Speaker Verification using Lasso based Sparse Total Variability Supervector with PLDA modeling

Ming Li*, Charley Lu†, Anne Wang† and Shrikanth Narayanan*
* Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, USA
E-mail: mingli@usc.edu, shri@sipi.usc.edu
† 3M Cogent, Inc
E-mail: charleylu@cogentsystems.com, annewang@cogentsystems.com

***Abstract*—In this paper, we propose a Lasso based framework to generate the sparse total variability supervectors (s-vectors). Rather than the factor analysis framework, which uses a low dimensional Eigenvoice subspace to represent the mean supervector, the proposed Lasso approach utilizes the l^1 norm regularized least square estimation to project the mean supervector on a pre-defined dictionary. The number of samples in this dictionary is appreciably larger than the typical Eigenvoice rank but the l^1 norm of the Lasso solution vector is constrained. Only a small number of samples in the dictionary are selected for representing the mean supervector, and most of the dictionary coefficients in the Lasso solution are 0. We denote these sparse dictionary coefficient vectors in the Lasso solutions as the s-vectors and model them using probabilistic linear discriminant analysis (PLDA) for speaker verification. The proposed approach generates comparable results to the conventional cosine distance scoring based i-vector system and improvement is achieved by fusing the proposed method with either the i-vector system or the joint factor analysis (JFA) system. Experiments results are reported on the female part of the NIST SRE 2010 task with common condition 5 using equal error rate (EER), norm old minDCF and norm new minDCF values. The norm new minDCF cost was reduced by 7.5% and 9.6% relative when fusing the proposed approach with the baseline JFA and i-vector systems, respectively. Similarly, 8.3% and 10.7% relative norm old minDCF cost reduction was observed in the fusion.** [1]

I. Introduction

The use of joint factor analysis (JFA) [1], [2], [3] has contributed to state of the art performance in text independent speaker verification and hence is being widely used. It is a powerful technique for compensating the variability caused by different channels and sessions.

Recently, total variability i-vector modeling has gained significant attention due to its excellent performance, low complexity and small model size [4]. In this modeling, first, a single factor analysis is used as a front end to generate a low dimensional total variability space which models both the speaker and channel variabilities [4]. Then, within this total variability vector space, channel variability compensation methods, such as Within-Class Covariance Normalization (WCCN) [5], Linear Discriminative analysis (LDA) and Nuisance Attribute Projection (NAP) [6], are performed to reduce the channel variability. Finally, two classification approaches, namely support vector machine (SVM) and cosine distance scoring (CDS), are proposed for the verification task [4]. It is also shown in [4] that LDA followed by WCCN achieved the best performance.

More recently, a sparse representation computed by l^1-minimization (to approximate the l^0-minimization) with equality constraints was proposed to replace the SVM in the GMM mean supervector modeling and by fusing the sparse representation based classification (SRC) method with SVM, the overall system performance was improved [7], [8]. This approach was extended in our previous work [9], [10] to handle the robust verification task against large session variabilities. First, the sparse representation is computed by l^1-minimization with quadratic constraints rather than equality constraints. Second, by adding a redundant identity matrix at the end of the original over-complete dictionary, the sparse representation is made more robust to variability and noise. Third, both the l^1 norm ratio and the background normalized (BNorm) l^2 residual ratio are used and shown to outperform the conventional l^2 residual ratio in the speaker verification task. In [10], SRC is employed to perform classification on the total variability i-vectors. Since the dimensionality of i-vectors is small, it requires fewer samples to construct the over-complete dictionary and the SRC approach becomes more efficient.

In the aforementioned approaches [7], [8], [9], [10], the sparse representation framework was used just as a kind of classification approach on various GMM supervectors. Since sparse representation solution needs to be calculated for every trial, it is computationally expensive for high dimensional supervectors and sometimes intractable to perform score normalization (ZT-norm). Therefore it is more efficient to utilize SRC to model the low dimensional supervectors, such as i-vectors and JFA speaker factors, rather than the mean supervectors. However, factor analysis based Eigenvoice modeling and sparse representation are generally similar in terms of projecting the supervector into a dictionary. The dictionary of Eigenvoice modeling is a low dimensional subspace which makes the factor vector low dimensional while the dictionary of SRC is over-complete which results in a sparse coefficient vector. Thus, this analogy motivates us to explore the sparse representation as a kind of front end representation framework which is similar to the factor analysis based Eigenvoice modeling in the i-vector modeling approach. In this case, the benefits are as follows. First, computing the sparse representation solution is required only once for each testing utterance which makes the score normalization efficient. Second, there is no need to use over-complete dictionary since it is adopted as a front end representation framework rather than the classification approach. Therefore, it can be performed on the high dimensional GMM mean supervectors.

In this work, we employ the Lasso based l^1 norm regularized weighted least square estimation to map the centered 1st order statistics vector on the UBM into a sparse factor vector which is denoted as sparse total variability supervector (s-vector). Although the number of elements in the dictionary is large, this representation only selects some of the dictionary elements to represent the mean supervector due to the l^1 constraint. Therefore, the selected dictionary elements are more likely to be more informative than

[1] This work was presented at 2011 NIST Speaker Recognition Workshop as a presentation without any official publication.

others. We applied the principal component analysis (PCA) on the centered 1st order statistics vector on the UBM and used the first 3000 eigenvectors (corresponding to the largest 3000 eigenvalues) to construct the dictionary. This preserves most of the energy and make the dictionary size and s-vector dimensionality small which is efficient for Lasso calculation. Furthermore, we can reuse the PCA training data for LDA, WCCN, PLDA and score normalization. Because if we construct the dictionary using raw mean supervector data samples, the Lasso solution for the same sample on the dictionary is a Kronecker delta vector. The speaker and channel variability information is encoded into the non-zero entries of this s-vector which servers as the feature vector in the subsequent probabilistic linear discriminant analysis (PLDA) modeling. Compared to i-vectors, the proposed s-vectors use information from lager subspace; but without performing matrix inversion, the Lasso solution is just an approximation under the l^1 norm constraint. So, we assume the proposed s-vectors carry complementary information to the i-vectors.

PLDA has been recently introduced to the speaker verification task to model the i-vectors and has demonstrated excellent performance in [11], [12], [13]. PLDA incorporates both within-speaker and between-speaker variations into modeling which is adopted to model the s-vectors in this system.

II. METHODS

A. Total variability i-vectors and cosine kernel modeling

In the total variability space, there is no distinction between the speaker effects and the channel effects. Rather than using the eigenvoice matrix V and the eigenchannel matrix U [1], the total variability space contains the speaker and channel variabilities simultaneously [4]. Given a C component GMM UBM model λ with $\lambda_c = \{p_c, \boldsymbol{\mu_c}, \boldsymbol{\Sigma_c}\}, c = 1, \cdots, C$ and an utterance with a L frame feature sequence $\{\boldsymbol{y_1}, \boldsymbol{y_2}, \cdots, \boldsymbol{y_L}\}$, the 0^{th} and centered 1^{th} order Baum-Welch statistics on the UBM are calculated as follows:

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

$$\boldsymbol{F_c} = \sum_{t=1}^{L} P(c|\boldsymbol{y_t}, \lambda)(\boldsymbol{y_t} - \boldsymbol{\mu_c}) \tag{2}$$

where $c = 1, \cdots, C$ is the GMM component index and $P(c|\boldsymbol{y_t}, \lambda)$ is the occupancy posterior probability for $\boldsymbol{y_t}$ on λ_c. The corresponding centered mean supervector $\tilde{\boldsymbol{F}}$ is generated by concatenating all the $\tilde{\boldsymbol{F_c}}$ together:

$$\tilde{\boldsymbol{F_c}} = \frac{\sum_{t=1}^{L} P(c|\boldsymbol{y_t}, \lambda)(\boldsymbol{y_t} - \boldsymbol{\mu_c})}{\sum_{t=1}^{L} P(c|\boldsymbol{y_t}, \lambda)}. \tag{3}$$

The speaker and channel dependent centered GMM mean supervector $\tilde{\boldsymbol{F}}$ can be written as follows:

$$\tilde{\boldsymbol{F}} = \boldsymbol{T}\boldsymbol{w}, \tag{4}$$

where $\boldsymbol{T}$ is a rectangular total variability matrix of low rank and $\boldsymbol{w}$ is the so-called i-vector [4]. Considering a C-component GMM and F dimensional acoustic features, the total variability matrix $\boldsymbol{T}$ is a $CF \times K$ matrix which can be estimated the same way as learning the eigenvoice matrix $\boldsymbol{V}$ in [14] except that here we consider that every utterance is produced by a new speaker [4].

Given the centered mean supervector $\tilde{\boldsymbol{F}}$ and total variability matrix $\boldsymbol{T}$, the i-vector is computed as follows [4]:

$$\boldsymbol{w} = (\boldsymbol{I} + \boldsymbol{T^t \Sigma^{-1} N T})^{-1} \boldsymbol{T^t \Sigma^{-1} N \tilde{F}} \tag{5}$$

where $\boldsymbol{N}$ is a diagonal matrix of dimension $CF \times CF$ whose diagonal blocks are $N_c\boldsymbol{I}, c = 1, \cdots, C$ and $\boldsymbol{\Sigma}$ is a diagonal covariance matrix of dimension $CF \times CF$ estimated in the factor analysis training step. It models the residual variability not captured by the total variability matrix $\boldsymbol{T}$ [4]. In our implementation, we only explored the 1^{th} order statistics and this $\boldsymbol{\Sigma}$ is the concatenated version of $\boldsymbol{\Sigma_c}$.

In this total variability space, two channel compensation methods, namely Linear Discriminant Analysis (LDA) and Within Class Covariance Normalization (WCCN) [5], are applied to reduce the variabilities. LDA attempts to transform the axes to minimize the intra-class variance due to the channel effects and maximize the variance between speakers while WCCN uses the inverse of the within-class covariance to normalize the cosine kernel. After LDA and WCCN steps, cosine distance scoring is used for i-vector modeling. The cosine kernel between two i-vectors $\boldsymbol{w_1}$ and $\boldsymbol{w_2}$ is defined as follows:

$$k(\boldsymbol{w_1}, \boldsymbol{w_2}) = \frac{< \boldsymbol{w_1}, \boldsymbol{w_2} >}{\|\boldsymbol{w_1}\|_2 \|\boldsymbol{w_2}\|_2} \tag{6}$$

B. Sparse total variability supervector extraction by Lasso

In the i-vector extraction equation (5), the identity matrix is the prior of the i-vector $\boldsymbol{w}$. The Maximum Likelihood (ML) solution is:

$$\boldsymbol{w} = (\boldsymbol{T^t \Sigma^{-1} N T})^{-1} \boldsymbol{T^t \Sigma^{-1} N \tilde{F}} \tag{7}$$

which is a weighted least square solution of equation (4). We define the normalized total variability matrix and normalized centered mean supervector as $\acute{\boldsymbol{T}}$ and $\acute{\boldsymbol{F}}$,

$$\acute{\boldsymbol{F}} = \tilde{\boldsymbol{F}} \boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{N}^{\frac{1}{2}} \tag{8}$$

$$\acute{\boldsymbol{T^k}} = \boldsymbol{T^k} \boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{N}^{\frac{1}{2}}, k = 1, \cdots, K, \tag{9}$$

where $\boldsymbol{T^k}$ is the k^{th} column in the matrix $\boldsymbol{T}$. Then equation (7) can be rewritten as a standard least square estimation

$$\boldsymbol{w} = (\acute{\boldsymbol{T}}^{\boldsymbol{t}} \acute{\boldsymbol{T}})^{-1} \acute{\boldsymbol{T}}^{\boldsymbol{t}} \acute{\boldsymbol{F}}. \tag{10}$$

In our Lasso based sparse total variability supervector extraction, the dictionary $\boldsymbol{T}$ is constructed by PCA on the centered GMM mean supervectors. Suppose we have D utterances to train the PCA, each column of the $CF \times D$ data matrix $\boldsymbol{A}$ is a centered GMM mean supervector. If $CF < D$, eigen decomposition can be performed directly on $(\boldsymbol{A}\boldsymbol{A^t})/D$ and $\boldsymbol{T}$ is the eigenvectors corresponding to the largest K eigenvalues. While if $CF > D$, we do the eigen decomposition on $(\boldsymbol{A^t}\boldsymbol{A})/D$ to generate eigenvectors matrix $\boldsymbol{V}$ and $\boldsymbol{T} = \boldsymbol{A}\boldsymbol{V}$.

Given the centered GMM mean supervector $\tilde{\boldsymbol{F}}$ from an input utterance and the total variability matrix $\boldsymbol{T}$ generated by PCA, we first normalize them into $\acute{\boldsymbol{F}}$ and $\acute{\boldsymbol{T}}$ by equations (8) and (9). Then the Lasso based l^1 norm regularized least square estimation is performed to calculate the s-vector $\acute{\boldsymbol{w}}$:

$$\min \qquad \|\acute{\boldsymbol{F}} - \acute{\boldsymbol{T}}\acute{\boldsymbol{w}}\|_2^2 \qquad \text{subject to} \qquad \|\acute{\boldsymbol{w}}\|^1 < \tau. \tag{11}$$

If we convert it back to the pre-normalization forms, equation (11) becomes minimizing the upper bound of KL divergence used to derive the GMM mean supervector kernel in [15]:

$$\min \sum_{c=1}^{C} N_c (\tilde{\boldsymbol{F}}_{\boldsymbol{c}} - \boldsymbol{T_c}\acute{\boldsymbol{w}}_{\boldsymbol{c}})^t {\boldsymbol{\Sigma_c}}^{-1} (\tilde{\boldsymbol{F}}_{\boldsymbol{c}} - \boldsymbol{T_c}\acute{\boldsymbol{w}}_{\boldsymbol{c}}) \text{subject to} \|\acute{\boldsymbol{w}}\|^1 < \tau. \tag{12}$$

From Fig.1, we can observe that the first 1000 largest eigenvalues cover majority of the total energy. The summation of the largest 3000 eigenvalues is around 72% of the total sum. Although the dictionary size K in the proposed s-vector framework is significantly larger than in the factor analysis approach, the l^1 norm constraint guarantee the s-vector to be sparse. As shown in Fig. 2, the l^0 norm of this s-vector is 753 which means only 753 coefficients are

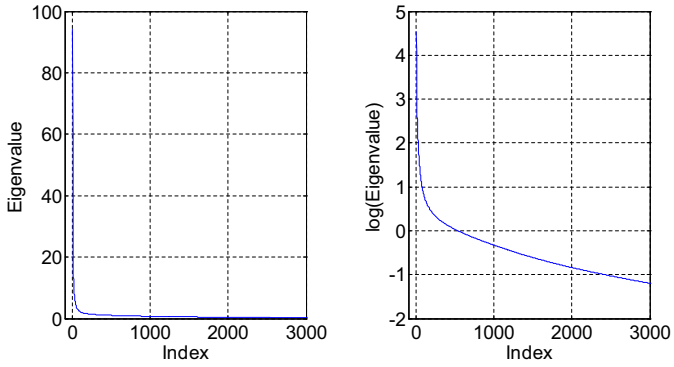

Fig. 1. The PCA eigenvalues of the centered GMM mean supervector space

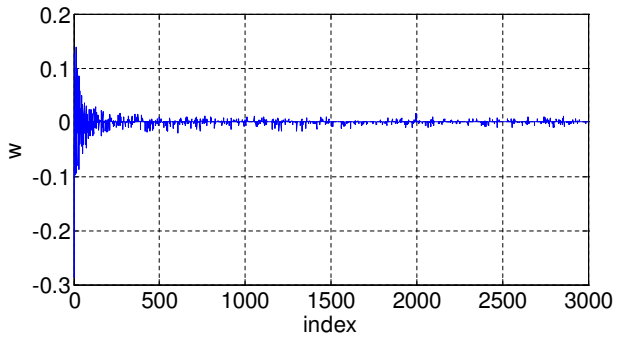

Fig. 2. S-vector of utterance fzzhwB with $\tau = 6$. $\|\acute{\boldsymbol{w}}\|_0 = 753$, $\|\acute{\boldsymbol{w}}\|_1 = 6$

non-zero in the 3000 dimensional s-vector. Generally, the selected dictionary elements are different for each utterance. Since every column of $\acute{\boldsymbol{F}}$ and $\acute{\boldsymbol{T}}$ is normalized to unit l^2 norm before Lasso is performed, the dimensions with larger eigenvalues tend to have higher coefficients in the s-vector which matches the case in Fig. 1.

C. Probabilistic linear discriminant analysis modeling

We assume that the training data consists of J utterances from I speakers and denote the j^{th} s-vector of the i^{th} speaker by $\boldsymbol{x_{ij}}$. We assume that the data are generated in the following way [16]:

$$\boldsymbol{x_{ij}} = \boldsymbol{\mu} + \boldsymbol{U}\boldsymbol{h_i} + \boldsymbol{G}\boldsymbol{w_{ij}} + \boldsymbol{\epsilon_{ij}}, \tag{13}$$

where the speaker term $\boldsymbol{\mu} + \boldsymbol{U}\boldsymbol{h_i}$ is only dependent on the speaker index and the variability term $\boldsymbol{G}\boldsymbol{w_{ij}} + \boldsymbol{\epsilon_{ij}}$ is different for every s-vector and used to model the within-speaker variances. The model parameters are estimated by employing Expectation Maximization (EM) algorithms on the training data. Given a pair of s-vectors $S(\boldsymbol{w_i}, \boldsymbol{w_j})$ for testing, the log likelihood ratio is computed based on a hypothesis testing $P(S|H_1)/P(S|H_0)$ where H_1 means it is a true trial and H_0 denotes a false trial [16]. Since the scoring is symmetric for the target and test s-vectors, symmetric normalization (Snorm) [17] is performed as the score normalization approach. The PLDA implementation is based on the UCL toolkit [16].

III. Experimental results

A. Corpus and baseline systems

We performed experiments on the NIST 2010 speaker recognition evaluation (SRE) corpus [18]. Our focus is the female part of the common condition 5 (a subset of tel-tel) in the core task. We used equal error rate (EER), the normalized old minimum decision cost value (norm old minDCF) and norm new minDCF as the metrics for evaluation [18].

TABLE I
CORPORA USED TO ESTIMATE THE UBM, TOTAL VARIABILITY MATRIX, JFA FACTOR LOADING MATRIX, WCCN, LDA, PLDA AND THE NORMALIZATION DATA FOR NIST 2010 TASK CONDITION 5.

	Switchboard	NIST04	NIST05	NIST06	NIST08
UBM		√	√		
T	√	√	√	√	√
JFA V	√				
JFA U		√	√	√	√
JFA D		√			
WCCN		√	√	√	√
LDA		√	√	√	√
PLDA		√	√	√	√
Znorm		√	√		
Snorm					√
Tnorm				√	

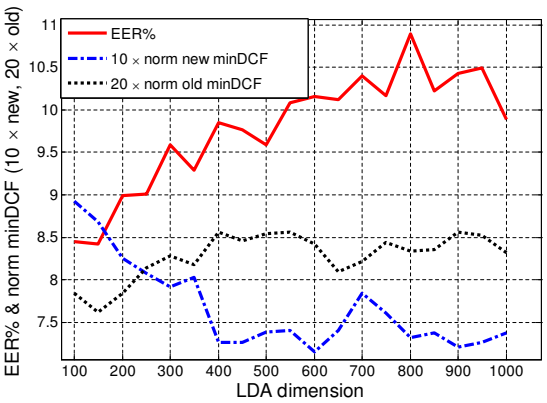

Fig. 3. Performance of the s-vector system using LDA and cosine distance raw scoring without Snorm

For cepstral feature extraction, a 25ms Hamming window with 10ms shifts was adopted. Each utterance was converted into a sequence of 36-dimensional feature vectors, each consisting of 18 MFCC coefficients and their first derivatives. We employed a Czech phoneme recognizer [19] to perform the voice activity detection (VAD) by simply dropping all frames that are decoded as silence or speaker noises. Feature warping is applied to mitigate channel effects.

The training data for NIST 2010 task included Switchboard II part1 to part3, NIST SRE 2004, 2005, 2006 and 2008 corpora on the telephone channel. The description of the dataset used in each step is provided in Table I. The gender-dependent GMM UBMs consist of 1024 mixture components, which were trained using EM with the data from NIST SRE 04 and 05 corpus. We used all of the training data for estimating the total variability space. The NIST SRE 2004, 2005, 2006 and 2008 datasets were used for training WCCN, LDA and PLDA matrix, and a data set chosen from SRE 2006 corpus was used for Tnorm score normalization, including 1325 female utterances. 256 female utterances from NIST 2008 were adopted as Snorm data.

The JFA baseline system is trained using the BUT toolkit [20] and linear common channel point estimate scoring [21] is adopted. The speaker factor size and channel factor size is 300 and 100, respectively. ZTnorm was applied on JFA subsystem while Snorm was employed in i-vector subsystem.

B. Results and discussion

Performance of the s-vector system using LDA with cosine distance raw scoring and PLDA modeling are shown in Fig.3 and Table II, respectively. Compared to the raw scoring (EER 15.23% in Table II), applying LDA on top of the s-vectors significantly

TABLE II
PERFORMANCE OF THE S-VECTOR SYSTEM USING PLDA MODELING (#EM DENOTES THE EM ITERATION NUMBER FOR PLDA TRAINING)

τ	LDA	WCCN	PLDA			Snorm	EER%	norm minDCF	
			U	G	#EM			new	old
6	×	×	×	×	×	×	15.23	0.96	0.65
6	600	×	×	×	×	×	10.16	0.72	0.42
6	600	√	×	×	×	×	11.79	0.71	0.44
6	600	×	100	100	10	×	11.22	0.97	0.55
6	100	×	100	100	10	×	8.97	0.85	0.43
6	×	×	100	100	10	×	11.02	0.90	0.52
6	×	×	100	100	20	×	8.44	0.83	0.42
6	×	×	100	100	20	√	4.80	0.62	0.27
8	×	×	100	100	20	√	4.86	0.65	0.26
6	×	×	150	50	20	×	8.73	0.75	0.41
6	150	×	150	50	20	√	7.61	0.85	0.34
6	×	×	**150**	**50**	**20**	√	**4.83**	**0.55**	**0.28**

TABLE III
PERFORMANCE OF THE S-VECTOR SYSTEM WHEN FUSING WITH THE JFA AND I-VECTOR BASELINE SYSTEMS

ID	Systems	EER(%)	norm minDCF	
			new	old
1	JFA linear scoring ZTnorm	3.62	0.41	0.193
2	I-vector LDA WCCN Cosine Snorm	5.04	0.52	0.241
3	S-vector PLDA Snorm	4.83	0.55	0.277
4	Fusion JFA + S-vector	3.37	0.38	0.177
5	Fusion I-vector + S-vector	4.14	0.47	0.215
6	Fusion JFA + I-vector + S-vector	3.09	0.37	0.157

improved the performance which might be because that majority of s-vector coefficients are zero. Furthermore, both EER and norm old minDCF cost continue to reduce by decreasing the LDA dimensionality while the norm new minDCF cost achieved the best result at 600. Adding WCCN on top of 600 dimensional LDA did not help which may suggest nonlinear session variabilities. Therefore, PLDA was applied to replace LDA and WCCN. PLDA modeling improved the system performance with small rank submatrixes (U and G). This matches the result in [11] that 90 rank U and G achieved the best performance. The Snorm score normalization achieved big improvements on both the EER and minDCF cost values. The best result was observed by using 150 eigenvoices and 50 eigen-channels which matches the parameter setting (300 eigen-voices and 100 eigen-channels) in the JFA framework. Furthermore, the proposed s-vector system is not very sensitive to the constraint τ values in equation (11). Larger τ can loose the l^1 norm constraint which results in a more accurate least square solution. However, a large norm constraint also slows the Lasso computation and may violate the sparse assumption of s-vector $\acute{\boldsymbol{w}}$. With a large τ, the proposed Lasso becomes the standard least squares estimator. On the other hand, setting τ to be very small increases the residue between $\acute{\boldsymbol{F}}$ and $\boldsymbol{TW}$ which may lead to non accurate s-vectors. Thus, a balanced τ is preferred.

It is shown in Table III that the proposed s-vector PLDA approach achieved comparable results to the conventionali-vector method. By fusing the s-vector system with JFA and ivector systems, the overall performance was enhanced by 7.5% - 10.7% relative EER and minDCF reduction. This supports our claim that the s-vector modeling is complementary to the conventional factor analysis framework based methods.

IV. CONCLUSIONS

We propose a l^1 norm regularized least square based Lasso approach to generate the sparse total variability supervectors (s-vectors) and use probabilistic linear discriminant analysis to model the s-vectors. Rather than the Eigenvoice modeling approach that projects the mean supervector into a low dimensional subspace, the proposed Lasso framework maps the mean supervector into a larger dimensional s-vector with l^1 norm constraint. The proposed approach generates comparable results to the conventional cosine distance scoring based i-vector system and improvement is achieved by fusing the proposed method with either the i-vector system or the joint factor analysis (JFA) system.

REFERENCES

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

[2] P. Kenny, G. Boulianne, P. Dumouchel, and P. Ouellet, "Speaker and Session Variability in GMM-Based Speaker Verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1448–1460, 2007.

[3] 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*, vol. 16, no. 5, pp. 980–988, 2008.

[4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, no. 99, pp. 1–1, 2010.

[5] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. Interspeech*, vol. 4, no. 2.2, 2006.

[6] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. ICASSP*, vol. 1, 2006, pp. 97–100.

[7] I. Naseem, R. Togneri, and M. Bennamoun, "Sparse Representation for Speaker Identification," in *Proc. ICPR*, 2010, p. 4460.

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

[9] M. Li and S. Narayanan, "Robust talking face video verification using joint factor analysis and sparse representation on GMM mean shifted supervectors," in *Proc. ICASSP*, 2011, pp. 1481–1484.

[10] M. Li, X. Zhang, Y. Yan, and S. Narayanan, "Speaker verification using sparse representations on total variability i-vectors," in *Proc. Interspeech*, 2011.

[11] P. Matejka, O. Glembek, F. Castaldo, M. Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocky, "Full-covariance ubm and heavy-tailed plda in i-vector speaker verification," in *Proc. ICASSP*, 2011, pp. 4828–4831.

[12] N. Dehak, Z. Karam, D. Reynolds, R. Dehak, W. Campbell, and J. Glass, "A channel-blind system for speaker verification," in *Proc. ICASSP*, 2011, pp. 4536–4539.

[13] 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," 2011.

[14] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 3, pp. 345–354, 2005.

[15] W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using gmm supervectors for speaker verification," *Signal Processing Letters, IEEE*, vol. 13, no. 5, pp. 308–311, 2006.

[16] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. ICCV*, 2007, pp. 1–8.

[17] N. Brümmer and A. Strasheim, "Agnitios speaker recognition system for evalita 2009," 2009.

[18] "The NIST Year 2010 Speaker Recognition Evaluation Plan," *http://www.itl.nist.gov/iad/mig/tests/spk/2010/index.html*.

[19] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical structures of neural networks for phoneme," in *Proc. ICASSP*, 2006, pp. 325–328, software available at http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context.

[20] L. Burget, M. Fapšo, and V. Hubeika, "But system description: Nist sre 2008," in *Proc. 2008 NIST Speaker Recognition Evaluation Workshop*, 2008, pp. 1–4, software available at http://speech.fit.vutbr.cz/software/joint-factor-analysis-matlab-demo.

[21] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny, "Comparison of scoring methods used in speaker recognition with joint factor analysis," in *Proc. ICASSP*, 2009, pp. 4057–4060.