined in Section 2 computed under the assumption that the data in the j^{th} block were generated by a Markov source governed by transition probabilities $\hat{q}_s(t,j)$ to be defined in the following iteration:

1. Choose common initial distributions $q_s^I(t)$ conforming to all tying conventions and set $\hat{q}_s^0(t,j) = q_s^I(t)$ for all $s\epsilon\mathcal{S}^I, t\epsilon\mathcal{T}$, and $j=1,2,\ldots,k$.

2. Set $\nu = 0$.

3. Compute $\hat{P}_i(t,b_1^m)$ for all t and i using transition probabilities $\hat{q}_s^\nu(t,j)$ in block j.

4. Compute tied counts $\hat{c}_j(\mathcal{T}^*_g,b_1^m)$ for all tied sets $\mathcal{S}_g$ and blocks j by the formula

$$\hat{c}_j(\mathscr{T}^*_g, b_1^m) = \sum_{t \in \mathscr{T}^*_g} \sum_{i=(j-1)\ell+1}^{j\ell} \hat{P}_i(t, b_1^m) \tag{27}$$

5. Determine new distributions $\hat{q}_s^{\nu}(t,j)$ *by*

$$\hat{q}_s^{\nu}(t,j) = \frac{\delta(\mathscr{S}^*_i, L^*(\mathscr{T}^*_g)) \sum_{\ell \neq j} \hat{c}_\ell(\mathscr{T}^*_g, b_1^m)}{\sum_{h=1}^{k} \delta(\mathscr{S}^*_i, L^*(\mathscr{T}^*_h)) \sum_{\ell \neq j} \hat{c}_\ell(\mathscr{T}^*_h, b_1^m)} \qquad t \in \mathscr{T}^*_g, s \in \mathscr{S}^*_i \tag{28}$$

6. Set $\nu = \nu + 1$.

7. Repeat from step 3 until convergence achieved.

8. For the final model use transition probabilities

$$\hat{q}_s(t) = \frac{\delta(\mathscr{S}^*_i, L^*(\mathscr{T}^*_g)) \sum_{\ell=1}^{n} \hat{c}_\ell(\mathscr{T}^*_g, b_1^m)}{\sum_{h=1}^{k} \delta(\mathscr{S}^*_i, L^*(\mathscr{T}^*_h)) \sum_{\ell=1}^{n} \hat{c}_\ell(\mathscr{T}^*_h, b_1^m)} \qquad t \in \mathscr{T}^*_g, s \in \mathscr{S}^*_i \tag{29}$$

Intuitively, this *relaxed deleted interpolator* algorithm is suitable because the estimation of the probability of any given transition taking place in the j^{th} block is based on the assumption that the parameters of the Markov source producing the j^{th} block are computed from data gathered everywhere but in the j^{th} block (see (27) and (28)). Thus the j^{th} block can intuitively be regarded as constituting new unseen data relative to the "old" retained data on which the transition probability estimates are based.

7. Some Generalizations

There are many aspects of the deleted interpolator method that have not been settled. For some of these even the possibility of analytical treatment seems doubtful. On the other hand, the point of view introduced is both powerful and fruitful in suggesting many generalizations.

First, it should be noted that the questions of desirable deletion block length ℓ, tying thresholds τ_i, tying of states in general, or initial selections of transitions distributions, are completely open and their solution depends on the concrete "physical" situation one is trying to model. Also, the tying that depends on the occupancy (26) is somewhat incomplete, since the values of λ_s should ideally depend on respective reliabilities of untied and tied distributions $\hat{q}_s$ and $^*\hat{q}_s$.

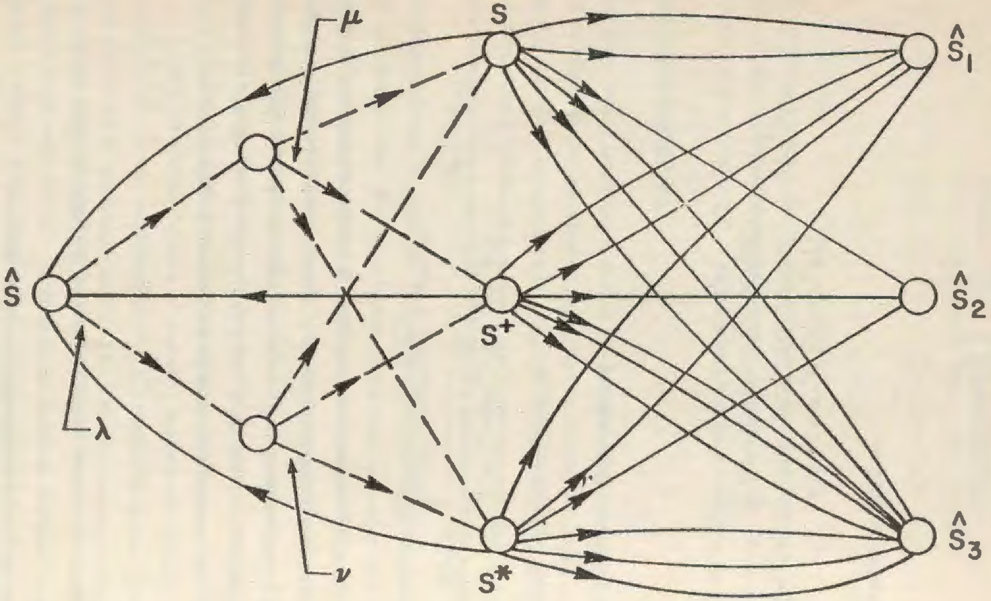

Figure 5

Segment of a Markov source corresponding to that to that of Figure 4 with a network of interpolation distributions.

Next, there is no reason why there should be only one tying partition $\mathscr{P}^*_1,\ldots,\mathscr{P}^*_r$. Several "competing" alternatives might be preferable, as might be hierarchical, nested tying where the sets $\mathscr{P}^*_i$ are themselves partitioned into subsets $\mathscr{P}^*_{i1},\ldots,\mathscr{P}^*_{ik_i}$. Such multiple k-fold tyings lead to λ vectors with components $\lambda_1,\ldots,\lambda_k$ $(\sum_{i=1}^{k} \lambda_i = 1)$. In these situations the question of appropriate λ ties is opened even further. In fact, because of the tying problem, networks of interpolation distributions might be useful, as is illustrated in Figure 5 where λ and μ distributions are tied with respect to individual state counts (as in the preceding sections), and ν is tied with respect to the + tying counts.

Since as discussed in Section 3, the F-B algorithm converges for a mixture of undetermined and fixed distributions, it is possible to mix the fixed and relaxed deleted interpolator approach by, e.g., using the interpolation distributions $\lambda_1,\lambda_2,\lambda_3,\lambda_4$ $(\sum_{i=1}^{4} \lambda_i = 1)$ where the first two transitions lead to relaxed distributions (to be determined together with these λ values) and the last two to fixed distributions. Indeed, the fixed distributions may be of any kind whatever, and need not have been determined from the given data b_1^m at all. In fact, the deleted interpolator method may be viewed as a general way of combining multiple disparate estimators and may be used for judging their relative worth: if the values of λ_i are consistently small relative to those of the other components λ_j then the i^{th} estimator may well be eliminated.

The idea of deleted estimation is not new in statistics, only its application in a maximum likelihood setting seems to be. Deleted estimation with usually (but not necessarily) quadratic loss functions is related to the *jack knife* principle discussed by Mosteller and Tukey [4] and to *cross validation* of predictions investigated by Stone [5]. The last reference contains an excellent formulation of the problem (that can in principle be generalized to cover our case) with many examples and an extensive biography. Some valuable analytical results related to deleted estimation have been obtained by Wagner and his colleagues [5, 6].

References

[1] L. E. Baum: *An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of Markov Processes*, Inequalities, Vol. 3, 1972, pp. 1-8.

[2] L. E. Baum and T. Petrie: *Statistical Inference for Probabilistic Functions of Finite State Markov Chains*, Annals of Math. Statistics, Vol. 37, No. 6, Dec. 1966, pp. 1554-1563.

[3] F. Jelinek, R. Mercer, L. Bahl: *Continuous Speech Recognition: Statistical Methods*, to appear in *Handbook of Statistics*, vol. 2, P. R. Krishnaiah and L. N. Canal Editors, North-Holland Publishers, Amsterdam.

[4] F. Mosteller and J. W. Tukey: *Data Analysis Including Statistics*, Handbook of Social Psychology, G. Lindsey and E. Aronson Eds., Vol. 2, Addison-Wesley, Reading, Mass., 1968.

[5] M. Stone: *Cross Validitory Choice and Assessment of Statistical Predictions*, Journal of Royal Statistical Society, Sec. B, 1974, pp. 111-147.

[6] T. J. Wagner: *Deleted Estimates of the Bayes Risk*, Annals of Statistics, Vol. 1, 1973, pp. 359-362.