A Bayesian Classification Approach with Application to Speech Recognition

Neri Merhav, *Member, IEEE,* and Yariv Ephraim, *Senior Member, IEEE*

***Abstract*—A Bayesian approach to classification of parametric information sources whose statistics are not explicitly given is studied and applied to recognition of speech signals based upon Markov modeling. A classifier based on generalized likelihood ratios, which depends only on the available training and testing data, is developed and shown to be optimal in the sense of achieving the highest asymptotic exponential rate of decay of the error probability. The proposed approach is compared to the standard classification approach used in speech recognition, in which the parameters for the sources are first estimated from the given training data, and then the maximum *a posteriori* (MAP) decision rule is applied using the estimated statistics.**

I. Introduction

THE problem of classifying parametric information sources, whose statistics are not explicitly given but rather available through training data, is of considerable importance in speech recognition applications (see, e.g., [1]) and in digital communications (see, e.g., [2], [3]). In this paper, a Bayesian classification approach is studied and compared to the standard approach commonly used in speech recognition.

Let $\lambda_i \in \Lambda$ be the parameter set of the ith source, $1 \le i \le M$, where $\Lambda \subset \boldsymbol{R}^N$ is the parameter space and $\boldsymbol{R}^N$ is the Nth dimensional Euclidean space. Let $\boldsymbol{\lambda} \triangleq (\lambda_1, \cdots, \lambda_M) \in \Lambda^M$, where Λ^M denotes the Mth Cartesian power of Λ. We treat the unknown $\boldsymbol{\lambda}$ as a random variable with prior probability density function (pdf) $p(\boldsymbol{\lambda})$. Let $y_i = \{y_{i,1}, \cdots, y_{i,n_i}\}$ be a training sequence from the ith source where $y_{i,t} \in \boldsymbol{U}$, $t = 1, \cdots, n_i$, and $\boldsymbol{U}$ is the alphabet of the source. The alphabet $\boldsymbol{U}$ may either be a finite set, the real line $\boldsymbol{R}$, or $\boldsymbol{R}^L$. Let $Y \triangleq (y_1, \cdots, y_M)$ be the set of given training sequences from the M sources. Let $x = \{x_1, \cdots, x_n\}$, where $x_t \in \boldsymbol{U}$, $t = 1, \cdots, n$, be a test sequence generated by one of the M sources, henceforth referred to as the active source. The index of the active source is unknown and considered a discrete random variable m taking values on $\{1, 2, \cdots, M\}$. Given x and Y, the classification problem is that of identifying the active source. A decision rule $\Omega(Y)$, derived from the training data Y, is a partition of the space of all possible test sequences $\boldsymbol{U}^n$ into M disjoint regions $\Omega_1(Y), \cdots, \Omega_M(Y)$ whose union equals $\boldsymbol{U}^n$. The test sequence x is classified as being generated by the ith source if $x \in \Omega_i(Y)$.

The conditional probability of error $P_\Omega(e \mid Y)$ associated with a decision rule $\Omega = \Omega(Y)$ is given by

$$P_\Omega(e \mid Y) = \sum_{i=1}^{M} p(H_i) \int_{\Omega_i^c(Y)} p(x \mid Y, H_i)\, dx \qquad (1.1)$$

where $\Omega_i^c(Y)$ is the complement set of $\Omega_i(Y)$, $p(H_i)$ is the prior probability of the ith source, namely, the probability that $m = i$, and $p(x \mid Y, H_i)$ is the conditional pdf of x given Y and that x was generated by the ith source H_i. The pdf $p(x \mid Y, H_i)$ is calculated from the joint pdf $p(x, Y, \boldsymbol{\lambda}, H_i)$ given by

$$p(x, Y, \boldsymbol{\lambda}, H_i) = p(x \mid Y, \boldsymbol{\lambda}, H_i) p(Y \mid \boldsymbol{\lambda}, H_i) p(\boldsymbol{\lambda} \mid H_i) p(H_i) \qquad (1.2)$$

using the following assumptions:

1) The test sequence and the training data are conditionally independent given the parameter sets and the active source, i.e.,

$$p(x \mid Y, \boldsymbol{\lambda}, H_i) = p(x \mid \lambda_i, H_i). \qquad (1.3)$$

2) The training sequences are conditionally independent of the active source given the parameter sets, and the conditional pdf of the training data has a product form

$$p(Y \mid \boldsymbol{\lambda}, H_i) = p(Y \mid \boldsymbol{\lambda}) = \prod_{i=1}^{M} p(y_i \mid \lambda_i). \qquad (1.4)$$

3) The parameter sets are independent and identically distributed (i.i.d.), and they are independent of the active source, i.e.,

$$p(\boldsymbol{\lambda} \mid H_i) = p(\boldsymbol{\lambda}) = \prod_{i=1}^{M} p(\lambda_i). \qquad (1.5)$$

4) The prior probability of the active source is assumed uniform, i.e.,

$$p(H_i) = \frac{1}{M}. \qquad (1.6)$$

Since we deal with parametric sources, the conditional pdf's $\{p(y_i \mid \lambda_i)\}$ and $p(x \mid \lambda_i, H_i)$ have an assumed known

Manuscript received December 27, 1989; revised September 12, 1990.

The authors are with the Speech Research Department, AT&T Bell Laboratories, Murray Hill, NJ 07974.

IEEE Log Number 9101858.

parametric form. From (1.3)–(1.6) we get

$$p(x|Y, H_i) = \frac{p(x, Y, H_i)}{p(Y, H_i)}$$

$$= \frac{\int_{\Lambda^M} p(x, Y, \lambda, H_i)\, d\lambda}{\int_{U^n}\int_{\Lambda^M} p(x, Y, \lambda, H_i)\, d\lambda\, dx}$$

$$= \frac{\int_{\Lambda} p(x|\lambda_i, H_i)p(y_i|\lambda_i)p(\lambda_i)\, d\lambda_i}{\int_{\Lambda} p(y_i|\lambda_i)p(\lambda_i)\, d\lambda_i}$$

$$= \frac{\int_{\Lambda} p(x|y_i, \lambda_i, H_i)p(y_i|\lambda_i, H_i)p(\lambda_i|H_i)\, d\lambda_i}{\int_{\Lambda} p(y_i|\lambda_i, H_i)p(\lambda_i|H_i)\, d\lambda_i}$$

$$= \frac{p(x, y_i|H_i)}{p(y_i|H_i)} = p(x|y_i, H_i). \tag{1.7}$$

On substituting (1.7) into (1.1) we obtain

$$P_\Omega(e|Y) = \frac{1}{M}\sum_{i=1}^{M}\int_{\Omega_i^c(Y)} p(x|y_i, H_i)\, dx. \tag{1.8}$$

The conditional probability of error $P_\Omega(e|Y)$ in (1.8) is minimized by the decision rule $\Omega^*(Y) \triangleq \{\Omega_i^*(Y)\}_{i=1}^M$ given by

$$\Omega_i^*(Y) = \{x: p(x|y_i, H_i) \geq p(x|y_j, H_j), \quad \forall j \neq i\} \tag{1.9}$$

where ties are broken arbitrarily. Following [4] this decision rule will be referred to as the Bayesian decision rule.

Note that if $\{\lambda_i\}$ were known, then the probability of error associated with a decision rule $\Omega \triangleq \{\Omega_i\}_{i=1}^M$ would be given by

$$P_\Omega(e|\lambda) = \frac{1}{M}\sum_{i=1}^{M}\int_{\Omega_i^c} p(x|\lambda_i, H_i)\, dx \tag{1.10}$$

where the minimizing decision rule is given by the standard MAP decision rule

$$\Omega_i^*(\lambda) = \{x: p(x|\lambda_i, H_i) \geq p(x|\lambda_j, H_j), \quad \forall j \neq i\}. \tag{1.11}$$

The Bayesian decision rule (1.9) uses the implicit knowledge about $\{\lambda_i\}_{i=1}^M$ given by the training sequences $\{y_i\}_{i=1}^M$. Following Nádas [4], the decision rule (1.9) can be interpreted as the MAP decision rule applied to conditional mean estimates of $p(x|\lambda_i, H_i)\}$. This follows from (1.3)–(1.6)

$$p(x|y_i, H_i) = \int_{\Lambda} p(x|\lambda_i, H_i)p(\lambda_i|Y)\, d\lambda_i$$

$$= E\{p(x|\lambda_i, H_i)|Y\} \tag{1.12}$$

where the expectation is taken over the ensemble of the parameter set λ_i. The Bayesian approach is different from the plug in (PI) approach, widely used in speech recognition, where $p(x|\lambda_i, H_i)$ is estimated by $p(x|\hat{\lambda}_i, H_i)$, where $\hat{\lambda}_i$ is an estimate of λ_i obtained from Y. The commonly used approaches for estimating $\{\lambda_i\}$, namely, the maximum likelihood (ML) approach [5]–[7], the minimum discrimination information (MDI) approach [8], [9], and the maximum mutual information (MMI) approach [10], [11], are summarized in [12].

The conditional pdf's $\{p(x|y_i, H_i)\}$ depend on the priors $\{p(\lambda_i)\}$ which normally are unavailable, and the integrals involved with the evaluation of these pdf's in (1.7) are not trivial. Hence, we develop an asymptotically optimal version of $\Omega^*(Y)$ which requires neither the knowledge of the priors $\{p(\lambda_i)\}$ nor the calculation of integrals. The proposed decision rule, henceforth referred to as the approximate Bayesian (AB) decision rule, is given by

$$\max_{1 \leq i \leq M} \frac{\max_{\lambda_i}[p(x|\lambda_i, H_i)p(y_i|\lambda_i)]}{\max_{\lambda_i} p(y_i|\lambda_i)}. \tag{1.13}$$

It is shown that if the probability of error associated with the Bayesian decision rule (1.9) decays exponentially as the number of observations n tends to infinity, then the error probability, associated with the AB decision rule also decays exponentially with the same asymptotic rate.

Test statistics similar to (1.13) were developed in [13] and [14] under the Neyman–Pearson criterion for unifilar Markov sources, namely, Markov sources for which the state sequence can be uniquely obtained from the observation sequence. The AB decision rule is generalized to classification of a noisy signals, given training data from the clean sources and the noise process. The AB decision rule and the PI decision rule $\Omega^*(\hat{\lambda}) = \{\Omega_i^*(\hat{\lambda})\}_{i=1}^M$ given by

$$\Omega_i^*(\hat{\lambda}) = \{x: p(x|\hat{\lambda}_i, H_i) \geq p(x|\hat{\lambda}_j, H_j)\} \tag{1.14}$$

where

$$\hat{\lambda}_i = \operatorname*{argmax}_{\lambda_i} p(y_i|\lambda_i) \tag{1.15}$$

are examined and compared in classification of Markov sources [15], [16] or hidden Markov models (HMM's), and in speech recognition based upon hidden Markov modeling of the acoustic signals.

The outline of the paper is as follows. In Section II the main theorem is stated and proved. In Section III this theorem and several extensions are discussed. In Section IV we provided simulation results on classification of computer generated Markov sources and speech signals. Finally, in Section V we summarize our conclusions.

II. Asymptotic Optimality of the AB Decision Rule

Let $\Lambda_n \subset \Lambda$ be a finite grid of parameter values. We make the following assumptions.

A1) Λ is a bounded set.

A2) The grid Λ_n becomes dense in Λ as n tends to infinity, i.e., for the Euclidean metric d: $\Lambda \times \Lambda \to \boldsymbol{R}^+$

$$\lim_{n\to\infty} \min_{\lambda' \in \Lambda_n} d(\lambda, \lambda') = 0, \qquad \forall \lambda \in \Lambda. \tag{2.1}$$

A3) For every λ_i, $1 \le i \le M$

$$p(\lambda_i) = p_n(\lambda_i) \triangleq \sum_{\lambda \in \Lambda_n} \beta_n(\lambda)\delta(\lambda_i - \lambda) \tag{2.2}$$

where $\delta(\cdot)$ is the Dirac function and $\{\beta_n(\lambda)\}_{\lambda\in\Lambda_n}$ are the prior probabilities of the grid points.

A4) The prior probabilities $\{\beta_n(\lambda)\}_{\lambda\in\Lambda_n}$ satisfy

$$\beta_n(\lambda) \ge 2^{-\epsilon_n n} \tag{2.3}$$

for some positive sequence $\{\epsilon_n\}_{n\ge 1}$, independent of λ and $\lim_{n\to\infty}\epsilon_n = 0$.

Note that A4 implies that the cardinality $|\Lambda_n|$ of the grid is upper bounded by $2^{\epsilon_n n}$ since

$$1 = \sum_{\lambda\in\Lambda_n} \beta_n(\lambda) \ge 2^{-\epsilon_n n}|\Lambda_n|. \tag{2.4}$$

Assumptions A1 and A2 are needed only to guarantee that the entire parameter space is well "covered" by Λ_n for large n. For the formal proof of the theorem below, only assumptions A3 and A4 will be needed. Let

$$\mu(x|y_i, H_i) \triangleq \frac{\max_{\lambda_i\in\Lambda_n}[p(x|\lambda_i, H_i)p(y_i|\lambda_i)]}{\max_{\lambda_i\in\Lambda_n} p(y_i|\lambda_i)}. \tag{2.5}$$

Define the AB decision rule Ω' by

$$\Omega_i'(Y) = \{x: \mu(x|y_i, H_i) \ge \mu(x|y_j, H_j), \qquad \forall j \ne i\}. \tag{2.6}$$

In other words, we choose the index i which maximizes $\mu(x|y_i, H_i)$. Note that the test statistics $\mu(x|y_i, H_i)$ defined in (2.5) can be viewed as a measure for "statistical similarity" between x and y_i. If x and y_i emerge from the same source then it is likely that a single parameter set can "fit" both x and y_i and result in a relatively high value of the joint maximum likelihood in the numerator of (2.5). On the other hand, if x and y_i do not emerge from the same source, then $\mu(x|y_i, H_i)$ should have a lower value.

Theorem 1: Under assumptions A3 and A4, for every Y

$$\lim_{n\to\infty}\left[\frac{1}{n}\log P_{\Omega'}(e|Y) - \frac{1}{n}\log P_{\Omega^*}(e|Y)\right] = 0.$$

The theorem states that the asymptotic behavior of $n^{-1}\log P_{\Omega'}(e|Y)$ is equivalent to that of $n^{-1}\log P_{\Omega^*}(e|Y)$ and hence optimal. Namely, if the parametric family and Y are such that $P_{\Omega^*}(e|Y)$ decays exponentially fast as $n \to \infty$, then $P_{\Omega'}(e|Y)$ also decays exponentially at the same rate. The theorem can be extended to continuous prior densities $p(\lambda_i) > 0$ whose support is Λ. This extension requires, however, additional regularity conditions concerning uniform continuity of the likelihood function at its maximum. The discrete prior considered here is more realistic since in practice λ_i can only be represented with finite precision.

Proof of Theorem 1: We first note that by A3

$$p(y_i|H_i) = \int_\Lambda p_n(\lambda_i)p(y_i|\lambda_i)\,d\lambda_i = \sum_{\lambda_i\in\Lambda_n}\beta_n(\lambda_i)p(y_i|\lambda_i) \le \max_{\lambda_i\in\Lambda_n} p(y_i|\lambda_i). \tag{2.7}$$

Let $\lambda^* = \operatorname{argmax}_{\lambda_i\in\Lambda_n} p(y_i|\lambda_i)$. Now by A3 and A4 we get

$$p(y_i|H_i) \ge \beta_n(\lambda^*)p(y_i|\lambda_i = \lambda^*) \ge 2^{-\epsilon_n n}\max_{\lambda_i\in\Lambda_n} p(y_i|\lambda_i). \tag{2.8}$$

In a similar manner, for $p(x, y_i|H_i)$ we have

$$\begin{aligned} 2^{-\epsilon_n n}\max_{\lambda_i\in\Lambda_n} p(x|\lambda_i)p(y_i|\lambda_i) &\le p(x, y_i|H_i) \\ &= \int_\Lambda p_n(\lambda_i)p(x|\lambda_i)p(y_i|\lambda_i)\,d\lambda_i \\ &\le \max_{\lambda_i\in\Lambda_n} p(x|\lambda_i)p(y_i|\lambda_i). \end{aligned} \tag{2.9}$$

Hence, by (1.7), (2.5), (2.7)–(2.9), we have that

$$2^{-\epsilon_n n}\mu(x|y_i, H_i) \le p(x|y_i, H_i) \le 2^{\epsilon_n n}\mu(x|y_i, H_i). \tag{2.10}$$

Define now an auxiliary cost function for any decision rule $\Omega(Y)$

$$\mu_\Omega(e|Y) \triangleq \frac{1}{M}\sum_{i=1}^{M}\int_{\Omega_i^c(Y)}\mu(x|y_i, H_i)\,dx \tag{2.11}$$

and note from (2.10) that

$$2^{-\epsilon_n n}\mu_\Omega(e|Y) \le P_\Omega(e|Y) \le 2^{\epsilon_n n}\mu_\Omega(e|Y). \tag{2.12}$$

Since Ω^* minimizes $P_\Omega(e|Y)$ and Ω' minimizes $\mu_\Omega(e|Y)$, it now follows that

$$\begin{aligned} P_{\Omega^*}(e|Y) \le P_{\Omega'}(e|Y) &\le 2^{\epsilon_n n}\mu_{\Omega'}(e|Y) \\ &\le 2^{\epsilon_n n}\mu_{\Omega^*}(e|Y) \le 2^{2\epsilon_n n}P_{\Omega^*}(e|Y) \end{aligned} \tag{2.13}$$

or, equivalently,

$$0 \le \frac{1}{n}\log P_{\Omega'}(e|Y) - \frac{1}{n}\log P_{\Omega^*}(e|Y) \le 2\epsilon_n. \tag{2.14}$$

Since $\epsilon_n \to 0$, this completes the proof of Theorem 1.

Note that if the priors $\{\beta_n(\lambda)\}_{\lambda\in\Lambda_n}$ are uniform, then inequalities (2.3) and (2.4) become equalities with $\epsilon_n = n^{-1}\log|\Lambda_n|$. In practice, however, $\log|\Lambda_n|$ is upper bounded by the number of bits available for representing λ_i in the given machine.

III. Discussion

In this section several theoretical aspects and generalizations of Theorem 1 are discussed. For simplicity, we shall adopt throughout the sequel the following abbrevi-

ated notation for the parametric conditional pdf's of x and y_i:

$$p(x|\lambda) \triangleq p(x|\lambda_i = \lambda, H_i) \tag{3.1a}$$

$$p(y_i|\lambda) \triangleq p(y_i|\lambda_i = \lambda). \tag{3.1b}$$

As mentioned in Section I, Theorem 1 is meaningful if the conditional error probability $P_{\Omega^*}(e|Y)$ decays exponentially as $n \to \infty$. It is demonstrated in the Appendix that this probability converges exponentially in the case where that $p(\cdot|\lambda)$ is i.i.d. and the length n_i, of each training sequence y_i, grows at least linearly with n. This property can be easily shown to extend to the case of Markov sources by using techniques from large deviations theory (see, e.g., [17], [33]). For wider classes of parametric densities, which can be approximated by block i.i.d. densities (or block Markovian densities), the same techniques can be applied. Specifically, let l divide n and denote $x_t^r = (x_t, x_{t+1}, \cdots, x_r)$ for $t \le r$. Let

$$G(l, \delta_l) = \left\{ x \in U^n: \left| \frac{1}{n} \log p(x|\lambda) - \frac{1}{n} \sum_{j=0}^{n/l-1} \log p(x_{jl+1}^{jl+l}|\lambda) \right| > \delta_l \right\}. \tag{3.2}$$

If there exists a sequence $\{\delta_l\}$ independent of n and x with $\lim_{l\to\infty} \delta_l = 0$ such that for all large n

$$\int_{G(l,\delta_l)} p(x|\lambda)\, dx \le e^{-\epsilon n} \tag{3.3}$$

for every fixed l and $\epsilon > 0$, then $p_{\Omega^*}(e|Y)$ decays exponentially. Equations (3.2) and (3.3) mean that the sample space can be divided into two subsets, where in the first, $p(\cdot|\lambda)$ behaves approximately like a block i.i.d. source and hence exponential decay of the error probability is attained, and the other subset has itself an exponentially small probability and hence can be always considered an "error region" without violating the exponential behavior of the total error probability. It is shown in the Appendix that if $p(\cdot|\lambda)$ is a hidden Markov model (HMM) with strictly positive transition probabilities (see Section IV-A below), then (3.3) holds with a particular choice of the sequence $\{\delta_l\}$.

The AB decision rule degenerates to the PI method if the training sequences $\{y_i\}$ are considerably longer than the test sequence x. In this case, the parameter set λ_i that maximizes the denominator of (2.5) is very close to the parameter set that maximizes the numerator. Hence, the factor $p(y_i|\lambda_i)$ in both numerator and denominator is essentially canceled. Alternatively, if $p(\lambda_i|y_i) = \delta(\lambda_i - \hat{\lambda}_i)$, namely, the posterior probability mass of λ_i is concentrated at the ML estimate $\hat{\lambda}_i$ obtained from y_i, then it is easy to see from (1.7) that the AB decision rule coincides with the PI decision rule.

It was shown in [14] that if a rejection option is allowed, then a test statistic similar to (2.5) attains the best tradeoff between the rejection probability and the false decision probability. Specifically, let

$$h(x, y_i) = \frac{1}{n} \log \frac{\max_{\lambda} [p(x|\lambda)p(y_i|\lambda)]}{[\max_{\lambda} p(x|\lambda)][\max_{\lambda} p(y_i|\lambda)]} \tag{3.4}$$

and consider a decision rule $\tilde{\Omega}$ that for a given $\xi > 0$, selects the index i if $h(x, y_i) + \xi \ge 0$ while $h(x, y_j) + \xi < 0$ for all $j \ne i$, and rejects whenever there is no unique index i with $h(x, y_i) + \xi \ge 0$. This decision rule has been shown [14] to attain the minimum asymptotic rejection probability among all decision rules for which the false decision probability is less than $e^{-\xi n}$.

Following Forney [18], Theorem 1 can be extended to Bayesian erasure and list decision schemes. These are generalized decision schemes where the decision regions $\{\Omega_i(Y)\}_{i=1}^M$ are not necessarily a partition of the sample space. Two possibilites are considered.

a) *Erasure* (rejection) schemes, which include an option of not making a decision. This is useful for erasure channels [18] where a feedback channel can be utilized for repetition requests. In this case, the union of all decision regions $\cup_{i=1}^M \Omega_i(Y)$ does not necessarily cover the entire sample space, and if $x \in [\cup_{i=1}^M \Omega_i(Y)]^c$, then a rejection is made. The best tradeoff between the rejection probability and the probability of false decision, is a desirable objective in this case.

b) *List* schemes, which include an option of providing more than one candidate estimate of m. This is applicable in speech recognition systems, where a second stage recognizer may attempt to correct errors with all possibilities on the lists, by using grammar rules or a language model. Here $\{\Omega_i(Y)\}_{i=1}^M$ may intersect with each other, and if x falls in the intersection of two or more regions $\Omega_i(Y)$, then a list of the corresponding indices is produced. Here the objective is to attain the best tradeoff between the expected number of incorrect items on list, and the list-error probability, that is, the probability that the entire list is incorrect.

Define the decision regions:

$$\Omega_i(Y) = \left\{ x: \frac{p(x|y_i, H_i)}{\sum_{j\ne i} p(x|y_j, H_j)} \ge e^{\xi n} \right\}, \qquad i = 1, \cdots, M \tag{3.5}$$

where ξ is an arbitrary real number. It is shown in [18] that for $\xi \ge 0$, (3.5) is an optimal erasure scheme in the sense of minimizing the erasure probability for a prescribed false decision probability, which is determined by ξ. For $\xi < 0$, (3.5) is an optimal list scheme which minimizes the expected number of erroneous candidates on list for a prescribed list error probability, which is controlled by ξ. Similar to (2.6), the decision rule (3.5) can be approximated by

$$\bar{\Omega}_i(Y) = \left\{ x: \frac{\mu(x|y_i, H_i)}{\max_{j\ne i} \mu(x|y_i, H_i)} \ge e^{\xi n} \right\} \tag{3.6}$$

with equivalent conditional asymptotic exponents in the sense of Theorem 1. Note that for $\xi = 0$, (3.6) agrees with (2.6). It should be pointed out that this is different from the erasure scheme $\tilde{\Omega}$ described in the previous paragraph, for two reasons. First, while here λ is considered a random variable, in [14] it is assumed fixed. Second, here we focus on the conditional probabilities of erasure and false decision given Y, while in [14] the unconditional probabilities of these events are considered.

Another possible generalization of Theorem 1 is associated with classification of noisy signals. Let $z = x + w$ be an observed test signal, where $x \in \boldsymbol{R}^n$ is a clean signal and $w \in \boldsymbol{R}^n$ is a sample function of the noise process, and assume that x and w are statistically independent. Assume further that the noise process can be modeled by a parametric pdf $q(\cdot \mid \nu)$, where the unknown parameter ν is considered a random variable. Suppose that in addition to the training data Y of the clean sources, a training sequence w' of the noise is available. Let

$$f(z \mid \lambda_i, \nu) = \int_{\boldsymbol{R}^n} p(x \mid \lambda_i) q(z - x \mid \nu)\, dx \tag{3.7}$$

denote the pdf of the noisy signal z given the ith source. In this case, the Bayesian decision rule analogous to (1.9) is to select the index i that maximizes the conditional density

$$f(z \mid w', y_i, H_i) = \frac{\int\int p(\lambda)\gamma(\nu) f(z \mid \lambda, \nu) p(y_i \mid \lambda) q(w' \mid \nu)\, d\lambda\, d\nu}{\int p(\lambda) p(y_i \mid \lambda)\, d\lambda \int \gamma(\nu) q(w' \mid \nu)\, d\nu} \tag{3.8}$$

where $\gamma(\nu)$ is a prior density of ν. Similar to (2.5), $f(z \mid w', y_i, H_i)$ can be approximated by

$$\mu(z \mid w', y_i, H_i) = \frac{\max\limits_{\lambda, \nu} [f(z \mid \lambda, \nu) p(y_i \mid \lambda) q(w' \mid \nu)]}{\max\limits_{\lambda} p(y_i \mid \lambda) \max\limits_{\nu} q(w' \mid \nu)}. \tag{3.9}$$

This approach is different from [19] where $\{\lambda_i\}$ and ν are first estimated from training data of clean signals and noise, and then used in the MAP decision rule.

Finally, it should be pointed out that the computational complexity associated with the AB decision rule is significantly larger than that of the PI decision rule. The reason is that the maximization of $[p(x \mid \lambda) p(y_i \mid \lambda)]$ over λ must be performed only upon observing the test sequence x. This is in contrast to the PI approach where the testing phase requires only likelihood computations using predesigned models $\{\hat{\lambda}_i\}$. Furthermore, the AB decision rule is more sensitive to local rather than global maximization as it constitutes the ratio of two global maxima.

IV. Experimental Results

In subsection A below we define the HMM in general, and then describe some specific types of HMM's which were used in our experimental studies. In subsection B the main numerical procedures are explained. In subsection C we examine the AB decision rule on computer generated HMM's and compare it to the PI approach. Finally, subsection D provides simulation results on speech signals.

A. Hidden Markov Models

Let $x = \{x_t, t = 1, \cdots, n\}$, $x_t \in \boldsymbol{U}$ be a sequence of observations. We consider HMM's with S states, where the state set is denoted by $\boldsymbol{S} = \{1, 2, \cdots, S\}$. Let $s = \{s_t, t = 1, \cdots, n\}$, $s_t \in \boldsymbol{S}$, be a sequence of states corresponding to x. The pdf of x is given by

$$p(x \mid \lambda) = \sum_s p(x, s \mid \lambda) = \sum_s p(s \mid \lambda) p(x \mid s, \lambda) \tag{4.1}$$

where $p(s \mid \lambda)$ is the probability of the sequence of states s, and $p(x \mid s, \lambda)$ is the probability of the given output sequence x given s. For first-order HMM's we have

$$p(s \mid \lambda) = \prod_{t=1}^{n} a_{s_{t-1} s_t} \tag{4.2}$$

where $a_{s_{t-1} s_t}$ denotes the transition probability from state s_{t-1} at time $(t-1)$ to state s_t at time t, and $a_{s_0 s_1} \triangleq \pi_{s_1}$ is the initial state probability. For $p(x \mid s, \lambda)$ we make the following standard assumption:

$$p(x \mid s, \lambda) = \prod_{t=1}^{n} p(x_t \mid s_t, \lambda) \triangleq \prod_{t=1}^{n} b(x_t \mid \theta_{s_t}) \tag{4.3}$$

where $b(x_t \mid \theta_{s_t})$ is the output parametric pdf associated with state s_t with parameter vector $\theta_{s_t} \in \Psi \subset \boldsymbol{R}^q$. The parameter set of an HMM is $\lambda \triangleq \{\pi, A, \Theta\}$, where $\pi = \{\pi_\alpha\}$, $A = \{a_{\alpha\beta}\}$, and $\Theta = \{\theta_\alpha\}$, $\alpha, \beta \in \boldsymbol{S}$, and $\lambda \in \Lambda$.

An HMM is called left–right [1, p. 266] if $\{a_{\alpha\beta}\}$ are constrained to vanish for all $\beta < \alpha$ and $\beta > \alpha + 1$, namely, the only allowed transitions from state $\alpha \in \boldsymbol{S}$ are the self-transition and a transition to the next state $\alpha + 1$. Clearly, once the last state $\alpha = S$ is visited, the process remains in that state. If no such constraints on A are imposed, the model is usually referred to as ergodic model.

Two types of output pdf's $b(\cdot \mid \theta_\alpha)$ are considered here.

1) Gaussian Autoregressive (AR) HMM's: In this case the observation vectors $x_t \in \boldsymbol{R}^K$ are frames of K successive waveform samples $x_t = \{x_t(0), \cdots, x_t(K-1)\}$ and

$$b(x_t \mid \theta_{s_t}) = [2\pi\sigma^2(s_t)]^{-K/2} \exp\left\{ -\frac{1}{2\sigma^2(s_t)} \cdot \sum_{r=0}^{K-1} \left[x_t(r) - \sum_{j=1}^{p} a_j(s_t) x_t(r-j) \right]^2 \right\}. \tag{4.4}$$

The initial conditions $x_t(-p), \cdots, x_t(-1)$ are either taken from the previous frame x_{t-1} or set to zero. Here $\theta_\alpha \triangleq (\sigma^2(\alpha), \cdots, a_p(\alpha))$ and hence $q = p + 1$.

2) Gaussian Cepstral HMM's: In this case the observations $x_t \in \boldsymbol{R}^p$ are vectors of cepstral coefficients obtained from frames of length K in the following manner. First, the pth order AR coefficients from each frame are estimated by the autocorrelation method [20]. Then, the coefficients of each frame, say the tth, $\{a_l(t)\}_{l=1}^p$ are transformed into cepstral coefficients $\{x_t(l)\}_{l=1}^p$ defined as

$$x_t(l) = -\frac{1}{2\pi}\int_{-\pi}^{\pi} d\omega\, e^{jl\omega} \log\left|1 - \sum_{r=1}^{p} a_r(t) e^{-jr\omega}\right|^2 \qquad l = 1, \cdots, p. \tag{4.5}$$

The cepstral coefficients at each state, say s_t, are assumed Gaussian independent random variables with means $\{\mu_l(s_t)\}_{l=1}^p$ and variances $\{\sigma_l^2(s_t)\}_{l=1}^p$, respectively, i.e.,

$$b(x_t|\theta_{s_t}) = \prod_{l=1}^{p} \left\{[2\pi\sigma_l^2(s_t)]^{-1/2} \cdot \exp\left[-\frac{(x_t(l) - \mu_l(s_t))^2}{2\sigma_l^2(s_t)}\right]\right\}. \tag{4.6}$$

Here $\theta_\alpha = (\mu_1(\alpha), \cdots, \mu_p(\alpha), \sigma_1^2(\alpha), \cdots, \sigma_p^2(\alpha))$ and hence $q = 2p$.

Note that although both AR HMM and cepstral HMM are based on AR modeling, there is an essential difference between them. While in the AR HMM approach the waveform from each state is assumed Gaussian and the AR coefficients are assumed deterministic, in cepstral HMM's the AR coefficients are considered random variables. Our experiments in subsection D below were performed for the four combinations: ergodic AR HMM's, ergodic cepstral HMM's, left–right AR HMM's, and left–right cepstral HMM's.

B. Numerical Procedures

The maximization of the likelihood functions in (1.15) and (2.5) was implemented by the segmental K-means algorithm [21], [22]. For a given observation sequence, say, $\nu = \{\nu_t, t = 1, \cdots, n\}$, this algorithm performs local joint maxmization of $p(\nu, s|\lambda)$ over s and λ. This maximization of $p(\nu, s|\lambda)$ by the segmental K-means algorithm, rather than of $p(\nu|\lambda)$ by the Baum algorithm [5], [6], [23] is numerically easier and results in similar λ estimates for large K [24]. The segmental K-means algorithm performs alternate maximization of $p(\nu, s|\lambda)$ once over the state sequences s for a given $\lambda \in \Lambda$ using the Viterbi algorithm, and then over λ for the resulting most likely state sequence s^* using reestimation formulas similar to those of the Baum algorithm. This algorithm was shown [22] to converge to a local maximum under certain regularity conditions. For ergodic AR and cepstral HMM's, the segmental K-means algorithm was initialized by a model obtained from vector quantization (VQ) [25]–[27] of $\{v_i\}$ into S code words representing the output pdf's from the S states θ_α, $\alpha = 1, \cdots, S$. The distortion measure used here was

$$d(\nu_t, \theta_\alpha) = -\log b(\nu_t|\theta_\alpha) \tag{4.7}$$

which for Gaussian pdf's $b(\cdot|\theta_\alpha)$ asymptotically coincides with the Itakura–Saito distortion measure [28]–[30]. The VQ provides initial estimates of $\{\theta_\alpha\}_{\alpha=1}^S$. The initial estimates for π and A results from first decoding the data using the designed vector quantizer, and then computing the relative frequencies at which each state was initially used and each state transition occurs. For left–right HMM's, initial model estimates are obtained from uniform segmentation of ν into S intervals and estimating θ_α, $\alpha \in S$, from the αth segment. Here π and A are initially estimated from the state sequence defined by this segmentation. The initial matrix A is therefore left–right and this structure is preserved in each iteration of the segmental K-means algorithm. (See [1, eq. (44)).

C. Classification of Computer Generated Gaussian AR HMM's

We first examined the AB and the PI decision rules on computer generated data from ergodic Gaussian AR HMM's. We used $M = 9$ sources whose parameter sets λ_i, $i = 1, \cdots, 9$, were estimated from spoken versions of the English E-set letters b, c, d, e, g, p, t, v, z. For each model λ_i, we used $S = 5$ states and an AR model of order $p = 8$. The dimension of the output vector x_t was $K = 256$. The training sequence y_i from each source comprised a fixed number $c_i = c$ of statistically independent utterances each of length n. From each source, 40 testing sequences of length n were generated. Table I provides comparative results of classification accuracy in percent for the PI and AB decision rules where $3 \le c \le 5$ and $10 \le n \le 14$.

As can be seen, the AB decision rule significantly outperforms the PI decision rule even for the small values of n considered here, although the result stated in Theorem 1 is merely asymptotic. As c and n increase, however, the differences in classification accuracy between the two decision rules decrease.

Comment: Note that the recognition accuracy associated with the AB classification rule does not grow monotonically with n. The reason for this effect is sensitivity to local maxima when calculating $\mu(x|y_i, H_i)$ (in particular, its numerator) by an iterative algorithm.

D. Speech Recognition Results

We next compared the PI and AB classification approaches in speaker dependent isolated word recognition of the English E-set words, recorded from 4 speakers (2 females and 2 males) through a telephone handset and sampled at 6.67 kHz. For each speaker and each word, 5 training utterances and 10 testing utterances were used.

Preemphasis and windowing of the speech signals were applied to obtain x and $y_1, \cdots, y_M$. The preemphasis filter was of the form $H(z) = 1 - \alpha z^{-1}$. We used a gen-

TABLE I
CLASSIFICATION RESULTS ON COMPUTER GENERATED AR HMM'S

n	10		11		12		13		14	
c	PI	AB	PI	AB	PI	AB	PI	AB	PI	AB
3	40.27	73.89	43.33	68.33	50.27	73.61	60.27	75.00	64.72	82.78
4	68.33	83.33	67.22	80.83	70.83	90.00	66.94	83.61	68.89	84.72
5	73.33	93.66	74.16	90.55	78.60	99.44	88.89	96.38	84.16	94.44

TABLE II
RECOGNITION RESULTS FOR LEFT–RIGHT MODELS

Type	Rule	K	p	S	α	β	1	2	3	4	Average
AR	AB	270	12	10	0.95	0.90	88.9	92.2	91.1	74.4	86.7
AR	PI	256	10	10	0.00	1.00	88.9	87.8	88.9	75.6	85.3
Ceps.	AB	256	10	10	0.98	0.58	93.3	91.1	93.3	73.3	87.8
Ceps.	PI	256	12	10	0.95	0.50	90.0	91.1	88.9	73.3	85.8

TABLE III
RECOGNITION RESULTS FOR ERGODIC MODELS

Type	Rule	K	p	S	α	β	1	2	3	4	Average
AR	AB	256	8	8	0.00	1.00	85.6	87.8	74.4	54.4	75.6
AR	PI	256	8	10	0.00	1.00	85.6	90.0	76.7	71.1	80.8
Ceps.	AB	256	12	11	0.95	0.50	72.2	72.2	45.6	62.2	63.1
Ceps.	PI	256	12	10	0.95	0.50	80.0	82.2	66.7	62.2	72.8

eralized Hanning window

$$w(k) = \beta - (1 - \beta)\cos\left(\frac{2\pi k}{K}\right) \qquad k = 0, 1, \cdots, K-1. \tag{4.8}$$

For cepstral HMM's, the cepstral coefficients of each frame were obtained from the AR coefficients of that frame using the recursive formula [31]

$$x_t(1) = a_1(t)$$
$$x_t(l) = \sum_{r=1}^{l-1}\left(1 - \frac{r}{l}\right)a_r(l)x_t(l-r) + a_l(t) \qquad 1 < l \le p. \tag{4.9}$$

Four experiments were performed, corresponding to left–right AR HMM's, left–right cepstral HMM's, ergodic AR HMM's, and ergodic cepstral HMM's. In each experiment the following design parameters were empirically optimized for both the AB and PI decision rules in order to obtain the best average recognition accuracy over the four speakers: the frame length K, the AR model order p, the number of states S, the preemphasis filter coefficient α, and the window parameter β.

Table II summarizes recognition accuracy results of the AB and PI decision rules for left–right AR HMM's and left–right cepstral HMM's. We present the best values found for the design parameters, the recognition accuracy in percent for each one of the four speakers, and the average accuracy over the four speakers.

The results indicate only a slight preference of the AB approach over the PI approach in both cepstral and AR types of HMM. The significance of these results is that the standard PI method performs close to the asymptotically optimal AB rule within the Bayesian framework of hidden Markov modeling. Note also that AR HMM's and cepstral HMM's provide very similar recognition performance. Table III presents results for the ergodic version of the two types of HMM's.

Comparing Tables II and III, we observe that left–right HMM's considered here are significantly superior to ergodic HMM's for speech recognition applications (see also [1, p. 266]). Furthermore, the performance of the AB decision rule is poor compared to the PI decision rule since the former is more sensitive to the model correctness assumption.

V. CONCLUSIONS

We analyzed and examined an asymptotically optimal version of Nádas's Bayesian decision rule, which requires neither explicit knowledge of the prior pdf's of the parameter sets of the sources nor integration over the parameter space. The proposed AB decision rule was compared to the standard PI approach commonly used in speech recognition, using HMM's for the acoustic signals. For computer generated HMM's the performance of the AB decision rule was found to be significantly better than the PI

approach. For speech signals, the AB approach provided only a minor, improvement compared to the PI method. This suggests that in speech recognition, the PI approach is fairly close to asymptotic optimum in the present framework of HMM's. Hence, we believe that improvement in speech recognition accuracy may be accomplished if better models and estimation approaches for these models, which are less sensitive to the model correctness assumption are found, rather than better decision rules for the present modeling framework. One modeling approach might be to allow a small model mismatch between training and testing. In practice, such a mismatch is likely to occur, especially if training and testing are performed at a different time or under different recording conditions. One can model this mismatch by allowing a small departure of the pdf that governs the test sequence from the pdf of the corresponding training sequence.

The experimental results on computer generated HMM's are meaningful for other classification applications, e.g., in digital communication systems, radar, and sonar, where the channel characteristics are unknown but can be observed empirically from training sequences.

Finally, the extensions of the proposed approach to Bayesian rejection and list schemes and to classification in a noisy environment, may be important for future research in speech recognition.

Appendix

1. Exponential Decay of $P_{\Omega^*}(e \mid Y)$ in the IID Case

Let $p(x \mid \lambda) = \Pi_{t=1}^{n} g(x_t \mid \lambda)$ and assume the following regularity conditions.

1) The lengths of the training sequences satisfy

$$\lim_{n\to\infty} \frac{n_i}{n} \triangleq c_i > 0. \tag{A.1}$$

2) The marginal entropy is finite, i.e.,

$$\left| \int_U g_\lambda(u) \log g_\lambda(u)\, du \right| < \infty. \tag{A.2}$$

3) For every two distinct parameter sets λ and λ'

$$D(\lambda \| \lambda') \triangleq \int_U g_\lambda(u) \log \frac{g_\lambda(u)}{g_{\lambda'}(u)}\, du > 0. \tag{A.3}$$

4) The maximum likelihood estimator

$$\hat{\lambda} \triangleq \operatorname{argmax}_{\lambda\in\Lambda} \Pi_{t=1}^{n} g_\lambda(x_t)$$

is strongly consistent. Furthermore, $\{\hat{\lambda}_i\}$ converge to distinct points $\{\lambda_i\}$ in Λ.

5) The marginal density $g_\lambda(u)$ is continuous in λ for almost all $u \in U$.

We now show that under these assumptions

$$\lim_{n\to\infty} \frac{1}{n} \log P_{\Omega^*}(e \mid Y) < 0 \quad \text{a.s.} \tag{A.4}$$

It is sufficient to show exponential convergence for the PI decision rule $\Omega^*(\hat{\lambda})$. To do this, we upper bound $P_{\Omega^*(\hat{\lambda})}(e \mid Y)$:

$$p_{\Omega^*(\hat{\lambda})}(e \mid Y) = \frac{1}{M} \sum_{i=1}^{M} \int_{[\Omega_i^*(\hat{\lambda})]^c} p(x \mid y_i, H_i)\, dx$$

$$\leq (M-1) \max_{j\neq i} \int_{\Omega_j^*(\hat{\lambda})} p(x \mid y_i, H_i)\, dx$$

by (2.9)

$$\leq 2^{\epsilon n n} (M-1) \max_{j\neq i} \int_{\Omega_j^*(\hat{\lambda})} \mu(x \mid y_i, H_i)\, dx$$

$$\leq 2^{\epsilon n n} (M-1) \max_{j\neq i} \left\{ [p(y_i \mid \lambda_i)]^{-1} \cdot \max_{\lambda} \left[p(y_i \mid \lambda_i) \int_{\Omega_j^*(\hat{\lambda})} p(x \mid \lambda)\, dx \right] \right\}$$

$$\leq \exp_2 \left\{ n \left[\epsilon_n + \frac{1}{n} \log (M-1) + \max_{j\neq i} \max_{\lambda} \cdot \left(\frac{1}{n} \log \frac{p(y_i \mid \lambda)}{p(y_i \mid \lambda_i)} + \frac{1}{n} \log \int_{\Omega_j^*(\hat{\lambda})} p(x \mid \lambda)\, dx \right) \right] \right\}. \tag{A.5}$$

As $n \to \infty$, we have $\epsilon_n \to 0$, $n^{-1} \log (M-1) \to 0$, $\hat{\lambda}_i \to \lambda_i$ almost surely, and by the ergodic theorem for densities [32]

$$\frac{1}{n} \log \frac{p(y_i \mid \lambda)}{p(y_i \mid \lambda_i)} \to c_i D(\lambda_i \| \lambda) \tag{A.6}$$

almost surely. Hence,

$$-\frac{1}{n} \log P_{\Omega^*(\hat{\lambda})}(e \mid Y) \to \min_{j\neq i} \min_{\lambda} \left[c_i D(\lambda_i \| \lambda) - \frac{1}{n} \log \int_{x: p(x\mid\lambda_j) > p(x\mid\lambda_i)} p(x \mid \lambda)\, dx \right] \tag{A.7}$$

almost surely. We now focus on the integral on the right-hand side of (A.7). By the Chernoff bound

$$\frac{1}{n} \log \int_{x: p(x\mid\lambda_j) > p(x\mid\lambda_i)} p(x \mid \lambda)\, dx \leq \frac{1}{n} \log \min_{\phi\geq 0} \int_{U''} p(x \mid \lambda) \left[\frac{p(x \mid \lambda_j)}{p(x \mid \lambda_i)} \right]^{\phi} dx$$

$$= \log \min_{\phi\geq 0} \int_U g(u \mid \lambda) \left[\frac{g(u \mid \lambda_j)}{g(u \mid \lambda_i)} \right]^{\phi} du \triangleq -R(\lambda, \lambda_i, \lambda_j) \leq 0. \tag{A.8}$$

Thus,

$$\lim_{n\to\infty} \left[-\frac{1}{n} \log P_{\Omega^*(\hat{\lambda})}(e \mid Y) \right] \geq \min_{j\neq i} \min_{\lambda} [c_i D(\lambda_i \| \lambda) + R(\lambda, \lambda_i, \lambda_j)]. \tag{A.9}$$

By (A.3), $D(\lambda_i \| \lambda)$ is strictly positive unless $\lambda = \lambda_i$. If $\lambda = \lambda_i$, however,

$$R(\lambda, \lambda_i, \lambda_j) = R(\lambda_i, \lambda_j, \lambda_i) = -\log \min_{\phi \geq 0} \int_U [g(u|\lambda_i)]^{1-\phi} \cdot [g(u|\lambda_j)]^{\phi} \, du > 0$$

for any $\lambda_i \neq \lambda_j$. Since $\{\lambda_i\}$ are assumed distinct, the right-hand side of (A.9) is strictly positive.

2. *HMM's Can be Approximated by Block IID Sources*

Let $p(\cdot|\lambda)$ be an HMM as defined in Section IV-A. Assume further that there exists $\delta > 0$ such that $\pi_\alpha \geq \delta$ and $\pi_{\alpha\beta} \geq \delta$ for all $\alpha, \beta \in \boldsymbol{S}$, and hence $\delta \cdot \pi_\gamma \leq a_{\alpha\beta} \leq \delta^{-1} \cdot \pi_\gamma$ for $\alpha, \beta, \gamma \in \boldsymbol{S}$. For $i < j$ let s_i^j denote a segment $(s_i, s_{i+1}, \cdots, s_j)$ of the state sequence $s = s_1^n$. Then, from (4.1)–(4.3) we have

$$\begin{aligned} p(x|\lambda) &= \sum_s \prod_{t=1}^{n} a_{s_{t-1}s_t} b(x_t|\theta_{s_t}) \\ &= \sum_s \prod_{j=0}^{n/l-1} a_{s_{jl}s_{jl+1}} b(x_{jl+1}|\theta_{s_{jl+1}}) \cdot \prod_{i=2}^{l} a_{s_{jl+i-1}s_{jl+i}} b(x_{jl+i}|\theta_{s_{jl+i}}) \\ &\leq \sum_s \prod_{j=0}^{n/l-1} \delta^{-1}\pi_{jl+1} b(x_{jl+1}|\theta_{s_{jl+1}}) \cdot \prod_{i=2}^{l} a_{s_{jl+i-1}s_{jl+i}} b(x_{jl+i}|\theta_{s_{jl+i}}) \\ &= \delta^{-n/l} \prod_{j=0}^{n/l-1} \sum_{s_{jl+1}^{jl+l}} \pi_{jl+1} b(x_{jl+1}|\theta_{s_{jl+1}}) \cdot \prod_{i=2}^{l} a_{s_{jl+i-1}s_{jl+i}} b(x_{jl+i}|\theta_{s_{jl+i}}) \\ &= \exp[nl^{-1} \log(1/\delta)] \prod_{j=0}^{n/l-1} p(x_{jl+1}^{jl+l}|\lambda). \end{aligned} \tag{A.10}$$

In the same manner, one obtains

$$p(x|\lambda) \geq \exp[-nl^{-1} \log(1/\delta)] \prod_{j=0}^{n/l-1} p(x_{jl+1}^{jl+l}|\lambda) \tag{A.11}$$

and hence by taking the logarithms of (A.10) and (A.11), and dividing by n we find that for all $x \in \boldsymbol{U}^n$

$$\left| \frac{1}{n} \log p(x|\lambda) - \frac{1}{n} \sum_{j=0}^{n/l-1} \log p(x_{jl+1}^{jl+l}|\lambda) \right| \leq \frac{1}{l} \log \frac{1}{\delta}. \tag{A.12}$$

Thus, for $\delta_l = l^{-1} \log 1/\delta$ (which is independent of x and n and tends to zero as $l \to \infty$), we obtain $\int_{G(l,\delta_l)} p(x|\lambda)\, dx = 0$, which is much stronger than (3.3).

Comment: The assumption that all state transition probabilities are strictly positive can be replaced by a weaker assumption that all states communicate within a finite number of steps, i.e., there exists a finite positive integer k such that all entries of A^k are strictly positive, where A is the state transition matrix.

ACKNOWLEDGMENT

The authors greatly appreciate the helpful comments made by the anonymous referees.

REFERENCES

[1] L. R. Rabiner, "A tutorial on hidden Markov Models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[2] J. Ziv, "Universal decoding for finite-state channels," *IEEE Trans. Inform. Theory*, vol. IT-31, no. 4, pp. 453–460, 1985.

[3] S. U. H. Qureshi, "Adaptive equalization," *Proc. IEEE*, vol. 73, no. 9, pp. 1349–1387, Sept. 1985.

[4] A. Nádas, "Optimal solution of a training problem in speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, no. 1, pp. 326–329, 1985.

[5] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Stat.*, vol. 41, no. 1, pp. 164–171, 1970.

[6] 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.

[7] 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.

[8] S. Kullback, *Information Theory and Statistics*. New York: Dover, 1968.

[9] Y. Ephraim, A. Dembo, and L. R. Rabiner, "A minimum discrimination information approach for hidden Markov modeling," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1001–1013, Sept. 1989.

[10] P. F. Brown, "The acoustic-modeling problem in automatic speech recognition," Ph.D. dissertation, Dep. Computer Sci., Carnegie-Mellon Univ., Pittsburgh, PA, 1987.

[11] 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. Acoust., Speech, Signal Processing*, Apr. 1986, pp. 49–52.

[12] Y. Ephraim and L. R. Rabiner, "On the relations between modeling approaches for speech recognition," *IEEE Trans. Inform. Theory*, vol. 36, pp. 372–380, Mar. 1990.

[13] J. Ziv, "On classification with empirically observed statistics and universal data compression," *IEEE Trans. Inform. Theory*, vol. IT-34, no. 2, pp. 278–286, 1988.

[14] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inform. Theory*, vol. 35, no. 2, pp. 401–408, Mar. 1989.

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

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

[17] I. Csiszár, T. M. Cover, and B.-S. Choi, "Conditional limit theorems under Markov conditioning," *IEEE Trans. Inform. Theory*, vol. IT-33, no. 6, pp. 788–801, May 1987.

[18] G. D. Forney, Jr., "Exponential error bounds for erasure, list, and decision feedback schemes," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 206–220, Mar. 1968.

[19] Y. Ephraim, "Gain adapted hidden Markov models for recognition of clean and noisy speech," *Trans. Signal Processing*, to be published.

[20] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. New York: Springer, 1976.

[21] Y. Ephraim, D. Malah, and B.-H. Juang, "On the application of hidden Markov models for enhancing noisy speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 12, pp. 1846–1856, Dec. 1989.

[22] B.-H. Juang and L. R. Rabiner, "The segmental *K*-means algorithm for estimating parameters of hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, no. 9, pp. 1639–1641, Sept. 1990.

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data," *J. Roy. Stat. Soc.*, vol. B-39, pp. 1–38, 1977.

[24] N. Merhav and Y. Ephraim, "Maximum likelihood hidden Markov modeling using a dominant sequence of states," vol. 39, no. 9, pp. 2111–2115, Sept. 1991.

[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] R. Billi, "Vector quantization and Markov source models applied to speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1982, pp. 574–577.

[27] A. Buzo, A. H. Gray, Jr., R. M. Gray, and J. D. Markel, "Speech coding based on vector quantization," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 5, pp. 562–574, Oct. 1980.

[28] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencies," *Electron. Commun. Japan*, vol. 53-A, no. 1, pp. 36–43, 1970.

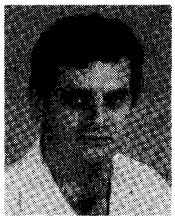
[29] K. Dzhaparidze, *Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series*. New York: Springer, 1986.

[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] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, no. 6, pp. 1304–1312, June 1974.

[32] A. R. Barron, "The strong ergodic theorem for densities: Generalized Shannon–McMillan–Breiman theorem," *Ann. Probability*, vol. 13, no. 4, pp. 1292–1303, 1985.

[33] S. Natarajan, "Large deviations, hypothesis testing, and source coding for finite Markov chains," *IEEE Trans. Inform. Theory*, vol. IT-31, no. 3, pp. 360–365, May 1985.



Neri Merhav (S'86–M'87) was born in Haifa, Israel, on March 16, 1957. He received the B.Sc., M.Sc., and D.Sc. degrees from the Technion, Israel Institute of Technology, Haifa, Israel, in 1982, 1985, and 1988, respectively, all in electrical engineering.

From 1982 to 1985 he was a Research Associate with the Israel IBM Scientific Center in Haifa, where he developed algorithms for speech coding, speech synthesis, and adaptive filtering of speech signals in array sensors. From 1988 to 1990 he was with the Speech Research Department of AT&T Bell Laboratories, Murray Hill, NJ, where he investigated and developed algorithms for speech recognition. He is currently with the Electrical Engineering Department of the Technion, and his research interests are statistical signal processing, information theory, and statistical communication.

Yariv Ephraim (S'82–M'84–SM'90) was born on September 9, 1951. He received the B.Sc., M.Sc., and D.Sc. degrees from the Technion, Israel Institute of Technology, in 1977, 1979, and 1984, respectively, all in electrical engineering.

During 1984–1985 he was a Rothschild Postdoctoral Fellow at the Information Systems Laboratory of Stanford University, CA. Since 1985 he has been a Member of the Technical Staff at the Speech Research Department of AT&T Bell Laboratories, Murray Hill, NJ.