SPEAKER VERIFICATION USING M-VECTOR EXTRACTED FROM MLLR SUPER-VECTOR

A. K. Sarkar, J. F. Bonastre and D. Matrouf

University of Avignon, LIA, Avignon, France

sarkar.achintya@gmail.com, {jean-francois.bonastre,driss.matrouf}@univ-avignon.fr

ABSTRACT

In this paper, we propose a speaker verification system called *m-vector* system, where speakers are represented by uniform segmentation of their Maximum Likelihood Linear Regression (MLLR) super-vectors, denoted *m-vectors*. The MLLR super-vectors are extracted with respect to Universal Background Model (UBM) with MLLR adaptation using the speakers data. Two criterion are followed to segment the MLLR super-vector: one is disjoint segmentation technique and other one is overlapped windows. Afterward, m-vectors are conditioned by our recently proposed [1] session variability compensation algorithm before calculating score during test phase. However, the proposed method is not based on any *total variability space* concept and uses simple MLLR transformation for extracting m-vector *without* considering any transcription of the speech segment. The proposed system shows promising performance compared to the conventional i-vector system. This indicates that session variability compensation plays an important role in speaker verification. Speakers can be represented by simpler way instead of generating i-vector in conventional system and able to achieve performance comparable to the i-vector based system. Experiment results are shown on NIST 2008 SRE core condition.

***Index Terms*—** m-vector, MLLR super-vector, LDA, WCCN, Speaker Verification

1. INTRODUCTION

Speaker verification is a binary classification problem. It either accepts or rejects the claimant speakers by analyzing his/her voice signal.

Most commonly used i-vector technique [2] becomes the state-of-the-art in speaker verification. In this technique, speakers are represented by vectors on *total variability space*. The vectors are called *i-vectors*. The *total variability space* is build using computationally heavy iterative algorithm and pooling data from many speakers over different channels/sessions. The i-vectors are then post-processed to account the session variability before scoring (during testing). Commonly, Linear Discriminant Analysis (LDA) [2, 3] is used to discriminate the speakers and Within Class Covariance Normalization (WCCN) [4] to account the session variability. Probabilistic (P)-LDA [5, 6] based generative modeling technique is shown to be useful in gender independent speaker verification task. Each of these post-processing methods has its own scoring technique between two i-vectors. Therefore, the i-vector system can be broadly divided into two parts: one is building *total variability space*, T and other is post-processing task i.e. *session variability modeling*. Recently, a conditioning algorithm called Eigen Factor Radial (EFR) was proposed [1] on the i-vector, which includes length normalization. An interesting performance is noticed [1] using the method on i-vector system. This arise some questions about the role of T on i-vector system. The performance of the i-vector system comes from the *total variability space*, T, and from the *session variability modeling*.

The motivation of this paper is to partially answer this question with experimentally approach. If we are able to achieve performance which is comparable to the conventional i-vector system with the same session variability and decision steps but using simpler vector (which is generated without concept of *total variability space*), it will evident that the key role in speaker verification is played by *session variability modeling* while different approaches can be applied for i-vector estimation.

In order to reach our objective, we are proposing a speaker data representation by vectors called *m-vectors*. The *m-vectors* are extracted by uniform segmentation of speakers Maximum Likelihood Linear Regression (MLLR) super-vectors. It is important to note that the m-vector extraction is performed independently for each speech segment with respect to Universal Background Model (UBM). No specific development data or heavy computing load are required here. Then m-vectors are post-processed using EFR conditioning algorithm. Performance of the proposed method is compared with an i-vector-based baseline.

This paper is organized as follows: Section 2 describes the concept of MLLR adaptation. Baseline system is described in Section 3 followed by post-processing method on i-vector (Section 4). Section 5 describes the proposed method. Experimental setup is described in Section 6. Results and discussion are presented in Section 7. Finally, the paper is concluded in Section 8.

2. MLLR ADAPTATION

MLLR [7] is a adaptation technique. It is generally used on Automatic Speech Recognition (ASR) system. It estimates a affine transformation, A with respect to Speaker Independent (SI) model for a given speech data. The transformation is then applied to Gaussian mean vectors of the SI model to get the adapted model:

$$\hat{\mu} = A\mu + b, \quad \hat{\Sigma} = \Sigma \tag{1}$$

where μ and Σ are the mean and co-variance matrix of the SI model. (A, b) are the MLLR transformation parameters. $\hat{\mu}$ and $\hat{\Sigma}$ are the parameters of the adapted model.

The concept of the MLLR super-vector was first proposed by Stolcke et al. [8] in speaker verification on Support Vector Machines (SVMs) framework. Several other speaker verification systems can be found in literature based on MLLR/Constraint (C)-MLLR super-vector, specially in [8, 9, 10]. They use MLLR super-vector as a feature in speaker verification system using SVMs. The main differences of our proposed method with [8, 9, 10] are: (i) we estimate only *single global* MLLR transformation to create speaker specific

© EURASIP, 2012 - ISSN 2076-1465

m-vector without any Automatic Speech Recognition (ASR) transcription in contrast to [8, 10] (ii) we do not use any SVM modeling technique on speakers MLLR super-vectors. Recently, the usefulness of the MLLR super-vector has also shown on speaker identification task in anchor modeling framework [11] using eigen voice concept.

3. CLASSICAL I-VECTOR SYSTEM

We consider classical i-vector system [2] as the baseline system. An i-vector w (of dimension R) is calculated using,

$$\hat{M}_{[CF\times 1]} = M_{[CF\times 1]} + T_{[CF\times R]} w_{[R\times 1]} \tag{2}$$

where T is the *total variability space*. C and F are respectively, the number of mixture in UBM and dimension of the feature vector. $\hat{M}$ and M indicate the Gaussian Mixture Model (GMM) super-vector of the speaker adapted model and UBM, respectively.

During training and test phases, i-vectors of the target speakers and test utterance are estimated from the training/test data respectively using Eq.(2). i-vectors are then post-processed before calculating score. The post-processing methods are described in the next section.

4. POST-PROCESSING METHOD AND SCORING

There are several post-processing techniques available in literature [1, 2, 5], which are applied on i-vector to discriminate the speakers and account the effect of the channel/session variability. LDA [3, 2]+WCCN [4] is the most commonly used technique. In this approach, i-vectors are first projected onto LDA space to discriminate the speakers and followed by WCCN is applied to accounting the session variability. Finally, LDA+WCCN projected i-vectors are used for scoring in test phase. Generally, cosine kernel fast scoring [2] is used in this domain.

Eigen Factor Radial (EFR), a new method for intersession compensation and scoring is recently proposed [1] on the i-vector space. In this approach, an iterative conditioning algorithm is applied on the i-vectors in order to handle the session variability as,

$$\hat{w} \leftarrow \frac{V^{-\frac{1}{2}}(w-\overline{w})}{\sqrt{(w-\overline{w})^t V^{-1}(w-\overline{w})}} \tag{3}$$

where V and $\overline{w}$ are the covariance matrix and mean vector of the training i-vectors respectively in successive iteration. During test, a Mahalanobis-based scoring function described in Eq.(4) is used for scoring between two i-vectors.

$$score(\hat{w}_1, \hat{w}_2) = (\hat{w}_1 - \hat{w}_2)^t W^{-1} (\hat{w}_1 - \hat{w}_2) \tag{4}$$

W is the within-class covariance matrix computed using development data set. Details about the technique can be found in [1].

5. PROPOSED M-VECTOR SYSTEM

In this section, we describe our proposed m-vector based speaker verification system, where m-vector is extracted from speaker specific MLLR super-vector.

5.1. Speaker specific MLLR Super-vector estimation

MLLR transformations are estimated with respect to Speaker Independent (SI) model for given speech segments as in Eq.(1). Then the elements of the transformation matrix are stacked one by one to form a vector called MLLR super-vector [8]. The UBM is considered as the SI model and bias (b) is not considered in our experiment. Single iteration is followed in MLLR adaptation process without using any transcription of speech data. We use 50 dimensional feature vector, which gives $50 \times 50 = 2500$ elements for each MLLR super-vector. *Algorithm 1* describes the steps involved in MLLR transformation estimation for r^{th} speaker. Fig.1 shows graphical illustration of MLLR super-vector estimation process of speaker, r.

Algorithm 1: MLLR transformation

Initial: Load UBM and obtain speaker, r training feature vectors, $X = \{x_1, \ldots x_N\}$

Step 1: Determine the probabilistic alignment, $\gamma_j\,(t)$ of X with respect to UBM $\sim \mathcal{N}(\omega, \mu, \Sigma)$ for Gaussian mixture j as,

$$\gamma_j\,(t) = p(j|x_t) = \frac{\omega_j b_j(x_t)}{\sum_{k=1}^{C} \omega_k b_k(x_t)} \tag{5}$$

Step 2: Compute two sufficient statistics for i^{th} dimension of feature vectors,

$$K^{(i)} = \sum_{j=1}^{C}\sum_{t=1}^{N} \gamma_j(t)\, \frac{1}{\sigma_{ji}^2}\, x_i(t)\, \mu_j^{'} \tag{6}$$

$$G^{(i)} = \sum_{j=1}^{C} \frac{1}{\sigma_{ji}^2} \mu_j \mu_j^{'} \sum_{t=1}^{N} \gamma_j\,(t) \tag{7}$$

μ_j, σ_{ji}^2 and C are j^{th} mean, i^{th} component of j^{th} covariance matrix and number of Gaussian components of UBM, respectively. The symbol $(.)'$ indicates matrix transpose operation.

Step 3: i^{th} row of the MLLR transformation is obtained,

$$A_i^r = K^{(i)} {G^{(i)}}^{-1} \tag{8}$$

Step 4: Repeat Step 2 to 3 upto feature vector dimension

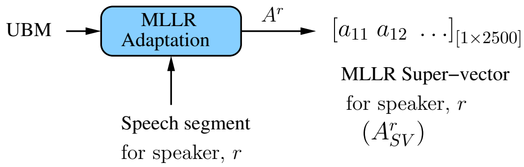

Fig. 1. *Illustration of MLLR super-vector estimation.*

5.2. Speaker specific m-vector extraction

After estimation of MLLR super-vector as described in previous section, MLLR super-vector is segmented into different part to form m-vector. We consider two criterion to extract m-vector as follows:

5.2.1. Method-I: Disjoint segmented m-vector system

Here, the MLLR super-vector of each target speaker is segmented into *equal disjoint* part i.e. $m_i^r \cap m_j^r = \Phi, \ \forall i \neq j$ (without any overlap between the segments). Each segment of the MLLR super-vector is considered as a m-vector of the particular target speaker. For example, m_1^r is a m-vector which is belonging to the first segment of the r^{th} speaker MLLR super-vector. Hence, each speaker is characterized by a number of m-vectors and each part of the m-vector constitutes a sub-system. In our experiment, we have 2500 elements in MLLR super vector. In case of m-vector dimension of 500 yields maximum 5 vectors to characterize each target speaker i.e. 5 sub-systems. Fig.2 graphically illustrates the m-vector extraction procedure of r^{th} target speaker from his/her MLLR super-vector.

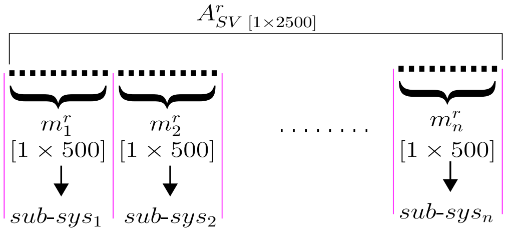

Fig. 2. *Illustration of m-vector extraction of r^{th} target speaker from his/her MLLR super-vector using disjoint segmented method.*

5.2.2. Method-II: Overlapped windowed m-vector system

In this case, target speaker specific MLLR super-vector is uniformly segmented using a moving window with 50% overlap of its previous adjacent segment as illustrated in Fig.3. The main motivation of this method is that it will be able to capture the speaker information lying in-between two adjacent segments (like feature extraction from speech signal with 50% overlap windows). Similarly to *method I*, the elements of the MLLR super-vector within each window is considered as a m-vector. The m-vector obtained for each window is constituted a separate system. The size of the window length controls the dimension of the m-vectors. By varying the length of window, various dimension of m-vectors are generated. In our experiment, window size of 500 elements yield best performance, which is presented later in this paper. This gives $[2 \times (2500/500) - 1] = 9$ m-vectors to characterize each speaker i.e. 9 sub-systems. Same notations are followed to represent the speakers in this system (in Fig.3) as *disjoint segmented m-vector system.*

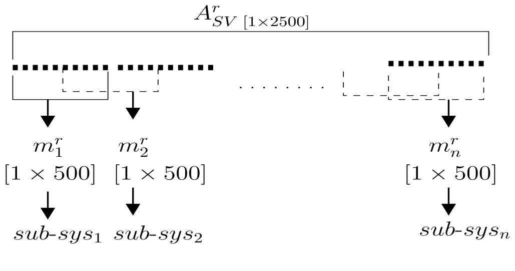

Fig. 3. *Illustration of m-vector extraction of speaker, r from his/her MLLR super-vector using overlapped windowed method.*

After extracting the m-vectors as described previous subsections, LDA is applied on m-vector of each sub-system to discriminant the speakers. Each sub-system has its own LDA transform. LDA transform is estimated pooling data from 890 non-target speakers.

5.3. Test Phase and Score Fusion

The m-vectors of the test utterance are extracted in similar manner as described in previous sub-sections. Then, m-vectors are projected on the LDA space by using the projection matrix belonging to their respective sub-systems. Finally, the projected m-vectors of the test utterance, say $\tilde{m}_1^{test}$ is scored against corresponding m-vector $\tilde{m}_1^r$ of the claimant speaker, r (obtained during training phase) and so on. Finally, the scores of the different m-vectors are fused (for particular LDA projected dimension across all the sub-systems) into single value. It can be expressed as,

$$S_i = score(\tilde{m}_i^r, \tilde{m}_i^{test}) \qquad (9)$$

$$Fusion\ score: Z = \frac{1}{N_{subsys}} \sum_{i=1}^{N_{subsys}} S_i \qquad (10)$$

where $score(.,.)$ indicates the function which defines the scoring between two m-vectors. Mahalanobis distance measure is used for scoring as Eq.(4). We use two iterations of Eigen Factor Radial (EFR) conditioning algorithm before calculating score between the two m-vectors. For fusion, equal weightage is given to all systems.

6. EXPERIMENTAL SETUP

All experiments are carried out on NIST 2008 SRE core condition (male speakers and det7 condition) as per NIST evaluation plan [12]. There are 1270 speech segments for training 1270 target models. Each speech segment consists approximately 2.5 minutes of speech in an average.

The male gender dependent UBM of 512 mixture components with diagonal covariance matrices, is trained using data from non-target speaker in NIST 2004 SRE. A 50 dimensional Linear Frequency Cepstral Coefficient (LFCC) feature vector (19 static, 19 Δ, 11 $\Delta\Delta$ and Δ energy) is extracted from speech signal at frame rate of 10 ms with 20 ms Hamming windowed over frequency brand 300-3400 Hz. Then Voice Activity Detection (VAD) is used to remove the less energize/silence frame from the feature vectors. Finally, silence-removed feature vectors are normalized to zero mean and unity variance normalization at utterance level.

For i-vector system, the *total variability space* T is trained using 12399 utterances from 890 non-target speakers (NIST 2004-05, Switchboard II part 1, 2 & 3; Switchboard cellular part 1 & 2, about 15 sessions per speaker). This data set is also used for implementing LDA, WCCN and Eigen Factor Radial (EFR) technique. 400 dimensional i-vectors are extracted from speech segments during training and testing phase.

The systems performance is evaluated using Equal Error Rate (EER) and Minimum Detection Cost Function (MinDCF) as per NIST plan [12].

7. RESULTS AND DISCUSSION

7.1. Selection of best m-vector system

In this section, we first evaluate the proposed m-vector systems and then best system is chosen from them. The best proposed system is used for comparing the performance with baseline system in next section.

Table 1 shows the speaker verification performance of the proposed systems for different length of m-vectors (column 2) and corresponding their optimal Linear Discriminant Analysis (LDA) projected dimension (column 3). It is better to mention that each subsystem has their own LDA projection matrix. For example, m-vector dimension of 500 yields 5 sub-systems in disjoint technique. The scores of all the sub-systems for particular LDA dimension are fused together (linear fusion), which is presented on the table.

Table 1. *EER and MinDCF of the proposed method for different size of m-vectors on NIST* 2008 *SRE core condition (male speakers and det7 condition).*

System	m-vector dim.	LDA Opt. dim.	EER (%)	MinDCF
Disjoint	2500	250	5.92	0.03296
	1250	200	5.69	0.0293
	500	**200**	**5.47**	**0.0268**
	250	200	5.69	0.0301
	125	100	7.52	0.0341
Overlapped	**500**	**350**	**4.78**	**0.0261**

From Table 1, the following observations can be drawn:

- *Disjoint segmentation method* shows best performance for the m-vector size of 500. This indicates that m-vector size of 500 is able to extract more speaker information from their MLLR super-vector compared to 2500, 1250, 250 and 125.
- *Overlapped window method* shows further reduction of EER and MinDCF compared to *disjoint segmentation technique*. This implies that *overlapped technique* is able to capture the speaker information in-between the adjacent two segments (i.e. two m-vectors), which is not captured in *disjoint case*.

Fig.4 shows the details performance of the sub-systems in *disjoint* (for m-vector of dimension=500) and *overlapped* case. From Fig.4, it can be noticed that the EER value of each sub-system is significantly higher than the fusion result of the sub-systems presented in Table 1. This indicates that each segments (i.e. m-vector) of the MLLR super-vector contains speaker specific information which are complementary for each another, and it is better to utilize the information available in different parts of MLLR super-vector separately.

7.2. Comparison of performance of Baseline system with Proposed method

In this section, we compare the speaker verification performance of the proposed best system (presented in Table 1) with baseline i-vector system.

Table 2 compares the performance of the best proposed method with baseline i-vector system. We can make the following observations from Table 2:

- Baseline system with LDA+WCCN performs better than WCCN alone. This indicates that post-processing task plays a significant role on i-vector system.

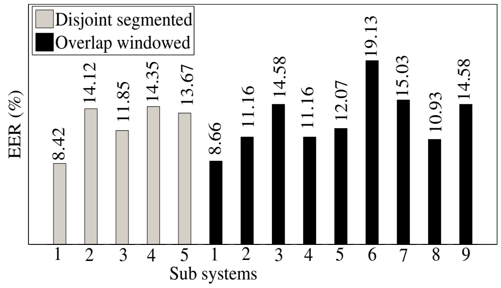

Fig. 4. *EER of each sub-system in* disjoint *(m-vector dim.*=500*) and* overlapped *techniques on NIST 2008 SRE core condition (male speakers only and det7 condition).*

Table 2. *Comparison of EER and MinDCF of baseline system with proposed best system on NIST 2008 SRE core condition (male speakers and det7 condition).*

System	EER (%)	MinDCF
Baseline:		
(**A2**) i-vector (dim.=400) + WCCN + Fast scoring	4.10	0.0185
(**A2**) i-vector (dim.=400) + LDA (Opt. dim.=200) + WCCN + Fast scoring	**3.63**	**0.0183**
Proposed:		
(**B**) Overlapped m-vector (dim.=500) + LDA (Opt. dim.=350)+EFR	**4.78**	**0.0261**
Fusion systems (**A2,B**)	**3.15**	**0.0155**

- Proposed method shows promising results comparable to the baseline i-vector system. However, it does not using any concept of *total variability space* and ASR transcription during MLLR super-vector estimation. This indicates that speakers can be characterized by simple vectors (i.e. m-vectors). Such system (with session variability compensation) can achieve comparable performance to the i-vector system.

  This also indicates that *total variability space* seems to play less important role in i-vector system compared to post-processing task (session variability compensation) on i-vectors.
- Fusion of systems **(A2)** (WCCN+LDA) with **(B)** further reduce the EER and MinDCF compared to baseline i.e. **(A2)**. This implies that the proposed system also contains complementary information for the i-vector system.

In Table 3, we present a m-vector system, where m-vectors are extracted by uniform segmentation of target speakers Gaussian Mixture Model (GMM) super-vectors [13]. This system is developed later submission of this paper. GMM super-vectors are derived from UBM with single iteration of Maximum a Posteriori (MAP) [14, 15] adaptation using speakers training data. During MAP adaptation, the value of relevance factor, 14 is considered. In our experiment, we use 50 dimensional feature vectors, which yields GMM super-vector size of 25600. In contrast to m-vector system using MLLR super-vector, Principal Component Analysis (PCA) is used to reduce the dimension of the m-vectors in this case. PCA gives the best performance. The performance of the system is shown in Table 3 for disjoint method. Since, overlapped method does not provide any further gain (system in this domain).

Table 3. *Comparison of EER and MinDCF of proposed m-vector system derived from GMM-super-vector with conventional i-vector system using EFR post-processing on NIST 2008 SRE core condition (male speakers and det7 condition).*

System	m-vector dim.	PCA Opt. dim.	EER (%)	MinDCF
Disjoint	1600	500	3.64	0.0221
	2560	900	3.42	0.0205
	3200	1000	3.42	0.0179
	6400	**800**	**2.96**	**0.0167**
	12800	800	3.18	0.0145
	25600	600	3.18	0.0158
i-vector (dim.=400)	-	-	2.05	0.0156

From Table 3, it is observed that proposed system (for m-vector size of 6400) gives performance which is comparable to i-vector system, even though both system uses same post-precessing and scoring techniques (i.e. EFR). It further indicates that post-processing task plays major rule in i-vector system than *total variability space*. Furthermore, advantage of the segmentation method is that it generates m-vectors which are much more smaller dimension than GMM super-vector. Hence, m-vector system is very easier to implement on device having less memory and computation power rather than i-vector system. It is important to note that LDA does not help in i-vector system.

It is to be noted that the results presented in this paper are obtained in the proposed method with two very simple *m-vector* extraction algorithms which possibly may not optimal. There may be lot of speaker related information in the MLLR super-vector, which is not capture by our simple algorithms.

8. CONCLUSION

In this paper, we proposed a m-vector based speaker verification system. The m-vectors are generated by uniform segmentation of the speakers MLLR super-vectors. An MLLR super-vector is estimated using *single* iteration of adaptation *without* considering ASR transcription and concept of the *total variability space* for a given speech data. m-vectors are then processed by conditioning algorithm to account the session variability. The proposed system shows promising performance compared to the conventional i-vector system on NIST 2008 SRE core condition, even though it is using a non-optimal *m-vector* extraction procedure. This indicates that speakers can be represented by simple m-vectors and post processing can yield performance comparable to the conventional i-vector system. Hence, *total variablity space* seems to play less important role in classical i-vector system. Besides, *total variablity space* training in i-vector system requires *computationally heavy iterative algorithm* and a large number of training data from different speakers over the various channels/sessions. However, the proposed system is very simple and can be easily implemented on small device having less memory and computing power for real-time application. Furthermore, fusion of the m-vector system with i-vector system allows a slight improvement of the performance compared to the i-vector system. We expect that better *m-vector* extraction technique will further improve the performance of the proposed method in future.

9. REFERENCES

[1] P. M. Bousquet, D. Matrouf, and J. F. Bonastre, “Intersession Compensation and Scoring Methods in the i-vectors Space for Speaker Recognition,” in *Proc. of Interspeech*, 2011, pp. 485–488.

[2] N. Dehak et al., “Front-End Factor Analysis for Speaker Verification,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, pp. 788–798, 2011.

[3] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, New York: John Wiley & Sons, 2001.

[4] A. Hatch, S. Kajarekar, and A. Stolcke, “Within-Class Covariance Normalization for SVM-Based Speaker Recognition,” in *Proc. of nt. Conf. Spoken Language Processing (ICSLP)*, 2006, pp. 1471–1474.

[5] M. Senoussaoui et al., “Mixture of PLDA Models in I-Vector Space for Gender-Independent Speaker Recognition,” in *Proc. of Interspeech*, 2011, pp. 25–28.

[6] Simon J.D. Prince, “Computer Vision: Models Learning and Inference,” in *Cambridge University Press, 2012, In press*.

[7] C. Leggetter and P. Woodland, “Maximum Likelihood Linear Regression for Speaker Adaptation of HMMs,” *Computer Speech and Language*, vol. 9, pp. 171–186, 1995.

[8] A. Stolcke et al., “MLLR Transforms as Features in Speaker Recognition,” in *Proc. of Eur. Conf. Speech Commun. and Tech. (EUROSPEECH)*, 2005, pp. 2425–2428.

[9] M. Ferras et al., “Constrained MLLR for Speaker Recognition,” in *Proc. of IEEE Int. Conf. Acoust. Speech Signal Processing (ICASSP)*, 2007, pp. 53–56.

[10] Z. N. Karam and W. M. Campbell, “A Multi-class MLLR Kernel for SVM Speaker Recognition,” in *Proc. of IEEE Int. Conf. Acoust. Speech Signal Processing (ICASSP)*, 2008, pp. 4117–4120.

[11] A. K. Sarkar and S. Umesh, “Eigen-voice Based Anchor Modeling System for Speaker Identification using MLLR Super-vector,” in *Proc. of Interspeech*, 2011, pp. 2357–2360.

[12] The NIST Year 2008 Speaker Recognition Evaluation Plan., “http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf,” .

[13] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, “Support Vector Machines using GMM Supervectors for Speaker Verification,” *IEEE Signal Process. Lett.*, vol. 13, pp. 308–311, 2006.

[14] J. L. Gauvain and C. H. Lee, “Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains,” *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[15] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker Verification using Adapted Gaussian Mixture Models,” *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.