DISCRIMINATIVELY TRAINED BAYESIAN SPEAKER COMPARISON OF I-VECTORS

Bengt J. Borgström[1] *and Alan McCree*[2]

[1]MIT Lincoln Laboratory, Lexington, MA
[2]Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, MD
jonas.borgstrom@ll.mit.edu, alan.mccree@jhu.edu

ABSTRACT

This paper presents a framework for fully Bayesian speaker comparison of i-vectors. By generalizing the train/test paradigm, we derive an analytic expression for the speaker comparison log-likelihood ratio (LLR), as well as solutions for model training and Bayesian scoring. This framework is useful for enrollment sets of any size. For the specific case of single-cut enrollment, it is shown to be mathematically equivalent to probabilistic linear discriminant analysis (PLDA). Additionally, we present discriminative training of model hyperparameters by minimizing the total cross entropy between LLRs and class labels. When applied to speaker recognition, significant performance gains are observed for various NIST SRE 2010 extended evaluation tasks.

Index Terms— Bayesian speaker comparison, speaker recognition, i-vector, discriminative training, cross entropy.

1. INTRODUCTION

Within the field of speaker recognition, the *i-vector* has been proposed as an effective method of extracting discriminative speaker and channel information in a manageable low-dimensional subspace [1]. Due to their low-dimensionality, i-vectors allow for more sophisticated channel compensation and scoring methods. In particular, the additive noise model has been used to develop probabilistic linear discriminant analysis (PLDA) scoring in the i-vector domain [2, 3].

The speaker comparison problem is to determine whether a test cut was generated by the same speaker as an enrollment set. In this paper, we present a framework for fully Bayesian speaker comparison with i-vectors, which we refer to as I-BSC. Using the additive noise model, we present an analytic expression for the log-likelihood ratio (LLR) of same-speaker to different-speaker hypotheses. Additionally, we derive accompanying solutions to model training and Bayesian scoring. When applied to speaker recognition, the single-cut enrollment case of the proposed framework is observed to be mathematically equivalent to the PLDA solutions in [2, 3]. The proposed framework, however, easily generalizes to speaker comparison with any number of enrollment cuts.

We then present discriminative training of system hyperparameters for Bayesian speaker comparison. We numerically optimize the across-class and within-class covariance matrices, along with the across-class mean, by minimizing the total cross entropy between LLRs and their underlying labels. As discussed in [3], this objective function directly addresses the aim of differentiating between same-speaker and different-speaker hypothesis. Discriminative training is shown to provide significant gains in speaker recognition performance, when Bayesian speaker comparison is applied to various NIST SRE 2010 tasks.

This paper is organized as follows. In Sec. 2, we present i-vector Bayesian speaker comparison, deriving solutions for model training and Bayesian scoring. Discriminative training of hyperparameters is discussed in Sec. 3. Sec. 4 presents experimental results for speaker recognition, and conclusions are provided in Sec. 5.

2. BAYESIAN SPEAKER COMPARISON OF I-VECTORS

In this section, we formulate the fully Bayesian speaker comparison problem, and present solutions to model training and scoring.

2.1. Statistical Framework

In this study, we assume the Gaussian additive noise model, as in [4]. Speakers are normally distributed with mean $\boldsymbol{\theta}$ and across-class covariance $\boldsymbol{\Phi}_s$

$$p(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu}; \boldsymbol{\theta}, \boldsymbol{\Phi}_s). \tag{1}$$

Observed i-vectors are degraded by an additive channel component with within-class covariance $\boldsymbol{\Phi}_c$, leading to the marginal distribution

$$p(\mathbf{w}_t) = \mathcal{N}(\mathbf{w}_t; \boldsymbol{\theta}, \boldsymbol{\Phi}_s + \boldsymbol{\Phi}_c). \tag{2}$$

In the Bayesian speaker comparison framework, an enrollment set of i-vectors from a known speaker is given as $\mathcal{D} = \{\mathbf{w}_1, \ldots, \mathbf{w}_N\}$, where $\mathbf{w}_i \in \mathbb{R}^K$ is the i^{th} i-vector provided for the known speaker. Conditioned on the model mean, the elements of $\mathcal{D}$ are assumed i.i.d., leading to the conditional distribution

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{i=1}^{N} p(\mathbf{w}_i|\boldsymbol{\mu}) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{w}_i; \boldsymbol{\mu}, \boldsymbol{\Phi}_c). \tag{3}$$

The goal of the speaker comparison problem is to determine whether a test i-vector $\mathbf{w}_t$ was produced by the given speaker. The possible hypotheses are

$\mathcal{H}_0$: $\mathcal{D}$ and $\mathbf{w}_t$ are produced by different speakers
$\mathcal{H}_1$: $\mathcal{D}$ and $\mathbf{w}_t$ are produced by the same speaker.

Using a Bayesian approach, this problem reduces to determining the log-likelihood ratio

$$\mathcal{L}(\mathbf{w}_t|\mathcal{D}) = \log \frac{p(\mathbf{w}_t|\mathcal{D}, \mathcal{H}_1)}{p(\mathbf{w}_t|\mathcal{D}, \mathcal{H}_0)}, \tag{4}$$

This work was sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

[1]now with Broadcom, Irvine CA (bjborgstrom@gmail.com)

which can be solved as

$$\mathcal{L}(\mathbf{w}_t|\mathcal{D}) = \log \frac{\int p(\mathcal{D}|\boldsymbol{\mu})p(\mathbf{w}_t|\boldsymbol{\mu})p(\boldsymbol{\mu})\partial\boldsymbol{\mu}}{p(\mathcal{D})p(\mathbf{w}_t)}. \tag{5}$$

Equivalently, using Bayes' rule, the LLR becomes [5]

$$\mathcal{L}(\mathbf{w}_t|\mathcal{D}) = \log \frac{\int p\left(\mathbf{w}_t|\boldsymbol{\mu}\right) p\left(\boldsymbol{\mu}|\mathcal{D}\right) \partial\boldsymbol{\mu}}{p\left(\mathbf{w}_t\right)}. \tag{6}$$

When expressed in this form, the speaker comparison LLR offers valuable intuition. The term $p\left(\boldsymbol{\mu}|\mathcal{D}\right)$ can be interpreted as a training step, wherein we obtain the distribution of the speaker model mean conditioned on the available data set. The term $p\left(\mathbf{w}_t|\boldsymbol{\mu}\right)$ can then be interpreted as testing, where the test cut is scored against the model mean. Finally, the denominator, $p\left(\mathbf{w}_t\right)$, represents the evaluation of a random speaker in a random channel.

2.2. Model Training

Model training consists of fitting a parametric model of the known speaker mean to the training set $\mathcal{D}$, and determining $p\left(\boldsymbol{\mu}|\mathcal{D}\right)$. The posterior distribution is given by

$$p\left(\boldsymbol{\mu}|\mathcal{D}\right) = \frac{p\left(\mathcal{D}|\boldsymbol{\mu}\right) p\left(\boldsymbol{\mu}\right)}{\int p\left(\mathcal{D}|\boldsymbol{\mu}\right) p\left(\boldsymbol{\mu}\right) \partial\boldsymbol{\mu}} \propto \left(\mathcal{D}|\boldsymbol{\mu}\right) p\left(\boldsymbol{\mu}\right). \tag{7}$$

Applying (1)-(3) leads to

$$\begin{aligned} p\left(\boldsymbol{\mu}|\mathcal{D}\right) &\propto \prod_{i=1}^{N} p\left(\mathbf{w}_i|\boldsymbol{\mu}\right) p\left(\boldsymbol{\mu}\right) \\ &\propto \exp\left(-\frac{1}{2}\left[\sum_{i=1}^{N} \left(\boldsymbol{\mu}-\boldsymbol{\mu}_{\mathcal{D}}\right)^T \boldsymbol{\Phi}_{\mathcal{D}}^{-1} \left(\boldsymbol{\mu}-\boldsymbol{\mu}_{\mathcal{D}}\right)\right]\right) \end{aligned} \tag{8}$$

where

$$\boldsymbol{\Phi}_{\mathcal{D}} = \frac{1}{N}\boldsymbol{\Phi}_s \left(\boldsymbol{\Phi}_s + \frac{1}{N}\boldsymbol{\Phi}_c\right)^{-1} \boldsymbol{\Phi}_c, \tag{9}$$

and

$$\begin{aligned} \boldsymbol{\mu}_{\mathcal{D}} =& \frac{1}{N}\boldsymbol{\Phi}_s \left(\boldsymbol{\Phi}_s + \frac{1}{N}\boldsymbol{\Phi}_c\right)^{-1} \sum_{i=1}^{N} \mathbf{w}_i \\ &+ \frac{1}{N}\boldsymbol{\Phi}_c \left(\boldsymbol{\Phi}_s + \frac{1}{N}\boldsymbol{\Phi}_c\right)^{-1} \boldsymbol{\theta}. \end{aligned} \tag{10}$$

Since $p\left(\boldsymbol{\mu}|\mathcal{D}\right)$ is a valid distribution, and must integrate to unity, it can be concluded to be normally distributed as

$$p\left(\boldsymbol{\mu}|\mathcal{D}\right) = \mathcal{N}\left(\boldsymbol{\mu}; \boldsymbol{\mu}_{\mathcal{D}}, \boldsymbol{\Phi}_{\mathcal{D}}\right). \tag{11}$$

Here, $\boldsymbol{\mu}_{\mathcal{D}}$ represents the mean of the conditional distribution $p\left(\boldsymbol{\mu}|\mathcal{D}\right)$, and $\boldsymbol{\Phi}_{\mathcal{D}}$ represents the uncertainty present when estimating $\boldsymbol{\mu}$ from the available data in $\mathcal{D}$.

2.3. Bayesian Scoring

Once the posterior distribution of model mean $\boldsymbol{\mu}$ is obtained, Bayesian scoring reduces to determining the LLR as in (6). The integral in the numerator of (6) can be interpreted as the sum of two independent normally distributed random variables, which itself is a normally distributed random variable

$$\int p\left(\mathbf{w}_t|\boldsymbol{\mu}\right) p\left(\boldsymbol{\mu}|\mathcal{D}\right) \partial\boldsymbol{\mu} = \mathcal{N}\left(\mathbf{w}_t; \boldsymbol{\mu}_{\mathcal{D}}, \boldsymbol{\Phi}_c + \boldsymbol{\Phi}_{\mathcal{D}}\right) \tag{12}$$

so that

$$\mathcal{L}(\mathbf{w}_t|\mathcal{D}) = \log \frac{\mathcal{N}\left(\mathbf{w}_t; \boldsymbol{\mu}_{\mathcal{D}}, \boldsymbol{\Phi}_c + \boldsymbol{\Phi}_{\mathcal{D}}\right)}{\mathcal{N}\left(\mathbf{w}_t; \boldsymbol{\theta}, \boldsymbol{\Phi}_s + \boldsymbol{\Phi}_c\right)} \tag{13}$$

Thus, the general case LLR can be expressed as the ratio of two Gaussian distributions.

For the single-cut training case, i.e. $N = 1$, this solution is mathematically equivalent to probabilistic linear discriminant analysis (PLDA) discussed in [2] and [3]. This can be shown by expressing the within-class covariance in (13) in terms of the total variability covariance, i.e. $\boldsymbol{\Phi}_c = \boldsymbol{\Phi}_t - \boldsymbol{\Phi}_s$, expanding the Gaussian expressions, and applying the matrix inversion lemma appropriately. For multiple enrollment cuts, [2] and [3] derive the PLDA solution by stacking i-vectors and covariance matrices, thereby requiring the analytic inversion of an $(N+1) \times (N+1)$ block matrix. This increases complexity and memory requirements with increasing N.

By contrast, the proposed solution in (13) has the same form regardless of the number of enrollment cuts. Note also that the numerator of this expression can be viewed as an application of Bayesian parameter estimation for Gaussian distributions as discussed in [5].

3. DISCRIMINATIVE TRAINING OF HYPERPARAMETERS

In Sec. 2, a generative model is used to derive the solution to Bayesian speaker comparison. In this section, we present discriminative training of model hyperparameters to better differentiate between hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$. Specifically, we wish to train the across-class mean, $\boldsymbol{\theta}$, along with the across-class and within-class covariance matrices, $\boldsymbol{\Phi}_s$ and $\boldsymbol{\Phi}_c$.

3.1. Total Cross Entropy

As the objective function, we use the total cross entropy (TCE) between output LLRs and the answer key, defined as [6]

$$\begin{aligned} TCE =& \frac{P\left(\mathcal{H}_0\right)}{|\mathcal{X}_0|} \sum_{\mathbf{w}_t \in \mathcal{X}_0} \log\left(1 + e^{\gamma + \mathcal{L}(\mathbf{w}_t|\mathcal{D})}\right) \\ &+ \frac{P\left(\mathcal{H}_1\right)}{|\mathcal{X}_1|} \sum_{\mathbf{w}_t \in \mathcal{X}_1} \log\left(1 + e^{-\gamma - \mathcal{L}(\mathbf{w}_t|\mathcal{D})}\right), \end{aligned} \tag{14}$$

where $\mathcal{X}_i$ represents the set of trials corresponding to hypothesis $\mathcal{H}_i$, $|\mathcal{X}_i|$ denotes the cardinality of $\mathcal{X}_i$, and γ is the log-ratio of priors, $\gamma = \log \frac{P(\mathcal{H}_1)}{P(\mathcal{H}_0)}$. Note that the use of priors in (14) allows the TCE to emphasize a specific operating point.

3.2. Evaluation of Gradients

To numerically optimize the objective function, we rely on the gradient descent method presented in [7] for maximizing mutual information. This algorithm requires the evaluation of the gradient of the TCE with respect to the hyperparameters $\boldsymbol{\theta}$, $\boldsymbol{\Phi}_s$, and $\boldsymbol{\Phi}_c$. The generalized gradient of TCE with respect to some vector $\boldsymbol{\psi}$ can be shown to be

$$\begin{aligned} \frac{\partial TCE}{\partial \boldsymbol{\psi}} =& -\frac{P\left(\mathcal{H}_0\right)}{|\mathcal{X}_0|} \sum_{\mathbf{w}_t \in \mathcal{X}_0} P\left(\mathcal{H}_1|\mathbf{w}_t\right) \frac{\partial \mathcal{L}(\mathbf{w}_t|\mathcal{D})}{\partial \boldsymbol{\psi}} \\ &- \frac{P\left(\mathcal{H}_1\right)}{|\mathcal{X}_1|} \sum_{\mathbf{w}_t \in \mathcal{X}_1} P\left(\mathcal{H}_0|\mathbf{w}_t\right) \frac{\partial \mathcal{L}(\mathbf{w}_t|\mathcal{D})}{\partial \boldsymbol{\psi}}. \end{aligned} \tag{15}$$

However, determining the gradient of the general case LLR, as given in (13), is mathematically difficult since $\boldsymbol{\mu}_{\mathcal{D}}$ and $\boldsymbol{\Phi}_{\mathcal{D}}$ are functions of the hyperparameters which we wish to train. Instead, we propose to approximate the exact gradient by assuming $\boldsymbol{\mu}_{\mathcal{D}}$ and $\boldsymbol{\Phi}_{\mathcal{D}}$ to be independent of $\boldsymbol{\theta}$, $\boldsymbol{\Phi}_s$, and $\boldsymbol{\Phi}_c$. The model parameters $\boldsymbol{\mu}_{\mathcal{D}}$ and $\boldsymbol{\Phi}_{\mathcal{D}}$ can then be re-estimated as a function of the updated hyperparameters after each optimization iteration. In this section we derive TCE gradients under this assumption.

Using (13), the gradient of the general case LLR with respect to the across-class mean can be approximated as

$$\frac{\partial \mathcal{L}(\mathbf{w}_t|\mathcal{D})}{\partial \boldsymbol{\theta}} = \left(\boldsymbol{\Phi}_s + \boldsymbol{\Phi}_c\right)^{-1}\left(\mathbf{w}_t - \boldsymbol{\theta}\right). \tag{16}$$

When optimizing the objective function with respect to $\boldsymbol{\Phi}_s$ and $\boldsymbol{\Phi}_c$, constraints must be applied to guarantee that the updated covariance matrices are symmetric and positive-definite. We accomplish this by updating only the corresponding eigenvalues. Applying eigendecompositions to $\boldsymbol{\Phi}_s$ and $\boldsymbol{\Phi}_c$ reveals

$$\begin{aligned}\boldsymbol{\Phi}_c &= \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T = \sum_{l=1}^{K}\lambda_l \mathbf{u}_l\mathbf{u}_l^T, \\ \boldsymbol{\Phi}_s &= \mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^T = \sum_{l=1}^{K}\sigma_l \mathbf{v}_l\mathbf{v}_l^T,\end{aligned} \tag{17}$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal bases with basis vectors $\mathbf{u}_l$ and $\mathbf{v}_l$, respectively, and $\boldsymbol{\Lambda}$ and $\boldsymbol{\Sigma}$ are diagonal matrices comprised of the eigenvalues λ_l and σ_l, respectively. Using (13) and the matrix chain rule [8], the derivative of the LLR with respect to a within-class covariance eigenvalue can be derived as

$$\begin{aligned}\frac{\partial \mathcal{L}(\mathbf{w}_t|\mathcal{D})}{\partial \lambda_l} =& -\frac{\partial/\partial\lambda_l \left|\boldsymbol{\Phi}_{\mathcal{D}} + \boldsymbol{\Phi}_c\right|}{2\left|\boldsymbol{\Phi}_{\mathcal{D}} + \boldsymbol{\Phi}_c\right|} + \frac{\partial/\partial\lambda_l \left|\boldsymbol{\Phi}_s + \boldsymbol{\Phi}_c\right|}{2\left|\boldsymbol{\Phi}_s + \boldsymbol{\Phi}_c\right|} \\ &+ \frac{1}{2}\mathrm{Tr}\left\{\left(\boldsymbol{\Phi}_{\mathcal{D}} + \boldsymbol{\Phi}_c\right)^{-1}\left(\mathbf{w}_t - \boldsymbol{\mu}_{\mathcal{D}}\right)\right. \\ &\qquad \left.\times\left(\mathbf{w}_t - \boldsymbol{\mu}_{\mathcal{D}}\right)^T\left(\boldsymbol{\Phi}_{\mathcal{D}} + \boldsymbol{\Phi}_c\right)^{-1}\mathbf{u}_l\mathbf{u}_l^T\right\} \\ &- \frac{1}{2}\mathrm{Tr}\left\{\left(\boldsymbol{\Phi}_{\mathcal{D}} + \boldsymbol{\Phi}_c\right)^{-1}\left(\mathbf{w}_t - \boldsymbol{\theta}\right)\right. \\ &\qquad \left.\times\left(\mathbf{w}_t - \boldsymbol{\theta}\right)^T\left(\boldsymbol{\Phi}_{\mathcal{D}} + \boldsymbol{\Phi}_c\right)^{-1}\mathbf{u}_l\mathbf{u}_l^T\right\} \\ =& -\frac{1}{2}\mathbf{u}_l^T\left(\boldsymbol{\Phi}_{\mathcal{D}} + \boldsymbol{\Phi}_c\right)^{-1}\mathbf{u}_l \\ &+ \frac{1}{2}\mathbf{u}_l^T\left(\boldsymbol{\Phi}_s + \boldsymbol{\Phi}_c\right)^{-1}\mathbf{u}_l \\ &+ \frac{1}{2}\left(\mathbf{u}_l^T\left(\boldsymbol{\Phi}_{\mathcal{D}} + \boldsymbol{\Phi}_c\right)^{-1}\left(\mathbf{w}_t - \boldsymbol{\mu}_{\mathcal{D}}\right)\right)^2 \\ &- \frac{1}{2}\left(\mathbf{u}_l^T\left(\boldsymbol{\Phi}_s + \boldsymbol{\Phi}_c\right)^{-1}\left(\mathbf{w}_t - \boldsymbol{\theta}\right)\right)^2.\end{aligned} \tag{18}$$

where the second step uses the cyclic property of matrix traces. Similarly, the derivative of the LLR with respect to an across-class covariance eigenvalue becomes

$$\begin{aligned}\frac{\partial \mathcal{L}(\mathbf{w}_t|\mathcal{D})}{\partial \sigma_l} =& \frac{1}{2}\mathbf{v}_l^T\left(\boldsymbol{\Phi}_s + \boldsymbol{\Phi}_c\right)^{-1}\mathbf{v}_l \\ &- \frac{1}{2}\left(\mathbf{v}_l^T\left(\boldsymbol{\Phi}_s + \boldsymbol{\Phi}_c\right)^{-1}\left(\mathbf{w}_t - \boldsymbol{\theta}\right)\right)^2.\end{aligned} \tag{19}$$

To avoid over-fitting to development data, the optimization of $\boldsymbol{\Phi}_s$ and $\boldsymbol{\Phi}_c$ can be further constrained by only updating matrix scalings. That is, the identities from (17) can be generalized as

$$\begin{aligned}\boldsymbol{\Phi}_c &= \alpha_c\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T = \alpha_c\sum_{l=1}^{K}\lambda_l\mathbf{u}_l\mathbf{u}_l^T, \\ \boldsymbol{\Phi}_s &= \alpha_s\mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^T = \alpha_s\sum_{l=1}^{K}\sigma_l\mathbf{v}_l\mathbf{v}_l^T,\end{aligned} \tag{20}$$

so that only the scaling factors α_c and α_s are updated. The derivative of the LLR with respect to these scaling factors are derived similarly to (18) and (19) as

$$\frac{\partial \mathcal{L}(\mathbf{w}_t|\mathcal{D})}{\partial \alpha_c} = \sum_{l=1}^{K}\lambda_l\frac{\partial \mathcal{L}(\mathbf{w}_t|\mathcal{D})}{\partial \lambda_l} \tag{21}$$

and

$$\frac{\partial \mathcal{L}(\mathbf{w}_t|\mathcal{D})}{\partial \alpha_s} = \sum_{l=1}^{K}\sigma_l\frac{\partial \mathcal{L}(\mathbf{w}_t|\mathcal{D})}{\partial \sigma_l} \tag{22}$$

4. EXPERIMENTAL RESULTS

This section presents experimental results for the proposed methods on NIST SRE 2010 extended evaluation tasks. The baseline system uses 600-dimensional i-vectors, with a further LDA dimension reduction to 200. Results are included for i-vector length-normalization, since it has been widely shown to provide performance improvements for speaker recognition [2]. However, since statistical modeling of i-vectors can be considered more straightforward without length normalization, results are also reported without this processing step. The background model is trained using Switchboard II as well as SRE telephone data from 2004, 2005, and 2006, and was based on 39-dimensional telephone-bandwidth cepstral features including deltas, with feature mean and variance normalization. Across-class and within-class covariance matrices are estimated from sample covariance matrices, using the same data.

Table 1 provides speaker recognition results for SRE 2010 extended evaluation single-cut enrollment telephone data. Results are reported in terms of equal error rate (EER), and the two minimum decision cost function (DCF) scores defined by [9] and [10]. The minimum DCF score from [10], referred to as minDCF, is normalized by 10^3, whereas the minimum DCF score from [9], referred to as oldDCF, is normalized by 10. Separate results are provided for male and female speakers, as well as for pooled scores. As discussed in Sec. 2.3, the PLDA baseline is mathematically equivalent to the proposed I-BSC scoring method for the single-cut enrollment case, and is therefore not included. Results are provided for I-BSC with discriminative training (DT) of the across-class mean, as well as the across-class and within-class covariance matrices. To avoid overfitting during training, covariance matrix updates included only scaling factors.

It can be observed in Table 1 that discriminative training of system hyperparameters provides significant improvements in speaker recognition for the case of i-vector length normalization, yielding approximately 10% relative improvement in EER. The effect of discriminative training is more apparent without length normalization, providing 45% relative improvement in EER. Furthermore, the use of DT seems to substantially reduce the performance gap between i-vector speaker recognition with and without length normalization.

Table 2 provides speaker recognition results for the multi-cut enrollment task. Results are provided for I-BSC scoring, with and

	Male Set			Female Set			Pooled Set		
Method	EER (%)	minDCF	oldDCF	EER (%)	minDCF	oldDCF	EER (%)	minDCF	oldDCF
With Length Normalization									
I-BSC	2.08	0.380	0.099	3.05	0.485	0.148	2.67	0.495	0.133
I-BSC + DT	1.84	0.344	0.092	2.74	0.477	0.140	2.36	0.473	0.121
Without Length Normalization									
I-BSC	5.14	0.526	0.230	5.37	0.554	0.223	5.24	0.541	0.228
I-BSC + DT	2.66	0.341	0.113	3.05	0.607	0.152	2.86	0.481	0.066

Table 1. Speaker Recognition Results for Single-Cut Enrollment Data

	Male Set			Female Set			Pooled Set		
Method	EER (%)	minDCF	oldDCF	EER (%)	minDCF	oldDCF	EER (%)	minDCF	oldDCF
With Length Normalization									
Average PLDA	0.48	0.121	0.029	1.49	0.218	0.048	1.12	0.193	0.042
I-BSC	0.48	0.156	0.036	1.86	0.256	0.068	1.36	0.252	0.057
I-BSC + DT	0.48	0.144	0.029	1.72	0.229	0.047	1.10	0.195	0.042
Without Length Normalization									
Average PLDA	2.86	0.277	0.092	2.59	0.258	0.100	2.71	0.282	0.098
I-BSC	3.33	0.375	0.113	3.45	0.324	0.135	3.39	0.346	0.126
I-BSC + DT	0.57	0.164	0.030	1.29	0.297	0.045	0.90	0.246	0.039

Table 2. Speaker Recognition Results for Multi-Cut Enrollment Data

without discriminative training of hyperparameters. It can be observed that DT provides significant performance improvements, especially when i-vector length normalization is not utilized.

It should be noted that the SRE10 multi-cut enrollment task is not completely consistent with the statistical framework assumed in this paper, since all 8 enrollment cuts for each speaker model are recorded from the same handset. Channel components can therefore not be assumed to be independently drawn according to a normal distribution with covariance $\mathbf{\Phi}_c$, so that (3) becomes invalid. For the case of multi-cut enrollment, PLDA is commonly implemented by averaging enrollment i-vectors, and is provided as a baseline. Table 2 shows I-BSC scoring to yield a performance degradation relative to the approximated PLDA baseline. This is most likely due to the previously discussed mismatch in statistical assumptions. However, the combination of I-BSC scoring with discriminative hyperparameter training overcomes this degradation for the pooled set case. Note that the use of enrollment cuts from identical handsets may not be realistic in many applications. Instead, I-BSC can be expected to perform well when enrollment cuts are sampled from various channels.

We note that [3] discriminatively trained the PLDA hyperparameters for the single-cut enrollment case using the same cross entropy metric, and reported similar gains as in Table 1. The solution presented here is more general since it includes the multiple cut enrollment case, and it also maintains the Gaussian form of the final solution.

5. CONCLUSIONS

In this paper, we have presented fully Bayesian speaker comparison with i-vectors. We derived the speaker comparison log-likelihood ratio, along with solutions for model training and Bayesian scoring. The framework is easily adaptable to the number of enrollment cuts available. Additionally, we have presented discriminative training of these model hyperparameters. When applied to speaker recognition, experiments on various NIST SRE 2010 extended evaluation tasks have shown the proposed methods to provide significant improvements in performance.

6. REFERENCES

[1] N. Dehak, P. Kenny, R. Dehak, P. Ouellet, and P. Dumouchel, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, pp. 788–798, May 2011.

[2] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech*, 2011, pp. 249–252.

[3] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka, and N. Brummer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification," in *Proc. ICASSP*, 2011, pp. 4832–4835.

[4] A. McCree, D. Sturim, and D. Reynolds, "A new perspective on GMM subspace compensation based on PPCA and Wiener filtering," in *Proc. Interspeech*, 2011, pp. 145–148.

[5] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley, 2001.

[6] N. Brummer and J. du Preez, "Application independent evaluation of speaker detection," *Speech Communication*, vol. 20, no. 2, pp. 230–275, Nov. 2006.

[7] C.-H. Lee, F. K. Song, and K. K. Paliwal, *Automatic Speech Recognition: Advanced Topics, Chapter 8*, Springer, 1996.

[8] K. B. Petersen and M. S. Pedersen, *The Matrix Cookbook*, 2006.

[9] "The NIST year 2008 speaker recognition evaluation plan," http://www.itl.nist.gov/iad/mig/tests/sre/2008.

[10] "The NIST year 2010 speaker recognition evaluation plan," http://www.itl.nist.gov/iad/mig/tests/sre/2010.