LINEAR PREDICTIVE HIDDEN MARKOV MODELS AND THE SPEECH SIGNAL

Alan B. Poritz
Institute for Defense Analyses
Communications Research Division
Princeton, New Jersey

ABSTRACT

A method for modelling time series is presented and then applied to the analysis of the speech signal. A time series is represented as a sample sequence generated by a finite state hidden Markov model with output densities parameterized by linear prediction polynomials and error variances. These objects are defined and their properties developed. The theory culminates in a theorem that provides a computationally efficient iterative scheme to improve the model. The theorem has been used to create models from speech signals of considerable length. One such model is examined with emphasis on the relationship between states of the model and traditional classes of speech events. A use of the method is illustrated by an application to the talker verification problem.

INTRODUCTION

We will investigate a statistical model for time series. A time series of length $T \times M$ is decomposed into a sequence of shorter segments of length M. Each segment is represented by a local model as the output of one of S previously selected all pole recursive filters driven by previously selected noise sources. The filters are defined by polynomials of some degree N with $N < M$. The order in which the filters appear in the sequence is controlled by a previously selected S state Markov chain, a component of the global model. For example, in the case of the speech signal, it is as if the vocal tract were capable of only S possible configurations that alternated in Markov fashion and produced whispered speech whose formant structure was piecewise constant (in time) and completely described by an all pole filter. While this model is only a cartoon view of speech production, we will show that it is useful for describing some of the global information in the speech signal. The underlying mathematical ideas, together with their efficient implementation, are an extension, using methods from linear algebra, of the work of Baum, Petrie, Soules and Weiss (1). The model will be outlined starting with the local theory.

THE LOCAL MODEL: PREDICTIVE DENSITIES

Let $Y = (y_0, \ldots, y_{M-1}) \in \Re^M$ be any time series; set

$$U(Y) = \begin{pmatrix} y_N & y_{N+1} & \cdots & y_{M-1} \\ y_{N-1} & y_N & \cdots & y_{M-2} \\ \vdots & \vdots & & \vdots \\ y_0 & y_1 & \cdots & y_{M-N-1} \end{pmatrix}$$

and let

$$R(Y) = U(Y)U(Y)^* \qquad (1)$$

where *denotes transpose. $R(Y)$ is a symmetric positive semi-definite $N+1$ by $N+1$ matrix. Let $A = (a_0, \ldots, a_N)$ be a vector in real $N+1$ space $\Re^{N+1}$; unless otherwise stated we assume that the leading coordinate value is one: $a_0 = 1$. $AR(Y)A^*$ is the sum of the squares of the errors when Y is predicted using A; if we define $\tilde{y}_\tau = -\sum_{k=1}^{N} a_k y_{\tau-k}$, then $AR(Y)A^* = \sum_{\tau=N}^{M}(y_\tau - \tilde{y}_\tau)^2$. Let σ be a positive number. We use the pair (A, σ) to define a real function on $\Re^M$ for $M > N$:

$$L(Y \mid A, \sigma) = \frac{2}{\sqrt{2\pi\sigma^2}} \exp(-AR(Y)A^*/2\sigma^2). \qquad (2)$$

Intuitively, this function measures how well a time series $Y \in \Re^M$ satisfies the recursion defined by A, computed in units of σ. We will call it a *predictive* or *recursive density of degree N determined by the model* (A, σ). It is a likelihood, i.e. probability density, on $\Re^M$ with respect to an appropriately defined σ-algebra and measure on $\Re^M$. The measure is obtained by pulling back Gauss$(0, \sigma^2)$ from $\Re^+$ to $\Re^M$ via the projection induced by A.

Given a time series $Y \in \Re^M$ and an integer $N < M$, we want to find a pair (A, σ) with $A \in \Re^{N+1}$

and $\sigma > 0$ that maximizes $L(Y \mid A, \sigma)$. The proof of the existence and uniqueness of such a maximum likelihood density follows from a lemma needed again later.

For $V = (v_0, v_1, \ldots, v_N) \in \Re^{N+1}$ set $V_N = (v_1, \ldots, v_N)$ so that $V = (v_0, V_N)$. Corresponding to this decomposition, $\Re^{N+1} = \Re \oplus \Re^N$, we can decompose any $N+1$ by $N+1$ symmetric matrix R:

$$R = \begin{pmatrix} B & C \\ C^* & D \end{pmatrix} \qquad (3)$$

where D is an N by N matrix.

Lemma. *Given a positive definite symmetric $N+1$ by $N+1$ matrix, R, define the smooth real valued function F with domain $\Re^N \times (0,\infty)$ by the formula: $F(A_N, \sigma) = \log(\sigma) + \frac{ARA^*}{2\sigma^2}$ where $A = (1, A_N)$. Then F has a unique minimum and this occurs at the point $(\hat{A}_N, \hat{\sigma})$ where $\hat{A}_N = -CD^{-1}$, $\hat{A} = (1, \hat{A}_N)$ and $\hat{\sigma}^2 = \hat{A}R\hat{A}^*$*

Proposition. *Let $Y \in \Re^M$ be a time series and suppose that $R(Y)$ is nonsingular. Then $L(Y \mid A, \sigma)$ has a unique maximum and this occurs at $(A, \sigma) = (\hat{A}, \hat{\sigma})$ where $\hat{A}_N = -C(Y)D(Y)^{-1}$, $\hat{A} = (1, \hat{A}_N)$ and $\hat{\sigma}^2 = \hat{A}R(Y)\hat{A}^*$.*

THE GLOBAL MODEL

For any positive integer S we use S also as the name of the set $\{1, \ldots, S\}$. We refer to this set as a ***state space*** and to its elements as ***states***. A ***first order S state Markov chain*** is a state space with S elements together with a stochastic vector $\{a_1, a_2, \ldots, a_S\}$, the ***initial state vector***, and an S by S row stochastic matrix

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1S} \\ a_{21} & a_{22} & \cdots & a_{2S} \\ \vdots & \vdots & & \vdots \\ a_{S1} & a_{S2} & \cdots & a_{SS} \end{pmatrix}$$

called the ***state transition matrix***. We recall that a stochastic vector has non-negative entries that sum to one. If G is any set an S ***state hidden Markov model for observations in G*** is an S state Markov chain together with a set of ***output densities***

$$\{L_s : G \mapsto \Re^+ \cup \{0\} \mid s \in S\}$$

each of which integrates to unity with respect to some measure on G.

When $G = \Re^M$ and each L_s is a predictive density of degree N, e.g. $L_s(Y) = L(Y \mid A(s), \sigma(s))$ as defined in the preceding section, we have a ***linear predictive*** hidden Markov model. The models studied below all have initial state probabilities set equal to $1/S$. Let λ denote the remaining parameters of one of these models: $\lambda = \{a_{sr}, A(s), \sigma(s) \mid s, r \in S\}$. Denote by $\Lambda(S, N, M)$, the set of all models λ for a fixed choice of S,N and M. A model will be said to be ***interior*** if its transition probabilities are all positive.

Let $\mathcal{Y}$ be a time series of $T \times M$ samples: $\mathcal{Y} = (y_0, \ldots, y_{TM-1})$. Given a model $\lambda \in \Lambda(S, N, M)$, we want to define the likelihood of $\mathcal{Y}$ given λ. Decompose $\mathcal{Y}$ into a sequence of T shorter segments, $Y(t)$, each of length M: $Y(t) = (y_{tM}, \ldots, y_{(t+1)M-1})$ for $t = 0, 1, \ldots, T-1$, so that $\mathcal{Y} = (Y(0), \ldots, Y(T-1))$. Let ω be any T long path (i.e. sequence of states) in the state space: $\omega = (s_0, \ldots, s_{T-1})$ and let Ω be the set of all such paths. The ***probability of ω given λ*** is given by the Markov chain: $P(\omega \mid \lambda) = \prod_{t=0}^{T-1} a_{s_{t-1}s_t}$ where by convention $a_{s_{-1}s_0}$ means a_{s_0}. Define the ***likelihood of $\mathcal{Y}$ given ω and λ*** as

$$L(\mathcal{Y} \mid \omega, \lambda) = \prod_{t=0}^{T-1} L(Y(t) \mid A(s_t), \sigma(s_t)).$$

We have $L(\mathcal{Y}, \omega \mid \lambda) = L(\mathcal{Y} \mid \omega, \lambda)P(\omega \mid \lambda)$ so that the ***likelihood of the time series $\mathcal{Y}$ given the model λ*** can now be defined as

$$L(\mathcal{Y} \mid \lambda) = \sum_{\omega \in \Omega} L(\mathcal{Y}, \omega \mid \lambda)$$

$$= \sum_{\omega \in \Omega} \prod_{t=0}^{T-1} a_{s_{t-1}s_t} \cdot L(Y(t) \mid A(s_t), \sigma(s_t)).$$

Although the computational burden in evaluating this likelihood appears to be exponential in T, the paper of Baum *et al.*(1) shows that with a recursive calculation, it is only linear in T. Define $\alpha_0(s) = a_s L_s(Y(0))$, $s = 1, \ldots, S$ and

$$\alpha_t(s) = \sum_{r=1}^{S} \alpha_{t-1}(r) a_{rs} L_s(Y(t))$$

for $s = 1, \ldots, S$ and $t = 1, \ldots, T-1$. Also, define $\beta_{T-1}(s) = 1$, $s = 1, \ldots, S$ and

$$\beta_t(s) = \sum_{r=1}^{S} \beta_{t+1}(r) a_{sr} L_r(Y(t+1))$$

for $s = 1, \ldots, S$ and $t = T-2, \ldots, 0$. These formulae satisfy $\alpha_t(s) \cdot \beta_t(s) = L(\mathcal{Y}, s_t = s \mid \lambda)$, the latter being the likelihood of $\mathcal{Y}$ and state s being used at time t, given the model. Thus for any t, we have $L(\mathcal{Y} \mid \lambda) = \sum_{s=1}^{S} \alpha_t(s)\beta_t(s)$ so that in particular, $\sum_{s=1}^{S} \alpha_{T-1}(s)$ is an expression for $L(\mathcal{Y} \mid \lambda)$ whose computation is linear in T.

GLOBAL MAXIMUM LIKELIHOOD

We want to find a model λ that maximizes the likelihood of a given time series $\mathcal{Y}$ among all models in $\Lambda(S, N, M)$. But, unlike the problem for the local model discussed above, no closed form solution is known for hidden Markov models. However, as shown in Baum *et al.*(1), for a certain class of hidden Markov models this problem can be attacked by an iterative hill climbing technique. By using the Lemma stated previously, that method can be extended to the class of linear predictive hidden Markov models. The theorem expressing this extension requires the following items. Let

$$\gamma_t(s) = \alpha_t(s) \cdot \beta_t(s)/L(\mathcal{Y} \mid \lambda)$$

with $s = 1, \ldots, S$ and $t = 0, \ldots, T-1$. $\gamma_t(s)$ is the *posterior probability of state s at time t based on $\mathcal{Y}$ and* λ. Also let

$$\gamma_t(s, r) = \alpha_t(s) a_{sr} L_r(Y(t+1))\beta_{t+1}(r)/L(\mathcal{Y} \mid \lambda)$$

with $r, s = 1, \ldots, S, t = 0, \ldots, T-2$. Let $R(Y(t)) = U(Y(t))U(Y(t))^*$ for $t = 0, \ldots, T-1$ as in Eq.(1). For each $s = 1, \ldots, S$, define

$$R(s) = \sum_{t=0}^{T-1} \gamma_t(s) R(Y(t))$$

Then, as in Eq.(3), we have matrix decompositions:

$$R(s) = \begin{pmatrix} B(s) & C(s) \\ C(s)^* & D(s) \end{pmatrix}$$

where each $D(s)$ is an N by N matrix. Finally define

$$R(\mathcal{Y}) = \sum_{t=0}^{T-1} R(Y(t))$$

Theorem. *Suppose $\mathcal{Y}$ is a time series with $T \times M$ samples and $R(\mathcal{Y})$ is nonsingular. For any interior model $\lambda = \{a_{sr}, A(s), \sigma(s) \mid s, r \in S\}$ in $\Lambda(S, N, M)$, define a new model $\hat{\lambda} = \{\hat{a}_{sr}, \hat{A}(s), \hat{\sigma}(s) \mid s, r \in S\}$ by*

$$\hat{a}_{sr} = \sum_{t=0}^{T-2} \gamma_t(s, r) / \sum_{t=0}^{T-2} \gamma_t(s),$$

$$\hat{A}_N(s) = -C(s)D(s)^{-1}, \hat{A}(s) = (1, \hat{A}_N(s))$$

$$\text{and } \hat{\sigma}(s)^2 = \hat{A}(s)R(s)\hat{A}(s)^* \;/ \sum_{t=0}^{T-1} \gamma_t(s).$$

Then $\hat{\lambda}$ is an interior model and if $\hat{\lambda}$ differs from λ then $L(\mathcal{Y} \mid \hat{\lambda}) > L(\mathcal{Y} \mid \lambda)$.

The invertibility of the matrices $D(s)$ is a consequence of the nonsingularity of $R(\mathcal{Y})$ and the fact that λ is an interior model. Other simple hypotheses also guarantee these generic conditions. The theorem gives an iterative algorithm; each output model is the next input model. Iteration continues until some convergence criterion is satisfied. The search for a maximum likelihood model consists of applying this method to starts distributed across $\Lambda(S, N, M)$. If in Eq.(2) the exponent in the denominator is $(M-N)/2$ instead of $1/2$, an entirely similar theory results and corresponds to the χ^2 assumption on the distribution of the predictive error.

APPLICATION OF THE MODEL TO SPEECH

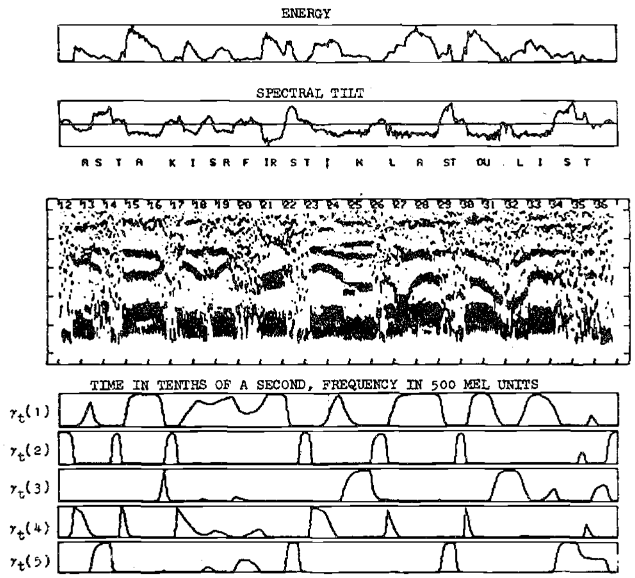
A time series $\mathcal{Y}$ was made from 40 seconds of 12 bit PCM speech that had been filtered at 5 kHz. and sampled at 10000 samples/second. With $M = 100$ so that each $Y(t)$ contains 10 milliseconds, this means a total of $T = 4000$ *frames*. There are 88 words and roughly 400 phones in the signal, 40% of which is silence.

The best model $\lambda_F \in \Lambda(5, 3, 100)$ found by the theorem in hill climbs from many random starts has the transition matrix:

$$\begin{pmatrix} .91 & .00 & .06 & .00 & .03 \\ .00 & .95 & .00 & .05 & .00 \\ .03 & .07 & .87 & .01 & .02 \\ :24 & .00 & .00 & .76 & .00 \\ .00 & .07 & .00 & .05 & .88 \end{pmatrix}$$

Its stationary probabilities are (.30, .36, .15, .10, .09). The vectors $A(s) = (a_0(s) = 1, a_1(s), a_2(s), a_3(s))$ and the error variances $\sigma(s)^2$ for the predictive densities L_s are:

s	$a_0(s)$	$a_1(s)$	$a_2(s)$	$a_3(s)$	$\sigma(s)^2$
1	1.00	−1.39	.93	−.25	.0860
2	1.00	.16	.30	−.08	.0004
3	1.00	−1.31	.85	−.29	.0094
4	1.00	−.25	.42	−.30	.1136
5	1.00	1.19	.79	.06	.0235



Figure

Figure 1 contains information about λ_F and the 2.5 seconds from $\mathcal{Y}$ during which the utterance was "a stack is a first in last out list". Each curve at the bottom of the figure is the time function that gives the posterior probability, $\gamma_t(s)$, of being in state s at each of the times t during the utterance.

LOG MAGNITUDE OF THE IMPULSE RESPONSE OF THE SYNTHESIS FILTER FOR EACH STATE OF THE MODEL λ_F IN LPHMM(5,3,100) FREQUENCY IN KHZ,ENERGY AT 10DB LEVELS

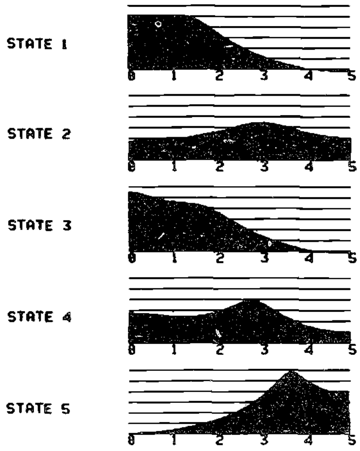


FIGURE 2

Figure 2 shows the log magnitudes of the power spectra for the all-pole filters corresponding to the state polynomials $A(s), s = 1, \ldots, 5$. We can make some rough characterizations of the states from the Figures and the parameters of the model.

State	*Characteristic*
1	*strong voicing*
2	*silence*
3	*nasal, liquid*
4	*stop burst, post silence*
5	*frication*

Models with other values of S also have suggestive interpretations. When S is small the models are not very sensitive to N. A *run of the state s* is a maximal interval I where $t \in I$ implies $\gamma_t(s) = \max_{r \in S} \gamma_t(r)$. The distribution of the durations of the runs of the states of λ_F on $\mathcal{Y}$ is given in the table below; notice that the number of runs is about the same as the number of phones.

time in msecs	$s = 1$	2	3	4	5	*sum*
0 — 50	22	21	32	64	5	144
50 — 100	27	25	19	16	18	105
100 — 150	17	3	10	4	13	47
150 — 200	18	2	3	1	4	28
200 — 300	9	1	5	1	0	16
300 — 500	6	4	2	0	0	12
> 500	1	12	1	0	0	14
sum	100	68	72	86	40	366
%	27	19	20	23	11	100

A talker verification experiment was conducted using linear predictive hidden Markov models. Speech was collected from 10 talkers; all reading the same passage. For each talker, j, a 40 second section was used to create a time series $\mathcal{Y}_j$ based on the same paragraph as the time series $\mathcal{Y}$ discussed above. Separate 15 second sections Z_j from each talker j were chosen for speech density. These shorter sections did not overlap the longer sections but several shared some content. A model $\lambda_j \in \Lambda(5,3,100)$ was made for talker j by iteratively applying the theorem to $\mathcal{Y}_j$ and the same starting model as was used to construct the model λ_F discussed above. The likelihoods $L(Z_j \mid \lambda_i)$ for $i, j = 1, \ldots, 10$ were computed. The results showed that in each of the 10 experiments the correct talker's model outscored the other 9.

The experiment is of course not definitive and larger ones are contemplated. Nonetheless it does suggest that important information about the speech signal is encapsulated in these small hidden Markov models. They may be of use both in research on the speech signal and in a variety of practical applications.

Reference:

(1)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.*, Vol. 41. No. 1,pp. 164-171,1970.