Exploring Sparse Representation Classification for Speaker Verification in Realistic Environment

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*—We address the problem of speaker verification (SV) by exploiting discriminative classification ability of the sparse representation. The proposed sparse representation based speaker verification (SR-SV) system uses dictionary created with the mean supervectors derived from adapted GMMs. For classification purpose, the sparse coefficients obtained by the l^1 minimization are used employing different scoring methods. The SV systems are developed using the speech data collected in realistic environments with multiple sensors. On comparing with standard 1024 mixture GMM-UBM system, we find that 128 mixture GMM based SR-SV system performs better for all of the four sensors data considered.**

***Index Terms*—speaker verification, sparse representation, GMM mean supervectors.**

I. Introduction

The goal of the speaker verification (SV) algorithms is to authenticate persons from their voice samples. Its main applications are in person authentication and in forensic science. The state-of-the-art SV systems are based on either the adapted Gaussian mixture models with universal background model (GMM-UBM) [1] or the support vector machines (SVM) over the GMM mean supervectors [2] for modeling the speakers. Mel-frequency cepstral coefficients (MFCC) are the most commonly used features and are also combined with supra-segmental informations such as prosody and speaking style for improved performance.

The recent works in this area are concentrated toward addressing the problem of mismatch in sensors, environment and language and changes across sessions. Different approaches developed to address this problem include various techniques like score/handset/test normalization [3], feature mapping [4], nuisance attribute projection [5], speaker model synthesis [6], and joint factor analysis (JFA) [7].

In last few years, there is a lot of interest generated about *sparse representation* and *compressive sensing* which provide a new directions to signal processing research. Recently the discriminative abilities of the sparse representation have also been exploited in various areas of the pattern recognition [8], [9]. In sparse representation, a target vector is expressed as a sparse linear combination of columns of an overcomplete matrix representing the target signal space. The overcomplete matrix is commonly referred to as the 'dictionary' and its columns as 'atoms' in the sparse representation literature. The dictionary can be created broadly in two ways either by the use of model of the data (e.g., wavelets, curvelets etc.) or by the use of data-driven approaches like k-SVD. A detailed survey of dictionary creation for sparse presentation can be found in [10]. The sparse solution to the linear system of equations formed by the dictionary and the target vector can be obtained by the convex optimization using l_1-norm minimization. The greedy search methods like matching pursuits or LASSO (least absolute shrinkage and selection operator) are used for this purpose. For classification, the dictionary consists of training examples from all classes. The classification is then performed by comparing the norm of sparse coefficients or the sparse representation error across different classes. Sparse representation based classification using GMM supervectors has recently been explored for speaker identification [11] and face video verification [12]. In this work, we explore the sparse representation based classification for speaker verification task on a database collected in realistic environments.

The organization of the paper is as follows: In the Section II, we review the recent works in the area of sparse representation based classification and describe implementation details of the speaker verification system. The details of the database and the experimental setup are given in Section III followed by the discussion of result of the proposed speaker verification method in Section IV. The paper is concluded in Section V.

II. Classification based on Sparse Representation

In this section we first briefly review the sparse representation based classification approach as proposed for the face identification task in [8] and its adaptation for face video verification task in [12]. It is followed by description of our proposed method for speaker verification.

Assume that we have k distinct classes and n_i examples available for training the i^{th} class. Let the m-dimensional feature vector be designated as $\mathbf{v}_{ij} \in R^m$, where i is the index of the class, $i = 1, 2, \ldots, k$ and j is the index of the training example, $j = 1, 2, \ldots, n_i$. All training examples of the i^{th} class are combined to form a matrix $A_i = [\mathbf{v}_{i1}, \mathbf{v}_{i2}, \ldots, \mathbf{v}_{in_i}] \in R^{m \times n_i}$. It is assumed that an example $\mathbf{y}$ belonging to i^{th} class can be approximated as:

$$\mathbf{y} = \alpha_{i1}\mathbf{v}_{i1} + \alpha_{i2}\mathbf{v}_{i2} + \cdots + \alpha_{in_i}\mathbf{v}_{in_i} \quad (1)$$

where α_{ij} are the real scalar quantities. For the classification purpose, an overcomplete dictionary is formed by concatenat-

ing all class matrices as follows:

$$\mathbf{A} = [A_1, A_2, \ldots, A_k] \in R^{m \times n}, \quad n = \sum_{i=1}^{k} n_i \qquad (2)$$

Now $\mathbf{y}$ is represented as linear combination of all n columns of the overcomplete dictionary $\mathbf{A}$:

$$\mathbf{y} = \mathbf{A}\mathbf{x} \qquad (3)$$

where $\mathbf{x} \in R^n$ is a vector of unknown coefficients. The solution to these coefficients is found using l_1-norm minimization [13]

$$(l^1): \quad \hat{\mathbf{x}} = \arg\min \|\mathbf{x}\|_1 \quad \text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{y} \qquad (4)$$

With the assumption given by Equation (II), $\hat{\mathbf{x}}$ will be sparse and ideally all nonzero coefficients should correspond to the atoms from the class of $\mathbf{y}$ only. But in practice the atoms other than those of the class of $\mathbf{y}$ also have nonzero coefficients. As a result the identification is done by finding the representation errors corresponding to each of the class matrix A_i.

$$\min_i \ \|\mathbf{y} - \mathbf{A}\delta_i(\hat{\mathbf{x}})\|_2 \qquad (5)$$

where $\delta_i(\mathbf{x})$ is a new vector whose only nonzero entries are the entries in $\mathbf{x}$ that are associated with the class i. Following the above mentioned procedure, the speaker identification is also reported using the GMM mean supervector as the feature vector on TIMIT database [11].

In addition to this, the GMM mean supervector based sparse classification for the face video verification is also explored [12]. In this work, the dictionary for verifying each claim is created as,

$$\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2] \qquad (6)$$

where $\mathbf{A}_1$ and $\mathbf{A}_2$ are formed by the target training supervectors and a set of fixed non-target background training supervectors, respectively. For verification purpose, the score of the testing samples were proposed to be computed with two decision criteria based on l^1 norm ratio and l^2 norm residual ratio as given below:

$$\mathbf{M1}: \ l^1 \text{ norm ratio} = \|\delta_{\mathbf{1}}(\hat{\mathbf{x}})\|_{\mathbf{1}} / \|\hat{\mathbf{x}}\|_{\mathbf{1}} \qquad (7)$$

$$\mathbf{M2}: \ l^2 \text{ residual ratio} = \frac{\|\mathbf{y} - \mathbf{A}\delta_{\mathbf{2}}(\hat{\mathbf{x}})\|_{\mathbf{2}}}{\|\mathbf{y} - \mathbf{A}\delta_{\mathbf{1}}(\hat{\mathbf{x}})\|_{\mathbf{2}}} \qquad (8)$$

A. Implementation details of speaker verification system

Following the above discussed verification approach we explored the speaker verification using GMM mean supervectors. In speaker verification systems, usually the speaker models are developed using large (1-2k) number of mixtures Gaussian models and also the features are of dimensions as high as 39. As a result, doing full greedy searches over such large size dictionary was not feasible with our limited compute resources. So we have tried an alternate method as discussed below.

Our implementation of the speaker verification system differs from the one discussed in Section II in the following two aspects:

- The verification of a target speaker is done in presence of all other remaining target speakers unlike using any fixed set of non-target background speakers and so our dictionary contains all the target speaker supervectors.
- For finding the sparse representation of a target, the maximum number of sparse coefficients is restricted to the number of training examples available for each speaker.

For scoring purpose we first tried the above described scoring methods *M1* and *M2*. It was noticed that the classification performances obtained with these scoring methods turned out to be quite imbalanced with respect to miss and false alarm probabilities. The reasons for this behavior is discussed in Section IV and we have also proposed an alternate scoring method to address this problem.

III. Database and Experimental Setup

For the experiments, we have used the IITG-MV speaker recognition database, a locally collected speech data to study the impact of different variabilities on the speaker recognition task due to multi-sensor, multi-lingual, multi-style and multi-environment conditions. The details of the IITG-MV speaker recognition database can be found in [14]. The database contains two sets each having two recording sessions. Each set contains 100 speakers, with 50 speakers common across both sets. The first set is collected in office-environment involving multiple sensors, multiple languages,and different speaking styles (conversational and read speech) and it is referred to as the IITG-MV Phase-I. The second set differs from the first one in data collected in uncontrolled environments such as laboratories, hostel rooms and corridors etc., while keeping the other variabilities unchanged and is referred to as the IITG-MV Phase-II. In Phase-I, the speech data collection was done in a small office room with electric fan and air conditioner switched on. The data was collected in parallel with a headset microphone connected to a tablet PC, the built-in microphone of another tablet PC, two mobile phones of different make with voice recording facility and one digital voice recorder (DVR). The Phase-II dataset contains speech from 70 male and 30 female subjects. For collecting the data, the same devices as in Phase-I were employed. In contrast to Phase-I data set, the Phase-II data set contains data which is recorded after passing through a wireless mobile channel. In this work we have used the MV Phase-II data.

The baseline UBM-GMM based speaker verification system was developed and tested using procedure similar to laid down for NIST SRE-03 evaluations [15]. The training data set contains 100 speech segments of 2 minutes each from the first session recordings. The test set contains 1000 segments of speech, derived from the second session recording of the same data set with 10 segments for each speaker. The segments are of length varying from 30 to 45 seconds. Similar to NIST SRE-03 protocol, each test segment is tested against 11 models out of which one is a true trial. This makes a total of 11,000 test trials with a true trial to false trial ratio equal to 0.1. For building the UBM, we have used 10 hours of speech data from 50 speakers from IITG-MV Phase-I data set which are who are not common with speakers of Phase-II data set.

The speaker models were created by adapting only the mean parameters of the UBM using maximum *a posteriori* (MAP)

approach with the speaker specific data. All speech data used is sampled at 8 kHz with 16 bits/sample resolution and was analyzed using a Hamming window of length 20 ms, frame rate of 100 Hz and pre-emphasis factor of 0.97. The MFCC feature vectors of 39 dimension were used to parameterize the speech data. Each feature vector comprised of C_1 to C_{13} static MFCC and their first and second order derivatives. To remove the non-speech portions from input data, an energy based voice activity detector with fixed threshold was used. The cepstral mean subtraction and variance normalization was applied on all features so as to reduce the effect of mismatch in channel. For sparse representation based speaker verification, the training speech of 2 minutes duration was further segmented into 4 parts to get sufficient number of training examples from the available data. Each of these training samples and the test samples were used to adapt the UBM to get corresponding supervectors. For finding out the performance of the SV systems, the detection error trade-off (DET) curves were plotted using the verification scores. The equal error rates (EER) noted from the DET curves are used to evaluate the speaker verification performance in all conditions. The GMM-UBM and sparse representation based speaker verification systems were developed with the help of the hidden Markov model toolkit.

IV. Results and Discussion

The initial sparse representation based speaker verification (SR-SV) system is developed using speaker supervectors derived from the MAP adapted UBM of size 1024 mixtures. The performances in terms of DET plots for the SR-SV systems using *M1* and *M2* scoring methods on headset recorded data are given in Figure 1. The figure also shows the performance of the standard GMM-UBM system with 1024 mixtures for contrast. It is to note that with both the scoring methods the SR-SV system has performed significantly poor compared to GMM-UBM system. Before addressing the poor performance obtained with the SR-SV system further, we first highlight the imbalanced performances with respect to to miss and false alarm probabilities obtained with *M1* and *M2* scoring methods. In case of *M1*, as the maximum number of sparse coefficients is limited to the number of training examples in our case, a large number of the impostor trials would result in a score equal to zero. In case of *M2*, for impostor trials the denominator error dominates and so the scores turns out to be a small positive number. As a result, the variance of the false scores in both of the cases are found to be very small compared to that of the true scores and this explains the imbalanced performance noted in the DET curve. The same can be verified from the histograms of true and false trails shown in Figure 2 for each of the DET plots in Figure 1. To address the above discussed issue, we also developed an alternate scoring method based on l^2 norm residual ratio as discussed in the following subsection.

A. Alternate scoring method

As noted, in case of scoring methods *M1* and *M2*, most of the false trials produce similar scores. To avoid this problem,

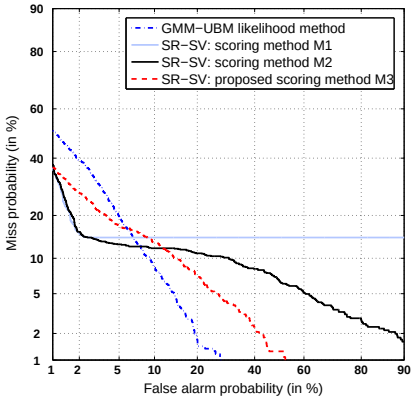

Fig. 1. *DET curves for SR-SV systems using different scoring methods along with that of GMM-UBM system on headset condition data using a UBM of 1024 mixtures.*

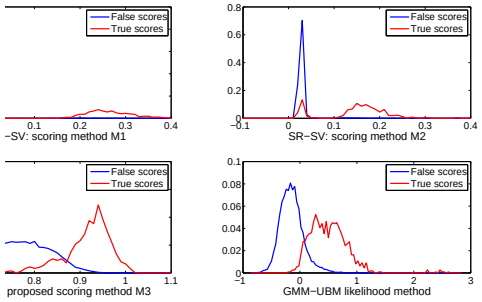

Fig. 2. *Histograms of true and false trial scores corresponding to the DET plots shown in Figure1.*

we define an alternate scoring method as the ratio between the two representation errors of the test vector $\mathbf{y}$ obtained by using the two dictionaries $\mathbf{A}$ and $\mathbf{A}_1$. The dictionary $\mathbf{A}$ contains the training supervectors all target speakers whereas the dictionary $\mathbf{A}_1$ contains the training supervectors of the claimed speaker only. The test vector $\mathbf{y}$ is represented using dictionaries $\mathbf{A}$ and $\mathbf{A}_1$ by finding greedy projections $\hat{\mathbf{x}}$ and $\hat{\mathbf{x}}_\mathbf{1}$ respectively. The proposed scoring metric is defined as:

$$\mathbf{M3}: \quad \text{Score} = \frac{\|\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}\|_\mathbf{2}}{\|\mathbf{y} - \mathbf{A_1}\hat{\mathbf{x}}_\mathbf{1}\|_\mathbf{2}} \tag{9}$$

In case of the above scoring method, for the true trials the representation error using dictionaries $\mathbf{A}$ and $\mathbf{A}_1$ are expected to be approximately same which gives a score near to 1 and for the false trials the representation error using $\mathbf{A}_1$ dominates which gives a smaller score. In Figure 1 and 2 the DET plot and histogram of scores for the SR-SV system with scoring method *M3* are also shown. It can be noted from the plot that, the DET curve corresponding to the *M3* scoring method is balanced with respect to miss and false alarm probabilities without any degradation in the EER value.

On further exploration about the cause of the poor performance of SR-SV system we noted that there is high coherence among the atoms of the dictionary. The high coherence is attributed to the fact that due to small amount of the data per speaker training segment, large number of the mixtures of

TABLE II
Speaker verification performance for SR-SV systems with supervectors created using UBMs of different size along with that of GMM-UBM system of different sizes for contrast purpose.

Train-Test Sensor Conditions	EER (in %)							
	Sparse Classification with GMM Supervectors						GMM-UBM	
Size of UBM →	32	64	128	256	512	1024	1024	128
Headset	9.29	8.86	**8.70**	9.36	10.48	11.83	**9.39**	11.02
Mobile	9.84	9.66	**9.57**	10.03	11.74	14.36	**14.27**	15.44
Tablet PC	13.82	13.18	**12.74**	12.92	13.91	16.17	**13.00**	13.27
Voice Recorder	13.95	12.82	**12.28**	12.74	13.37	16.06	**12.37**	14.00

the UBM model remain unadapted. To verify this hypothesis we trained smaller size UBMs and then derived the speaker supervectors by their adaptation. With the smaller number of mixture it is expected that most of them would get adapted thus result in reduced coherence among the atoms. The maximal coherence across speakers in the dictionary for different dimension supervectors are given in Table I. It is to note that, as expected the coherence does reduce with reducing size of the supervectors. Further the performance of the SR-SV system is evaluated for these different size supervectors and is given in Table II for different sensor condition datasets. We note that the best performance is obtained with SV system developed from 128 size UBM model which turn out to be better than those for GMM-UBM of 1K mixture size, for all sensor conditions. For the contrast, we have also given the performance for GMM-UBM system with 128 mixture size which is expectedly significantly degraded compared to 128 mixture based SR-SV system.

TABLE I
Table showing the maximal mutual coherence among speaker supervectors created using UBMs of different size.

Size of UBM	32	64	128	256	512	1024
Max. mut. coh.	0.91	0.91	0.92	0.94	0.96	0.98

V. Conclusions

In this paper we have reported the performance of a speaker verification system developed using the sparse representation framework as the classifier. The training and test speech samples with variable duration were represented as fixed length GMM mean supervectors created by adaptation of a UBM. The test supervectors were represented in terms of the dictionary created using all training supervectors. The sparse solution to the corresponding coefficients was used to identify the class of the test vector. The SV systems are developed using the speech data collected in realistic environments with multiple sensors and the results were compared against that of a standard GMM-UBM system in terms of EER. for all of the four sensors data considered, the SR-SV system based on 128 mixture GMM was found working better than the standard GMM-UBM systems developed using mixtures up to 1024.

In future we would like to explore the methods of designing dictionaries with lesser coherence and compare the performance of the proposed method with SVM based speaker verification system employing the GMM mean supervectors. In addition to this we would also like to explore the use of joint factor analysis methods to improve the performance for mixed sensor cases.

VI. Acknowledgements

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.

References

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

[2] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, pp. 308–311, 2006.

[3] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42–54, 2000.

[4] D. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. ICASSP*, vol. 2, 2003, pp. II–53–6.

[5] A. Solomonoff, W. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. ICASSP*, vol. 1, 2005, pp. 629 – 632.

[6] R. Teunen, B. Shahshahani, and L. Heck, "A model-based transformational approach to robust speaker recognition," in *Proc. ICSLP*, vol. 2, 2000, pp. 495–498.

[7] S. C. Yin, R. Rose, and P. Kenny, "A joint factor analysis approach to progressive model adaptation in text-independent speaker verification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 1999–2010, 2007.

[8] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[9] K. Huang and S. Aviyente, "Sparse representation for signal classification," in *Neural Information Processing Systems*, 2006.

[10] R. Rubinstein, A. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045 –1057, 2010.

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

[12] M. Li and S. Narayanan, "Robust talking face video verification using joint factor analysis and sparse representation on gmm mean shifted supervectors," *to appear in Proc. ICASSP*, 2011.

[13] D. L. Donoho and M. Elad, "Optimally sparse representation in general (non-orthogonal) dictionaries via l_1 minimization," in *Proc. National Academy of Science of the USA*, 2003, pp. 2197–2202.

[14] Haris B C, G. Pradhan, A. Misra, S. Shukla, R. Sinha, and S. R. M. Prasanna, "Multi-variability speech database for robust speaker recognition," in *Proc. NCC*, 2011, pp. 1 –5.

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