Geometric Discriminant Analysis for I-vector Based Speaker Verification

Can Xu, Xianhong Chen, Liang He* and Jia Liu
Department of Electronic Engineering, Tsinghua University, Beijing, China
E-mail: {xucan,chenxianhong,heliang,liuj}@mail.tsinghua.edu.cn

***Abstract*—Many i-vector based speaker verification use linear discriminant analysis (LDA) as a post-processing stage. LDA maximizes the arithmetic mean of the Kullback-Leibler (KL) divergences between different pairs of speakers. However, for speaker verification, speakers with small divergence are easily misjudged. LDA is not optimal because it does not emphasize on enlarging small divergences. In addition, LDA makes an assumption that the i-vectors of different speakers are well modeled by Gaussian distributions with identical class covariance. Actually, the distributions of different speakers can have different covariances. Motivated by these observations, we explore speaker verification with geometric discriminant analysis (GDA), which uses geometric mean instead of arithmetic mean when maximizing the KL divergences. It puts more emphasis on enlarging small divergences. Furthermore, we study the heteroscedastic extension of GDA (HGDA), taking different covariances into consideration. Experiments on i-vector machine learning challenge indicate that, when the number of training speakers becomes smaller, the relative performance improvement of GDA and HGDA compared with LDA becomes larger. GDA and HGDA are better choices especially when training data is limited.**

I. Introduction

Speaker verification aims to verify a person's identity from voice characteristics. I-vector based speaker verification system [1] provides a method to map a variable length utterance into a low dimensional fixed length i-vector in total variability subspace. An i-vector is able to retain the essential information of an utterance, such as speaker, language [2], and age [3]. Thus, it can be viewed as a front-end feature for further back-end classifiers.

The state-of-the-art back-end classifier for i-vector based speaker verification is linear discriminant analysis-probabilistic linear discriminant analysis (LDA-PLDA) [4]. LDA transforms i-vectors to a lower dimensional space to eliminate the undesired information in the total variability subspace. Such a transformation attempts to maximize the between-class scatter and minimize the within-class variation [5]. Theoretically, it is equivalent to maximize the arithmetic mean of the Kullback-Leibler (KL) divergences between different classes' distributions under the assumption that these distributions are Gaussian with different means but an identical covariance [6]. However, LDA faces two problems. 1), For speaker verification, non-target speakers which have small divergences with target speakers are more easily misjudged. Arithmetic mean treats all divergence equally and does not emphasis the pair of speakers with small divergence. LDA may merge two speakers with small divergence to preserve large divergence. 2), the speakers' i-vectors are not necessary Gaussian distributions with an identical covariance.

To solve the first problem, some algorithms have been proposed. One is weighted LDA (WLDA) [7], [8]. When calculating the between-class scatter matrix, it adds a weighting function to speaker pairs. The weighting function is defined to weight the speakers with small distance more heavily. It can increase the influence of speakers with small divergence and reduce the speaker confusion. Another approach is local pairwise LDA (LPLDA) [9]. It builds local confusable data for each class and uses the pairs to compute between-class scatter matrix. A third approach is locality preserving projection (LPP) [10], [11]. LPP only focuses on the local data structure of each i-vector with its K nearest within- and between-speaker i-vectors. It attempts to ensure that nearby within-class i-vectors are kept closer, while nearby between-class i-vectors are mapped farther apart. Such a design can reduce the influence of speakers with large divergence because the i-vector pairs with large distance are ignored. All these three methods emphasis distinguishing the confusable i-vectors from different speakers with small distance. But ignore the confusable i-vectors from the same speaker with large distance. Considering this, distance-dependent metric learning (DDML) [12] is proposed. When the i-vectors have small distance, it focuses more on expanding the negative pairs. When the i-vectors have large distance, it focuses more on shrinking the positive pairs. Nevertheless, all the above methods need to calculate the distance between i-vector pairs and increase the computation complexity.

For the i-vector distribution, in [13], heavy-tailed distribution in place of Gaussian distribution is used to diminish the effect of outlying data. The resulting better performance provides evidence for non-Gaussian behavior of i-vectors. Based on this non-Gaussian behavior, nonparametric discriminant analysis (NDA) [14], [15] calculates both within and between class scatter matrices on a local basis using a nearest neighbor rule. Non-Gaussian behavior make the i-vector modeling more complex. Therefore paper [16] performs a non-linear transformation (length normalization) of the i-vectors and finds that length normalized i-vectors can be approximately modeled as a Gaussian distribution. For short utterance speaker verification, paper [17] propose short utterance variance normalization (SUVN) technique and an utterance variance (SUV) modelling to improve system performance. Even though, the covariances of different speakers are still different.

In this paper, we propose to investigate a new discriminant analysis technique, named as geometric discriminant analysis (GDA), for i-vector based speaker verifiaction. It is based on geometric mean instead of arithmetic mean used in LDA. Geometric mean is introduced in subspace selection in [6]. It amplifies the effects of small divergences and reduces the effects of large divergences. We further extend GDA to heteroscedastic version to take the influence of different speaker covariances into consideration.

The remainder of this paper is organized as follows. Section 2 gives a brief review of LDA. Section 3 presents the motivation of this paper, and proposes GDA and heteroscedastic GDA algorithms. Section 4 describes the experiments and discusses the results. In section 5, a conclusion is made.

II. Linear Discriminant Analysis

LDA aims to find a low-dimensional subspace, in which different classes are well separated. Suppose the sth class contains n_s samples $\{\mathbf{x}_{s,k}\}_{k=1}^{n_s}$ and has an expected mean of $\boldsymbol{\mu}_s$. The between-class scatter matrix and within-class scatter matrix are defined as [5], [18]:

$$\begin{aligned}\mathbf{S}_b &= \frac{1}{n}\sum_{s=1}^{S} n_s(\boldsymbol{\mu}_s-\boldsymbol{\mu})(\boldsymbol{\mu}_s-\boldsymbol{\mu})^{\mathrm{t}},\\ \mathbf{S}_w &= \frac{1}{n}\sum_{s=1}^{S}\sum_{k=1}^{n_s}(\mathbf{x}_{s,k}-\boldsymbol{\mu}_s)(\mathbf{x}_{s,k}-\boldsymbol{\mu}_s)^{\mathrm{t}},\end{aligned} \tag{1}$$

where $n=\sum_{s=1}^{S} n_s$ is the size of the training set, S is the total class number, $\boldsymbol{\mu}=\frac{1}{n}\sum_{s=1}^{S}\sum_{k=1}^{n_s}\mathbf{x}_{s,k}$ is the overall sample mean. LDA can be formulated as an optimization problem to find a subspace that maximizes the ratio of the between-class scattering to the within-class scattering, as

$$\mathbf{W} = \underset{\mathbf{W}}{\operatorname{argmax}}\,\mathrm{tr}\left((\mathbf{W}^{\mathrm{t}}\mathbf{S}_w\mathbf{W})^{-1}\mathbf{W}^{\mathrm{t}}\mathbf{S}_b\mathbf{W}\right) \tag{2}$$

Theoretically, maximizing the ratio of the between-class scattering to the within-class scattering is equal to maximizing the arithmetic mean of the KL divergences between different classes under the assumption that different class data distributions are Gaussian distributions with different means but identical covariance. It is proved in [6] and simply described below.

Proof: Suppose the ith class data follows Gaussian distribution $p_i = \mathcal{N}(\mathbf{x};\mu_i,\Sigma)$ and the data after projection is $\boldsymbol{y} = \mathbf{W}^{\mathrm{t}}\mathbf{x}$. Then the KL divergence between two projected classes can be calculated as

$$D_{\mathbf{W}}(p_i||p_j) = \frac{1}{2}\mathrm{tr}\Big((\mathbf{W}^{\mathrm{t}}\boldsymbol{\Sigma}\mathbf{W})^{-1}(\mathbf{W}^{\mathrm{t}}\boldsymbol{\Sigma}_{ij}\mathbf{W})\Big) + \mathrm{c} \tag{3}$$

where $\boldsymbol{\Sigma}_{ij} = (\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^{\mathrm{t}}$, c represents a constant.

To maximizing the arithmetic mean of the KL divergences between different classes, the projection matrix can be found by solving the object function:

$$\mathbf{W} = \underset{\mathbf{W}}{\operatorname{argmax}} \sum_{1\le i\ne j\le S} \frac{q_i q_j}{\sum\limits_{1\le m\ne n\le S} q_m q_n} D_{\mathbf{W}}(p_i||p_j), \tag{4}$$

where $q_i = n_i/n$. It can be further transformed into:

$$\begin{aligned}\mathbf{W} &= \underset{\mathbf{W}}{\operatorname{argmax}} \sum_{1\le i\ne j\le S} q_i q_j D_{\mathbf{W}}(p_i||p_j)\\ &= \underset{\mathbf{W}}{\operatorname{argmax}} \sum_{1\le i\ne j\le S} q_i q_j \mathrm{tr}\Big((\mathbf{W}^{\mathrm{t}}\boldsymbol{\Sigma}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{t}}\boldsymbol{\Sigma}_{ij}\mathbf{W}\Big)\\ &= \underset{\mathbf{W}}{\operatorname{argmax}}\,\mathrm{tr}\Big((\mathbf{W}^{\mathrm{t}}\boldsymbol{\Sigma}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{t}}(\sum_{i=1}^{S-1}\sum_{j=i+1}^{S} q_i q_j \boldsymbol{\Sigma}_{ij})\mathbf{W}\Big).\end{aligned}$$

Submit $\mathbf{S}_b = \sum_{i=1}^{S-1}\sum_{j=i+1}^{S} q_i q_j \boldsymbol{\Sigma}_{ij}$ [19] and $\mathbf{S}_w = \boldsymbol{\Sigma}$ into the above equation, we can get (2). Thus, it is proved that LDA is to maximizing the arithmetic mean of the KL divergences between different classes under the assumption that different classes have a same covariance.

III. Geometric Discriminant Analysis

A. Geometric Mean

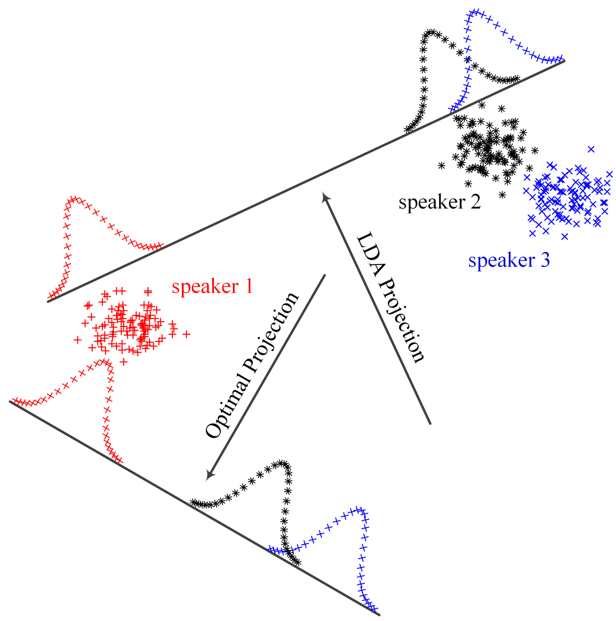

Fig. 1. Compare LDA projection with optimal projection.

1) Motivation: LDA maximizes the arithmetic mean of the KL divergences between different pairs of speakers. As shown in Fig. 1, it may preserve large divergence and neglect small divergence to obtain the arithmetic mean maximization. Then, speaker 2 and speaker 3 merge with each other in the projected subspace and introduce mistake. In fact, when more attention is payed to expand the speakers with small divergence, this mistake can be avoided, such as the optimal projection in Fig. 1. The effect of speakers with small divergence should be emphasized.

2) Realization: To solve the this problem, when maximizing the KL divergences between different speakers, we use geometric mean [6] instead of arithmetic mean.

$$\mathbf{W} = \underset{\mathbf{W}}{\operatorname{argmax}} \prod_{1\le i\ne j\le S} D_{\mathbf{W}}(p_i||p_j)^{\frac{q_i q_j}{\sum_{1\le m\ne n\le S} q_m q_n}} \tag{5}$$

where $D_{\mathbf{W}}(p_i||p_j)$ is the same as (3). To simplify the calculation, we apply a log function, which is a strict monotonic real-valued increasing function, to (5). Then obtain the loss function:

$$\begin{aligned} L(\mathbf{W}) &= -\log \prod_{1\leq i\neq j\leq S} D_W(p_i||p_j)^{\frac{q_i q_j}{\sum_{1\leq m\neq n\leq S} q_m q_n}} \\ &= -\sum_{1\leq i\neq j\leq S} \frac{q_i q_j}{\sum_{1\leq m\neq n\leq S} q_m q_n} \log D_{\mathbf{W}}(p_i||p_j) \end{aligned} \tag{6}$$

Compared (6) with (4), we can find that, in (6), the divergence $D_{\mathbf{W}}(p_i||p_j)$ is handled with an additional log function. As shown in Fig. 2, function $\log(x)$ has the characteristic that the larger x is, the smaller its derivative becomes. When x is large enough, the increase of x will bring little in the increase of $\log(x)$. That means, increasing a small $D_{\mathbf{W}}(p_i||p_j)$ will bring more benefit to minimizing the loss function than increasing a large $D_{\mathbf{W}}(p_i||p_j)$. Therefore, the effect of large divergences is reduced and the effect of small divergences is amplified.

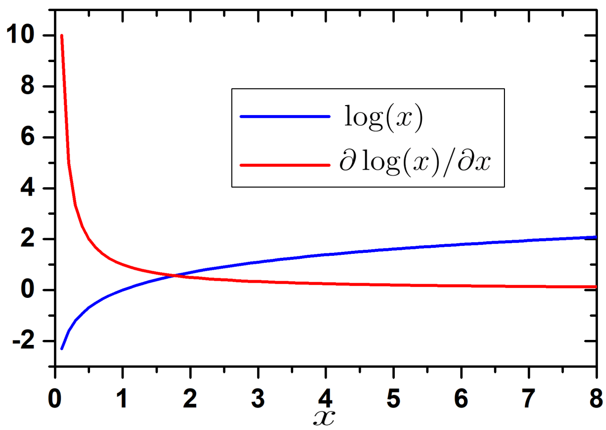

Fig. 2. Log function and its derivative when $x > 0$.

The projection matrix $\mathbf{W}$ can be trained by gradient descent algorithm with the derivative computed as

$$\frac{\partial L(\mathbf{W})}{\partial \mathbf{W}} = -\sum_{1\leq i\neq j\leq S} \frac{q_i q_j D_{\mathbf{W}}^{-1}(p_i||p_j)}{\sum_{1\leq m\neq n\leq S} q_m q_n} \frac{\partial D_{\mathbf{W}}(p_i||p_j)}{\partial \mathbf{W}} \tag{7}$$

where

$$\begin{aligned} \frac{\partial D_{\mathbf{W}}(p_i||p_j)}{\partial \mathbf{W}} &= D_{ij}\mathbf{W}(\mathbf{W}^{\mathrm{t}}\mathbf{\Sigma}\mathbf{W})^{-1} \\ &- \mathbf{\Sigma}\mathbf{W}(\mathbf{W}^{\mathrm{t}}\mathbf{\Sigma}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{t}}D_{ij}\mathbf{W}(\mathbf{W}^{\mathrm{t}}\mathbf{\Sigma}\mathbf{W})^{-1}. \end{aligned} \tag{8}$$

B. Heteroscedastic Extension

1) Motivation: Given a speaker s, its GMM supervector $\boldsymbol{m}_s$ can be represented as:

$$\boldsymbol{m}_s = \boldsymbol{m}_{\text{ubm}} + \mathbf{T}\boldsymbol{x}_s \tag{9}$$

where $\boldsymbol{m}_{\text{ubm}}$ is the universal background model (UBM) supervector. $\mathbf{T}$ is the total variability matrix. I-vector $\boldsymbol{x}_s$ and its posterior covariance matrix $\mathbf{\Lambda}_s^{-1}$ can be estimated as:

$$\begin{aligned} \boldsymbol{x}_s &= \mathbf{\Lambda}_s^{-1}\mathbf{T}^{\mathrm{t}}\mathbf{\Sigma}_{\text{ubm}}^{-1}\boldsymbol{F}_s \\ \mathbf{\Lambda}_s &= \mathbf{I} + \mathbf{T}^{\mathrm{t}}\mathbf{\Sigma}_{\text{ubm}}^{-1}\mathbf{N}_s\mathbf{T} \end{aligned} \tag{10}$$

where $\mathbf{\Sigma}_{\text{ubm}}$ is the covariance matrix of UBM. $\mathbf{N}_s$ and $\boldsymbol{F}_s$ are the zero and first Baum-Welch statistics of speaker s, respectively.

For each speaker, $\mathbf{N}_s$ is different, resulting in different covariance matrix $\mathbf{\Lambda}_s^{-1}$. However, both LDA and GDA algorithms assume that different speakers have the same covariance $\mathbf{\Sigma}$. This assumption loses speaker information and degrades system performance.

2) Realization: We extend the GDA to its heteroscedastic version (HGDA). The loss function of HGDA is the same as GDA in (6), but the representation of $D_{\mathbf{W}}(p_i||p_j)$ is different. In HGDA assumption, the ith speaker follows Gaussian distribution $p_i = \mathcal{N}(\mathbf{x}; \mu_i, \Sigma_i)$. That means different speakers have different means and covariances. The KL divergence between two speakers after projection is

$$\begin{aligned} D_{\mathbf{W}}(p_i||p_j) =& \frac{1}{2}\Big(\ln|\mathbf{W}^{\mathrm{t}}\mathbf{\Sigma}_j\mathbf{W}| - \ln|\mathbf{W}^{\mathrm{t}}\mathbf{\Sigma}_i\mathbf{W}| \\ &+ \mathrm{tr}\big((\mathbf{W}^{\mathrm{t}}\mathbf{\Sigma}_j\mathbf{W})^{-1}(\mathbf{W}^{\mathrm{t}}(\mathbf{\Sigma}_i + \mathbf{\Sigma}_{ij})\mathbf{W})\big)\Big) \end{aligned} \tag{11}$$

The derivative of KL divergence with respect to $\mathbf{W}$ is also different from (8) and computed as

$$\begin{aligned} \frac{\partial D_{\mathbf{W}}(p_i||p_j)}{\partial \mathbf{W}} &= \mathbf{\Sigma}_j\mathbf{W}(\mathbf{W}^{\mathrm{t}}\mathbf{\Sigma}_j W)^{-1} - \mathbf{\Sigma}_i\mathbf{W}(\mathbf{W}^{\mathrm{t}}\mathbf{\Sigma}_i\mathbf{W})^{-1} \\ &+ (\mathbf{\Sigma}_i + \mathbf{\Sigma}_{ij})\mathbf{W}(\mathbf{W}^{\mathrm{t}}\mathbf{\Sigma}_j\mathbf{W})^{-1} \\ &- \mathbf{\Sigma}_j\mathbf{W}(\mathbf{W}^{\mathrm{t}}\mathbf{\Sigma}_j\mathbf{W})^{-1}\mathbf{W}^{\mathrm{t}}(\mathbf{\Sigma}_i + \mathbf{\Sigma}_{ij})\mathbf{W}(\mathbf{W}^{\mathrm{t}}\mathbf{\Sigma}_j\mathbf{W})^{-1} \end{aligned} \tag{12}$$

Submit (11) and (12) to (7), we can get the derivative of the loss function and carry out the gradient descent algorithm.

IV. Experiments

A. Dataset

The proposed methods are evaluated on the i-vector machine learning challenge [20]. This challenge takes i-vectors instead of speech as input to examine the backend of speaker verification system. The enrollment set contains 1306 speakers, each of which is enrolled with 5 i-vectors. There are totally 9634 test i-vectors and 12582004 trials. The trials are randomly divided into a progress subset and an evaluation subset. In addition, a development set containing 36572 labeled i-vectors and totally 4958 speakers are also provided. All the i-vectors have 600 dimensions.

B. Configuration

The cosine system is one baseline system with the following processes: 1) Use the development data to estimate a global mean and covariance. 2) Center and whiten all the i-vectors based on the estimated mean and covariance. 3) Do the length normalization for all the i-vectors. 4) For each model, average its five i-vectors and then do the length normalization for the resulting average-model i-vector. 5) Compute the inner product between all the average-model i-vectors and test i-vectors.

PLDA system has the same processes of cosine system, except in step 5), it uses PLDA instead of inner product to make scores. The dimension of PLDA is 100.

LDA/GDA/HGDA-cosine and LDA/GDA/HGDA-PLDA systems are the same as cosine and PLDA systems, respectively. Except that before step 5), all the 600-dimensional i-vectors are reduced to 200-dimensional by LDA/GDA/HGDA and then their length are re-normalized. The covariance matrix in GDA is set to be full. While in HGDA, the covariance matrix of each speaker is diagonal. Mini-batch gradient descent with momentum is used to train GDA and HGDA. The batch size is 128 and the weight of last gradient is 0.8. As the concavity of $L(\mathbf{W})$ cannot be guaranteed, we randomly initial the transform matrix 5 times and select the one with best performance. It should be noted that, after the training of GDA and HGDA, $\mathbf{W}$ needs to be orth-normalized.

Performance is evaluated using the equal error rate (EER) and minimum decision cost function (MDCF) defined in [20].

C. Results Analysis

Three groups of experiments are carried out in the i-vector machine learning challenge. Each group explores the effect of LDA, GDA, and HGDA dimension reduction methods in combination with cosine or PLDA scoring method. In the first group, all the data in development and enrollment set, totally 6264 speakers, are used to train LDA, GDA, HGDA, and PLDA. In the second group, only 3000 speakers in development set are chosen as training data. For the third group, only 1500 speakers in development set are used for training.

TABLE I
EXPERIMENT RESULTS OF DIFFERENT SYSTEMS. ALL DATA ARE USED AS TRAINING DATA.

Methods	Evaluation Set		Progress Set	
	EER[%]	MDCF	EER[%]	MDCF
cosine	4.49	0.378	5.16	0.386
LDA-cosine	3.61	0.310	4.14	0.319
GDA-cosine	3.52	0.304	3.78	0.315
HGDA-cosine	3.39	0.301	3.67	0.310
PLDA	2.92	0.299	3.22	0.308
LDA-PLDA	2.44	0.280	2.56	0.292
GDA-PLDA	2.35	0.274	2.50	0.286
HGDA-PLDA	2.34	0.272	2.47	0.285

The results of these three groups are list in Table I, II, and III, respectively. From each of the three tables, we can see that, with a same cosine or PLDA scoring method, GDA has better performance than LDA. We think the reason for the improvement is that LDA may merge two speakers with a small divergence in the projected subspace so that large divergences can be preserved as much as possible to meet the goal of arithmetic mean maximization. In contrast, GDA focuses on maximizing geometric mean, which emphasizes enlarging small divergence, and can better separate easily-confused speakers. HGDA further improves the system performance. This demonstrates the effectiveness of modeling different speakers with different covariances. The relative performance improvement of HGDA-PLDA compared with LDA-PLDA is smaller than that of HGDA-cosine compared with LDA-cosine. It maybe because PLDA also models different speakers with a same covariance. Its assumption is in contradiction with that of HGDA.

TABLE II
EXPERIMENT RESULTS COMPARISON. ONLY 3000 SPEAKERS IN DEVELOPMENT SET ARE CHOSEN AS TRAINING DATA.

Methods	Evaluation Set		Progress Set	
	EER[%]	MDCF	EER[%]	MDCF
cosine	4.49	0.378	5.16	0.386
LDA-cosine	3.80	0.326	4.23	0.337
GDA-cosine	3.63	0.309	3.84	0.324
HGDA-cosine	3.49	0.304	3.76	0.318
PLDA	3.37	0.329	3.53	0.345
LDA-PLDA	2.98	0.318	3.07	0.333
GDA-PLDA	2.86	0.301	2.95	0.318
HGDA-PLDA	2.81	0.297	2.86	0.315

TABLE III
EXPERIMENT RESULTS COMPARISON. ONLY 1500 SPEAKERS IN DEVELOPMENT SET ARE CHOSEN AS TRAINING DATA.

Methods	Evaluation Set		Progress Set	
	EER[%]	MDCF	EER[%]	MDCF
cosine	4.49	0.378	5.16	0.386
LDA-cosine	3.97	0.335	4.41	0.352
GDA-cosine	3.74	0.316	3.91	0.329
HGDA-cosine	3.55	0.308	3.79	0.324
PLDA	3.40	0.345	3.73	0.363
LDA-PLDA	3.03	0.331	3.34	0.346
GDA-PLDA	2.89	0.305	3.10	0.320
HGDA-PLDA	2.82	0.306	3.02	0.323

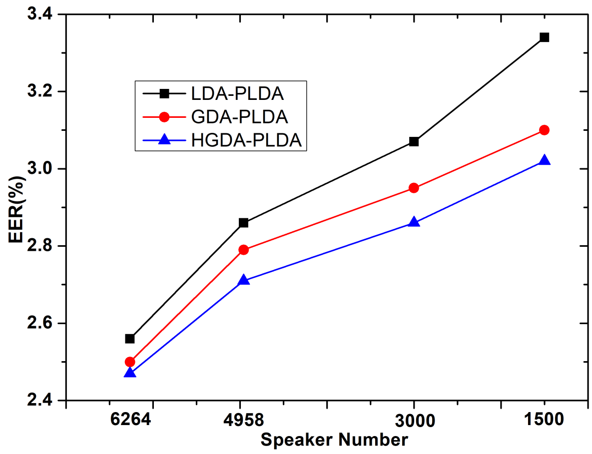

Fig. 3. The Variations of EER with Speaker Number.

In order to compare the effect of different dimension reduction methods more intuitive, the variations of EER with training speaker number of LDA-PLDA, GDA-PLDA, and HGDA-PLDA systems in progress set are given in Fig. 3. It can be seen that when the training speaker number becomes smaller, the performance of LDA degrades more severely than GDA and HGDA. The relative performance improvement of GDA and HGDA goes larger. We suspect that when the speakers number is smaller, the speakers tend to distribute in the space more sparsely. And the situation in Fig. 1 occurs more often. Thus, GDA and HGDA are more effective.

V. Conclusion

In this paper, we explore the application of geometric discriminant analysis (GDA) to i-vector based speaker verification. It maximizes the geometric mean of KL divergences instead of the arithmetic mean used in LDA. Thus it can put more emphasis on separating the speakers with small KL divergences. GDA is further extended to heteroscedastic version (HGDA). HGDA can model the i-vector distributions of different speakers more precisely. Experiment results indicate that, when the training speaker number becomes smaller, the relative performance improvement of GDA and HGDA compared with LDA becomes larger. GDA and HGDA are better choices especially when training data is limited.

In our future work, we will explore the performances of GDA and HGDA for x-vector based speaker verification system.

VI. Acknowledgment

This work was supported by the China Postdoctoral Science Foundation under Grant No. 2018M641355, and the National Natural Science Foundation of China under Grant No. 61403224 and No. U1836219.

References

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, May 2011.

[2] N. Dehak, P. A.Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association, Interspeech 2011*, 2011, pp. 857–860.

[3] S. O. Sadjadi, S. Ganapathy, and J. W. Pelecanos, "Speaker age estimation on conversational telephone speech using senone posterior based i-vectors," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5040–5044.

[4] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proceedings of the IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[5] A. M. Martinez and A. C. Kak, "Pca versus lda," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, Feb 2001.

[6] D. Tao, X. Li, X. Wu, and S. J. Maybank, "Geometric mean for subspace selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 260–274, Feb 2009.

[7] M. Loog, R. P. W. Duin, and R. Haeb-Umbach, "Multiclass linear dimension reduction by weighted pairwise fisher criteria," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 7, pp. 762–766, July 2001.

[8] A. Kanagasundaram, D. Dean, R. Vogt, M. McLaren, S. Sridharan, and M. Mason, "Weighted lda techniques for i-vector based speaker verification," in *Acoustics, Speech, and Signal Processing (ICASSP), 2014IEEE International Conference on*, 2012, pp. 4781 – 4784.

[9] L. He, X. Chen, C. Xu, J. Liu, and M. T. Johnson, "Local pairwise linear discriminant analysis for speaker verification," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1575–1579, 2018.

[10] S. Zhong-Hua, P. Yong-Hui, and W. Shi-Tong, "A supervised locality preserving projection algorithm for dimensionality reduction," *Pattern Recognition and Artificial Intelligence*, vol. 21, no. 2, p. 233C239, 2008.

[11] L. You, W. Guo, Y. Song, and S. Zhang, "Improved supervised locality preserving projection for i-vector based speaker verification," in *Interspeech 2018*, 2018, pp. 62–66.

[12] X. Chen, L. He, C. Xu, and J. Liu, "Distance-dependent metric learning," *IEEE Signal Processing Letters*, vol. 26, no. 2, pp. 357–361, Feb 2019.

[13] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Odyssey 2010: The Speaker and Language Recognition Workshop*, 2010.

[14] S. O. Sadjadi, J. W. Pelecanos, and W. Zhu, "Nearest neighbor discriminant analysis for robust speaker recognition," in *INTERSPEECH*, 2014, pp. 1860–1864.

[15] S. O. Sadjadi, S. Ganapathy, and J. W. Pelecanos, "The ibm 2016 speaker recognition system," in *Odyssey 2016: The Speaker and Language Recognition Workshop*, 2016, pp. 174–180.

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

[17] A. Kanagasundaram, D. Dean, S. Sridharan, J. Gonzalez-Dominguez, J. Gonzalez-Rodriguez, and D. Ramos, "Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques," *Speech Communication*, vol. 59, pp. 69 – 82, 2014.

[18] P. M. Bousquet, A. Larcher, D. Matrouf, J. F. Bonastre, and O. Plchot, "Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis," *Odyssey 2012: The Speaker and Language Recognition Workshop*, 2012.

[19] M. Loog, *Approximate Pairwise Accuracy Criteria for Multiclass Linear Dimension Reduction: Generalizations of the Fisher Criterion.* Delft University Press, 1999.

[20] "The 2013-2014 speaker recognition i-vector machine learning challenge," https://www.nist.gov/itl/iad/mig/i-vector-machine-learning-challenge.