Kernel partial least squares for speaker recognition

Balaji Vasan Srinivasan[1], *Daniel Garcia-Romero*[2], *Dmitry N. Zotkin*[1], *Ramani Duraiswami*[1]

[1]Perceptual Interfaces and Reality Laboratory, Department of Computer Science
[2]Speech Communication Laboratory, Department of Electrical and Computer Engineering
[1,2]University of Maryland, College Park, MD, USA

balajiv@umiacs.umd.edu, dgromero@umd.edu, dz@umiacs.umd.edu, ramani@umiacs.umd.edu

Abstract

I-vectors are a concise representation of speaker characteristics. Recent advances in speaker recognition have utilized their ability to capture speaker and channel variability to develop efficient recognition engines. Inter-speaker relationships in the i-vector space are non-linear. Accomplishing effective speaker recognition requires a good modeling of these non-linearities and can be cast as a machine learning problem. In this paper, we propose a kernel partial least squares (kernel PLS, or KPLS) framework for modeling speakers in the i-vectors space. The resulting recognition system is tested across several conditions of the NIST SRE 2010 extended core data set and compared against state-of-the-art systems: Joint Factor Analysis (JFA), Probabilistic Linear Discriminant Analysis (PLDA), and Cosine Distance Scoring (CDS) classifiers. Improvements are shown.

Index Terms: kernel partial least squares, speaker recognition, i-vectors.

1. Introduction

Speaker recognition [1] deals with the task of verifying a speaker's claimed identity from a sample utterance based on a number of training utterances for which the speaker is known. Apart from carrying the speaker-specific characteristics, the speech data also encapsulates phonemic content, channel variability, and session variability. It is also often subject to noise and reverberation, making the problem of speaker recognition challenging. Over the past decade, the field has made substantial progress in addressing these issues. The commonly used feature space is the set of mel-cepstral coefficients along with their deltas and double-deltas.

State-of-the-art speaker recognition systems use a Gaussian mixture model (GMM) to represent each speaker. To account for limited training data available, the problem is cast into a framework in which differences from a universal background model (UBM) are used to adapt speaker-specific GMMs [2]. Recently, several approaches have been tested to make the GMM-based speaker recognition robust to session and channel variabilities, including JFA technique [3] [4] and the i-vectors framework [5]. The i-vectors are smaller in dimension compared to the GMM-supervectors and thus provide an abridged representation of the utterance.

The key problem for the i-vector representation is to develop learning techniques that distinguish target and non-target trials in the i-vector space. Generative PLDA models [6], discriminative SVMs [7], and CDS classifiers [8] have been studied for speaker recognition using i-vectors. Earlier, we have introduced a linear PLS framework for supervector-based speaker discrimination [9]; however, the best performance with i-vectors is achieved with non-linear discrimination functions such as PLDA [6] or CDS [8]. In the current work, we explore a kernelized version of the PLS for speaker recognition.

The paper is organized as follows. In Section 2, i-vectors are introduced and their extraction is detailed. KPLS framework is introduced and adapted to the speaker recognition problem in Section 3. Section 4 discusses the results of the KPLS evaluation on NIST SRE 2010 data and compares them against several state-of-the-art systems. Finally, section 5 concludes the paper.

2. i-vectors

The key idea in the JFA technique [3] [4] is to find two subspaces that best capture the speaker and the channel variabilities in the feature space. JFA has been quite successful in terms of performance; however, Dehak et al. [5] observed that the channel subspace still contains some information about the speaker and vice-versa. Therefore, he proposed using a combined subspace to capture both variabilities and called it the *total variability space*. In this formulation, a speaker- and channel-dependent supervector s is modeled as

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

where m is a speaker- and channel-independent supervector (usually the UBM supervector), T is a low-rank matrix representing the basis of the total variability space, and w is a normal-distributed vector representing the coordinates of the speaker in that space. The vector w is called the *identity vector*, or *i-vector*. Typically, the number of dimensions of w is three orders of magnitude smaller than that of the supervectors (e.g. 400 vs 10^5).

3. Kernel Partial least squares (KPLS)

A brief description of kernel PLS is provided here; more detailed analysis is available in [10]. Denote a d-dimensional feature by x (the i-vector from a single speech utterance in our case) and the corresponding speaker label by y. KPLS considers the mapping of the features x to a higher dimensional space, given by $\Phi : R^d \Rightarrow R^{\bar{d}}$. Assume momentarily that such a Φ is defined and known. Let the total number of speakers be N and denote the $N \times \bar{d}$ matrix of feature vectors by $\Phi(X)$ and the $N \times 1$ vector of labels ($+1$ for speaker and -1 for imposter) by Y. Given the variable pairs $\{\Phi(x_i), y_i\}, i = 1, \ldots, N$ ($\Phi(x_i) \in R^{\bar{d}}, y_i \in R$), KPLS aims at modeling the relationship between x and y using projection into latent spaces by decomposing $\Phi(X)$ and Y as

$$\Phi(X) = TP^T + E, \tag{2}$$

$$Y = UQ^T + F, \tag{3}$$

where T and U ($N \times p$) are the latent vectors, P ($\bar{d} \times p$) and Q ($1 \times p$) are the loading vectors, and E ($N \times \bar{d}$) and F ($N \times 1$) are residual matrices.

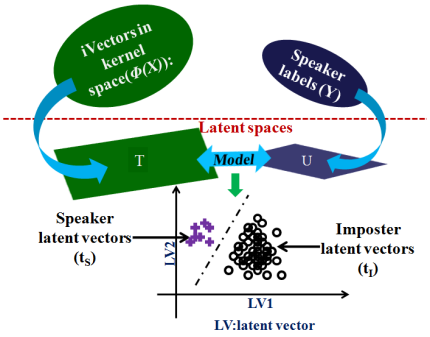

Figure 1: (color) *Non-linear mapping and the corresponding subspaces learnt via Kernel Partial Least Squares (KPLS).*

KPLS is usually solved via the *nonlinear iterative partial least squares (NIPALS) algorithm* [11], which constructs a set of weight vectors $W = \{w_1, w_2, \ldots, w_p\}$ such that

$$\max[\text{cov}(t_i, u_i)]^2 = \max_{|w_i|=1} [\text{cov}(\Phi(X)w_i, Y)]^2, \quad (4)$$

where t_i and u_i are the i^{th} columns of T and U respectively and $\text{cov}(t_i, u_i)$ indicates the sample covariance between latent vectors t_i and u_i. Maximizing the covariance in the latent vector space is equivalent to maximizing discrimination in the same space; in other words, for a particular speaker, KPLS learns a subspace in which the speaker latent vectors t_S are well separated from the imposter latent vector t_I as illustrated in Figure 1. Thus, KPLS learns a unique latent space for each speaker.

It has been shown [11] that the NIPALS algorithm is equivalent to iteratively finding the dominant eigenvectors of the problem

$$[\Phi(X)^T yy^T \Phi(X)]w_i = \lambda w. \quad (5)$$

The $\Phi(X)$-scores t_i are then obtained as $t_i = \Phi(X)w_i$. Rosipal et al. [10] modify this eigenproblem as

$$[\Phi(X)\Phi(X)^T yy^T]t = \gamma t. \quad (6)$$

Using the "kernel" trick [12], $\Phi(X)\Phi(X)^T$ can be defined as a kernel matrix K leading to the final eigenproblem

$$[Kyy^T]t = \gamma t. \quad (7)$$

A key advantage of this kernelization is that an explicit definition of the mapping function Φ is not required and it suffices to define a kernel function between pairs of feature vectors. This modified version of the NIPALS algorithm for KPLS has been detailed in [10].

After extraction of latent vectors t_i and u_i, the kernel matrix K is deflated by removing any information captured by t_i and u_i from K:

$$K \leftarrow (I_n - tt^T)K(I_n - tt^T). \quad (8)$$

The process is repeated till a sufficient number (determined via standard cross-validation) of latent vectors is obtained.

KPLS Regression: We use the KPLS in the regression framework [9] for speaker scoring. Substituting the w from Eq. (4) in Eq. (2), we get

$$\Phi(X)W = TP^T W + E \Rightarrow T = \Phi(X)W(P^T W)^{-1}. \quad (9)$$

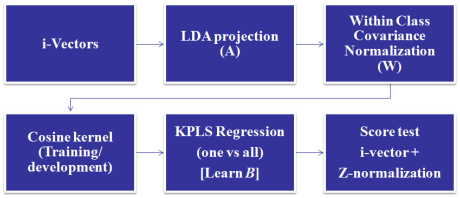

Figure 2: (color) *Kernel Partial Least Squares (KPLS) schematic for speaker recognition.*

Now, U can be written in terms of T as $U = TD + H$, where D is a diagonal matrix and H is the residual [11]. Eq. (3) now becomes

$$Y = TDQ^T + HQ^T + F = \Phi(X)W(P^T W)^{-1}DQ^T + \bar{F}.$$

Using $P = \Phi(X)^T T$ and $W = \Phi(X)^T U$ from [10] in the above equation, we generate the score for a test i-vector:

$$\text{score}_{\text{KPLS}} = \Phi(X_t)\Phi(X)^T U(T^T \Phi(X)\Phi(X)^T U)^{-1} DQ^T.$$

This leads to the KPLS regression:

$$\text{score}_{\text{KPLS}} = K_t B; \quad B = U(T^T KU)^{-1} DQ^T, \quad (10)$$

where B is the set of PLS regression coefficients, K_t is the kernel matrix between training data and testing data i-vectors, and K is the kernel matrix between the training data i-vectors only. This regression framework provides a direct method to compute the matching score for seamless speaker discrimination, eliminating the need for a separate classifier. Note that the regression coefficients are unique to each speaker.

Choice of Kernel Matrix: The key choice in any kernel method is the kernel function. We use the cosine kernel similar to the one used in [5]. A Linear Discriminant Analysis (LDA) based subspace is first learnt on the i-vectors, and training and testing i-vectors are projected into this space. Let A denote the LDA projection matrix. Then, a within-class covariance normalization matrix W is learnt on LDA-projected space. Finally, given two i-vectors w_1 and w_2, the kernel function $k(w_1, w_2)$ is defined as

$$\frac{(Aw_1)^T W^{-1}(Aw_2)}{\sqrt{(Aw_1)^T W^{-1}(Aw_1)}\sqrt{(Aw_2)^T W^{-1}(Aw_2)}}. \quad (11)$$

Present approach: In our experiments, we use separate development sets for male and female speakers. All speakers from the development set constitute the negative examples in the KPLS framework (one-vs-all approach). For each target speaker, the corresponding i-vector is assigned a label of $+1$; all negative examples are assigned a label of -1; the KPLS is trained; and the speaker-specific regression coefficients B are learnt according to Eq. (10). The final score is obtained using Eq. (10) with the kernel built using testing data i-vectors against the development data and training data i-vectors. *Note that the final score in this case is a linear combination of the cosine scores between the testing data i-vector and the combination of target data and development data i-vectors and that this linear combination (B) is unique to each speaker.* These steps are summarized in Figure 2.

4. Experiments

We performed experimental evaluation of the proposed method on the *extended core set* of the NIST SRE 2010 evaluation

data set, which is grouped into 9 trial conditions[1]. Our development data consisted of NIST SRE 2004, 2005, 2006, and 2008 data; Switchboard data set, phases 2 and 3; Switchboard-Cellular data set, parts 1 and 2; and Fisher data set (total of 17319 male and 22256 female utterances). A gender-dependent 2048-center UBM with diagonal covariance was trained using the standard 38 MFCC features, and the gender-dependent total variability matrix T of dimension 400 was also learnt. All available development data were used as negative examples in the KPLS framework. KPLS output scores were Z-normalized [1].

4.1. Systems compared

The proposed KPLS based speaker recognition was compared against several state-of-the-art systems, specifically JFA [3] [4], PLDA [6], and CDS [14]. We describe these systems briefly here.

Joint Factor Analysis: JFA provides an explicit mechanism to model the undesired variability in the speech signal. It decomposes the speaker supervector as

$$s = m + Ux + Vy + Dz, \tag{12}$$

where $\{m, U, V, D\}$ are the hyper-parameters of the JFA model, which are estimated via Expectation Maximization (EM). In our experiments, we use the JFA as described in [15]. The U and V matrices are learnt with 300 and 100 dimensions respectively. Defining $\Phi_{JFA} = [UVD]$ and $\beta = [xyz]^T$, we obtain compensated training and testing supervectors as $\eta_{train} = \Phi_{JFA}\beta - Ux_{train}$ and $\eta_{test} = \Phi_{JFA}\beta - Ux_{test}$. The final score is given by

$$\text{score}_{\text{JFA}} = \frac{1}{N}\eta_{train} W_{test} \eta_{test}, \tag{13}$$

where W_{test} is defined in [15] and N is the number of frames in the test segment. The JFA scores are then ZT-normalized [1].

Probabilistic Linear Discriminant Analysis: PLDA facilitates the comparison of i-vectors in a verification trial. A special *two-covariance* PLDA model is generally used for speaker recognition in the i-vector space. The speaker variability and session variability are modeled using across-class and within-class covariance matrices (Σ_{ac} and Σ_{wc} respectively) in the PLDA setup. A latent vector y representing the speakers is assumed to be normally distributed $\mathcal{N}(y; \mu, \Sigma_{ac})$, and for a given speaker represented by this latent vector, the i-vector distribution is assumed to be $p(w|y) = \mathcal{N}(w; y, \Sigma_{wc})$.

Given two i-vectors w_1 and w_2, PLDA defines two hypotheses $\mathcal{H}_s$ and $\mathcal{H}_d$ indicating that they belong to the same speaker or to different speakers respectively. The score is then defined as $\log \frac{p(w_1,w_2|\mathcal{H}_s)}{p(w_1,w_2|\mathcal{H}_d)}$. Marginalization of the two distributions with respect to the latent vectors leads to

$$\text{score}_{\text{PLDA}} = \log \frac{\mathcal{N}\left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}; \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \Sigma_{tot}\Sigma_{ac} \\ \Sigma_{ac}\Sigma_{tot} \end{bmatrix}\right)}{\mathcal{N}\left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}; \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \Sigma_{tot}0 \\ 0\Sigma_{tot} \end{bmatrix}\right)}$$

In our experiments, we found that using a Σ_{ac} of rank 200 along with a full-rank (rank 400) matrix Σ_{wc} produced the best results. The scores were S-normalized only for those conditions that involve telephone speech (all except C1, C2 and C4, where S-norm was found to be detrimental for both EER and DCF). The S-norm is defined in [6] and can be interpreted loosely as a symmetric version of Z-norm [1].

Cosine Distance Scoring: The CDS classifier has been used by Dehak et al. [5] and Senoussaoui et al. [8]. Improved performance has been reported over the corresponding SVM-based approach. The CDS classifier defines the score for the trial as a cosine similarity function between two i-vectors after projecting them to an LDA subspace (learnt on the development data) to remove the session variability. If w_1 and w_2 are the training data and the testing data i-vectors and A is the LDA projection matrix, the CDS score is given by

$$\text{score}_{\text{CDS}} = \frac{(Aw_1)^T(Aw_2)}{\sqrt{(Aw_1)^T(Aw_1)}\sqrt{(Aw_2)^T(Aw_2)}}. \tag{14}$$

In out experiments, the CDS scores were Z-normalized [1].

4.2. Results

We compared the performance of the KPLS based speaker recognition against the JFA, PLDA, and CDS systems. The corresponding equal error rate (EER) and detection cost function (DCF) values across each condition are tabulated in Table 1 and are shown graphically in Figure 3. The DCF is defined as for NIST SRE 2010 "core" and "8conv/core" conditions.

The PLDA and JFA systems belong to the class of generative methods for speaker recognition. Between them, the PLDA is better in most of the conditions. In contrast, KPLS and CDS belong to the class of discriminative methods, and KPLS outperforms CDS in most of the conditions (in terms of EER). It is well known that discriminative methods perform better when several training utterances are available per speaker. In our evaluation data set, only a single training utterance per speaker is provided; despite that, KPLS performance was better than PLDA performance in three of nine testing conditions and is comparable in two of the remaining ones.

Given than the PLDA and KPLS perform consistently better than other systems, we explored the possibility of score fusion between these approaches. We computed the fused score by combining the output scores with linear weights, which were trained using a small subset of development data. The results are also shown in Table 1 and Figure 3. The fused scores yield the best EERs in all conditions, suggesting the complementary nature of PLDA and KPLS in capturing speaker characteristics. More sophisticated fusion strategy is a subject of further research.

5. Conclusions

In this paper, we have proposed a kernel partial least squares framework for speaker recognition in the i-vector space. The proposed framework was compared against several state-of-the-art systems on the NIST SRE 2010 extended core data set. The KPLS system outperforms the state-of-the-art in several conditions and provides complementary information, resulting in further improved performance using simple linear score combination as a score fusion technique.

6. Acknowledgement

This research was partially funded by the Office of the Director of National Intelligence (ODNI) and Intelligence Advanced Research Projects Activity (IARPA) through the Army Research

[1] www.itl.nist.gov/iad/mig/tests/sre/2010/

	Number of trials		JFA		PLDA		CDS		KPLS		KPLS + PLDA	
	TGT	NTGT	EER	DCF	EER	DCF	EER	DCF	EER	DCF	EER	DCF
C1	4,304	795,995	2.67	0.502	**1.77**	**0.247**	2.27	0.328	1.80	0.289	**1.66**	**0.238**
C2	15,084	2,789,534	4.34	0.573	**3.09**	**0.480**	4.06	0.535	3.41	0.485	**2.87**	**0.451**
C3	3,989	637,850	4.06	0.575	**3.00**	**0.551**	3.71	0.562	4.41	0.591	**2.98**	**0.546**
C4	3,637	756,775	3.65	0.593	2.85	0.412	3.48	0.478	**2.82**	**0.374**	**2.68**	0.377
C5	7,169	408,950	3.55	0.551	**2.59**	**0.438**	4.18	0.545	4.10	0.551	**2.59**	**0.438**
C6	4,137	461,438	7.04	0.889	**5.43**	**0.789**	6.36	0.815	6.77	0.854	**5.27**	**0.789**
C7	359	82,551	8.16	0.944	8.06	0.805	8.46	**0.767**	**7.34**	0.838	**7.26**	0.808
C8	3,821	404,848	3.11	0.495	**2.51**	0.516	2.99	0.527	2.85	**0.495**	**2.27**	0.496
C9	290	70,500	2.13	0.482	2.17	0.375	1.96	**0.290**	**1.82**	0.326	**1.60**	0.333

Table 1: *Equal error rate (EER) and detection cost function (DCF) values obtained using Joint Factor Analysis, Probabilistic Linear Discriminant Analysis, Cosine Discriminative Scoring, and Kernel Partial Least Squares for the NIST SRE 2010 extended core data set.*

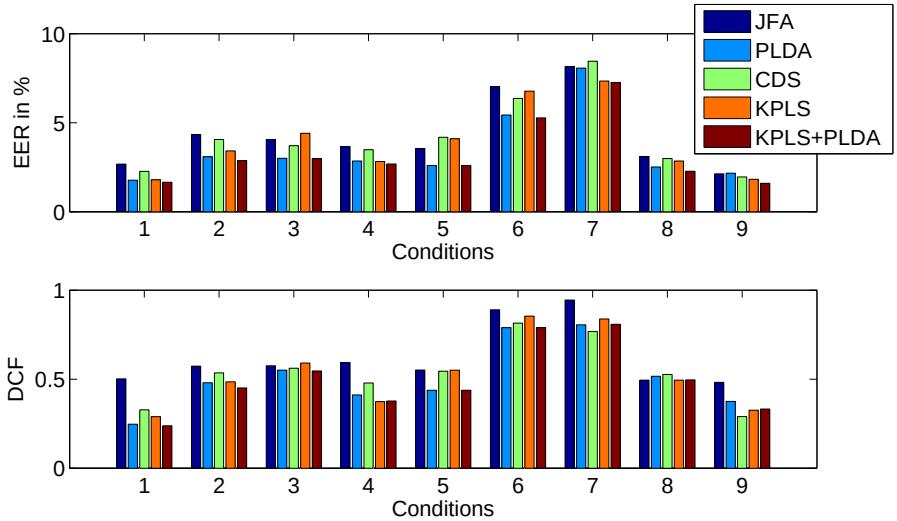

Figure 3: (color) *Performance of JFA, PLDA, CDS, and KPLS on the NIST SRE 2010 extended core data set.*

Laboratory (ARL). All statements of fact, opinion, or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of ODNI, the IARPA, or the U. S. Government.

7. References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, 52:12–40, 2010.

[2] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, 10:19–41, 2000.

[3] L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. Cernocky, "Analysis of feature extraction and channel compensation in a GMM speaker recognition system," *IEEE Transactions on Audio, Speech, and Language Processing*, 15:1979–1986, 2007.

[4] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, 16:980–988, 2008.

[5] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," *Proc. INTERSPEECH 2009*, September 2009, 1559–1562.

[6] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Proc. Odyssey Speaker and Language Recognition Workshop*, June 2010.

[7] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition", *Computer Speech and Language*, 20:210–229, 2006.

[8] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, "An i-vector extractor suitable for speaker recognition with both microphone and telephone speech," *Proc. Odyssey Speaker and Language Recognition Workshop*, June 2010.

[9] B. Srinivasan, D. N. Zotkin, and R. Duraiswami, "A partial least squares framework for speaker recognition," *Proc. IEEE ICASSP 2011*, May 2011, in print.

[10] R. Rosipal and L. Trejo, "Kernel partial least squares regression in reproducing kernel Hilbert space," *Journal Machine Learning Research*, 2:97–123, 2002.

[11] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," *Lecture Notes in Computer Science: Subspace, Latent Structure, and Feature Selection Techniques*, Springer, New York, 2006.

[12] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[13] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," *Proc. INTERSPEECH 2006 – ICSLP*, September 2006, 1471–1474.

[14] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, "Cosine similarity scoring without score normalization techniques," *Proc. Odyssey Speaker and Language Recognition Workshop*, June 2010.

[15] D. Garcia-Romero and C. Espy-Wilson, "Joint factor analysis for speaker recognition reinterpreted as signal coding using overcomplete dictionaries," *Proc. Odyssey Speaker and Language Recognition Workshop*, June 2010.