A Minimum Discrimination Information Approach for Hidden Markov Modeling

YARIV EPHRAIM, MEMBER, IEEE, AMIR DEMBO,
AND LAWRENCE R. RABINER, FELLOW, IEEE

Abstract—**An iterative approach for minimum discrimination information (MDI) hidden Markov modeling of information sources is proposed. The approach is developed for sources characterized by a given set of partial covariance matrices and for hidden Markov models (HMM's) with Gaussian autoregressive output probability distributions (PD's). The proposed approach aims at estimating the HMM which yields MDI with respect to all sources that could have produced the given set of partial covariance matrices. Each iteration of the MDI algorithm generates a new HMM as follows. First, a PD for the source is estimated by minimizing the discrimination information measure with respect to the old model over all PD's which satisfy the given set of partial covariance matrices. Then, a new model that decreases the discrimination information measure between the estimated PD of the source and the PD of the old model is developed. The problem of estimating the PD of the source is formulated as a standard constrained minimization problem in the Euclidean space. The estimation of a new model given the PD of the source is done by a procedure that generalizes the Baum algorithm. The MDI approach is shown to be a descent algorithm for the discrimination information measure and its local convergence is proved.**

I. Introduction

IN MANY statistical signal processing problems, the probability distributions (PD's) of the sources being processed are not known, yet the application of optimal procedures requires a priori knowledge of such PD's. For example, optimal detection theory (see, e.g., [1]) could be applied to speech recognition if the probability of any word in the recognizer's vocabulary and the PD of the corresponding acoustic signal were known. Similarly, optimal minimum average distortion estimation approaches could be successfully applied for enhancing noisy speech if the PD's of the speech signal and the noise process were known [2], [3]. A common practice is, therefore, that of replacing each unknown PD by its estimate obtained from a long training sequence from the source. The estimation of the PD of the source is usually done by attributing to the source a parametric PD and estimating the parameters of this PD from the given training sequence. Thus a parametric estimation problem results. This problem, however, is not a standard parametric estimation problem, since the PD of the source producing the training sequence is not necessarily that of the assumed parametric form. A better formulation of this estimation problem is given in terms of modeling of the original source, or of its PD, by a parametric PD which constitutes the model.

A useful class of models is that of Markov sources, also called probabilistic functions of Markov chains or hidden Markov models (HMM's) [4], [5, pp. 63–70], [6], which have been proven successful in speech recognition (see, e.g., [7], [8]) and speech enhancement [3] applications. The parameters of these models are usually estimated by a maximum likelihood (ML) approach developed by Baum *et al.* [9], [10], and extended in [11]–[13]. Recently, an alternative method for estimating the parameters of HMM's by a maximum mutual information (MMI) approach [14, p. 262] was proposed by Bahl *et al.* [15], [16]. The ML estimate is obtained by maximizing the logarithm of the probability density function (pdf) of the HMM over its parameter set for the given training sequence from the source. This is done by an iterative estimation-maximization (EM) procedure which converges locally [9], [17], [18]. The MMI estimate is obtained by maximizing the average mutual information between two dependent sources, the first of which has an HMM PD conditioned on the second source, and the second an assumed known PD. The average of the mutual information is calculated using the empirical conditional distribution of the first source as obtained from given training sequences from that source. In speech recognition applications which motivated the MMI hidden Markov modeling approach, the first source is the acoustic signal from a given word and the second source represents the words in the vocabulary. An MMI estimate is obtained using any standard optimization procedure, e.g., the steepest-descent method [15].

The ML and the MMI modeling approaches can be justified as being standard ML and MMI estimation approaches only if the source producing the training sequences is itself a Markov source. In this case, the ML estimation may have under certain conditions, asymptotically (large-sample) optimal properties (see, e.g., [49], [50]). Otherwise, the theoretical justification for these approaches can only be given on the basis of a model

Manuscript received February 9, 1987; revised December 1, 1988. The material in this paper was partially presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Dallas, TX, April 6–9, 1987.

Y. Ephraim and L. R. Rabiner are with the Speech Research Department, AT&T Bell Laboratories, Murray Hill, NJ 07974.

A. Dembo was with the Communications Analysis Research Department, AT&T Bell Laboratories. He is now with the Information Systems Laboratory, Stanford University, Stanford, CA 94305.

IEEE Log Number 8930403.

0018-9448/89/0900-1001$01.00

correctness assumption according to which the PD of the source is assumed to be that of the model. A less conservative interpretation for the ML approach, which can also be adopted for the MMI approach [19], was proposed by Csiszár and Tusnady [20]. They showed that the ML estimate results from minimizing the discrimination information measure[1] between a PD concentrated in the training sequence from the source and the PD of the model. In this interpretation, the source being modeled and the model itself are treated as independent entities and hence no model correctness assumption is needed. This interpretation, however, implies that the ML and MMI modeling approaches might be sensitive to the specific training sequences from the source which are used for performing the modeling.

We propose an alternative approach for hidden Markov modeling which is based on the minimum discrimination information (MDI) modeling approach proposed by Kullback [21, chs. 3, 5] and Kupperman [22]. This approach assumes knowledge of a partial set of moments for each vector in the training sequence from the source and aims at estimating the parameter set of the model which yields MDI with respect to all PD's which satisfy the given moments. The MDI modeling approach was proven by Shore and Johnson [23] to be the only correct inference approach, when the source is characterized by moment constraints, in the sense of satisfying a set of consistency axioms. Any other inference approach will either provide the same estimate as the MDI approach or will lead to inconsistency. It can also be argued that the MDI approach should be more robust than the ML and the MMI modeling approaches, since here we do not attribute to the source any specific PD but rather consider all PD's which satisfy the given set of moments.

The MDI approach was applied in [21, p. 83], [22], [24] to classification problems of sources characterized by a given set of moments and models (or hypotheses) that comprise an exponential family of PD's. Using appropriate sample average estimates for the moments being specified, it was shown that the MDI approach results in the same classification rule as the traditional ML approach. A similar relation between the MDI and the ML modeling approaches was found in [25], [26] for the particular case of sources characterized by given partial covariance matrices and Gaussian models, e.g., autoregressive (AR) and autoregressive moving average (ARMA) sources. It was shown that asymptotic MDI modeling (as the frame length approaches infinity) is equivalent to asymptotic ML modeling [27, ch. 1], [28], achieved by minimizing the Itakura–Saito distortion measure [29, p. 134], [30] between the power spectral densities of the source and the model. For HMM's with Guassian output PD's and sources characterized by a given set of partial covariance matrices, it is shown here that the MDI modeling approach approximately becomes an ML modeling approach when the MDI measure is assumed to be concentrated in a single sequence of states of the model.

The general approach for MDI modeling is first to minimize the discrimination information measure, between the PD of the source and the PD of the model, over all source PD's which satisfy the given set of moments. This results in an estimate of the PD of the source, called the MDI PD with respect to the model or the I-projection of the model on the set of PD's that satisfy the given moments. The MDI PD depends on the parameter set of the model and the given moments of the source. The discrimination information measure between the MDI PD and the PD of the model is called the MDI measure with respect to the model. The modeling is achieved by minimizing the MDI measure over all parameter sets of the model. Unfortunately, hidden Markov modeling cannot be done in this way since no explicit expression for the MDI PD, and hence also for the MDI measure, in terms of the given moments and the parameter set of the model, is known. The MDI PD depends on a set of Lagrange multipliers which must be chosen so that the given moments from the source are satisfied. MDI hidden Markov modeling can, however, be iteratively performed by alternating minimization of the discrimination information measure once over all PD's which satisfy the given set of moments assuming that an HMM is given, and then over all HMM's assuming that the MDI PD with respect to the old model is given. If each iteration comprises the estimation of the MDI PD for a given model and the estimation of a new model for the MDI PD with respect to the old model, then the algorithm effectively generates a sequence of HMM's with nonincreasing values of the MDI measure. Note that the discrimination information measure need not be strictly minimized in each step of the iterative algorithm, since any procedure which alternatively reduces the value of this measure can be used without affecting the descent nature of the algorithm.

The alternating minimization of the discrimination information measure was first proposed by Csiszár and Tusnady [20]. They considered the general problem of minimizing the discrimination information measure over two sets of PD's and gave geometric conditions for global convergence of the iterative procedure. Furthermore, they showed that these conditions are satisfied if the two sets of PD's are convex sets of measures. In our case the set of PD's satisfying the given moments is convex, but the set of HMM PD's is not. Since the geometric conditions are difficult to verify when either of the two sets of PD's is not convex, we shall prove here only local convergence using a variant of the convergence theorem from [31, p. 187] given in [32, lemma 1].

We develop the iterative algorithm and prove its convergence for MDI modeling of sources characterized by a given set of partial covariance matrices and HMM's with zero mean Guassian output PD's. Such models will be referred to as zero mean Gaussian HMM's. In addition, we shall be focusing on the subset of AR processes of this class, which have been shown to be useful in speech

[1]The discrimination information measure is also known as the cross entropy, relative entropy, directed divergence, I-divergence, and Kullback–Leibler number.

recognition (see, e.g., [8], [12]) and speech enhancement [3] applications. These models will be referred to as zero mean Guassian AR HMM's. We show that the estimation of the MDI PD with respect to a given HMM can be formulated as a unimodal minimization problem in a subset of the Euclidean space, which can be solved by any standard constrained optimization procedure. Furthermore, a new model that decreases (or keeps constant) the discrimination information measure between the MDI PD with respect to the old model and the PD of the old model can be efficiently estimated by a procedure which generalizes the Baum reestimation algorithm using "forward–backward" formulas [9], [10].

The proposed algorithm has the intuitive interpretation that in each iteration it first removes the existing "mismatch" between the source (as characterized by the given partial covariance matrices) and the current estimate of the model, and then it improves the modeling. Thus this modeling approach continuously improves the estimation of the PD of the source and the PD of the model. Note that the MDI modeling approach used here is different from the statistical inference philosophy of Shore and Johnson (see, e.g., [33], [34]). In their work, the model is treated as given prior information about the source, and as such, it is not changed during the inference process. Hence, Shore and Johnson's approach constitutes only one iteration in our algorithm. For the priors used in Shore and Johnsons' work, however, further iterations are not useful since it can be shown that the algorithm reaches a fixed point after the first iteration.

In Section II we derive the iteration algorithm for performing the MDI modeling. In Section III we prove local convergence of the algorithm. In Section IV we establish a relation between the MDI and ML modeling approaches. In Section V we consider principal implementation aspects of the MDI modeling approach. All proofs of theorems, corollaries, and lemmas are given in the Appendix.

II. Descent Algorithm for MDI Hidden Markov Modeling

A. Problem Formulation

Let $y \triangleq \{y_0, y_1, \cdots, y_T\}$ be a set of zero mean observations, $y_t \in R^N$, where R^N is the N-dimensional Euclidean space. Let $R_t^\dagger \triangleq E_{Q^\dagger}\{y_t y_t^\#\}$, where # denotes vector transpose and $Q^\dagger$ is the true PD of the source, be the full covariance matrix of y_t. Assume that for each t we are given an $N \times N$ matrix R_t whose elements within some given symmetric band, say B, are the elements of $R_t^\dagger$. Such a matrix will be referred to as a partial covariance matrix of y_t. The band B can, for example, be an upper left block or the main diagonal along with some off diagonals of the full covariance matrix. In practice, we may need to estimate the partial covariance matrix R_t from y_t. In this case the estimate will be treated as if it were the true partial covariance matrix. If the source is stationary and ergodic, and the frame length is large enough, then a good estimate of the true partial covariance matrix results. Alternatively, for any value of N, we can consider any given estimate of the partial covariance matrix of the source as being the characterization of the source, regardless of how well it estimates the true partial covariance matrix. In either case, we assume that the partial covariance matrix R_t, given or estimated, is consistent with some valid $N \times N$ covariance matrix called an extension of R_t. If this extension is positive definite, then it is called a positive definite extension.

Let P_λ be the PD of an M state zero mean Gaussian HMM, where λ is the parameter set of the model. $\lambda \triangleq (\pi, A, S)$, where $\pi \triangleq (\pi_1, \pi_2, \cdots, \pi_M)$ is the initial state probability vector, $A \triangleq \{a_{\alpha,\beta},\ \alpha, \beta = 1, \cdots, M\}$ is the state transition probability matrix, and $S \triangleq \{S_\beta,\ \beta = 1, \cdots, M\}$ is the set of positive definite covariance matrices of the output processes from the different states. The pdf corresponding to P_λ is given by

$$p_\lambda(y) = \sum_x \prod_{t=0}^{T} a_{x_{t-1}x_t} b(y_t|x_t) \tag{1}$$

$$b(y_t|x_t = \beta) = \frac{\exp\left(-\frac{1}{2} y_t^\# S_\beta^{-1} y_t\right)}{(2\pi)^{N/2} \det^{1/2}(S_\beta)}, \qquad \beta = 1, 2, \cdots, M$$

where $x \triangleq \{x_0, x_1, \cdots, x_T\}$ is a sequence of states and $x_t \in \{1, 2, \cdots, M\}$, $a_{x_{t-1}x_t}$ is the transition probability from the state x_{t-1} (at time $t-1$) to the state x_t (at time t), $a_{x_{-1}x_0} \triangleq \pi_{x_0}$ is the probability of the initial state x_0, and $b(y_t|x_t)$ is the output pdf on R^N corresponding to the state x_t.

Let $R \triangleq \{R_0, R_1, \cdots, R_T\}$ be the set of given partial covariance matrices. Let $\Omega(R)$ be the set of all PD's Q which satisfy

$$R_t = E_Q\{y_t y_t^\#\} \quad \text{within the band } B \tag{2}$$

for all $0 \le t \le T$. The general MDI modeling problem is that of finding the parameter set λ which minimizes the MDI measure defined as

$$\nu(R, P_\lambda) \triangleq \inf_{Q \in \Omega(R)} D(Q\|P_\lambda), \tag{3}$$

where $D(Q\|P_\lambda)$ is the discrimination information measure between Q and P_λ. The discrimination measure between two PD's Q and P can be evaluated as

$$D(Q\|P) \triangleq \begin{cases} \int q_P \ln q_P \, dP, & \text{if } Q \ll P \\ +\infty, & \text{otherwise} \end{cases} \tag{4a}$$

where q_P is the Radon–Nikodym derivative of Q with respect to P, and $Q \ll P$ means that Q is absolutely continuous with respect to P. If P and Q are absolutely continuous with respect to the Lebesgue measure, then their pdf's p and q, respectively, exist and

$$D(Q\|P) = \int q(y) \ln(q(y)/p(y))\, dy \tag{4b}$$

with the convention that $\ln 0 = -\infty$, $\ln(c/0) = \infty$, where c is any positive number, and $0 \ln 0 = 0$.

As explained in Section I, the implementation of the MDI approach for hidden Markov modeling must be performed iteratively since no explicit expression for the MDI measure $\nu(R, P_\lambda)$ is known. Each iteration of the proposed MDI algorithm comprises the following two steps. Starting from a given HMM P_λ, the MDI PD Q_λ with respect to P_λ is first estimated by

$$\inf_{Q \in \Omega(R)} D(Q\|P_\lambda). \tag{5a}$$

It is shown in the second part of this section that, under certain conditions, the MDI PD Q_λ exists and hence the infimum in (5a) is a minimum. In this case, the resulting discrimination information measure is the MDI measure with respect to P_λ given by $\nu(R, P_\lambda) = D(Q_\lambda\|P_\lambda)$. Given the MDI PD Q_λ, a new model $P_{\lambda'}$ which decreases the discrimination information $D(Q_\lambda\|P_\lambda)$ or at least keeps its value constant, is estimated. Thus

$$D(Q_\lambda\|P_{\lambda'}) \le D(Q_\lambda\|P_\lambda). \tag{5b}$$

Such a model is often easier to calculate than a model that minimizes $D(Q_\lambda\|P_{\lambda'})$ over all possible $P_{\lambda'}$, as is demonstrated in the third part of this section for HMM's. The fact that we only calculate a model that decreases the discrimination information, rather than minimizing it, does not affect the descent nature of the algorithm but it may slow its convergence. The proposed MDI algorithm iterates the two steps described in (5a) and (5b) until some convergence criterion is satisfied. For example, the algorithm can be stopped if the difference in values of the MDI measure (5a) in two consecutive iterations is smaller than or equal to a given threshold, say ϵ_{stop}, i.e., if

$$\nu(R, P_\lambda) - \nu(R, P_{\lambda'}) \le \epsilon_{\text{stop}}. \tag{5c}$$

The implementation of (5a) and (5b) is discussed in parts B and C of this section, respectively. Local convergence of the proposed algorithm is proved in Section III.

B. Source MDI PD Estimation

The estimation of the MDI PD in (5a) incorporates an infimum rather than a minimum, since the minimum may not exist. The following theorem, however, provides conditions for the existence and uniqueness of this MDI PD. The theorem and its proof are a straightforward extension of the results derived by Csiszár [35] and developed by Gray *et al.* [25] for the case where the model is a single Gaussian process.

Theorem 1: Let P_λ be a zero mean Gaussian HMM as in (1), and let $R = \{R_t,\ t = 0, \cdots, T\}$ be the sequence of given partial covariance matrices, specified within a band B, for a zero mean source. Let $\Omega(R)$ be the set of all PD's Q which satisfy (2). Let $\Psi = \{\Psi_t,\ t = 0, \cdots, T\}$ be a sequence of real symmetric matrices which vanish outside the band B. Let $\delta_{P_\lambda} \triangleq \{\Psi\colon S_\beta^{-1} + \Psi_t$ is positive definite for each $t = 0, \cdots, T$ and $\beta = 1, \cdots, M\}$.

a) If for some t, R_t does not have any positive definite extension, then $D(Q\|P_\lambda) = \infty$ for all $Q \in \Omega(R)$ and hence $\nu(R, P_\lambda) = \infty$.

b) If each R_t has any positive definite extension, then a unique PD Q_λ exists that minimizes $D(Q\|P_\lambda)$ over $Q \in \Omega(R)$. The pdf of Q_λ is given by

$$\begin{aligned} q_\lambda(y) &= Cp_\lambda(y)\exp\left(-\frac{1}{2}\sum_{\tau=0}^{T} y_\tau^{\#}\Lambda_\tau y_\tau\right) \\ &= \sum_x \frac{\prod_{\tau=0}^{T} a_{x_{\tau-1}x_\tau}\det^{-1/2}\left(I + S_{x_\tau}\Lambda_\tau\right)}{\sum_{x'}\prod_{\tau=0}^{T} a_{x'_{\tau-1}x'_\tau}\det^{-1/2}\left(I + S_{x'_\tau}\Lambda_\tau\right)} \\ &\quad \cdot \prod_{\tau=0}^{T} \frac{\exp\left(-\frac{1}{2}y_\tau^{\#}\left(S_{x_\tau}^{-1} + \Lambda_\tau\right)y_\tau\right)}{(2\pi)^{N/2}\det^{-1/2}\left(S_{x_\tau}^{-1} + \Lambda_\tau\right)} \end{aligned} \tag{6}$$

where C is a finite normalization factor which makes $\int dy\, q_\lambda(y) = 1$ and $\Lambda \in \delta_{P_\lambda}$ is a sequence of Lagrange multiplier matrices corresponding to the covariance constraints R. The PD Q_λ is the MDI PD with respect to P_λ, or the I projection of P_λ on $\Omega(R)$ [35], and it yields a finite MDI measure given by

$$\begin{aligned} \nu(R, P_\lambda) &= D(Q_\lambda\|P_\lambda) \\ &= -\ln\left[\sum_x \prod_{\tau=0}^{T} a_{x_{\tau-1}x_\tau}\det^{-1/2}\left(I + S_{x_\tau}\Lambda_\tau\right)\right] \\ &\quad - \frac{1}{2}\operatorname{tr}\sum_{\tau=0}^{T}(R_\tau\Lambda_\tau). \end{aligned} \tag{7}$$

Note that, for each $0 \le t \le T$, the trace of $R_t\Lambda_t$ in (7) depends only on the elements of R_t which are within the band B, since Λ_t vanishes outside this band.

Using (6), it is easy to see that the Lagrange multiplier matrices must be chosen to satisfy the following set of equations:

$$\begin{aligned} R_t^* &\triangleq E_{Q_\lambda}\{y_t y_t^{\#}\} \\ &= \sum_{\beta=1}^{M} q_t(\beta)\left(S_\beta^{-1} + \Lambda_t\right)^{-1} \\ &= R_t \text{ within the band } B, \qquad 0 \le t \le T \end{aligned} \tag{8}$$

where R_t^* is called the MDI extension of R_t with respect to P_λ and

$$q_t(\beta) \triangleq \frac{\sum_{\{x:\, x_t = \beta\}}\prod_{\tau=0}^{T} a_{x_{\tau-1}x_\tau}\det^{-1/2}\left(I + S_{x_\tau}\Lambda_\tau\right)}{\sum_{\beta=1}^{M}\sum_{\{x:\, x_t=\beta\}}\prod_{\tau=0}^{T} a_{x_{\tau-1}x_\tau}\det^{-1/2}\left(I + S_{x_\tau}\Lambda_\tau\right)} \tag{9}$$

is the probability, induced by the MDI pdf (6), of being in state β at time t (see (20) and the discussion after (27)). Kullback [21, p. 38] and Csiszár [35, secs. 1, 3] have shown that if $Q_\lambda \in \Omega(R)$ with pdf q_λ as in (6) exists, then it is the unique I projection of P_λ on $\Omega(R)$. This means that the set of (8) must have a unique solution for Λ within δ_{P_λ}, for if not, we could find two sets of solutions, say Λ and Λ',

and construct by (6) two I projections for P_λ on $\Omega(R)$. Equations (8), however, are difficult to solve in any straightforward manner. The following corollary of Theorem 1 provides an alternative way to estimate the Lagrange multiplier matrices by replacing the algebraic problem in (8) by a constrained minimization problem in the Euclidean space. A similar approach was suggested in [25], [36].

Corollary 1: Let P_λ, R, Ψ, and δ_{P_λ} be as in Theorem 1. Define

$$d(R;\Psi,\lambda) \triangleq \ln\left[\sum_x \sum_{\tau=0}^{T} a_{x_{\tau-1}x_\tau} \det^{-1/2}\left(I + S_{x_\tau}\Psi_\tau\right)\right] + \frac{1}{2}\operatorname{tr}\sum_{\tau=0}^{T}(R_\tau\Psi_\tau). \tag{10}$$

If each R_t has any positive definite extension, then $d(R;\Psi,\lambda)$ is a unimodal function of Ψ on δ_{P_λ}, and

$$\nu(R,P_\lambda) = -d(R;\Lambda,\lambda) = -\min_{\Psi\in\delta_{P_\lambda}} d(R;\Psi,\lambda). \tag{11}$$

Note that, due to the unimodality of $d(R;\Psi,\lambda)$ on δ_{P_λ}, the minimization in (11) can be done by any standard constrained optimization procedure in the Euclidean space.

C. HMM Estimation

Assume that Q_λ, the MDI PD with respect to P_λ, is given. We now show how a new HMM $P_{\lambda'}$ which reduces $D(Q_\lambda\|P_\lambda)$, or at least keeps it constant, is estimated. Let

$$p_\lambda(x,y) \triangleq \prod_{t=0}^{T} a_{x_{t-1}x_t} b(y_t|x_t) \tag{12}$$

and

$$q_\lambda(x,y) \triangleq C p_\lambda(x,y)\exp\left(-\frac{1}{2}\sum_{t=0}^{T} y_t^{\#}\Lambda_t y_t\right) \tag{13}$$

be, respectively, the joint pdf's of states and observations sequences for the old model P_λ and the corresponding MDI PD Q_λ. Using Jensen's inequality, (6), (12), and (13), we have that

$$\begin{aligned} D(Q_\lambda\|P_\lambda) - D(Q_\lambda\|P_{\lambda'}) &= \int q_\lambda(y)\ln\left[\sum_x p_{\lambda'}(x,y)/p_\lambda(y)\right]dy \\ &= \int q_\lambda(y)\ln\sum_x \frac{p_\lambda(x,y)}{p_\lambda(y)}\frac{p_{\lambda'}(x,y)}{p_\lambda(x,y)}\,dy \\ &\geq \sum_x \int \frac{q_\lambda(y)}{p_\lambda(y)} p_\lambda(x,y)\ln\frac{p_{\lambda'}(x,y)}{p_\lambda(x,y)}\,dy \\ &= \sum_x \int q_\lambda(x,y)\ln\frac{p_{\lambda'}(x,y)}{p_\lambda(x,y)}\,dy \\ &\triangleq \phi(\lambda') - \phi(\lambda) \end{aligned} \tag{14}$$

where

$$\phi(\lambda') \triangleq \sum_x \int q_\lambda(x,y)\ln p_{\lambda'}(x,y)\,dy. \tag{15}$$

Equality in (14) holds if and only if $p_{\lambda'}(x,y) = p_\lambda(x,y)$ almost everywhere with respect to $q_\lambda(x,y)$ (a.e. Q_λ). Otherwise, if $\phi(\lambda') > \phi(\lambda)$, then $D(Q_\lambda\|P_{\lambda'}) < D(Q_\lambda\|P_\lambda)$. Hence a new model which decreases the MDI measure, or at least keeps its value constant, can be found by maximizing $\phi(\lambda')$ over all feasible parameter sets λ'.

This approach for estimating a new model through maximization of the auxiliary function $\phi(\lambda')$ is a generalization of the Baum algorithm for ML hidden Markov modeling [9]. In the latter case, the auxiliary function which must be maximized so that the likelihood associated with the new model is equal to or greater than that associated with the old model is given by

$$\sum_x p_\lambda(x|y)\ln p_{\lambda'}(x,y) \tag{16}$$

where y are the observations from the source. Formal comparison of (15) and (16) shows that the MDI and the ML hidden Markov modeling approaches result in the same model estimate, when starting from the same initial model, if

$$q_\lambda(x,z) = p_\lambda(x|z)\,\delta(z-y) \tag{17}$$

where $\delta(\cdot)$ is a Dirac function. This condition, however, cannot be satisfied by the MDI modeling algorithm, since the MDI pdf (13) does not approach the pdf in (17) for any value of the Lagrange multipliers Λ. Further discussion on the relation between MDI and ML hidden Markov modeling will be given in Section IV.

The above procedure for estimating a new model resembles a single iteration of the *EM* algorithm [17], where the evaluation of $\phi(\cdot)$ corresponds to the *E*-step and the maximization of $\phi(\cdot)$ corresponds to the *M*-step of this algorithm. This procedure results in the so-called reestimation formulas since the new set of parameters is given in terms of the old set of parameters.

The maximization of (15) over λ' is done as follows. On substituting $p_{\lambda'}(x,y)$ from (12) into (15) the reestimation problem becomes

$$\begin{aligned} \max_{\lambda'}\Bigg\{ &\sum_x \ln\pi'_{x_0}\int q_\lambda(x,y)\,dy \\ &+ \sum_x\sum_{t=1}^{T}\ln a'_{x_{t-1}x_t}\int q_\lambda(x,y)\,dy \\ &- \frac{1}{2}\operatorname{tr}\sum_x\sum_{t=0}^{T} S'^{-1}_{x_t}\int q_\lambda(x,y)\,y_t y_t^{\#}\,dy \\ &+ \frac{1}{2}\sum_x\sum_{t=0}^{T}\ln\det S'^{-1}_{x_t}\int q_\lambda(x,y)\,dy\Bigg\} \end{aligned} \tag{18}$$

or, equivalently,

$$\max_{\lambda'}\Bigg\{\sum_{\beta=1}^{M}\ln\pi'_\beta\sum_{\{x:\,x_0=\beta\}}\int q_\lambda(x,y)\,dy + \sum_{\alpha,\beta=1}^{M}\ln a'_{\alpha\beta}\sum_{t=1}^{T}\sum_{\left\{x:\ \substack{x_{t-1}=\alpha\\ x_t=\beta}\right\}}\int q_\lambda(x,y)\,dy - \frac{1}{2}\operatorname{tr}\sum_{\beta=1}^{M}S'^{-1}_\beta\sum_{t=0}^{T}\sum_{\{x:\,x_t=\beta\}}\int q_\lambda(x,y)\,y_t y_t^{\#}\,dy + \frac{1}{2}\sum_{\beta=1}^{M}\ln\det S'^{-1}_\beta\sum_{t=0}^{T}\sum_{\{x:\,x_t=\beta\}}\int q_\lambda(x,y)\,dy\Bigg\}. \tag{19}$$

From (12), (13), and (9), it can be shown that

$$q_t(\beta)=\sum_{\{x:\,x_t=\beta\}}\int q_\lambda(x,y)\,dy,\qquad 0\le t\le T$$

$$q_t(\alpha,\beta)\triangleq\sum_{\{x:\,x_{t-1}=\alpha x_t=\beta\}}\int q_\lambda(x,y)\,dy = \frac{\displaystyle\sum_{\left\{x:\ \substack{x_{t-1}=\alpha\\ x_t=\beta}\right\}}\prod_{\tau=0}^{T}a_{x_{\tau-1}x_\tau}\det^{-1/2}\left(I+S_{x_\tau}\Lambda_\tau\right)}{\displaystyle\sum_{\alpha,\beta=1}^{M}\sum_{\left\{x:\ \substack{x_{t-1}=\alpha\\ x_t=\beta}\right\}}\prod_{\tau=0}^{T}a_{x_{\tau-1}x_\tau}\det^{-1/2}\left(I+S_{x_\tau}\Lambda_\tau\right)},\qquad 0<t\le T \tag{20}$$

and clearly $q_t(\beta)=\sum_{\alpha=1}^{M}q_t(\alpha,\beta)$ for $0<t\le T$. Similarly,

$$\sum_{\{x:\,x_t=\beta\}}\int q_\lambda(x,y)\,y_t y_t^{\#}\,dy = q_t(\beta)\left(S_\beta^{-1}+\Lambda_t\right)^{-1}\triangleq R_t(\beta),\qquad 0\le t\le T. \tag{21}$$

Hence, using (20) and (21) we can rewrite (19) as

$$\max_{\lambda'}\Bigg\{\sum_{\beta=1}^{M}\ln\pi'_\beta\,q_0(\beta) + \sum_{\alpha,\beta=1}^{M}\ln a'_{\alpha\beta}\sum_{t=1}^{T}q_t(\alpha,\beta) - \frac{1}{2}\sum_{\beta=1}^{M}\left[\operatorname{tr}\left(S'^{-1}_\beta\sum_{t=0}^{T}R_t(\beta)\right) - \ln\det S'^{-1}_\beta\sum_{t=0}^{T}q_t(\beta)\right]\Bigg\}. \tag{22}$$

The maximization problem can now be summarized as follows. The initial state probability vector of the Markov chain is obtained from

$$\max_{\pi'}\sum_{\beta=1}^{M}\ln\pi'_\beta q_0(\beta)\quad\text{subject to}\quad\sum_{\beta=1}^{M}\pi'_\beta=1,\qquad \pi'_\beta\ge 0,\quad \beta=1,\cdots,M. \tag{23}$$

The state transition probability matrix of the Markov chain is obtained from

$$\max_{a'}\sum_{\alpha,\beta=1}^{M}\ln a'_{\alpha\beta}\sum_{t=1}^{T}q_t(\alpha,\beta)\quad\text{subject to}\quad\sum_{\beta=1}^{M}a'_{\alpha\beta}=1,\qquad \alpha=1,\cdots,M,\quad a'_{\alpha\beta}\ge 0. \tag{24}$$

The covariance matrices of the output processes corresponding to the different states of the Markov chain are obtained from

$$\min_{S'_\beta}\left\{\operatorname{tr}\left(S'^{-1}_\beta\sum_{t=0}^{T}R_t(\beta)\right)-\ln\det S'^{-1}_\beta\sum_{t=0}^{T}q_t(\beta)\right\}\quad\text{subject to } S'_\beta\text{ is positive definite},\qquad \beta=1,\cdots,M. \tag{25}$$

The maximization in (23) results in

$$\pi'_\beta=q_0(\beta),\qquad \beta=1,\cdots,M. \tag{26}$$

Similarly, the maximization in (24) results in

$$a'_{\alpha\beta}=\frac{\displaystyle\sum_{t=1}^{T}q_t(\alpha,\beta)}{\displaystyle\sum_{\beta=1}^{M}\sum_{t=1}^{T}q_t(\alpha,\beta)},\qquad \alpha,\beta=1,\cdots,M, \tag{27}$$

provided that

$$\sum_{\beta=1}^{M}\sum_{t=1}^{T}q_t(\alpha,\beta)>0.$$

If not, then $\sum_{t=1}^{T}q_t(\alpha,\beta)=0$, and any $a'_{\alpha\beta}$ that satisfies the constraints in (24) can be chosen without affecting the value of (24). Note that $\sum_{t=1}^{T}q_t(\alpha,\beta)=0$ if and only if $q_t(\alpha,\beta)=0$ for all $0<t\le T$. From (20), however, $q_t(\alpha,\beta)$ is the joint probability, under $q_\lambda(x,y)$, of being in state α at time $t-1$ and in state β at time t. Hence $a'_{\alpha\beta}$ is arbitrarily chosen, up to the constraints in (24), for the forbidden states α and β.

The minimization in (25) is considered for zero mean Gaussian AR HMM's. Suppose first that $\sum_{t=0}^{T}q_t(\beta)>0$. The problem then becomes

$$\min_{S'_\beta}\left\{\operatorname{tr}\left(R(\beta)S'^{-1}_\beta\right)-\ln\det S'^{-1}_\beta\right\},\qquad \beta=1,2,\cdots,M \tag{28}$$

where $R(\beta)$ is a positive definite covariance matrix defined by

$$R(\beta) \triangleq \sum_{t=0}^{T} R_t(\beta) \Big/ \sum_{t=0}^{T} q_t(\beta), \qquad \beta = 1, \cdots, M. \quad (29)$$

This is exactly the problem that arises in ML estimation of structured covariance matrices given a measured covariance matrix [37], [38]. In our case we are interested in estimating the covariance matrix S'_β of an rth-order AR process given $R(\beta)$. S'_β is given by $S'_\beta = \sigma_\beta^2 (L_\beta^{\#} L_\beta)^{-1}$, where σ_β^2 is a gain constant and L_β is an $N \times N$ lower triangular matrix whose i, jth element is given by

$$l_\beta(i, j) = \begin{cases} f_\beta(i-j), & 0 \le i-j \le r \\ 0, & \text{otherwise,} \end{cases}$$

$f_\beta(0) = 1$, and $f_\beta(i)$, $i = 1, \cdots, r$, are the coefficients of the AR process. Since $R(\beta)$ is positive definite, the set of all AR covariance matrices S'_β is a closed subset of the set of positive semidefinite symmetric matrices, and the set of all inverses of AR covariance matrices with $\sigma_\beta^2 > 0$ is convex, there exists a unique positive definite matrix S'_β that minimizes (28) [38, theorem 2]. Since $\det(L_\beta) = 1$, the coefficients $f_\beta(\cdot)$ are obtained from the minimization of $\mathrm{tr}(R(\beta) L_\beta^{\#} L_\beta)$. From [25, corollary 2], this is done by minimizing the quadratic form

$$\epsilon \triangleq \sum_{n=0}^{r} \sum_{m=0}^{r} f_\beta(n) f_\beta(m) \frac{1}{N} \sum_{k=\max(n,m)}^{N-1} r_\beta(k-n, k-m) \quad (30)$$

where $r_\beta(\cdot, \cdot)$ are the elements of $R(\beta)$. This results in a set of linear equations similar to that obtained in the "covariance method" for linear prediction analysis [29, p. 14]. The gain constant σ_β^2 which minimizes (28) equals the minimal value of ϵ in (30).

If $\sum_{t=0}^{T} q_t(\beta)$ in (25) equals zero, then $q_t(\beta) = 0$ for all t, and from (21) $R_t(\beta) = 0$. Hence any positive definite AR covariance matrix S'_β can be chosen since its value does not affect (25).

We now show how $q_t(\alpha, \beta)$ and $q_t(\beta)$ in (20) and $d(R; \Psi, \lambda)$ in (10) can be efficiently calculated using the forward–backward formulas. Define

$$F_t(\alpha) \triangleq \sum_{\substack{\{x_0, \cdots, x_{t-1}\} \\ x_t = \alpha}} \prod_{\tau=0}^{t} a_{x_{\tau-1} x_\tau} \det{}^{-1/2}(I + S_{x_\tau} \Lambda_\tau) \quad (31)$$

$$B_t(\beta) \triangleq \sum_{\substack{\{x_{t+1}, \cdots, x_T\} \\ x_t = \beta}} \prod_{\tau=t+1}^{T} a_{x_{\tau-1} x_\tau} \det{}^{-1/2}(I + S_{x_\tau} \Lambda_\tau) \quad (32)$$

$B_T(\beta) \triangleq 1$, and note that

$$F_0(\alpha) = \pi_\alpha \det{}^{-1/2}(I + S_\alpha \Lambda_0)$$

$$F_t(\alpha) = \sum_{\gamma=1}^{M} F_{t-1}(\gamma) a_{\gamma\alpha} \det{}^{-1/2}(I + S_\alpha \Lambda_t), \qquad 0 < t \le T, \quad (33)$$

$$B_t(\beta) = \sum_{\gamma=1}^{M} B_{t+1}(\gamma) a_{\beta\gamma} \det{}^{-1/2}(I + S_\gamma \Lambda_{t+1}), \qquad 0 \le t < T. \quad (34)$$

From (9), (20), (31), and (32) we have

$$q_t(\beta) = \frac{F_t(\beta) B_t(\beta)}{\sum_{\beta=1}^{M} F_t(\beta) B_t(\beta)}, \quad 0 \le t \le T,$$

$$q_t(\alpha, \beta) = \frac{F_{t-1}(\alpha) B_t(\beta) a_{\alpha\beta} \det{}^{-1/2}(I + S_\beta \Lambda_t)}{\sum_{\alpha=1}^{M} \sum_{\beta=1}^{M} F_{t-1}(\alpha) B_t(\beta) a_{\alpha\beta} \det{}^{-1/2}(I + S_\beta \Lambda_t)}, \qquad 0 < t \le T. \quad (35)$$

The argument of the logarithm in (10) has the same form as the denominator of $q_t(\beta)$ in (9) for any $0 \le t \le T$. This argument can therefore be calculated similarly to the denominator of $q_t(\beta)$ in (35) with Λ_t in (33), (34) replaced by Ψ_t.

III. Convergence Analysis

In this section we analyze the MDI algorithm for hidden Markov modeling developed in Section II and prove its local convergence. As we have seen, in each iteration of this algorithm, the MDI PD with respect to a given model is first estimated, and then a new model which reduces the resulting MDI measure, or at least keeps its value constant, is estimated. For sources characterized by a given set of partial covariance matrices such that each of them has a positive definite extension and zero mean Gaussian AR HMM's, we have shown that a unique MDI PD with respect to a given model exists, and a new model which reduces the MDI measure, or at least keeps its value constant, can always be found. The new model is, however, not unique since some of its parameters (those corresponding to forbidden state transitions) can be arbitrarily chosen from the feasible set of parameters (see, e.g., the discussion following (27)).

Suppose that each given partial covariance matrix for the source has a positive definite extension. Let P_λ be, as above, the PD of a given model and Q_λ be the MDI PD with respect to P_λ. Then, for any PD $Q \in \Omega(R)$ we have the following inequality:

$$D(Q \| P_\lambda) \ge D(Q_\lambda \| P_\lambda) = \nu(R, P_\lambda) \quad (36)$$

where, due to the uniqueness of Q_λ, equality holds if and only if $Q = Q_\lambda$. Now, given Q_λ, the new model $P_{\lambda'}$ is

chosen so that

$$D(Q_\lambda \| P_\lambda) \geq D(Q_\lambda \| P_{\lambda'}). \tag{37}$$

Since

$$D(Q_\lambda \| P_\lambda) - D(Q_\lambda \| P_{\lambda'}) = \int dy\, q_\lambda(y) \ln\left(p_{\lambda'}(y)/p_\lambda(y)\right), \tag{38}$$

equality in (37) holds if and only if $P_\lambda = P_{\lambda'}$ a.e. Q_λ. Combining (36) and (37), we obtain the following inequality:

$$\begin{aligned} \nu(R, P_\lambda) &= D(Q_\lambda \| P_\lambda) \geq D(Q_\lambda \| P_{\lambda'}) \\ &\geq D(Q_{\lambda'} \| P_{\lambda'}) = \nu(R, P_{\lambda'}). \end{aligned} \tag{39}$$

Thus the MDI measure associated with the new model $P_{\lambda'}$ is lower than or equal to that associated with the initial model P_λ. If $\nu(R, P_\lambda) = \nu(R, P_{\lambda'})$, then from (39) we have that $D(Q_\lambda \| P_\lambda) = D(Q_\lambda \| P_{\lambda'}) = D(Q_{\lambda'} \| P_{\lambda'})$, which by (36) and (37) implies that $P_\lambda = P_{\lambda'}$ a.e. Q_λ. Based on this discussion we have the following lemma.

Lemma 1: Assume that each given partial covariance matrix for the source has a positive definite extension. Let P_λ be a given HMM, Q_λ be the MDI PD with respect to P_λ, and $P_{\lambda'}$ be an estimated new HMM. Then

$$\nu(R, P_\lambda) \geq \nu(R, P_{\lambda'}) \tag{40}$$

and equality holds if and only if $P_\lambda = P_{\lambda'}$ a.e. Q_λ.

Lemma 1 shows that the algorithm generates a sequence of HMM's, say P_{λ_n}, for which $\nu(R, P_{\lambda_n})$ is a strictly decreasing sequence, unless $\nu(R, P_{\lambda_{n+1}}) = \nu(R, P_{\lambda_n})$. In the latter case $P_{\lambda_n} = P_{\lambda_{n+1}}$ a.e. Q_{λ_n}, where Q_{λ_n} is the MDI PD with respect to P_{λ_n}, and a fixed point of the algorithm is reached. Since $\nu(R, P_{\lambda_n}) \geq 0$, the limit $\lim_{n\to\infty} \nu(R, P_{\lambda_n})$ exists. Unfortunately, however, this neither guarantees the convergence of the model sequence P_{λ_n} to a fixed point nor that a fixed point should ever be reached. Hence convergence of the model sequence should be examined. Note that since P_λ is a continuous function of λ (see (1)), and the corresponding MDI PD Q_λ is a continuous function of λ and Λ (see (6)), convergence can be equivalently considered in terms of either $(P_{\lambda_n}, Q_{\lambda_n})$ or (λ_n, Λ_n).

Let

$$\zeta(P_{\lambda_n})\colon\ P_{\lambda_n} \to (P_{\lambda_n}, Q_{\lambda_n}) \tag{41}$$

be the "point-to-point" mapping from the model P_{λ_n} to itself and its MDI PD Q_{λ_n}. This mapping is exactly determined by the procedure provided by Corollary 1. Let

$$\mu(P_{\lambda_n}, Q_{\lambda_n})\colon (P_{\lambda_n}, Q_{\lambda_n}) \to \{P_{\lambda_{n+1}}\}_{Q_{\lambda_n}} \tag{42}$$

be the "point-to-set" mapping from the pair of PD's $(P_{\lambda_n}, Q_{\lambda_n})$ to the set of Q_{λ_n} equivalence models $P_{\lambda_{n+1}}$. Each of these models results from the maximization of the auxiliary function,

$$g(P_{\lambda_n}, Q_{\lambda_n}; P_{\lambda_{n+1}}) \triangleq \sum_x \int q_{\lambda_n}(x, z) \ln p_{\lambda_{n+1}}(x, z)\, dz, \tag{43}$$

over all λ_{n+1}, as was shown in Section II-C. The algorithm is now defiend as the composition of these two mappings as follows:

$$T_R(P_{\lambda_n})\colon\ P_{\lambda_n} \to \{P_{\lambda_{n+1}}\}_{Q_{\lambda_n}} \qquad T_R(P_{\lambda_n}) = \mu\left(\zeta(P_{\lambda_n})\right). \tag{44}$$

We have the following theorem.

Theorem 2: Assume that each given partial covariance matrix has a positive definite extension. Let P_{λ_0} be an initially given zero mean Gaussian AR HMM, and let $P_{\lambda_{n+1}} \in T_R(P_{\lambda_n})$, $n \geq 0$. Let $\Gamma \triangleq \{P_\lambda\colon\ P_\lambda = T_R(P_\lambda)$ a.e. $Q_\lambda\}$ be the set of fixed points of T_R, where Q_λ is the MDI PD with respect to P_λ. If all parameters of AR models generated by T_R are in a compact subset of the Euclidean space, then

1) each accumulation point P_{λ^*} of $[P_{\lambda_n}]_{n=0}^\infty$ is a fixed point, i.e., $P_{\lambda^*} \in \Gamma$;
2) $\rho(P_{\lambda_n}, \Gamma) \to 0$, where ρ is the usual distance in the Euclidean space;
3) $\nu(R, P_{\lambda_n}) \to \nu(R, P_{\lambda^*})$.

The theorem says that the limit of any convergent subsequence of $[P_{\lambda_n}]_{n=0}^\infty$ is a fixed point of the algorithm T_R. Since this sequence lies in a compact space, the existence of at least one convergence subsequence is guaranteed, and hence the set Γ is not empty. Furthermore, the theorem states that the sequence of HMM's generated by T_R approaches the set of fixed points of T_R and that the MDI sequence approaches the MDI which corresponds to some fixed point. The theorem, however, does not guarantee the convergence of the model sequence to any specific fixed point of T_R.

The following Lemma establishes the relations between a fixed point of the algorithm and a stationary point of the MDI measure.

Lemma 2: Assume that each given partial covariance matrix has a positive definite extension. Let T_R be an algorithm as in Theorem 2. Then any fixed point of T_R is a stationary point of the MDI measure.

IV. ML and MDI Modeling Approaches

In this section a relation between the MDI approach and an ML approach for hidden Markov modeling is established based upon the results obtained in [26] for Gaussian models. Consider the MDI measure (7), which can be written as

$$\nu(R, P_\lambda) = -\ln \sum_x \left[\prod_{\tau=0}^{T} a_{x_{\tau-1}x_\tau} \det^{-1/2}\left(I + S_{x_\tau}\Lambda_\tau\right) \cdot \exp\left\{1/2 \operatorname{tr}\left(R_\tau \Lambda_\tau\right)\right\} \right] \tag{45}$$

and assume that there exists a unique sequence of states, say x^*, which dominates the sum in (45). In this case, the

MDI measure can be approximated by

$$\nu(R, P_\lambda) \approx -\ln \max_x \left[\prod_{\tau=0}^{T} a_{x_{\tau-1}x_\tau} \det^{-1/2}(I + S_{x_\tau}\Lambda_\tau) \cdot \exp\{1/2 \operatorname{tr}(R_\tau \Lambda_\tau)\} \right]$$

$$= -\sum_{\tau=0}^{T} \ln a_{x_{\tau-1}^* x_\tau^*} - \frac{1}{2}\sum_{\tau=0}^{T} \operatorname{tr}(R_\tau\Lambda_\tau) - \ln\det(I + S_{x_\tau^*}\Lambda_\tau). \quad (46)$$

Furthermore, from (20), we have for $0 < t \le T$ that

$$q_t(\alpha, \beta) \approx \begin{cases} 1, & \alpha = x_{t-1}^*, \ \beta = x_t^* \\ 0, & \text{otherwise} \end{cases} \quad (47)$$

and from (9) we have for all $0 \le t \le T$ that $q_t(\beta) \approx 1$ if $\beta = x_t^*$ and $q_t(\beta) \approx 0$ otherwise. Hence, from (8) an approximate Lagrange multiplier matrix for each time t is obtained from the solution of the following set of equations.

$$R_t^* = \left(S_{x_t^*}^{-1} + \Lambda_t\right)^{-1}. \quad (48)$$

Note that the approximate Lagrange multiplier matrix Λ_t depends on the state x_t^*, not only on t as is the case in (8). If for each t the given partial covariance matrix comprises the $J \times J$ ($J \le N$) principal leading block of the original covariance matrix of the source, $R_t^\dagger$, then a closed-form solution exists for the approximate Lagrange multiplier matrix Λ_t, which upon substitution in (46) gives [26, theorem 1];

$$\nu(R, P_\lambda) \approx -\sum_{\tau=0}^{T} \ln a_{x_{\tau-1}^* x_\tau^*} + \frac{1}{2}\sum_{\tau=0}^{T} \left[\operatorname{tr}\left(R_{J,\tau} S_{J,x_\tau^*}^{-1}\right) - \ln\det\left(R_{J,\tau} S_{J,x_\tau^*}^{-1}\right) - J\right] \quad (49)$$

where $R_{J,\tau}$ and S_{J,x_τ^*} are the $J \times J$ principal leading blocks of $R_\tau^\dagger$ and $S_{x_\tau^*}$, respectively. Hence, given the dominant sequence of states x^*, approximate MDI hidden Markov modeling can be performed by minimizing (49) over all feasible parameter sets λ. This can be done in a way similar to the maximization in (22) using $q_t(\alpha, \beta)$ and $q_t(\beta)$ from (47). The value of J should be at least the order of the AR models plus one for the normal equations which result from (30) to have a unique solution.

Given a model P_λ, the dominant sequence of state x^* can be estimated by examining the value of

$$\prod_{\tau=0}^{T} a_{x_{\tau-1}x_\tau} \det^{-1/2}(I + S_{x_\tau}\Lambda_\tau) \exp\{1/2 \operatorname{tr}(R_\tau\Lambda_\tau)\} \quad (50)$$

or of its logarithm, for each possible sequence of states x. Since each sequence is considered independently, (47) and (48) apply to the examined sequence, say x, and the dominant sequence of states x^* is obtained from

$$\max_x \left\{ \sum_{\tau=0}^{T} \ln a_{x_{\tau-1}x_\tau} - \frac{1}{2}\left[\operatorname{tr}\left(R_{J,\tau} S_{J,x_\tau}^{-1}\right) - \ln\det\left(R_{J,\tau} S_{J,x_\tau}^{-1}\right) - J\right] \right\}. \quad (51)$$

The maximization of (51) can be efficiently performed by applying the Viterbi algorithm [39] using the path metric

$$\ln a_{x_{\tau-1}x_\tau} - \frac{1}{2}\left[\operatorname{tr}\left(R_{J,\tau} S_{J,x_\tau}^{-1}\right) - \ln\det\left(R_{J,\tau} S_{J,x_\tau}^{-1}\right) - J\right], \quad \tau = 0, \cdots, T. \quad (52)$$

The above discussion suggests that approximate MDI hidden Markov modeling of sources for which the given partial covariance matrices comprise the $J \times J$ principal leading blocks of the full covariance matrices can be achieved by alternating minimization of

$$-\sum_{\tau=0}^{T} \ln a_{x_{\tau-1}x_\tau} - \frac{1}{2}\left[\operatorname{tr}\left(R_{J,\tau} S_{J,x_\tau}^{-1}\right) - \ln\det\left(R_{J,\tau} S_{J,x_\tau}^{-1}\right) - J\right], \quad (53)$$

once over all sequences of states assuming that the model is known, and then over all HMM's assuming that the minimizing sequence of states is available. Both minimizations can be efficiently performed, the first by the Viterbi algorithm and the second by a variant of the Baum algorithm. This results in a descent algorithm for the approximate MDI measure (49). This procedure is equivalent to the so-called segmental k-means algorithm [40], [3] in the speech recognition area, which aims at

$$\max_{x,\lambda} \ln p_\lambda(x, y) \quad (54)$$

where $p_\lambda(x, y)$ is given in (12). Note that (54) is an approximation of the original ML approach developed by Baum *et al.* [9], [10] which aims at

$$\max_\lambda \ln \sum_x p_\lambda(x, y). \quad (55)$$

For sources that exhibit stationary properties, e.g., asymptotically weakly stationary (AWS) sources [26], we can approximate the second term in (49) (and similarly in (51)–(53)) by its asymptotic form [26, theorem 2] obtained for $J \to \infty$ and get for large J

$$\nu(R, P_\lambda) \approx -\sum_{\tau=0}^{T} \ln a_{x_{\tau-1}^* x_\tau^*} + \frac{J}{2}\sum_{\tau=0}^{T} \int_0^{2\pi} \left[\frac{f_\tau(\theta)}{g_{x_\tau^*}(\theta)} - \ln\frac{f_\tau(\theta)}{g_{x_\tau^*}(\theta)} - 1\right]\frac{d\theta}{2\pi},$$

$$= -\sum_{\tau=0}^{T} \left[\ln a_{x_{\tau-1}^* x_\tau^*} - \frac{J}{2} d_{IS}\left(f_\tau(\theta), g_{x_\tau^*}(\theta)\right)\right] \quad (56)$$

where $f_t(\theta)$ and $g_{x_t^*}(\theta)$ are, respectively, the power spec-

tral densities associated with the given covariance of the source at time t and the covariance of the output Gaussian process from state x_t^*, and $d_{IS}(f_t(\theta), g_{x_t^*}(\theta))$ is the well-known Itakura–Saito distortion measure between $f_t(\theta)$ and $g_{x_t^*}(\theta)$ [30].

V. Comments

We have proposed a new information theoretic approach, which is optimal in the MDI sense, for Gaussian AR hidden Markov modeling of sources characterized by a given set of partial covariance matrices. The modeling is performed by alternating minimization of the discrimination information measure over the set of all PD's which satisfy the given partial covariance matrices from the source, and the set of all zero mean Gaussian AR HMM's. The algorithm aims at finding a pair of PD's, one in each set, that are closest to each other. We have shown that, for a given model and given partial covariance matrices where each one has a positive definite extension, the estimation of the PD of the source can be done by any standard constrianed minimization procedure in the Euclidean space. Furthermore, for a given PD of the source, the estimation of a new model can be efficiently done by a procedure which generalizes the Baum algorithm. Local convergence of the algorithm was proved under the mild assumption that the estimated parameters of the AR models are all in a compact (or bounded) subset of the Euclidean space. Finally, it was shown that the MDI modeling approach approximately becomes an ML modeling approach when the MDI measure is assumed to be concentrated in a single sequence of states of the model.

In principle, the MDI modeling approach can be applied, for any source, using any sequence of partial covariance matrices from that source. In practice, however, each partial covariance matrix R_t has to be estimated from the observation vector y_t. Since y_t is N-dimensional and the number of given elements in R_t is proportional to $N(N+1)/2$, the class of sources which can be modeled is restricted to those sources whose covariance characterization is compact in the sense that it is specified by a relatively small number of parameters: for example, vector stationary sources for which each R_t is Toeplitz; wide-sense asymptotically mean stationary (AMS) sources characterized by the averages of the elements along each diagonal of the covariance matrix [41], [42]; and sources in which the covariance matrices are circulant. In all of these examples, the partial covariance R_t is characterized by at most N elements and ergodic theorems that guarantee the consistency of the sample covariance estimator exist (see, e.g., [41], [43]). The case of sources with circulant covariance matrices is particularly interesting since, as can be seen from (8), if we also use Gaussian models with circulant covariance matrices, then the resulting Lagrange multiplier matrices must be circulant. This, of course, significantly simplifies the implementation of the MDI algorithm as each Lagrange multiplier matrix is characterized by at most $N/2+1$ elements, and all matrix operations can be performed using the FFT algorithm [44]. The circulant approximation of the covariance matrix of each AR output process of the model is commonly done in practice.

The MDI hidden Markov modeling approach proposed here can be extended without any principal difficulties to HMM's with output PD's other than Gaussian and sources characterized by any appropriate set of moments; for example, HMM's with mixtures of Gaussian AR output PD's [11], [8], or HMM's that are supplemented by time durational probabilistic models [13], [45]. The case of Gaussian AR hidden Markov modeling, given second-order statistics from the source, was chosen here to demonstrate the procedure and also because of its particular importance in speech recognition and enhancement applications. The extension of the MDI algorithm to sources which are characterized by multiple sequences of partial covariance matrices, as is the case in speech recognition when modeling is done from several utterances of the word being modeled, and for HMM's with mixtures of Gaussian AR output PD's can be found in [46].

The expected performance of the proposed MDI approach for hidden Markov modeling is as yet unknown since it has not been fully implemented or studied. The major difficulty in implementing the MDI approach is the calculation of the Lagrange multipliers which requires application of constrained minimization optimization methods. A theoretical investigation of the performance of the MDI approach may be possible only for sources that are of the same class of the models, i.e., Markov sources.

Acknowledgment

The authors wish to thank Dr. O. Zeitouni for some interesting discussions during this work. The authors also wish to thank the anonymous reviewers for critical reading of the manuscript and their helpful comments both with regard to interpreting some of the results and in clarifying the presentation.

Appendix

Lemmas A1 and A2 below are straightforward applications of [25, lemmas 1 and 2] to HMM's.

Lemma A1: Let $\Psi \triangleq \{\Psi_t,\ t=0,\cdots,T\}$, where Ψ_t is any real symmetric matrix which vanishes outside the band B. Define

$$\Theta_P \triangleq \left\{ \Psi:\ E_P\left\{ \exp\left(-\frac{1}{2} \sum_{t=0}^{T} y_t^{\#} \Psi_t y_t \right) \right\} < \infty \right\} \qquad \text{(A1)}$$

where E_P is the expectation with respect to the PD P whose pdf is given in (1). Then $\Psi \in \Theta_P$ if and only if $S_\beta^{-1} + \Psi_t$ is positive definite for every $t=0,1,\cdots,T$ and $\beta=1,\cdots,M$. Furthermore, Θ_P is open in R^K, where $K=(T+1)|B|$ and $|B|$ is the cardinality of the band B.

Proof:

$$E_P\left\{\exp\left(-\frac{1}{2}\sum_{t=0}^{T} y_t^{\#}\Psi_t y_t\right)\right\} = \sum_x \prod_{t=0}^{T} a_{x_{t-1}x_t} \frac{1}{(2\pi)^{N/2}\det^{1/2}(S_{x_t})} \cdot \int dy \prod_{t=0}^{T} \exp\left(-\frac{1}{2} y_t^{\#}\left(S_{x_t}^{-1}+\Psi_t\right) y_t\right). \quad \text{(A2)}$$

If $S_{x_t}^{-1}+\Psi_t$ is positive definite, then

$$\int dy_t \exp\left(-\frac{1}{2} y_t^{\#}\left(S_{x_t}^{-1}+\Psi_t\right) y_t\right) = (2\pi)^{N/2}\det^{-1/2}\left(S_{x_t}^{-1}+\Psi_t\right) < \infty$$

and hence

$$E_P\left\{\exp\left(-\frac{1}{2}\sum_{t=0}^{T} y_t^{\#}\Psi_t y_t\right)\right\} = \sum_x \prod_{t=0}^{T} a_{x_{t-1}x_t} \det^{-1/2}\left(S_{x_t}^{-1}+\Psi_t\right)\det^{-1/2}\left(S_{x_t}\right) < \infty. \quad \text{(A3)}$$

If

$$\int dy \prod_{t=0}^{T} \exp\left(-\frac{1}{2} y_t^{\#}\left(S_{x_t}^{-1}+\Psi_t\right) y_t\right) < \infty,$$

then from Fubini's theorem [47, p. 150],

$$\int dy_t \exp\left(-\frac{1}{2} y_t^{\#}\left(S_{x_t}^{-1}+\Psi_t\right) y_t\right) < \infty$$

for every t. This, however, happens if and only if $S_{x_t}^{-1}+\Psi_t$ is positive definite as was shown by Gray *et al.* [25, lemma 1].

This discussion shows that Θ_P contains precisely all sequences Ψ for which $S_{x_t}^{-1}+\Psi_t$ is positive definite for every t. As such, Θ_P is open in R^K since the eigenvalues of a matrix are continuous functions of its elements, and hence any small perturbation of matrix elements results in a small perturbation of its eigenvalues. Note also that Θ_P is not empty since, for example, $\Psi \equiv 0 \in \Theta_P$.

Lemma A2: Let R and $\Omega(R)$ be as in Theorem 1, and P be as in Lemma A1. Define

$$\Phi_P \triangleq \{ R\text{: for which there exists a PD } Q \in \Omega(R) \text{ with } D(Q\|P) < \infty \}. \quad \text{(A4)}$$

Then $R \in \Phi_P$ if and only if each R_t has a positive definite extension. Furthermore, Φ_P is open in R^K, where K is as in Lemma A1.

Proof: Let $V_t \triangleq E_Q\{y_t y_t^{\#}\}$ be an extension of R_t which results from a PD Q. Since V_t is symmetric, it can be written as $V_t = U_t \Psi_t U_t^{\#}$, where U_t is a unitary matrix and Ψ_t is a diagonal matrix which contains the eigenvalues of V_t. If R_t does not have any positive definite extension, then each extension of R_t has at least one eigenvalue, say the ith, which equals zero. Define $F \triangleq \{ y_t\text{: } U_{t_i} y_t = 0\}$, where U_{t_i} is the ith row of U_t corresponding to the zero eigenvalue, and note that $E_Q\{(U_{t_i} y_t)^2\} = 0$. This means that $Q(F) = 1$ while $P(F) = 0$. Hence Q is not absolutely continuous with respect to P and from (4a) we have that $D(Q\|P) = \infty$ for any Q, which implies that $\nu(R, P) = \infty$.

Suppose now that each R_t has a positive definite extension, say R_t^*, and consider the PD Q_{R^*} whose pdf is given by

$$q_{R^*}(y) = \prod_{\tau=0}^{T} \frac{\exp\left(-\frac{1}{2} y_\tau^{\#} R_\tau^{*-1} y_\tau\right)}{(2\pi)^{N/2}\det^{1/2}(R_\tau^*)}. \quad \text{(A5)}$$

For this PD, and the prior (1) which is written as

$$p(y) = \sum_x \prod_{\tau=0}^{T} a_{x_{\tau-1}x_\tau} p_S(x, y) \quad \text{(A6)}$$

where

$$p_S(x, y) \triangleq \prod_{\tau=0}^{T} \frac{\exp\left(-\frac{1}{2} y_\tau S_{x_\tau}^{-1} y_\tau\right)}{(2\pi)^{N/2}\det^{1/2}(S_{x_\tau})},$$

we have that

$$\begin{aligned} D(Q_{R^*}\|P) &= \int dy\, q_{R^*}(y)\ln q_{R^*}(y) - \int dy\, q_{R^*}(y)\ln p(y) \\ &\le \int dy\, q_{R^*}(y)\ln q_{R^*}(y) \\ &\quad - \int dy\, q_{R^*}(y)\ln p_S(x', y) - \sum_{\tau=0}^{T} \ln a_{x'_{\tau-1}x'_\tau} \\ &= D(Q_{R^*}\|P_S) - \sum_{\tau=0}^{T} \ln a_{x'_{\tau-1}x'_\tau} \\ &= \frac{1}{2}\sum_{\tau=0}^{T}\left[\operatorname{tr}\left(R_\tau^* S_{x'_\tau}^{-1}\right)\right. \\ &\quad \left. - \ln\det\left(R_\tau^* S_{x'_\tau}^{-1}\right) - N - 2\ln a_{x'_{\tau-1}x'_\tau}\right] \\ &< \infty \end{aligned} \quad \text{(A7)}$$

where P_S is the PD corresponding to p_S, $(x'_0, x'_1, \cdots, x'_T)$ is any Markov chain with strictly positive probability (there always exists at least one such chain), and the last equation in (A7) results from [21, p. 189]. Thus we have proved that Φ_P contains precisely all sequences R of partial covariance matrices for which each R_t has any positive definite extension. Φ_P can be shown to be open in R^K by arguments similar to these used in the proof of Lemma A1.

Proof of Theorem 1: a) Results from Lemma A2. b) In Csiszár [35, theorem 3.3] the existence is ensured of a PD Q with pdf

$$q(y) = Cp(y)\exp\left(-\frac{1}{2}\sum_{\tau=0}^{T} y_\tau^{\#}\Lambda_\tau y_\tau\right)$$

where $C < \infty$ is a normalization factor which makes $\int q(y)\,dy = 1$ and Λ is defined in a way similar to Ψ in Lemma A1, which minimizes $D(Q\|P)$ over all $Q \in \Omega(R)$, provided that Θ_P is open and R is an inner point of Φ_P. Lemmas A1 and A2 prove the validity of these hypotheses for P as in (1) and any sequence R for which each R_t has a positive definite extension. The minimizing PD Q is unique by Lemma A2, the convexity of $D(Q\|P_\lambda)$ in Q, and the convexity of $\Omega(R)$ [35, sec. 1]. Since

$$\int dy\, p(y)\exp\left(-\frac{1}{2}\sum_{\tau=0}^{T} y_\tau \Lambda_\tau y_\tau\right) = C^{-1} < \infty,$$

$\Lambda \in \Theta_P$, and hence $S_{x_t}^{-1}+\Lambda_t$ is positive definite for every $t =$

$0,1,\cdots,T$. The MDI measure (7) is finite and is obtained by substituting (1) and (6) into (4b).

Proof of Corollary 1: The first equality in (11) follows by definition. Define $F_\tau \triangleq I + S_{x_\tau}\Psi_\tau$, and use $\det F_\tau = \exp(\mathrm{tr}\ln F_\tau)$ and $\partial \ln F_\tau/\partial f_{ij} = F_\tau^{-1}\,\partial F_\tau/\partial f_{ij}$, where f_{ij} are the elements of F_τ, to get the gradient of $d(R;\Psi,\lambda)$, $\Psi \in \delta_{P_\lambda}$, with respect to each element of Ψ. This shows that the equation set $\nabla_\Psi d(R;\Psi,\lambda)|_{\Psi=\Lambda} = 0$ coincides with (8), and hence, from the discussion which follows Theorem 1, it has a unique solution $\Lambda \in \delta_{P_\lambda}$. Furthermore, $d(R;\Psi,\lambda) \in C^1$, the set of functions with continuous first-order derivatives. Since δ_{P_λ} is an open subset of the Euclidean space (see Lemma A1), we have from the corollary in [31, p. 169] that, if $d(R;\Psi,\lambda)$ has a minimum point within δ_{P_λ}, this point must be at $\Psi = \Lambda$. Now since $d(R;\Lambda,\lambda) < \infty$ and $d(R;\Psi,\lambda) \to \infty$ as Ψ approaches the boundaries of δ_{P_λ}, $d(R;\Psi,\lambda)$ must have a minimum in δ_{P_λ}, and by the preceding argument the minimization point is obtained at $\Psi = \Lambda$. The unimodality of $d(R;\Psi,\lambda)$ within δ_{P_λ} follows from the uniqueness of the solution of the gradient equations and the existence of a minimum point within δ_{P_λ}.

Proof of Theorem 2: The proof follows from the global convergence thoerem developed by Sabin and Gray [32], provided that 1) there exists a continuous function, called a descent function, which strictly decreases outside the solution set Γ and does not increase inside Γ; and 2) the algorithm T_R is closed. By Lemma 1 and the continuity of $\nu(R,P_\lambda)$ that follows from (10) and the representation (11), $\nu(R,P_\lambda)$ is a descent function. The closedness of T_R results from Luenberger [31, corollary 2, p. 187] if $\zeta(\cdot)$ is continuous and $\mu(\cdot)$ is closed. The mapping $\zeta(\cdot)$ is continuous if Λ_n is a continuous function of λ_n. This is, however, implied by the continuity of $\nu(R,P_\lambda)$ within δ_{P_λ}, or more directly, by using the implicit function theorem (see the proof of Lemma 2). To show that $\mu(\cdot)$ is closed, let $(P_{\lambda_n},Q_{\lambda_n}) \to (P_\lambda,Q_\lambda)$, $P_{\lambda'_n} \to P_{\lambda'}$, where $P_{\lambda'_n} \in \mu(P_{\lambda_n},Q_{\lambda_n})$. By definition,

$$\mu(P_{\lambda_n},Q_{\lambda_n}) = \left\{ P\colon\ g(P_{\lambda_n},Q_{\lambda_n};P) \ge g\left(P_{\lambda_n},Q_{\lambda_n};P_{\lambda_n}\right)\right\}. \quad \text{(A8)}$$

Due to the continuity of $g(P_{\lambda_n},Q_{\lambda_n};\cdot)$ for $\Lambda_n \in \delta_{P_{\lambda_n}}$, which can be seen from (22), (9), (20), and (21), the set in (A8) becomes

$$\{P\colon\ g(P_\lambda,Q_\lambda;P) \ge g(P_\lambda,Q_\lambda;P_\lambda)\} = \mu(P_\lambda,Q_\lambda) \quad \text{(A9)}$$

when $n \to \infty$. Hence

$$P_{\lambda'} = \lim_{n\to\infty} P_{\lambda'_n} \in \mu(P_\lambda,Q_\lambda) \quad \text{(A10)}$$

and $\mu(\cdot)$ is closed. This completes the proof.

Proof of Lemma 2: By definition, if P_λ is a fixed point of T_R, then Q_λ is optimal for P_λ ($(P_\lambda,Q_\lambda) = \zeta(P_\lambda)$) and P_λ is optimal for Q_λ ($P_\lambda \in \mu(P_\lambda,Q_\lambda)$ a.e. Q_λ). The significance of Q_λ being optimal for P_λ is that $D(Q_{\lambda'}\|P_\lambda)$, $Q_{\lambda'} \in \Omega(R)$ with pdf $q_{\lambda'}(y)$ as in (6), attains its minimum at $\lambda' = \lambda$. Hence, if $D(Q_{\lambda'}\|P_\lambda)$ is differentiable with respect to λ' at $\lambda' = \lambda$, then

$$\nabla_{\lambda'} D(Q_{\lambda'}\|P_\lambda)|_{\lambda'=\lambda} = 0. \quad \text{(A11)}$$

To prove that $D(Q_{\lambda'}\|P_\lambda)$ is differential with respect to λ' we first expand the integrand of the discrimination information meausre into a Taylor series [21, p. 15],

$$\int q_{\lambda'}(y)\ln\frac{q_{\lambda'}(y)}{p_\lambda(y)}\,dy = \frac{1}{2}\int\frac{[r_{\lambda'}(y)-1]^2}{h_{\lambda'}(y)}\,dP_\lambda(y) \triangleq \int f_y(\lambda')\,dP_\lambda(y) \quad \text{(A12)}$$

where $r_{\lambda'}(y) \triangleq q_{\lambda'}(y)/p_\lambda(y)$ and $h_{\lambda'}(y)$ lies between $r_{\lambda'}(y)$ and 1. Since $0 < r_{\lambda'}(y) < \infty$ a.e., we have that $0 < h_{\lambda'}(y) < \infty$ a.e. and hence $f_y(\lambda') > 0$ a.e. For a given $P_{\lambda'}$ and $\{R_t\}$ which have positive definite extensions, Theorem 1 and the discussion which follows it guarantee the existence and uniqueness of $\Lambda(\lambda') \in \delta_{P_{\lambda'}}$ which satisfies (8). Using the implicit function theorem [31, p. 462] it can be shown that $\Lambda(\lambda') \in C^1$. Hence $f_y(\lambda')$ is differentiable with respect to λ'. The Gateaux differential [48, p. 171] of this integrand with respect to λ' is given by

$$\delta f_y(\lambda;u) = \lim_{\alpha\downarrow 0}\left[f_y(\lambda'+\alpha u) - f_y(\lambda')\right]/\alpha. \quad \text{(A13)}$$

The existence of the Gateaux differential of $D(Q_{\lambda'}\|P_\lambda)$ with respect to λ' is now shown using (A13) and Lebesgue's monotone convergence theorem [47, p. 22].

$$\begin{aligned}\delta D(\lambda';u) &\triangleq \lim_{\alpha\downarrow 0}\left[D(Q_{\lambda'+\alpha u}\|P_\lambda) - D(Q_{\lambda'}\|P_\lambda)\right]/\alpha \\ &= \lim_{\alpha\downarrow 0}\int\left[f_y(\lambda'+\alpha u) - f_y(\lambda')\right]/\alpha\,dP_\lambda(y) \\ &= \int\lim_{\alpha\downarrow 0}\left[f_y(\lambda'+\alpha u) - f_y(\lambda')\right]/\alpha\,dP_\lambda(y) \\ &= \int\delta f_y(\lambda';u)\,dP_\lambda(y). \end{aligned} \quad \text{(A14)}$$

The significance of P_λ being optimal for Q_λ is that

$$\nabla_{\lambda'} g(P_\lambda,Q_\lambda;P_{\lambda'})|_{\lambda'=\lambda} = 0 \quad \text{(A15)}$$

where $g(P_\lambda,Q_\lambda;P_{\lambda'})$ is given in (43). This condition, however, is equivalent to

$$\nabla_{\lambda'} D(Q_\lambda\|P_{\lambda'})|_{\lambda'=\lambda} = 0, \quad \text{(A16)}$$

since

$$\nabla_{\lambda'} D(Q_\lambda\|P_{\lambda'})|_{\lambda'=\lambda} = -\nabla_{\lambda'} g(P_\lambda,Q_\lambda;P_{\lambda'})|_{\lambda'=\lambda}$$

as is easy to verify. This, in fact, is a property of the EM algorithm [9, proposition 2.1] whose single iteration is being applied here in calcualting P_λ for a given Q_λ, as we saw in Section II. Combining (A11) and (A16), we have that

$$\nabla_{\lambda',\lambda''} D(Q_{\lambda'}\|P_{\lambda''})|_{\lambda'=\lambda,\lambda''=\lambda} = 0, \quad \text{(A17)}$$

which implies that the directional derivative of the MDI measure $\nu(R;P_\lambda) = D(Q_\lambda\|P_\lambda)$ with repect to λ is zero.

REFERENCES

[1] E. L. Lehmann, *Testing Statistical Hypotheses*. New York: Wiley, 1959.

[2] Y. Ephraim and R. M. Gray, "A unified approach for encoding clean and noisy sources by means of waveform and autoregressive model vector quantization," *IEEE Trans. Inform. Theory*, vol. IT-34, no. 4, pp. 826–834, July 1988.

[3] Y. Ephraim, D. Malah, and B.-H. Juang, "On the application of hidden Markov models for enhancing noisy speech," to appear in *IEEE Trans. Acoust., Speech, Signal Processing*, Dec. 1989.

[4] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[5] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[6] A. B. Poritz, "Hidden Markov models: A guided tour," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Apr. 1988, pp. 7–13.

[7] L. R. Bahl, F. Jelinek, and R. L. Mercer, "A maximum likelihood approach to continuous speech recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-5, no. 2, pp. 179–190, Mar. 1983.

[8] B. H. Juang and L. R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, no. 6, pp. 1404–1413, Dec. 1985.

[9] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic func-

tions of Markov chains," *Ann. Math. Statist.*, vol. 41, no. 1, pp. 164–171, 1970.
[10] L. E. Baum, "An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov processes," *Inequalities*, vol. 3, no 1, pp. 1–8, 1972.
[11] L. A. Liporace, "Maximum likelihood estimation for multivariate observations of Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-28, no. 5, pp. 729–734, Sept. 1982.
[12] A. B. Poritz, "Linear predictive hidden Markov models and the speech signal," in *Proc. Symp. Applications of Hidden Markov Models to Text and Speech*, J. D. Ferguson, Ed., IDA-CRD, Princeton, NJ, 1980, pp. 88–142 (summarized in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Apr. 1982, pp. 1291–1294).
[13] J. D. Ferguson, "Variable duration models for speech," in *Proc. Symp. Applications of Hidden Markov Models to Text and Speech*, J. D. Ferguson, Ed., IDA-CRD, Princeton, NJ, 1980, pp. 143–179.
[14] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. London: Prentice-Hall International, 1982.
[15] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 49–52, Apr. 1986.
[16] P. F. Brown, "The acousting-modeling problem in automatic speech recognition," Ph.D. dissertation, Dep. Comput. Sci., Carnegie–Mellon Univ., Pittsburgh, PA, 1987.
[17] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the *EM* algorithm," *J. Roy. Statist. Soc. B*, vol. 39, pp. 1–38, 1977.
[18] C. F. J. Wu, "On the convergence properties of the EM algorithm," *Ann. Statist.*, vol. 11, no. 1, pp. 95–103, 1983.
[19] Y. Ephraim and L. R. Rabiner, "On the relations between modeling approaches for speech recognition," to appear in *IEEE Trans. Inform. Theory*.
[20] I. Csiszár and G. Tusnady, "Information geometry and alternating minimization procedures," Math. Inst. Hungarian Academy of Sciences, ACAD SCI., Budapest, Preprint 35, 1983.
[21] S. Kullback, *Information Theory and Statistics*. New York: Dover, 1968.
[22] M. Kupperman, "Probabilities of hypotheses and information-statistics in sampling from exponential-class populations," *Ann. Math. Statist.*, vol. 29, pp. 571–575, 1958.
[23] J. E. Shore and R. W. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inform. Theory*, vol. IT-26, no. 1, pp. 26–37, Jan. 1980 (cf. comments and corrections, *IEEE Trans. Inform. Theory*, vol. IT-29, no. 6, pp. 942–943, Nov. 1983).
[24] J. E. Shore, "On a relation between maximum likelihood classification and minimum relative-entropy classification," *IEEE Trans. Inform. Theory*, vol. IT-30, no. 6, pp. 851–854, Nov. 1984.
[25] R. M. Gray, A. H. Gray, Jr., G. Rebolledo, and J. E. Shore, "Rate-distortion speech coding with a minimum discrimination information distortion measure," *IEEE Trans. Inform. Theory*, vol. IT-27, no. 6, pp. 708–721, Nov. 1981.
[26] Y. Ephraim, H. Lev-Ari, and R. M. Gray, "Asymptotic minimum discrimination information measure for asymptotically weakly stationary processes," *IEEE Trans. Inform. Theory*, vol. IT-34, no. 5, pp. 1033–1040, Sept. 1988.
[27] K. Dzhaparidze, *Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series*. New York: Springer-Verlag, 1986.
[28] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and format frequencies," *Electron. Commun. Japan*, vol. 53-A, no. 1, pp. 36–43, 1970.
[29] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. New York: Springer-Verlag, 1976.
[30] R. M. Gray, A. Buzo, A. H. Gray, Jr., and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 4, pp. 367–376, Aug. 1980.
[31] D. G. Luenberger, *Linear and Non-Linear Programming*. Reading, MA: Addison-Wesley, 1984.
[32] M. J. Sabin and R. M. Gray, "Global convergence and empirical consistency of the generalized Lloyd algorithm," *IEEE Trans. Inform. Theory*, vol. IT-32, no. 2, pp. 148–155, Mar. 1986.
[33] J. E. Shore, "Minimum cross-entropy spectral analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, no. 2, pp. 230–237, Apr. 1981.
[34] R. W. Johnson and J. E. Shore, "Minimum cross-entropy spectral analysis of multiple signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, no. 3, pp. 574–582, June 1983.
[35] I. Csiszár, "*I*-divergence geometry of probability distributions and minimization problems," *Ann. Prob.*, vol. 3, no. 1, pp. 146–158, 1975.
[36] P. L. Brockett, A. Charnes, and K. H. Paick, "Computation of minimum cross entropy spectral estimates: An unconstrained dual convex programming method," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 236–242, Mar. 1986.
[37] J. P. Burg, D. G. Luenberger, and D. L. Wenger, "Estimation of structured covariance matrices," *Proc. IEEE*, vol. 70, no. 9, pp. 963–974, Sept. 1982.
[38] A. Dembo, "The relation between maximum likelihood estimation of structured covariance matrices and periodograms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, no. 6, pp. 1661–1662, Dec. 1986.
[39] G. D. Forney, Jr., "The Viterbi algorithm," *Proc. IEEE*, vol. 61, no. 3, pp. 268–278, Mar. 1973.
[40] L. R. Rabiner, J. G. Wilpon, and B.-H. Juang, "A segmental *k*-means training procedure for connected word recognition," *AT&T Tech. J.*, pp. 21–40, May–June 1986.
[41] E. Parzen, "Spectral analysis of asymptitically stationary time series," *Bull. Int. Statist. Inst.*, vol. 39, Livraison 2, pp. 87–102, 1962.
[42] W. A. Gardner, *Introduction to Random Processes*. New York: Macmillan, 1986.
[43] R. M. Gray, *Probability, Random Processes, and Ergodic Properties*. New York: Springer-Verlag, 1988.
[44] R. M. Gray, "Toeplitz and Circulant Matrices: II," Stanford Electron. Lab., Stanford CA: Tech. Rep. 6504-1, Apr. 1977.
[45] S. E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Comput. Speech Language*, vol. 1, pp. 29–45, Nov. 1986.
[46] Y. Ephraim, A. Dembo, and L. R. Rabiner, "Extensions of the minimum discrimination information approach for hidden Markov modeling," unpublished work.
[47] W. Rudin, *Real and Complex Analysis*, 2nd ed. New York: McGraw-Hill, 1974.
[48] D. G. Luenberger, *Optimization by Vector Space Methods*. New York: Wiley, 1969.
[49] J. C. Kiefer, *Introduction to Statistical Inference*. New York: Springer-Verlag, 1987.
[50] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite-state Markov chains," *Ann. Math. Statist.*, vol. 37, pp. 1554–1563, 1966.