I-VECTOR/PLDA SPEAKER RECOGNITION USING SUPPORT VECTORS WITH DISCRIMINANT ANALYSIS

Fahimeh Bahmaninezhad, John H.L. Hansen

Center for Robust Speech Systems (CRSS), University of Texas at Dallas, Richardson, TX 75080

{fahimeh.bahmaninezhad,john.hansen}@utdallas.edu

ABSTRACT

i-Vector feature representation with probabilistic linear discriminant analysis (PLDA) scoring in speaker recognition system has recently achieved effective permanence even on channel mismatch conditions. In general, experiments carried out using this combined strategy employ linear discriminant analysis (LDA) after the i-Vector extraction phase to suppress irrelevant directions, such as those introduced by noise or channel distortions. However, speaker-related and -non-related variability present in the data may prevent LDA from finding the best projection matrix. In this study, we exclusively use support vectors of each class to find the optimum linear transformation. Post-processing of the i-Vectors by discriminant analysis via support vectors (SVDA) and traditional LDA is evaluated on NIST2010 speaker recognition evaluation (SRE) core and extended core (coreext) conditions. In addition, truncated coreext test data is used to examine the performance of the system for both long and short duration test segments. Computed equal error rate (EER) and minimum detection cost function (minDCF) criteria confirm consistent improvement of SVDA over traditional LDA. The relative improvement in terms of EER and minDCF with SVDA are about 32% and 9%, respectively.

Index Terms— i-Vector/PLDA speaker recognition, discriminant analysis, support vectors.

1. INTRODUCTION

Over the past decade, the importance of intra-speaker (such as emotional and stress conditions, health, aging) and inter-speaker (noise, session, or channel affects) variability have made speaker recognition state-of-the-art gradually migrate from Gaussian mixture model (GMM)-universal background model (UBM) [1] to joint factor analysis (JFA) [2] and i-Vector [3] solutions with cosine distance scoring or support vector machine (SVM) classification [3]; and finally to i-Vector with probabilistic discriminant analysis (PLDA) scoring [4, 5]. Extracted i-Vectors usually are post-processed by linear discriminant analysis (LDA) in order to reduce the dimensionality based on the Fisher criterion [6] that in turn will help to compensate for channel variability.

The traditional LDA finds the transformation that minimizes the ratio of the within to between class scatters. LDA assumes speaker classes have a Gaussian distribution and share the same covariance matrix. Many variations of discriminant analysis have been proposed to partly relax the LDA assumptions. Kernel discriminant analysis or generalized discriminant analysis (GDA) [7, 8] finds a non-linear transformation, heterocedastic LDA (HLDA) [9] employs different covariance matrices for different classes, mixture discriminant analysis (MDA) [10] assumes the distribution of each class is a mixture of Gaussians.

In the i-Vector based system, the effectiveness of various discriminant analysis methods has been studied. [12] employed non-parametric or nearest neighbor discriminant analysis (NDA). The experimental results show that NDA outperforms LDA especially when data are multimodal [13]. In addition, [14] used source-normalized LDA (SN-LDA); and [15] employed weighted LDA (WLDA) and weighted SN-LDA which are shown to be more effective in especial conditions. Moreover, we studied GDA in [11] previously.

Here, we apply another variation of LDA named discriminant analysis via support vectors (SVDA) into the i-Vector/PLDA system. SVDA calculates the within and between class covariance matrices using only the support vectors. In contrast to LDA, SVDA captures the boundary of classes (which is important in classification), and performs well for small sample size problem which is present in SRE2010 task (i.e. when the dimensionality is greater than sample size). The idea of using support vectors with discriminant analysis has been previously introduced in [16] which made significant improvement over LDA. In this study, the effectiveness of SVDA in the i-Vector/PLDA system has been evaluated on NIST2010 speaker recognition evaluation (SRE) [17] task with the telephony condition for both long and short duration test segments. Compared to the above-mentioned methods, from the aspect of the number of hyper-parameters, training time, and also the equal error rate (EER) and minimum detection cost function (minDCF) criteria, SVDA is shown to be effective.

This project was funded by AFRL under contract FA8750-15-1-0205 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen.

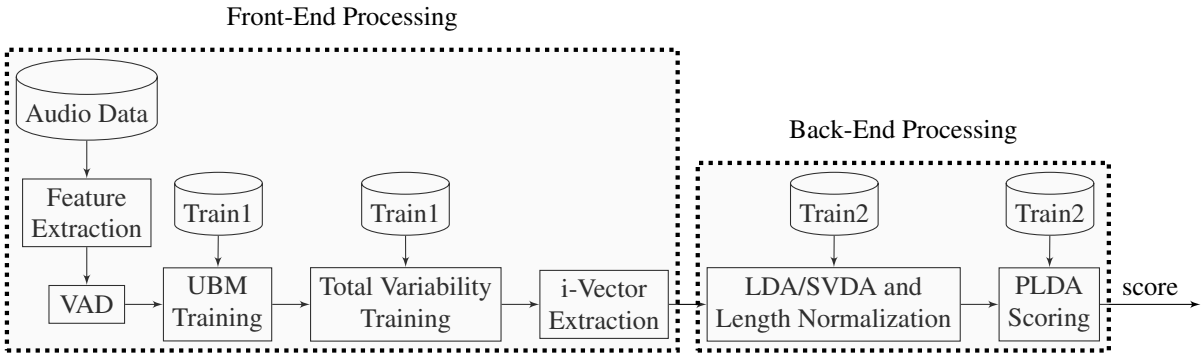

Fig. 1. *Overview of the i-Vector/PLDA system. Train data used for modeling UBM, total variability (TV) matrix, LDA (or SVDA) and PLDA, enrollment and test data used for evaluation of the system [11].*

In this paper, first an overview of our i-Vector/PLDA system is discussed in Sec. 2. Next, Sec. 3, 4 introduce LDA and SVDA, respectively. The experimental setup and results are presented in Sec.5, and finally conclusions and future work are summarized in Sec. 6.

2. I-VECTOR/PLDA SPEAKER RECOGNITION

The overall block-diagram of i-Vector/PLDA speaker recognition used in this study is depicted in Fig. 1. The speaker- and channel-dependent GMM supervector in the i-Vector configuration is factorized as [3],

$$M = m + Tw, \tag{1}$$

where m is the UBM speaker- and channel-independent supervector, T is the low rank total variability matrix (TV) that maps the high dimensional GMM supervector into lower dimensional i-Vector representation w.

Using training data, the UBM and TV matrix will be modeled by expectation maximization (EM) method. In the E-step, w is considered as a latent variable with normal prior distribution $N(0, I)$. Eventually, the i-Vectors will be estimated as the mean of posterior distribution of w, that is [3],

$$\hat{w}(u) = (I + T^T\Sigma^{-1}N(u)T)^{-1}T^T\Sigma^{-1}S(u), \tag{2}$$

where for utterance u, the terms $N(u)$ and $S(u)$ represent zeroth and centralized first order Baum-Welch statistics respectively, and Σ is the covariance matrix of UBM. Thereafter, i-Vectors are post-processed by applying LDA/SVDA and normalizing their lengths [18]. Finally, in the recognition phase given two i-vectors $\hat{w}_1$ and $\hat{w}_2$ we need to verify that these i-Vectors have been produced by the same speaker (target) or not (nontarget), that can be identified using the following log likelihood ratio,

$$\log -likelihood = log\frac{p(\hat{w}_1, \hat{w}_2|target)}{p(\hat{w}_1, \hat{w}_2|nontarget)} \tag{3}$$

3. LINEAR DISCRIMINANT ANALYSIS (LDA)

LDA is a widely-used statistical method for dimensionality reduction in classification and pattern recognition problems. It finds the exact optimal linear transformation when each class has a Gaussian distribution with a common covariance matrix. Traditional LDA defines the speaker class separation criterion in direction of A as,

$$\lambda = \frac{A^T S_b A}{A^T S_w A}, \tag{4}$$

where S_b and S_w represent between and within class covariance matrices. The projection matrix A that contains the k eigenvectors corresponding to the k largest eigenvalues of $S_w^{-1}S_b$ is the solution for LDA optimization problem.

For feature vectors x, the between and within class scatters are calculated by,

$$S_b = \sum_{c=1}^{C} n_c(\mu_c - \mu)(\mu_c - \mu)^T \tag{5}$$

$$S_w = \sum_{c=1}^{C}\sum_{k\in c}(x_k - \mu_c)(x_k - \mu_c)^T, \tag{6}$$

where C is the total number of speaker classes, n_c is the number of samples in class c, μ is the total mean of all samples, μ_c is the mean of samples in class c.

LDA imposes strict assumptions in finding the linear transformation. The following items express our main concerns regarding LDA that will be revised by the proposed alternative discriminant analysis approach. First, assuming Gaussian distribution for speaker classes is simplifying. Second, the imbalanced classes (one class has many samples while the other may contain a few ones) always was a problem in pattern recognition area; here, we will attempt to balance the samples in each class as far as possible. Third, small sample size problem that is present in NIST SRE2010 task cause difficulties for LDA. Forth, assuming discriminatory

Table 1. *statistics of data used for training the models and evaluating the system. Trials, enrollment and data used for training LDA/SVDA/PLDA are restricted to male speakers.*

Enrollment/Test	UBM-TV		LDA/SVDA/PLDA		Enrollment	Trials	
	Spkrs	Segments	Spkrs	Segments	Spkrs	Target	nonTarget
Core/Core	5756	57273	1115	13605	2426	353	13707
Coreext/3,5,10,20,40s,full	5756	57273	1115	13605	5237	3465	175873

information is all included in the centroid of classes is not applicable to real-world problems; covariance or boundary structure of classes should not be disregarded completely.

Therefore, for accounting the above-mentioned limitations, discriminant analysis via support vectors (SVDA) [16] has been adopted into the i-Vector/PLDA structure. SVDA only uses support vectors to calculate the within and between covariance matrices. In addition, with SVDA the degree of generalization could be controlled in solving SVM problem which has a valuable advantage.

4. DISCRIMINANT ANALYSIS VIA SUPPORT VECTORS (SVDA)

The class separation measure for SVDA is similar to the LDA; however, only distinct support vectors will be used to calculate the within class and between class covariance matrices. More specifically, if we define $w_{c_1c_2} = \sum_{i=1}^{l} y_i\alpha_i x_i$ as the optimal direction to classify two classes c_1 and c_2 by a linear SVM (y_i represents target value (+1 for first class and -1 for second class) of learning pattern x_i and α_i is its coefficient), then the between class covariance matrix will be updated as,

$$V_b = \sum_{1\leq c_1\leq c_2\leq C} w_{c_1c_2} w_{c_1c_2}^T. \tag{7}$$

Also, let $\hat{X} = [\hat{x}_1, \hat{x}_2, ..., \hat{x}_{\hat{N}}]$ be all the support vectors and $\hat{N}$ represents their number. Then, the within class covariance matrix will be formulated as,

$$V_w = \sum_{c=1}^{C} \sum_{i\in\hat{I}_c} (\hat{x}_i - \hat{\mu}_c)(\hat{x}_i - \hat{\mu}_c)^T \tag{8}$$

where $\hat{I}_c$ includes the index of support vectors in class c, and $\hat{\mu}_c$ denotes the mean of them. Finally, similar to LDA, the optimum transformation $\hat{A}$ will contain the k eigenvectors corresponding to the k largest eigenvalues of $V_w^{-1}V_b$

From the aspect of classification, [16] shows SVM performs better than LDA. For multi-class problems, Fisher criterion in LDA finds the subspace that gives well-separated classes more importance than those are closer. Therefore, the classes that are already distinct will move away further from each other; however, the closer classes will not be treated the same. In contrast, SVM focuses more on hardly-separable classes. From this viewpoint, we expect that the transformation find by the SVDA will work better for closer classes.

Moreover, the within class and between class covariance matrices calculated by SVDA only uses support vectors instead of using all the training samples. Obviously, SVDA finds the discriminatory directions using the boundary structure of classes; and also SVM is a well-known method for small sample size problem [16]. On the other hand, while solving SVM problem we can adjust the tolerance of classification error; therefore, generalization can be controlled conveniently in SVDA rather than LDA.

5. EXPERIMENTS

5.1. Experimental setup

The extracted feature vectors contain 19 Mel-frequency features as well as the frame energy appended with delta and delta-delta coefficients. The window length and shift size are 25-ms and 10-ms, respectively. In addition, 3-s sliding window cepstral mean normalization has beed applied on feature vectors. Non-speech frames are also discarded using energy-based voice activity detection (VAD).

2048-mixture full covariance UBM and total variability matrix have been trained using both male and female data collected from SRE2004, 2005, 2006, 2008 and Switchboard II phase 2,3 and Switchboard Cellular Part1 and Part2. Next, 600-dimensional i-Vectors have been extracted. The dimension of i-Vectors then reduced to 400 using LDA/SVDA technique. Data used for training LDA, SVDA and PLDA has been restricted to the male speakers (for the sake of tractability) from SRE2004, SRE2005, SRE2006, and 2008. To evaluate the system, we used male trials of core and extended core conditions of SRE2010. All the experiments are carried out on telephony condition (condition 5) of NIST SRE2010.

The enrollment/test condition combinations used in the experiments and the statistics of training and enrollment data, as well as trials are provided in Table 1. In addition, to evaluate the performance of the system on short duration test segments, after applying VDA, the first 3, 5, 10, 20 and 40s of extended core test data have been extracted. For training SVM the publically available LIBSVM [19] toolkit has been used.

5.2. Experimental results

This subsection provides the experimental results comparing SVDA and LDA. We used equal error rate (EER) and mini-

Table 2. *EER/minDCF results comparing LDA and SVDA. The dimension of i-Vectors is reduced from 600 to 400.*

Enrollment/Test	**LDA**	**SVDA**		
		traditional 1-vs-1	weighted 1-vs-1	1-vs-rest
Core/Core	1.66 / .037	**1.13** / .0399	1.25 / **.0364**	1.42 / .0368
Coreext/Coreext	1.5 / .0297	1.35 / .0308	**1.3** / **.0287**	1.39 / .029
Coreext/Coreext3sec	14.5 / .0984	14.22 / .0974	14.23 / **.0974**	**14.2** / .0975
Coreext/Coreext5sec	9.71 / .0924	9.64 / .0915	**9.55** / **.0909**	9.81 / .092
Coreext/Coreext10sec	5.61 / .0759	**5.58** / .0749	5.60 / **.0737**	5.72 / .076
Coreext/Coreext20sec	3.17 / .0585	**3.12** / .0574	3.17 / **.0573**	3.35 / .0593
Coreext/Coreext40sec	2.48 / .0448	2.4 / .0423	**2.37** / **.0407**	2.42 / .0411

mum detection cost function (minDCF) defined as,

$$C_{Det} = C_{Miss} \times P_{Miss|Target} \times P_{Target} + C_{FalseAlarm} \times P_{FalseAlarm|NonTarget} \times (1 - P_{Target})$$

to evaluate our system. Based on the SRE2010 task $C_{Miss} = C_{FalseAlarm} = 1$ and $P_{Target} = 1/1000$.

Table 2 summarizes the performance of i-Vector/PLDA speaker recognition comparing SVDA against LDA. To calculate the between and within class covariance matrices using SVDA three strategies have been considered: traditional one-versus-one, weighted one-versus-one and one-versus-rest. In one-versus-one strategy, the SVM will be applied to just two classes, therefore we need to model $C(C-1)/2$ SVMs; in contrast to the one-versus-rest approach that each class will be classified against all data from all other speakers (need to train C SVM classifiers). As stated before, C represents the number of speaker classes. It is worth mentioning that the one-versus-one strategy is more appropriate for imbalance problem. The weighted one-versus-one has been designed to punish the classes that does not have enough samples to define their structure (or may have noisy or random samples). In other words, some of the classes do not have well-defined structure and when we apply SVM, all the samples in the class will be recognized as the support vectors. Therefore, by giving these types of classes smaller weight for their contribution in just calculating Vb in Eq. 7, the SVM classifier will be forced to emphasize more on well-defined classes.

The results prove that SVDA consistently improves LDA in terms of both EER and minDCF. For SVDA, weighted one-versus-one strategy is approximately working better than the traditional one (and both are better than one-versus-rest); these results meet our expectation that: first, the imbalance problem present in data (some classes have less than 10 samples and some around 94) will be partly solved with SVDA. More specifically, EER and minDCF is considerably improved by 32% and 5.6% (respectively) with traditional one-versus-one strategy. In addition, with weighted one-versus-one approach and punishing those classes that are not well-distinguishable with SVM (and probably contain noisy and error-full data), the relative improvement attained is 25% and 9% for EER and minDCF, respectively. Second, the capability of SVM for small sample size problem has been confirmed (i-Vectors are 600 dimensional but there is not any class with more than 100 samples in training set).

Table 3. *Speaker recognition results comparing LDA and SVDA in terms of EER/minDCF without dimension reduction.*

Enrollment/Test	**LDA**	**SVDA**	
		traditional 1-vs-1	weighted 1-vs-1
Core/Core	1.58 / .039	**1.45** / **.038**	1.46 / .04
Coreext/Coreext	1.46 / .0302	1.37 / **.0301**	**1.36** / .0302

Table 3 reports results comparing LDA and SVDA without dimension reduction for core and extended core conditions. In terms of EER, SVDA outperforms LDA significantly; however, in terms of minDCF there is just marginal improvement. In summary, with regard to the number of hyper-parameters, computation time, and the performance (approximate 32% relative improvement in EER) SVDA works really well.

6. CONCLUSION AND FUTURE WORK

In this paper, the effectiveness of SVDA in the i-Vector/PLDA speaker recognition has been studied. The EER and minDCF scores achieved from the experiments carried out on NIST SRE2010 task prove that SVDA consistently works better than LDA. In contrast to LDA that limits the discriminatory information to the centroid of classes, SVDA captures the boundary structure of them. In addition, small sample size problem is well treated with SVDA.

Although SVDA has a considerable improvement for longer duration test segments, the decrease in EERs and minDCF is less for short duration test segments. In continue, we would like to introduce the uncertainty of i-Vectors into the SVDA to surpass the impact of short duration test data. In addition, the application of kernel SVM instead of traditional linear SVM in SVDA will be studied later. Moreover, we will compare SVDA and LDA for other conditions of NIST SRE2010 for male, female and pooled speakers.

7. REFERENCES

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.

[2] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

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

[4] 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.

[5] G. Liu and J. H.L. Hansen, "An investigation into back-end advancements for speaker recognition in multi-session and noisy enrollment scenarios," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1978–1992, 2014.

[6] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[7] B. Scholkopft and K.R. Mullert, "Fisher discriminant analysis with kernels," *Neural Networks for Signal Processing IX*, vol. 1, no. 1, pp. 1, 1999.

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

[9] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition," *Speech Communication*, vol. 26, no. 4, pp. 283–297, 1998.

[10] T. Hastie and R. Tibshirani, "Discriminant analysis by gaussian mixtures," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 155–176, 1996.

[11] F. Bahmaninezhad and J. H.L. Hansen, "Generalized discriminant analysis (GDA) for improved i-vector based speaker recognition," in *Proceedings of INTERSPEECH*, 2016.

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

[13] S.O. Sadjadi, J. Pelecanos, and S. Ganapathy, "The ibm speaker recognition system: Recent advances and error analysis," in *Proceedings of INTERSPEECH*, 2016.

[14] M. Mclaren and D. Van Leeuwen, "Source-normalized LDA for robust speaker recognition using i-vectors from multiple speech sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 755–766, 2012.

[15] A. Kanagasundaram, D. Dean, R. Vogt, M. McLaren, S. Sridharan, and M. Mason, "Weighted LDA techniques for i-vector based speaker verification," in *Proceedings of ICASSP*. IEEE, 2012, pp. 4781–4784.

[16] S. Gu, Y. Tan, and X. He, "Discriminant analysis via support vectors," *Neurocomputing*, vol. 73, no. 10, pp. 1669–1675, 2010.

[17] "The NIST year 2010 speaker recognition evaluation plan," http://www.nist.gov/itl/iad/mig/upload/NIST_SRE10_evalplan-r6.pdf, 2010.

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

[19] C.C. Chang and C.J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27, 2011.