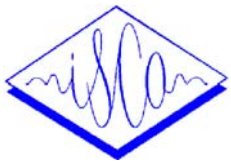

ODYSSEY04 - The Speaker and
Language Recognition Workshop
Toledo, Spain
May 31 - June 3, 2004

Bayes Factor Scoring of GMMs for Speaker Verification

Robbie Vogt and Sridha Sridharan

Speech and Audio Research Laboratory,
Queensland University of Technology
GPO Box 2434, George St, Brisbane, AUSTRALIA, 4001.
{r.vogt, s.sridharan}@qut.edu.au

Abstract

This paper implements and assesses the Bayes factor as a replacement verification criterion to the likelihood-ratio test in the context of GMM-based speaker verification. An advantage of the Bayesian method is that model parameters are considered random variables, allowing for the incorporation of prior information and uncertainty of parameter estimates into the scoring process. A novel development of Bayes factors for GMMs is presented based on incremental adaptation that is well-suited to inclusion in existing state-of-the-art GMM-UBM systems. Experiments on the 1999 NIST Speaker Recognition Evaluation corpus demonstrate improved performance over expected log-likelihood ratio scoring particularly when combined with the feature mapping technique.

1. Introduction

Over the past decade, speaker recognition technology has advanced to the extent that it is sufficiently accurate for use in real applications. However, to date the range of these applications falls well short of the extensive possibilities for the technology.

While current state-of-the-art text-independent speaker verification systems are capable of equal error rates (EER) of 5-10%, most applications require an EER in the order of 0.1%. It is clear that there are still significant improvements required.

Many of the techniques used in current speaker verification technology require vast amounts of contextual acoustic data to adapt the system to a particular situation or application of interest. Most advances in speaker recognition in recent times (in very general terms) have been developments that find ways to utilise more data to train, adapt or otherwise fortify speaker recognition systems in adverse conditions. Techniques that fall in this category include the introduction of Universal Background Models (UBM) [1], handset type and test-segment normalisation (HNorm and TNorm) [2].

This paper presents an improved scoring method for GMM-based speaker verification systems by employing a Bayesian approach to analysing the underlying verification problem. The resulting technique replaces the commonly used likelihood-ratio test (LRT) with a criterion based on Bayes factors [3]. Hypothesis testing using Bayes factors has several advantages over non-Bayesian approaches including the ability to evaluate evidence *in favour* of the null hypothesis and to incorporate prior information into the scoring process analogous to *maximum a posteriori* (MAP) adaptation for model training.

The work presented herein was motivated by the application of Bayes factor scoring to speaker verification championed by Jiang and Deng [4] and while it adopts their central theme several significant implementational choices differentiate this work from its predecessors. Firstly, an incremental Bayes learning approach is used for calculating Bayes factors for GMMs instead of a Viterbi approximation method. Secondly, the method presented is more suited to current state-of-the-art systems based on a GMM-UBM approach and MAP adaptation; it is effectively a drop-in replacement scoring method. Results are also presented on the combination of Bayes factor scoring and the feature mapping technique [5] to reduce the impact of telephone handset mismatch.

Section 2 presents speaker verification (and the verification problem in general) in terms of a statistical hypothesis test, proceeding to develop the decision criterion for verification under a Bayesian framework, resulting in the Bayes factor. The traditional LRT is presented to contrast the approaches.

In Section 3 the Bayes factor scoring of GMMs is derived and the implementational aspects of the speaker verification system used for experimental comparison are presented. The feature mapping technique is also reviewed in this section.

Sections 4 and 5 detail the experiments performed and results achieved when comparing the LRT based speaker verification system to the proposed Bayes factor scored system and the Bayes factor with feature mapping system. These experiments target conversational telephony data and are based on the NIST 1999 Speaker

Recognition Evaluation protocol.

2. Bayes Factors

Speaker verification, and verification problems generally, can be considered in the framework of statistical hypothesis testing. In the case of speaker verification, the hypothesis under scrutiny, H_1, is that an utterance was produced by the claimant speaker. The null hypothesis, H_0, is simply that the utterance was produced by another speaker. Under this scenario, the appropriate statistic for testing the hypotheses is the posterior odds of H_1 given by

$$\frac{P(H_1|\boldsymbol{D})}{P(H_0|\boldsymbol{D})} \tag{1}$$

where $\boldsymbol{D}$ is the available data evidence and $P(H_k|\boldsymbol{D})$ is the *a posteriori* probability of the hypothesis H_k given this evidence. Applying Bayes theorem to numerator and denominator, (1) becomes

$$\frac{P(H_1|\boldsymbol{D})}{P(H_0|\boldsymbol{D})} = \frac{P(H_1)}{P(H_0)} \times \frac{P(\boldsymbol{D}|H_1)}{P(\boldsymbol{D}|H_0)} \tag{2}$$

It can be readily seen that the posterior odds are the prior odds scaled by a factor dependent on the evidence. This scaling factor is the *Bayes factor* [3], denoted B_{10},

$$B_{10} = \frac{P(\boldsymbol{D}|H_1)}{P(\boldsymbol{D}|H_0)} \tag{3}$$

The Bayes factor can be used directly as a decision criterion for verification, with an easily interpreted threshold if the prior odds are known.

Typically, the available evidence $\boldsymbol{D}$ consists of the test utterance, $\boldsymbol{y}$, and sample information from each class (i.e., training data), represented by $\boldsymbol{X_1}$ for the claimant and $\boldsymbol{X_0}$ to represent the "all other speakers" class (commonly referred to as the background speaker). Incorporating this data, the posterior odds become

$$B_{10} = \frac{P(\boldsymbol{y}, \boldsymbol{X_1}, \boldsymbol{X_0}|H_1)}{P(\boldsymbol{y}, \boldsymbol{X_1}, \boldsymbol{X_0}|H_0)} \tag{4}$$

For this paper we are particularly concerned with the solution of (4) incorporating a parametric model structure for each class (Gaussian mixtures). Under a Bayesian framework, the model parameters are considered *unknown random variables* with a probability distribution, allowing for the case of incomplete data and uncertainty in parameter estimates. Thus the densities $P(\boldsymbol{D}|H_k)$ in (3) are calculated by integration over the model parameter space (instead of maximising).

$$P(\boldsymbol{D}|H_k) = \int p(\boldsymbol{D}|\lambda_k, H_k) p(\lambda_k|H_k) d\lambda_k \tag{5}$$

where λ_k is the vector of unknown parameters for the model representing class k. Assuming independence of the training and test data and utilising Bayesian incremental learning [6], (4) becomes

$$B_{10} = \frac{\int p(\boldsymbol{y}, \boldsymbol{X_1}|\lambda_1) p(\lambda_1) d\lambda_1 \cdot \int p(\boldsymbol{X_0}|\lambda_0) p(\lambda_0) d\lambda_0}{\int p(\boldsymbol{X_1}|\lambda_1) p(\lambda_1) d\lambda_1 \cdot \int p(\boldsymbol{y}, \boldsymbol{X_0}|\lambda_0) p(\lambda_0) d\lambda_0} = \frac{\int p(\boldsymbol{y}|\lambda_1) p(\lambda_1|\boldsymbol{X_1}) d\lambda_1}{\int p(\boldsymbol{y}|\lambda_0) p(\lambda_0|\boldsymbol{X_0}) d\lambda_0} \tag{6}$$

In this paper, the factor in (6) is used as the criterion for verification. Although this Bayes factor requires integration over the entire parameter space (comprising thousands of dimensions in the high-order GMM case), a method for efficiently calculating an approximation is presented in Section 3.1. The next section derives the likelihood ratio test (LRT) as a special case of the Bayes factor approach highlighting the constraints and assumptions implied by the LRT.

2.1. The Likelihood Ratio: a Special Case

Under the assumption that both hypotheses are represented by probability distributions with no free parameters, (4) resolves to the familiar likelihood ratio—this is known as the "simple-vs-simple" case [3]. Additionally, under the strong condition that the probability distributions are exactly known, the Neyman-Pearson Lemma suggests that the likelihood ratio is in fact the most powerful criterion. Therefore, (6) becomes

$$B_{10} = \frac{p(\boldsymbol{y}, \boldsymbol{X_1}|\lambda_1) p(\boldsymbol{X_0}|\lambda_0)}{p(\boldsymbol{y}, \boldsymbol{X_1}|\lambda_1) p(\boldsymbol{X_0}|\lambda_0)} = \frac{p(\boldsymbol{y}|\lambda_1)}{p(\boldsymbol{y}|\lambda_0)} = \Lambda_{10} \tag{7}$$

where λ_1 and λ_0 are the estimated parameters for the claimant and background models respectively.

It follows from these conditions that by using the likelihood ratio we are assuming that our model parameters are constant and estimated perfectly.

3. GMM Speaker Verification using Bayes Factor Scoring

This section describes the incorporation of Bayes factor scoring into an existing speaker verification system [7] based on the GMM-UBM [1] structure. Section 3.1 derives the Bayes factor scoring criteria for Gaussian mixture models while Section 3.2 describes some of the practical implementation issues and efficiency improvements used in this research. Finally, Section 3.3 describes the use of feature mapping to reduce observed mismatch.

3.1. Bayes Factor Scoring for GMMs

To evaluate Bayes factors for GMMs it is necessary to evaluate the Bayesian predictive density (5)

$$p(\boldsymbol{X}|H) = \int p(\boldsymbol{X}|\lambda) p(\lambda) d\lambda \tag{8}$$

with the model density function given by

$$p(\boldsymbol{X}|\lambda) = \prod_{t=1}^{T} \sum_{i=1}^{N} w_i g(\boldsymbol{x}_t|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \quad (9)$$

with the constraint of diagonal covariance matrices

$$g(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\sigma_{id}^2}} \exp\left\{-\frac{(x_d - \mu_{id})^2}{2\sigma_{id}^2}\right\} \quad (10)$$

Following from common practice in MAP adaptation of GMMs and supporting experimental evidence, only the component Gaussian means are considered for adaptation in this work. Consequently the prior distribution for $\lambda = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \dots \boldsymbol{\mu}_N\}$ is [8]

$$p(\lambda) = \prod_{i=1}^{N} g(\boldsymbol{\mu}_i|\Theta_i) \quad (11)$$

where $\Theta_i = \{\tau_i, \boldsymbol{m}_i\}$ are the set of hyperparameters with $\tau_i > 0$ and $\boldsymbol{m}_i$ is a D-dimensional vector and $g(\boldsymbol{\mu}_i|\Theta_i)$ is given by

$$g(\boldsymbol{\mu}_i|\Theta_i) = \prod_{d=1}^{D} \sqrt{\frac{\tau_i}{2\pi\sigma_{id}^2}} \exp\left\{-\frac{\tau_i(\mu_{id} - m_{id})^2}{2\sigma_{id}^2}\right\} \quad (12)$$

Jiang and Deng [4] approximate the solution of (8) by performing the Viterbi approximation of [9], effectively assigning each sample to a single component Gaussian. In contrast, we adopt an incremental approach by updating the model prior density after each observation using incremental Bayesian learning. Hence, (8) simplifies to the iterative evaluation of

$$p(\boldsymbol{X}|H) = \prod_{t=1}^{T} \int p(\boldsymbol{x}_t|\lambda) p(\lambda|\boldsymbol{X}^{(t-1)}) d\lambda \quad (13)$$

where $\boldsymbol{X}^{(t-1)} = \{\boldsymbol{x}_1, \boldsymbol{x}_2 \dots \boldsymbol{x}_{t-1}\}$ is the set of observation vectors preceding $\boldsymbol{x}_t$. Under this interpretation, $\int p(\boldsymbol{x}_t|\lambda) p(\lambda|\boldsymbol{X}^{(t-1)}) d\lambda$ simplifies to a weighted sum of integrals over the component Gaussians,

$$\int p(\boldsymbol{x}|\lambda) p(\lambda|\boldsymbol{X}) d\lambda = \sum_{i=1}^{M} w_i \int p(\boldsymbol{x}|\boldsymbol{\mu}_i) p(\boldsymbol{\mu}_i|\boldsymbol{X}) d\boldsymbol{\mu}_i \quad (14)$$

where

$$\int p(\boldsymbol{x}|\boldsymbol{\mu}_i) p(\boldsymbol{\mu}_i|\boldsymbol{X}) d\boldsymbol{\mu}_i = \prod_{d=1}^{D} \sqrt{\frac{\tau_i}{2\pi\sigma_{id}^2(\tau_i + 1)}} \exp\left\{-\frac{\tau_i(x_{id} - m_{id})^2}{2(\tau_i + 1)\sigma_{id}^2}\right\} \quad (15)$$

The update equations for the prior distribution hyperparameters are equivalent to the MAP update equations for GMMs

$$\tau_i' = \tau_i + P(i|\boldsymbol{x}) \quad (16)$$

$$\boldsymbol{m}_i' = \frac{\tau_i \boldsymbol{m}_i + P(i|\boldsymbol{x})\boldsymbol{x}}{\tau_i + P(i|\boldsymbol{x})} \quad (17)$$

where τ_i' and $\boldsymbol{m}_i'$ are the updated hyperparameters after observing $\boldsymbol{x}$ and

$$P(i|\boldsymbol{x}) = \frac{w_i g(\boldsymbol{x}_t|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{p(\boldsymbol{x}|\lambda)}$$

is posterior probability of mixture component i producing the observation. From the above equations, it can be seen that Bayes factor scoring can in fact be implemented as incremental MAP adaptation with adjusted variances to compensate for uncertainty in the component means. It should be noted that both the claimant and background models are scored in this fashion.

3.2. Implementation in a Speaker Verification System

Several issues remain with respect to the practical implementation of Bayes factor scoring within a speaker verification system.

Firstly, the discussion above doesn't mention the initial values for the prior distribution hyperparameters, $\Theta = \{\boldsymbol{m}_i, \tau_i | i = 1, 2 \dots M\}$. For all models the initial values of the hyperparameters are the same; the prior means are derived from the UBM (as is the case with MAP adaptation) and all τ_i are set to the MAP adaptation "relevance factor," τ. These values are then updated as a result of the training procedure; the prior means become the MAP adapted means and τ_i is the sum of the relevance factor and the probabilistic count for mixture component i. The probabilistic counts from model training must therefore be recorded. The model training procedure consequently has a slightly different interpretation under this scheme as it adapts the *prior* distribution hyperparameters to be speaker dependent rather than estimating a speaker dependent model directly.

For testing, the background model hyperparameters are set to the initial values. An interpretation of this is that, at the start of a test utterance the background model effectively represents *no* speaker in contrast to the usual interpretation of representing many unknown speakers. To be verified a claimant speaker model has to be *more like* the test utterance than *no* speaker as the background model will adapt more rapidly toward the test utterance than a trained model.

Secondly, for efficient evaluation of the Bayes factor a *top-N* scoring strategy is employed that works similarly to the *top-N* expected log-likelihood ratio (ELLR) scoring [1]. This also implies that only the N highest contributing components of a model are updated by an obser-

vation; a positive side-effect of this is the reduced potential for numerical accuracy issues in the update step. All experiments in this study use $N = 10$. It should be noted that even with *top*-N scoring Bayes factor scoring is more computationally expensive than ELLR scoring due to the extra effort in incrementally adapting the means.

3.3. Feature Mapping

The recently published feature mapping technique [5] is a handset type normalisation technique similar in approach to Speaker Model Synthesis [10] however it is interpreted as a feature-space approach. This has the advantage of reducing the apparent mismatch in features observed by the speaker models.

Feature mapping learns a set of non-linear transformations from a known set of contexts to a common, context-independent feature space. The transformation for each context is derived by adapting a context-dependent GMM from a context-independent root model using labeled data.

Given an utterance the most likely context dependent model is detected; each feature vector is then transformed to the context-independent space by applying the mapping

$$\boldsymbol{y} = \left(\boldsymbol{x} - \boldsymbol{\mu}_i^{CD}\right)\left(\boldsymbol{\Sigma}_i^{CD}\right)^{-1}\boldsymbol{\Sigma}_i^{CI} + \boldsymbol{\mu}_i^{CI} \qquad (18)$$

where i is the top scoring component in the selected context-dependent model.

For the purposes of this study the relevant contexts represent the different telephone handset types encountered in the 1999 NIST Evaluation corpus which correspond to the electret and carbon-button transducer types. The root context-independent model was also used as the verification system UBM (although this is not a requirement) and the context-dependent models were trained using MAP adaptation of mixture component means and variances.

4. Experiments

The recognition system used in this study utilises fully coupled GMM-UBM modelling using iterative MAP adaptation and feature-warped MFCC features with appended delta coefficients, as described in [7]. An adaptation relevance factor of $\tau = 8$ and 512-component models are used throughout.

For this evaluation, the NIST 1999 Speaker Recognition Evaluation database was used. (For further information see [11].) This database is an excerpt of the SWITCHBOARD-II Phase 3 telephone speech corpus including a collection of 230 male and 309 female target speakers, each providing approximately two minutes of enrollment speech. There are 1448 male and 1972 female test segments of up to one minute in length. Of particular interest with this database is the emphasis placed on the varying levels of mismatch represented. The results are categorised into three subsets; SNST, DNST and DNDT in order of increasing mismatch between training and testing conditions. The SNST or Same Number Same Type represents the least mismatched case. Here Type refers to the telephone handset type, either electret or carbon-button transducer. DNST is Different Number Same Type, and DNDT is the most mismatched and poorest performing set indicating that different handset types were used for training and testing.

5. Results and Discussion

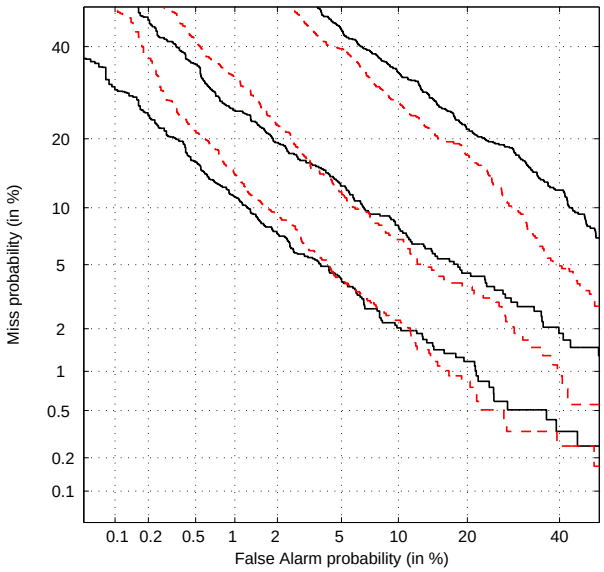

Figure 1: *DET plot of NIST99 results compares Bayes factor (solid) and ELLR (dashed) scoring for the SNST, DNST and DNDT conditions.*

Figure 1 compares the detection error trade-off (DET) curves of Bayes factor and ELLR scoring for the NIST99 data with equal error rate (EER) and minimum detection cost function (DCF) [11] presented in Tables 1 and 2. Generally improved performance in the low false alarm region is attained with the Bayes factor method, with up to a 17% improvement in DCF observed in the SNST case. The DET plots demonstrate a trend of a counter-clockwise rotation of the Bayes factor curves compared to ELLR scoring which is reflected by the improvements in the minimum DCF results reported in Table 1. Assuming Gaussian output score distributions, the observed reduction in DET curve slope would indicate a proportional reduction in the ratio of standard deviations of impostor to target trial score distributions [12] (termed the σ-ratio). This however does not seem to be the case as Bayes factor scoring results in increased σ-ratios (by 16% in the SNST case) as presented in Table 3. Further investigation of the negentropy [12] statistics suggests that Bayes factor scoring produces target score distributions

	ELLR	Bayes	FM Bayes
SNST	0.0240	**0.0199**	0.0206
DNST	0.0412	**0.0349**	0.0362
DNDT	0.0713	0.0741	**0.0690**

Table 1: *Minimum DCF results for NIST99 of ELLR scoring and Bayes factor scoring with and without feature mapping applied for the SNST, DNST and DNDT conditions.*

	ELLR	Bayes	FM Bayes
SNST	**4.5%**	**4.5%**	4.6%
DNST	**8.2%**	9.1%	8.9%
DNDT	18.9%	21.1%	**18.8%**

Table 2: *EER results for NIST99 of ELLR scoring and Bayes factor scoring with and without feature mapping applied for the SNST, DNST and DNDT conditions.*

that exhibit a better match to the Gaussian assumption. The negentropy statistics are also presented in Table 3 for both the target and impostor distributions, with lower values indicating distributions closer to being Gaussian.

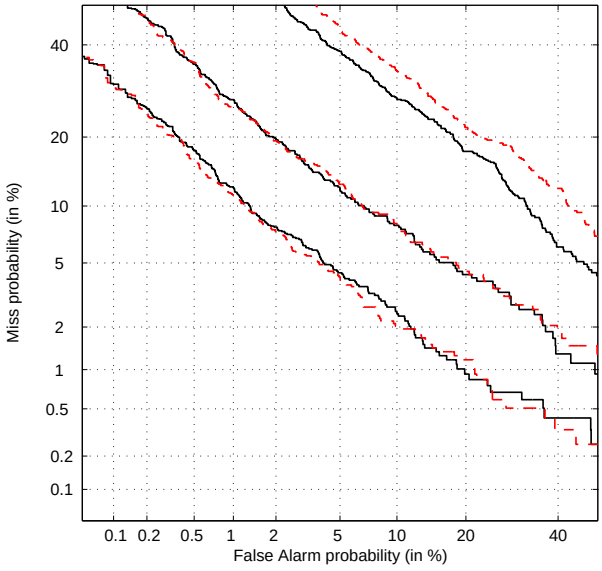

Figure 2: *DET plot of NIST99 results for Bayes factor scoring, with (solid) and without (dashed) feature mapping applied for the SNST, DNST and DNDT conditions.*

It is also noted that the results indicate a reducing effectiveness of Bayes factor scoring as mismatch increases resulting in worse performance in the DNDT case compared to standard ELLR. It is hypothesised that while the Bayes scoring method is more effective than ELLR scoring at discriminating between speaker classes, it is significantly more affected by (the common case of) mismatched features. Motivated by this hypothesis, the recently published feature mapping technique [5] was applied to reduce the apparent mismatch in features observed by the speaker models.

	$J \times 10^{-2}$ *(target/impostor)*		σ-ratio	
	ELLR	Bayes	ELLR	Bayes
SNST	9.4/5.6	8.3/6.0	0.63	0.73
DNST	5.8/5.6	5.0/6.0	0.73	0.82
DNDT	7.4/6.0	5.0/6.3	0.80	0.81

Table 3: *Negentropy (J) and σ-ratio distribution statistics for NIST99 of ELLR and Bayes factor scoring for the SNST, DNST and DNDT conditions.*

Figure 2 and the right-most columns of Tables 1 and 2 present the results of applying feature mapping to the Bayes factor scoring system. It can be seen that feature mapping does in fact improve the Bayes factor performance especially for the more mismatched scenario. A slight degradation in performance is observed however for the matched type conditions. This degradation can be attributed to errors in the handset type classification performed as a preliminary step in feature mapping, thus causing an *increase* in the observed mismatch for misclassified utterances.

Future research will re-evaluate incremental MAP adaptation as utilised in the Bayes factor scoring implementation presented for the purpose of speaker model training. Potentially, performance on par with the current iterative adaptation approach could be achieved with a single-pass algorithm. The elegance of a unified approach to training and scoring is also appealing. This technique would have obvious extensions to on-line adaptation of speaker models.

6. Conclusion

This study presented an application of Bayes factor scoring to speaker verification. The general Bayesian approach to verification was reviewed, highlighting the ability of the approach to incorporate prior information into the scoring process and to allow for uncertainty in model parameters. It was then applied to the specific case of Gaussian mixture models using a novel incremental learning derivation resulting in a drop-in replacement for ELLR scoring.

Experiments conducted on the 1999 NIST Speaker Recognition Evaluation corpus demonstrated generally improved performance of Bayes factor scoring over ELLR scoring particularly in better matched conditions and in the low false alarm operating region. Further improved performance was subsequently achieved for the mismatched case with the application of feature mapping.

7. Acknowledgments

This research was supported by the Office of Naval Research (ONR) under grant N000140310662.

8. References

[1] D. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Eurospeech*, 1997, vol. 2, pp. 963–966.

[2] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1/2/3, pp. 42–54, 2000.

[3] R.E. Kass and A.E. Raftery, "Bayes factors," *Journal of the American Statistical Society*, vol. 90, no. 430, pp. 773–795, 1995.

[4] H. Jiang and L. Deng, "A Bayesian approach to the verification problem: applications to speaker verification," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 874–884, 2001.

[5] D. Reynolds, "Channel robust speaker verification via feature mapping," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003, vol. 2, pp. II–53–6.

[6] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, John Wiley and Sons Inc, New York City, New York, USA, 2001.

[7] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *A Speaker Odyssey, The Speaker Recognition Workshop*, 2001, pp. 213–218.

[8] J-L. Gauvain and C-H. Lee, "Bayesian adaptive learning and map estimation of hmm," in *Automatic Speech and Speaker Recognition: Advanced Topics*, C-H. Lee, F.K. Soong, and K.K. Paliwal, Eds., pp. 83–107. Kluwer Academic, Boston, Mass, 1996.

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

[10] R. Teunen, B. Shahshahani, and L. Heck, "A model-based transformational approach to robust speaker recognition," in *International Conference on Spoken Language Processing*, 2000, p. Paper 1642.

[11] National Institute of Standards and Technology, "NIST speech group website," http://www.nist.gov/speech, 2003.

[12] J. Navratil and G. Ramaswamy, "The awe and mystery of t-norm," in *Eurospeech*, 2003, pp. 2009–2012.