SUPERVECTOR BAYESIAN SPEAKER COMPARISON

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

In this paper we propose fully Bayesian speaker comparison of supervectors, which we refer to as SV-BSC, as a method for estimating whether a test cut was generated by the same speaker as an enrollment set. We derive the SV-BSC log-likelihood ratio of same-speaker to different-speaker hypotheses, and present solutions for model training and Bayesian scoring. We then show that if speaker and channel variability are assumed to inhabit a total variability subspace, SV-BSC scoring reduces to a form which requires only low-computation subspace operations. Finally, we show that common speaker recognition techniques such as Joint Factor Analysis (JFA) and i-vector Probabilistic Linear Discriminant Analysis (PLDA) are approximations to this full solution under certain additional assumptions. Experiments on the NIST 2010 SRE show SV-BSC to outperform a PLDA system.

Index Terms— Bayesian speaker comparison, speaker recognition, total variability, supervector, i-vector.

1. INTRODUCTION

Traditional speaker recognition systems extracted speaker-specific information using Gaussian mixture models (GMMs) [1]. Subsequent work assumed speaker and/or channel variability to inhabit lower dimension subspaces of the *supervector* of concatenated GMM means, allowing scoring to emphasize discriminative directions in supervector space [2]. Building on the subspace assumption, the *i-vector* was introduced as a low dimensional speaker-specific feature extracted from the supervector domain [3]. Due to their low-dimensionality, i-vectors allow for sophisticated modeling and scoring, and have become widely used. A drawback of the i-vector, however, is that it is derived as a point-estimate, and excludes uncertainty due to observation noise, thereby ignoring cut duration.

In this paper, we propose fully Bayesian speaker comparison of supervectors. As opposed to the use of i-vectors, SV-BSC retains duration information of enrollment and test cuts throughout modeling and scoring in the form of observation noise. Under the subspace assumption, we show SV-BSC scoring to reduce to a low computation subspace operation. We then discuss the role of SV-BSC as a generalized framework for many speaker recognition techniques, and show that certain data approximations result in existing methods such as PLDA [4], [5] and full Gaussian scoring (FGS) [6] or JFA [2]. Experimentation on the NIST 2010 SRE reveals SV-BSC to outperform PLDA.

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.

2. SUPERVECTOR BAYESIAN SPEAKER COMPARISON

In this section, we present fully Bayesian speaker comparison in the supervector space, which we refer to as SV-BSC. We introduce our statistical framework, derive the SV-BSC log-likelihood ratio, and present solutions for model training and scoring.

2.1. Statistical Framework

In this study, we build upon the use of maximum a posteriori (MAP) adaptation of GMMs[1]. We assume speaker GMMs to differ only with respect to means, and alignment to Gaussian mixtures is determined using the universal background model (UBM). We use the additive noise model, as in [7]. Speaker supervectors are normally distributed with mean $\boldsymbol{\theta}$ and across-class covariance $\boldsymbol{\Sigma}_s$, so that $p(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu}; \boldsymbol{\theta}, \boldsymbol{\Sigma}_s)$. An observed supervector $\mathbf{x}_t \in \mathbb{R}^K$ is degraded by additive channel and observation noise. The channel component is normally distributed with zero mean and within-class covariance $\boldsymbol{\Sigma}_c$. The cut-specific observation noise is normally distributed with zero mean and covariance $\boldsymbol{\Sigma}_{n,t}$, leading to the marginal distribution

$$p(\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_t; \boldsymbol{\theta}, \boldsymbol{\Sigma}_s + \boldsymbol{\Sigma}_c + \boldsymbol{\Sigma}_{n,t}). \quad (1)$$

The observation noise is defined as $\boldsymbol{\Sigma}_{n,t} = \mathbf{N}_t^{-1}\boldsymbol{\Sigma}_0$, where $\mathbf{N}_t$ is a diagonal matrix comprised of mixture counts for $\mathbf{x}_t$, and $\boldsymbol{\Sigma}_0$ is the Universal Background Model (UBM) covariance matrix [1].

In the speaker comparison framework, an enrollment set of observed supervectors is given from a known speaker, denoted by $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. Here, $\mathbf{x}_i$ is the i^{th} supervector for the given enrollment set, and is observed with noise $\boldsymbol{\Sigma}_{n,i}$. Conditioned on the model mean of the speaker, $\boldsymbol{\mu}$, the elements of $\mathcal{D}$ are assumed i.i.d., leading to the conditional distribution

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{i=1}^{N} p\left(\mathbf{x}_i|\boldsymbol{\mu}\right) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}_c + \boldsymbol{\Sigma}_{n,i}). \quad (2)$$

The goal of the speaker comparison problem is to determine whether a supervector $\mathbf{x}_t$ was produced by the same speaker as the enrollment set. The possible hypotheses are

$$\begin{aligned} \mathcal{H}_0 &: \mathcal{D} \text{ and } \mathbf{x}_t \text{ are produced by different speakers} \\ \mathcal{H}_1 &: \mathcal{D} \text{ and } \mathbf{x}_t \text{ are produced by the same speaker.} \end{aligned}$$

Using a Bayesian approach, the speaker comparison problem reduces to determining the log-likelihood ratio (LLR)

$$\mathcal{L}(\mathbf{x}_t|\mathcal{D}) = \log \frac{p\left(\mathbf{x}_t|\mathcal{D}, \mathcal{H}_1\right)}{p\left(\mathbf{x}_t|\mathcal{D}, \mathcal{H}_0\right)}, \quad (3)$$

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

which can be solved in a straight-forward manner via marginalization of the speaker model mean

$$\mathcal{L}(\mathbf{x}_t|\mathcal{D}) = \log \frac{\int p(\mathcal{D}|\boldsymbol{\mu})p(\mathbf{x}_t|\boldsymbol{\mu})p(\boldsymbol{\mu})d\boldsymbol{\mu}}{p(\mathcal{D})p(\mathbf{x}_t)}. \quad (4)$$

Equivalently, by applying Bayes' rule as in [8], the LLR can be expressed as

$$\mathcal{L}(\mathbf{x}_t|\mathcal{D}) = \log \frac{\int p\left(\mathbf{x}_t|\boldsymbol{\mu}\right) p\left(\boldsymbol{\mu}|\mathcal{D}\right) d\boldsymbol{\mu}}{p\left(\mathbf{x}_t\right)}. \quad (5)$$

When given in this form, the LLR offers valuable insight into the speaker comparison problem. The term $p\left(\boldsymbol{\mu}|\mathcal{D}\right)$ can be interpreted as an initial model training step. The term $p\left(\mathbf{x}_t|\boldsymbol{\mu}\right)$ can then be interpreted as Bayesian scoring. Finally, the denominator represents the likelihood of a random speaker in a random channel.

2.2. Model Training

Model training consists of fitting a parametric model to the training set $\mathcal{D}$, and determining the posterior probability distribution of the enrollment set speaker mean, $p\left(\boldsymbol{\mu}|\mathcal{D}\right)$. To simplify notation, sufficient statistics can be defined to fully characterize set $\mathcal{D}$. We define the 0^{th} and 1^{st} order statistics, respectively, as

$$\begin{aligned}\mathbf{A}_{\mathcal{D},0} &= \boldsymbol{\Sigma}_c \sum_{i=1}^{N} \left(\boldsymbol{\Sigma}_c + \boldsymbol{\Sigma}_{n,i}\right)^{-1} \quad (6)\\ \mathbf{A}_{\mathcal{D},1} &= \boldsymbol{\Sigma}_c \sum_{i=1}^{N} \left(\boldsymbol{\Sigma}_c + \boldsymbol{\Sigma}_{n,i}\right)^{-1} \mathbf{x}_i.\end{aligned}$$

The reason for these exact formulas will be made clear in this section.

The posterior distribution of the enrollment set speaker mean is given by $p\left(\boldsymbol{\mu}|\mathcal{D}\right) \propto p\left(\mathcal{D}|\boldsymbol{\mu}\right) p\left(\boldsymbol{\mu}\right)$. Applying (1), and (2) leads to

$$\begin{aligned}p\left(\boldsymbol{\mu}|\mathcal{D}\right) &\propto \prod_{i=1}^{N} p\left(\mathbf{x}_i|\boldsymbol{\mu}\right) p\left(\boldsymbol{\mu}\right) \quad (7)\\ &\propto \exp\left(-\frac{1}{2}\left[\sum_{i=1}^{N}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_{\mathcal{D}}\right)^T \boldsymbol{\Sigma}_{\mathcal{D}}^{-1}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_{\mathcal{D}}\right)\right]\right),\end{aligned}$$

where

$$\boldsymbol{\Sigma}_{\mathcal{D}} = \boldsymbol{\Sigma}_s\left(\boldsymbol{\Sigma}_s + \mathbf{A}_{\mathcal{D},0}^{-1}\boldsymbol{\Sigma}_c\right)^{-1}\mathbf{A}_{\mathcal{D},0}^{-1}\boldsymbol{\Sigma}_c, \quad (8)$$

and

$$\begin{aligned}\boldsymbol{\mu}_{\mathcal{D}} =& \boldsymbol{\Sigma}_s\left(\boldsymbol{\Sigma}_s + \mathbf{A}_{\mathcal{D},0}^{-1}\boldsymbol{\Sigma}_c\right)^{-1}\mathbf{A}_{\mathcal{D},0}^{-1}\mathbf{A}_{\mathcal{D},1} \quad (9)\\ &+ \mathbf{A}_{\mathcal{D},0}^{-1}\boldsymbol{\Sigma}_c\left(\boldsymbol{\Sigma}_s + \mathbf{A}_{\mathcal{D},0}^{-1}\boldsymbol{\Sigma}_c\right)^{-1}\boldsymbol{\theta}.\end{aligned}$$

Since $p\left(\boldsymbol{\mu}|\mathcal{D}\right)$ is a valid distribution, and must integrate to unity, it can be concluded that $p\left(\boldsymbol{\mu}|\mathcal{D}\right) = \mathcal{N}\left(\boldsymbol{\mu};\boldsymbol{\mu}_{\mathcal{D}},\boldsymbol{\Sigma}_{\mathcal{D}}\right)$. Here, $\boldsymbol{\mu}_{\mathcal{D}}$ represents the mean of the conditional distribution $p\left(\boldsymbol{\mu}|\mathcal{D}\right)$, and $\boldsymbol{\Sigma}_{\mathcal{D}}$ represents the uncertainty present when estimating $\boldsymbol{\mu}$ from the available data in $\mathcal{D}$. Note that $\mathbf{A}_{\mathcal{D},0}$ and $\mathbf{A}_{\mathcal{D},1}$ are the only speaker-dependent terms in (8) and (9), and can be considered sufficient statistics since they fully parameterize the enrollment set $\mathcal{D}$.

It is interesting to note that (7)-(9) represent a generalized version of Bayesian parameter estimation discussed in [8], where our framework includes nonstationary, cut-specific observation noise. For the special case where $\boldsymbol{\Sigma}_{n,i} = \mathbf{0}$, the two are equivalent.

2.3. Bayesian Scoring

Once the posterior distribution of the enrollment speaker model mean $\boldsymbol{\mu}$ is obtained, Bayesian scoring reduces to determining the LLR in (5). The integral in the numerator of (5) can be interpreted as the sum of two independent normally distributed random variables, which itself is a normally distributed random variable

$$\int p\left(\mathbf{x}_t|\boldsymbol{\mu}\right) p\left(\boldsymbol{\mu}|\mathcal{D}\right) d\boldsymbol{\mu} = \mathcal{N}\left(\mathbf{x}_t;\boldsymbol{\mu}_{\mathcal{D}},\boldsymbol{\Sigma}_c + \boldsymbol{\Sigma}_{\mathcal{D}} + \boldsymbol{\Sigma}_{n,t}\right) \quad (10)$$

so that

$$\mathcal{L}(\mathbf{x}_t|\mathcal{D}) = \log \frac{\mathcal{N}\left(\mathbf{x}_t;\boldsymbol{\mu}_{\mathcal{D}},\boldsymbol{\Sigma}_c + \boldsymbol{\Sigma}_{\mathcal{D}} + \boldsymbol{\Sigma}_{n,t}\right)}{\mathcal{N}\left(\mathbf{x}_t;\boldsymbol{\theta},\boldsymbol{\Sigma}_s + \boldsymbol{\Sigma}_c + \boldsymbol{\Sigma}_{n,t}\right)}. \quad (11)$$

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

3. SV-BSC UNDER THE SUBSPACE ASSUMPTION

In the previous section, Bayesian speaker comparison is derived in the supervector space. Due to the high dimensionality of supervectors, the computational load required by such a system may be prohibitively large. Instead, many studies have assumed speaker and channel variabilities to lie within subspaces [2], [7]. In this section, we derive supervector Bayesian speaker comparison for the case when speaker and channel variabilities are assumed to inhabit a reduced dimension subspace.

When performing analysis of subspaces, it is often helpful to use the Woodbury matrix inversion lemma, which can be used to derive the following identity

$$\begin{aligned}&\mathbf{V}^T\left(\mathbf{C} + \mathbf{V}\mathbf{D}\mathbf{V}^T\right)^{-1} \quad (12)\\ &\quad = \mathbf{D}^{-1}\left(\mathbf{D}^{-1} + \mathbf{V}^T\mathbf{C}^{-1}\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{C}^{-1}\end{aligned}$$

for invertible matrices $\mathbf{C}$ and $\mathbf{D}$. By substituting $\mathbf{C} = \epsilon\mathbf{I}$ into (12), we obtain

$$\begin{aligned}&\lim_{\epsilon\to 0}\mathbf{V}^T\left(\epsilon\mathbf{I} + \mathbf{V}\mathbf{D}\mathbf{V}^T\right)^{-1} \quad (13)\\ &\quad = \lim_{\epsilon\to 0}\frac{1}{\epsilon}\mathbf{D}^{-1}\left(\mathbf{D}^{-1} + \frac{1}{\epsilon}\mathbf{V}^T\mathbf{V}\right)^{-1}\mathbf{V}^T = \mathbf{D}^{-1}\mathbf{V}^+,\end{aligned}$$

where the $^+$ operator denotes the Moore-Penrose pseudoinverse.

3.1. The Subspace Assumption

Under the subspace assumption, variability covariances become

$$\begin{aligned}\boldsymbol{\Sigma}_s &= \mathbf{U}\boldsymbol{\Phi}_s\mathbf{U}^T \quad (14)\\ \boldsymbol{\Sigma}_c &= \mathbf{U}\boldsymbol{\Phi}_c\mathbf{U}^T.\end{aligned}$$

where $\mathbf{U} \in \mathbb{R}^{K\times Q}$ defines the total variability subspace [3], and where $\boldsymbol{\Phi}_c \in \mathbb{R}^{Q\times Q}$ and $\boldsymbol{\Phi}_s \in \mathbb{R}^{Q\times Q}$ denote the subspace within-class and across-class covariance matrices, respectively. In order to derive certain results in this section, the speaker variability model will at times be extended to include a diagonal tail $\boldsymbol{\Sigma}_s = \mathbf{U}\boldsymbol{\Phi}_s\mathbf{U}^T + \epsilon\mathbf{I}$ in the limit $\epsilon \to 0$. It is assumed that the across-class mean exists within the total variability subspace, so that $\boldsymbol{\theta} = \mathbf{U}\boldsymbol{\gamma}$, where $\boldsymbol{\gamma}$ is the across-class mean in subspace defined by $\mathbf{U}$. Furthermore, it is assumed that $\boldsymbol{\Phi}_c$ is full rank and invertible.

We derive subspace versions of the sufficient statistics $\mathbf{A}_{\mathcal{D},0}$ and $\mathbf{A}_{\mathcal{D},1}$ by projecting them into the total variability space. Using (12) and (14), the subspace sufficient statistics are given by

$$\begin{aligned}\mathbf{B}_{\mathcal{D},0} &= \mathbf{U}^{+}\mathbf{A}_{\mathcal{D},0}\mathbf{U} \\ &= \sum_{i=1}^{N}\left(\mathbf{\Phi}_c^{-1} + \mathbf{U}^T\mathbf{\Sigma}_{n,i}^{-1}\mathbf{U}\right)^{-1}\mathbf{U}^T\mathbf{\Sigma}_{n,i}^{-1}\mathbf{U},\end{aligned} \tag{15}$$

and

$$\begin{aligned}\mathbf{B}_{\mathcal{D},1} &= \mathbf{U}^{+}\mathbf{A}_{\mathcal{D},1} \\ &= \sum_{i=1}^{N}\left(\mathbf{\Phi}_c^{-1} + \mathbf{U}^T\mathbf{\Sigma}_{n,i}^{-1}\mathbf{U}\right)^{-1}\mathbf{U}^T\mathbf{\Sigma}_{n,i}^{-1}\mathbf{x}_i.\end{aligned} \tag{16}$$

3.2. Subspace SV-BSC

Under the subspace assumption, and using (12) and (13), the model mean posterior covariance simplifies to

$$\mathbf{\Sigma}_{\mathcal{D}} = \mathbf{U}\mathbf{\Phi}_s\left(\mathbf{\Phi}_s + \mathbf{B}_{\mathcal{D},0}^{-1}\mathbf{\Phi}_c\right)^{-1}\mathbf{B}_{\mathcal{D},0}^{-1}\mathbf{\Phi}_c\mathbf{U}^T. \tag{17}$$

and the posterior mean can be expressed as

$$\begin{aligned}\boldsymbol{\mu}_{\mathcal{D}} = \mathbf{U}\Big[&\mathbf{\Phi}_s\left(\mathbf{\Phi}_s + \mathbf{B}_{\mathcal{D},0}^{-1}\mathbf{\Phi}_c\right)^{-1}\mathbf{B}_{\mathcal{D},0}^{-1}\mathbf{B}_{\mathcal{D},1} \\ &+\mathbf{B}_{\mathcal{D},0}^{-1}\mathbf{\Phi}_c\left(\mathbf{\Phi}_s + \mathbf{B}_{\mathcal{D},0}^{-1}\mathbf{\Phi}_c\right)^{-1}\boldsymbol{\gamma}\Big].\end{aligned} \tag{18}$$

Note that the posterior mean and covariance both exist in the low dimensional subspace defined by $\mathbf{U}$. Subspace versions of these parameters are defined as

$$\mathbf{\Phi}_{\mathcal{D}} = \mathbf{\Phi}_s\left(\mathbf{\Phi}_s + \mathbf{B}_{\mathcal{D},0}^{-1}\mathbf{\Phi}_c\right)^{-1}\mathbf{B}_{\mathcal{D},0}^{-1}\mathbf{\Phi}_c, \tag{19}$$

and

$$\begin{aligned}\boldsymbol{\xi}_{\mathcal{D}} =&\mathbf{\Phi}_s\left(\mathbf{\Phi}_s + \mathbf{B}_{\mathcal{D},0}^{-1}\mathbf{\Phi}_c\right)^{-1}\mathbf{B}_{\mathcal{D},0}^{-1}\mathbf{B}_{\mathcal{D},1} \\ &+\mathbf{B}_{\mathcal{D},0}^{-1}\mathbf{\Phi}_c\left(\mathbf{\Phi}_s + \mathbf{B}_{\mathcal{D},0}^{-1}\mathbf{\Phi}_c\right)^{-1}\boldsymbol{\gamma},\end{aligned} \tag{20}$$

so that $\mathbf{\Sigma}_{\mathcal{D}} = \mathbf{U}\mathbf{\Phi}_{\mathcal{D}}\mathbf{U}^T$ and $\boldsymbol{\mu}_{\mathcal{D}} = \mathbf{U}\boldsymbol{\xi}_{\mathcal{D}}$. In the above expressions, the term $\mathbf{B}_{\mathcal{D},0}^{-1}\mathbf{\Phi}_c$ represents uncertainty due both to channel variability and observation noise of the enrollment set. For long duration cuts, when observation noise is small, this term approaches $\mathbf{\Phi}_c$. Conversely, for shorter cuts, the term represents an amplified version of the within-class covariance matrix.

Having shown that the subspace assumption leads the posterior mean and covariance matrix to both exist within the total variability subspace, we now prove that the SV-BSC LLR also exists within this subspace. Expansion of the Gaussian distributions in (11) leads to

$$\begin{aligned}\mathcal{L}(\mathbf{x}_t|\mathcal{D}) = &-\frac{1}{2}\log\frac{|\mathbf{\Sigma}_{n,t} + \mathbf{\Sigma}_c + \mathbf{\Sigma}_{\mathcal{D}}|}{|\mathbf{\Sigma}_{n,t} + \mathbf{\Sigma}_c + \mathbf{\Sigma}_s|} \\ &-\frac{1}{2}\boldsymbol{\mu}_{\mathcal{D}}^T\left(\mathbf{\Sigma}_c + \mathbf{\Sigma}_{\mathcal{D}} + \mathbf{\Sigma}_{n,t}\right)^{-1}\left(\boldsymbol{\mu}_{\mathcal{D}} - 2\mathbf{x}_t\right) \\ &+\frac{1}{2}\boldsymbol{\theta}^T\left(\mathbf{\Sigma}_s + \mathbf{\Sigma}_c + \mathbf{\Sigma}_{n,t}\right)^{-1}\left(\boldsymbol{\theta} - 2\mathbf{x}_t\right) \\ &-\frac{1}{2}\mathbf{x}_t^T\left[\left(\mathbf{\Sigma}_s + \mathbf{\Sigma}_c + \mathbf{\Sigma}_{n,t}\right)^{-1}\right. \\ &\left.-\left(\mathbf{\Sigma}_c + \mathbf{\Sigma}_{\mathcal{D}} + \mathbf{\Sigma}_{n,t}\right)^{-1}\right]\mathbf{x}_t.\end{aligned} \tag{21}$$

Applying (12), (14), and Sylvester's determinant theorem [9], the first term becomes

$$-\frac{1}{2}\log\frac{|\mathbf{\Sigma}_{n,t} + \mathbf{\Sigma}_c + \mathbf{\Sigma}_{\mathcal{D}}|}{|\mathbf{\Sigma}_{n,t} + \mathbf{\Sigma}_c + \mathbf{\Sigma}_s|} = -\frac{1}{2}\log\frac{\left|\mathbf{\Phi}_{\mathcal{D}} + \mathbf{B}_{\mathcal{D},0}^{-1}\mathbf{\Phi}_c\right|}{\left|\mathbf{\Phi}_s + \mathbf{B}_{\mathcal{D},0}^{-1}\mathbf{\Phi}_c\right|}. \tag{22}$$

Using (12) and (14), the second term in (21) reduces to

$$\begin{aligned}&-\frac{1}{2}\boldsymbol{\mu}_{\mathcal{D}}^T\left(\mathbf{\Sigma}_c + \mathbf{\Sigma}_{\mathcal{D}} + \mathbf{\Sigma}_{n,t}\right)^{-1}\left(\boldsymbol{\mu}_{\mathcal{D}} - 2\mathbf{x}_t\right) \\ &\quad= -\frac{1}{2}\boldsymbol{\xi}_{\mathcal{D}}^T\left(\mathbf{\Phi}_{\mathcal{D}} + \mathbf{B}_{t,0}^{-1}\mathbf{\Phi}_c\right)^{-1}\left(\boldsymbol{\xi}_{\mathcal{D}} - 2\mathbf{z}_t\right),\end{aligned} \tag{23}$$

where

$$\mathbf{z}_t = \mathbf{B}_{t,0}^{-1}\mathbf{B}_{t,1} = \left(\mathbf{U}^T\mathbf{\Sigma}_{n,t}^{-1}\mathbf{U}\right)^{-1}\mathbf{U}^T\mathbf{\Sigma}_{n,t}^{-1}\mathbf{x}_t. \tag{24}$$

Note that $\mathbf{z}_t$ is equivalent to the pseudoinverse form of the total variability i-vector proposed in [6]. Similarly, the third term is simplified as

$$\begin{aligned}&\frac{1}{2}\boldsymbol{\theta}^T\left(\mathbf{\Sigma}_s + \mathbf{\Sigma}_c + \mathbf{\Sigma}_{n,t}\right)^{-1}\left(\boldsymbol{\theta} - 2\mathbf{x}_t\right) \\ &\quad=\frac{1}{2}\boldsymbol{\gamma}^T\left(\mathbf{\Phi}_s + \mathbf{B}_{t,0}^{-1}\mathbf{\Phi}_c\right)^{-1}\left(\boldsymbol{\gamma} - 2\mathbf{z}_t\right)\end{aligned} \tag{25}$$

Finally, applying (12) and (14) reduces the fourth term to

$$\begin{aligned}&-\frac{1}{2}\mathbf{x}_t^T\left[\left(\mathbf{\Sigma}_s + \mathbf{\Sigma}_c + \mathbf{\Sigma}_{n,t}\right)^{-1} - \left(\mathbf{\Sigma}_c + \mathbf{\Sigma}_{\mathcal{D}} + \mathbf{\Sigma}_{n,t}\right)^{-1}\right]\mathbf{x}_t \\ &\quad= -\frac{1}{2}\mathbf{z}_t^T\left[\left(\mathbf{\Phi}_{\mathcal{D}} + \mathbf{B}_{t,0}^{-1}\mathbf{\Phi}_c\right)^{-1} - \left(\mathbf{\Phi}_s + \mathbf{B}_{t,0}^{-1}\mathbf{\Phi}_c\right)^{-1}\right]\mathbf{z}_t\end{aligned} \tag{26}$$

By substituting (22)-(26) into (21), and grouping terms into Gaussian distributions, the LLR from (11) reduces to

$$\mathcal{L}(\mathbf{x}_t|\mathcal{D}) = \log\frac{\mathcal{N}\left(\mathbf{z}_t; \boldsymbol{\xi}_{\mathcal{D}}, \mathbf{\Phi}_{\mathcal{D}} + \mathbf{B}_{t,0}^{-1}\mathbf{\Phi}_c\right)}{\mathcal{N}\left(\mathbf{z}_t; \boldsymbol{\gamma}, \mathbf{\Phi}_s + \mathbf{B}_{t,0}^{-1}\mathbf{\Phi}_c\right)}. \tag{27}$$

Thus, if speaker and channel variability exits within a subspace, the SV-BSC LLR represents a low-computation operation involving subspace parameters, hyperparameters, and sufficient statistics.

4. RELATIONSHIP TO EXISTING TECHNIQUES

In this section we show that SV-BSC serves as a generalized framework for many well-known speaker recognition methods. Under certain data approximations, subspace SV-BSC reduces to specific existing scoring techniques.

4.1. The Long-Duration Approximation: I-vector BSC

There may exist scenarios when adequately long enrollment and test cuts are available for speaker comparison. In such cases, each GMM mixture of the UBM is sampled a large number of times during MAP adaptation of supervectors. The definition of observation noise from Sec. 2.1 can be expressed alternatively as $\mathbf{\Sigma}_{n,i} = \frac{1}{T}\bar{\mathbf{N}}_i^{-1}\mathbf{\Sigma}_0$, where T is the number of frames used to obtain supervector $\mathbf{x}_i$, and $\bar{\mathbf{N}}$ contains the frame-averaged mixture counts. In the limit as $T \to \infty$, corresponding to long-duration cuts, observation noise disappears, i.e. $\mathbf{\Sigma}_{n,i} = \mathbf{0}$.

	Male Set			Female Set			Pooled Set		
Method	EER (%)	minDCF	oldDCF	EER (%)	minDCF	oldDCF	EER (%)	minDCF	oldDCF
With Length Normalization									
PLDA	2.08	0.380	0.099	3.05	0.485	0.148	2.67	0.495	0.133
SV-BSC	1.94	0.372	0.094	2.81	0.481	0.142	2.45	0.474	0.124
Without Length Normalization									
PLDA	5.14	0.526	0.230	5.37	0.554	0.223	5.24	0.541	0.228
SV-BSC	4.48	0.496	0.207	4.89	0.529	0.206	4.70	0.526	0.208

Table 1. Speaker Recognition Results for SV-BSC

If enrollment and test cuts are assumed long in duration, so that observation noise approaches zero, then the 0^{th} order statistics become $\mathbf{B}_{\mathcal{D},0} = N\mathbf{I}$ and $\mathbf{B}_{t,0} = \mathbf{I}$. Using the long duration approximation, the subspace hyperparameters from (19) and (20) reduce to

$$\boldsymbol{\Phi}_{\mathcal{D}} = \frac{1}{N}\boldsymbol{\Phi}_s \left(\boldsymbol{\Phi}_s + \frac{1}{N}\boldsymbol{\Phi}_c\right)^{-1} \boldsymbol{\Phi}_c, \tag{28}$$

and

$$\begin{aligned}\boldsymbol{\xi}_{\mathcal{D}} =& \frac{1}{N}\boldsymbol{\Phi}_s \left(\boldsymbol{\Phi}_s + \frac{1}{N}\boldsymbol{\Phi}_c\right)^{-1} \mathbf{B}_{\mathcal{D},1} \\ &+ \frac{1}{N}\boldsymbol{\Phi}_c \left(\boldsymbol{\Phi}_s + \frac{1}{N}\boldsymbol{\Phi}_c\right)^{-1} \boldsymbol{\gamma}.\end{aligned} \tag{29}$$

and the speaker comparison LLR becomes

$$\mathcal{L}(\mathbf{x}_t|\mathcal{D}) = \log \frac{\mathcal{N}\left(\mathbf{B}_{t,1}; \boldsymbol{\xi}_{\mathcal{D}}, \boldsymbol{\Phi}_{\mathcal{D}} + \boldsymbol{\Phi}_c\right)}{\mathcal{N}\left(\mathbf{B}_{t,1}; \boldsymbol{\gamma}, \boldsymbol{\Phi}_s + \boldsymbol{\Phi}_c\right)}. \tag{30}$$

For the case of single-cut enrollment, i.e. $N = 1$, the scoring method defined by (30) is similar to PLDA [4], [5]. If the term $\mathbf{B}_{t,1}$ is replaced by the total variability i-vector from [3], the two methods are equivalent.

4.2. The Large Enrollment Set Approximation: Full Gaussian Scoring

The commonly used train-test paradigm for speaker recognition assumes a perfectly known model, which theoretically requires the availability of infinite enrollment data. This corresponds to the proposed SV-BSC framework in the limit $N \to \infty$, so that $\boldsymbol{\Phi}_{\mathcal{D}} = \mathbf{0}$ and $\boldsymbol{\xi}_{\mathcal{D}} = \mathbf{z}_{\mathcal{D}}$. The LLR then reduces to

$$\mathcal{L}(\mathbf{x}_t|\mathcal{D}) = \log \frac{\mathcal{N}\left(\mathbf{z}_t; \mathbf{z}_{\mathcal{D}}, \mathbf{B}_{t,0}^{-1}\boldsymbol{\Phi}_c\right)}{\mathcal{N}\left(\mathbf{z}_t; \boldsymbol{\gamma}, \boldsymbol{\Phi}_s + \mathbf{B}_{t,0}^{-1}\boldsymbol{\Phi}_c\right)}. \tag{31}$$

Note that the scoring technique defined by (31) is equivalent to full Gaussian scoring (FGS) proposed in [6], or to JFA using a point estimate for speaker factors and integrating over channel factors [2].

5. EXPERIMENTAL RESULTS

This section presents experimental results for SV-BSC on the NIST SRE 2010 extended evaluation[10]. The baseline speaker recognition system uses 39-dimensional telephone-bandwidth cepstral features including deltas, with feature mean and variance normalization. The background model is trained using Switchboard II as well as SRE telephone data from 2004, 2005, and 2006. As a baseline system, we use full-rank PLDA scoring with 600 dimensional i-vectors and a further LDA dimension reduction to 200. Total variability $\mathbf{U}$ is estimated using the same data as for the background model, as are the subspace across-class and within-class sample covariance matrices $\boldsymbol{\Phi}_s$ and $\boldsymbol{\Phi}_c$. Some results include i-vector length normalization [5]. Since statistical modeling of i-vectors can be considered more straightforward without length normalization, results are also reported without this processing step. For the case of SV-BSC, length normalization is performed by normalizing $\mathbf{B}_{t,1}$ to unit length.

Table 1 provides speaker recognition results for SRE 2010 telephone data with single-cut enrollment. Results are reported in terms of equal error rate (EER), as well as the 2010 NIST SRE minimum decision cost function (minDCF) score, normalized by 10^3. It can be observed that SV-BSC yields improved performance across the reported conditions, relative to the baseline. Specifically, SV-BSC provides 8%-12% relative improvements in EER.

6. CONCLUSION

This paper has proposed fully Bayesian speaker comparison of supervectors as a method for estimating whether a test cut was generated by the same speaker as an enrollment set. We have derived the same-speaker to different-speaker log-likelihood ratio, and presented solutions for model training and Bayesian scoring. Under the subspace assumption, SV-BSC scoring is shown to exist within the total variability subspace, requiring only low-computation subspace operations. Common speaker recognition techniques such as JFA and i-vector PLDA are approximations to this full solution under certain additional assumptions. Finally, experiments on the NIST 2010 SRE show the promise of this new technique.

7. REFERENCES

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[2] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," Tech. Rep. CRIM-06/08-13, CRIM, 2005.

[3] 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.

[4] 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.

[5] 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.

[6] A. McCree and B. J. Borgstrom, "Supervector LDA: A new approach to reduced-complexity i-vector language recognition," in *Proc. Interspeech*, 2012.

[7] 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.

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

[9] A. Laub, *Matrix Analysis for Scientists and Engineers*, SIAM, 2004.

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