A hybrid Factor Analysis and Probabilistic PCA-based system for Dictionary Learning and Encoding for Robust Speaker Recognition

Srikanth Madikeri

Department of Computer Science and Engineering
Indian Institute of Technology Madras
mrsri@cse.iitm.ac.in

Abstract

Probabilistic Principal Component Analysis (PPCA) based low dimensional representation of speech utterances is found to be useful for speaker recognition. Although, performance of the FA (Factor Analysis)-based total variability space model is found to be superior, hyperparameter estimation procedure in PPCA is computationally efficient. In this work, recent insight on the FA-based approach as a combination of dictionary learning and encoding is explored to use its encoding procedure in the PPCA framework. With the use of an alternate encoding technique on dictionaries learnt using PPCA, performance of state-of-the-art FA-based i-vector approach is matched by using the proposed procedure. A speed up of 4x is obtained while estimating the hyperparameter at the cost of 0.51% deterioration in performance in terms of the Equal Error Rate (EER) in the worst case. Compared to the conventional PPCA model, absolute improvements of 2.1% and 2.8% are observed on two telephone conditions of NIST 2008 SRE database. Using Canonical Correlational Analysis, it is shown that the i-vectors extracted from the conventional FA model and the proposed approach are highly correlated.

1. Introduction

Low dimensional representation of speech utterances based on factor analysis has become a part of state-of-the-art speaker recognition systems [1]. A variable-length speech pattern is projected onto a low-dimensional linear subspace. The basis vectors of this subspace are estimated from the EM algorithm given in [2] . This low dimensional representation of a speech utterance is termed as the i-vector (identity vector). The conventional i-Vector estimates are obtained from the first and second order statistics of the feature vectors with respect to a UBM (Universal Background Model) [3]. Further, for robust speaker recognition, channel effects are removed from the estimated i-vectors using techniques such as WCCN (Within Class Covariance Normalization) or LDA (Linear Discriminant Analysis) or PLDA (Probabilistic Linear Discriminant Analaysis).

A recent insight into factor analysis (FA)-based approaches for speaker recognition allows us to consider the total variability matrix as an overcomplete dictionary [4]. The dictionary refers to the basis vectors representing the low dimensional linear subspace. Subsequently, estimating the i-vector from a speech utterance is considered as encoding. That is, the i-vector is the code obtained for the speech utterance using the dictionary. Such a perspective suggests possibilities of looking at better ways to perform both stages of the process - dictionary building (basis vectors/hyperparameter estimation) and encoding (i-vector estimation). In most subspace estimation algorithms, the dictionary building and encoding are closely related. The overall system can be viewed as shown in Figure 1. In [5], it is argued and shown that the encoding procedure used while building the codebooks and encoding evaluation data (training and testing) need not be the same. However, in their case, only sparse encoding schemes were tested. It is observed that similar or better performance can be achieved by encoding data using a different scheme.

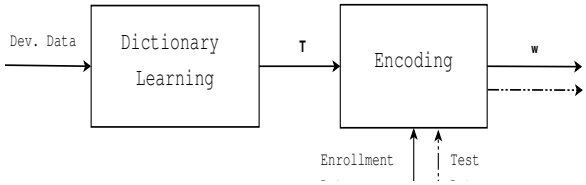

Figure 1: Block diagram of the total variability space model from the perspective of dictionary building and encoding. $\mathbf{T}$ refers to the matrix whose columns are the dictionary elements. $\mathbf{w}$ is, in conventional terms, the i-vector. Difference in line spacings are used to illustrate the difference in input/output pairs for the encoding process.

In this work, it is shown that while the dictionary building algorithm may be varied, the encoding procedure is important. The PPCA framework provides a suitable dictionary learning procedure. The comparison of PPCA to the total variability model is given in [6]. While the performance of the conventional PPCA technique ([7]) is not as good as that of the conventional i-vector technique, it is shown in this paper that learning dictionaries through the former approach and encoding i-vectors through the latter approach can prove to be beneficial. An important benefit from this is that the hyperparameter estimation can be expedited without compromising the performance of the speaker recognition system. The hyperparmeter that is referred here is usually the matrix containing the basis vectors of the total variability space.

The rest of the paper is organized as follows: in Section 2, the general framework of total variability space based approaches is discussed. This is followed by the discussion of the Factor Analysis-based i-vector approach (Section 2.1) and PPCA approaches (Section 2.2). Their dictionary building and encoding stages are detailed. The advantage of varying the encoding scheme is hypothesized in Section 2.3. In Section 2.4, a method to analyse different i-vector representations is reviewed. Analyses and results of the experiments on NIST 2008 SRE's core condition are discussed in Section 3.

2. Dictionary Learning and Encoding Procedures

In the total variability space model, a speech utterance is represented using a low dimensional vector. This is modelled as

$$\mathbf{s} = \boldsymbol{\mu} + \mathbf{T}\mathbf{w} \tag{1}$$

In Eq. 1, $\mathbf{T}$ consists of the basis vectors of the total variability space, $\mathbf{s}$ is the observed feature and $\mathbf{w}$ is the i-vector (identity vector). An observation $\mathbf{s}$ is obtained from the feature vectors of dimension F of a speech utterance by aligning each utterance with respect to a UBM. The clustered feature vectors are mean centered and variance normalized with respect to their corresponding mixtures in the UBM with C mixtures. This process is represented as follows

$$s_c = \boldsymbol{\Sigma}_c^{-\frac{1}{2}} \left(\sum_{t \in c} \mathbf{f}_t - \eta_c \boldsymbol{\mu}_c \right) \tag{2}$$

where s_c is the c^{th} $F \times 1$ block of the supervector such that

$$\mathbf{s} = [\mathbf{s}_1^t \mathbf{s}_2^t ... \mathbf{s}_C^t]^t$$

$\boldsymbol{\mu}_c$ is the mean vector of the c^{th} mixture of the UBM, η_c is the effective number of feature vectors aligned with the mixture and $\{\mathbf{f}_t\}$ are the sequence of feature vectors obtained from the speech utterance. This definition of supervector in Eq 1 is used throughout this work because it fits well with the assumption that the utterance was generated by a GMM (Gaussian Mixture Model).

To obtain a representation of a speech utterance with respect to the model in Eq 1, estimate of $\mathbf{T}$ needs to be obtained. Estimation of $\mathbf{T}$ can be considered as a dictionary building procedure as the columns of $\mathbf{T}$ are the basis vectors of the subspace. Given $\mathbf{T}$ and $\mathbf{s}$, estimating $\mathbf{w}$ can be considered as encoding. Several procedures for subspace estimation have been established. In the context of speaker recognition, a derivative of the procedure proposed in [2] is used in [1]. In this paper, this system will be referred to as the FA system. The state-of-the-art speaker recognition systems utilize this framework to provide superior performance in benchmark evaluations. However, other subspace estimation procedures also fit into this framework. PPCA is one such method. The effectiveness of PPCA in the context of speaker recognition has already been established [6]. While PPCA is certainly effective, its performance is not observed to be as good as that of the conventional i-vector approach. This is established in our experiments. PPCA's dictionary building procedure is, however, computationally simple, as opposed to the i-vector method (analysed in Sections 2.1, 2.2 and 3.3.1). This motivates us to study the combination of hyperparameter estimation of PPCA with FA-based i-vector extractor. It must be noted that all entries in a dictionary obtained from different procedures belong to the same feature space (supervector space in the case of speaker recognition).

In the next subsections, the individual techniques of FA-based approach and PPCA are briefly explained. The dictionary building and encoding phases are outlined. Following this, the combination of the two different phases are explained.

2.1. FA-based i-vector extraction

The EM algorithm provided in [2] is used to estimate $\mathbf{T}$ assuming all training examples come from different speakers [1]. The E- and M-steps in the algorithm are as follows:

E step:

$$\mathbf{w}_i = \mathbf{L}_i^{-1} \mathbf{T}^t s_i \tag{3}$$

where

$$\mathbf{L}_i = (\mathbf{I} + \mathbf{T}^t \mathbf{N}_i \mathbf{T}) \tag{4}$$

and $\mathbf{N}_i$ is the block diagonal matrix containing $\eta_c \mathbf{I}_F$ for $c = 1$ to C (defined in Eq 2). $\mathbf{I}_F$ is an identity matrix of size F.

M-step:

$$\mathbf{T}^{(c)} = \mathbf{C}^{(c)} (\mathbf{A}^{(c)})^{-1} \tag{5}$$

where $\mathbf{T}^{(c)}$ is the c^{th} FxR block of the $\mathbf{T}$ matrix, where R is the dimensionality of $\mathbf{w}$, such that

$$\mathbf{T} = [\mathbf{T}^{(1)t} \ \ \mathbf{T}^{(2)t} \ \ ... \ \ \mathbf{T}^{(C)t}]^t \tag{6}$$

Blocks $\mathbf{C}^{(c)}$ and $\mathbf{A}^{(c)}$ of matrices $\mathbf{C}$ and $\mathbf{A}$, respectively, are similarly arranged. $\mathbf{C}$ and $\mathbf{A}$ are defined as follows

$$\mathbf{C} = (\sum_i \mathbf{s}_i \mathbf{w}_i^t) \tag{7}$$

and

$$\mathbf{A}^{(c)} = \left(\sum_i \eta_{c,i} \left(\left(\mathbf{I} + \mathbf{T}^t \mathbf{N}_i \mathbf{T} \right)^{-1} + \mathbf{w}_i \mathbf{w}_i^t \right) \right)^{-1} \tag{8}$$

In Eq 8, $\eta_{c,i}$ refers to η_c for i^{th} utterance. In the E-step, MAP estimates of $\mathbf{w}$ are obtained for the supervector with respect to the current estimate of $\mathbf{T}$. In the M-step, the ML estimate of $\mathbf{T}$ is obtained. The value of $\mathbf{T}$ after convergence of the above EM algorithm can be considered as the dictionary. Equation 3 is used to extract i-vectors and can be considered as the encoding phase

Eq. 3 has a computational complexity of $O(CFR + CR^2 + R^3)$ [8]. Eq 5 is computationally intensive. The $\mathbf{T}$ matrix is estimated blockwise in sizes of $F \times R$. The right hand side is adjusted accordingly for blockwise estimation. The complexity of the entire re-estimation process is $O(\mathbb{N}CFR + C(\mathbb{N}R^2 + R^3))$. $\mathbb{N}$ is defined as the number of examples in the training set (size of development data). CFR refers to the cost of computing the first term in RHS of Eq 5. For each of the C blocks, the matrix needs to be computed and inverted.

2.2. PPCA

The PPCA algorithm is similar to a conventional factor analysis algorithm ([9]) that assumes isotropic covariance on the residual (unexplained) variabilities. The computation of covariance matrix in PCA [10] can be avoided using the EM algorithm to find the principal components. The E- and the M-steps follow:

E step:

$$\mathbf{w}_i = (\sigma^2 \mathbf{I} + \mathbf{T}^t \mathbf{T})^{-1} \mathbf{T}^t \mathbf{s} \tag{9}$$

M step:

$$\mathbf{T} = \left(\sum_i \mathbf{s}_i \mathbf{w}_i^t \right) \left(\sum_i \mathbf{w}_i \mathbf{w}_i^t \right)^{-1} \tag{10}$$

$$\sigma^2 = \frac{1}{\mathbb{N}F} \sum_i \left(|\mathbf{s}_i|^2 - 2\mathbf{w}_i^t \mathbf{T}^t \mathbf{s}_i + tr(\mathbf{w}_i \mathbf{w}_i^t (\mathbf{T}^t \mathbf{T})) \right) \tag{11}$$

The complexity of computing $\mathbf{w}$ in the E-step is $O(CFR)$ as the supervector independent terms can be precomputed. The complexity of recomputing $\mathbf{T}$ is just $O(\mathbb{N}CFR + NR^2 + R^3)$. This is certainly less compute intensive compared to the hyperparameter estimation procedure in Section 2.1.

The encoding procedure after having estimated the $\mathbf{T}$ matrix can be simplified from MAP (Maximum a Posteriori) to ML (Maximum Likelihood) estimates without any changes to the result. The ML estimate is given as

$$\mathbf{w} = (\mathbf{T}^t\mathbf{T})^{-1}\mathbf{T}^t\mathbf{s} \tag{12}$$

2.3. Proposed Method

The linear subspace estimation techniques discussed earlier result in basis vectors that span the subspace in which all supervectors lie. Assuming both techniques are effective enough in estimating the bases, the i-vectors extracted using them should be related through a linear transformation. This is because, the bases lie in the same physical space and i-vectors are just factor loadings (in Factor Analysis parlance). For a given set of basis vectors it will be interesting to study if a particular encoding scheme would perform better. In sparse coding schemes for example, soft thresholding-based encoding has been observed to provide consistently better performance regardless of the training method used for building dictionaries [5]. In such a case, it would be easier to build systems whose hyperparameter estimation procedures are computationally simpler without significant deterioration in performance.

Thus, in this work, we aim to use the computationally simple hyperparameter estimation technique of PPCA and show that when FA-based encoding scheme is used, the performance of the conventional i-vector system can be matched. That is, the hyperparameter $\mathbf{T}_P$ is randomly initialized and re-estimated by iterating through Equations 13,15 and 14 until convergence (presented for completeness)

$$\mathbf{w}_i = (\sigma^2\mathbf{I} + \mathbf{T}_p^t\mathbf{T}_p)^{-1}\mathbf{T}_p^t\,\mathbf{s}_i \tag{13}$$

$$\mathbf{T}_p = \left(\sum_i \mathbf{s}_i\mathbf{w}_i^t\right)\left(\sum_i \mathbf{w}_i\mathbf{w}_i^t\right)^{-1} \tag{14}$$

$$\sigma^2 = \frac{1}{CF}\sum_i \left(|\mathbf{s}_i|^2 - 2\mathbf{w}^t\mathbf{T}_p^t\mathbf{s}_i + tr(\mathbf{w}\mathbf{w}^t(\mathbf{T}_p^t\mathbf{T}_p))\right) \tag{15}$$

After estimating the $\mathbf{T_p}$ matrix, the i-vector $\mathbf{w}_i$ of a speaker i with supervector $\mathbf{s}_i$ is obtained using Eq 16.

$$\mathbf{w}_i = \mathbf{L}_{p_i}^{-1}\mathbf{T}^t\mathbf{s}_i \tag{16}$$

where

$$\mathbf{L}_{p_i} = \left(\mathbf{I} + \mathbf{T}_p^t\mathbf{N}_i\mathbf{T}_p\right) \tag{17}$$

The continued reference to factor loadings as i-vectors is because they still aim to represent the speaker class. This system is referred to as PPCA-ivec throughout this work. The overall system organization is shown in Fig 2.

Canonical Correlational Analysis (CCA) is used to study the relation between the i-vectors produced using the proposed and existing methods. The CCA method is briefly described in the next subsection.

2.4. Canonical Correlational Analysis

CCA is a useful analysis method to study linear relationships between two different representations of data samples [11, 12]. If for a set of samples there are two different representations, CCA finds directions in which the correlations of the vectors projected onto these directions from both representations are maximized. The correlation value obtained as a result has been

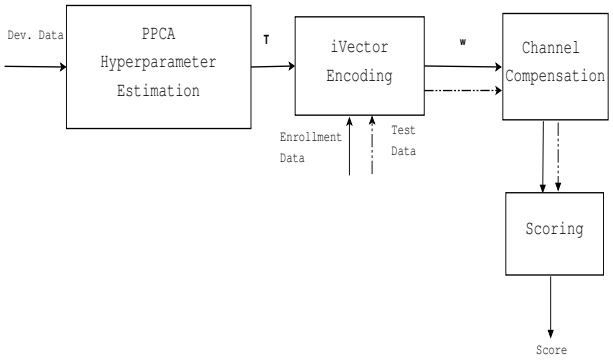

Figure 2: Block of diagram of system that uses PPCA algorithm for building dictionaries and mixing i-vector encoding scheme

shown to be a function of mutual information between the two given representations [13].

If x and y are two different representations, and there are a set of such pairs of examples, the correlation matrices $\mathbf{S}_{xx}$ and $\mathbf{S}_{yy}$ for each representation and the cross-correlation matrix $\mathbf{S}_{xy}$ are computed. To find directions $\boldsymbol{\omega}_x$ and $\boldsymbol{\omega}_y$ such that one representation can be transformed to another, it is required to optimize

$$\rho = \arg\max_{\boldsymbol{\omega}_x\boldsymbol{\omega}_y} \frac{\boldsymbol{\omega}_x\mathbf{S}_{xy}\boldsymbol{\omega}_y}{\sqrt{\boldsymbol{\omega}_x^t\mathbf{S}_{xx}\boldsymbol{\omega}_x\boldsymbol{\omega}_y^t\mathbf{S}_{yy}\boldsymbol{\omega}_y}} \tag{18}$$

$\boldsymbol{\omega}_x$ is the direction that transforms representation x and $\boldsymbol{\omega}_y$ is the corresponding transformation for representation y. The absolute value of the correlation co-efficient ρ shows the extent of linear relationship between two different representations.

An extension of CCA to analyse non-linear relationships is kernel CCA (KCCA) ([14],[15]). KCCA applies the kernel trick to estimate the extent of linearity in higher dimensional spaces, thereby discovering non-linear relations between the different representations. Using a similarity measure from a Mercer kernel, the dot product between the inputs is obtained in the kernel-induced feature space. The input feature vectors are assumed to have a linear relation in the kernel-induced space. In this paper, CCA and KCCA are used to analyse the nature of the relationships between different i-vector representations.

2.5. Physical Significance

The relation between the i-vectors obtained from the two different encoding schemes is not immediately apparent. If $\mathbf{T}_p$ is the hyperparameter estimated using PPCA and $\mathbf{w}_F$ and $\mathbf{w}_p$ are the i-vectors obtained using PPCA (Eq 3) and PPCA-ivec methods (Eq 12) , respectively, their relationship is given as

$$\mathbf{L}\mathbf{w}_f = \left(\mathbf{T}_p^t\mathbf{T}_p\right)\mathbf{w}_p \tag{19}$$

In the conventional PPCA framework, the cosine distance scores between two i-vectors $\mathbf{w}_{p_1}$ and $\mathbf{w}_{p_2}$ is given by

$$k_{cos}(\mathbf{w}_{p_1}, \mathbf{w}_{p_1}) = \frac{\mathbf{w}_{p_1}^t\mathbf{w}_{p_2}}{||\mathbf{w}_{p_1}||\,||\mathbf{w}_{p_2}||} \tag{20}$$

In Equation 20, $||\ .\ ||$ refers to the L_2 norm. Combining Equations 19 and 20, it can be inferred that the similarity between two i-vectors $\mathbf{w}_{f_1}$ and $\mathbf{w}_{f_2}$ estimated using the proposed

method are weighted distance measures between the PPCA i-vectors $\mathbf{w}_{p_1}$ and $\mathbf{w}_{p_2}$:

$$k_{cos}\left(\mathbf{w}_{f_1}, \mathbf{w}_{f_2}\right) = \frac{\mathbf{w}_{p_1}^t \left(\mathbf{T}_p^t \mathbf{T}_p\right) \mathbf{L}_{p_1}^{-1} \mathbf{L}_{p_2}^{-1} \left(\mathbf{T}_p^t \mathbf{T}_p\right) \mathbf{w}_{p_2}}{||\mathbf{w}_{p_1}^t \left(\mathbf{T}^t \mathbf{T}\right) \mathbf{L}_{p_1}^{-1}|| \, ||\mathbf{L}_{p_2}^{-1} \left(\mathbf{T}^t \mathbf{T}\right) \mathbf{w}_{p_2}||} \tag{21}$$

Eq 21 can be considered as a weighted similarity measure with the weights being class-specific. The weights are related to the covariance estimate $\mathbf{L}$ of the posterior distribution of the i-vector in FA model. Therefore, a difference in performance when using the two different coding schemes under consideration would only be due to the application of these weights. It is shown in Sec 3.5 that it has a positive effect in improving the discriminability amongst speakers.

Alternately, the term $\left(\mathbf{T}_p^t \mathbf{T}_p\right) \mathbf{L}_{p_1}^{-1} \mathbf{L}_{p_2}^{-1} \left(\mathbf{T}_p^t \mathbf{T}_p\right)$ can be considered as a normalization factor used in scoring as it is a positive semi-definite matrix (psd) by virtue of its constituents being psd.

3. Experimentation

The experiments are performed on the benchmark NIST 2008 SRE database. The telephone-telephone conditions, namely C6 and C8, are used for evaluation [16]. Conditition C8 refers to the the subset where only native American English speakers are considered for modelling and evaluation. Condition C6 refers to the entire telephone based evaluation that includes multiple languages. There are 648 male speakers and 1200 female speakers. Details of MFCC feature extraction is given in Section 3.1. Gender-dependent 1024 mixture UBMs are built using the NIST 99 and 2003 SRE databases [17]. These databases are derived from Switchboard Cellular and Switchboard Phase II corpora.

3.1. Feature extraction

MFCC (Mel Frequency Cepstral Co-efficients) feature vectors [18] with 22 cepstral co-efficients are computed from 40 log filter bank energies for both male and female speakers from 25 ms window with 15 ms overlap. Velocity features are computed over 7 consecutive frames and appended to the MFCC feature vectors. On utterances for which ASR transcripts were unavailable, a tri-Gausssian-based Voice Activity Detector (VAD) is used to remove non-speech frames. Short term Gaussianization (STG) is performed over 300 frames after VAD [19].

3.2. Factor Analysis system

Gender-dependent FA system is developed with i-vector dimensionality being 500. The development data set consists of - NIST SRE 1999, NIST SRE 2003, NIST SRE 2004, Switchboard Cellular Part 2, Switchboard II Phase 1, Fisher Database Part 1 [1] [2]. This amounts to 640 hours of data (after VAD) for males and 840 hours of data for female subsets.

3.2.1. Length Normalization

In [20], it is observed that length normalized i-vectors after whitening are extremely useful in Gaussianizing the i-vector distribution. This is a simpler form of Radial Gaussianization [21]. It is extremely important in dealing with non-Gaussian nature of the i-vectors, if present [22]. Thus, i-vectors obtained are length normalized before applying channel compensation techniques. If $\mathbf{W}$ is the whitening matrix obtained from the data, each i-vector $\mathbf{w}$ is processed as

$$\tilde{\mathbf{w}} = \frac{\mathbf{W}\,\mathbf{w}}{||\mathbf{W}\,\mathbf{w}||} \tag{22}$$

[1] The choice of development data has to do with its availability to the author.

[2] Transcripts of only Fisher Part 1 and Switchboard II Phase 1 databases were available at the time of experimentation.

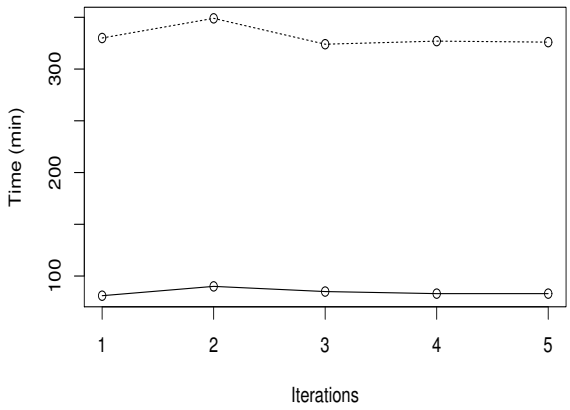

Figure 3: Analysis of time taken to complete an iteration in FA and PPCA systems. Solid: PPCA, dotted: FA

3.3. PPCA systems

The development data used to compute the projection matrix in the PPCA case is the same as that used for the FA system. The basis vector set's size is 500. It is important to note that the supervector (Eq 2) that is projected here is the same as that projected in the i-vector system. As discussed earlier, two kinds of encoding schemes are employed to estimate the i-vector - the conventional PPCA based encoding and the i-vector based encoding scheme. Length normalization is applied only on the i-vectors obtained using the latter scheme. This system is referred to as PPCA-iVec-LN. Length normalization on the i-vectors obtained from the conventional PPCA technique did not improve the performance.

3.3.1. Duration analysis on Dictionary Learning

To emphasize further that the hyperparameter estimation procedure of a PPCA system is much faster than that in the FA system, the duration of each iteration is analysed. Figure 3 compares the duration of each iteration for the first 5 iterations while estimating the dictionary for the male data set. The duration measure is given in minutes. Clearly, there is a 4x difference in speed. It should be noted that the implementations of PPCA and FA techniques have been optimized to take advantage of a heterogeneous distributed computing environment using Map-Reduce techniques. GNU parallel is used as a tool to facilitate this [23].

3.4. Channel Compensation and Scoring

After estimating i-vectors and length-normalizing them, channel compensation is performed using LDA followed by WCCN. Only speakers with at least 6 examples were considered for LDA computation. The constraint enhances the performance of LDA.

Given a set of i-vectors $\{\mathbf{w}_i\}_{i=1}^{I}$ containing J classes with each class containing n_j examples, the LDA projection matrix $\mathbf{A}_{lda}$ is estimated by optimizing the Fisher Discriminant function [10] where the inter-class distance is maximized while minimizing the intra-class distance. If $\mathbf{S}_w$ is the intraclass scatter matrix and $\mathbf{S}_b$ is the interclass scatter matrix, the projection matrix is the solution to the generalized eigenvalue problem given below

$$\mathbf{S}_b e = \lambda \mathbf{S}_w \mathbf{e} \tag{23}$$

where

$$\mathbf{S}_w = \sum_{j=1}^{J} \frac{1}{n_j} \sum_{i=1}^{n_j} (\mathbf{w}_{j,i} - \bar{\mathbf{w}}_j)^t (\mathbf{w}_{j,i} - \bar{\mathbf{w}}_j) \tag{24}$$

and

$$\mathbf{S}_b = \sum_{j=1}^{J} \frac{1}{n_j} \sum_{i=1}^{n_j} \mathbf{w}_{i,j} \mathbf{w}_{i,j}^t \tag{25}$$

where $\bar{w}_j$ is the class-specific mean. In Equation 25, it is assumed that the mean of all i-vectors is $\mathbf{0}$. The estimation procedure of WCCN matrix is related to Eq 24 through a normalization factor J. The projection matrix after WCCN is given as

$$\mathbf{B} = chol(J \ \mathbf{S}_w^{-1}) \tag{26}$$

where $chol$ refers to Cholesky decomposition.

The LDA matrix was used to remove 150 dimensions according to the eigenvalues obtained after discriminant analysis. To test a claim attached to an utterance during evaluation, cosine distance scoring is used. 200 T-Norm speakers were used for score normalization [24]. If $\mathbf{A}_{lda}$ is the projection matrix obtained from LDA and $\mathbf{B}$ is obtained on LDA projected data , cosine distance score is given by

$$k_{cos}(\mathbf{w}_{trn}, \mathbf{w}_{tst}) = \frac{\left(\mathbf{B}^t \mathbf{A}_{lda}^t \mathbf{w}_{train}\right)^t \left(\mathbf{B}^t \mathbf{A}_{lda}^t \mathbf{w}_{tst}\right)}{||\mathbf{B}^t \mathbf{A}_{lda}^t \mathbf{w}_{train}||||\mathbf{B}^t \mathbf{A}_{lda}^t \mathbf{w}_{tst}||} \tag{27}$$

3.5. Discriminability Analysis

Before evaluating the system, a study on the discriminability of the i-vectors is performed to observe if the proposed schemes may provide any benefit. The interspeaker and intraspeaker variabilities of i-vectors obtained for different systems are compared. All analyses are performed on i-vectors obtained without channel compensation. To study interspeaker and intraspeaker variabilities cosine similarity measure is used (Eq 27). i-vectors extracted from the FA system, PPCA system, PPCA-iVec system, and PPCA-iVec-LN system are analysed.

The means and variances of interspeaker and intraspeaker variabilities are summarised in Table 1. The measures are obtained from the speakers in core condition of NIST 2004 SRE data set. It can be observed that the FA system provides better separability than PPCA. Moreover, the variances of the measures in the PPCA system are too high. This accounts for the confusability in recognition. The separability in PPCA-iVec is similar to that of the conventional i-vector system. Notably, confusability has been reduced compared to PPCA system. In particular, the difference with respect to PPCA emphasizes the role of class-specific information used (as hypothesized in Section 2.5) to improve discriminability.

Table 1: Interspeaker and Intraspeaker variabilities of speakers in NIST 2004 SRE when using different approaches discussed. Means and Variances of Cosine distance scores without normalization are used for comparison.

System	Interspeaker		Intraspeaker	
	Mean	Var	Mean	Var
FA	2.1E-5	6.27E-5	0.01	1.6E-3
PPCA	2	2.71	3.1	2.6
PPCA-iVec	6E-4	1E-4	8.6E-3	2.4E-3
PPCA-iVec-LN	2.71E-2	4E-3	0.4164	0.05

Table 2: CCA and KCCA-based analyses of nature of relationships between i-vectors obtained through different extraction procedures. All non-linear relationships are observed in the space induced by a polynomial kernel of degree 4

Comparison	Linear(l)/ Non-linear (n) Relation	Correlation Co-efficient
FA vs PPCA	l	0.9819
FA vs PPCA-ivec	n	1.0
PPCA vs PPCA-ivec	n	1.0
PPCA vs PPCA-ivec-LN	n	0.99
FA vs PPCA-ivec-LN	n	**1.0**

3.6. Relationship between i-vectors

The relationship between various i-vectors extracted through different methods discussed is interesting to study. CCA and KCCA are used for this purpose. The former is used to detect the presence of linear relationships while the latter is used for non-linear relationships. In this study, a polynomial kernel is used [25, 11]. A polynomial kernel of degree d is defined as

$$k_{poly}(\mathbf{x}_a, \mathbf{x}_b) = (\mathbf{x}_a^t \mathbf{x}_b + c)^d \tag{28}$$

where $\mathbf{x}_a$ and $\mathbf{x}_b$ are two data points whose similarity is being measured, and c is a parameter of the kernel. c value is set to 0 in the experiments. Absolute values of correlation are reported. The i-vectors from the male subset of development set are compared before channel compensation. The following observations are made: the relationship between FA and PPCA is linear with a co-efficient of 0.9819. However, the relationship between FA and PPCA-ivec is non-linear. A polynomial kernel of degree 4 detected a non-linear relationship with a co-efficient of exactly **1.0**. This result provides interesting scope for using other similarity measures for scoring. In this work, however, we use only the cosine distance measure for scoring. All relevant comparisons are given in Table 2.

An important inference from these results is that even though the i-vectors across different representations are related, the orientation of the subspace and/or the projection method are important for better discriminability. The discriminability provided by these methods have already been quantified in Section 3.5.

3.7. Results

The FA system with length-normalized i-vectors obtained from the conventional i-vector estimation procedure followed by channel compensation constitutes the i-vector baseline. Simi-

Table 3: Performance of the baseline systems and PPCA systems using i-vector encoding scheme (PPCA-iVec). LN: Length Normalization

System	C8	C6
Baseline		
FA + LN + LDA + WCCN	5.37%	8.75%
PPCA + LDA + WCCN	8.11%	12.08%
Proposed		
PPCA-iVec	9.5%	12.5%
PPCA-iVec + LDA + WCCN	7.12%	10.06%
PPCA-iVec-LN + LDA + WCCN	**5.76**%	**9.26**%

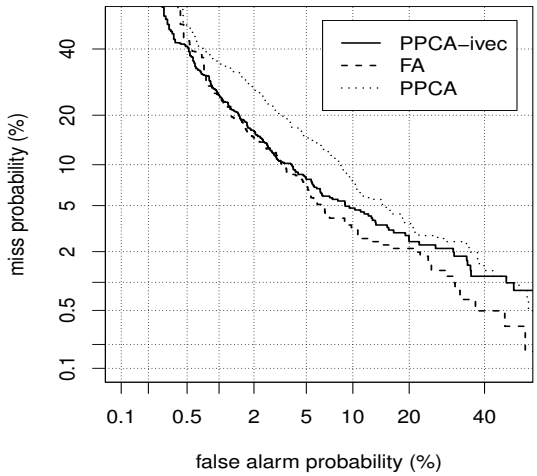

Figure 4: DET curve comparing the baseline and the PPCA-ivec-LN systems on condition C8

larly, the PPCA-baseline is built using the conventional procedure (discussed in Sec 2.2) along with channel compensation. The Equal Error Rates of the FA and PPCA systems are reported in Table 3. Even though the development data set used to estimate hyperparameters is the same, the FA system performs better than the PPCA system.

The results of the proposed scheme with and without channel compensation and length normalization are also given in Table 3. All results are superior to the PPCA baseline. The results on $C8$ condition of the PPCA-ivec-LN system closely matches that of the FA system. This clearly shows the positive effect of the encoding scheme used. A similar inference can be made from the result on $C6$. An absolute difference of 1.6% with respect to the FA system is present. This, however, is still better than PPCA system by approximately 2.8%. The advantage is clearly introduced by the efficient FA-based encoding. Thus, computational efficiency of the PPCA system is utilized to obtain the basis vectors while still matching the performance of FA system through the use of its efficient encoding scheme. Also, the efficiency of extracting i-vectors using this encoding scheme is realized. The DET curves [26] corresponding to the C8 and C6 systems are given in Fig 4 and Fig 5. Only the best performing, namely PPCA-iVec-LN, system is compared with baseline. The DET curves illustrate that PPCA-iVec-LN is better than the PPCA system and closer to the FA system in terms of performance.

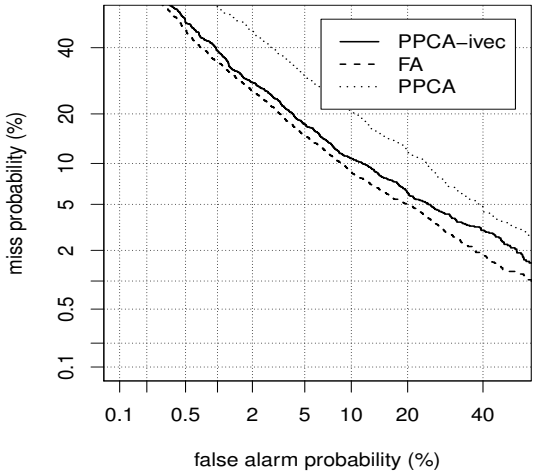

Figure 5: DET curve comparing the baseline and the PPCA-ivec-LN systems on condition C6

4. Summary and Future Work

PPCA framework for subspace modelling is used to build dictionaries that could be used to extract i-vectors. To provide performance efficiency similar to that of FA-based i-vectors, the i-vector extraction algorithm in the FA framework for speaker recognition is used. A $4\times$ speed up in time is obtained while building the system for speaker recognition. The performance of the system matches that of the FA system with just an absolute difference of 0.39% on C8 condition and 0.51% on C6 condition. The performance compared to that of the PPCA baseline is relatively superior by 2.8% in terms of EER.

The relationship between the i-vectors obtained in FA system and PPCA system is non-linear and related by a polynomial kernel of 4^{th} degree. The i-vectors have a correlation coefficient of 1.0 in the higher dimensional space. This definitely suggests a scope for new scoring techniques that take advantage of this relationship to further the performance of the current system.

5. References

[1] Najim Dehak, Patrick Kenny, Rda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Tran. on Audio, Speech and Language Processing*, pp. 788–798, 2011.

[2] Patrick Kenny, G Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 3, pp. 335–354, May 2005.

[3] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[4] Daniel Garcia-Romero and Carol Y Espy-Wilson, "Joint factor analysis for speaker recognition reinterpreted as signal coding using overcomplete dictionaries," *In Odyssey, paper 022*, 2010.

[5] Adam Coates and Andrew Y. Ng, "The importance of

encoding versus training with sparse coding and vector quantization," *ICML*, 2010.

[6] Yun Lei and John Hansen, "Speaker recognition using supervised probabilistic principal component analysis," *In Proc. of Interspeech*, pp. 382–385, 2010.

[7] Michael E. Tipping and Christopher M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 661–622.

[8] O Glembek and et al., "Simplification and optimization of i-vector extraction," *In Proc. of ICASSP*, 2011.

[9] Donald Rubin and Dorothy Thayer, "Em algorithms for ml factor analysis," *Psychometrika*, vol. 47, no. 1, pp. 69–76, March 1982.

[10] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*, Wiley India, 2007.

[11] John Shawe-Taylor and Nello Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.

[12] David R. Hardoon, Sandor Szedmak, Or Szedmak, and John Shawe-taylor, "Canonical correlation analysis; an overview with application to learning methods," Tech. Rep., 2007.

[13] Magnus Borga, "Learning multidimensional signal processing," *PhD Thesis*, 1998.

[14] Shotaro Akaho, "A kernel method for cannoical correlation analysis," *International Meeting of Psychometric Society IMPS2001*, 2001.

[15] Simon Haykin, *Artificial Neural Networks: A Comprehensive Foundation*, Prentice-Hall Inc., 3rd edition, 2007.

[16] "The NIST year 2008 speaker recognition evaluation plan," http://www.itl.nist.gov/iad/mig/tests/sre/2008/index.html.

[17] "The NIST year 2003 speaker recognition evaluation plan," http://www.itl.nist.gov/iad/mig/tests/sre/2003/index.html.

[18] Davis, S. and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, pp. 357–366, 1980.

[19] Jason Pelecanos and Sridha Sridharan, "Feature warping for robust speaker verification," *In Proc. of Speaker Odyssey*, 2001.

[20] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," *In Proc. of Interspeech, Florence, Italy*, pp. 249–252, August 2011.

[21] S Lyu and E. P. SimonCelli, "Nonlinear extraction of independent components of natural images using radial gaussianization," *Neural Computation*, vol. 21, no. 6, June 2009.

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

[23] O. Tange, "GNU parallel - the command-line power tool," *login: The USENIX Magazine*, pp. 42–47, 2011.

[24] Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

[25] Bernhard Schlkopf and Alexander J.Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond (adaptive computation and machine learning)*, MIT Press, Cambridge, 2002.

[26] A Martin, G Doddington, T Kamm, M Ordowski, and M Przybocki, "The DET curve in assessment of detection task performance," *in Proc. of EUROSPEECH, 1997*, pp. 1895–1898.