THE IITG SPEAKER VERIFICATION SYSTEMS FOR NIST SRE 2012

Haris B C, G. Pradhan, R. Sinha and S. R. M. Prasanna

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

ABSTRACT

In this paper, we describe the speaker verification (SV) systems developed by Indian Institute of Technology Guwahati (IITG) for the NIST 2012 speaker recognition evaluations. The primary submission consists of five gender dependent SV systems combined at score level. Among the five systems two are based on sparse representation over learned and exemplar dictionaries, and the remaining are based on the generic i-vector and its variants obtained by vowel and non-vowel conditioning. The exemplar dictionary based system in particular exploits the new evaluation rule allowing the knowledge of all targets in each detection trial. The performance of the system is presented for the NIST SRE 2012 core task.

1. INTRODUCTION

The NIST SRE 2012 is the most recent one in the on going series of speaker recognition evaluations (SREs) conducted by the national institute of standards and technology (NIST). NIST has introduced several changes in the SRE 2012 compared to the earlier ones. In previous evaluations, the evaluation data set, which is released at the beginning of the evaluation period, has contained both training and test data. In SRE 2012, most of the target speakers are taken from previous SRE datasets and the corresponding training data was provided to the participants well in advance of the evaluation period. Furthermore, in SRE 2012 the training data for each such target speaker includes a fairly large number of speech segments taken from multiple recording sessions. In SRE 2012 the knowledge of all targets is allowed in computing the detection score of each trial. To examine the effect of these new conditions on systems' performance, test segments from non target speakers are also included in the test data set which forms *un-known* non-target trials in addition to the *known* non-target trials. The decision cost continues as the primary performance measure and it's computation method is modified to accommodate the known and unknown non-target trials [1].

Recently proposed i-vector representation [2] forms the basis for most of the state-of-the-art speaker recognition systems. i-vectors can be considered as a compact and fixed dimension representation of speech utterances. i-vectors are being used with classifiers like support vector machines, cosine kernel or with Bayesian methods like PLDA to perform speaker verification. One of the major challenges in speaker recognition research is the mismatch due to session and channel variability. Various session and channel compensation methods used with modern speaker recognition stems include LDA, WCCN, NAP and JFA.

In last few years, the discriminative abilities of the sparse representation techniques have also been exploited for speaker recognition. In [3], Kua et. al. proposed a speaker identification system which uses sparse representation classification (SRC) with an exemplar dictionary created using GMM mean supervectors. Later speaker verification (SV) tasks using the SRC with exemplar dictionary created using GMM mean supervectors and total variability i-vectors were also reported in [4] and [5] respectively. In our earlier work [6], 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 in [7] we have presented a speaker verification system employing sparse representation of centered GMM mean supervectors over a dictionary learned using the KSVD algorithm. We have extended this work with the use of discriminatively learned dictionaries in [8, 9] 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 SRE 2003 dataset, the proposed system with discriminatively learned dictionary found to outperform all other SV systems considered both with and without session/channel variability compensation. Motivated by this we have used the SRC over learned dictionary based SV system as a subsystem for our NIST SRE 2012 submission. To exploit the availability of multiple training segments for most of the speakers, and the new rule allowing the usage of allowing the knowledge of all targets in each detection trial, we have also used an SRC system with exemplar dictionary.

Recent works from our group [10, 11] have explored the use of vowel and nonvowel like regions of speech for SV by building systems exclusively for vowel like and nonvowel like regions and combining them at score level. Experiments conducted with both GMM-UBM and i-vector based systems on NIST SRE 2003 data set showed significant improvements in performance especially in degraded conditions. The drawback of this method was the computational complexity involved in the segmentation of training and test speech signals into vowel and nonvowel like regions which made it difficult to use with very large data sets like SRE 2012. Motivated by the improved results obtained by the said conditioning, we have developed a similar method which avoids the segmentation of training and test speech. This novel idea uses a total variability (T) matrix conditioned with vowel and nonvowel like speech data for extracting the i-vector representation for training and test speech signals. Based on the methods adopted for building the conditioned T matrix we have built two variants of the vowel-nonvowel conditioned i-vector systems for the NIST 2012 evaluations.

In addition to the four systems mentioned above, we have also developed a generic i-vector based system for the evaluation. The primary system for the 'core' task is the score level fusion of two sparse representation based systems, two vowel-nonvowel conditioned i-vector based systems and a generic i-vector based systems. The alternate-1 system is the fusion of only sparse representation based systems while the alternate-2 system is the fusion of the two vowel-nonvowel conditioned i-vector systems.

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

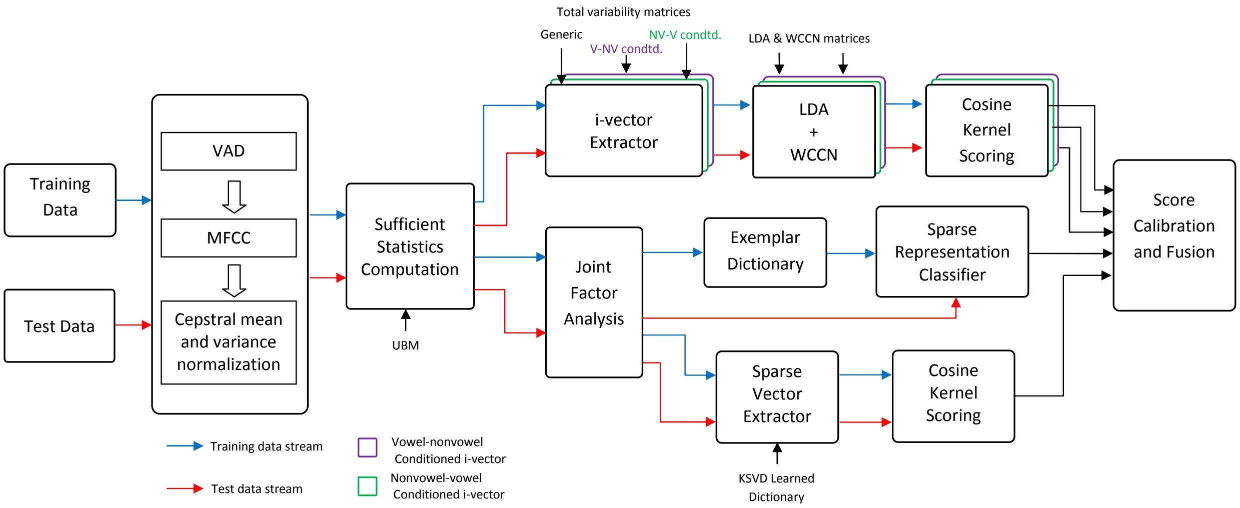

Fig. 1. Block diagram of the IITG primary system for the NIST SRE 2012

2. DATA PROCESSING, FEATURE EXTRACTION AND UBM

In NIST SRE 2012, the training and the test datasets include speech data recorded in three different conditions viz. telephone recorded phone call, microphone recoded interview and microphone recorded phone call. As the training data is mostly derived from the training and test sets of the SRE'06, SRE'08 and SRE'10 evaluations, we have used data from same datasets for the system development.

In NIST SRE 2012, multiple training segments are available for most of the speakers. As the duration of training segments provided are varying from a few seconds to minutes, we have redistributed the data into segments of approximately equal duration prior to using it for modeling the speakers. Standard MFCC features of 13 dimensions with their first and second derivatives are used as the base features for all systems. To remove the non-speech portions from input data, an energy based voice activity detector is used. The cepstral mean subtraction and variance normalization are also applied on feature vectors so as to reduce the effect of mismatch due to channel.

Two gender-dependent universal background models (UBMs) of 1024 Gaussian mixtures are created and are used in all the developed systems. The male UBM is created using approximately 40 hours of telephone recorded speech data of 725 speakers taken from the development data. Similarly the female UBM is learned using 50 hours of telephone recorded speech data of 1099 speakers taken from the development data. The MFCC feature extraction and UBM building are done using the HTK toolkit[12].

3. SESSION AND CHANNEL COMPENSATION

In this section, the various session and channel compensation techniques used are described.

3.1. Joint factor analysis

Joint factor analysis (JFA) is used for removing the session and channel factors from the GMM mean supervectors by modeling the speaker and session and channel subspaces. The gender dependent UBMs are used to collect the 0^{th} and 1^{st} order statistic for training gender dependent JFA systems. We have used two types of JFA implementations to use with our SV systems and are described below.

JFA type-I: The type-I implementation of JFA closely follows the method described in [13]. First, the eigenvoices are trained on the telephone speech from the development dataset. Then a set of eigenchannels is trained on telephone speech only after removing the earlier learned speaker factors. It is then followed by learning another set of eigenchannels on microphone speech while removing the speaker and telephone channel factors.

JFA type-II: In the type-II implementation of JFA, eigenvoices and eigenchannels are learned by pooling the telephone and microphone speech data from the development dataset.

3.2. Linear discriminant analysis and Within Class Covariance Normalization:

Standard Linear discriminant analysis (LDA) [2] and Within Class Covariance Normalization (WCCN) [14] techniques are used to perform session/channel compensation for i-vectors.

4. SPEAKER VERIFICATION SYSTEMS

Figure 1 shows the block diagram of the IITG primary system for the NIST SRE 2012. This shows the fusion of five of our sub-systems which are described in the following subsections

4.1. SRC over exemplar dictionary based SV system

In this system, the speaker verification is done by finding the sparse representation of the vector representing the test data over a exemplar dictionary $\boldsymbol{D}$ created using the vectors representing the training data of all the target speakers[6]. This system in particular exploits the new evaluation rule allowing the knowledge of all targets in each detection trial. Given M target speakers each having multiple training examples represented as $\{\boldsymbol{y}_{mi}\}_{i=1}^{N_M}$, the exemplar dictionary $\boldsymbol{D}$ is constructed as follows,

$$\boldsymbol{D} = [\boldsymbol{y}_{1,1}, \ldots, \boldsymbol{y}_{1,N_1}, \boldsymbol{y}_{2,1}, \ldots, \boldsymbol{y}_{2,N_2}, \ldots, \ldots, \boldsymbol{y}_{M,N_M}] \quad (1)$$

The session and channel compensated GMM mean shifted supervectors are used to represent speech utterances in a vector form. JFA type-1 with 300 eigen voices, 100 telephone eigen channels and 100 microphone eigen channels is used for removing the session and channel factors from the supervectors in case of telephone-telephone trials. Similarly, JFA type-II with 300 eigen voices and 200 telephone eigen channels is used with microphone-microphone and microphone-telephone trials. For the classification, the sparse projection $\boldsymbol{x}$ of the test vector $\boldsymbol{y}$ over the dictionary $\boldsymbol{D}$ is obtained using the OMP algorithm with sparsity constraint of 50. The method followed for computing the verification score is described in the Sec. 5

4.2. SRC over learned dictionary based SV system

This system employs sparse representation of channel compensated GMM mean shifted supervectors over a learned dictionary for modeling speakers [8]. The GMM mean shifted supervector $\boldsymbol{y}$ is modeled using the sparse representation with a learned dictionary $\boldsymbol{D}$ as,

$$\boldsymbol{y} = \boldsymbol{D}\boldsymbol{x} \quad (2)$$

We have used JFA type-II with 300 eigen voices and 200 telephone eigen channels for removing the session and channel factors from all of the mean shifted supervectors as a pre-processing. A gender dependent dictionary of 1000 atoms is learned on the development data using the KSVD algorithm [15]. The sparse representation vector $\boldsymbol{x}$ for both training and test data are estimated using the orthogonal matching pursuit (OMP) algorithm with a sparsity constraint of 50. The verification score for a given trial is computed using cosine kernel as described in Sec. 5.

4.3. Generic i-vector based SV system

This is an implementation of the generic i-vector system suitable for telephone and microphone speech proposed in [16]. Here the GMM mean supervector $\boldsymbol{s}$ for a speaker is represented as,

$$\boldsymbol{s} = \boldsymbol{m} + [\boldsymbol{T}_{phn}|\boldsymbol{T}_{mic}]\boldsymbol{w} \quad (3)$$

where, $\boldsymbol{m}$ is the speaker-independent UBM mean supervector, $\boldsymbol{T}_{phn}$ is the total variability matrix learned using telephone data, $\boldsymbol{T}_{mic}$ is the total variability matrix learned using microphone data and $\boldsymbol{w}$ is the i-vector. LDA and WCCN are used to compensate the session and channel effects in the i-vectors. The method followed for finding the verification score is described in the Sec. 5.

4.4. Vowel and nonvowel like regions conditioned i-vector based SV system

In this system, the columns of the total variability matrix learned using telephone and microphone data are further conditioned with vowel-like regions (VLRs) and nonvowel like regions (nonVLRs) of the development speech data. With this, the Eq. 3 will be modified as,

$$\boldsymbol{s} = \boldsymbol{m} + [\boldsymbol{T}_{phn,vl}|\boldsymbol{T}_{phn,nvl}|\boldsymbol{T}_{mic,vl}|\boldsymbol{T}_{mic,nvl}]\boldsymbol{w} \quad (4)$$

where, $\boldsymbol{T}_{phn,vl}$ and $\boldsymbol{T}_{phn,nvl}$ represent the subspaces corresponding to the VLRs and nonVLRs learned using telephone data and $\boldsymbol{T}_{mic,vl}$ and $\boldsymbol{T}_{mic,nvl}$ represent the subspaces corresponding to the VLRs and nonVLRs learned using microphone data. The VLRs are detected following the similar procedure as that given in [11]. The nonVLRs are selected by excluding the VLRs from the speech regions detected using an energy based VAD. The conditioned T matrix for this system is learned in a similar fashion to that of the generic i-vector system described in the Sec. 4.3. Two approaches are considered with emphasize on either VLRs or nonVLRs as explained below:

Table 1. *Performances of individual and fused SV systems on the development data set in terms of DCF values*

No.	System	Act. DCF	Min. DCF
i	SRC with exemplar dictionary	0.16	0.13
ii	SRC with learned dictionary	0.31	0.27
iii	Generic i-vector	0.29	0.26
iv	V-NV conditioned i-vector	0.31	0.25
v	NV-V conditioned i-vector	0.28	0.28
vi	Fusion: i to v (Primary)	0.14	0.13
vii	Fusion: i & ii (Alternate -1)	0.15	0.12
viii	Fusion: iv & v (Alternate -2)	0.29	0.25

4.4.1. VLRs emphasized i-vector system

In this system, the subspaces representing VLRs are learned prior to that representing the nonVLRs for both telephone and microphone speech cases. We have learned 450 eigen vectors corresponding to the VLR and 350 eigen vectors corresponding to the nonVLR.

4.4.2. nonVLRs emphasized i-vector system

In this, the estimation of nonVLR subspaces is followed by that of the VLR subspaces. We have learned 450 eigen vectors corresponding to the nonVLR and 350 eigen vectors corresponding to the VLR.

5. SCORING, CALIBRATION AND FUSION

In this section we describe the different scoring methods used for the developed SV systems and the calibration and fusion of the scores to get the final likelihood ration scores for submission.

5.1. Scoring methods

In NIST 2012 SRE, multiple training segments are available for most of the speakers which includes both telephone and microphone recorded data. For all speakers, at least one telephone recorded training utterance is available. Based on the test data recording channel and the availability of telephone/microphone models for the claimed speaker, we have used different test strategies for different systems as given below.

1. **i-vector based systems:** The telephone and microphone speaker models for each speaker are created by taking the mean of the i-vectors of the telephone and microphone training utterances available for that speaker, respectively. All telephone test segments are tested against telephone models. In case of microphone test segments, microphone models are used for scoring if available and telephone models are used otherwise. Scores for a trial is found using cosine kernel as given below.

$$\text{Score=} \; \frac{< \hat{\boldsymbol{x}}_{clm}.\, \hat{\boldsymbol{x}}_{tst} >}{\|\hat{\boldsymbol{x}}_{clm}\| \; \|\hat{\boldsymbol{x}}_{tst}\|} \quad (5)$$

   where $\hat{\boldsymbol{x}}_{clm}$ and $\hat{\boldsymbol{x}}_{tst}$ represent the model vector and test vector, respectively.

Table 2. *Performance of the primary, alternate-1 (SRC based) and alternate-2 (Vowel-Nonvowel conditioned i-vector based) systems for the common evaluation conditins of the NIST SRE 2012 core task [1]* in terms of actual and minimum DCF

Evaluation condition		Systems submitted to NIST SRE 2012					
		Primary (all fused)		Alt.-1 (Sparse only)		Alt.-2 (Cond. i-vect. only)	
		Act. DCF	Min. DCF	Act. DCF	Min. DCF	Act. DCF	Min. DCF
Phone call	No noise	0.42	0.33	0.40	0.37	0.56	0.37
	Added noise	0.51	0.35	0.48	0.42	0.69	0.41
	Noisy env.	0.45	0.29	0.42	0.36	0.60	0.33
Interview	No noise	0.44	0.43	0.47	0.44	0.53	0.48
	Added noise	0.62	0.45	0.84	0.48	0.58	0.48

2. **SRC over learned dictionary based SV system:** Telephone and microphone speaker models for each speaker are created by taking the weighted average of the sparse vectors representing the training data for a given speaker. While creating the telephone speaker models, a weight of 0.7 is given to vectors representing telephone speech and a weight of 0.3 is given to the sparse vectors of microphone speech. All telephone test segments are tested against telephone models. In case of microphone test segments, microphone models are used for scoring if available and telephone models are used otherwise. Scores for a trial is found using cosine kernel similar to the i-vector systems case.

3. **SRC over exemplar dictionary based SV system:** The exemplar dictionary is created by pooling the supervectors created using the telephone training data for all speakers for both telephone and microphone test cases with the following exemption. The score for a given trial is found from the sparse representation vector $\boldsymbol{x}$ as the l_1 norm $\|\delta_1(\hat{\boldsymbol{x}})\|_1$ where, $\delta_1(\hat{\boldsymbol{x}})$ is a vector whose nonzero entries are the only entries corresponding to the training vectors of the claimed speaker in the exemplar dictionary.

5.2. System tuning using development trials

For finding the optimal parameters for the SV systems, we have used a set of development data set and trials. The system training was done using a part of the actual SRE 2012 training data. A test data set of about 4000 segments created using the other part of the actual training data was used to perform about 80000 trials. The various system parameters were tuned using this development trials.

5.3. Calibration and fusion

Mapping of the scores generated by sub systems to log-likelihood ratios (LLR) and fusion of these LLRs were performed the BOSARIS [17] toolkit which uses the linear logistic regression method for the same. The scores from the development trials which is described in the previous subsection was used to train the calibration and fusion process. The primary submission for the core-core task is the score level fusion of all five subsystems described in the above section. The first alternate submission for the core-core task is the score level combination of the two sparse representation based systems i.e. the SRC over learned dictionary based and SRC over exemplar dictionary based SV systems. The second alternate submission for the core-core task is the score level fusion of the two vowel-nonvowel constrained SV systems.

6. RESULTS

The Table 1 shows the performance in terms of actual and minimum detection costs [1] for various SV systems on the development trials. As the development test set does not contain speech segment from any 'unknown' speaker, the $P_{unknown}$ was set to 0 while computing the detection cost. It can be noted that the SRC with exemplar dictionary based SV system which exploits the new evaluation rule allowing a closed set speaker verification, performs significantly better than all other systems in consideration. The SRC with learned dictionary based and the two variants of i-vector systems are observed to be performing comparable to that of the generic i-vector system. We have also observed that all other systems contribute improvements while fused with the best performing system. The table also shows the performance of the three different system combinations submitted as the primary and two alternate systems evaluated using the development trials. These scores were used to train the fusion and calibration of the corresponding systems on the actual evaluation trials.

The performance of the primary and alternate systems on the actual evaluation trials of the NIST SRE 2012 for the five common conditions [1] of the core task are summarized in the Table 2. On comparing the performance of the alternate systems with that of the primary one, it can be noted that the sparse representation based systems have performed better than the i-vector based ones in all the cases except the noise added interview case, in terms of actual DCF. We are not able to report the performance of the individual subsystems and its analysis here as the new evaluation tools from the NIST is not available at the time of submission of this paper.

7. SUMMARY

In this paper, we have described the speaker verification (SV) systems developed by the Indian Institute of Technology Guwahati for the NIST SRE 2012. The five parallel gender dependent subsystems developed include two sparse representation based SV systems, two vowel-nonvowel conditioned i-vector SV systems and a generic i-vector SV system. The performances evaluated on the development trials and on the actual evaluation trials shows the potential of the sparse representation over exemplar dictionary based SV systems which exploits the new evaluation rule allowing the closed set speaker verification.

8. ACKNOWLEDGEMENT

This work has been supported by the ongoing project titled "Development of speech based multilevel person authentication system" sponsored by the Department of Information Technology, Government of India.

9. REFERENCES

[1] The NIST Year 2012 Speaker Recognition Evaluation Plan, www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf.

[2] N. Dehak, P.J. 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.

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

[4] Jia Min Karen Kua, Eliathamby Ambikairajah, Julien Epps, and Roberto Togneri, "Speaker verification using sparse representation classification," in *ICASSP 2011*, may 2011, pp. 4548–4551.

[5] Ming Li, Xiang Zhang, Yonghong Yan, and Shrikanth Narayanan, "Speaker verification using sparse representations on total variability i-vectors," in *Interspeech 2011*, may 2011, pp. 4548–4551.

[6] 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.

[7] Haris B C and R Sinha, "Speaker verification using sparse representation over KSVD learned dictionary," in *Proc. 18th National Conference on Communications 2012*, Feb. 2012.

[8] Haris B C and R Sinha, "Sparse representation over learned and discriminatively learned dictionaries for speaker verification," in *Proc. ICASSP 2012*, March. 2012.

[9] Haris B C and R Sinha, "On exploring the similarity and fusion of i-vector and sparse representation based speaker verification systems," in *Odyssey 2012: The Speaker and Language Recognition Workshop, Singapore*, 2012.

[10] S. R. M. Prasanna and G. Pradhan, "Significance of vowel-like regions for speaker verification under degraded condition," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2552–2565, May 2011.

[11] G. Pradhan and S.R.M. Prasanna, "Speaker verification by vowel and nonvowel like segmentation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 854–867, April 2013.

[12] S.J. Young, G. Evermann, M.J.F. Gales, D. Kershaw, G. Moore, J.J. Odell, D.G. Ollason, D. Povey, V. Valtchev, and P.C. Woodland, *The HTK Book version 3.4*, Cambridge University Engineering Department, Cambridge, U. K., 2006.

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

[15] 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.

[16] Mohammed Senoussaoui, Patrick Kenny, Najim Dehak, and Pierre Dumouchel, "An i-vector extractor suitable for speaker recognition with both microphone and telephone speech," 2010.

[17] The BOSARIS Toolkit, www.https://sites.google.com/site/bosaristoolkit/.