A Bayesian Approach to the Verification Problem: Applications to Speaker Verification

Hui Jiang, *Member, IEEE,* and Li Deng, *Senior Member, IEEE*

***Abstract*—In this paper, we study the general verification problem from a Bayesian viewpoint. In the Bayesian approach, the verification decision is made by evaluating *Bayes factors* against a critical threshold. The calculation of the *Bayes factors* in turn requires the computation of several Bayesian predictive densities. As a case study, we apply the method to speaker verification based on the Gaussian mixture model (GMM). We propose an efficient algorithm to calculate the *Bayes factors* for the GMM, where the Viterbi approximation is adopted in the computation of joint Bayesian predictive densities. We evaluate the proposed method for the *NIST98* speaker verification evaluation data. Experimental results show that new Bayesian approach achieves moderate improvements over a well-trained baseline system using the conventional likelihood ratio test.**

***Index Terms*—Bayes factors, Bayesian prediction, equal error rate (EER), Gaussian mixture model (GMM), likelihood ratio test, outlier verification, speaker verification, sufficient statistics, verification problem.**

I. Introduction

During the past few decades, the verification problem has been attracting considerable research attention in the speech research community. The verification problem encompasses all problems which require a binary answer: *yes* or *no*. In speech technology, speaker verification and utterance verification are two most active areas due to their increasing importance in many practical applications. In speaker verification, based on a user's voice, the goal is to make the decision of whether to accept or to reject the identity claimed by the speaker. Successful speaker verification will enable an automatic device to use the user's voice to verify their identity and to control access to various services. These applications include voice dialing, banking over a telephone network, telephone shopping, database access services, voice mail, security control for confidential information, and remote access to computers. Utterance verification, on the other hand, aims to equip speech recognition systems with the ability to detect whether the input speech does not contain any of the words in the recognizer's vocabulary set [19], [23]. As speech recognition technology migrates from the laboratory to many services and products, the role of utterance verification has become increasingly essential. Utterance verification is especially important in the design of user-friendly systems because such systems should be able to reject speech utterances either with no valid keywords or with valid keywords but are incorrectly recognized by the speech recognizer.

In principle, both speaker verification and utterance verification, as well as all other verification problems, can be addressed in a unified theoretical framework. As we will show later, in theory, every verification problem can be cast as a problem of "statistical hypothesis testing." According to Neyman–Pearson's Lemma, under certain conditions, the optimal solution to hypothesis testing is the so-called "likelihood ratio test" (LRT). Many researchers in speech technology have introduced the LRT to utterance verification [19], [23] and speaker verification [15]. The LRT technique has achieved significant success in both of these areas. The aim of the research work presented in this paper is to extend the LRT to the new, Bayesian framework.

Generally speaking, automatic speech and speaker recognition aim to solve two different types of problems: *classification* and *verification*. In the classification problem, the objective is to classify an input speech segment or utterance into one of a predefined set of categories $\{C_l | l = 1, \ldots, L\}$ based on the theory of statistical pattern recognition. For a given speech segment X, if the conditional probability $p(X|C_l)$ and the *a prior* probability $p(C_l)$ are assumed known, then the optimal class decision $\hat{C}(X)$ that minimizes the classification error is the Bayes decision rule that maximizes the *a posteriori* probability such that

$$\hat{C}(X) = \arg\max_{l=1}^{L} p(C_l|X) = \arg\max_{l=1}^{L} p(X|C_l) \cdot p(C_l). \quad (1)$$

Speech recognition and speaker identification are two classification problems that have attracted much research effort [13]. The verification problem, on the other hand, generalizes all problems which require a binary answer, and does not involve a predefined set of categories. The verification problem is usually formulated as a problem of "statistical hypothesis testing" [7], where the problem formulation is to test the *null hypothesis* H_0 against the alternative hypothesis H_1. If the probabilities of the null and the alternative hypotheses are known exactly, according to Neyman–Pearson's Lemma, the optimal hypothesis test involves the evaluation of a likelihood ratio such that the null hypothesis, H_0, is accepted if

$$LR = \frac{f(X|H_0)}{f(X|H_1)} > \tau \quad (2)$$

Manuscript received September 20, 2000; revised July 25, 2001. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Philip C. Loizou.

H. Jiang was with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada N2L 3G1. He is now with the Dialogue Systems Research, Multimedia Communication Research Lab, Bell Labs, Lucent Technologies Inc., Murray Hill, NJ 07974 USA (e-mail: hui@research.bell-labs.com).

L. Deng was with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada N2L 3G1. He is now with Microsoft Research, Redmond, WA 98052 USA (e-mail: deng@microsoft.com).

Publisher Item Identifier S 1063-6676(01)09666-3.

where τ is a predefined critical threshold and $f(\cdot|H_0)$ and $f(\cdot|H_1)$ denote the probability distributions under hypotheses H_0 and H_1, respectively.

For testing simple hypotheses where the pdfs of H_0 and H_1 are known exactly, the LRT is known to be the most powerful test for a given level of significance. However, in any practical speech-related verification problem (either speaker verification or utterance verification), it is impossible to obtain the exact pdfs for either the null hypothesis or the alternative hypothesis. Under the practical condition of imprecise pdfs, a feasible strategy is to estimate both $f(X|H_0)$ and $f(X|H_1)$ by assuming a parametric form of the distribution under each hypothesis. Clearly, any assumption of a parametric distribution may cause a mismatch between the "true" and estimated conditional distributions. This possible mismatch, as well as possible estimation errors due to insufficient training data, invalidate the commonly held optimality of the LRT implied by Neyman–Pearson's Lemma. It is this "mismatch" problem that motivates us to search for a superior solution. A solution we found is under the Bayesian framework.

In this paper, we address the verification problem strictly under the Bayesian framework. The Bayesian approach to "hypothesis testing" involves evaluation of the so-called *Bayes factors*. The calculation of the *Bayes factors* in turn requires the computation of several Bayesian predictive densities. Specifically, we propose a Bayesian solution to the general outlier verification problem, where the judgment is made through evaluating the value of *Bayes factors* against some preset threshold. Although the novel Bayesian approach is equally applicable to many practical verification problems, as a case study reported in this paper, it is applied only to speaker verification based on the Gaussian mixture model (GMM). Due to the *missing data* problem in the GMM, we adopt the Viterbi approximation as in [10] to calculate the *Bayes factors* and derive an efficient algorithm to perform speaker verification based on the Bayesian approach. This approach consists of the following three steps:

1) collecting sufficient statistics for all GMMs from all available training data;
2) computing Bayesian predictive densities based on the sufficient statistics only;
3) performing Bayes-factors-based testing. Furthermore, we also study heuristic methods for estimating proper prior pdfs used in speaker verification based on the empirical Bayes method.

In order to examine the viability of the proposed algorithms, they have been evaluated on the *NIST98* speaker recognition evaluation data under the NIST evaluation framework. For several training and testing conditions, the proposed Bayesian approach has been compared with the conventional LRT. The experimental results demonstrate the effectiveness and efficiency of the Bayesian approach. Some moderate improvements over a well-trained baseline system which is employing the conventional LRT have been observed in all training and testing conditions examined.

The remainder of the paper is organized as follows. We briefly introduce the concept of *Bayes factors* in Section II. In Section III, we investigate the outlier verification problem from both non-Bayesian and Bayesian viewpoints and show that the Bayesian approach forms a novel solution to the verification problem. Next, in Section IV, as an example, we apply the Bayesian method to the speaker verification problem and propose an efficient algorithm for the GMM-based speaker verification. Then, we investigate some heuristic methods to estimate prior pdfs for speaker verification, as reported in Section V. Further, the proposed method is evaluated on *NIST98* speaker recognition data and the experimental setup and the results are reported in Section VI. Finally, we conclude the paper with our findings in Section VII.

II. Bayes Factor

The verification problem has traditionally formulated as "statistical hypothesis testing," where two complementary hypotheses, H_0 and H_1, are used, each corresponding to one of the yes and no answers accordingly. Within this traditional, non-Bayesian framework, under some conditions, which are often invalid in practice, the Neyman–Pearson's Lemma would give the optimal test as shown in (2). To remove these impractical conditions, we are interested in investigating a novel solution to the hypothesis testing problem strictly under the Bayesian framework.

As shown in [12], the Bayesian approach to hypothesis testing involves the calculation of the so-called *Bayes factors*. Given the observation data y, the Bayes factors are computed as

$$B = \frac{\hat{p}(y|H_0)}{\hat{p}(y|H_1)} = \frac{\int f(y|\Lambda_0, H_0) \cdot p(\Lambda_0|H_0)\, \mathrm{d}\Lambda_0}{\int f(y|\Lambda_1, H_1) \cdot p(\Lambda_1|H_1)\, \mathrm{d}\Lambda_1} \quad (3)$$

where for

$k = 0, 1$, Λ_k	model parameter under H_k;
$p(\Lambda_k\|H_k)$	prior density;
$p(y\|\Lambda_k, H_k)$	likelihood function of Λ_k under H_k.

Bayes factors offer a way to evaluate evidence in favor of the null hypothesis H_0 because the Bayes factor is the ratio of the posterior odds[1] of H_0 to its prior odds, regardless of the value of the prior odds [12].

Therefore, during testing, Bayes factors can be compared with a preset threshold, much like the likelihood ratio in Neyman–Pearson Lemma, to make a decision with regards to H_0. In other words, if $B > \tau$, where τ is a preset critical threshold, then we accept H_0; otherwise we reject H_0.

III. Outlier Verification

In Section II, we have outlined the Bayesian approach to general hypothesis testing problems. In this section, we will consider a specific example of the outlier verification problem in statistical pattern recognition, and we will investigate the optimal solutions to this problem from the non-Bayesian and Bayesian viewpoints, respectively.

[1]Any probability can be converted to the odds scale, i.e., odds = probability/(1 − probability). Thus, $\Pr(H_0|y)/\Pr(H_1|y)$ is called the posterior odds in favor of H_0, and $\Pr(H_0)/\Pr(H_1)$ is prior odds in favor of H_0.

In a typical problem of statistical pattern recognition, given L classes $\{C_l | l = 1, 2, \ldots, L\}$ and an unknown observation y, we consider categorizing y into one of these classes $C_l (l = 1, 2, \ldots, L)$. If y actually arises from any one of these classes, the optimal solution is given in (1). However, in many practical situations, y sometimes belongs to none of these classes, which is called an *outlier* in this case. The outliers should be rejected in the practical applications (*outlier rejection*). Most statistical pattern recognition techniques do not have an explicit mechanism to determine whether y is an outlier or not. Therefore, a verification process is proposed here to combine with conventional pattern classification methods in order to detect outliers.

We will adopt a two-pass strategy for outlier rejection. In the first phase, which we call *pattern classification*, we first classify y into one most likely class, e.g., C_l $(1 \le l \le L)$. In the second phase, which we call *outlier verification*, we verify whether y actually comes from the class C_l (accept) or y is just an outlier (reject). The first stage is a typical pattern recognition problem which has been well studied. In this section, we will focus on the second stage, namely outlier verification.

In practice, we usually have no knowledge about each class C_l $(1 \le l \le L)$ except we can collect a set of representative samples for each class C_l. Suppose, for any class $C_l (1 \le l \le L)$, we have collected samples $\mathbf{X}_l = \{X_{l1}, X_{l2}, \ldots, X_{lM_l}\}$, where M_l denotes the total number of samples for class C_l. Meanwhile, we also assume another special outlier class C_0, which includes all possible outliers. Usually it is possible to collect another set of samples of outliers: $\mathbf{X}_0 = \{X_{01}, X_{02}, \ldots, X_{0T_0}\}$ for the outlier class C_0.

Based on these assumptions, given an unknown observation y, we first classify it to the most likely class, say C_l $(1 \le l \le L)$, by using the conventional pattern classification method. The problem here is to verify whether the observation y actually comes from the class C_l or not. In general, this can be formulated as a problem of hypothesis testing. By explicitly taking into account the information available in the training data $\mathbf{X}_l$ and $\mathbf{X}_0$, the *null* hypothesis, which corresponds to the case where y is not an outlier, is proposed as $\mathbf{H_0}$: Both $\mathbf{X}_l$ and y come from C_l while $\mathbf{X}_0$ comes from C_0. Accordingly, the alternative hypothesis is $\mathbf{H_1}$: $\mathbf{X}_l$ comes from C_l while both $\mathbf{X}_0$ and y come from C_0. The solution to this hypothesis testing problem is as follows.

If

$$\eta = \frac{p(\mathbf{X}_l, y, \mathbf{X}_0 | H_0)}{p(\mathbf{X}_l, y, \mathbf{X}_0 | H_1)} = \frac{p(\mathbf{X}_l, y | C_l) \cdot p(\mathbf{X}_0 | C_0)}{p(\mathbf{X}_l | C_l) \cdot p(\mathbf{X}_0, y | C_0)} > \xi \quad (4)$$

then we accept y; otherwise, reject y as an outlier, where ξ is a fixed threshold, and $p(\cdot|\cdot)$ denotes the conditional probability density.

In the following, we will study this outlier verification problem from both non-Bayesian and Bayesian viewpoints. Two different solutions will be derived from these different viewpoints.

A. Non-Bayesian Method (Neyman–Pearson Approach)

In non-Bayesian statistics, we usually assume that the form of a parametric model for each class Ω_l is known in advance, leaving its parameters Λ_l to be estimated from the training samples $\mathbf{X}_l$. Thus, parameters Λ_l are viewed as an estimator from the samples $\mathbf{X}_l$ under the non-Bayesian framework. Because likelihood functions play a central role in non-Bayesian statistics, all probability densities in (4) are viewed as the corresponding likelihood functions of model parameters. If all data are independent, the hypothesis testing problem expressed in (4) reduces to the conventional LRT in Neyman–Pearson Lemma

$$\eta = \frac{f(\mathbf{X}_l | \Lambda_l) \cdot f(y | \Lambda_l) \cdot f(\mathbf{X}_0 | \Lambda_0)}{f(\mathbf{X}_l | \Lambda_l) \cdot f(y | \Lambda_0) \cdot f(\mathbf{X}_0 | \Lambda_0)} = \frac{f(y | \Lambda_l)}{f(y | \Lambda_0)} > \xi \quad (5)$$

where $f(\cdot|\cdot)$ denotes the likelihood function, and Λ_l and Λ_0 are the estimators of the model parameters for the class C_l and C_0, based on the samples $\mathbf{X}_l$ and $\mathbf{X}_0$, respectively.

Therefore, the non-Bayesian method for outlier verification consists of two separate steps. First, we estimate the model parameters for a certain data class and for the outlier class. Second, verification is performed as an LRT based on the estimated model parameters.

B. Bayesian Approach

In the Bayesian framework, all model parameters Λ_l $(l = 0, 1, \ldots, L)$ are treated as random variables, and we assume that all knowledge about each class C_l $(l = 0, 1, \ldots, L)$ is contained in a prior pdf $\rho_l(\Lambda_l)$ $(l = 0, 1, \ldots, L)$.

According to Section II, the Bayesian solution to the hypothesis testing of (4) involves the computation of the Bayes factor

$$\begin{aligned} B &= \frac{\hat{p}(\mathbf{X}_l, y, \mathbf{X}_0 | H_0)}{\hat{p}(\mathbf{X}_l, y, \mathbf{X}_0 | H_1)} = \frac{\hat{p}_l(\mathbf{X}_l, y | C_l) \cdot \hat{p}_0(\mathbf{X}_0 | C_0)}{\hat{p}_l(\mathbf{X}_l | C_l) \cdot \hat{p}_0(\mathbf{X}_0, y | C_0)} \\ &= \frac{\int f(\mathbf{X}_l, y | \Lambda_l) \cdot \rho_l(\Lambda_l)\, d\Lambda_l \cdot \int f(\mathbf{X}_0 | \Lambda_0) \cdot \rho_0(\Lambda_0)\, d\Lambda_0}{\int f(\mathbf{X}_l | \Lambda_l) \cdot \rho_l(\Lambda_l)\, d\Lambda_l \cdot \int f(\mathbf{X}_0, y | \Lambda_0) \cdot \rho_0(\Lambda_0)\, d\Lambda_0} \end{aligned} \quad (6)$$

where

- $\hat{p}_l(\cdot)$ joint Bayesian predictive density for class C_l;
- $\rho(\cdot)$ prior pdf;
- $f(\cdot|\cdot)$ likelihood function.

Differing from the non-Bayesian approach, in the current Bayesian approach we first estimate a prior pdf for each class. Then the decision is made based on the value of this *Bayes factor*. As for the prior pdf estimation, we adopt the principle of partial Bayes factors [18]. Concretely, the initial prior pdf is set as a noninformative prior pdf at the beginning. Then, a portion of the samples are used to estimate the prior pdf based on Bayesian learning or according to the empirical Bayes method. Finally, the Bayes factor is calculated based on the remaining samples.

As a remark, if we adopt the following incremental method to calculate the joint predictive density

$$\hat{p}_l(\mathbf{X}_l | \Lambda_l) = \prod_{i=1}^{M_l} \int f(X_{li} | \Lambda_l) \cdot p\left(\Lambda_l \left| \mathbf{X}_l^{(i-1)} \right.\right) d\Lambda_l \quad (7)$$

where $\mathbf{X}_l^{(i-1)} = \{X_{l1}, X_{l2}, \ldots, X_{li-1}\}$ and the prior/posterior pdf $p(\Lambda_l|\mathbf{X}_l^{(i-1)})$ is calculated based on incremental Bayesian learning at each step, then this hypothesis testing in (6) can be significantly simplified to

$$\eta = \frac{\int f(y|\Lambda_l) \cdot p(\Lambda_l|\mathbf{X}_l)\, \mathrm{d}\Lambda_l}{\int f(y|\Lambda_0) \cdot p(\Lambda_0|\mathbf{X}_0)\, \mathrm{d}\Lambda_0} \tag{8}$$

where $p(\Lambda_l|\mathbf{X}_l)$ is the *posteriori* density after observing data $\mathbf{X}_l$, which can be obtained from the Bayesian learning method.

IV. Bayesian Approach to GMM-Based Speaker Verification

We have presented the general formulation of Bayesian approach to the verification problem. As a specific case, in this section, the Bayesian approach is applied to text-independent speaker verification based on the GMM. We also derive an efficient algorithm to perform Bayes-factors-based hypothesis testing for the GMM. Note that, although the algorithm in this section is derived for the GMM, it is straightforward to extend it to the HMM case.

The typical scenario in speaker verification looks as follows: User first claims an identity and the system prompts the user to say some utterances in order to verify whether the user actually is the claimed identity or not. Obviously, speaker verification can be viewed as an outlier verification problem, where in every verification step, the claimed speaker identity is the target speaker and all other speakers are thought to be the outliers (or imposters). When building the system, each target speaker is required to provide some speech data to construct some models for the speaker. In verification, the new utterances are used to match with the target speaker's model in order to make the verification decision. It is well known that GMM and/or HMM have become the most effective models for speaker identification and speaker verification. In this paper, we adopt GMM as the basic model for speaker verification, i.e., each speaker is represented by one GMM. We also use another GMM model to represent outliers for all nontarget speakers.

Suppose we have L target speakers in the pool, and a total of $L+1$ GMM models, denoted as $\{\Lambda^{(l)}|l = 0, 1, \ldots, L\}$, where $\Lambda^{(l)}$ $(l > 0)$ represents the lth speaker model and $\Lambda^{(0)}$ the outlier model. Given a feature vector x, the likelihood function of GMM $\Lambda^{(l)}$ can be expressed as

$$f\left(x\middle|\Lambda^{(l)}\right) = \sum_{i=1}^{N} w_i^{(l)} \cdot \mathcal{N}\left(x\middle|m_i^{(l)}, r_i^{(l)}\right) \tag{9}$$

where

N	total number of Gaussians in the model;
$w_i^{(l)}$	weight of ith mixture component with the constraint $\sum_{i=1}^{N} w_i^{(l)} = 1$;
$\mathcal{N}(x\|m_i^{(l)}, r_i^{(l)})$	ith multivariate Gaussian mixture component with the mean vector $m_i^{(l)}$ and the precision matrix $r_i^{(l)}$.

If we assume all precision matrices are diagonal, we have

$$\mathcal{N}\left(x\middle|m_i^{(l)}, r_i^{(l)}\right) = \prod_{d=1}^{D} \sqrt{\frac{r_{id}^{(l)}}{2\pi}}\, e^{-(r_{id}^{(l)}/2)(x_d - m_{id}^{(l)})^2} \quad (1 \le i \le N) \tag{10}$$

where D is the dimension of the vector. Hereafter, for conciseness, we will omit all superscripts (l) when no confusion occurs.

When we observe an utterance $X = \{x_1, x_2, \ldots, x_T\}$ with length T, the likelihood function of GMM Λ can be expressed as

$$f(X|\Lambda) = \sum_{\mathcal{P}} f(X, \mathcal{P}|\Lambda) \tag{11}$$

where $\mathcal{P}$ is called a path (over GMM Λ), which is a sequence of mixture component labels: $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_T\}$. This summation is carried out over the entire path space.

In this paper, among all of the GMM parameters, we assume, for simplicity purposes, that only Gaussian mean vectors are random while the remaining parameters are deterministic and known. We further adopt the idea of the natural conjugate prior [4], and choose the prior pdfs for each GMM Λ according to

$$\begin{aligned}\rho(\Lambda) &= \rho(m_1, m_2, \ldots, m_N) = \prod_{i=1}^{N} \mathcal{N}(m_i|\mu_i, \tau_i) \\ &= \prod_{i=1}^{N}\prod_{d=1}^{D} \sqrt{\frac{\tau_{id}}{2\pi}}\, e^{-(\tau_{id}/2)(m_{id} - \mu_{id})^2}\end{aligned} \tag{12}$$

where $\{\mu_{id}, \tau_{id} | 1 \le i \le N, 1 \le d \le D\}$ denote all hyperparameters related to the GMM Λ.

The Bayesian approach to hypothesis testing involves the calculation of Bayes factors. The computation of the Bayes factors in turn requires the computation of several Bayesian predictive densities, and this latter computation is not straightforward in many cases. As for the GMM, due to its *missing-data* nature, some approximations are needed to calculate the Bayesian predictive value in a feasible way. Here we adopt the same Viterbi approximation to Bayesian prediction as that used in [10] and [11].

Given an utterance $X = \{x_1, x_2, \ldots, x_T\}$, the Bayesian predictive density $\hat{p}(X)$ for model Λ is computed as follows:

$$\begin{aligned}\hat{p}(\mathbf{X}) &= \int f(\mathbf{X}|\Lambda) \cdot \rho(\Lambda)\, \mathrm{d}\Lambda \\ &= \int \sum_{\mathcal{P}} f(\mathbf{X}, \mathcal{P}|\Lambda) \cdot \rho(\Lambda)\, \mathrm{d}\Lambda \\ &= \sum_{\mathcal{P}} \int f(\mathbf{X}, \mathcal{P}|\Lambda) \cdot \rho(\Lambda)\, \mathrm{d}\Lambda \\ &\approx \max_{\mathcal{P}} \int f(\mathbf{X}, \mathcal{P}|\Lambda) \cdot \rho(\Lambda)\, \mathrm{d}\Lambda.\end{aligned} \tag{13}$$

We term the path which maximizes this integration as the *optimal path*, denoted as $\mathcal{P}^* = \{\mathcal{P}_1^*, \mathcal{P}_2^*, \ldots, \mathcal{P}_T^*\}$, i.e.,

$$\mathcal{P}^* = \arg\max_{\mathcal{P}} \int f(\mathbf{X}, \mathcal{P}|\Lambda) \cdot \rho(\Lambda)\, \mathrm{d}\Lambda. \tag{14}$$

Thus, the approximate Bayesian predictive density $\hat{p}(\mathbf{X})$ can be expressed as

$$\begin{aligned}\hat{p}(\mathbf{X}) &\approx \int f(X, \mathcal{P}^*|\Lambda) \cdot \rho(\Lambda)\, \mathrm{d}\Lambda \\ &= \int \prod_{t=1}^{T} w_{\mathcal{P}_t^*} \cdot \mathcal{N}(x_t|m_{\mathcal{P}_t^*}, r_{\mathcal{P}_t^*}) \cdot \rho(\Lambda)\, \mathrm{d}\Lambda \\ &= \prod_{i=1}^{N} w_i^{v_i} \prod_{d=1}^{D} \frac{\tau_{id}^{1/2} \cdot r_{id}^{v_i/2}}{(2\pi)^{v_i/2} \cdot (\tau_{id}+r_{id}v_i)^{1/2}} \\ &\quad \cdot \exp\left\{-\frac{r_{id}v_i}{2(\tau_{id}+r_{id}v_i)}\left[\tau_{id}\overline{(x-\mu)_{id}^2}+r_{id}v_i(\overline{x_{id}^2}-\overline{x}_{id}^2)\right]\right\}\end{aligned} \tag{15}$$

where

$$v_i = \sum_{t=1}^{T} \delta(\mathcal{P}_t^* - i) \tag{16}$$

$$\overline{(x-\mu)_i^2} = \frac{\sum_{t=1}^{T}(x_t-\mu_i)^2 \cdot \delta(\mathcal{P}_t^* - i)}{\sum_{t=1}^{T}\delta(\mathcal{P}_t^* - i)} \tag{17}$$

$$\overline{x}_i = \frac{\sum_{t=1}^{T} x_t \cdot \delta(\mathcal{P}_t^* - i)}{\sum_{t=1}^{T}\delta(\mathcal{P}_t^* - i)} \tag{18}$$

$$\overline{x_i^2} = \frac{\sum_{t=1}^{T} x_t^2 \cdot \delta(\mathcal{P}_t^* - i)}{\sum_{t=1}^{T}\delta(\mathcal{P}_t^* - i)}. \tag{19}$$

Here, we have obtained the result for the Viterbi approximation to the Bayesian predictive density for a single observation X. It is straightforward to extend the method to compute the joint Bayesian predictive density for multiple observations $\mathbf{X}$, as required in (6). From (6), we see that the brute-force calculation of the Bayes factor requires all training data $\mathbf{X}$. This means that we have to save all training data in order to verify any unknown utterance y with the Bayesian method. This obviously is not practical for most applications. In the following, we derive an efficient algorithm to calculate *Bayes factors* for multiple observations $\mathbf{X}$, which does not require using all training data in each verification step.

In this algorithm we first collect the so-called *sufficient statistics* for all models based on all available training data $\mathbf{X}$. Then the sufficient statistics, as well as all prior pdfs, are stored in order to carry out verification. When presented with an unknown utterance, the Bayes factor can be calculated strictly from these statistics only, without using the original data. Below, we first present an algorithm to calculate the joint Bayesian predictive value. Then, we present the Bayesian approach to verification based on these joint Bayesian predictive values.

Given a set of (training) observations $\mathbf{X} = \{X_1, X_2, \ldots, X_M\}$, a testing observation Y, and the prior pdf $\rho(\Lambda)$ for the model parameters Λ, the algorithm for calculating the joint Bayesian predictive densities $\hat{p}(\mathbf{X})$ and $\hat{p}(\mathbf{X}, Y)$ is presented as steps A and B in the following.

A. Collect Sufficient Statistics

- Attach a set of "global" statistics to each Gaussian mixture component in the model Λ. For the ith component $\mathcal{N}(m_i, r_i)$, its "global" statistics consists of a scalar Υ_i, and three vectors denoted as $\overline{\mathbf{X}}_i$, $\overline{X^2}_i$, and $\overline{(X-\mu)^2}_i$, respectively.
- Initialize all "global" statistics

$$\overline{\mathbf{X}}_i = \overline{X^2}_i = \overline{(X-\mu)^2}_i = \Upsilon_i = 0 \qquad (1 \le i \le N). \tag{20}$$

- For each observation $X_s = \{x_{s1}, x_{s2}, \ldots, x_{sT_s}\}$ $(1 \le s \le M)$ in $\mathbf{X}$ we do the following.
  - — Perform the Viterbi Bayesian predictive classification (VBPC) search [10] for X_s, based on $\rho(\Lambda)$, and obtain one optimal path $\mathcal{P}_s^*$ over the model Λ.
  - — Based on the path $\mathcal{P}_s^*$, collect statistics for each Gaussian mixture $\mathcal{N}(m_i, r_i)$ $(i = 1, 2, \ldots, N)$. Assuming its related prior pdf to be $\rho(m_i) = \mathcal{N}(\mu_i, \tau_i)$ as in (12), we collect the following "local" statistics for all mixture components. For $i = 1, 2, \ldots, N$:

$$v_i = \sum_{t=1}^{T_s} \delta(\mathcal{P}_{st}^* - i) \tag{21}$$

$$\overline{x}_i = \frac{1}{v_i}\sum_{t=1}^{T_s} x_{st} \cdot \delta(\mathcal{P}_{st} - i) \tag{22}$$

$$\overline{x^2}_i = \frac{1}{v_i}\sum_{t=1}^{T_s} x_{st}^2 \cdot \delta(\mathcal{P}_{st} - i) \tag{23}$$

$$\overline{(x-\mu)^2}_i = \frac{1}{v_i}\sum_{t=1}^{T_s}(x_{st}-\mu_i)^2 \cdot \delta(\mathcal{P}_{st} - i) \tag{24}$$

where

$$\delta(\mathcal{P}_{st} - i) = \begin{cases} 1, & \text{if path } \mathcal{P}_l \text{ lies in the mixture component } i \text{ at time } t \\ 0, & \text{otherwise.} \end{cases} \tag{25}$$

  - — Use these "local" statistics to update all "global" statistics as follows: for all $i = 1, 2, \ldots, N$

$$\overline{\mathbf{X}}_i \Leftarrow \frac{\overline{\mathbf{X}}_i \cdot \Upsilon_i + \overline{x}_i \cdot v_i}{\Upsilon_i + v_i} \tag{26}$$

$$\overline{X^2}_i \Leftarrow \frac{\overline{X^2}_i \cdot \Upsilon_i + \overline{x^2}_i \cdot v_i}{\Upsilon_i + v_i} \tag{27}$$

$$\overline{(X-\mu)^2}_i \Leftarrow \frac{\overline{(X-\mu)^2}_i \cdot \Upsilon_i + \overline{(x-\mu)^2}_i \cdot v_i}{\Upsilon_i + v_i} \tag{28}$$

$$\Upsilon_i \Leftarrow \Upsilon_i + v_i. \tag{29}$$

B. Calculate Joint Bayesian Prediction Using Sufficient Statistics

- Calculate the Bayesian predictive value $\hat{p}(\mathbf{X})$ according to

$$\hat{p}(\mathbf{X}) \approx \prod_{i=1}^{N} w_i^{\Upsilon_i} \prod_{d=1}^{D} \frac{\tau_{id}^{1/2} \cdot r_{id}^{\Upsilon_i/2}}{(2\pi)^{\Upsilon_i/2} \cdot (\tau_{id}+r_{id}\Upsilon_i)^{1/2}} \cdot \exp\left\{ -\frac{r_{id}\Upsilon_i}{2(\tau_{id}+r_{id}\Upsilon_i)} \cdot \left[\tau_{id}\overline{(X-\mu)^2}_{id} + r_{id}\Upsilon_i(\overline{X^2}_{id} - \overline{\mathbf{X}}_{id}^2) \right] \right\}. \quad (30)$$

- Calculate the Bayesian predictive value $\hat{p}(\mathbf{X}, Y)$:
  - — Keep all "global" statistics related to $\mathbf{X}$ as collected from the previous step.
  - — Collect the "local" statistics for Y alone as in (21)–(24), which are denoted as $\{\upsilon_i, \overline{y}_i, \overline{y^2}_i, \overline{(y-\mu)^2}_i | 1 \le i \le N\}$.
  - — Calculate $\hat{p}(\mathbf{X}, Y)$ as shown in (31) at the bottom of the page.

C. Hypothesis Testing

Let the *null* hypothesis be that Y comes from the lth speaker's model $\Lambda^{(l)}$ and let the alternative hypothesis be that Y is an outlier ($\Lambda^{(0)}$). The Bayesian solution to such a hypothesis testing problem is then based on the following Bayes factor:

If

$$\eta = \frac{\hat{p}_l(\mathbf{X}_l, Y) \cdot \hat{p}_0(\mathbf{X}_0)}{\hat{p}_l(\mathbf{X}_l) \cdot \hat{p}_0(\mathbf{X}_0, Y)} > \xi \quad (32)$$

then we accept the claimed speaker identity, otherwise reject it as an imposter. If the training data $\mathbf{X}_0$ and $\mathbf{X}_l$ are unchanged during the verification procedure, given any Y, this test can be simplified to

$$\eta = \frac{\hat{p}_l(\mathbf{X}_l, Y)}{\hat{p}_0(\mathbf{X}_0, Y)} > \xi' \quad (33)$$

where ξ' is a new threshold.

V. Prior Estimation for Speaker Verification

In this section, we consider estimating the prior pdfs for the calculation of Bayes factors in speaker verification. We assume that all priors have the functional form as shown in (12) and that their hyperparameters be estimated from the data. Here we propose to use the empirical Bayes method to estimate a speaker-independent (SI) prior pdf as the initial step. Then the data from a specific speaker are used to update the SI priors to obtain speaker-dependent (SD) priors for each speaker based on Bayesian learning.

Given a total of L speakers in a system pool, the collected data $\mathbf{X}_l$ $(1 \le l \le L)$ correspond to each target speaker C_l $(1 \le l \le L)$ and the data $\mathbf{X}_0$ represent all other nontarget speakers.

A. Estimate SI Prior Based on Empirical Bayes Method

First, all available data $\mathbf{X}_l$ $(l = 0, 1, 2, \ldots, L)$ are used to estimate a speaker-independent GMM model $\Lambda^{(SI)}$ based on the expectation-maximization (EM) method (see the Appendix for details). $\Lambda^{(SI)}$ consists of $\{w_i^{(SI)}, m_i^{(SI)}, r_i^{(SI)} | 1 \le i \le N\}$. Then, we construct the SI prior $\overline{\rho}(\Lambda)$ from the SI model $\Lambda^{(SI)}$ based on the empirical Bayes method which proceeds as follows. Assuming that the SI prior $\overline{\rho}(\Lambda)$ has the functional form of (12), then its hyperparameters are estimated as follows:

$$\overline{\mu}_{id} = m_{id}^{(SI)} \qquad (1 \le i \le N \text{ and } 1 \le d \le D) \quad (34)$$

$$\overline{\tau}_{id} = \epsilon \cdot r_{id}^{(SI)} \cdot c_i \qquad (1 \le i \le N \text{ and } 1 \le d \le D) \quad (35)$$

where $\epsilon > 0$ is a weighting coefficient for adjusting the shape of the prior pdf, and c_i is a weight count accumulated for the ith mixture component during the training procedure for the SI GMM $\Lambda^{(SI)}$, i.e., $c_i = \sum_{l=0}^{L} \sum_{s=1}^{M} \sum_{t=1}^{T_s} \xi_{lst}(i)$, where $\xi_{lst}(i)$ is defined in the Appendix.

B. Estimate SD Priors Based on Bayesian Learning

For every individual speaker $l = 1, \ldots, L$, the data $\mathbf{X}_l$ from that speaker are used to update the SI prior $\overline{\rho}(\Lambda)$ to obtain an SD priors based on Bayesian learning. In other words, we treat SI prior $\overline{\rho}(\Lambda)$ as the prior pdf in Bayesian learning, and the derived posterior pdf becomes the SD prior pdf (which is needed for the calculation of the Bayes factors). This gives

$$\rho_l(\Lambda) \propto \overline{\rho}(\Lambda) \cdot f(\mathbf{X}_l | \Lambda) \quad (36)$$

where $f(\mathbf{X}_l | \Lambda)$ is the likelihood function.

Due to the *missing data* problem in GMM, it is not possible to implement this Bayesian learning accurately [9]. Some approximations are needed. In this paper, we adopt the Viterbi approximation to derive the so-called segmental Bayesian learning as in [8], [9]. This approximation gives

$$\rho_l(\Lambda) \propto \overline{\rho}(\Lambda) \cdot \sum_{\mathcal{P}} f(\mathbf{X}_l, \mathcal{P} | \Lambda) \approx \overline{\rho}(\Lambda) \cdot f(\mathbf{X}_l, \mathcal{P}^* | \Lambda) \quad (37)$$

where $\mathcal{P}^*$ is the optimal path as defined in (14).

$$\hat{p}(\mathbf{X}, Y) \approx \prod_{i=1}^{N} w_i^{\Upsilon_i+\upsilon_i} \prod_{d=1}^{D} \frac{\tau_{id}^{1/2} \cdot r_{id}^{(\Upsilon_i+\upsilon_i)/2}}{(2\pi)^{(\Upsilon_i+\upsilon_i)/2} \cdot [\tau_{id} + (\upsilon_i + \Upsilon_i) r_{id}]^{1/2}} \cdot \exp\left\{ -\frac{r_{id}\tau_{id}\Upsilon_i\overline{(X-\mu)^2}_{id} + r_{id}\tau_{id}\upsilon_i\overline{(y-\mu)^2}_{id} + r_{id}^2(\Upsilon_i+\upsilon_i)(\Upsilon_i\overline{X^2}_{id} + \upsilon_i\overline{y^2}_{id}) - r_{id}^2(\Upsilon_i\overline{\mathbf{X}}_{id} + \upsilon_i\overline{y}_{id})^2}{2[\tau_{id} + (\upsilon_i + \Upsilon_i) r_{id}]} \right\} \quad (31)$$

Denoting $\mathbf{X}_l = \{X_1, X_2, \ldots, X_M\}$, and for each s $(s = 1, \ldots, M)$, $X_s = \{x_{s1}, x_{s2}, \ldots, x_{sT_s}\}$, where T_s is the length of X_s. Because we choose the nature conjugate prior for $\overline{p}(\Lambda)$ as in (12), the SD prior $\rho_l(\Lambda)$ will have the same form as that in the right hand side of (12). Then Bayesian learning gives the estimated hyperparameters of

$$\mu'_{id} = \frac{\overline{\mu}_{id}\overline{\tau}_{id} + \mu^*_{id}\tau^*_{id}}{\overline{\tau}_{id} + \tau^*_{id}} \qquad (1 \le i \le N \text{ and } 1 \le d \le D) \tag{38}$$

$$\tau'_{id} = \overline{\tau}_{id} + \tau^*_{id} \qquad (1 \le i \le N \text{ and } 1 \le d \le D) \tag{39}$$

where

$$\mu^*_{id} = \frac{\sum_{s=1}^{M}\sum_{t=1}^{T_s} x_{std} \cdot \delta(\mathcal{P}^*_{st} - i)}{\sum_{s=1}^{M}\sum_{t=1}^{T_s} \delta(\mathcal{P}^*_{st} - i)} \qquad (1 \le i \le N \text{ and } 1 \le d \le D) \tag{40}$$

$$\tau^*_{id} = r_{id}\sum_{s=1}^{M}\sum_{t=1}^{T_s} \delta(\mathcal{P}^*_{st} - i) \qquad (1 \le i \le N \text{ and } 1 \le d \le D) \tag{41}$$

and $\delta(\cdot)$ is the Kronecker delta function.

C. Priors for Outliers (Nontarget Speakers)

In speaker verification, a simplest strategy which we have implemented is to use the SI prior pdf as an approximation to the outlier prior pdf. A second method which we have implemented is to collect all data which represent all possible outliers (imposters). In speaker verification, this data would include $\{\mathbf{X}_l | 0 \le l \le L, l \ne l^t\}$, where l^t is the target speaker. Then, the empirical Bayes method described in Section V-A can be used to construct the outlier prior directly from the data, or the previous data can be used to update the SI priors to obtain the outlier prior pdf based on Bayesian learning.

VI. Speaker Verification Experiments

In order to examine the viability of the proposed new approach, we have applied it to some speaker verification tasks. In this paper, we report the experimental results from the *NIST98* speaker recognition evaluation data. In our experiments, the Bayesian approach was compared with the baseline system which uses the conventional LRT. Experimental results demonstrate the effectiveness and efficiency of the Bayesian approach in speaker verification based on GMM. Under all the training and testing conditions we have examined, the Bayesian approach has achieved moderate but consistent improvements over the well-trained baseline system.

A. Database and Evaluation Method

The *NIST98* evaluation data is used in all of our experiments. The *NIST98* data is drawn from the SwitchBoard-2 phase-2 corpus, which includes a large amount of spontaneous conversations between two speakers over telephone lines. There are 250 female and 250 male speakers in *NIST98*. In our experiments, only those data from 250 female speakers are used. These speakers serve both as target speakers and nontarget (impostor) speakers. The handset type label is supplied by NIST but we did not use this information in our experiments. In the experiments, we use two different training conditions for each target speaker:

- "**One-session**" training (denoted as **s1**): The training data are two minutes of the speech data taken from only one conversation session.
- "**Two-session**" training (denoted as **s2**) : Equal amounts (two minutes) of training data are taken from two different conversations collected from the same phone number. The training data consist of one-minute data from one conversation, plus additional one-minute data from a different session, with the same phone number.

Our experiments include three different testing conditions according to different test speech segment durations. These durations are approximately 3 s, 10 s, and 30 s, which are denoted as **T3**, **T10**, and **T30**, respectively.

The *NIST98* data include both training and testing sets. There is an average of about 10 test segments for each target speaker and for each test duration. This makes up a total of 2500 test segments for each of the three test durations. For each of the test segments, a total of ten speaker identities are assigned as test hypotheses. Each of these hypotheses is then required to be judged as true or false, and a decision score is also needed for each judgment.

B. Baseline System

In both training and testing, the speech waveforms are digitized and preprocessed by a front-end unit that extracts a set of 12 mel-frequency cepstral coefficients (MFCC) and log energy from each frame of data. Cepstral mean normalization is applied to each utterance to eliminate some of the spectral shaping occurring in the telephone channel. Each 39-dimension feature vector consists of 12 MFCCs and log-energy, as well as their delta and delta–delta coefficients.

In our baseline system, a single GMM is built for each speaker. The number of Gaussians, N, is the same for all speakers, and is fixed as 256 in all of our experiments. The construction of the SD GMMs for all speakers is made up of two separate phases. First, we use all available training data from all speakers to train an SI GMM based on maximum likelihood [16]. The generation of GMM starts with a random vector quantization (VQ) codebook. The codebook is then iteratively updated using the LBG algorithm. After convergence, the Gaussians are fitted to the codewords, and their parameters are adjusted using a few EM iterations (see the Appendix for details). Second, starting from the SI GMM, the SD GMM for each individual speaker is estimated from the specific speaker's speech data based on maximum *a posteriori* (MAP) estimation [8].

When presented with a speech segment X and a specific speaker, the verification decision in this baseline system is made based on the conventional LRT, i.e., if

$$LR = \frac{f\left(X|\Lambda^{(SD)}\right)}{f\left(X|\Lambda^{(SI)}\right)} > \xi \tag{42}$$

then we accept the speaker identity, and otherwise reject it; $f(\cdot|\cdot)$ denotes the likelihood function and ξ is the preset critical threshold.

C. Bayesian Approach to Speaker Verification

We use identical feature vectors and the GMM structure in the new Bayesian approach to those in the baseline system. A speaker verification system based on the Bayesian approach is established similarly as the baseline system. First, an SI prior pdf is constructed from all available data from all speakers based on the empirical Bayes method described in Section V-A. Here we set the control parameter $\epsilon = 1/\mathcal{M}$, where $\mathcal{M}$ denotes the total number of utterances in the training set. Second, based on the Bayesian learning method described in Section V-B, the SI prior pdf is updated by using the speech data from each individual speaker to obtain an SD prior pdf for that speaker. The SI prior pdf is used as the prior pdf for outliers (nontarget speakers) in our experiments. Third, for each speaker, in addition to its SD prior pdf, we also collect a set of "sufficient statistics" based on the training speech data from the speaker as described in Section IV-A. As for the outlier class, the SI prior pdf is used as its prior pdf and a corresponding set of "sufficient statistics" is similarly collected from all available data of all speakers.

When given a test speech segment and a speaker identity, without using all original data set, the verification decision is based on Bayes factors which are easily computed from the corresponding prior pdfs and the sufficient statistics.

D. Experimental Results and Discussions

In the experiments, for each testing condition we randomly selected 500 test segments out of the total 2500 segments which NIST supplies for the 1998 evaluation. For each test segment, ten speaker identities are provided as hypotheses. Thus, for each testing condition, we have a total of 5000 pairs of test speech segments and hypothesized speakers which require 5000 verification decisions. We classify these 5000 pairs into two categories: the first class is called "*in-set*" where the speech segment is actually uttered by the hypothesized speaker. The other class is called "*out-set*" where the speech segment is not uttered by the hypothesized speaker, who is an imposter. We count the false rejection error in "*in-set*" and the false acceptance error in "*out-set*." The comparative results of equal error rate (EER) between the new Bayesian method and the baseline system are given in Table I.

From these results, we observe that the **s2** training condition generally gives better performance than **s1**. The main reason is that the enrollment data comes from different conversation sessions under **s2** training condition, which may include more information about the telephone channel variations. In addition, for the same training condition, the longer test segment duration gives better verification performance. We also observe that by use of the novel Bayesian approach, moderate improvements have been achieved in the EER over the baseline system for all the training and test conditions we have examined as listed in Table I.

The next set of results are obtained when we change the critical threshold continuously, and then draw the verification ROC curve. Each point in the ROC curve represents a certain value of the threshold, which is also called an operating point. The ROC curves obtained for all the experiments we have conducted are shown in Figs. 1–6.

TABLE I
COMPARATIVE RESULTS OF EQUAL ERROR RATE (EER) BETWEEN BAYESIAN APPROACH (**BAYES**) AND THE LIKELIHOOD-RATIO-TEST-BASED BASELINE SYSTEM (**LRT**)

training condition	test condition	**LRT**	**Bayes**
s1	T3	23.2%	22.0%
s1	T10	18.8%	17.6%
s1	T30	16.0%	14.6%
s2	T3	21.6%	20.8%
s2	T10	19.0%	17.2%
s2	T30	16.0%	14.4%

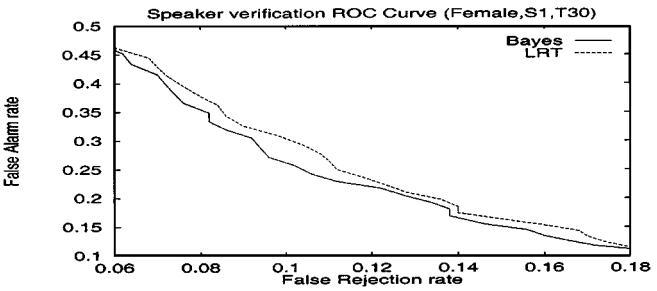

Fig. 1. Comparison of speaker verification ROC curves for female speakers under training condition **s1** and test condition **T30**. Dotted-line: likelihood ratio test (LRT), solid-line: Bayesian approach (Bayes).

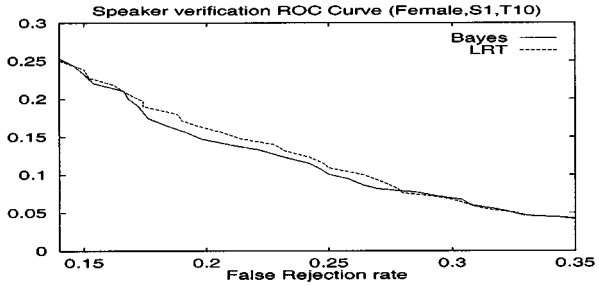

Fig. 2. Comparison of speaker verification ROC curves for female speakers under training condition **s1** and test condition **T10**. Dotted-line: likelihood ratio test (LRT), solid-line: Bayesian approach (Bayes).

From these ROC curves, we observe that generally the Bayesian approach is superior to the baseline system under all examined conditions. In particular, the advantage of the Bayesian approach is larger for the training condition of **s1** than that of **s2**. The possible reason is that the GMMs estimated in the condition of **s1** are relatively poor due to the telephone channel variations in the speech data, and it is well known that the Bayesian approach usually achieves larger gains than the likelihood-based non-Bayesian method when the estimated model is not sufficiently accurate.

In general, the advantage of the Bayesian approach lies in the following two aspects. First, the "mismatch" problem caused by incorrect model assumptions and by insufficient training data often invalidates the optimality of the LRT procedure implied by Neyman–Pearson's Lemma. The Bayesian approach is much

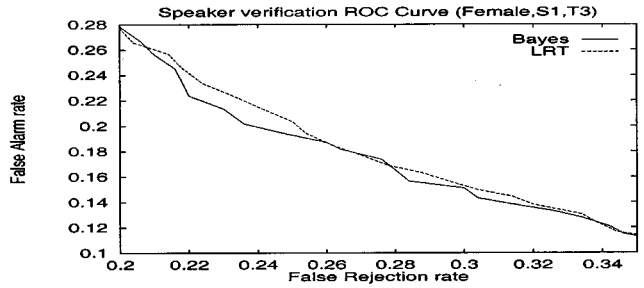

Fig. 3. Comparison of speaker verification ROC curves for female speakers under training condition **s1** and test condition **T3**. Dotted-line: likelihood ratio test (LRT), solid-line: Bayesian approach (Bayes).

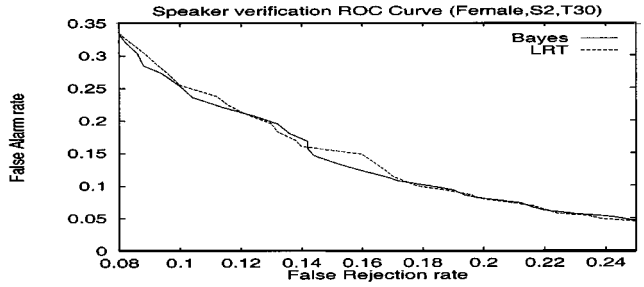

Fig. 4. Comparison of speaker verification ROC curves for female speakers under training condition **s2** and test condition **T30**. Dotted-line: likelihood ratio test (LRT), solid-line: Bayesian approach (Bayes).

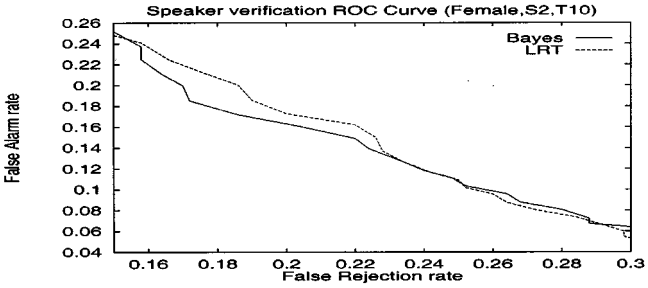

Fig. 5. Comparison of speaker verification ROC curves for female speakers under training condition **s2** and test condition **T10**. Dotted-line: likelihood ratio test (LRT), solid-line: Bayesian approach (Bayes).

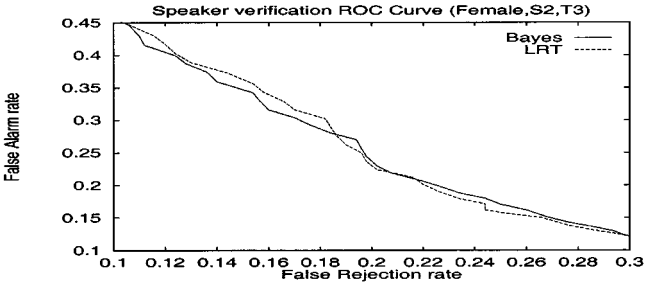

Fig. 6. Comparison of speaker verification ROC curves for female speakers under training condition **s2** and test condition **T3**. Dotted-line: likelihood ratio test (LRT), solid-line: Bayesian approach (Bayes).

less sensitive to model inaccuracy because the model parameters are integrated over the entire space based on the prior pdf. Second, as shown in (30), the joint Bayesian predictive log value $\ln \hat{p}(\mathbf{X})$ can be decomposed into two terms, namely $\tau_{id}\overline{(X-\mu)_i^2}$ and $r_{id}\Upsilon_i(\overline{X^2}_{id}-\overline{\mathbf{X}}^2_{id})$. The first term represents the information from the model and the second term represents the information obtained directly from the data. The Bayesian approach optimally combines the two different sources of information in making the verification decision. The optimal balance between these two sources is automatically adjusted by the prior pdf; that is, the decision is made to rely more on the data if the model is less accurate and more on the model if the model is more precise. Both aspects of the strength of the Bayesian approach over the conventional likelihood based approach have been demonstrated in the experimental results presented in this section.

As far as the computational complexity is concerned, in comparison with the Gaussian density computation in the conventional method, the Bayesian approach requires slightly more computation in calculating the predictive density, as shown in (31). However, this does not cause any significant computation increase in the overall computing loads of the system. One major weakness of the Bayesian approach, compared with the conventional method, is the relatively large requirement in storage, either in memory or in disc, for saving the sufficient statistics for all model parameters.

VII. Summary and Conclusion

The non-Bayesian approach based on the LRT has dominated speech technology while dealing with the verification problem. In this paper, we present a novel approach to solving the verification problem. We address the verification problem from a Bayesian viewpoint. The Bayesian approach to verification involves the evaluation of a quantity called the *Bayes factor* against a critical threshold. The calculation of the *Bayes factor* involves evaluating several Bayesian predictive densities.

Specifically, we report in this paper our study of the outlier verification problem in pattern recognition. While the approach presented is applicable to many practical verification problems, in this study, we apply it to speaker verification based on the GMM. For this application, we propose an efficient algorithm to calculate the *Bayes factor* for the GMM and an effective strategy to perform speaker verification based on the *Bayes factor*. The proposed approach is investigated and evaluated using the *NIST98* speaker recognition evaluation data. Experimental results demonstrate the consistent effectiveness and efficiency of the Bayesian approach in speaker verification.

Finally, we discuss some possible future research directions along the line of the work described in this paper. First, although the work in the paper is derived for speaker verification based on the GMM, it is straightforward to extend it to the HMM. The basic verification principle is also applicable to all other types of verification problems. Thus, as a direct extension, we also can apply the Bayesian approach to another important verification problem in speech recognition, i.e., utterance verification based on HMM. Second, as discussed in Section III-B, using the approximation of (7), we can obtain a much simpler method for computing the *Bayes factor*. It will be very interesting to see how well this highly efficient approximate method will work in practice.

APPENDIX
ESTIMATION OF GMM BASED ON EM ALGORITHM

We consider the parameter estimation problem of the GMM $\Lambda = \{m_i, r_i, w_i | 1 \le i \le N\}$ from multiple observations $\mathbf{X}$ by using the EM algorithm. We denote $\mathbf{X} = \{X_s | s = 1, 2, \ldots, M\}$, where each observation is a sequence denoted by $X_s = \{x_{s1}, x_{s2}, \ldots, x_{sT_s}\}$.

Given the initial GMM parameters $\Lambda^0 = \{m_i^0, r_i^0, w_i^0 | 1 \le i \le N\}$

E-Step:

$$\begin{aligned} Q(\Lambda|\Lambda^0) &= E_{\mathcal{P}}[\ln f(\mathbf{X}, \mathcal{P}|\Lambda)|\mathbf{X}, \Lambda^0)] \\ &= \sum_{s=1}^{M}\sum_{t=1}^{T_s}\sum_{i=1}^{N} \\ &\quad \cdot \left\{\sum_{d=1}^{D}\left[-\frac{r_{id}}{2}(x_{st}-m_{id})^2 + \frac{1}{2}\ln r_{id}\right] + \ln w_i\right\} \\ &\quad \cdot \xi_{st}(i) + \text{const.} \end{aligned} \tag{43}$$

where $\xi_{st}(i)$ denotes the probability of x_{st} residing in the ith mixture component, i.e.,

$$\xi_{st}(i) = \frac{w_i^0 \cdot \mathcal{N}(x_{st}|m_i^0, r_i^0)}{\sum_{i=1}^{N} w_i^0 \cdot \mathcal{N}(x_{st}|m_i^0, r_i^0)}. \tag{44}$$

M-Step: For $i = 1, 2, \ldots, N$ and $d = 1, 2, \ldots, D$,

$$\frac{\partial Q(\Lambda|\Lambda^0)}{\partial m_{id}} = \sum_{s=1}^{M}\sum_{t=1}^{T_s}(x_{st} - m_{id}) \cdot \xi_{st}(i) = 0. \tag{45}$$

Thus

$$m_{id} = \frac{\sum_{s=1}^{M}\sum_{t=1}^{T_s} x_{st} \cdot \xi_{st}(i)}{\sum_{s=1}^{M}\sum_{t=1}^{T_s}\xi_{st}(i)} \tag{46}$$

$$\frac{\partial Q(\Lambda|\Lambda^0)}{\partial r_{id}} = \sum_{s=1}^{M}\sum_{t=1}^{T_s}\left[\frac{1}{r_{id}} - (x_{st} - m_{id})^2\right] \cdot \xi_{st}(i) = 0. \tag{47}$$

Thus

$$r_{id} = \frac{\sum_{s=1}^{M}\sum_{t=1}^{T_s}\xi_{st}(i)}{\sum_{s=1}^{M}\sum_{t=1}^{T_s}\xi_{st}(i) \cdot (x_{st} - m_{id})^2}. \tag{48}$$

As for w_i, we have the constraint of $\sum_{i=1}^{N} w_i = 1$. Using the Lagrange multiplier method, we have

$$\frac{\partial Q(\Lambda|\Lambda^0)}{\partial w_i} - \lambda = \sum_{s=1}^{M}\sum_{t=1}^{T_s}\frac{1}{w_i} \cdot \xi_{st}(i) - \lambda = 0. \tag{49}$$

Thus

$$w_i = \frac{\sum_{s=1}^{M}\sum_{t=1}^{T_s}\xi_{st}(i)}{\sum_{i=1}^{N}\sum_{s=1}^{M}\sum_{t=1}^{T_s}\xi_{st}(i)}. \tag{50}$$

REFERENCES

[1] M. Aitkin, "Posterior Bayes factors," *J. R. Statist. Soc. B*, vol. 53, no. 1, pp. 111–142, 1991.

[2] R. Auckenthaler, E. S. Parris, and M. J. Carey, "Improving a GMM speaker verification system by phonetic weighting," in *Proc. ICASSP'99*, Phoenix, AZ, Mar. 1999, pp. I-313–316.

[3] J. Cernocky, *et al.*, "A segmental approach to text-independent speaker verification," in *Proc. Eurospeech'99*, Budapest, Hungary, 1999, pp. 2207–2210.

[4] M. H. DeGroot, *Optimal Statistical Decisions*. New York: McGraw-Hill, 1970.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, ser. B, vol. 39, no. 1, pp. 1–38, 1977.

[6] A. C. Emmanuel, M. Newman, B. Peskin, L. Gillick, and R. Roth, "Progress in speaker recognition at Dragon systems," in *Proc. 1998 Int. Conf. Spoken Language Processing*, Sydney, Australia, 1998, pp. 1355–1358.

[7] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York: Academic, 1972.

[8] J.-L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observation of Markov chains," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 291–298, 1994.

[9] Q. Huo and C.-H. Lee, "On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 161–172, Mar. 1997.

[10] H. Jiang, K. Hirose, and Q. Huo, "Robust speech recognition based on Bayesian prediction approach," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 426–440, July 1999.

[11] ——, "Improving Viterbi Bayesian predictive classification via sequential Bayesian learning in robust speech recognition," *Speech Commun.*, vol. 28, no. 4, pp. 313–326, Aug. 1999.

[12] R. E. Kass and A. E. Raftery, "Bayes factors," *J. Amer. Statist. Assoc.*, vol. 90, no. 430, pp. 773–795, June 1995.

[13] C.-H. Lee, F.-K. Soong, and K.-K. Paliwal, Eds., *Automatic Speech and Speaker Recognition: Advanced Topics*. Norwell, MA: Kluwer, 1996.

[14] Q. Li, B.-H. Juang, Q. Zhou, and C.-H. Lee, "Automatic verbal information verification for user authentication," IEEE Trans. Speech Audio Processing, vol. 8, pp. 585–596, Sept. 1999, submitted for publication.

[15] C.-S. Liu, H.-C. Wang, and C.-H. Lee, "Speaker verification using normalized log-likelihood score," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 56–60, Jan. 1996.

[16] G. J. McLachlan and K. E. Basford, *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker, 1988.

[17] H. A. Murthy, F. Beaufays, L. P. Heck, and M. Weintraub, "Robust text-independent speaker identification over telephone channels," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 554–568, Sept. 1999.

[18] A. O'Hagan, "Fractional Bayes factors for model comparison," *J. R. Statist. Soc. B*, vol. 57, no. 1, pp. 99–138, 1995.

[19] M. G. Rahim, C.-H. Lee, and B.-H. Juang, "Discriminative utterance verification for connected digits recognition," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 266–277, May 1997.

[20] D. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Commun.*, vol. 17, pp. 91–108, Aug. 1995.

[21] ——, "Comparison of background normalization methods for text-independent speaker verification," in *Proc. Eurospeech'97*, vol. 2, Sept. 1997, pp. 963–966.

[22] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, no. 1-3, pp. 19–41, Jan. 2000.

[23] R. A. Sukkar and C.-H. Lee, "Vocabulary independent discriminant utterance verification for nonkeyword rejection in subword based speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 420–429, Nov. 1996.

[24] K. Yu and J. Mason, "On-line incremental adaptation for speaker verification using maximum likelihood estimates of CDHMM parameters," in *Proc. Int. Conf. Spoken Language Processing 1996*, 1996, pp. 1752–1755.

Hui Jiang (M'00) received the B.Eng. and M.Eng. degrees from the University of Science and Technology of China (USTC), Hefei, in July 1992 and December 1994, respectively, and the Ph.D. degree from the University of Tokyo, Tokyo, Japan in September 1998, all in electrical engineering.

From 1992 to 1994, he worked on large vocabulary Chinese speech recognition at USTC. From October 1998 to April 1999, he was a Researcher with the University of Tokyo. From April 1999 to June 2000, he was with Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, as a Postdoctoral Fellow. Since June 2000, he has been with Dialogue Systems Research, Multimedia Communication Research Lab, Bell Labs, Lucent Technologies Inc., Murray Hill, NJ. His current research interests include all issues related to speech recognition and understanding, especially robust speech recognition, utterance verification, adaptive modeling of speech, spoken language systems, and speaker recognition/verification.

Li Deng (S'83–M'86–SM'91) received the B.S. degree from the University of Science and Technology of China in biophysics in 1982, the Master degree from University of Wisconsin-Madison in electrical engineering in 1984, and the Ph.D. degree from University of Wisconsin-Madison in electrical engineering in 1986.

He worked on large vocabulary automatic speech recognition for telecommunications in Montreal, PQ, Canada, from 1986 to 1989. In 1989, he joined the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, as Assistant Professor; he became Full Professor in 1996. From 1992 to 1993, he conducted sabbatical research at the Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, and from 1997 to 1998 with ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan. In December 1999, he joined Microsoft Research, Redmond, WA, as Senior Researcher. His research interests include acoustic-phonetic modeling of speech, speech and speaker recognition, speech synthesis and enhancement, speech production and perception, auditory speech processing, statistical methods and machine learning, nonlinear signal processing and system theory, spoken language systems, and human–computer interaction.