A MINIMUM DISCRIMINATION INFORMATION APPROACH FOR HIDDEN MARKOV MODELING

Yariv Ephraim, Amir Dembo, and Lawrence R. Rabiner*

AT&T Bell Laboratories
Speech Research Department
Murray Hill, NJ 07974

*Brown University
Division of Applied Mathematics
Providence, RI 02912

Abstract

A new iterative approach for hidden Markov modeling of information sources which aims at minimizing the discrimination information (or the cross-entropy) between the source and the model is proposed. This approach does not require the commonly used assumption that the source to be modeled is a hidden Markov process. The algorithm is started from the model estimated by the traditional maximum likelihood (ML) approach and alternatively decreases the discrimination information over all probability distributions of the source which agree with the given measurements and all hidden Markov models. The proposed procedure generalizes the Baum algorithm for ML hidden Markov modeling. The procedure is shown to be a descent algorithm for the discrimination information measure and its local convergence is proved.

1. Introduction

Commonly used approaches (e.g., see [1]-[2]) for hidden Markov modeling of information sources assume that the observations were generated by some hidden Markov source, and attempt to find a maximum likelihood (ML) [1] or a maximum mutual information (MMI) [2] estimate of the parameters of that source. This assumption is, however, not necessarily true, especially for speech signals for which hidden Markov models (HMM's) have been recently extensively applied.

We propose an alternative approach for doing the modeling in which the model and the observations are matched in an information theoretic way. We do not assume that the *true* probability distribution (PD) of the source to be modeled is that of a hidden Markov source or has any other explicitly given form, as this PD is unknown. The idea here is first to find a PD for the source which agrees with the given measurements and is optimal in the sense of minimizing the discrimination information** with respect to the HMM. Then, the resulting minimum discrimination information (MDI) measure, which depends on the given observations and the model's parameters, is minimized over all HMM's.

Unfortunately, in the case of hidden Markov modeling the resulting MDI measure cannot be made explicit and it is implicitly dependent on the Lagrange multipliers corresponding to the measurements. We therefore have designed an iterative descent algorithm for implementing the MDI modeling. We start from the hidden Markov model estimated from the ML approach, and alternatively decrease the discrimination information over the PD's of the source and the model.

*This work was done while A. Dembo was with AT&T Bell Laboratories, Communications Analysis Research Department.

**The discrimination information is also known as Kullback-Leibler number, cross-entropy, relative entropy, I-divergence or directed divergence.

For a given HMM, we minimize the discrimination information over all PD's for the source which are consistent with the measurements. Then, we fix the PD of the source and estimate an HMM which decreases the discrimination information with respect to the given PD. Thus, each iteration produces a new HMM for the source with a lower (or at least the same) discrimination information for the given measurements.

In particular, we consider the class of HMM's for which the output process from each state is a zero mean Gaussian process. 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 recognition applications. These models will be referred to as zero mean, Gaussian, AR HMM's.

We show that given an HMM, the estimation of the PD of the source, which agrees with the given measurements, can be formulated as a minimization problem, which can be implemented by any standard optimization procedure in the Euclidean space. In addition, given a PD for the source, the estimation of the model's parameters can be done by a procedure similar to that of Baum [1], using the Forward-Backward formulas. Thus, despite the rather more complicated modeling approach used here as compared with the ML approach, the efficiency of the original Baum algorithm is maintained; However, an additional optimization procedure which compensates for the existing "mismatch" between the measurements and the model, is performed.

The global convergence of procedures for alternating optimization of the discrimination information measure has been studied by Csiszar and Tusnady [3]. They gave geometric conditions for convergence and proved that these conditions are satisfied if both classes of PD's of the source and model are convex. In our case, however, the set of HMM's is not convex and the verification of the geometric conditions for convergence is not straightforward. We therefore prove local convergence using a variant of the classic convergence theorem of Luenberger [4].

Proofs of theorems, lemmas, and corollaries are not given here and can be found in [5].

2. Descent algorithm for MDI modeling

2.1 Problem Formulation

Let $\{y_0, y_1, \ldots, y_T\}$ be a set of observations, $y_t \in R^N$, where, R^N is the N-dimensional Euclidean space. Assume that each observation y_t has zero mean and that it is characterized by a set of covariance samples given by

$$R_t(i, j) = E_{Q^*}\{y_t(i)\, y_t(j)\}, \qquad i, j \in B \tag{1}$$

where B is any symmetric band of the original covariance of y_t and Q^* is the true PD of $\{y_0, y_1, \ldots, y_T\}$. We assume that the given covariance R_t at each time t is consistent with some $N \times N$ valid covariance matrix called an extension of R_t.

CH2396-0/87/0000-0025 $1.00

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, and $p_\lambda(z)$ be the corresponding pdf. $\lambda \triangleq (\pi, A, S)$ is the parameter set of the HMM, where, $\pi \triangleq (\pi_1, \pi_2, \cdots, \pi_M)$, is the vector of initial probabilities; $A \triangleq \{a_{\alpha,\beta}\}$, $\alpha, \beta=1, \cdots, M$, is the transition matrix; and $S \triangleq \{S_\beta, \beta=1, \ldots, M\}$ is the set of positive definite covariance matrices of the output processes from the different states.

$$p_\lambda(z) = \sum_x \prod_{t=0}^{T} a_{x_{t-1}x_t} b(z_t \mid x_t), \tag{2}$$

where, $b(z_t \mid x_t)$ is the output pdf on R^N corresponding to the state x_t,

$$b(y_t \mid x_t = \beta) = \frac{\exp(-\frac{1}{2} y_t^{\#} S_\beta^{-1} y_t)}{(2\pi)^{N/2} \det^{1/2}(S_\beta)}, \qquad \beta=1, 2, \ldots, M;$$

$a_{x_{t-1}x_t}$ is the transition probability from the state x_{t-1} at time $t-1$ to the state x_t at time t, and $x_t \in \{1, 2, \ldots, M\}$ for every $t=0, 1, \ldots, T$; $a_{x_{-1}x_0} \triangleq \pi_{x_0}$ is the probability of the initial state x_0; $z \triangleq (z_0, z_1, \ldots, z_T)^{\#}$ with # being vector transpose; and the summation in (2) is taken over all possible sequences of $x \triangleq \{x_0, x_1, \ldots, x_T\}$.

Let $\Omega(R)$ be the set of all PD's Q which satisfy the constraints (1), where, $R \triangleq \{R_t, t=0, \ldots, T\}$. The modeling problem is that of finding the parameter set $\lambda = (\pi, A, S)$ which minimizes the MDI measure given by

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

where, $D(Q \parallel P_\lambda)$ is the discrimination information between Q and P_λ. The discrimination information between two PD's Q and P, with pdf's q and p, respectively, can be evaluated as

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

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

2.2 Estimation of the source PD

The definition of the MDI measure (3) incorporates an infimum rather than a minimum, since the minimum may not exist. The following theorem, however, provides conditions for the existence of a PD which minimizes the discrimination information with respect to a zero mean Gaussian HMM over all PD's which agree with the given second order statistics of the source. The theorem, and its proof, are a straightforward extension of the results developed by Csiszar [6], and by Gray *et al* [7] for the case where the model is a single Gaussian process.

Theorem 1: Let P_λ be a zero mean Gaussian HMM as in (2), $R \triangleq \{R_t, t=0, \ldots, T\}$ be the sequence of given covariance matrices for a zero mean source, and $\Omega(R)$ be the set of all PD's Q which satisfy (1).

(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 $v(R, P_\lambda) = \infty$.

(b) If each R_t has any positive definite extension, then there exists a unique sequence of matrices $\Lambda \triangleq \{\Lambda_t, t=0, \ldots, T\}$, where Λ_t is symmetric and vanishes outside the band B, such that $S_\beta^{-1} + \Lambda_t$ is positive definite for every $t=0, \ldots, T$ and $\beta=1, \ldots, M$, and the pdf

$$q_\lambda(y) = C\, p_\lambda(y) \exp\left[-\frac{1}{2} \sum_{\tau=0}^{T} y_\tau^{\#} \Lambda_\tau y_\tau\right] \tag{5}$$

yields the MDI measure given by

$$v(R, P_\lambda) = -\ln\left[\sum_x \prod_{\tau=0}^{T} a_{x_{\tau-1}x_\tau} \det^{-1/2}(I + S_{x_\tau}\Lambda_\tau)\right] - \frac{1}{2} \operatorname{tr} \sum_\tau (R_\tau \Lambda_\tau). \tag{6}$$

C is a finite normalization factor. The PD Q_λ which corresponds to q_λ is called the MDI PD with respect to P_λ. In this case, the infimum in (3) is a minimum and $v(R, P_\lambda)=D(Q_\lambda \| P_\lambda)$. □

The set of Lagrange multiplier matrices Λ can be obtained from the unique solution of the following equation set which must be satisfied *within* the band B.

$$R_t \triangleq E_{Q_\lambda}\{y_t y_t^{\#}\} = \sum_{\alpha,\beta=1}^{M} q_t(\alpha, \beta)(S_\beta^{-1} + \Lambda_t)^{-1} \tag{7}$$

where,

$$q_t(\alpha, \beta) \triangleq \frac{\sum_{\left\{x:\ x_{t-1}=\alpha,\ x_t=\beta\right\}} \prod_{\tau=0}^{T} a_{x_{\tau-1}x_\tau} \det^{-1/2}(I + S_{x_\tau}\Lambda_\tau)}{\sum_{\alpha,\beta=1}^{M} \sum_{\left\{x:\ x_{t-1}=\alpha,\ x_t=\beta\right\}} \prod_{\tau=0}^{T} a_{x_{\tau-1}x_\tau} \det^{-1/2}(I + S_{x_\tau}\Lambda_\tau)} \tag{8}$$

These equations, however, are difficult to solve in any straightforward manner. The following corollary of Theorem 1 provides an alternative way to evaluate the MDI measure by replacing the algebraic problem in (7) by a minimization problem which can be iteratively solved by any standard minimization procedure.

Corollary 1: Let R and P_λ be as in Theorem 1. Let $\Psi = \{\Psi_t, t=0, \ldots, T\}$ be a sequence of symmetric matrices which vanish outside the band B. Define

$$d(R; \Psi, \lambda) = \ln\left[\sum_x \prod_{\tau=0}^{T} a_{x_{\tau-1}x_\tau} \det^{-1/2}(I + S_{x_\tau}\Psi_\tau)\right] + \frac{1}{2} \operatorname{tr} \sum_\tau (R_\tau \Psi_\tau). \tag{9}$$

If each R_t has a positive definite extension, then

$$v(R, P_\lambda) = -d(R; \Lambda, \lambda) = -\min_\Psi d(R; \Psi, \lambda) \quad \square \tag{10}$$

The MDI measure (6) cannot be made explicit in terms of the given measurements R and the parameters of the HMM λ. Hence, MDI hidden Markov modeling cannot be implemented as a direct minimization of the MDI measure over all HMM's. The MDI modeling can, however, be iteratively performed when starting from some initial HMM. Each iteration consists of first estimating the MDI PD with respect to the given HMM, as outlined above, and then improving the model by decreasing the discrimination information with respect to the estimated MDI PD. We now show how a new HMM is estimated given an MDI PD with respect to the old HMM.

2.3 Hidden Markov modeling

Suppose that Q_λ, the MDI PD with respect to P_λ, has been estimated. Let λ' be the parameter set of the new HMM to be estimated. Since $D(Q_\lambda \| P_{\lambda'}) = \int q_\lambda(y) \ln(q_\lambda(y)/p_{\lambda'}(y))\, dy$, and Q_λ is given, the minimization of $D(Q_\lambda \| P_{\lambda'})$ over all $P_{\lambda'}$s is equivalent to maximizing $\phi(\lambda') \triangleq \int q_\lambda(y) \ln p_{\lambda'}(y)\, dy$ over

all $P_{\lambda'}$'s. Let $p_\lambda(y) \triangleq \sum_x p_\lambda(x, y)$, where $p_\lambda(x, y)$ is the joint pdf of x and y as given in (2). Using Jensen's inequality we have that

$$\begin{aligned}\phi(\lambda') - \phi(\lambda) &= \int q_\lambda(y) \ln \sum_x p_{\lambda'}(x, y)/\sum_x p_\lambda(x, y)\, dy \\ &= \int q_\lambda(y) \ln \sum_x \frac{p_\lambda(x, y)}{p_\lambda(y)} \frac{p_{\lambda'}(x, y)}{p_\lambda(x, y)}\, dy \\ &\geq \sum_x \int \frac{q_\lambda(y)}{p_\lambda(y)} p_\lambda(x, y) \ln \frac{p_{\lambda'}(x, y)}{p_\lambda(x, y)}\, dy, \quad (11)\end{aligned}$$

where equality holds if and only if $p_{\lambda'}(x, y) = p_\lambda(x, y)$ almost everywhere with respect to Q_λ (a.e. Q_λ). Hence, the value of $\phi(\lambda')$ can be increased by

$$\max_{\lambda'} \left\{ \sum_x \int [q_\lambda(y)/p_\lambda(y)]\, p_\lambda(x, y) \ln p_{\lambda'}(x, y)\, dy \right\}. \quad (12)$$

On substituting $q_\lambda(y)/p_\lambda(y)$ from (5) into (12), and denoting

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

we arrive at the following maximization problem

$$\max_{\lambda'} \sum_x \int q_\lambda(x, z) \ln p_{\lambda'}(x, z)\, dz. \quad (14)$$

The function to be maximized in (14) plays here the same role as the auxiliary function proposed by Baum in the ML estimation of the parameters of HMM's [1]. There the problem is

$$\max_{\lambda'} \sum_x p_\lambda(x, y) \ln p_{\lambda'}(x, y)\,, \quad (15)$$

where y are the observations from the source. Comparing (14) and (15) shows that the MDI and the ML hidden Markov modeling approaches result in the same model estimate if and only if

$$q_\lambda(x, z) = p_\lambda(x, z)\, \delta(z - y), \quad (16)$$

where $\delta(\cdot)$ is the Dirac function. This happens when the source producing the observations is the HMM itself whose parameters are being estimated, since then the MDI PD is the PD of the model and the resulting MDI is zero.

On substituting (2) into (14) we obtain

$$\begin{aligned}\max_{\lambda'} \Bigg\{ &\sum_{\beta=1}^{M} \ln \pi'_\beta \sum_{\{x:\, x_0=\beta\}} \int q_\lambda(x, y)\, dy \\ &+ \sum_{\alpha,\, \beta=1}^{M} \ln a'_{\alpha\beta} \sum_{t=1}^{T} \sum_{\left\{x:\ \substack{x_{t-1}=\alpha \\ x_t=\beta}\right\}} \int q_\lambda(x, y)\, dy \\ &+ \frac{1}{2} \sum_{\beta=1}^{M} \ln \det S'^{-1}_\beta \sum_{t=0}^{T} \sum_{\{x:\, x_t=\beta\}} \int q_\lambda(x, y)\, dy \\ &- \frac{1}{2} \operatorname{tr}\left\{ \sum_{\beta=1}^{M} S'^{-1}_\beta \sum_{t=0}^{T} \sum_{\{x:\, x_t=\beta\}} \int q_\lambda(x, y)\, y_t y_t^{\#}\, dy \right\} \Bigg\}. \quad (17)\end{aligned}$$

Note from (2), (8) and (13) that

$$\sum_{\left\{x:\ \substack{x_{t-1}=\alpha \\ x_t=\beta}\right\}} \int q_\lambda(x, y)\, dy = q_t(\alpha, \beta) \quad (18)$$

$$\sum_{\{x:\, x_t=\beta\}} \int q_\lambda(x, y) y_t y_t^{\#}\, dy = (S_\beta^{-1} + \Lambda_t)^{-1} \sum_{\alpha=1}^{M} q_t(\alpha, \beta)$$

$$\triangleq R_t(\beta). \quad (19)$$

Hence, using (18) and (19) we can rewrite (17) as

$$\begin{aligned}\max_{\lambda'} \Bigg\{ &\sum_{\beta=1}^{M} \ln \pi'_\beta\, q_0(\alpha, \beta) + \sum_{\alpha,\, \beta=1}^{M} \ln a'_{\alpha\beta} \sum_{t=1}^{T} q_t(\alpha, \beta) \quad (20) \\ &- \frac{1}{2} \sum_{\beta=1}^{M} \left[\operatorname{tr}\left(S'^{-1}_\beta \sum_{t=0}^{T} R_t(\beta) \right) - \ln \det S'^{-1}_\beta \sum_{t=0}^{T} \sum_{\alpha=1}^{M} q_t(\alpha, \beta) \right] \Bigg\}\end{aligned}$$

The maximization of (20) over π'_β results in

$$\pi'_\beta = q_0(\alpha, \beta), \qquad \beta = 1, \ldots, M. \quad (21)$$

Similarly, the maximization of (20) over $a'_{\alpha\beta}$ results in

$$a'_{\alpha\beta} = \frac{\sum_{t=1}^{T} q_t(\alpha, \beta)}{\sum_{\beta=1}^{M} \sum_{t=1}^{T} q_t(\alpha, \beta)}, \qquad \alpha, \beta = 1, \ldots, M, \quad (22)$$

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 $a'_{\alpha,\beta}$ can be arbitrarily chosen since its value does not affect (20). The maximization of (20) over S'_β is considered for zero mean Gaussian AR HMM's. Suppose first that $\sum_{t=0}^{T} \sum_{\alpha=1}^{M} q_t(\alpha, \beta) > 0$. The problem then becomes

$$\min_{S'_\beta} \left\{ \operatorname{tr}\left[R(\beta) S'^{-1}_\beta \right] - \ln \det S'^{-1}_\beta \right\}, \quad (23)$$

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

$$R(\beta) \triangleq \frac{\sum_{t=0}^{T} R_t(\beta)}{\sum_{t=0}^{T} \sum_{\alpha=1}^{M} q_t(\alpha, \beta)}, \qquad \beta = 1, \ldots, M. \quad (24)$$

This is exactly the problem arisen in ML estimation of structured covariance matrices given a measured covariance matrix [8]. In our case we are interested in estimating the covariance matrix S'_β of an r-th order AR process given $R(\beta)$. S'_β is given by $S'_\beta = \sigma'^2_\beta (L^{\#}_\beta L_\beta)^{-1}$, where, σ_β is a gain constant, L_β is a lower triangular matrix given by

$$L_\beta = \{l_{ij}, \qquad i, j = 0, 1, \ldots, N-1\}$$

$$L_\beta \triangleq \begin{cases} f_\beta(i-j) & 0 \leq i-j \leq r \\ 0 & \text{otherwise,} \end{cases}$$

$f_\beta(0) = 1$, and $f_\beta(i)$, $i = 1, \ldots, r$, are the coefficients of the AR process. Since $R(\beta)$ is positive definite, the set of all AR matrices S'_β is a closed subset of the set of positive semidefinite matrices, and the set of all inverses of AR matrices is convex, there exists a unique positive definite matrix S'_β which minimizes (23) [8]. Since $\det L_\beta = 1$, the coefficients $f_\beta(\cdot)$ are obtained from the minimization of $\operatorname{tr}(R(\beta) L^{\#}_\beta L_\beta)$. From [7, Corollary 2], this is done by minimizing the quadratic form

$$\varepsilon \triangleq \sum_{n=0}^{r} \sum_{m=0}^{r} f_\beta(n) f_\beta(m) \frac{1}{N} \sum_{k=\max(n,\, m)}^{N-1} r_\beta(k-n,\, k-m) \quad (25)$$

where $r_\beta(\cdot, \cdot)$ are the elements of $R(\beta)$. This results in a linear set of equations similar to that obtained in the "covariance method" for linear prediction analysis. The gain which

minimizes (23) is easily shown to be $\sigma_{\beta}^{2}=\varepsilon$.

If $\sum_{t=0}^{T}\sum_{\alpha=1}^{M} q_t(\alpha, \beta)$ in (20) equals zero, then $q_t(\alpha, \beta) = 0$, and therefore from (19) $R_t(\beta) = 0$. Hence, S'_{β} can be arbitrarily chosen since its value does not affect (20).

We now show how $q_t(\alpha, \beta)$ can be efficiently calculated using the forward-backward formulas. Define

$$F_t(\alpha) \triangleq \sum_{\substack{\{x_0, \ldots, x_{t-1}\} \\ x_t = \alpha}} \prod_{\tau=0}^{t} a_{x_{\tau-1}x_\tau} \det^{-1/2}(I + S_{x_\tau}\Lambda_\tau) \tag{26}$$

$$B_t(\beta) \triangleq \sum_{\substack{\{x_{t+1}, \ldots, x_T\} \\ x_t = \beta}} \prod_{\tau=t+1}^{T} a_{x_{\tau-1}x_\tau} \det^{-1/2}(I + S_{x_\tau}\Lambda_\tau) \tag{27}$$

$$F_{-1}(\alpha) = B_T(\beta) = 1$$

and note that

$$F_t(\alpha) = \sum_{\gamma=1}^{M} F_{t-1}(\gamma)\, a_{\gamma\alpha} \det^{-1/2}(I + S_\alpha\Lambda_t) \tag{28}$$

$$B_t(\beta) = \sum_{\gamma=1}^{M} B_{t+1}(\gamma)\, a_{\beta\gamma} \det^{-1/2}(I + S_\gamma\Lambda_{t+1}) \tag{29}$$

From (8) we have

$$q_t(\alpha, \beta) = \frac{F_{t-1}(\alpha)\, B_t(\beta)\, a_{\alpha\beta} \det^{-1/2}(I + S_\beta\Lambda_t)}{\sum_{\alpha=1}^{M}\sum_{\beta=1}^{M} F_{t-1}(\alpha)B_t(\beta)\, a_{\alpha\beta}\det^{-1/2}(I + S_\beta\Lambda_t)} \tag{30}$$

3. Convergence analysis

Suppose that each given 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.

$$D(Q \parallel P_\lambda) \geq D(Q_\lambda \parallel P_\lambda) = \nu(R, P_\lambda), \tag{31}$$

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

$$D(Q_\lambda \parallel P_\lambda) \geq D(Q_\lambda \parallel P_{\lambda'}). \tag{32}$$

Since

$$D(Q_\lambda \parallel P_\lambda) - D(Q_\lambda \parallel P_{\lambda'}) = \int dy\, q_\lambda(y) \ln p_{\lambda'}(y)/p_\lambda(y), \tag{33}$$

equality in (32) holds if and only if $p_\lambda = p_{\lambda'}$ a.e. Q_λ. Combining equations (31) and (32) we obtain the following inequality.

$$\begin{aligned} \nu(R, P_\lambda) = D(Q_\lambda \parallel P_\lambda) &\geq D(Q_\lambda \parallel P_{\lambda'}) \\ &\geq D(Q_{\lambda'} \parallel P_{\lambda'}) = \nu(R, P_{\lambda'}). \end{aligned} \tag{34}$$

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 (34) we have that $D(Q_\lambda \parallel P_\lambda) = D(Q_\lambda \parallel P_{\lambda'}) = D(Q_{\lambda'} \parallel P_{\lambda'})$, which by (31) and (32) 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 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

$$\nu(R, P_\lambda) \geq \nu(R, P_{\lambda'}) \tag{35}$$

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 each HMM is a continuous function of λ (see (2)), and the corresponding MDI PD is a continuous function of λ and Λ (see (5)), convergence can be equivalently considered in terms of either $(P_{\lambda_n}, Q_{\lambda_n})$ or $(\lambda_n, \Lambda_{\lambda_n})$.

Let

$$\zeta(P_{\lambda_n}): P_{\lambda_n} \to (P_{\lambda_n}, Q_{\lambda_n}) \tag{36}$$

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

$$\mu(P_{\lambda_n}, Q_{\lambda_n}): (P_{\lambda_n}, Q_{\lambda_n}) \to P_{\lambda_{n+1}} \tag{37}$$

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 following function

$$g(\Lambda_{\lambda_n}, \lambda_{n+1}) \triangleq \sum_x \int q_{\lambda_n}(x, z) \ln p_{\lambda_{n+1}}(x, z)\, dz \tag{38}$$

over all λ_{n+1}, as is required in (14). As we have shown in Section 2, this maximization reduces the value of the MDI measure with respect to Q_{λ_n}. The algorithm is now defined as the composition of these two mappings as follows.

$$T_R(P_{\lambda_n}): P_{\lambda_n} \to P_{\lambda_{n+1}}, \quad \text{and} \quad T_R(P_{\lambda_n}) = \mu(\zeta(P_{\lambda_n})). \tag{39}$$

We have the following theorem.

Theorem 2: Assume that each given 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 $\Gamma \triangleq \{P_\lambda: p_\lambda = T_R(p_\lambda)$ 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

(i) Each accumulation point P_{λ^*} of P_{λ_n} is a fixed point, i.e., $P_{\lambda^*} \in \Gamma$.

(ii) $\rho(P_{\lambda_n}, \Gamma) \to 0$, where ρ is the usual distance in the Euclidean space.

(iii) $\nu(R, P_{\lambda_n}) \to \nu(R, P_{\lambda^*})$. □

It can also be shown that any fixed point of T_R is a stationary point of $\nu(R, P_\lambda)$ [5].

References

[1] L. E. Baum, "An inequality and associated maximization techniques in statistical estimation of probabilistic functions of Markov processes," *Inequalities*, vol. 3, pp. 1-8, 1972.

[2] 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," *ICASSP' 86, pp. 49-52.*

[3] I. Csiszar and G. Tusnady, "Information geometry and alternating minimization procedure," Preprint no. 35, 1983, Math-Inst. Hungar, ACAD SCI., Budapest.

[4] 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, March 1986.

[5] Y. Ephraim, A. Dembo, and L. R. Rabiner, "A minimum discrimination information approach for hidden Markov modeling," submitted for publication.

[6] I. Csiszar, "I-Divergence geometry of probability distributions and minimization problems," *Ann. Prob.*, vol. 3, pp. 146-158, 1975.

[7] R. M. Gray, A. H. Gray, Jr., G. Rebolledo, and J. E. Shore, "Rate distortion speech coding with a minimum discrimination information measure," *IEEE Trans. Inform. Theory*, vol. IT-27, No. 6, Nov. 1981.

[8] A. Q. Nguyen, "On the uniqueness of the maximum likelihood estimate of structured covariance matrices," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 1249-1251, Dec. 1984.