Odyssey 2012
The Speaker and Language Recognition Workshop
25-28 June 2012, Singapore

On exploring the similarity and fusion of i-vector and sparse representation based speaker verification systems

Haris B C and R. Sinha

Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati
Guwahati -781039, India
email: {haris, rsinha}@iitg.ernet.in

Abstract

The total variability based i-vector has become one of the most dominant approaches for speaker verification. In addition to this, recently the sparse representation (SR) based speaker verification approaches have also been proposed and are found to give comparable performance. In SR based approach, the dictionary used for sparse representation is either exemplar or learned from data using the KSVD algorithms and its variants. Recently the use of the total variability matrix of the i-vector system as the dictionary for the SR based approach has also been reported. Motivated by these, in this work, we first highlight the similarity between the i-vector and the learned dictionary SR based approaches for speaker verification. It is followed by the exploration about various kinds of learned dictionaries, their sizes and the sparsity constraint in context of SR based speaker verification. Further we have explored the feature level as well as the scores level fusions of these two approaches.
Index Terms: speaker verification, sparse representation, learned dictionaries, total variability space.

1. Introduction

Speaker verification systems are predominently based on either generative or discriminative modeling techniques. The most commonly used generative modeling method is the Gaussian mixture model-universal background model (GMM-UBM) where as the most successful discriminative techniques are based on support vector machines (SVM). The SVM based approach requires a fixed dimension representation of the speaker utterance for the classification. There are a number of methods to achive this from a utterances having varying numer of features, but the most sccessfull one is based on the concatenation of the mean vectors of the GMM-UBM derived from speech utterances which are commonly reffered to as GMM mean supervectors. Later, various mehtods like LDA, WCCN, NAP and JFA were proposed to futher improve the performance of the SVM based systems by intersession variability compensation. In recent years, the total variability i-vector based speaker verification [1] has attained large popularity because of its excellent performance with reduced complexity compared to the GMM supervector based approaches. It removes the less significant dimensions from the supervector representation of speech utterance by projecting it to a low dimensional space called the total variability space.

In last few years, the *sparse representation* and its properties have been actively for signal processing applications. The sparse representation involves the representation of a target vector in terms of a sparse linear combination of the columns of a redundant matrix representing the target signal space. In the sparse representation literature, the redundant matrix is commonly referred to as the 'dictionary' and its columns as 'atoms'. Recently the discriminative abilities of the sparse representation have been exploited for various pattern recognition tasks including speaker recognition and verification. In [2], Kua et. al. proposed a speaker recognition system which uses sparse representation classification (SRC) with an exemplar dictionary created using GMM mean supervectors. Later speaker verification tasks using the SRC with exemplar dictionary created using GMM mean supervectors and total variability i-vectors were also reported in [3] and [4] respectively.

In our recent work [5], we have explored the use of exemplar dictionary based SRC for the speaker verification task in realistic environment, which gave an improved performance compared to the conventional GMM-UBM based system. Later, motivated by the fact that the learned dictionaries not only outperform the exemplar ones but also are more data-independent, in [6] we have presented a speaker verification system employing sparse representation of centered GMM mean supervectors over a redundant dictionary learned using the KSVD algorithm. This system uses the sparse representation of centered GMM mean supervectors over the learned dictionary as a speaker representation, and is referred to as SRSV system. We have extended this work with the use of discriminatively learned dictionaries in [7] and the proposed system was compared to the SRC over exemplar dictionary based SV system as well as the existing i-vector based SV system. On NIST 2003 SRE dataset, the proposed system with discriminatively learned dictionary found to outperform all other SV systems considered both with and without session/channel variability compensation.

In [8], an SV system which employs sparse representation of centered GMM mean supervectors using the total variability matrix of the i-vector based system as the dictionary has been explored. The system was reported to give comparable performance to the i-vector based system and a score level fusion of it with the i-vector based system resulted in an improved performance showing the complementary information carried by these systems. In this work, we compare the i-vector and the learned dictionary based SRSV systems in a consistent setup and also study the different aspects of the dictionary used for the SRSV system. We also present a novel method of combining the i-vector and SRSV sparse representation over KSVD learned dictionary based SV systems at the feature representation level. Here, the centered GMM mean supervectors are smoothed using the total variability matrix of the i-vector system prior to the sparse representation over the dictionary. We also report the results of the score level fusion of various sys-

tems considered in this paper.

The organization of the paper is as follows: In the Section 2, we describe the total variability i-vector based SV system. In Section 3 the SRSV system employing sparse representation of GMM supervectors is explained. In Section 4, we explain the proposed smoothing of GMM mean shifted supervectors. The various session/channel compensation methods used in this work are explained in Section 5. The details of the database and the experimental setup are given in Section 6 followed by the discussion of results in Section 7. The paper is concluded in Section 8.

2. Total variability i-vector SV system

In this section we describe the total variability i-vector based speaker verification system. In this system, the centered GMM mean supervectors are projected to a low rank matrix to get the i-vector representation. The low rank projection matrix represents the dominant speaker and channel variabilities simultaneously and hence is called the total variability matrix. For a given total variability matrix $\boldsymbol{T}$, the i-vector $\boldsymbol{w}$ can be related to the centered GMM mean supervector $\boldsymbol{y}$ as,

$$\boldsymbol{y} = \boldsymbol{T}\boldsymbol{w} \tag{1}$$

Given a GMM-UBM λ consisting of C components with mean $\boldsymbol{\mu}_c$, and variance $\boldsymbol{\Sigma}_c$, where $c = 1, 2, \ldots, C$ and a sequence of L speech feature vectors $\{\boldsymbol{f_1}, \boldsymbol{f_2}, \ldots, \boldsymbol{f_L}\}$, the centered GMM mean supervector $\boldsymbol{y}$ is formed by concatenating the component specific centered GMM mean vectors $\boldsymbol{y}_c$ which are computed as,

$$\boldsymbol{y}_c = \frac{\boldsymbol{F}_c}{N_c} \tag{2}$$

where, N_c and $\boldsymbol{F}_c$ are the 0^{th} order and the 1^{st} order statistics of the speech frames on the c^{th} component of the GMM-UBM which are given by,

$$N_c = \sum_{t=1}^{L} P(c|\boldsymbol{f}_t, \lambda) \tag{3}$$

$$\boldsymbol{F_c} = \sum_{t=1}^{L} P(c|\boldsymbol{f}_t, \lambda)(\boldsymbol{f}_t - \boldsymbol{\mu}_c) \tag{4}$$

The matrix $\boldsymbol{T}$ is learned using probabilistic PCA (PPCA) method using the centered GMM mean supervectors of suitable development data as described in [1]. For a given $\boldsymbol{T}$ and $\boldsymbol{y}$, the estimated i-vector $\hat{\boldsymbol{w}}$ is computed as,

$$\hat{\boldsymbol{w}} = (\boldsymbol{I} + \boldsymbol{T'}\boldsymbol{\Sigma^{-1}}\boldsymbol{N}\boldsymbol{T})^{-1}\boldsymbol{T'}\boldsymbol{\Sigma}\boldsymbol{N}\boldsymbol{y} \tag{5}$$

In the training and testing phases, speech utterances are represented in the form of i-vectors. Let $\hat{\boldsymbol{w}}_{clm}$ and $\hat{\boldsymbol{w}}_{tst}$ represent the i-vectors of the claimed and the test speakers utterances respectively, then the verification of a claim is performed by comparing the cosine kernel score between these two i-vectors to a threshold γ as given below.

$$\frac{< \hat{\boldsymbol{w}}_{clm}. \hat{\boldsymbol{w}}_{tst} >}{\|\hat{\boldsymbol{w}}_{clm}\| \|\hat{\boldsymbol{w}}_{tst}\|} \lessgtr \gamma \ \text{(Threshold)} \tag{6}$$

3. Sparse representation based SV system

In this method, the supervector $\boldsymbol{y}$ derived from a speaker utterances is modeled using the sparse representation with a dictionary $\boldsymbol{D}$ as,

$$\boldsymbol{y} = \boldsymbol{D}\boldsymbol{x} \tag{7}$$

The dictionary $\boldsymbol{D}$ is of $\boldsymbol{M} \times \boldsymbol{N}$ size where $\boldsymbol{M}$ corresponds to the dimension of supervector and $\boldsymbol{N}$ is the number of atoms. Assuming $\boldsymbol{D}$ to be sufficiently redundant, the sparse solution $\hat{\boldsymbol{x}}$, estimated using suitable algorithms can be considered as a representation of the speaker. In the training phase, the sparse representation for all speakers are derived. During the testing phase, the sparse representation of the test utterance is also derived in a similar fashion and is compared with the sparse representation of the claimed speaker using an appropriate metric.

In our previous work [6] the mean supervector for an utterances was derived from the GMM obtained by relevance MAP adaptation [9] of means of the UBM by that utterance. Then the centered GMM mean supervectors were derived by subtracting the UBM mean supervector from the adapted mean supervector. The dictionary $\boldsymbol{D}$ is learned using KSVD [10] algorithm with the centered GMM mean supervectors of a large set of utterances from the development database. The sparse representation $\hat{\boldsymbol{x}}$ was estimated using the orthogonal matching pursuit (OMP) algorithm which minimizes representation error with a constraint on the l_0-norm as,

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \|\boldsymbol{y} - \boldsymbol{D}\boldsymbol{x}\|_2^2 \ \text{ such that, } \ \|\boldsymbol{x}\|_0 < s \tag{8}$$

where, s is the constraint on the number of atoms selected for representation. The cosine kernel metric can be used for finding the similarity between the claimed and the test sparse vectors and that is compared with a threshold for the verification purpose as given in Equation 9. To be consistent with the i-vector based system which does not involve any relevance factor in creation of the supervectors, in this work the supervectors for the SRSV system are derived with relevance factor set to zero. With zero relevance factor, the MAP adaptation method and the formulation of centered GMM mean supervectors described in Section 1 become identical.

3.1. Dictionaries for SR-SV system

The choice and design of the dictionary plays a crucial role in sparsifying the signal and hence in the success of the sparse representation based classification. In the following subsections, we describe various dictionary learning algorithms that are used for learning the dictionaries for the SRSV system.

3.1.1. Dictionary learned using KSVD

The KSVD [10] is one of the most widely used algorithms for learning redundant dictionaries for sparse representations. It is a generalization of the well known K-means clustering algorithm. KSVD algorithm constructs a dictionary of K atoms that leads to the best possible representation for each member of the training examples with a minimum sparsity constraint. The dictionary learning problem is represented as,

$$\min_{\boldsymbol{D},\boldsymbol{X}} \left\{\|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|_2^2\right\} \quad \text{subject to } \|\boldsymbol{x}_i\|_0 \leq T_0 \quad \forall i \tag{9}$$

where, $\boldsymbol{Y}$ is the set of dictionary training vectors, $\boldsymbol{D}$ is the dictionary, $\boldsymbol{X}$ is the set of sparse vectors corresponding to $\boldsymbol{Y}$ and T_0 is the constraint on sparsity. The learning is an iterative process and each iteration has two stages: the sparse coding stage and the dictionary update stage. In the sparse coding stage, any of the pursuit methods such as OMP can be used for finding the sparse representation of the given set of examples based on the current dictionary. The update of the dictionary atoms is done jointly with an update of the sparse representation coefficients related to it, thus resulting in accelerated convergence.

The SRSV system which uses KSVD dictionary is referred to as KSVD-SRSV system in this paper.

3.1.2. Dictionary learned using S-KSVD

The SKSVD [11] is a *supervised* version of the KSVD algorithm for learning discriminative dictionary. It uses *class supervised simultaneous* OMP (CSSOMP) in the sparse coding stage of the dictionary learning process which differs from OMP in two aspects: (i) CSSOMP uses the same set of atoms from the dictionary to represent all examples from a given class and so attempts to extract the common internal structure of that class whereas OMP treats each example independently (ii) In addition to the original reconstruction criterion of minimum squared error used in OMP, CSSOMP also uses a discrimination measure which increases the separability among classes. The sparse discriminant dictionary learning problem is represented as,

$$\max_{\boldsymbol{D},\boldsymbol{X}} \left\{ \theta . J\left(\left\{ \left\{ \boldsymbol{x}_i^j \right\}_{i=1}^{n_j} \right\}_{j=1}^{c} \right) - \sum_{j=1}^{c} \sum_{i=1}^{n_j} \left\| \boldsymbol{y}_i^j - \boldsymbol{D}\boldsymbol{x}_i^j \right\|_2^2 \right\}$$

$$\text{subject to} \quad \left\| \boldsymbol{x}_i^j \right\|_0 \leq T_0, \quad \forall i, j \tag{10}$$

The function $J(.)$ represents the discriminant measure defined as $:= \frac{trace(\boldsymbol{B})}{trace(\boldsymbol{W})}$ where $\boldsymbol{B}$ and $\boldsymbol{W}$ are the *between-class* and the *within-class* covariance matrices of the learning data, respectively. $\boldsymbol{D}$ is the learned dictionary, $\boldsymbol{y}_i^j$ is i^{th} example vector of j^{th} class from a set of dictionary training data having c classes with $n_j,\ 1 \leq j \leq c$ examples per class. $\boldsymbol{x}_i^j$ is the sparse coefficient vector corresponding to $\boldsymbol{y}_i^j$. θ is a parameter controlling the trade-off between discriminative and re-constructive terms in the learning criterion. The SKSVD dictionary based SRSV system is referred to as SKSVD-SRSV system in this paper.

3.2. Total-variability dictionary for the SRSV system

In [8] the total variability matrix $\boldsymbol{T}$ of the i-vector based system is used as the dictionary for sparse representation in a setup similar to that explain in Section 3. For the purpose of comparison, we have created an SRSV system which uses the total variability matrix derived using PPCA as the dictionary. This system is referred to as the T-SRSV system in this work. It is to note that, we have used the OMP algorithm for finding the sparse representation while LASSO algorithm has been used for the same purpose in [8].

4. Total variability matrix smoothing of supervectors

Motivated by the improvement reported with the score level fusion of i-vector and SRSV based SV systems in [8], we recently explored a novel feature level fusion of these two systems[12]. In our proposed approach, we basically smooth the centered GMM mean supervectors using the total variability matrix prior to dictionary learning and sparse representation. For this purpose, the centered GMM mean supervectors are first projected using the total variability matrix using the equation 5 to get the i-vector representations. The resulting i-vectors are then multiplied with the $\boldsymbol{T}$ matrix to re-synthesize the centered GMM mean supervectors as given below.

$$\tilde{\boldsymbol{y}} = \boldsymbol{T}\hat{\boldsymbol{w}} \tag{11}$$

As $\boldsymbol{T}$ is a low rank matrix, the re-synthesized supervectors are the smoothed version of the original supervectors. These smoothed supervectors are then used for the speaker verification using sparse representation as described in section 2. This system is referred to as the *T-smoothed SRSV* system in this work. The smoothing of supervectors with the total variability matrix would result in the removal of smaller nuisance variations which is hypothesized to be from the intra-speaker variabilities. To verify this, we have performed a study with five speakers each having five utterances taken from the Switchboard corpus (development data). The centered GMM mean supervector for each of these utterances were derived and these supervectors were then smoothed using the total variability matrix which is learned from the full development data. To understand the effect of smoothing, we have computed the cosine kernel based similarity among all supervectors for both cases. The similarity scores among all the supervectors with and without smoothing as an intensity plot is shown in Figure 1. On comparing the two plots, we note that with smoothing the similarity scores among supervectors have improved in general. Also, the improvement in similarity score in case of intra-speaker cases is more than that in case of inter-speaker cases. It is noted that the average improvement in similarity score due to smoothing is 154% in case of intra-speaker cases whereas it is 85% in case of inter-speaker cases. This can be interpreted as higher reduction in intra-speaker variations compared to that in inter-speaker variations due to the smoothing.

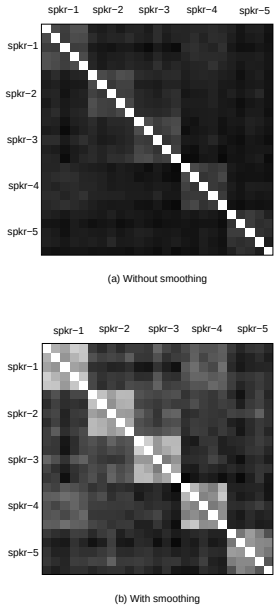

Figure 1: Intensity plot showing the similarity among supervectors of five speakers with five utterances per speaker (a) without smoothing and (b) with smoothing

5. Session/channel variability compensation

The various session/channel variability compensation methods that are applied to different SV systems considered in this work are briefly described in this section.

5.1. Joint factor analysis

In joint factor analysis (JFA) [13], the centered GMM mean supervector $\boldsymbol{y}$ for a speaker is represented as the sum of three factors as,

$$\boldsymbol{y} = \boldsymbol{U}\boldsymbol{u} + \boldsymbol{V}\boldsymbol{v} + \boldsymbol{D}\boldsymbol{d} \quad (12)$$

where $\boldsymbol{U}$ is the session/channel subspace matrix, $\boldsymbol{V}$ is the speaker subspace matrix, and $\boldsymbol{D}$ represents the diagonal residual matrix. The vectors $\boldsymbol{u}$, $\boldsymbol{v}$ and $\boldsymbol{d}$ are the projections of $\boldsymbol{y}$ in their respective subspaces. The session/channel compensated centered GMM mean supervector is given by $\tilde{\boldsymbol{y}} = \boldsymbol{V}\hat{\boldsymbol{v}} + \boldsymbol{D}\hat{\boldsymbol{d}}$, where $\hat{\boldsymbol{v}}$ and $\hat{\boldsymbol{d}}$ are the estimated projections to the respective subspaces. In our implementation, we have used $\boldsymbol{V}\hat{\boldsymbol{v}}$ factor only ignoring the residual factor.

5.2. Linear discriminant analysis

Linear discriminant analysis (LDA) is a commonly used method for dimensionality reduction and is widely used in pattern recognition applications. In LDA, the feature vectors are projected down to a set of new orthogonal axises where the discrimination between different classes is maximum. The projection matrix is composed by the eigen vectors corresponding to the best eigen values of the eigen analysis equation, $(\boldsymbol{W}^{-1}\boldsymbol{B})\boldsymbol{v} = \boldsymbol{\lambda}\boldsymbol{v}$, where $\boldsymbol{W}$ is the within-class covariance matrix, $\boldsymbol{B}$ is the between-class covariance matrix, $\boldsymbol{v}$ is an arbitrary vector, and $\boldsymbol{\lambda}$ is the diagonal matrix of eigen values [1].

5.3. Within class covariance normalization

In within class covariance normalization (WCCN) method, the feature vectors are transformed using a matrix which minimizes the upper bounds on the classification error metric and hence minimizes the classification error [14]. The transformation matrix $\boldsymbol{B}$ is obtained by Cholesky decomposition of the inverse of the within-class covariance matrix $\boldsymbol{W}$ as, $\boldsymbol{W}^{-1} = \boldsymbol{B}\boldsymbol{B}^{t}$.

6. Experimental Setup

The experiments are performed using the NIST 2003 SRE database. It contains speech data of 356 target speakers collected over cellular phone network. The evaluation of the system is done as per the NIST 2003 SRE evaluation plan for primary task [15]. This experimental setup contains 24981 trials for verification task including true and false trials. The standard MFCC feature vectors of 39-dimensions with cepstral mean and variance normalization are used. An energy based VAD is used for selecting the speech frames. The Switchboard Cellular Part 2 corpus is used as the development data for all the systems. A gender-independent UBM model of 1024 Gaussian mixtures created using approximately 10 hours of the development speech data is used for all the systems. The GMM supervectors are created by adapting only the mean parameters of the UBM using maximum *a posteriori* (MAP) approach with the speaker specific data. The total variability matrix of the i-vector based system and the dictionaries for the SRSV systems are created using 1872 speech utterances taken from the development database. θ of value 0.7 is used for learning the discriminative dictionary. The JFA is made up of 300 speaker factors and 100 channel factors without the residual factor. The LDA and WCCN matrices are created using the same development data which is used for learning the total variability matrix and the dictionaries. The LDA for the i-vector system uses 250 top dimensions where as the proposed SRC based system uses LDA of 375 top dimensions. All the above mentioned parameters are chosen out of experimentation. The performance of the SV systems are evaluated using the equal error rate (EER) and the minimum detection cost function (minDCF).

7. Experimental Results and Discussions

In this section, we first explain the tuning experiments done for the KSVD-SRSV system to get an optimal sparse representation. Followed by this we compare the SRSV and the i-vector based SV systems and then study the sparsity of representation in SRSV system with different kinds of dictionaries. It is followed by exploration of the total variability smoothing of supervectors in SRSV system. Motivated by recent usage of PLDA in i-vector based systems [16], a discriminatively learned dictionary based SRSV system is also presented for contrast purpose. The performance of various systems are evaluated with session/channel compensation and their score level fusion is also reported.

7.1. Tuning of the SRSV system

In a KSVD learned dictionary based SRSV system, there are three main tuning parameters : i) the number of atoms in the dictionary, ii) the number of atoms selected while learning the dictionary, iii) the number of atoms selected while representation of target data. The significance of the first parameter is obvious. For explaining the significance of the other two parameters, recall that the KSVD dictionary learning process involves two stages: the sparse representation of the development data and the dictionary updation. Unlike the sparse representation of the unseen training and test data in an SV system, the dictionary learning process involves the sparse representation of the seen development data. Due to this fact, there is a scope of tuning the other two parameters for optimal system performance. To be consistent with the i-vector dimension reported in literature, all KSVD based learned dictionaries are chosen to be of 400 atoms. The performances of the system obtained while tuning the other two parameters are shown in Figure 5. It is noted that the system with dictionary created with with the best performance is obtained with selection of 5 atoms while dictionary learning and 50 atoms while representation of training and testing supervectors. We have used these parameter values for KSVD based learned dictionary unless specified otherwise.

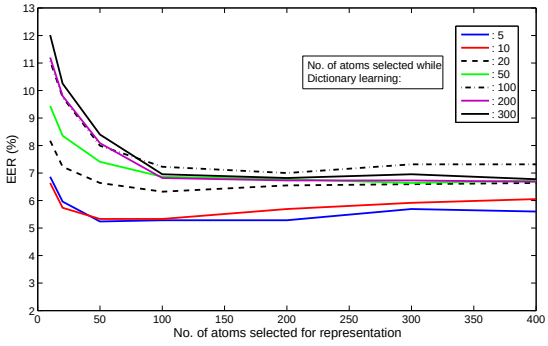

Figure 2: Figure showing the effect of number of atoms selected while dictionary learning and the number of atoms selected for representation for KSVD-SRSV system

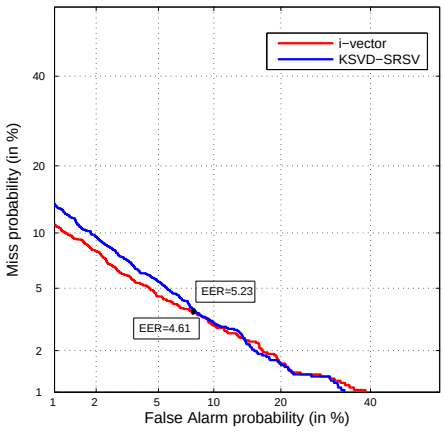

Figure 3: DET plots showing performances of i-vector based and KSVD-SRSV systems

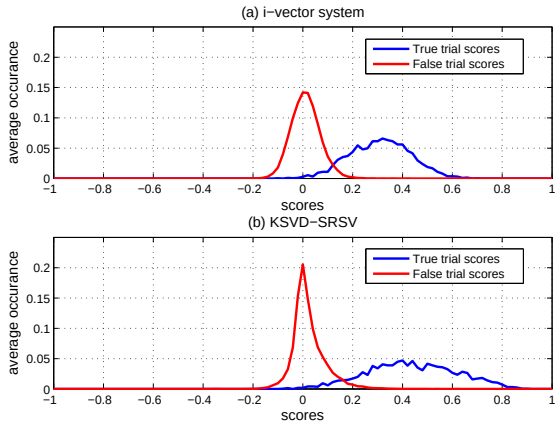

Figure 4: Histogram of true and false trial scores of (a) i-vector based and (b) KSVD-SRSV systems

7.2. Comparison of KSVD-SRSV system with i-vector based SV system

On comparing Eq. 1 and Eq. 10, it is obvious to note that the SRSV and the i-vector based systems are having some broad similarities. The dimensions of matrices $\boldsymbol{D}$ and $\boldsymbol{T}$ are the same and their respective projections $\boldsymbol{x}$ and $\boldsymbol{w}$ of a given centered GMM mean supervector are used for classification with the same scoring metric. Thus the main differences between the two approaches lie in the different criteria used for learning those matrices and the nature of the projections derived from them. The matrix $\boldsymbol{D}$ in the SRSV system is designed to be redundant having non-orthogonal columns and hence the projections with respect to that are sparse. In case of the i-vector based system, the columns of the matrix $\boldsymbol{T}$ are orthogonal to each other and as a result the projections with respect to that are generally non-sparse.

The performances of similar complexity i-vector based and KSVD-SRSV systems are shown in Figure 3 in the form of DET curves. It can be noted that the i-vector based system performs slightly better than the KSVD-SRSV system with an EER of 4.61 % against 5.2 %. The nature of the i-vector and the sparse vector representations are quite different. In case of the i-vector the energy is distributed to a large number of coefficients whereas for the sparse representation vector the energy is concentrated in a few coefficients only. So the sparse representation vectors are expected to be more sensitive to session/channel variability than the i-vectors and we hypothesis that this is the cause of the slightly inferior performance of the KSVD-SRSV system compared to the i-vector based system. To analyze the performance further, the histogram for the true and false trial scores of the KSVD-SRSV and i-vector based systems are plotted which is shown in Figure 4. On comparing the histograms, we note that the false trial scores are centered around zero for both the i-vector and KSVD-SRSV systems but, the spread is less in case of the KSVD-SRSV system. Although the mean of the true trial scores are more right shifted for KSVD-SRSV system compared to that of the i-vector system, but its spread is much more for the KSVD-SRSV case. As a result, the KSVD-SRSV system ended up giving poorer performance compared to the i-vector system.

7.3. Total variability dictionary for SRSV system

In this subsection we explore the use of the total variability matrix of the i-vector approach as a dictionary (*T-dictionary*)for the SRSV system. The performance of the SRSV system with 400 atoms T-dictionary is evaluated with varying number of atoms selected for representations. For comparison purpose, an SRSV system with KSVD dictionary of 400 atoms with varying number of atoms selected for representations is also evaluated. These performances are shown in Figure 5 along with the performance of a 400 dimension i-vector based system for contrast purpose. It can be noted that the SR-SV system with T-dictionary gives very poor performance when small number of atoms are selected (say 10 atoms). With increasing number of atoms, the performance significantly improves and for all 400 atoms selected, it matches that of the 400-dimension i-vector based system. On comparing with the KSVD-SRSV system, we note that the T-dictionary based SRSV system (T-SRSV) gives slightly better performance for more than 200 atoms selected, but for smaller number of atoms it is found to be significantly degraded. As reported in [8], for improved performance of an SRSV system with T-dictionary, a bigger size dictionary (i.e. more number of columns) and more number of atoms (selected for sparse representation) are to be used. This is counter intuitive to sparse representation and this aspect is explored in the following subsection.

7.3.1. Effect of size of T-dictionary in SRSV

To explore the effect of the size of the T-SRSV system, dictionaries of 200, 300, 400, 600, 700 and 900 columns are created. The T-SRSV systems with these dictionaries are evaluated for three different numbers of selected atoms viz. 50, 200 and 400. The performance of these systems are shown in Figure 6 along with that of the corresponding i-vector and KSVD-SRSV systems. For all the systems considered, the performance is found to degrade consistently for dictionary sizes beyond 400. For the T-SRSV system with higher number of atoms selected for representations, the performance is found to be improving and closely become comparable to that of the i-vector based system. It is interesting to note that for the three types of systems considered, the best performances corresponds to dictionary sizes between 300-400.

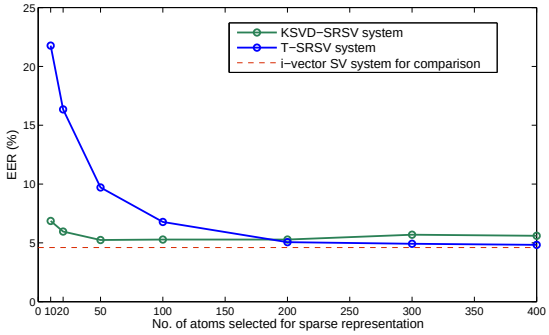

Figure 5: Figure showing the effect of number of atoms selected for sparse representation in various SRSV systems

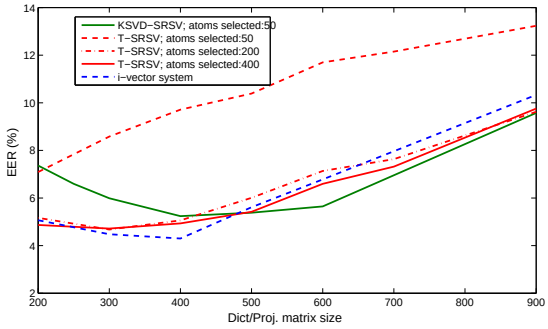

Figure 6: Figure showing the effect of dictionary size for various SRSV systems

7.4. SKSVD based SRSV system

In our previous work [7] we had explored a discriminative version of the KSVD algorithm, referred to as the SKSVD algorithm. The SKSVD algorithm is characterized by inclusion of discriminative term in the learning criterion. The SKSVD-SRSV system was found to result in significantly improved performance compared to the simple KSVD-SRSV system. In recent literature we also find the use of PLDA to improve the performance of the i-vector based system [16]. In this work also we have evaluated the performance the SKSVD-SRSV system which is given in the given in Table 1 along with that of the i-vector, KSVD-SRSV, and T-SRSV SV systems. Note that the performance of the T-SRSV system is evaluated with 200 atoms selected for sparse representation. It can be observed that the SKSVD-SRSV system hugely outperforms all other above mentioned systems. The main motivation of considering this system in this work is to evaluate the significance of the fusion of the i-vector and sparse representation based approaches explored in the next subsection.

7.5. T-smoothed supervector based SR-SV system

As already shown in Section 4 with help of a controlled experiment that T-smoothing (TS) of supervectors results in reduction of intra-speaker variations. To evaluate its effect on SV system, the performance of KSVD- and SKSVD-SRSV systems with TS applied to supervectors are evaluated and are also given in Table 1. It can be observed that with the inclusion of TS about 30 % relative improvement is obtained for both KSVD- and SKSVD-SRSV systems over their respective baselines.

Table 1: *Performances comparison of i-vector and various SRSV systems on NIST 2003 SRE dataset*

Systems	EER (%)	minDCF
i-vector	4.21	0.072
KSVD-SRSV	5.23	0.097
T-SRSV	5.05	0.088
SKSVD-SRSV	2.87	0.042
TS + KSVD-SRSV	3.70	0.065
TS + SKSVD-SRSV	1.96	0.031

Table 2: *Performances comparison of i-vector and various SRSV systems with appropriate session/channel compensation on NIST 2003 SRE dataset*

Sys. No.	Kind of system and session/channel compensation	EER (%)	min DCF
1	i-vector + LDA-WCCN	2.21	0.040
2	KSVD-SRSV + LDA-WCCN	3.61	0.065
3	T-SRSV + LDA-WCCN	3.43	0.063
4	SKSVD-SRSV + LDA-WCCN	1.98	0.036
5	TS + KSVD-SRSV + LDA-WCCN	2.53	0.045
6	TS + SKSVD-SRSV + LDA-WCCN	1.59	0.032
7	JFA + KSVD-SRSV	1.56	0.031
8	JFA + SKSVD-SRSV	1.53	0.031
9	Fusion: 1 + 2 + 3	2.26	0.042
10	Fusion: 1 + 2 + 3 + 5	2.08	0.038
11	Fusion: 1 + 2 + 3 + 5 + 7	1.63	0.027
12	Fusion: 4 + 6	1.27	0.022
13	Fusion: 4 + 6 + 8	0.99	0.018

7.6. Channel/session compensation

To explore the effectiveness of the proposed SV system in presence of the session/channel variability compensation, we have applied suitable methods among JFA, LDA and WCCN to different SV systems considered. The performance of various systems considered in this work with appropriate channel/session compensation technique applied are shown in the Table 2. For the i-vector based system, LDA+WCCN is applied as suggested in [1]. For the comparison purpose, the KSVD- and T-SRSV systems were applied with LDA+WCCN compensation. It can be noted that the channel/session compensation using LDA+WCCN has resulted in about 50 % improvement for i-vector SV system while about 30% improvement for KSVD- and T-SRSV systems over their uncompensated baselines. Similarly for SKSVD-SRSV system, the LDA+WCCN compensation also results in about 30% relative improvement over its uncompensated baseline. For TS based KSVD- and SKSVD-SRSV systems the LDA+WCCN compensation results in 30% and 19% relative improvement over its uncompensated baseline. In [7] we have already reported that with the JFA cleaning of supervectors prior to the sparse representation, the learned dictionary based SRSV systems are found to give largely improved performance. The performances of KSVD- and SKVD-SRSV systems with JFA cleaning are also evaluated for further gains in system combination and shown in the Table 2.

To explore the complementary information among the i-vector and SRSV systems in particularly including the T-

smoothing and JFA cleaning of supervectors, the score level combination of various systems are explored. On score level fusion of session/channel compensated i-vector, KSVD-SRSV and T-SRSV we do not see any further improvement. But, on adding the TS KSVD-SRSV to the above combination results in a small improvement which further gets improved to 1.16 % EER on inclusion of JFA cleaned KSVD-SRSV system. Similarly the fusion of SKSVD and T-smoothing based SKSVD does not result in any further improvement but with the inclusion of JFA cleaned SKSVD system to this combination, some significant improvement is noted.

8. Conclusions

In this work we have studied the various aspects of the SV systems using sparse representation over learned dictionaries. The work compares the similarity between the SRSV and i-vector based systems and study in detail the use of the T-matrix of the i-vector based system as a dictionary for the SRSV system. The study shows that the T-dictionary based SRSV system can give a comparable performance to that of the KSVD learned dictionary based SRSV system but with significantly more atoms selected for the representation of the target supervectors. From the experiments conducted to tune the KSVD-SRSV system, it is noted that in the sparse representation stage of dictionary learning, very small number of atoms are to be selected whereas, in sparse representation of target supervectors, a relatively larger number of atoms need to be selected. We have also explored the use of T matrix for smoothing of supervectors prior to sparse representation. The performance of the discriminative dictionary based SKSVD-SRSV system is also evaluated which turns to be the single best system among different types of SV systems considered in this work. The score level fusion of the three SRSV systems incorporating the discriminative dictionary resulted in a performance of 0.99% EER.

9. Acknowledgement

This work has been supported by the ongoing project grant No. 12(4)/2009-ESD sponsored by the Department of Information Technology, Government of India.

10. References

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

[2] I. Naseem, R. Togneri, and M. Bennamoun, "Sparse representation for speaker identification," in *Proc. International Conference on Pattern Recognition*, 2010, pp. 4460–4463.

[3] J. M. K. Kua, E. Ambikairajah, J. Epps, and R. Togneri, "Speaker verification using sparse representation classification," in *ICASSP 2011*, may 2011, pp. 4548–4551.

[4] M. Li, X. Zhang, Y. Yan, and S. Narayanan, "Speaker verification using sparse representations on total variability i-vectors," in *Interspeech 2011*, may 2011, pp. 4548–4551.

[5] Haris B C and R. Sinha, "Exploring sparse representation classification for speaker verification in realistic environment," in *Proc. Centenary Conference, Electrical Engineering, Indian Institute of Science*, 2011.

[6] ——, "Speaker verification using sparse representation over KSVD learned dictionary," in *Proc. 18th National Conference on Communications 2012 (to appear)*, Feb. 2012.

[7] ——, "Sparse representation over learned and discriminatively learned dictionaries for speaker verification," in *Proc. ICASSP 2012 (accepted for)*, March. 2012.

[8] M. Li, C. Lu, A. Wang, and S. Narayanan, "Speaker verification using lasso based sparse total variability supervector and probabilistic linear discriminant analysis," in *NIST Speaker Recognition Workshop, Atlanta*, 2011.

[9] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[10] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, nov. 2006.

[11] F. Rodriguez and G. Sapiro, "Sparse representations for image classification: Learning discriminative and reconstructive non-parametric dictionaries," University of Minnesota, Tech. Rep., December 2007.

[12] Haris B C and R. Sinha, "Sparse representation of total variability smoothed GMM mean supervectors for speaker verification," in *(Communicated to) International Conference on Signal Processing and Communications (SPCOM), 2012*.

[13] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in gmm-based speaker verification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1448 –1460, may 2007.

[14] A. O. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for svm-based speaker recognition," in *Proc. of ICSLP*, 2006, p. 14711474.

[15] NIST 2003 Speaker Recognition Evaluation Plan, www.itl.nist.gov/iad/ mig/tests/sre/2003/2003-spkrec-evalplan-v2.2.pdf.

[16] M. Senoussaoui, P. Kenny, N. Brummer, E. de Villiers, and P. Dumouchel, "Mixture of plda models in i-vector space for gender-independent speaker recognition," in *Interspeech 2011*, May 2011.