Generalized Fisher Discriminant Analysis as A Dimensionality Reduction Technique

Yuechi Jiang
Electronic and Information Engineering
The Hong Kong Polytechnic University
Hong Kong, China
yuechi.jiang@connect.polyu.hk

Frank H. F. Leung
Electronic and Information Engineering
The Hong Kong Polytechnic University
Hong Kong, China
frank-h-f.leung@polyu.edu.hk

***Abstract*—Fisher Discriminant Analysis (FDA) has been widely used as a dimensionality reduction technique. Its application varies from face recognition to speaker recognition. In the past two decades, there have been many variations on the formulation of FDA. Different variations adopt different ways to combine the between-class scatter matrix and the within-class scatter matrix, which are two basic components in FDA. In this paper, we propose the Generalized Fisher Discriminant Analysis (GFDA), which provides a general formulation for FDA. GFDA generalizes the standard FDA as well as many different variants of FDA, such as Regularized Linear Discriminant Analysis (R-LDA), Regularized Kernel Discriminant Analysis (R-KDA), Inverse Fisher Discriminant Analysis (IFDA), and Regularized Fisher Discriminant Analysis (RFDA). GFDA can also degenerate to Principal Component Analysis (PCA). Four special types of GFDA are then applied as dimensionality reduction techniques for speaker recognition, in order to investigate the performance of different variants of FDA. Basically, GFDA provides a convenient way to compare different variants of FDA by simply changing some parameters. It makes it easier to explore the roles that the between-class scatter matrix and the within-class scatter matrix play.**

Keywords—Generalized Fisher discriminant analysis; dimensionality reduction; speaker recognition

I. INTRODUCTION

Fisher Discriminant Analysis (FDA) has been widely used as a feature dimensionality reduction technique or used for classification. It has been extensively applied to face recognition to reduce the dimensionality of the input face image, such as the usage of Fisherfaces [1]-[3]. It is also found to be applicable to speaker recognition, such as the usage of Fishervoices [4]-[7], and the usage as a projection technique to suppress unwanted information existing in the speech signal [8]. It has also been used together with Sparse Representation (SR) to increase the discrimination of the features [9][10], or used together with factor analysis to build a Probabilistic Linear Discriminant Analysis (PLDA) model for signal classification [11][12].

FDA aims at mapping the original features to a projected feature space, where features within the same class are closer to each other while features of different classes are farther from each other. FDA includes two versions, one is the linear version called Fisher Linear Discriminant Analysis (LDA), and the other is the kernel version called Fisher Kernel Discriminant Analysis (KDA). KDA is the generalization of LDA, which can involve an implicit feature mapping before projection. When the mapping is linear, KDA degenerates to LDA. Equipped with a kernel, the performance of KDA is comparable to that of other general-purpose classifiers, such as Support Vector Machine (SVM) [13][14]. However, a potential issue about applying FDA is the Small Sample Size (SSS) problem, where the dimensionality of the feature is larger than the number of features, such as face recognition [15]. This SSS problem may cause the matrix being inverted in FDA to be singular. One solution to SSS is to first utilize Principal Component Analysis (PCA) to reduce the dimensionality of the original feature, and then utilize FDA for further projection [1][3]. However, PCA may discard useful discriminative information embedded in the original feature, which degrades the effectiveness of the further FDA process [16]. Another solution is to add a regularization term so that the matrix being inverted is not singular [13][17]. However, the difficulty lies in how to choose a proper regularization parameter.

In the literature, different variations on the original formulation of FDA have been proposed and evaluated, such as Regularized Linear Discriminant Analysis (R-LDA) [16], Regularized Kernel Discriminant Analysis (R-KDA) [18], Doubly Regularized Linear Discriminant Analysis (D-RLDA) and Doubly Regularized Kernel Discriminant Analysis (D-RKDA) [19], Inverse Linear Discriminant Analysis (ILDA) [20], Inverse Kernel Discriminant Analysis (IKDA) [21], and Regularized Fisher Discriminant Analysis (RFDA) [22]. Basically, FDA comprises a between-class scatter matrix and a within-class scatter matrix, and different variations on FDA mainly differ in the ways to combine the two scatter matrices.

In this paper, we propose the Generalized Fisher Discriminant Analysis (GFDA), which aims at generalizing different variations on FDA to a single formula. Both the non-kernel version (called GLDA) and the kernel version (called GKDA) of GFDA are given. We then apply four special types of GFDA as dimensionality reduction techniques for speaker recognition. Regarding speaker recognition, Gaussian Supervector (GSV) and i-vector are two most widely used features [8]. Nevertheless, GSV is more general and computationally efficient than i-vector, and thus has wider

The work described in this paper was substantially supported by a grant from The Hong Kong Polytechnic University (Project Account Code: RUG7).

applications in speech-related recognition tasks, such as speech recording device recognition, including microphone recognition [22][23] and mobile phone recognition [24][25]. Therefore, we adopt GSV as the feature for speaker recognition in this paper, and then apply the four dimensionality reduction techniques to investigate their effectiveness. Linear Support Vector Machine (SVM) is a widely used linear classifier for high-dimension signal classification, especially for speech-related classification tasks [8][22]-[24][26]. Thus, we employ linear SVM as the classifier for speaker recognition.

This paper is organized as follows. In Section II, the formulation of GLDA as well as the relationship between GLDA and several variants of LDA is given. In Section III, the formulation of GKDA is given, together with its relationship with several variants of KDA. In Section IV, the speech dataset and experimental setting are briefly described. In Section V, the experimental results and discussions are presented. In Section VI, a conclusion is drawn.

II. Non-kernel Version GFDA

A. Non-kernel Version GFDA (GLDA)

Suppose we have N training feature vectors denoted as $\{x_1, x_2 \ldots x_N\}$. Each feature vector is associated with a class label. Assume there are totally K different classes, with the k-th class denoted as C_k, containing N_k training feature vectors. Obviously, we have $N = \sum_{k=1}^{K} N_k$. Based on these N training feature vectors, the proposed GLDA targets at finding a projection matrix W, by optimizing the objective function $J(W)$ defined in (1), where α_1, β_1, γ_1, α_2, β_2, γ_2 are pre-defined parameters, S_B is the between-class scatter matrix defined in (2), S_W is the within-class scatter matrix defined in (3), I is an identity matrix. In (2) and (3), m_k is the mean vector of the training vectors in class C_k as given by (4), and m is the mean vector of all the training vectors as given by (5) [27].

$$J(W) = \frac{\left|W^T\left(\alpha_1 S_B + \beta_1 S_W + \gamma_1 I\right)W\right|}{\left|W^T\left(\alpha_2 S_B + \beta_2 S_W + \gamma_2 I\right)W\right|} \tag{1}$$

where

$$S_B = \sum_{k=1}^{K} N_k \left(m_k - m\right)\left(m_k - m\right)^T \tag{2}$$

$$S_W = \sum_{k=1}^{K} \sum_{x_n \in C_k} \left(x_n - m_k\right)\left(x_n - m_k\right)^T \tag{3}$$

$$m_k = \frac{1}{N_k} \sum_{x_n \in C_k} x_n \tag{4}$$

$$m = \frac{1}{N} \sum_{n=1}^{N} x_n \tag{5}$$

Based on the results of LDA [27], W is the collection of the eigenvectors of $\left(\alpha_2 S_B + \beta_2 S_W + \gamma_2 I\right)^{-1}\left(\alpha_1 S_B + \beta_1 S_W + \gamma_1 I\right)$, meaning that the i-th column w_i in W is namely the i-th eigenvector of $\left(\alpha_2 S_B + \beta_2 S_W + \gamma_2 I\right)^{-1}\left(\alpha_1 S_B + \beta_1 S_W + \gamma_1 I\right)$. $\gamma_1 I$ and $\gamma_2 I$ can be used to increase the rank of $\left(\alpha_1 S_B + \beta_1 S_W + \gamma_1 I\right)$ and $\left(\alpha_2 S_B + \beta_2 S_W + \gamma_2 I\right)$ respectively, so as to increase the number of orthogonal eigenvectors. In addition, $\gamma_2 I$ can also be used to prevent $\left(\alpha_2 S_B + \beta_2 S_W + \gamma_2 I\right)$ to be singular during matrix inversion. Having W, for a feature vector x_t, the projected feature y_t is then calculated using (6).

$$y_t = W^T x_t \tag{6}$$

B. Relationship between GLDA and Other Variants of LDA

When (α_1, β_1, γ_1, α_2, β_2, γ_2)=(1, 0, 0, 0, 1, 0), GLDA becomes standard LDA, which aims to maximize the objective function defined in (7), whose solution is the collection of the eigenvectors of $S_W^{-1} S_B$ [1].

$$J(W) = \frac{\left|W^T S_B W\right|}{\left|W^T S_W W\right|} \tag{7}$$

When (α_1, β_1, γ_1, α_2, β_2, γ_2)=(1, 0, 0, η, 1−η, 0), GLDA becomes Regularized LDA (R-LDA), which aims to maximize the objective function defined in (8), where η is a weighting parameter. The solution of R-LDA is the collection of the eigenvectors of $\left(\eta S_B + (1-\eta) S_W\right)^{-1} S_B$ [16]. In fact, R-LDA becomes standard LDA when η=0.

$$J(W) = \frac{\left|W^T S_B W\right|}{\left|W^T\left(\eta S_B + (1-\eta) S_W\right)W\right|} \tag{8}$$

When (α_1, β_1, γ_1, α_2, β_2, γ_2)=(1, 0, ε_1, 0, 1, ε_2), GLDA becomes Doubly Regularized LDA (D-RLDA), which aims to maximize the objective function defined in (9), where ε_1 and ε_2 are regularization parameters. The solution of D-RLDA is the collection of the eigenvectors of $\left(S_W + \varepsilon_2 I\right)^{-1}\left(S_B + \varepsilon_1 I\right)$ [19]. In fact, D-RLDA becomes standard LDA when ε_1=0 and ε_2=0.

$$J(W) = \frac{\left|W^T\left(S_B + \varepsilon_1 I\right)W\right|}{\left|W^T\left(S_W + \varepsilon_2 I\right)W\right|} \tag{9}$$

When (α_1, β_1, γ_1, α_2, β_2, γ_2)=(0, 1, 0, 1, 0, 0), GLDA becomes Inverse LDA (ILDA), which aims to minimize the objective function defined in (10), whose solution is the collection of the eigenvectors of $S_B^{-1} S_W$ [20]. Maximizing (7) is equivalent to minimizing (10), meaning that ILDA and LDA are theoretically equivalent.

$$J(W) = \frac{\left|W^T S_W W\right|}{\left|W^T S_B W\right|} \tag{10}$$

When (α_1, β_1, γ_1, α_2, β_2, γ_2)=(1, −λ, 0, 0, 0, 1), GLDA becomes Regularized FDA (RFDA), which aims to maximize the objective function defined in (11), where λ is a weighting parameter. The solution of RFDA is the collection of the eigenvectors of $S_B - \lambda S_W$ [22].

$$J(W) = \left|W^T S_B W\right| - \lambda \left|W^T S_W W\right| \tag{11}$$

When (α_1, β_1, γ_1, α_2, β_2, γ_2)=(1, 1, 0, 0, 0, 1), GLDA becomes Principal Component Analysis (PCA), which aims to maximize the objective function defined in (12), where S_T is the total scatter matrix [1], which is the summation of S_B and S_W [28]. The solution to (12) is the collection of the eigenvectors of S_B+S_W.

$$J(W)=\left|W^T S_T W\right|=\left|W^T\left(S_B+S_W\right)W\right| \tag{12}$$

III. Kernel Version GFDA

A. Kernel Version GFDA (GKDA)

It has been shown in [13] that a projection direction w_i can be expressed as a linear combination of all the training vectors, as given by (13), where v_i is an N×1 column vector and $(v_i)_n$ is its n-th element, and X is a matrix consisting of all the training vectors $\{x_1, x_2 \ldots x_N\}$ whose n-th column vector is namely x_n.

$$w_i=\sum_{n=1}^{N} x_n\left(v_i\right)_n = Xv_i \tag{13}$$

By defining a new matrix V whose i-th column vector is v_i, we can rewrite (13) into a more compact expression as in (14).

$$W=XV \tag{14}$$

With the help of (14), we can then express $W^T m_k$, $W^T m$, $W^T x_n$ in terms of V, as given by (15) ~ (17), where we define three new notations l_k, l and q_n to be N×1 column vectors whose j-th element is given by (18) ~ (20).

$$W^T m_k = V^T X^T\left(\frac{1}{N_k}\sum_{x_n\in C_k} x_n\right)=V^T\frac{1}{N_k}X^T\sum_{x_n\in C_k} x_n = V^T l_k \tag{15}$$

$$W^T m = V^T X^T\left(\frac{1}{N}\sum_{n=1}^{N} x_n\right)=V^T\frac{1}{N}X^T\sum_{n=1}^{N} x_n = V^T l \tag{16}$$

$$W^T x_n = V^T X^T x_n = V^T q_n \tag{17}$$

where

$$\left(l_k\right)_j=\frac{1}{N_k}x_j^T\sum_{x_n\in C_k} x_n=\frac{1}{N_k}\sum_{x_n\in C_k} x_j^T x_n=\frac{1}{N_k}\sum_{x_n\in C_k} k\left(x_j,x_n\right) \tag{18}$$

$$\left(l\right)_j=\frac{1}{N}x_j^T\sum_{n=1}^{N} x_n=\frac{1}{N}\sum_{n=1}^{N} x_j^T x_n=\frac{1}{N}\sum_{n=1}^{N} k\left(x_j,x_n\right) \tag{19}$$

$$\left(q_n\right)_j = x_j^T x_n = k\left(x_j,x_n\right) \tag{20}$$

In (18) ~ (20) above, $k(x_j, x_n)$ is the kernel function. The kernel function defines the inner product of two vectors, while the explicit expression of the vector is unnecessary to be known. This is the so-called "kernel trick", which enables people to define a kernel function involving an implicit feature mapping before projection. For example, if we define the kernel function like $k(x_j, x_n)=\varphi(x_j)^T\varphi(x_n)$, this kernel function involves an implicit feature mapping φ: $x\rightarrow\varphi(x)$.

Based on (15) ~ (20), $W^T S_B W$ and $W^T S_W W$ can be expressed in terms of V, as given by (21) and (22), where in (21) and (22), U_B and U_W are N×N matrices defined in (23) and (24) [22].

$$\begin{aligned} W^T S_B W &= \sum_{k=1}^{K} N_k W^T\left(m_k-m\right)\left(m_k-m\right)^T W \\ &= \sum_{k=1}^{K} N_k V^T\left(l_k-l\right)\left(l_k-l\right)^T V = V^T U_B V \end{aligned} \tag{21}$$

$$\begin{aligned} W^T S_W W &= \sum_{k=1}^{K}\sum_{x_n\in C_k} W^T\left(x_n-m_k\right)\left(x_n-m_k\right)^T W \\ &= \sum_{k=1}^{K}\sum_{x_n\in C_k} V^T\left(q_n-l_k\right)\left(q_n-l_k\right)^T V = V^T U_W V \end{aligned} \tag{22}$$

where

$$U_B=\sum_{k=1}^{K} N_k\left(l_k-l\right)\left(l_k-l\right)^T \tag{23}$$

$$U_W=\sum_{k=1}^{K}\sum_{x_n\in C_k}\left(q_n-l_k\right)\left(q_n-l_k\right)^T \tag{24}$$

Adopting the new expressions in (21) and (22), (1) can be rewritten as (25), which is the objective function of GKDA.

$$J(V)=\frac{\left|V^T\left(\alpha_1 U_B+\beta_1 U_W+\gamma_1 I\right)V\right|}{\left|V^T\left(\alpha_2 U_B+\beta_2 U_W+\gamma_2 I\right)V\right|} \tag{25}$$

The solution V of optimizing the objective function defined in (25) is then the collection of the eigenvectors of $\left(\alpha_2 U_B+\beta_2 U_W+\gamma_2 I\right)^{-1}\left(\alpha_1 U_B+\beta_1 U_W+\gamma_1 I\right)$. After solving (25) for v_i, for a feature vector x_t, the projected feature y_t is calculated using (26), where P is the number of eigenvectors.

$$\begin{aligned} y_t &= W^T x_t = \left(XV\right)^T x_t = V^T X^T x_t \\ &= \begin{bmatrix} \sum_{n=1}^{N}\left(v_1\right)_n x_n^T x_t \\ \sum_{n=1}^{N}\left(v_2\right)_n x_n^T x_t \\ \vdots \\ \sum_{n=1}^{N}\left(v_P\right)_n x_n^T x_t \end{bmatrix} = \begin{bmatrix} \sum_{n=1}^{N}\left(v_1\right)_n k\left(x_n,x_t\right) \\ \sum_{n=1}^{N}\left(v_2\right)_n k\left(x_n,x_t\right) \\ \vdots \\ \sum_{n=1}^{N}\left(v_P\right)_n k\left(x_n,x_t\right) \end{bmatrix} \end{aligned} \tag{26}$$

B. Relationship between GKDA and Other Variants of KDA

When (α_1, β_1, γ_1, α_2, β_2, γ_2)=(1, 0, 0, 0, 1, 0), GKDA becomes KDA, which aims to maximize the objective function defined in (27), whose solution is the collection of the eigenvectors of $U_W^{-1}U_B$ [13][14].

$$J(V)=\frac{\left|V^T U_B V\right|}{\left|V^T U_W V\right|} \tag{27}$$

When (α_1, β_1, γ_1, α_2, β_2, γ_2)=(1, 0, 0, η, 1−η, 0), GKDA becomes Regularized KDA (R-KDA), which aims to maximize the objective function defined in (28), where η is a weighting parameter. The solution of R-KDA is the collection of the eigenvectors of $\left(\eta U_B+\left(1-\eta\right)U_W\right)^{-1}U_B$ [18].

$$J(V)=\frac{\left|V^T U_B V\right|}{\left|V^T\left(\eta U_B+\left(1-\eta\right)U_W\right)V\right|} \tag{28}$$

When $(\alpha_1, \beta_1, \gamma_1, \alpha_2, \beta_2, \gamma_2)=(1, 0, \varepsilon_1, 0, 1, \varepsilon_2)$, GKDA becomes Doubly Regularized KDA (D-RKDA), which aims to maximize the objective function defined in (29), where ε_1 and ε_2 are regularization parameters. The solution of D-RKDA is the collection of the eigenvectors of $(U_W+\varepsilon_2 I)^{-1}(U_B+\varepsilon_1 I)$ [19].

$$J(V)=\frac{\left|V^T(U_B+\varepsilon_1 I)V\right|}{\left|V^T(U_W+\varepsilon_2 I)V\right|} \tag{29}$$

When $(\alpha_1, \beta_1, \gamma_1, \alpha_2, \beta_2, \gamma_2)=(0, 1, 0, 1, 0, 0)$, GKDA becomes Inverse KDA (IKDA), which aims to minimize the objective function defined in (30), whose solution is the collection of the eigenvectors of $U_B^{-1}U_W$ [21].

$$J(V)=\frac{\left|V^T U_W V\right|}{\left|V^T U_B V\right|} \tag{30}$$

When $(\alpha_1, \beta_1, \gamma_1, \alpha_2, \beta_2, \gamma_2)=(1, -\lambda, 0, 0, 0, 1)$, GKDA becomes Kernel RFDA (K-RFDA), which aims to maximize the objective function defined in (31), where λ is a weighting parameter. The solution of K-RFDA is the collection of the eigenvectors of $U_B-\lambda U_W$ [22].

$$J(V)=\left|V^T U_B V\right|-\lambda\left|V^T U_W V\right| \tag{31}$$

When $(\alpha_1, \beta_1, \gamma_1, \alpha_2, \beta_2, \gamma_2)=(1, 1, 0, 0, 0, 1)$, GKDA becomes Kernel PCA (KPCA), which aims to maximize the objective function defined in (32), whose solution is the collection of the eigenvectors of U_B+U_W [3].

$$J(V)=\left|V^T U_T V\right|=\left|V^T(U_B+U_W)V\right| \tag{32}$$

IV. Speech Dataset and Experimental Setting

GSV is utilized as the original feature, which is calculated by adapting a 32-mixture Universal Background Model (UBM) with a relevance factor of 5 [29][30]. Linear SVM is employed as the classifier, which is implemented using LIBSVM [31]. Four special types of GFDA are then applied on GSV as the dimensionality reduction techniques, including two non-kernel projection techniques R-LDA, RFDA and two kernel-based projection techniques R-KDA, K-RFDA. R-LDA is a generalization of LDA, thus can be a good representative. RFDA is quite different from LDA regarding the formulation, thus can also be a representative. So do the kernel-based projection techniques. When applying the kernel-based techniques, only the linear kernel is used, meaning that $k(a,b)=a^T b$. In this way, no implicit feature mapping is involved before projection, thus the non-kernel and kernel-based techniques are both linear and hence comparable.

The speaker recognition dataset is the Ahumada-25 speech corpus [32]. There are totally 25 speakers to be recognized. Each speaker contributes approximately the same number of speeches, recorded using different microphones. Totally there are 1440 speeches used for training, 1440 speeches used for testing, and 1019 speeches used for UBM.

When applying R-LDA and R-KDA, to prevent the matrix being inverted to be singular, an extra regularization term εI is added, where ε is a regularization parameter and I is an identity matrix. Then the solution to R-LDA and R-KDA becomes the collection of the eigenvectors of $(\eta S_B+(1-\eta)S_W+\varepsilon I)^{-1}S_B$ and $(\eta U_B+(1-\eta)U_W+\varepsilon I)^{-1}U_B$ respectively. Different values of η and ε are investigated in the experiments. While on applying RFDA and K-RFDA, it is unnecessary to add an extra regularization term, as matrix inversion is not involved. The solution to RFDA and K-RFDA is the collection of the eigenvectors of $S_B-\lambda S_W$ and $U_B-\lambda U_W$, where λ is the parameter being investigated in the experiments.

V. Experimental Results and Discussions

In this section, we compare the dimensionality reduction capabilities of the projection techniques when training data are sufficient (i.e. all the 1440 training data are involved in calculating the projection matrix) and insufficient (i.e. only 720 training data are involved in calculating the projection matrix). Results are shown in Figs. 1 ~ 4. In these figures, the x-axis is the reduced dimensionality of the projected features. Since each projection direction is an eigenvector, the projection directions can be sorted according to the absolute value of the corresponding eigenvalues, and the larger the eigenvalue, the more significant the projection direction is. Therefore, by choosing only a few significant projection directions, the dimensionality can be reduced.

A. Dimensionality Reduction Ability of R-LDA and RFDA

Results on applying R-LDA and RFDA are shown in Figs. 1 and 2. When the training data are sufficient (Fig. 1), the performance of R-LDA and RFDA degrades with the decrease of the dimensionality of the projected feature, and the latter degrades even faster. Nevertheless, when the dimensionality is high enough (e.g. 20), both R-LDA and RFDA work well. When the training data are insufficient (Fig. 2), the performance of R-LDA is quite stable under different dimensionalities, while that of RFDA is still highly affected by the dimensionality. Even when the dimensionality is high enough (e.g. 768), R-LDA still outperforms RFDA a lot. Comparing Fig. 1 and Fig. 2, the performance of R-LDA and RFDA with insufficient training data is worse than that with sufficient training data. When training data are insufficient, there is no enough information to find good projection directions. On applying RFDA, the performance is highly dependent on parameter λ. On applying R-LDA, while η is relatively small, ε should be relatively large; while η is relatively large, ε should be relatively small. Although both R-LDA and RFDA can work well if the dimensionality is high enough, R-LDA is less affected by the dimensionality.

B. Dimensionality Reduction Ability of R-KDA and K-RFDA

Results on applying R-KDA and K-RFDA with linear kernel are shown in Figs. 3 and 4. When the training data are sufficient (Fig. 3), the performance of R-KDA and K-RFDA is good when the dimensionality is high enough (e.g. 20), but degrades with the decrease of the dimensionality. When the training data are insufficient (Fig. 4), the performance of R-

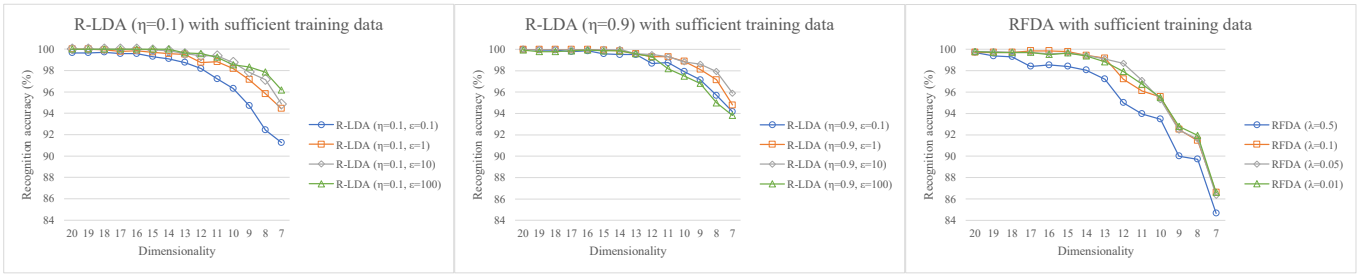

Fig. 1. Recognition accuracy with respect to dimensionality, using R-LDA and RFDA, when training data are sufficient.

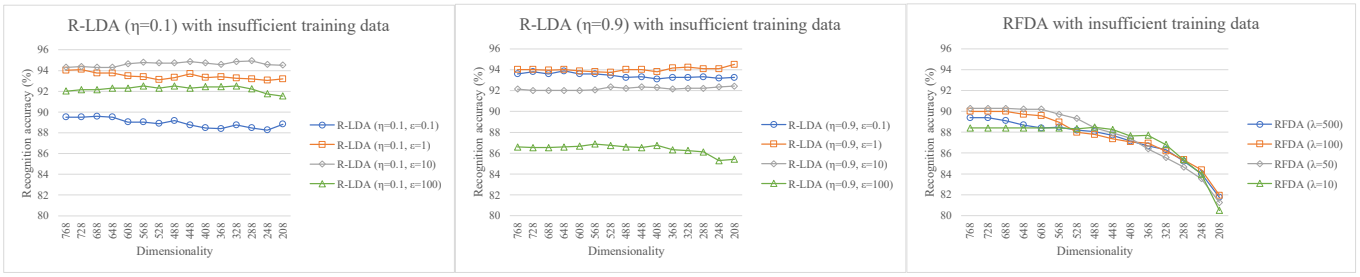

Fig. 2. Recognition accuracy with respect to dimensionality, using R-LDA and RFDA, when training data are insufficient.

KDA is quite stable with respect to different dimensionalities, while K-RFDA is not. Comparing Fig. 3 and Fig. 4, the performance of R-KDA and K-RFDA is worse when the training data are insufficient. Besides, the performance of R-KDA is more dependent on ε than η.

C. Brief Comparison

In this part, we make a brief comparison between non-kernel projection techniques (i.e. R-LDA and RFDA) and kernel-based projection techniques (i.e. R-KDA and K-RFDA). When the training data are sufficient (Figs. 1 and 3), both R-LDA and R-KDA are insensitive to the dimensionality, while K-RFDA is more sensitive to the dimensionality than RFDA. The highest accuracies achieved by the non-kernel and kernel-based projection techniques are almost the same. When training data are insufficient (Figs. 2 and 4), R-KDA is less sensitive to ε and η than R-LDA. In addition, the highest accuracies achieved by the kernel-based projection techniques can be higher than those achieved by the non-kernel projection techniques. The non-kernel techniques and the kernel-based techniques can be complementary to each other. Suppose the original dimensionality of the training data is D and the number of training data is N. Then for the non-kernel projection techniques, the objective is to find the eigenvectors of $(\alpha_2 S_B + \beta_2 S_W + \gamma_2 I)^{-1}(\alpha_1 S_B + \beta_1 S_W + \gamma_1 I)$, whose dimensionality is $D \times D$. While for the kernel-based projection techniques, the objective is then to find the eigenvectors of $(\alpha_2 U_B + \beta_2 U_W + \gamma_2 I)^{-1}(\alpha_1 U_B + \beta_1 U_W + \gamma_1 I)$, whose dimensionality is $N \times N$. So, if $D<N$, we can apply the non-kernel techniques because calculating the eigenvectors of a $D \times D$ matrix is more efficient than an $N \times N$ matrix, otherwise we resort to the kernel-based techniques.

VI. Conclusion

In this paper, we propose Generalized Fisher Discriminant Analysis (GFDA), which is the generalization of many different variants of Fisher Discriminant Analysis (FDA). We describe both the non-kernel version GFDA (i.e. GLDA) and the kernel version GFDA (i.e. GKDA). Then four special types of GFDA are compared as dimensionality reduction techniques for speaker recognition, including both the non-kernel projection techniques and the kernel-based projection techniques. Experimental results demonstrate that the performances of the non-kernel and kernel-based techniques are similar, which provides the flexibility to choose a proper projection technique. For example, if the dimensionality of the training data is lower than the number of training data, it is more efficient to use the non-kernel projection techniques, otherwise we had better resort to the kernel-based projection techniques. The proposed GFDA provides a general formulation that summarizes different variants of FDA, making comparing different variants more convenient.

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711-720, 1997.

[2] M. H. Yang, "Kernel eigenfaces vs. kernel fisherfaces: face recognition using kernel methods," in *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2002.

[3] J. Yang, A. F. Frangi, J. Y. Yang, D. Zhang, and Z. Jin, "KPCA plus LDA: a complete kernel fisher discriminant framework for feature extraction and recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 230-244, 2005.

[4] S. M. Chu, H. Tang, and T. S. Huang, "Fishervoice and semi-supervised speaker clustering," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009, pp. 4089-4092.

[5] Z. Li, W. Jiang, and H. Meng, "Fishervoice: a discriminant subspace framework for speaker recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 4522-4525.

[6] C. Zhang and T. F. Zheng, "A fishervoice based feature fusion method for short utterance speaker recognition," in *Proc. IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP)*, 2013, pp. 165-169.

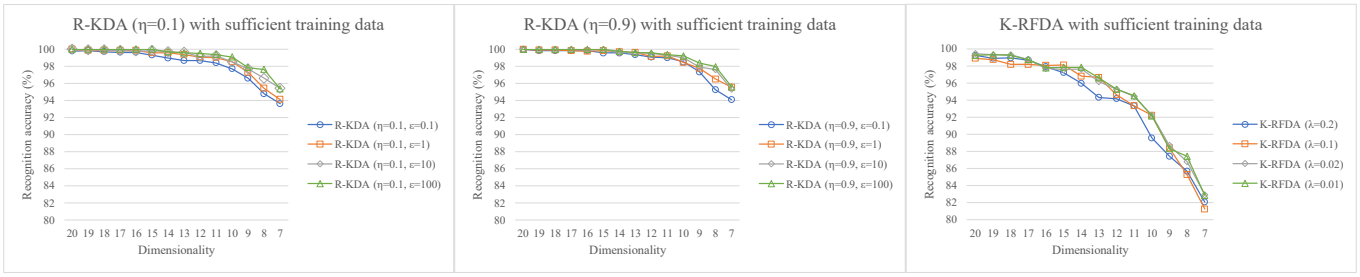

Fig. 3. Recognition accuracy with respect to dimensionality, using R-KDA and K-RFDA with linear kernel, when training data are sufficient.

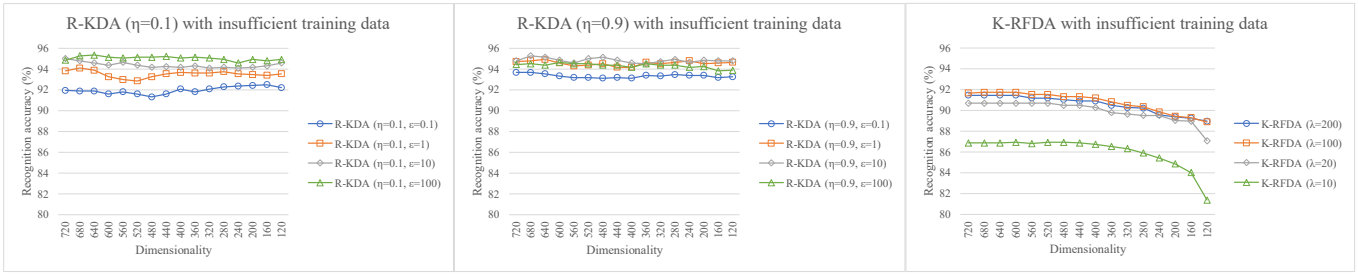

Fig. 4. Recognition accuracy with respect to dimensionality, using R-KDA and K-RFDA with linear kernel, when training data are insufficient.

[7] N. Li, W. Jiang, H. Meng, and Z. Li, "Clustering similar acoustic classes in the fishervoice framework," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 7726-7730.

[8] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: a tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74-99, 2015.

[9] K. Huang and S. Aviyente, "Sparse representation for signal classification," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2007, pp. 609-616.

[10] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2011, pp. 543-550.

[11] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2007, pp. 1-8.

[12] Y. Jiang and F. H. F. Leung, "The scalable version of probabilistic linear discriminant analysis and its potential as a classifier for audio signal classification," in *Proc. IEEE Int. Joint Conf. on Neural Networks (IJCNN)*, 2018.

[13] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. R. Muller, "Fisher discriminant analysis with kernels," in *Proc. IEEE Signal Processing Society Workshop*, 1999, pp. 41-48.

[14] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Computation*, vol. 12, no. 10, pp. 2385-2404, 2000.

[15] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Face recognition using kernel direct discriminant analysis algorithms," *IEEE Trans. On Neural Networks*, vol. 14, no. 1, pp. 117-126, 2003.

[16] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition," *Pattern Recognition Letters*, vol. 26, no. 2, pp. 181-191, 2005.

[17] T. V. Bandos, L. Bruzzone, and G. C. Valls, "Classification of hyperspectral images with regularized linear discriminant analysis," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 47, no. 3, pp. 862-873, 2009.

[18] J. Lu, K. N. Plataniotis, A. N. Venetsanopoulos, and J. Wang, "An efficient kernel discriminant analysis method," *Pattern Recognition*, vol. 38, no. 10, pp. 1788-1790, 2005.

[19] Y. Jiang and F. H. F. Leung, "Using double regularization to improve the effectiveness and robustness of Fisher discriminant analysis as a projection technique," in *Proc. IEEE Int. Joint Conf. on Neural Networks (IJCNN)*, 2018.

[20] X. S. Zhuang and D. Q. Dai, "Improved discriminate analysis for high-dimensional data and its application to face recognition," *Pattern Recognition*, vol. 40, no. 5, pp. 1570-1578, 2007.

[21] Z. Sun, J. Li, and C. Sun, "Kernel inverse Fisher discriminant analysis for face recognition," *Neurocomputing*, vol. 134, pp. 46-52, 2014.

[22] Y. Jiang and F. H. F. Leung, "Using regularized Fisher discriminant analysis to improve the performance of Gaussian supervector in session and device identification," in *Proc. IEEE Int. Joint Conf. on Neural Networks (IJCNN)*, 2017, pp. 705-712.

[23] D. G. Romero and C. Y. E. Wilson, "Automatic acquisition device identification from speech recordings," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 1806-1809.

[24] C. L. Kotropoulis, "Source phone identification using sketches of features," *IET Biometrics*, vol. 3, no. 2, pp. 75-83, 2014.

[25] L. Zou, Q. He, and J. Wu, "Source cell phone verification from speech recordings using sparse representation," *Digital Signal Processing*, vol. 62, pp. 125-136, 2017.

[26] Y. Jiang and F. H. F. Leung, "Mobile phone identification from speech recordings using weighted support vector machine," in *Proc. IEEE 42nd Annual Conf. on Industrial Electronics (IECON)*, 2016, pp. 963-968.

[27] C. M. Bishop, "Linear models for classification," in *Pattern Recognition and Machine Learning*, Springer, 2006, ch. 4, pp. 179-224.

[28] K. Fukunaga, "Feature Extraction and Linear Mapping for Classification," in *Introduction to Statistical Pattern Recognition*, 2nd Ed., San Diego, California: Academic Press, 1990, ch. 10, pp. 441-507.

[29] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308-311, 2006.

[30] D. Reynolds, "Gaussian mixture models," in *Encyclopedia of Biometrics*, 2015, pp. 827-832.

[31] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1-27, 2011.

[32] J. O. Garcia, J. G. Rodriguez, and V. M. Aguiar, "AHUMADA: a large speech corpus in Spanish for speaker characterization and identification," *Speech Communication*, vol. 31, no. 2, pp. 255-264, 2000.