COMPUTATIONALLY EFFICIENT SPEAKER IDENTIFICATION USING FAST-MLLR BASED ANCHOR MODELING

A. K. Sarkar[1], *S. Umesh*[2] *and J. F. Bonastre*[1]

[1]Universite D'Avignon (LIA), France; [2]IIT-Madras, India

sarkar.achintya@gmail.com, umeshs@ee.iitm.ac.in, jean-francois.bonastre@univ-avignon.fr

ABSTRACT

In this paper, we propose a computationally efficient method to identify a speaker from a large population of speakers. The proposed method is based on our earlier [1] Fast Maximum Likelihood Linear Linear Regression (MLLR) anchor modeling technique which provides performance comparable to the conventional anchor modeling system and yet reduces computation time significantly by computing likelihood efficiently using sufficient statistics of data and anchor specific MLLR matrix. However, both these systems still require a Gaussian Mixture Model-Universal Background Model (GMM-UBM) based back-end system to choose the optimal speaker, which is computationally heavy. In our proposed method, we show that applying Linear-Discriminant Analysis (LDA) and Within-Class-Covariance Normalization (WCCN) on the Speaker characterization Vector (SCV) of our recently proposed Fast-MLLR method, we can combine the computational efficiency and the discriminant capability to have a system that uses simple cosine-distance measure to identify speakers and yet has significantly superior performance compared to both full-blown GMM-UBM system and the anchor-model system. More importantly, there is no need of the "back-end" system. Experimental result on NIST 2004 SRE shows that the proposed method reduces identification error rate by an *absolute* 2% and takes *only* $2/3$ of the time taken by efficient Fast-MLLR system and *only* 20% of the time taken by the stand-alone GMM-UBM system.

Index Terms: Fast MLLR, WCCN, LDA, anchor model, speaker identification

1. INTRODUCTION

Speaker Identification (SI) is the task of identifying a speaker from a (closed) set of speakers in a database. This is in contrast to the binary-hypothesis problem of speaker-verification where we have to accept or reject a claimant speaker. Speaker identification is done by matching the test utterance with the known registered speaker models in the database. It can be mathematically expressed as,

$$\hat{S} = \arg \max_{1 \leq S \leq L} P(X|\lambda_S) \quad (1)$$

It is quite common to build speaker-models, λ_S, in a GMM-UBM frame-work. From Eqn.(1) it can be seen that the computation time of the system increases as the number of speakers in database, L, increases. This is especially problematic in identifying speakers in a large population.

Sturim *et al.* [2] proposed an approach called Cascade Anchor-modeling system to reduce computation as well as get performance comparable to GMM-UBM system of Eqn.(1) for speaker indexing in a database of large population. This is illustrated in Fig.1. In this cascade approach, the computationally-efficient front-end system selects the N-most probable speakers for the back-end GMM-UBM system to find the best speaker from the reduced set. Recently, we introduced a computationally efficient anchor modeling technique based on MLLR and sufficient statistics called *Fast-MLLR anchor system* [1]. However, the disadvantage of the both these anchor-modeling techniques [1, 2] is that they still need the computationally expensive GMM-UBM based back-end system.

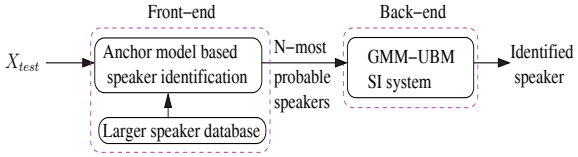

Fig. 1. *Cascade speaker identification system using anchor modeling technique.*

The motivation of the method proposed in this paper is to:

- eliminate the computationally heavy back-end GMM-UBM system in anchor modeling technique
- exploit the advantage of computational efficiency of our previously proposed Fast MLLR based anchor modeling technique [1]
- obtain performance that is better or comparable to the standalone GMM-UBM based system or cascade anchor modeling system

We show that by exploiting the discriminant ability of Linear Discriminant Analysis (LDA) and Within-class Covariance Normalization (WCCN) combined with the computational efficiency of Fast-MLLR, we can achieve all the above objectives.

Several techniques have been proposed in literature to reduce the computation cost of speaker-identification systems. The most commonly used GMM-UBM framework speaker identification approach is described in [3], where speaker models are adapted from GMM-UBM using Maximum a Posteriori (MAP) adaptation. Therefore, there is a correspondence between Gaussian components of the GMM-UBM and the speaker models. During test, the utterance is first aligned with respect to GMM-UBM to find the top-C best mixture components per feature vector. These same top-C components are then traversed through speaker models in the database to calculate the likelihood of the speaker. The saving in computation comes from avoiding less important mixture components during testing. Recently i-vector concept has shown great sucessful in speaker verification task [10]. Since the proposed method is based on without concept of *total variability space*, i-vector system is not considered as the scope of this paper.

Some of the other methods include pruning [4, 5], speaker cluster selection based method [6], pre-quantization [7] and Hash model [8] in GMM-UBM based speaker identification system. In most of these methods [4, 5, 6], the accuracy of their method and computation time are compared with calculating the likelihood from the speaker models *considering all Gaussian components* (i.e. without top-C fast scoring method). Further, none of these methods [4, 5, 6] give better accuracy as well as provide saving in computation time.

The paper is organized as follows: Section 2 and 3 describe the conventional and Fast-MLLR anchor modeling techniques. Our proposed method is described in Section 4. Experimental setup and baseline systems are described in Section 5. Section 6 describes the selection of optimal LDA dimension. Results and discussion are presented in Section 7. Finally, in Section 8 we provide our conclusions.

2. CONVENTIONAL ANCHOR SYSTEM

In Speaker identification using anchor modeling technique [2, 9], during training, evaluation speakers are represented by Speaker Characterization Vector (SCV) with respect to anchor models, i.e.

$$SCV_S = [\tilde{p}(X|\lambda_1)\ \tilde{p}(X|\lambda_2)\ \ldots\ \tilde{p}(X|\lambda_E)] \tag{2}$$

$\tilde{p}(X|\lambda_E)$ is the normalized log-likelihood ratio of the data X (of T feature vectors) with respect to E^{th} anchor model (λ_E) and GMM-UBM, i.e.

$$\tilde{p}(X|\lambda_E) = \frac{1}{T}\ [log\ p(X|\lambda_E) - log\ p(X|\lambda_{UBM})] \tag{3}$$

Therefore, the normalized likelihood is calculated only for the anchor speakers, whose number are usually significantly lower than the number of speakers in the population providing large gain in computation time. A similar concept is used in eigen-voices for speaker-adaptation and i-vector in speaker verification [10].

In test phase, the SCV, SCV_t, corresponding to the utterance from unknown speaker is compared to all speaker specific SCVs, SCV_S (obtained during training) in the database using a *simple* cosine angle similarity measure:

$$\hat{S} = \arg \min_{1 \le S \le L}\ arc\ cosine(SCV_t, SCV_S) \tag{4}$$

where $\hat{S}$ is the identified speaker of unknown test utterance. Since only a simple but computationally inexpensive measure is used, the identification accuracy of this stage is low. Therefore, N-most probable speakers are selected from this stage to find the optimal speaker using GMM-UBM based back-end system on this *reduced* set. This combination gives the advantage of less computational cost of anchor system as well as greater accuracy of GMM-UBM based system.

3. FAST MLLR ANCHOR SYSTEM

In this method, anchor speakers are represented by MLLR [11] matrices instead of GMMs. The anchor specific MLLR matrix is estimated with respect to GMM-UBM using data from the anchor speaker. More details of this approach can be found in [1]. For E anchor speakers, E number of MLLR matrices ($W_1, \cdots, W_E$) are computed during training phase. The SCV of a speech segment is *efficiently* calculated using anchor specific MLLR matrix and sufficient statistics accumulated from the data. Following steps are involved in likelihood calculation:

Initialization: Load the MLLR matrices of all anchor speakers

Step1: Determine the probabilistic alignment, $\gamma_j\,(t)$ of the training or test feature vectors, $X = \{x_1, x_2, \ldots, x_T\}$ for j^{th} components of GMM-UBM.

Step2: Accumulate the two sufficient statistics for i^{th} dimension of the feature vectors as,

$$K^{(i)} = \sum_{j=1}^{M}\sum_{t=1}^{T}\gamma_j(t)\ \frac{1}{\sigma_{j,i}^2}\ x_i(t)\ \mu_j^{'} \tag{5}$$

$$G^{(i)} = \sum_{j=1}^{M}\frac{1}{\sigma_{j,i}^2}\mu_j\mu_j^{'}\sum_{t=1}^{T}\gamma_j\,(t) \tag{6}$$

where $(.)^{'}$ indicates matrix transpose operation. K and G are the sufficient statistics estimated from speech segments and are *not specific to any particular speaker*. σ_j, μ_j indicate the variance and mean of j^{th} mixture respectively.

Step3: The likelihood of the speech sample is calculated using MLLR matrix W_E and sufficient statistics as follows:

$$p(X; W_E) = \left\{ -\frac{1}{2}\left\{ \sum_{i=1}^{D}(w_{e,i}G^{(i)}w_{e,i}^{'} - 2K^{(i)}w_{e,i}^{'}) \right\} \right\} \tag{7}$$

where, $w_{e,i}$ is the i^{th} row of MLLR matrix (W_E) and D is the dimension of the feature vector. This is computationally very efficient since likelihood calculation involves matrix multiplication with sufficient statistics as seen in Eqn.(7).

The SCV of the Fast MLLR anchor system is formed similar to Eqn.(2), with the elements being $\tilde{p}(X; W_E)$, where,

$$\tilde{p}(X; W_E) = \frac{1}{T}\ [log\ p(X; W_E) - log\ p(X; W_{glb})]\,. \tag{8}$$

W_{glb} is the global MLLR matrix estimated by pooling data from all the training speakers used to build the GMM-UBM.

4. PROPOSED METHOD OF FAST MLLR+LDA+WCCN

4.1. Linear Discriminant Analysis

In anchor-modeling frame-work, each speaker is characterized by the SCV vector analogous to the i-vector used in speaker-verification. Therefore, we can apply Linear-Discriminant Analysis (LDA) to reduce the dimension and increase discriminability among speaker classes. LDA is applied on the SCV, which is calculated using the computationally efficient Fast MLLR anchor system discussed in the previous section. LDA projection matrix (A) of SCVs is found by maximizing the ratio of between-class scatter (S_B) and within-class scatter (S_W) matrices, i.e.

$$\max_A\ J(A) = \frac{A^{'}S_BA}{A^{'}S_WA} \tag{9}$$

The S_W and S_B are defined for c speaker classes as follows,

$$S_W = \sum_{k=1}^{c}\sum_{z\in k}(z - m_k)(z - m_k)^{'} \tag{10}$$

$$S_B = \sum_{k=1}^{c} n_k\ (m_k - m)(m_k - m)^{'} \tag{11}$$

$$m_k = \frac{1}{n_k}\sum_{z\in k} z\ ; \quad m = \frac{1}{n}\sum_{z\forall c} z$$

where, m_k and m are the mean of k^{th} class and global mean taking data from all speaker classes respectively. z represents the SCV and n_k is the number of SCV examples that belong to k^{th}speaker class. The solution of Eqn.(9) reduces to the problem of maximum eigenvalue of $S_W^{-1}S_B$ and the optimal columns of A are the eigenvectors corresponding to the largest eigenvalues. In our experiment, we set n_k to unity in Eqn.(11) to gives equal weight to all speaker classes. SCV examples of each speaker are considered as single class during estimation of matrix A.

4.2. WCCN

LDA projection assumes equal within-speaker-class covariance matrices. To further minimize the effect of different within-class covariances, we applied WCCN [12] on LDA projected SCV. In WCCN, a projection matrix B is found by *cholesky decomposition* of W^{-1}=$BB^{'}$. W is the within-class scatter matrix and is expressed as,

$$W = \frac{1}{c}\sum_{k=1}^{c}\sum_{z\in k}(Az-\hat{m}_k)(Az-\hat{m}_k)^{'} \quad (12)$$

$$\hat{m}_k = \frac{1}{n_k}\sum_{z\in k}Az$$

where, $\hat{m}_k$ is the mean of LDA projected SCV of k^{th} speaker class. The projected $\overline{SCV}$ with LDA of dimension 150 followed by WCCN is,

$$\overline{SCV}_{[150\times 1]} = B_{[150\times 150]}\ A_{[150\times 346]}\ SCV_{[346\times 1]} \quad (13)$$

In our proposed method, speakers are finally represented by their LDA+WCCN project SCV i.e. $\overline{SCV}$ during training. Similarly during test, the LDA+WCCN projected SCV of the test utterance is used for cosine angle similarity.

$$\hat{S} = \arg\min_{1\le S\le L} arc\ cosine(\overline{SCV_t}, \overline{SCV_S}) \quad (14)$$

Note that we do not use any *further* back-end system.

5. EXPERIMENT SETUP

We compare the performance of our proposed method for speaker identification with stand-alone GMM-UBM based system with top-C scoring [3, 13], conventional [2] and Fast MLLR [1] based cascade anchor systems. The speaker models in case of GMM-UBM based system and anchor modeling systems, are adapted from the GMM-UBM with MAP adaptation using speaker's training data. Similarly, the anchor models for conventional anchor system are derived from GMM-UBM using MAP. Only mean parameters of the GMM-UBM are adapted during MAP adaptation in all cases. The value of relevance factor is set to 16 in all cases. During test phase, top-C=15 mixtures per feature vector are considered for likelihood calculation from the speaker models to create the SCV of test data. We used C=15 since it gives the best results in our experiment setup.

All speaker-identification experiments are performed using speakers from NIST 2004 SRE core condition (i.e. 1-side training and 1-side test condition) as belonging to the population. The database contains 310 speakers. There are 306 speakers having both training and test examples. Therefore, for the closed set speaker identification task, we consider these 306 speakers who have both training and test utterances. The experimental setup results in 1163 utterances for test.

1346 (655 male, 691 female) anchor speakers are taken from NIST-1999, 2001 SRE and speakers in training data of GMM-UBM. This ensures that they can cover a large acoustic space.

39 dimensional MFCC feature vectors (C_1 to C_{13} with Δ and $\Delta\Delta$ excluding C_0) are extracted from speech signal sampled at 8 kHz with 10 ms frame-rate and 20 ms Hamming window using the frequency band 300-3400 Hz. Two different frame removal techniques are followed [14] to remove the silence/less energy frames. Bi Gaussian modeling of energy components of the frames is applied for NIST 1999, 2001, 2002 SRE and Switchboard-1 Release-2, and tri Gaussian modeling of normalized energy components of the frames for NIST 2004 SRE. Silence-removed feature vectors are normalized to zero-mean and unit-variance at utterance level.
The GMM-UBM with 2048 mixture components of diagonal covariance matrices is trained using data from NIST 2002 and Switchboard-1 Release-2.

For discriminant analysis to calculate the transformation matrix, we consider data of the evaluation speakers from 3, 8, 16 training sides condition of the database. The silence-removed and normalized feature vectors are then segmented into 30 seconds intervals to estimate a set of SCVs for a particular speaker. This yields about 15-100 SCVs per speaker class.

Experiments are run on a desktop computer having Intel core $i7$ CPU 860 @2.80 GHz and 8 GB RAM. The program are implemented in Matlab software in contrast to [1]. To assess the computation complexity, we measure the relative time taken to process the data on identical computer setup by the different approaches.

The optimal anchor speaker set is selected from 1346 anchor speakers in the database. It was found in [1] that 346 anchor speakers provide the best representation in the space for the same experimental setup. More details can be found in [1]. Hence, in our experiments the size of the SCV ($SCV_{[346\times 1]}$) is 346 *without* LDA+WCCN.

6. SELECTION OF OPTIMAL LDA DIMENSION

In this section, we find the optimal LDA projection matrix, A which yields the best discrimination among the speaker classes. The optimal dimension is chosen based on the speaker Identification Error Rate (IER) of system which is defined as $(100 - accuracy)\%$.

Table 1. *Speaker identification error rate for different LDA projected dimension of SCV in proposed method.*

	LDA projected dimension of SCV					
	50	100	150	200	250	300
IER (%)	48.93	45.74	**45.40**	49.01	50.82	58.56

Table.1 shows the IER for different LDA dimensions. It is observed from Table.1 that LDA projection dimension of 150 gives the lowest IER and hence considered as the optimal LDA projection dimension. The LDA matrix corresponding to dimension 150 i.e. $A_{[150\times 346]}$ is selected as the best transformation matrix.

The LDA-projected SCV i.e. $A_{[150\times 346]}SCV_{[346\times 1]}$ is used for estimation of WCCN transformation/projection matrix, $B_{[150\times 150]}$ using Eqn.(12).Finally, speakers are represented by LDA+WCCN projected vector, $\overline{SCV}_{[150\times 1]}$.

7. RESULTS AND DISCUSSION

Fig.2 shows the comparison of speaker Identification Error Rate (IER) of the proposed method against stand-alone GMM-UBM and cascade anchor-model systems. Although our proposed and stand-alone GMM-UBM (using Eqn.(1)) directly return the optimal speaker, the cascade-systems first find the N-most probable speakers and then find the optimal speaker using GMM-UBM system. For the cascade systems, we show the performance for N=5 and N=10. The performance of the cascade anchor systems approach that of GMM-UBM system as N increases. However, the cascade anchor system never performs better than the standalone GMM-UBM system since their back-end system *is the* GMM-UBM system. Fig.2 also shows the performance when LDA is applied on the SCV of front-end in the cascade systems. From the figure, it can be seen that the proposed method significantly reduces IER compared to *all* the other systems. Fig.3 compares the average computation time required to identify the speaker using the proposed method and the other systems. The stand-alone GMM-UBM is computationally expensive because the

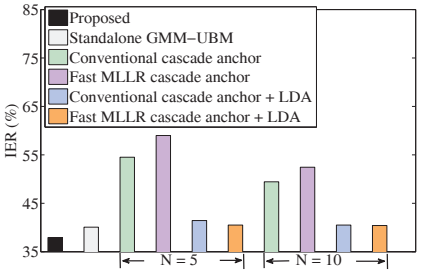

Fig. 2. *Comparison speaker identification error rate of proposed method with baseline systems.*

likelihood evaluations are done with respect to *all* the speaker models. For the cascade anchor systems the computation increases as number of N-most probable speakers increases. In our experiments, it so happens that the number of anchor models (i..e 346) are larger than the evaluation speakers (i.e. 306). Hence, conventional cascade system takes more time when compared GMM-UBM stand-alone system. The real benefit of conventional cascade becomes obvious when the database of evaluation speakers is very large [2] (say, 10,000).

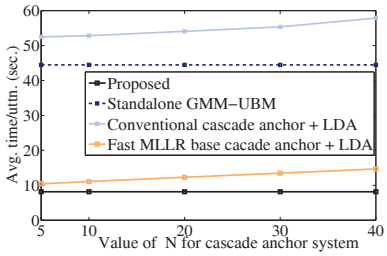

Fig. 3. *Comparison computation time required to identify the speaker of proposed method with baseline systems.*

From Fig. 2 and Fig. 3 we observe that the proposed method significantly reduces IER as well as provides significant gain in computation time. The main reason for the gain in computation time of the proposed method, is that there is no need of a GMM-UBM back-end system.

Table.2 summarizes the results and discussion of the above figures. It can be observed that proposed method shows *absolute* IER reduction of more that 2% compared to other systems, as well as takes only $2/3$ of the time taken by the already efficient Fast-MLLR system.

Table 2. *Comparison of different systems (N=10).*

System	IER (%)	IER Reduc. by **(a)** over	Avg. time/ uttn. (sec.)	time saving by **(a)** over
Proposed **(a)**	37.92	-	8.15	-
Standalone GMM-UBM	40.07	2.15	44.47	36.32
Con. cas. anchor+LDA	40.50	2.58	52.83	44.68
Fast MLLR cas. anchor+LDA	40.41	2.49	11.07	2.92

It is to be noted that all the systems i.e. stand-alone GMM-UBM, cascade and Fast MLLR cascade anchor systems require *only one* alignment of test data to identify the speaker.

8. CONCLUSION

In this paper, we have combined the discriminant ability of LDA and WCCN with the computational efficiently of our recently proposed Fast-MLLR system and shown that we can get better performance than stand-alone GMM-UBM or conventional anchor-modeling systems. More importantly, since only a simple cosine-distance measure is used and no GMM-UBM back-end, it is even more computationally efficient than our recently proposed Fast-MLLR system. The discriminant analysis is done on the speaker-characterization vector obtained from the front-end of the Fast-MLLR system. Cosine-similarity is used on these features to find the optimal speaker. Experimental result on NIST 2004 SRE shows that the proposed method reduces identification error rate by an *absolute* 2% and takes *only* $2/3$ of the time taken by efficient Fast-MLLR system and *only* 20% of the time taken by the stand-alone GMM-UBM system.

9. REFERENCES

[1] A. K. Sarkar and S. Umesh, "Fast Computation of Speaker Characterization Vector using MLLR and Sufficient Statistics in Anchor Model Framework," in *Interspeech*, 2010.

[2] D. Sturim et al., "Speaker Indexing in Large Audio Databases using Anchor Models," in *Proc. of ICASSP*, 2001, pp. 429–432.

[3] D. A. Reynolds, T. F. Quatieria, and R. B. Dunn, "Verification using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[4] B. L. Pellom and J. H. L. Hansen, "An Efficient Scoring Algorithm for Gaussian Mixture Model Based Speaker Identification," *IEEE Signal Proc. Lett.*, vol. 5, pp. 281–284, 1998.

[5] T. Kinnunen, E. Karpov, and P. Franti, "A Speaker Pruning Algorithm for Real-Time Speaker Identification," in *Proc. Audio- and Video-Based Biometric Authentication*, 2003, pp. 639–646.

[6] V. R. Apsingekar and P. L. De Leon, "Speaker Model Clustering for Efficient Speaker Identification in Large Population Applications," *IEEE Trans. Speech, Audio, Lang. Proc.*, vol. 17, pp. 848–853, 2009.

[7] J. McLaughlin, D. A. Reynolds, and T. Gleason, "A Study of Computation Speed-ups of the GMM-UBM Speaker Recognition System," in *Proc. of Eurospeech*, 1999, pp. 1215–1218.

[8] R. Auckenthaler and J. S. Mason, "Gaussian Selection applied to Text-Independent Speaker Verification," in *Proc. of Odyssey*, 2001, pp. 83–88.

[9] Y. Mami and D. Charlet, "Speaker Recognition by Location in the Space of Reference Speakers," *Speech Communication*, vol. 48, pp. 127–141, 2006.

[10] N. Dehak et al., "Front-End Factor Analysis for Speaker Verification," *IEEE Trans. on Speech, Audio, Lan. Proc.*, vol. 19, pp. 788–798, 2011.

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

[12] A. Hatch et al., "Within-Class Covariance Normalization for SVM-Based Speaker Recognition," in *Proc. of ICSLP*, 2006.

[13] A. K. Sarkar, S. Umesh, and S. P. Rath, "Computationally Efficient Speaker Identification for Large Population Tasks using MLLR and Sufficient Statistic," in *Proc. of Odyssey*, 2010.

[14] J. F. Bonastre et al., "Nist'04 Speaker Recognition Evaluation Campaign: New LIA Speaker Detection Plateform based on ALIZE Toolkit," in *Proc. of NIST 2004 Speaker Recognition Workshop*, 2004.