Discriminative Training of PLDA for Speaker Verification with X-vectors

Bengt J. Borgström

***Abstract*—This paper proposes a novel approach to discriminative training of probabilistic linear discriminant analysis (PLDA) for speaker verification with x-vectors. Model over-fitting is a well-known issue with discriminative PLDA (D-PLDA) for speaker verification. As opposed to prior approaches which address this by limiting the number of trainable parameters, the proposed method parameterizes the discriminative PLDA (D-PLDA) model in a manner which allows for intuitive regularization, permitting the entire model to be optimized. Specifically, the within-class and across-class covariance matrices which comprise the PLDA model are expressed as products of orthonormal and diagonal matrices, and the structure of these matrices is enforced during model training. The proposed approach provides consistent performance improvements relative to previous D-PLDA methods when applied to a variety of speaker recognition evaluations, including the Speakers in the Wild Core-Core, SRE16, SRE18 CMN2, SRE19 CMN2, and VoxCeleb1 Tasks. Additionally, when implemented in Tensorflow using a modern GPU, D-PLDA optimization is highly efficient, requiring less than 20 minutes.**

***Index Terms*—Speaker Verification, Probabilistic Linear Discriminant Analysis, Discriminative Training, X-vectors**

I. Introduction

Probabilistic linear discriminant analysis (PLDA) is a likelihood ratio test between same-class and different-class hypotheses in a verification task, and has become the standard practice for state-of-the-art speaker verification [1], [2]. By separately modeling across-class and within-class variability, PLDA emphasizes important speaker-specific information while de-emphasizing confusable information such as the acoustic channel. For many years, PLDA scoring was successfully used in combination with i-vectors [3]. Recently, however, x-vectors have been proposed as an alternative form of speaker embedding, and have shown impressive performance particularly in difficult acoustic channels [4].

Typically PLDA is trained as a generative model, using e.g. the expectation-maximization (EM) algorithm [5]. However, PLDA can alternatively be trained to directly optimize a cost function which is more closely related to verification performance. Several studies have explored discriminative training of PLDA (D-PLDA) for speaker verification [2], [6]–[10]. Some of these approaches reformulate PLDA scoring as logistic regression with a non-linear basis function whose form is derived from the PLDA log-likelihood ratio (LLR). While showing promise [2], [7], [8], such techniques are prone to overfitting. To address this issue, other approaches to D-PLDA have reduced the number of trainable parameters [6], [9], [10]. Such techniques, however, may limit the potential effectiveness of discriminative training.

In this paper, we propose a novel approach to discriminative training of PLDA. The PLDA model is parameterized so that the within-class and across-class covariance matrices are expressed as products solely of orthonormal and diagonal matrices. The structure of these matrices can be enforced during D-PLDA optimization, serving to naturally regularize the model. In this way, the entire D-PLDA model can be updated during training, as opposed to limiting the number of trainable parameters as in previous approaches. Additionally, important properties of the underlying PLDA covariance matrices, such as symmetry and positive definiteness, are easily guaranteed by applying parameter constraints during model training. The proposed method achieves consistent performance improvements relative to baseline D-PLDA approaches when used in combination with recently proposed x-vectors [4] and applied to the Speakers in the Wild (SITW) Core-Core [11], SRE16 [12], SRE18 CMN2 [13], SRE19 CMN2 [14], and VoxcCeleb1 [15] Tasks.

This paper is organized as follows: the statistical framework of PLDA is discussed in Sec. II. Sec. III presents the proposed D-PLDA model, and the associated discriminative training method. Experimental results are presented in Sec. IV, and Sec. V provides conclusions.

II. Probabilistic Linear Discriminant Analysis

A. The General Solution

In this section, the statistical framework for PLDA is reviewed in the context of speaker verification with x-vectors. The additive model is assumed, $\mathbf{x}$=$\mathbf{s}$+$\mathbf{c}$, where $\mathbf{x} \in \mathbb{R}^D$ is the observed x-vector, and $\mathbf{s}$ and $\mathbf{c}$ are the underlying speaker

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.

```
Algorithm 1: Diagonalizing $\mathbf{\Sigma}_w$ and $\mathbf{\Sigma}_a$
  Input: Covariance matrices, $\mathbf{\Sigma}_w$ and $\mathbf{\Sigma}_a$
  Output: The diagonalizing matrix $\mathbf{U}$
  Perform the eigendecomposition $\mathbf{\Sigma}_w\mathbf{H}$=$\mathbf{HS}$
  Define $\mathbf{M}$=$\mathbf{HS}^{-\frac{1}{2}}$, so $\mathbf{M}^T\mathbf{\Sigma}_w\mathbf{M}$=$\mathbf{I}$
  Perform the eigendecomposition $\left(\mathbf{M}^T\mathbf{\Sigma}_a\mathbf{M}\right)\mathbf{V}$=$\mathbf{VA}$
  Define $\mathbf{U}$=$\mathbf{HS}^{-\frac{1}{2}}\mathbf{V}$, so $\mathbf{U}^T\mathbf{\Sigma}_w\mathbf{U}$=$\mathbf{I}$ and $\mathbf{U}^T\mathbf{\Sigma}_a\mathbf{U}$=$\mathbf{A}$
```

and channel components. Speaker and channel components are drawn from Gaussian distributions

$$p\left(\mathbf{s}\right) = \mathcal{N}\left(\mathbf{s}; \boldsymbol{\mu}, \mathbf{\Sigma}_a\right), \tag{1}$$

$$p\left(\mathbf{x} \mid \mathbf{s}\right) = \mathcal{N}\left(\mathbf{x}; \mathbf{s}, \mathbf{\Sigma}_w\right). \tag{2}$$

Given two x-vectors, $\mathbf{x}_i$ and $\mathbf{x}_j$, PLDA provides the log-likelihood ratio (LLR) between the same-speaker and different-speaker hypotheses, $\mathcal{H}_1$ and $\mathcal{H}_{-1}$. The PLDA LLR is given by

$$\begin{aligned}\mathcal{L}\left(\mathbf{x}_i, \mathbf{x}_j\right) &= \log \frac{p\left(\mathbf{x}_i, \mathbf{x}_j \mid \mathcal{H}_1\right)}{p\left(\mathbf{x}_i, \mathbf{x}_j \mid \mathcal{H}_{-1}\right)} \\ &= \log \frac{\int \mathcal{N}\left(\mathbf{x}_i; \mathbf{s}, \mathbf{\Sigma}_w\right) \mathcal{N}\left(\mathbf{x}_j; \mathbf{s}, \mathbf{\Sigma}_w\right) \mathcal{N}\left(\mathbf{s}; \boldsymbol{\mu}, \mathbf{\Sigma}_a\right) d\mathbf{s}}{\mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}, \mathbf{\Sigma}_w + \mathbf{\Sigma}_a\right) \mathcal{N}\left(\mathbf{x}_j; \boldsymbol{\mu}, \mathbf{\Sigma}_w + \mathbf{\Sigma}_a\right)}.\end{aligned} \tag{3}$$

Given the LLR, the posterior probability of the same-speaker hypothesis is expressed as

$$P\left(\mathcal{H}_1 \mid \mathbf{x}_i, \mathbf{x}_j\right) = \sigma\left(\mathcal{L}\left(\mathbf{x}_i, \mathbf{x}_j\right)\right), \tag{4}$$

where σ is the logistic function, $\sigma\left(t\right) = \left(1 + \exp\left(-t\right)\right)^{-1}$. It was shown in [1] that the solution from (3) can be expressed equivalently as

$$\begin{aligned}\mathcal{L}\left(\mathbf{x}_i, \mathbf{x}_j\right) &= -\frac{1}{2}\log f + \frac{1}{2}\left(\mathbf{x}_i - \boldsymbol{\mu}\right)^T \mathbf{Q}\left(\mathbf{x}_i - \boldsymbol{\mu}\right) \\ &+ \frac{1}{2}\left(\mathbf{x}_j - \boldsymbol{\mu}\right)^T \mathbf{Q}\left(\mathbf{x}_j - \boldsymbol{\mu}\right) + \left(\mathbf{x}_i - \boldsymbol{\mu}\right)^T \mathbf{P}\left(\mathbf{x}_j - \boldsymbol{\mu}\right),\end{aligned} \tag{5}$$

where

$$\begin{aligned}f &= \frac{\left|\mathbf{\Sigma}_t - \mathbf{\Sigma}_a\right|\left|\mathbf{\Sigma}_t + \mathbf{\Sigma}_a\right|}{\left|\mathbf{\Sigma}_t\right|^2}, \\ \mathbf{Q} &= \mathbf{\Sigma}_t^{-1} - \left(\mathbf{\Sigma}_t - \mathbf{\Sigma}_a\mathbf{\Sigma}_t^{-1}\mathbf{\Sigma}_a\right)^{-1}, \\ \mathbf{P} &= \mathbf{\Sigma}_t^{-1}\mathbf{\Sigma}_a\left(\mathbf{\Sigma}_t - \mathbf{\Sigma}_a\mathbf{\Sigma}_t^{-1}\mathbf{\Sigma}_a\right)^{-1}.\end{aligned} \tag{6}$$

and where $\mathbf{\Sigma}_t$=$\mathbf{\Sigma}_w$+$\mathbf{\Sigma}_a$. The set $\{\boldsymbol{\mu}, \mathbf{\Sigma}_a, \mathbf{\Sigma}_w\}$ parameterizes PLDA, and can be trained as a generative model using the expectation-maximization (EM) algorithm [5].

B. Diagonalizing the Covariance Matrices

The expression in (5) can be simplified if the x-vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ are first transformed so that their within-class and across-class covariances are jointly diagonalized, as in the Kaldi Toolkit [16]. Algorithm 1 derives the matrix $\mathbf{U}$ which diagonalizes both $\mathbf{\Sigma}_w$ and $\mathbf{\Sigma}_a$, so that $\mathbf{U}^T\mathbf{\Sigma}_w\mathbf{U}$=$\mathbf{I}$ and $\mathbf{U}^T\mathbf{\Sigma}_a\mathbf{U}$=$\mathbf{A}$, where $\mathbf{A}$=diag$\left\{\mathbf{a}\right\}$. By pre-processing x-vectors according to

$$\mathbf{y}_i = \mathbf{U}^T\left(\mathbf{x}_i - \boldsymbol{\mu}\right), \tag{7}$$

the matrices $\mathbf{Q}$ and $\mathbf{P}$ are diagonalized with diagonal vectors $\mathbf{q}$ and $\mathbf{p}$, respectively. The LLR from (5) then reduces to

$$\begin{aligned}&\mathcal{L}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \\ &-\frac{1}{2}\log f + \frac{1}{2}\sum_{d=1}^{D}\left(q_d\left(\mathbf{y}_i^2\left(d\right) + \mathbf{y}_j^2\left(d\right)\right) + 2p_d\mathbf{y}_i\left(d\right)\mathbf{y}_j\left(d\right)\right),\end{aligned} \tag{8}$$

where

$$\begin{aligned}f &= \prod_{d=1}^{D}\frac{1 + 2a_d}{\left(1 + a_d\right)^2}, \\ q_d &= \frac{-a_d^2}{\left(1 + a_d\right)\left(1 + 2a_d\right)}, \\ p_d &= \frac{a_d}{1 + 2a_d},\end{aligned} \tag{9}$$

and where subscripts are used to index elements within vectors. In this way, the LLR is expressed solely in terms of scalar operations.

III. D-PLDA Optimization

The generative PLDA model discussed in Sec. II has become a standard method for scoring speaker embeddings in state-of-the-art speaker verification systems. However, improved results can be expected if the PLDA model is trained to directly optimize a cost that is more closely tied to verification performance. In this section, the proposed approach to discriminative PLDA is presented.

A. Parameterization

D-PLDA has previously been explored as a way to improve speaker verification performance. However, discriminative training of the full PLDA parameter set, $\{\boldsymbol{\mu}, \mathbf{\Sigma}_a, \mathbf{\Sigma}_w\}$, has been observed to over-fit the model to training data, resulting in degraded verification performance [6], [10]. Previous approaches to D-PLDA have therefore proposed various ways to reparameterize the PLDA model in order to reduce the number of trainable parameters, thereby regularizing the optimization process. For example, in [6], only scaling factors for the within-class and across-class covariance matrices are optimized. That is, the covariance matrices are expressed as

$$\begin{aligned}\mathbf{\Sigma}_w &\Leftarrow \lambda_w\mathbf{\Sigma}_w, \\ \mathbf{\Sigma}_a &\Leftarrow \lambda_a\mathbf{\Sigma}_a,\end{aligned} \tag{10}$$

and the parameter set $\{\lambda_w, \lambda_a\}$ is updated during D-PLDA training. In [9], the parameter set $\{\mu, \lambda_w, \mathbf{a}\}$ is instead updated.[1]

Such approaches to D-PLDA reduce the number of trainable parameters, making optimization less prone to over-fitting. However, such techniques may limit the potential effectiveness of discriminative training. This paper proposes to update the entire PLDA model during discriminative training, but to reparameterize the model so that the structure of the trainable

[1] In [9], the PLDA covariance matrices aren't strictly diagonalized as in Algorithm 1. Instead, $\mathbf{\Sigma}_a$ is diagonalized, but $\mathbf{\Sigma}_w$ is approximated as diagonal based on properties of unit-normalized vectors [17].

parameters can be enforced during optimization. From Algorithm 1, the diagonalizing matrix is defined as $\mathbf{U}$=$\mathbf{H}\mathbf{S}^{-\frac{1}{2}}\mathbf{V}^{T}$. Here $\mathbf{S}$ is a diagonal matrix, i.e. $\mathbf{S}$=diag$\{\mathbf{s}\}$, and $\mathbf{H}$ and $\mathbf{V}$ are orthonormal matrices. The within-class and across-class covariance matrices can then be expressed as products solely of diagonal and orthonormal matrices,

$$\begin{aligned}\mathbf{\Sigma}_w &= \mathbf{H}\mathbf{S}\mathbf{H}^T, \\ \mathbf{\Sigma}_a &= \mathbf{H}\mathbf{S}^{\frac{1}{2}}\mathbf{V}\mathbf{A}\mathbf{V}^T\mathbf{S}^{\frac{1}{2}}\mathbf{H}^T,\end{aligned} \tag{11}$$

where these matrices are tied across $\mathbf{\Sigma}_w$ and $\mathbf{\Sigma}_a$. In this paper, the PLDA model is parameterized as $\{\mathbf{H}, \mathbf{V}, \mathbf{s}, \mu, \mathbf{a}\}$, and the diagonal structures of $\mathbf{S}$ and $\mathbf{A}$ and orthonormal properties of $\mathbf{H}$ and $\mathbf{V}$ are enforced during model optimization, as will be discussed in Sec. III-D.

B. The Generalized Cost Function

To optimize the D-PLDA model, we can minimize some discriminative cost function over a training set of x-vectors. The general form of the cost function is given by

$$C = \sum_{m\in\{-1,1\}} \frac{1}{N_m} \sum_{(i,j)\in\mathcal{H}_m} l_m\left(\mathcal{L}\left(\mathbf{x}_i, \mathbf{x}_j\right)\right), \tag{12}$$

where $l_{-1}\left(\mathcal{L}\left(\mathbf{x}_i, \mathbf{x}_j\right)\right)$ and $l_1\left(\mathcal{L}\left(\mathbf{x}_i, \mathbf{x}_j\right)\right)$ represent the losses associated with $\mathcal{L}\left(\mathbf{x}_i, \mathbf{x}_j\right)$ for labels -1 and 1, respectively. Additionally, the notation (i, j) denotes a verification trial with inputs $\mathbf{x}_i$ and $\mathbf{x}_j$. Finally, N_m denotes the number of training trials for class $\mathcal{H}_m$.

C. The Loss Function

There exists a variety of loss functions that can be used in the generalized cost from (12), and the choice of $l_m\left(\mathcal{L}\left(\mathbf{x}_i, \mathbf{x}_j\right)\right)$ can be made based on the intended application. A commonly used loss function for discriminative PLDA is the Log Loss,

$$l_m\left(\mathcal{L}\left(\mathbf{x}_i, \mathbf{x}_j\right)\right) = -\log\left(\sigma\left(m\mathcal{L}\left(\mathbf{x}_i, \mathbf{x}_j\right)\right)\right), \tag{13}$$

but other approaches have utilized the Brier, Hinge, and 0-1 Losses [2], [6]–[8], [18]. In this paper, the Log and 0-1 Losses are studied. Since the 0-1 Loss is not differentiable, the sigmoid approximation from [19] is instead used

$$l_m\left(\mathcal{L}\left(\mathbf{x}_i, \mathbf{x}_j\right)\right) = \sigma\left(-m\mathcal{L}\left(\mathbf{x}_i, \mathbf{x}_j\right)\right). \tag{14}$$

The expression in (13) or (14) is substituted into (12) during D-PLDA optimization.

As discussed in [20], loss functions can typically be decomposed into two components: discrimination between target and non-target trials, and score calibration. When optimizing the D-PLDA cost, reductions in (12) may be attributed to a combination of changes in these two components. Since calibration can easily be addressed after PLDA scoring via e.g. logistic regression, the main goal of D-PLDA is to improve discrimination between target and non-target trials. However, reductions in (12) during optimization may focus on improving score calibration at the cost of discrimination. We therefore propose to calibrate LLRs prior to D-PLDA training, which using logistic regression leads to

$$\mathcal{L}\left(\mathbf{x}_i, \mathbf{x}_j\right) \Leftarrow \alpha\mathcal{L}\left(\mathbf{x}_i, \mathbf{x}_j\right) + \beta. \tag{15}$$

If this mapping is optimized with respect to the same data set used to train the D-PLDA model, negligible improvements in calibration can be expected during D-PLDA training, and reductions in (12) are more likely attributed to improved discrimination between target and non-target trials.

D. Constraints and Regularization

A benefit of the proposed approach is that the original PLDA statistical framework can be extracted from the D-PLDA hyperparameters using (11), which is not possible in approaches such as [2], [7], [8]. In this way, important properties of the PLDA covariance matrices can be guaranteed throughout optimization by setting parameter constraints. For example, it is clear from (11) that $\mathbf{\Sigma}_w$ and $\mathbf{\Sigma}_a$ are symmetric. Furthermore, $\mathbf{\Sigma}_w$ and $\mathbf{\Sigma}_a$ are guaranteed non-singular if $\mathbf{s}$ and $\mathbf{a}$ are constrained to be positive, respectively. Alternatively, $\mathbf{a}$ can be constrained to be non-negative, which is consistent with Simplified PLDA [21]. In the proposed training, the constraints $\mathbf{s} > 0$ and $\mathbf{a} \geq 0$ are applied.

Previous approaches to D-PLDA have explored various regularization techniques in order to avoid model over-fitting. As mentioned in Sec.III-A, the proposed approach achieves regularization by simply enforcing the diagonal structure of $\mathbf{S}$ and $\mathbf{A}$, and the orthonormal property of $\mathbf{H}$ and $\mathbf{V}$. The diagonal structure of a matrix is easily guaranteed during optimization. On the otherr hand, the orthonormal property of $\mathbf{H}$ and $\mathbf{V}$ can be enforced via the regularization term

$$\gamma\left(\left\|\mathbf{H}\mathbf{H}^T - \mathbf{I}\right\|_F^2 + \left\|\mathbf{V}\mathbf{V}^T - \mathbf{I}\right\|_F^2\right), \tag{16}$$

where γ is a constant.

IV. Results

This section presents experimental speaker verification results. The baseline verification system used x-vectors generated according to [4]. LDA dimension reduction to 250 was first performed, followed by global whitening and length normalization [1]. The baseline system used conventional PLDA for scoring, which was trained using the EM algorithm. Note that the baseline system in this study closely resembles that from [14].

The proposed D-PLDA system was used to model x-vectors after LDA dimension reduction, and was initialized using the EM algorithm. During D-PLDA training, Gradient Descent was performed using the Adam optimization technique [22], with γ=10^4 and θ=0.05, and logistic regression calibration from (15) was included. Mini-batches of $4,096$ trials were randomly selected from all possible combinations of x-vectors, where mini-batches were evenly comprised of target and non-target samples, and a total of 50.0 M trials were used. When run on a NVIDIA K80 GK210 GPU, D-PLDA optimization took less than 20 minutes.

The first set of experiments addresed telephony speech, and the proposed system was applied to the SRE16 Cantonese,

TABLE I
SPEAKER VERIFICATION RESULTS FOR TELEPHONY SPEECH

Model	Parameterization	Loss Function	SRE16 Cantonese		SRE16 Tagalog		SRE18 CMN2		SRE19 CMN2	
			EER (%)	mindcf	EER (%)	mindcf	EER (%)	mindcf	EER (%)	mindcf
PLDA	–	–	8.02	0.577	20.60	0.997	11.61	0.687	11.62	0.672
D-PLDA [6]	$\{\lambda_w, \lambda_a\}$	Log	7.94	0.570	20.43	0.994	11.54	0.682	11.63	0.670
D-PLDA [9]	$\{\lambda_w, \mathbf{a}, \boldsymbol{\mu}\}$	Log	7.56	0.560	20.17	0.994	11.19	0.674	11.19	0.657
D-PLDA	$\{\lambda_w, \mathbf{a}, \boldsymbol{\mu}\}$	0-1	7.26	0.556	19.86	0.994	10.89	0.672	10.51	0.633
D-PLDA	$\{\mathbf{w}, \mathbf{a}, \boldsymbol{\mu}\}$	0-1	7.18	0.555	19.69	0.993	10.70	0.670	10.66	0.647
D-PLDA	$\{\mathbf{H}, \mathbf{V}, \mathbf{s}, \mathbf{a}, \boldsymbol{\mu}\}$	0-1	**6.76**	**0.540**	**19.40**	**0.970**	**10.17**	**0.647**	**10.55**	**0.634**

TABLE II
SPEAKER VERIFICATION RESULTS FOR MICROPHONE SPEECH

Model	Parameterization	Loss Function	SITW Core-Core		VoxCeleb1	
			EER (%)	mindcf	EER (%)	mindcf
PLDA	–	–	5.33	0.545	7.40	0.583
D-PLDA [6]	$\{\lambda_w, \lambda_a\}$	Log	5.33	0.541	7.27	0.590
D-PLDA [9]	$\{\lambda_w, \mathbf{a}, \boldsymbol{\mu}\}$	Log	5.14	0.527	7.16	0.565
D-PLDA	$\{\lambda_w, \mathbf{a}, \boldsymbol{\mu}\}$	0-1	5.06	0.514	**6.93**	**0.555**
D-PLDA	$\{\mathbf{w}, \mathbf{a}, \boldsymbol{\mu}\}$	0-1	4.95	0.520	7.07	0.565
D-PLDA	$\{\mathbf{H}, \mathbf{V}, \mathbf{s}, \mathbf{a}, \boldsymbol{\mu}\}$	0-1	**4.92**	**0.511**	7.18	**0.555**

the SRE16 Tagalog, the SRE18 CMN2, and the the SRE19 CMN2 Tasks. The PLDA and D-PLDA training set included data from the NIST SRE04-SRE10 along with Mixer 6, and was extended using data augmentation as in [4], with noise and reverberation from [23], resulting in 151k cuts. The second set of experiments addressed microphone speech, and the proposed system as applied to the SITW Core-Core and VoxCeleb1 Tasks. Here, the PLDA and D-PLDA training set included data from the NIST SRE04-SRE10 and Mixer 6, along with a subset of the VoxCeleb2 corpus. Again, data augmentation was performed according to [4], resulting in 200k total cuts.

Tables I and II provide speaker verification results for the previously described experiments, in terms of equal error rate (EER) and the minimum decision cost function (mindcf) with $P(\mathcal{H}_1)$=10^{-2}. The tables are presented as ablation studies, summarizing a variety of D-PLDA parameterizations and loss functions, of which specific cases correspond the baseline systems from [6] and [9]. Results in bold indicate the best performance for each verification task.

It can be observed in Tables I and II that the baseline system from [6] provided minimal performance improvements relative to conventional PLDA, which is most likely due to the stringent constraints implied by its PLDA parameterization. The baseline system from [9] resulted in more substantial improvements across the various speaker verification tasks. The use of the approximated 0-1 Loss function provided further performance benefits across almost all tasks, relative to the Log Loss function. Finally, the proposed D-PLDA parameterization, in combination with the approximated 0-1 Loss function, resulted in significant performance benefits across almost all speaker verification tasks. Specifically, D-PLDA provides upto 15% relative improvement in EER and upto 6% relative improvement in mindcf compared to [9], and provides 4%-19% relative improvement in EER and upto 9% relative improvement in mindcf compared to using conventional PLDA. Throughout experimentation, the use of score calibration during D-PLDA training provided significant performance improvements, and omitting this step often yielded performance worse than the baseline PLDA system.

V. CONCLUSIONS

This paper proposed a novel approach to discriminative PLDA for speaker verification with x-vectors. Parameterizing the PLDA model in terms of diagonal and orthonormal matrices allows these properties to be enforced during D-PLDA optimization, and allows for intuitive constraints and regularization to be used. The proposed method provides performance improvements on the SRE16, SRE18, SRE19, SITW, and VoxCeleb1 Tasks, relative to other D-PLDA approaches. Additionally, the proposed technique is efficient, requiring less than 20 minutes to train on a modern GPU. Although introduced in the context of x-vectors from [4], the proposed D-PLDA method can likely be applied to many other types of speaker embeddings.

REFERENCES

[1] Daniel Garcia-Romero and Carol Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Interspeech*, 2011, pp. 249–252.

[2] Lucas Burget, Oldrich Pichot, Sandro Cumani, Ondrej Glembek, Patel Matejka, and Niko Brummer, "Discriminatively trained probabilistic linear discriminat analysis analysis for speaker verification," in *ICASSP*, 2011, pp. 4832–4835.

[3] Najim Dehak, Patrick Kenny, Rema Dehak, Pierre Ouellet, and Pierre Dumouchel, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 788–798, 2011.

[4] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," *ICASSP*, 2018.

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

[6] Bengt J. Borgström and Alan McCree, "Discriminatively trained bayesian speaker comparison of i-vectors," in *ICASSP*, 2013, pp. 7659–7662.

[7] Johan Rohdin, Sangeeta Biswas, and Koichi Shinoda, "Constrained discriminative plda training for speaker verification," in *ICASSP*, 2014, pp. 1670–1674.

[8] Johan Rohdin, Sangeeta Biswas, and Koichi Shinoda, "Discriminative plda training with application-specific loss functions for speaker verification," in *Odyssey, The Speaker and Language Recognition Workshop*, 2014.

[9] Pierre-Michel Bousquet and Jean-Francois Bonastre, "Constrained discriminative speaker verification specific to normalized i-vectors," in *Proceedings of Odyssey*, 2016, pp. 53–59.

[10] Johan Rohdin, Sangeeta Biswas, and Koichi Shinoda, "Robust discriminative training against data insufficiency in plda-based speaker verification," *Computer Speech & Language*, vol. 35, pp. 32–57, 2016.

[11] Mitchell McLaren, Luciana Ferrer, Diego Castan, and Aaron Lawson, "The 2016 speakers in the wild speaker recognition evaluation.," in *INTERSPEECH*, 2016, pp. 823–827.

[12] Seyed Omid Sadjadi, Timothée Kheyrkhah, Audrey Tong, Craig Greenberg, Elliot Singer Reynolds, Lisa Mason, and Jaime Hernandez-Cordero, "The 2016 nist speaker recognition evaluation," 2017.

[13] Seyed Omid Sadjadi, Timothée Kheyrkhah, Audrey Tong, Craig Greenberg, Elliot Singer Reynolds, Lisa Mason, and Jaime Hernandez-Cordero, "The 2018 nist speaker recognition evaluation," 2017.

[14] Seyed Omid Sadjadi, Craig Greenberg, Elliot Singer, Douglas Reynolds, Lisa Mason, and Jaime Hernandez-Cordero, "The 2019 nist speaker recognition evaluation cts challenge," 2017.

[15] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[16] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. 2011, IEEE Signal Processing Society.

[17] Pierre-Michel Bousquet, Anthony Larcher, Driss Matrouf, Jean-François Bonastre, and Oldřich Plchot, "Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis," 2012.

[18] Sandro Cumani, Niko Brümmer, Lukáš Burget, Pietro Laface, Oldřich Plchot, and Vasileios Vasilakakis, "Pairwise discriminative speaker verification in the I-vector space," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1217–1227, 2013.

[19] Tan Nguyen and Scott Sanner, "Algorithms for direct 0–1 loss optimization in binary classification," in *International Conference on Machine Learning*, 2013, pp. 1085–1093.

[20] Niko Brümmer and Johan Du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[21] Daniel Garcia-Romero and Alan McCree, "Supervised domain adaptation for i-vector based speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4047–4051.

[22] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.