Frame-Weighted Bayes Factor Scoring for Speaker Verification

Robbie Vogt and Sridha Sridharan

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

Abstract

In this paper, the Bayes factor is considered as a replacement verification criterion to the likelihood-ratio test in the context of GMM-based speaker verification. An advantage of this Bayesian method is that it allows for the incorporation of prior information and uncertainty of parameter estimates into the *scoring* process, complementing the Bayesian adaptation used in training. A development of Bayes factors for GMMs is presented based on incremental adaptation that is well-suited to inclusion in existing GMM-UBM systems. This method is extended to include the weighting of test frames to account for their statistical dependencies. Experiments on the 1999 NIST Speaker Recognition Evaluation corpus demonstrate improved performance over expected log-likelihood ratio scoring. These findings are supported with results from a modified version of the NIST Extended Data corpus of 2003.

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 1-10%, many 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 broad 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) (Reynolds 1997), handset type and test-segment normalisation (H-Norm and T-Norm) (Auckenthaler, Carey, and Lloyd-Thomas 2000).

In contrast, 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 (Kass and Raftery 1995). 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 (2001) 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. It extends the work presented in Vogt and Sridharan (2004) with a novel frame-weighted adaptation variant of Bayes factor scoring to compensate for highly correlated acoustic features commonly used in speaker verification.

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 and resulting in the Bayes factor.

In Section 3. Bayes factor scoring of GMMs is derived and the implementational aspects of the speaker verification system used for experimental comparison are presented. Section 3.2. also presents a novel enhancement of the presented Bayesian methods specific to acoustic speaker verification by compensating for the highly correlative nature of commonly used acoustic features via frame weighting.

Section 4. details the experiments performed and results achieved when comparing the LRT based speaker verification system to the proposed Bayes factor scored system. These experiments target conversational telephony data and are based on both the NIST 1999 Speaker Recognition Evaluation protocol (Section 4.1.) and a modified version of the NIST 2003 Extended Data Task protocol (Section 4.2.).

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 pro-

Proceedings of the 10th Australian International Conference on Speech Science & Technology
Macquarie University, Sydney, December 8 to 10, 2004. Copyright, Australian Speech Science & Technology Association Inc.

Accepted after full paper review

duced 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 the 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* (Kass and Raftery 1995), denoted B_{10} or simply B,

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

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 training data for the claimant, represented by $\boldsymbol{X}$. Incorporating this data, the Bayes factor becomes

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

For this paper we are particularly concerned with the solution of (4) incorporating a parametric model structure to represent a speaker or class (Gaussian mixtures). Under a Bayesian framework, the model parameters are considered *unknown random variables* which themselves have a probability density distribution, allowing for the case of incomplete data and uncertainty in parameter estimates. Thus to calculate $P(\boldsymbol{D}|H_k)$ in (3), we must integrate the densities $p(\lambda|H_k)$ over the model parameter space (rather than determining parameter estimates that maximise it).

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

where λ is the vector of unknown parameters for the model representing the claimant. Under this framework, (4) can be expressed as

$$B_{10} = \frac{\int p(\boldsymbol{y}, \boldsymbol{X}|\lambda) p(\lambda) d\lambda}{\int p(\boldsymbol{y}|\lambda_2) p(\lambda_2) d\lambda_2 \cdot \int p(\boldsymbol{X}|\lambda_1) p(\lambda_1) d\lambda_1} \tag{6}$$

where the numerator evaluates the likelihood of the evidence ($\boldsymbol{y}$ and $\boldsymbol{X}$) coming from a *single* class, while the denominator evaluates the likelihood of the $\boldsymbol{y}$ coming from a *different* class to that of $\boldsymbol{X}$ (this difference is emphasised by the subscripted λ).

Assuming independence of the training and test data and utilising Bayesian incremental learning (Duda, Hart, and Stork 2001), (6) can be expressed as

$$\begin{aligned} B_{10} &= \frac{\int p(\boldsymbol{y}|\lambda_2) p(\lambda_2|\boldsymbol{X}) d\lambda_2 \cdot \int p(\boldsymbol{X}|\lambda_1) p(\lambda_1) d\lambda_1}{\int p(\boldsymbol{y}|\lambda_2) p(\lambda_2) d\lambda_2 \cdot \int p(\boldsymbol{X}|\lambda_1) p(\lambda_1) d\lambda_1} \\ &= \frac{\int p(\boldsymbol{y}|\lambda) p(\lambda|\boldsymbol{X}) d\lambda}{\int p(\boldsymbol{y}|\lambda) p(\lambda) d\lambda} \end{aligned} \tag{7}$$

In this paper, the factor in (7) 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..

2.1. Modelling the null hypothesis

From (6) it can be seen that we are in fact evaluating a ratio of likelihoods as our verification criterion although it is not the familiar likelihood ratio commonly used in speaker verification systems. Of particular note is the difference in the modelling of the null hypothesis.

The Bayes factor approach outlined above elegantly removes the issue of modelling the background population that has been a significant issue in the history of speaker verification research. Early in this history the background population, represented in the denominator of the likelihood ratio, was ignored and verification decisions were based solely on the likelihood of the claimant's model producing the test utterance; the particular words spoken and the acoustic environment of the recording were significant sources unwanted variability in these scores. To reduce these dependencies, a cohort of background speakers were introduced and combined to model a background population in the denominator (Rosenberg, Delong, Lee, Juang, and Soong 1992). This approach raised the question of choosing an appropriate set of speakers to form this cohort: Should the cohorts be near, far or evenly distributed? How many cohort speakers are required?

The introduction of the universal background model (UBM) and Bayesian adaptive model estimation (Reynolds 1997) allowed for more detailed and robust models while replacing the background cohort with a single model. The UBM in this approach plays a dual role by providing a prior distribution for the claimant model parameters and a "rest of the world" model as the denominator of the LRT.

The Bayes factor approach presented in this paper removes this dual role of the UBM as it is used solely for providing a prior distribution for model parameters. It is simply unnecessary under this approach to provide a model for "all other speakers;" the denominator of the ratio in (6) evaluates the likelihood of a *different* model to the claimant producing the test utterance. In this way the Bayes factor is capable of evaluating the evidence *in favour* of the null hypothesis, rather than introducing a model to represent a background population.

3. Speaker verification using Bayes factor scoring

This section describes the incorporation of Bayes factor scoring into an existing speaker verification system (Pelecanos and Sridharan 2001) based on the GMM-UBM structure (Reynolds 1997). Section 3.1. derives the Bayes factor scoring criteria for Gaussian mixture models and Section 3.2. extends this derivation to compensate for a highly correlated feature set. Section 3.3. describes some of the practical implementation issues and efficiency improvements used in this research.

3.1. Bayes factor scoring for GMMs

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

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

with the model density function given by

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

with the constraint of diagonal covariance matrices

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

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 (Gauvain and Lee 1996)

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

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

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

Jiang and Deng (2001) approximate the solution of (8) by performing the Viterbi approximation of Jiang, Hirose, and Huo (1999), 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

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

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, $\int p(\boldsymbol{x}_t|\lambda)p\big(\lambda|\boldsymbol{X}^{(t-1)}\big)d\lambda$ simplifies to a weighted sum of integrals over the component Gaussians,

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

where

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

The prior distribution $p\big(\lambda|\boldsymbol{X}^{(t-1)}\big)$ can be determined with an incremental update approach. The update equations for the prior distribution hyperparameters are equivalent to the MAP update equations for GMMs for a single observation

$$\begin{aligned} \tau_i' &= \tau_i + P(i|\boldsymbol{x}) && (16) \\ \boldsymbol{m}_i' &= \frac{\tau_i\boldsymbol{m}_i + P(i|\boldsymbol{x})\boldsymbol{x}}{\tau_i + P(i|\boldsymbol{x})} && (17) \end{aligned}$$

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

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

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 while scoring with adjusted variances to compensate for uncertainty in the component means. It should be noted that both hypotheses are evaluated in this fashion.

3.2. Test frame weighting

Acoustic features commonly used for speaker verification, such as MFCCs, exhibit high levels of correlation between consecutive observation frames. This is essentially by definition, considering that the short-term spectra and cepstra typically calculated for consecutive frames share two-thirds of their waveform samples and that delta cepstra explicitly average over a number of frames.

This correlation obviously voids the commonly sited assumption of statistically independent and identically distributed feature vectors. Although not stated explicitly, much of the preceding discussion also invokes this assumption which leads to overly confident adaptation during the Bayes factor scoring process. Particularly in the case of extreme mismatch, such as mismatched telephone handset types, this ultimately leads to degraded performance.

To prevent over confident adaptation during scoring a *frame weighted* adaptation can be employed. Adding a weighting factor β to the update equations (16) and (17), we have

$$\begin{aligned} \tau_i' &= \tau_i + \beta P(i|\boldsymbol{x}) && (18) \\ \boldsymbol{m}_i' &= \frac{\tau_i\boldsymbol{m}_i + \beta P(i|\boldsymbol{x})\boldsymbol{x}}{\tau_i + \beta P(i|\boldsymbol{x})} && (19) \end{aligned}$$

where typically $0 < \beta \leq 1$. Intuitively, β represents how dependent each observation vector is from its predecessor; a value of 1 implies statistical independence and reducing values indicate increasing correlation (and, consequently, less information).

3.3. Implementation

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

Firstly, the discussion above does not mention the initial values for the prior distribution hyperparameters, $\Theta = \{\boldsymbol{m}_i, \tau_i | i = 1, 2 \ldots M\}$. 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. For the numerator, these values are then updated as a result of the speaker enrolment/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. As a practical note, the probabilistic counts determined from model training must therefore be recorded.

Under this scheme, the speaker enrolment procedure consequently has a slightly different interpretation as it *adapts the prior distribution hyperparameters to be speaker dependent* rather than estimating a speaker model directly.

For the denominator, the prior distribution hyperparameters are left as their initial speaker independent values. An interpretation of this is that, at the start of a test utterance the denominator effectively represents *no* speaker in contrast to the usual interpretation of representing many unknown speakers with a UBM. To be verified a claimant speaker model has to be *more like* the test utterance than *no* speaker as the speaker independent prior distribution will adapt more rapidly toward the test utterance than the speaker dependent prior.

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 (Reynolds 1997). This also implies that only the N highest contributing components of a model are updated by an observation; 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 prior distributions.

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 by Pelecanos and Sridharan (2001). An adaptation relevance factor of $\tau = 8$ and 512-component models are used throughout.

4.1. NIST 1999 experiments

For this evaluation, the NIST 1999 Speaker Recognition Evaluation database was used. (For further information see (National Institute of Standards and Technology 2004).) 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 enrolment speech. There are a total of 57,310 trials of up to 45 seconds in length, with 4,828 of these being target trials. Of particular interest with this database is the emphasis placed on the levels of mismatch represented. As well as overall performance, our results are categorised into two subsets distinguished by the level of mismatch; *Same* and *Different* handset type trials.[1] In this corpus, the telephone handset type is either electret or carbon-button transducer. A trial is categorised as *Same* type if the training and testing segments were both recorded on the same telephone type; representing moderate mismatch. *Different* type trials are significantly more mismatched with consequently poorer system performance.

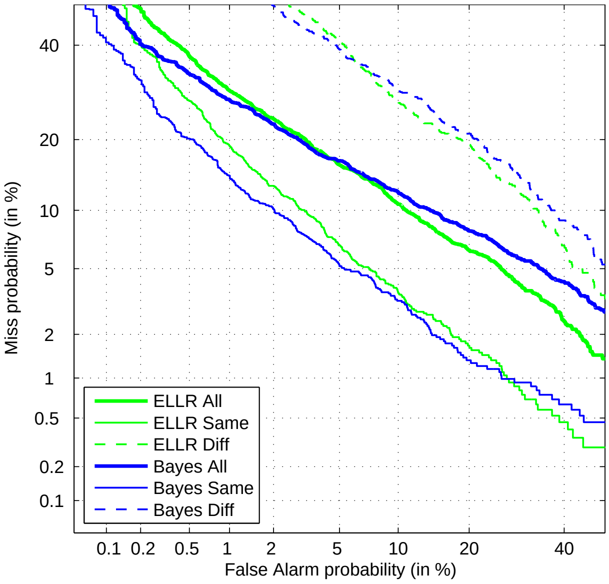


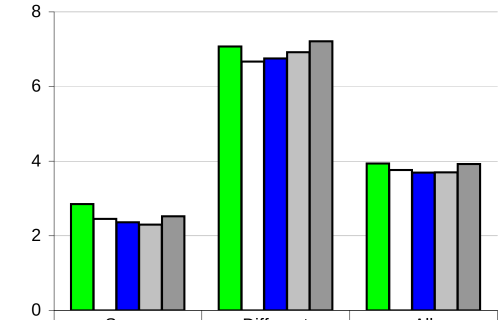
Figure 1: *DET plot of NIST '99 baseline results comparing ELLR and Bayes factor scoring ($\beta = 0.25$) for the All, Same and Different handset type conditions.*

Figure 1 compares the detection error trade-off (DET) curves of Bayes factor and ELLR scoring for the NIST '99 data with equal error rate (EER) and minimum detection cost function (DCF) (Martin and Przybocki 2000) presented in Figures 2 and 3 respectively. Improved performance in the low false alarm region is attained with the Bayes factor method, with reductions in the observed DCF for all conditions; up to a 19.3% in the *Same* case and 6.3% overall. Mixed results were observed at the EER operating point with improvements in the *Same* condition and degradations in the *All* and *Different* cases.

The DET plots demonstrate a trend of a counter-clockwise rotation of the Bayes factor curves compared to ELLR scoring. 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 termed the σ*-ratio* (Navratil and Ramaswamy 2003). This was indeed observed with the Bayes factor scoring reducing the σ-ratio by 5% overall

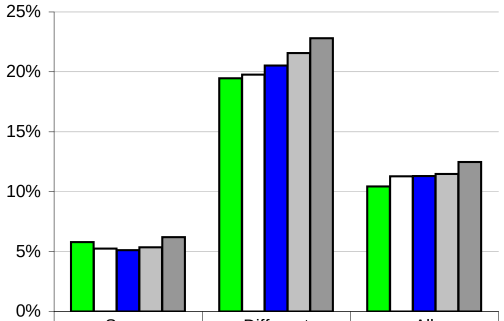
[1] The *Different* category corresponds directly to the *DNDT* condition commonly used for the NIST '99 corpus, however the *Same* condition combines the *SNST* and *DNST* conditions. This approach was chosen to improve the clarity of plots and the meaningfulness of the results presented.

Accepted after full paper review



	Same	Different	All
🟩 ELLR	2.85	7.07	3.94
□ Bayes β = 0.125	2.45	6.67	3.76
🟦 Bayes β = 0.25	2.36	6.75	3.69
▨ Bayes β = 0.5	2.30	6.92	3.70
■ Bayes β = 1.0	2.52	7.21	3.92

Figure 2: *Minimum DCF values ($\times 10^{-2}$) for NIST '99 baseline results comparing ELLR to Bayes factor scoring with varying β-values for the All, Same and Different handset type conditions.*



	Same	Different	All
🟩 ELLR	5.80%	19.47%	10.44%
□ Bayes β = 0.125	5.25%	19.78%	11.28%
🟦 Bayes β = 0.25	5.13%	20.53%	11.31%
▨ Bayes β = 0.5	5.37%	21.57%	11.47%
■ Bayes β = 1.0	6.21%	22.81%	12.47%

Figure 3: *EER for NIST '99 baseline results comparing ELLR to Bayes factor scoring with varying β-values for the All, Same and Different handset type 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 *Different* 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 more adversely affected by mismatched features. Figures 2 and 3 do, however, indicate the positive effect of incorporating frame-weighted Bayes factor scoring (compared to the unweighted version with $\beta = 1$), with $\beta = 0.125$ giving the best Bayes factor results for both DCF and EER in the *Different* case. Overall a β value of 0.25 gives the most consistent results.

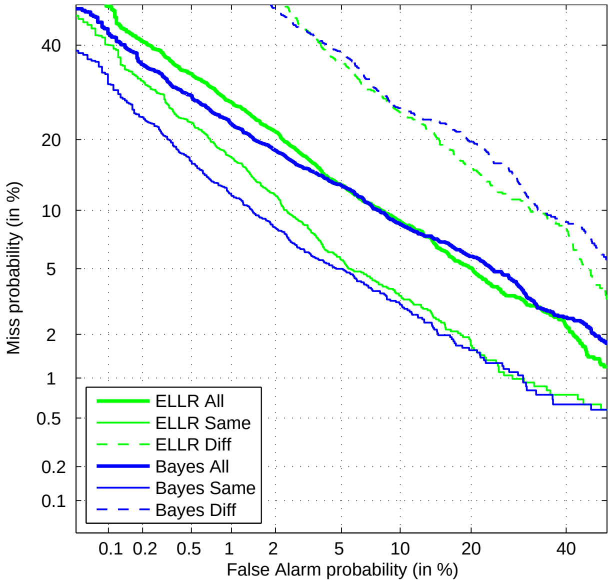


Figure 4: *DET plot of NIST '99 H-Norm results comparing ELLR and Bayes factor scoring ($\beta = 0.25$) for the All, Same and Different handset type conditions.*

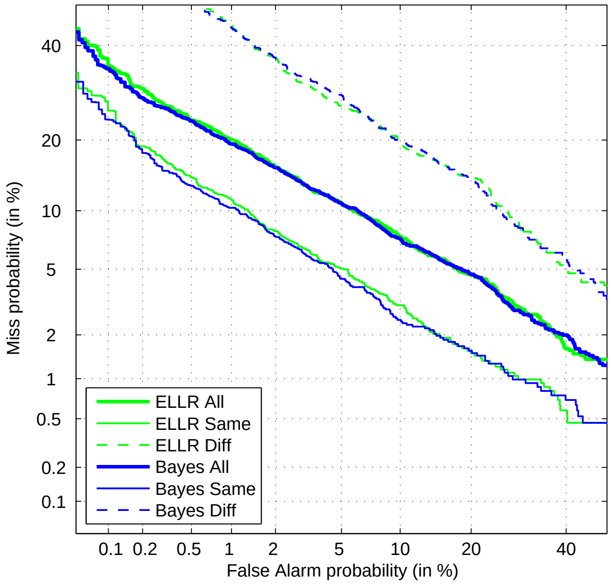


Figure 5: *DET plot of NIST '99 HT-Norm results comparing ELLR and Bayes factor scoring ($\beta = 0.25$) for the All, Same and Different handset type conditions.*

Figures 4 and 5 depict DET performance incorporating H-Norm and T-Norm (Auckenthaler, Carey, and Lloyd-Thomas 2000). H-Norm provides a significant boost for the Bayes factor method with an overall DCF improvement of 11.8% and EER imprvement of 2.7% in favour of the proposed method. The use of HT-Norm (Figure 5) almost nullifies the differences between the methods, however the Bayes factor approach has a small overall advantage in both DCF and EER.

4.2. QUT EDT 2003 experiments

The Bayes factor approach was further evaluated and compared using data from the NIST 2003 Speaker Recognition Evaluation Extended Data Task (EDT) (National Institute of Standards and Technology 2004). The evaluation data is a subset of the Switchboard-II Phase 2 and 3 databases. This study aimed at examining the performance of the approach in extended training scenarios. To mirror the NIST 2004 evaluation conditions, the NIST EDT '03 evaluation procedure was restructured to include three training length conditions: one, three and eight conversation sides. The training and testing lists for the new 1- and 3-side conditions were derived from the existing four conversation side lists. More impostor trials were also added to the evaluation to better reflect the minimum DCF operating region. We refer to this modified protocol as the QUT EDT '03.

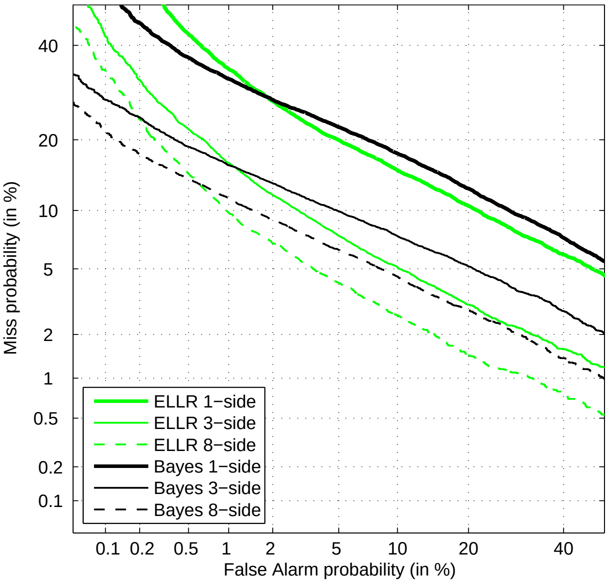


Figure 6: *DET plot of QUT EDT '03 baseline results comparing ELLR and Bayes factor scoring ($\beta = 0.25$) for the 1-side, 3-side and 8-side training conditions.*

The results for this task presented in Figure 6 support the results for the NIST '99 corpus. The Bayesian scoring method provided improved performance as measured by the minimum DCF with 6.3%, 8.5% and 2.1% relative improvement in the 1-, 3- and 8-side training conditions with degraded results at the EER operating point. These plots also confirm the trend of counter-clockwise DET curve rotation observed in the previous section.

It is clear however that the Bayes factor approach shows decreasing usefulness as the length of training data is increased. Undoubtedly this is due to the increased confidence in the model parameter estimates that can be expected from these extended quantities of training data.

5. 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 and an extended 2003 NIST 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.

6. Acknowledgments

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

The authors would like to acknowledge the fruitful discussion and input from Niko Brummer and Jason Pelecanos.

References

Auckenthaler, R., M. Carey, and H. Lloyd-Thomas (2000). Score normalization for text-independent speaker verification systems. *Digital Signal Processing 10*(1/2/3), 42–54.

Duda, R., P. Hart, and D. Stork (2001). *Pattern Classification*. New York City, New York, USA: John Wiley and Sons Inc.

Gauvain, J.-L. and C.-H. Lee (1996). Bayesian adaptive learning and MAP estimation of HMM. In C.-H. Lee, F. Soong, and K. Paliwal (Eds.), *Automatic Speech and Speaker Recognition: Advanced Topics*, pp. 83–107. Boston, Mass: Kluwer Academic.

Jiang, H. and L. Deng (2001). A Bayesian approach to the verification problem: Applications to speaker verification. *IEEE Transactions on Speech and Audio Processing 9*(8), 874–884.

Jiang, H., K. Hirose, and Q. Huo (1999). Robust speech recognition based on Bayesian prediction approach. *IEEE Transactions on Speech and Audio Processing 7*, 426–440.

Kass, R. and A. Raftery (1995). Bayes factors. *Journal of the American Statistical Society 90*(430), 773–795.

Martin, A. and M. Przybocki (2000). The nist 1999 speaker recognition evaluation—an overview. *Digital Signal Processing 10*(1-3), 1–18.

National Institute of Standards and Technology (2004). NIST speech group website. http://www.nist.gov/speech.

Navratil, J. and G. Ramaswamy (2003). The awe and mystery of t-norm. In *Eurospeech*, pp. 2009–2012.

Pelecanos, J. and S. Sridharan (2001). Feature warping for robust speaker verification. In *A Speaker Odyssey, The Speaker Recognition Workshop*, pp. 213–218.

Reynolds, D. (1997). Comparison of background normalization methods for text-independent speaker verification. In *Eurospeech*, Volume 2, pp. 963–966.

Rosenberg, A., J. Delong, C. Lee, B. Juang, and F. Soong (1992). The use of cohort normalized scores for speaker verification. In *International Conference on Spoken Language Processing*, pp. 599–602.

Vogt, R. and S. Sridharan (2004). Bayes factor scoring of GMMs for speaker verification. In *Odyssey: The Speaker and Language Recognition Workshop*, pp. 173–178.