INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Unsupervised Bayesian Adaptation of PLDA for Speaker Verification

Bengt J. Borgström

MIT Lincoln Laboratory, USA
jonas.borgstrom@ll.mit.edu

Abstract

This paper presents a Bayesian framework for unsupervised domain adaptation of Probabilistic Linear Discriminant Analysis (PLDA). By interpreting class labels as latent random variables, Variational Bayes (VB) is used to derive a maximum *a posterior* (MAP) solution of the adapted PLDA model when labels are missing, referred to as VB-MAP. The VB solution iteratively infers class labels and updates PLDA hyperparameters, offering a systematic framework for dealing with unlabeled data. While presented as a general solution, this paper includes experimental results for domain adaptation in speaker verification. VB-MAP estimation is applied to the 2016 and 2018 NIST Speaker Recognition Evaluations (SREs), both of which included small and unlabeled in-domain data sets, and is shown to provide performance improvements over a variety of state-of-the-art domain adaptation methods. Additionally, VB-MAP estimation is used to train a fully unsupervised PLDA model, suffering only minor performance degradation relative to conventional supervised training, offering promise for training PLDA models when no relevant labeled data exists.

1. Introduction

Probabilistic linear discriminant analysis is a likelihood ratio test between same-class and different-class hypotheses in a verification task, and is commonly used for applications such as face verification [1, 2] and speaker verification [3–5]. Training a PLDA model requires a data set which captures across-class and within-class variability, and includes corresponding class labels. Typically the expectation-maximization (EM) algorithm is used [1] .

It has been widely shown that the performance of speaker verification systems degrades when facing unseen conditions [6–10]. This is caused in part by a mismatch between the out-of-domain data used to train PLDA hyperparameters, and the in-domain data encountered during enrollment and testing. While large labeled data sets may exist for some commonly encountered domains, labeled data may be difficult to obtain for other conditions. In such cases, unsupervised domain adaptation can be used to mitigate performance degradation due to domain mismatch.

Various methods have been explored for unsupervised adaptation of PLDA in speaker verification. Early approaches relied on clustering of training samples to find approximated speaker labels [9,11]. Other studies designed transformations to align the second-order statistics of data across domains [12–15]. Finally, Bayesian approaches have been proposed [10, 16]. An overview of unsupervised domain adaptation of PLDA for speaker verification is provided in [17]

This paper proposes a Bayesian framework for unsupervised adaptation of PLDA. By interpreting class labels as random variables, a Variational Bayes solution is derived for PLDA estimation when labels are missing. While presented as a general solution, the method is studied in the context of speaker verification, and is shown to provide significant performance improvements on a variety of tasks. Specifically, it is applied to the 2016 and 2018 NIST SREs, which both included small and unlabeled in-domain data sets, and is shown to outperform a variety of state-of-the-art approaches. Additionally, the proposed method is used to train a fully unsupervised PLDA model, suffering only minor performance degradation relative to conventional supervised training.

This paper is organized as follows: Section 2 introduces a generative model for unlabeled data. Section 3 derives VB-MAP estimation for unsupervised PLDA adaptation. Experimental results are presented in Section 4, and conclusions are provided in Section 5.

2. Statistical Framework

In this section, a generative model for unlabelled speech data is presented. Let $\mathbf{x} \in \mathbb{R}^D$ denote a speaker embedding. The additive model is assumed, $\mathbf{x}$=$\mathbf{y} + \mathbf{c}$, where $\mathbf{y}$ and $\mathbf{c}$ are the speaker and channel components of $\mathbf{x}$, respectively. Speaker and channel components are drawn from Normal distributions, which is consistent with the two-covariance PLDA model as in [3, 4, 7, 18]. A set of training data, $\mathcal{X}$, is generated by first drawing a set of speakers, $\mathcal{Y}$, according to

$$p\left(\mathcal{Y} \mid \boldsymbol{\mu}, \mathbf{B}\right) = \prod_{m=1}^{M} \mathcal{N}\left(\mathbf{y}_m; \boldsymbol{\mu}, \mathbf{B}^{-1}\right), \tag{1}$$

where $\boldsymbol{\mu}$ and $\mathbf{B}$ are the mean and precision matrix of the speaker distribution.

Since $\mathcal{X}$ is unlabeled, speaker labels are interpreted as latent random variables. Specifically, let $\mathbf{z}_n$ be a Categorical random variable which indicates the latent speaker label of $\mathbf{x}_n$ so that $z_{n,m}$=1 implies that $\mathbf{x}_n$ belongs to speaker m. Let $\mathcal{Z}$=$\{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ denote the set of latent labels corresponding to $\mathcal{X}$. Channel components are then drawn as

$$p\left(\mathcal{X} \mid \mathcal{Y}, \mathcal{Z}, \mathbf{W}\right) = \prod_{n=1}^{N} \prod_{m=1}^{M} \mathcal{N}\left(\mathbf{x}_n \mid \mathbf{y}_m, \mathbf{W}^{-1}\right)^{z_{n,m}}, \tag{2}$$

where $\mathbf{W}$ is the within-class precision matrix of the channel distribution. In the proposed generative model, $\mathcal{X}$ are the observations, $\mathcal{U}$=$\{\mathcal{Y}, \mathcal{Z}\}$ are the latent random variables, and $\mathcal{M}$=$\{\boldsymbol{\mu}, \mathbf{B}, \mathbf{W}\}$ is the set of PLDA hyperparameters. Figure 1 provides the graphical model representation of the proposed statistical framework.

2.1. Model Prior Distributions

In this paper, Bayesian estimation is leveraged for training a PLDA model, and all hyperparameters in $\mathcal{M}$ are therefore interpreted as latent random variables. A Normal-Wishart prior is assumed for $\{\boldsymbol{\mu}, \mathbf{B}\}$,

$$p\left(\boldsymbol{\mu}, \mathbf{B}\right) = \mathcal{N}\left(\boldsymbol{\mu}; \mathbf{0}, (\beta\mathbf{B})^{-1}\right) \mathcal{W}\left(\mathbf{B}; \mathbf{B}_o/\beta, \beta\right), \tag{3}$$

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-33

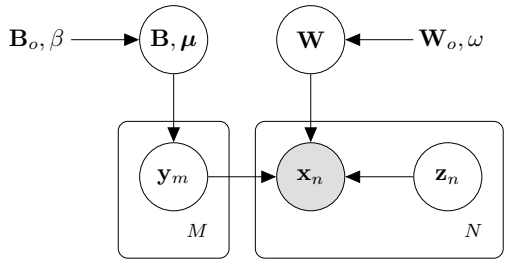

Figure 1: *The Graphical Model for the Proposed Statistical Framework with PLDA Model Parameters* $\mathcal{M}$=$\{\boldsymbol{\mu}, \mathbf{B}, \mathbf{W}\}$

where scale parameters are chosen so that $E\{\boldsymbol{\mu}\} = \mathbf{0}$ and $E\{\mathbf{B}\} = \mathbf{B}_o$[1]. Similarly, a Wishart prior is assumed for $\mathbf{W}$,

$$p(\mathbf{W}) = \mathcal{W}(\mathbf{W}; \mathbf{W}_o/\omega, \omega), \tag{4}$$

where scale parameters are chosen so that $E\{\mathbf{W}\} = \mathbf{W}_o$. Let $\mathcal{M}_o$=$\{\mathbf{0}, \mathbf{B}_o, \mathbf{W}_o\}$ denote the set of hyperparameters from the prior distributions. Finally, a non-informative prior is assumed for $\mathcal{Z}$, since information regarding the likelihoods of individual speakers in the training set is rarely relevant.The total data log-likelihood of the observations and latent variables is

$$\begin{aligned}\log p(\mathcal{X}, \mathcal{U}, \mathcal{M}) = &\log p(\mathcal{X} \mid \mathcal{Y}, \mathcal{Z}, \mathbf{W}) \\ &+ \log p(\mathcal{Y} \mid \boldsymbol{\mu}, \mathbf{B}) \\ &+ \log p(\boldsymbol{\mu}, \mathbf{B}) \\ &+ \log p(\mathbf{W}) + \text{const.},\end{aligned} \tag{5}$$

where the additive constant is due to the non-informative prior of $\mathcal{Z}$. Note that the statistical framework assumed in this paper is a special case of that proposed in [19].

3. VB-MAP Estimation

This section presents a method for Bayesian Estimation of the PLDA model introduced in Section 2. PLDA hyperparameters are determined as their MAP estimates

$$\hat{\mathcal{M}} = \underset{\mathcal{M}}{\operatorname{argmax}}\ p(\mathcal{M} \mid \mathcal{X}). \tag{6}$$

The VB-MAP estimate of $\mathcal{M}$ requires the joint posterior distribution of the hyperparameters, which is intractable. Instead, the joint distribution is approximated using Variational Bayes [20], leading to the factorized form

$$\begin{aligned}p(\mathcal{M} \mid \mathcal{X}) &\propto p(\mathcal{U}, \mathcal{M} \mid \mathcal{X}) \\ &\approx q(\mathcal{Y})\, q(\mathcal{Z})\, q(\boldsymbol{\mu}, \mathbf{B})\, q(\mathbf{W}),\end{aligned} \tag{7}$$

where the factors are determined by the following iterations

$$\begin{aligned}&1)\ \log q(\mathcal{Y}) = E_{\mathcal{U}\setminus\mathcal{Y},\mathcal{M}\mid\mathcal{X},\mathcal{Y}}\{\log p(\mathcal{X}, \mathcal{U}, \mathcal{M})\} \\ &2)\ \log q(\mathcal{Z}) = E_{\mathcal{U}\setminus\mathcal{Z},\mathcal{M}\mid\mathcal{X},\mathcal{Z}}\{\log p(\mathcal{X}, \mathcal{U}, \mathcal{M})\} \\ &3)\ \log q(\mathbf{W}) = E_{\mathcal{U},\mathcal{M}\setminus\mathbf{W}\mid\mathcal{X},\mathbf{W}}\{\log p(\mathcal{X}, \mathcal{U}, \mathcal{M})\} \\ &4)\ \log q(\boldsymbol{\mu}, \mathbf{B}) = E_{\mathcal{U},\mathcal{M}\setminus\{\boldsymbol{\mu},\mathbf{B}\}\mid\mathcal{X},\boldsymbol{\mu},\mathbf{B}}\{\log p(\mathcal{X}, \mathcal{U}, \mathcal{M})\}.\end{aligned} \tag{8}$$

Once the approximated posterior distribution of $\mathcal{M}$ from (7) has been obtained, MAP estimates of the hyperparameters are determined by maximizing the respective factors,

$$\hat{\mathcal{M}} = \underset{\mathcal{M}}{\operatorname{argmax}}\ q(\boldsymbol{\mu}, \mathbf{B})\, q(\mathbf{W}). \tag{9}$$

[1]Training samples can be globally centered prior to PLDA estimation, allowing for the assumption $E\{\boldsymbol{\mu}\} = \mathbf{0}$.

The steps outlined in (8)-(9) are discussed in the next sections. Each factor is derived via VB, which involves substitution of (1)-(5) into (8), and manipulating the resulting expressions into proper distributions. For these derivations, the conditional moments of the latent variables and hyperparameters are required. Let the $\langle\cdot\rangle$ operator denote the expectation of a given random variable, conditioned on all other variables, e.g $\langle\mathbf{y}_m\rangle$=$E_{\mathcal{Y}\mid\mathcal{X},\mathcal{M},\mathcal{Z}}\{\mathbf{y}_m\}$. Additionally, the following statistics, accumulated across the set $\mathcal{X}$, are defined to simplify notation,

$$\begin{aligned}N_m &= \sum_{n=1}^{N} \langle z_{n,m}\rangle,\ \ 1 \le m \le M \\ \mathbf{r}_{x,m} &= \sum_{n=1}^{N} \langle z_{n,m}\rangle \mathbf{x}_n,\ \ 1 \le m \le M \\ \mathbf{r}_y &= \sum_{m=1}^{M} \langle \mathbf{y}_m \rangle \\ \mathbf{R}_x &= \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^T \\ \mathbf{R}_y &= \sum_{m=1}^{M} N_m \langle \mathbf{y}_m \mathbf{y}_m^T \rangle \\ \mathbf{R}_{yy} &= \sum_{m=1}^{M} \langle \mathbf{y}_m \mathbf{y}_m^T \rangle \\ \mathbf{R}_{xy} &= \sum_{m=1}^{M} \mathbf{r}_{x,m} \langle \mathbf{y}_m \rangle^T.\end{aligned} \tag{10}$$

3.1. The factor $q(\mathcal{Y})$

The factor $q(\mathcal{Y})$ is derived by substituting (1)-(5) and (10) into (8) and collecting terms involving $\mathcal{Y}$, leading to

$$\begin{aligned}\log q(\mathcal{Y}) = -\frac{1}{2} \sum_{m=1}^{M} \Big(&\mathbf{y}_m^T \left(\langle\mathbf{B}\rangle + N_m \langle\mathbf{W}\rangle\right) \mathbf{y}_m \\ &-2\mathbf{y}_m^T \left(\langle\mathbf{B}\boldsymbol{\mu}\rangle + \langle\mathbf{W}\rangle \mathbf{r}_{x,m}\right) \Big) + \text{const.}\end{aligned} \tag{11}$$

By completing the square and using the identity $\langle\mathbf{B}\boldsymbol{\mu}\rangle$=$\langle\mathbf{B}\rangle\langle\boldsymbol{\mu}\rangle$ [20], some mathematical manipulation leads to the product of Normal distributions

$$q(\mathcal{Y}) = \prod_{m=1}^{M} \mathcal{N}\left(\mathbf{y}_m; \boldsymbol{\theta}_m, \boldsymbol{\Phi}_m^{-1}\right) \tag{12}$$

where

$$\begin{aligned}\boldsymbol{\Phi}_m &= \langle\mathbf{B}\rangle + N_m \langle\mathbf{W}\rangle \\ \boldsymbol{\theta}_m &= \boldsymbol{\Phi}_m^{-1} \left(\langle\mathbf{B}\rangle\langle\boldsymbol{\mu}\rangle + \langle\mathbf{W}\rangle \mathbf{r}_{x,m}\right)\end{aligned} \tag{13}$$

so that

$$\begin{aligned}\langle\mathbf{y}_m\rangle &= \boldsymbol{\theta}_m, \\ \langle\mathbf{y}_m \mathbf{y}_m^T\rangle &= \boldsymbol{\Phi}_m^{-1} + \langle\mathbf{y}_m\rangle\langle\mathbf{y}_m\rangle^T.\end{aligned} \tag{14}$$

3.2. The factor $q(\mathcal{Z})$

Following similar steps for the factor $q(\mathcal{Z})$ leads to

$$\log q(\mathcal{Z}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{M} z_{n,m}\Big(\mathrm{Tr}\Big\{\langle\mathbf{W}\rangle\Big(\mathbf{x}_n\mathbf{x}_n^T - \mathbf{x}_n\langle\mathbf{y}_m\rangle^T - \langle\mathbf{y}_m\rangle\mathbf{x}_n^T + \langle\mathbf{y}_m\mathbf{y}_m^T\rangle\Big)\Big\}\Big) + \text{const.} \tag{15}$$

By substituting (14) into (15), $q(\mathcal{Z})$ is shown to be a product of Categorical distributions

$$q(\mathcal{Z}) = \prod_{n=1}^{N}\prod_{m=1}^{M}\left(\frac{i_{n,m}}{\sum_j i_{n,j}}\right)^{z_{n,m}}, \tag{16}$$

where

$$i_{n,m} = \mathcal{N}\left(\mathbf{x}_n; \langle\mathbf{y}_m\rangle, \langle\mathbf{W}\rangle^{-1}\right) \times \exp\left(-\frac{1}{2}\mathrm{Tr}\left\{\langle\mathbf{W}\rangle\mathbf{\Phi}_m^{-1}\right\}\right) \tag{17}$$

so that

$$\langle z_{n,m}\rangle = \frac{i_{n,m}}{\sum_{j=1}^{M} i_{n,j}}. \tag{18}$$

Note that if $\mathrm{Tr}\left\{\langle\mathbf{W}\rangle\mathbf{\Phi}_m^{-1}\right\}$=0, then (17) and (18) become an application of Bayes' rule to cluster $\mathbf{x}_n$ around speakers in $\mathcal{Y}$, but in general, the last term in (17) accounts for the uncertainty in the distribution of $\mathbf{y}_m$ captured by the precision matrix $\mathbf{\Phi}_m$.

3.3. The factor $q(\mathbf{W})$

The factor $q(\mathbf{W})$ is determined by again substituting (1)-(5) and (10) into (8), leading to

$$\log\ q(\mathbf{W}) = \frac{1}{2}(N+\omega-D-1)\log|\mathbf{W}| - \frac{1}{2}\mathrm{Tr}\left\{\mathbf{W}\left(\mathbf{R}_x - \mathbf{R}_{xy} - \mathbf{R}_{xy}^T + \mathbf{R}_y + \omega\mathbf{W}_o^{-1}\right)\right\} + \text{const.}, \tag{19}$$

which can recognized as the Wishart distribution

$$q(\mathbf{W}) = \mathcal{W}\left(\mathbf{W}; \left(\mathbf{R}_x - \mathbf{R}_{xy} - \mathbf{R}_{xy}^T + \mathbf{R}_y + \omega\mathbf{W}_o^{-1}\right)^{-1}, \omega+N\right), \tag{20}$$

so that

$$\langle\mathbf{W}\rangle^{-1} = (\omega+N)^{-1}\left(\mathbf{R}_x - \mathbf{R}_{xy} - \mathbf{R}_{xy}^T + \mathbf{R}_y\right) + \omega(\omega+N)^{-1}\mathbf{W}_o^{-1}. \tag{21}$$

Maximizing $q(\mathbf{W})$, and assuming $N \gg D$, yields $\hat{\mathbf{W}} \approx \langle\mathbf{W}\rangle$. Note that the estimated covariance matrix $\hat{\mathbf{W}}^{-1}$ is expressed as a linear interpolation of the covariance matrices of the prior distribution and the observed data, where $N(\omega+N)^{-1} \in [0,1]$ is the interpolation weight. The hyperparameter ω can be tuned to control the emphasis of the prior during adaptation.

3.4. The factor $q(\boldsymbol{\mu}, \mathbf{B})$

The factor $q(\boldsymbol{\mu}, \mathbf{B})$ is approximated in a similar manner, leading to

$$\begin{aligned}\log q(\boldsymbol{\mu},\mathbf{B}) &= \frac{1}{2}(M+\beta-D)\log|\mathbf{B}| \\ &\quad - \frac{1}{2}\mathrm{Tr}\left\{\mathbf{B}\left(\mathbf{R}_{yy} - \mathbf{r}_y\boldsymbol{\mu}^T - \boldsymbol{\mu}\mathbf{r}_y^T + (M+\beta)\boldsymbol{\mu}\boldsymbol{\mu}^T + \beta\mathbf{B}_o^{-1}\right)\right\} + \text{const.} \\ &= \frac{1}{2}(M+\beta-D)\log|\mathbf{B}| \\ &\quad - \frac{M+\beta}{2}\left(\boldsymbol{\mu} - (M+\beta)^{-1}\mathbf{r}_y\right)^T \times \mathbf{B}\left(\boldsymbol{\mu} - (M+\beta)^{-1}\mathbf{r}_y\right) \\ &\quad - \frac{1}{2}\mathrm{Tr}\left\{\mathbf{B}\left(\mathbf{R}_{yy} - (M+\beta)^{-1}\mathbf{r}_y\mathbf{r}_y^T + \beta\mathbf{B}_o^{-1}\right)\right\} \\ &\quad + \text{const.},\end{aligned} \tag{22}$$

which can be recognized as the Normal-Wishart distribution

$$q(\boldsymbol{\mu},\mathbf{B}) = \mathcal{N}\left(\boldsymbol{\mu}; (\beta+M)^{-1}\mathbf{r}_y, ((\beta+M)\mathbf{B})^{-1}\right) \times \mathcal{W}\left(\mathbf{B}; \left(\mathbf{R}_{yy} - (\beta+M)^{-1}\mathbf{r}_y\mathbf{r}_y^T + \beta\mathbf{B}_o^{-1}\right)^{-1}, \beta+M\right) \tag{23}$$

so that

$$\begin{aligned}\langle\boldsymbol{\mu}\rangle &= (\beta+M)^{-1}\mathbf{r}_y \\ \langle\mathbf{B}\rangle^{-1} &= (\beta+M)^{-1}\mathbf{R}_{yy} + \beta(\beta+M)^{-1}\mathbf{B}_o^{-1} - \langle\boldsymbol{\mu}\rangle\langle\boldsymbol{\mu}\rangle^T.\end{aligned} \tag{24}$$

The MAP estimates of $\{\boldsymbol{\mu}, \mathbf{B}\}$ can be determined by maximizing $q(\boldsymbol{\mu}, \mathbf{B})$, which for $M \gg D$ leads to $\hat{\boldsymbol{\mu}}$=$\langle\boldsymbol{\mu}\rangle$ and $\hat{\mathbf{B}} \approx \langle\mathbf{B}\rangle$. Note that $\hat{\mathbf{B}}^{-1}$ is expressed as a linear interpolation with weight $M(\beta+M)^{-1}$, so that β can be tuned to control the effect of the prior.

Table 1: *Speaker Verification Results for Unsupervised PLDA Domain Adaptation*

Adaptation Method	SRE16 Cantonese EER (%)	SRE16 Cantonese mindcf	SRE16 Tagalog EER (%)	SRE16 Tagalog mindcf	SRE18 CMN2 EER (%)	SRE18 CMN2 mindcf
None	4.55	0.253	16.48	0.745	7.81	0.373
Whiten	4.27	0.234	11.58	0.584	6.68	0.312
Kaldi [12]	3.25	**0.183**	10.22	0.495	6.03	0.279
CORAL [13]	3.82	0.212	10.80	0.527	6.25	0.302
Bayesian [16]	3.82	0.213	10.13	0.492	6.34	0.294
VB-MAP	**3.18**	0.185	**9.96**	**0.490**	**5.31**	**0.259**

Table 2: *Speaker Verification Results for a Fully Unsupervised PLDA Model*

Training Set	M	SRE18 CMN2 EER (%)	SRE18 CMN2 mindcf
Supervised	-	7.81	0.373
Unsupervised	3k	8.66	0.410
	5k	8.01	0.386
	10k	8.39	0.384
	15k	8.93	0.417

3.5. PLDA Adaptation

VB-MAP estimation can be used for unsupervised domain adaptation of PLDA. In many situations, a rich set of labeled out-of-domain data may exist. Additionally, an in-domain data set may be available, but without speaker labels. In this case, the scale parameters $\mathbf{B}_o$ and $\mathbf{W}_o$ can be chosen as the precision matrices estimated from the out-of-domain set. The VB-MAP estimates derived in Section 3.3 and 3.4 then provide the adapted PLDA model, where the shape parameters β and ω can be tuned to control the reliance on the prior estimates. As an iterative solution, the steps in (8) require the model $\mathcal{M}$ and latent variables $\mathcal{Z}$ to be initialized. During experimentation, $\mathcal{M}$ was initialized as $\mathcal{M}_o$, and for each n, $\langle z_{n,m} \rangle$ was randomly drawn from a M-category Dirichlet distribution.

4. Experimental Results

This section presents experimental results for domain adaptation in speaker verification. The baseline verification system used 256 dimensional x-vectors extracted from a ResNet-34 model similar to [21]. The model was trained on data from the 2004-2012 NIST SREs [22–25], along with the Switchboard [26–31] and VoxCeleb2 [32] corpora, and used data augmentation with the MUSAN set [33]. LDA dimension reduction to 150 was first performed, followed by global centering, whitening and length normalization [3]. Verification scores were generated with PLDA, which was trained with the conventional EM algorithm [1] using 90k telephony speech samples from the 2004-2012 NIST SREs, referred to as the *sre-tel* data set.

4.1. Unsupervised PLDA Adaptation

In the domain adaptation experiments, VB-MAP estimation was applied to the 2016 NIST SRE (SRE16) [34] and 2018 NIST SRE (SRE18) CMN2 [35] Tasks. The SRE16 Task contained speech in two languages, Cantonese and Tagalog, and the SRE18 CMN2 Task contained speech in Tunisian Arabic. Additionally, both tasks included data captured on foreign communication networks. In this way, the tasks introduced domain mismatch relative to the *sre-tel* set. To allow for domain adaptation, each of the evaluations provided small, unlabeled in-domain data sets with roughly 2k samples each.

In the experiments, the set $\mathcal{M}_o$ was trained on the *sre-tel* data, and VB-MAP estimation was used to adapt the PLDA model to the respective unlabeled in-domain data sets. Prior distribution scale parameters were set to ω=$2N$ and β=$2M$, resulting in interpolation weights of 0.67. Table 1 provides speaker verification results for the SRE16 and SRE18 CMN2 Tasks. Along with the proposed adaptation method, the table includes results for no domain adaptation, denoted by *None*, for adaptation of the whitening matrix and centering mean only, denoted by *Whiten*, and for the following state-of-the-art methods:

1. Unsupervised adaptation in Kaldi [12] distributes excess variance from the in-domain total covariance matrix to the PLDA model covariance matrices.
2. Correlation Alignment (CORAL) [13] designs a linear feature transform to align the second-order statistics of the out-of-domain data to the in-domain data.
3. [16] applies Bayesian adaptation, with the assumption that in-domain samples are speaker-independent.

For VB-MAP, M=800 speakers were used, but the approach was relatively insensitive to the value of M.

In Table 1, results are provided in terms of equal error rate (EER) and the minimum decision cost (mindcf) with target prior 0.05, and bold entries denote the best result for each task. It can be observed that Kaldi and VB-MAP significantly outperform the remaining baseline methods on the SRE16 Tasks. On the SRE18 CMN2 Task, VB-MAP provides 10%-16% and 7%-12% relative improvement in EER and mindcf, respectively, compared to each of the state-of-the-art baseline systems.

4.2. Fully Unsupervised PLDA Estimation

The VB-MAP solution was also used to estimate a fully unsupervised PLDA model. If non-informative priors are utilized for the PLDA hyperparameters, i.e. ω=0 and β=0, VB-MAP reduces to unsupervised estimation of $\mathcal{M}$. Table 2 provides speaker verification results for unsupervised estimation. The first row includes results for the baseline PLDA model trained with the labeled *sre-tel* corpus. The following rows provide results for training $\mathcal{M}$ with the *sre-tel* set without labels, thereby representing a fully unsupervised model, for a variety of speaker numbers M, ranging from an average of roughly 5 to 30 samples per speaker. Note that both the LDA dimension reduction and PLDA model were trained without labels. It can be observed in the table that using the VB-MAP method without speaker labels can lead to less than 5% performance degradation relative to the supervised system, and is relatively insensitive to the parameter M. Although speaker labels were required for training the embedding extractor, the results show the proposed VB-MAP method to be promising for training a PLDA model in a domain for which no relevant labeled data exists.

5. Conclusions

This paper proposed a Bayesian framework for unsupervised PLDA domain adaptation. Speaker labels are interpreted as latent random variables, and a MAP solution of the PLDA model is derived using Variational Bayes. VB-MAP estimation iteratively infers class labels and updates PLDA hyperparameters. The proposed method is general but is discussed in the context of speaker verification. Experimental results show the benefit of the proposed method on a variety of speaker verification tasks. It is used for domain adaptation in the NIST SRE16 and SRE18 CMN2 Tasks, outperforming baseline techniques. The proposed method is also used to train a fully unsupervised PLDA model, leading to only a small performance degradation relative to conventional supervised training.

6. Acknowledgements

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported by the Department of Defense under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Department of Defense.

Delivered to the U.S. Government with Unlimited Rights, as defined in DFARS Part 252.227-7013 or 7014 (Feb 2014). Notwithstanding any copyright notice, U.S. Government rights in this work are defined by DFARS 252.227-7013 or DFARS 252.227-7014 as detailed above. Use of this work other than as specifically authorized by the U.S. Government may violate any copyrights that exist in this work.

7. References

[1] S. Ioffe, "Probabilistic linear discriminant analysis," in *European Conference on Computer Vision*, 2006, pp. 531–542.

[2] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *ICCV*. IEEE, 2007, pp. 1–8.

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

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

[5] B. J. Borgström and A. McCree, "Discriminatively trained bayesian speaker comparison of I-vectors," in *ICASSP*, 2013, pp. 7659–7662.

[6] C. Vaquero, "Dataset shift in PLDA based speaker verification," in *Odyssey*, 2012.

[7] J. Villalba and E. Lleida, "Bayesian adaptation of PLDA based speaker recognition to domains with scarce development data," in *Odyssey*, 2012.

[8] D. Garcia-Romero and A. McCree, "Supervised domain adaptation for I-vector based speaker recognition," in *ICASSP*, 2014, pp. 4047–4051.

[9] D. Garcia-Romero, A. McCree, S. Shum, N. Brummer, and C. Vaquero, "Unsupervised domain adaptation for I-vector speaker recognition," in *Odyssey*, 2014.

[10] J. Villalba and E. Lleida, "Unsupervised adaptation of PLDA by using Variational Bayes methods," in *ICASSP*, 2014, pp. 744–748.

[11] S. H. Shum, D. A. Reynolds, D. Garcia-Romero, and A. McCree, "Unsupervised clustering approaches for domain adaptation in speaker recognition systems," in *Odyssey*, 2014, pp. 265–272.

[12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *ASRU*, 2011.

[13] M. J. Alam, G. Bhattacharya, and P. Kenny, "Speaker verification in mismatched conditions with frustratingly easy domain adaptation." in *Odyssey*, 2018, pp. 176–180.

[14] K. A. Lee, Q. Wang, and T. Koshinaka, "The CORAL+ algorithm for unsupervised domain adaptation of plda," in *ICASSP*, 2019, pp. 5821–5825.

[15] Q. Wang, K. Okabe, K. A. Lee, and T. Koshinaka, "A generalized framework for domain adaptation of plda in speaker recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6619–6623.

[16] B. J. Borgstrom, D. A. Reynolds, E. Singer, and O. Sadjadi, "Improving the effectiveness of speaker verification domain adaptation with inadequate in-domain data," in *Interspeech*, 2017, pp. 1557–1561.

[17] P.-M. Bousquet and M. Rouvier, "On robustness of unsupervised domain adaptation for speaker recognition." in *Interspeech*, 2019, pp. 2958–2962.

[18] N. Brümmer and E. de Villiers, "The speaker partitioning problem," in *Odyssey*, 2010, pp. 194–201.

[19] B. J. Borgström and P. Torres-Carrasquillo, "Bayesian estimation of PLDA with noisy training labels, with applications to speaker verification," in *ICASSP*, 2020, pp. 7594–7598.

[20] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2013.

[21] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, L. P. García-Perera, F. Richardson, R. Dehak *et al.*, "State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations," *Computer Speech & Language*, vol. 60, p. 101026, 2020.

[22] NIST, "NIST 2004 speaker recognition evaluation plan," https://www.nist.gov/sites/default/files/documents/2017/09/26/sre-04evalplan-v1a.pdf, 2004.

[23] ——, "NIST 2006 speaker recognition evaluation plan," https://www.nist.gov/sites/default/files/documents/2017/09/26/sre-06evalplan-v9.pdf, 2006.

[24] ——, "NIST 2008 speaker recognition evaluation plan," https://www.nist.gov/sites/default/files/documents/2017/09/26/sre-08evalplanrelease4.pdf, 2008.

[25] ——, "NIST 2010 speaker recognition evaluation plan," https://www.nist.gov/system/files/documents/itl/iad/mig/NIST-SRE10evalplan-r6.pdf, 2010.

[26] J. Godfrey and E. Holliman, "Switchboard-1 Release 2," https://catalog.ldc.upenn.edu/LDC97S62, 1993.

[27] D. Graff, A. Canavan, and G. Zipperlen, "Switchboard-2 Phase I," https://catalog.ldc.upenn.edu/LDC98S75, 1998.

[28] D. Graff, K. Walker, and A. Canavan, "Switchboard-2 Phase II," https://catalog.ldc.upenn.edu/LDC99S79, 1999.

[29] D. Graff, D. Miller, and K. Walker, "Switchboard-2 Phase III," https://catalog.ldc.upenn.edu/LDC2002S06, 2002.

[30] D. Graff, K. Walker, and D. Miller, "Switchboard Cellular Part 1 Audio," https://catalog.ldc.upenn.edu/LDC2001S13, 2001.

[31] ——, "Switchboard Cellular Part 2 Audio," https://catalog.ldc.upenn.edu/LDC2004S07, 2004.

[32] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *Interspeech*, pp. 1086–1090, 2018.

[33] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*, 2018, pp. 5329–5333.

[34] NIST, "NIST 2016 speaker recognition evaluation plan," https://www.nist.gov/file/325336, 2016.

[35] ——, "NIST 2018 speaker recognition evaluation plan," https://www.nist.gov/document/sre18evalplan2018-05-31v6pdf, 2018.