Eigen-voice Based Anchor Modeling System for Speaker Identification using MLLR Super-vector

A. K. Sarkar and S. Umesh

Department of Electrical Engineering, Indian Institute of Technology Madras, India
sarkar.achintya@gmail.com, umeshs@ee.iitm.ac.in

Abstract

In this paper, we propose an anchor modeling scheme where instead of conventional “anchor” speakers, we use eigenvectors that span the Eigen-voice space. The computational advantage of conventional Anchor-modeling based speaker identification system comes from representing all speakers in a space spanned by a small number of anchor speakers instead of having separate speaker models. The conventional “anchor” speakers are usually chosen using data-driven clustering and the number of such speakers are also empirically determined. The use of proposed eigen-voice based anchors provide a more systematic way of spanning the speaker-space and in determining the optimal number of anchors. In our proposed method, the eigenvector space is built using the Maximum Likelihood Linear Regression (MLLR) super-vectors of non-target speakers. Further, the proposed method does not require calculation of the likelihood with respect to anchor speaker models to create the speaker-characterization vector as done in conventional anchor systems. Instead, speakers are characterized with respect to eigen-space by projecting the speaker’s MLLR-super vector onto the eigen-voice space. This makes the method computationally efficient. Experimental results show that the proposed method consistently performs better than conventional anchor modeling technique for different number of anchor speakers.

Index Terms: Eigen voice, anchor model, speaker identification, MLLR super-vector

1. Introduction

Anchor modeling technique has been shown to be useful in speaker/audio indexing [1], speaker recognition [2, 3] and language identification [4]. In this approach, speakers/audio files are represented with respect to a set of pre-defined speaker/audio models. The pre-defined speaker models are called *reference speakers* or *anchor models*. The computational advantage comes from using a small number of anchor models instead of separate models for the entire population. In this paper, we focus on the speaker-identification problem.

Most implementations of anchor-modeling systems [1, 2, 3, 4] maintain a set of anchor speaker models and represent the speakers by calculating likelihood with respect to the anchor models, i.e. each speaker is represented by a Characterization Vector (CV) given by,

$$CV^k = [\tilde{p}(X|\lambda_1)\; \tilde{p}(X|\lambda_2)\; \ldots\; \tilde{p}(X|\lambda_E)] \tag{1}$$

CV^k represents the CV of speaker, k and $\tilde{p}(X|\lambda_E)$ is the normalized log-likelihood ratio of the data X (of T feature vectors) with respect to E^{th} anchor model and the Gaussian Mixture Model-Universal Background Model (GMM-UBM), λ_{UBM}, i.e.

$$\tilde{p}(X|\lambda_E) = \frac{1}{T}\,[log\; p(X|\lambda_E) - log\; p(X|\lambda_{UBM})] \tag{2}$$

Recently, we [3] proposed a variant of the above method using MLLR and sufficient statistics, where anchor speakers are represented by MLLR matrices instead of conventional GMMs. The characterization vector consisting of likelihoods is efficiently calculated using anchor specific MLLR matrix and sufficient statistics accumulated from the speaker data by aligning with respect to the GMM-UBM.

In both of these approaches, the selection of the anchor models or space (i.e. number of anchor models) plays a great role in system performance. There are several approaches available in literature to select the set of anchor speakers either using *iterative clustering algorithm* [2] or *random approach* [2, 3]. However, these are based on heuristic principles and may not provide the optimal solution for selecting the anchor speaker models and the space.

In this paper, we propose to represent the speakers in an orthogonal eigenvector space built using eigen value decomposition of pooled MLLR-super-vectors [5] from a number of non-target speakers. The number of eigen vectors chosen can be thought of as the number of anchor speakers used in conventional anchor modeling technique. In our proposed method, speaker k is represented by characterization vector, y^k, which is estimated by solving Eqn.(4) as,

$$\begin{aligned} M^{sup^k} &= By^k && (3)\\ &= [e_1\, e_2\, \ldots\, e_E]\, y^k && (4) \end{aligned}$$

where, B is a matrix of dimension $[DD \times E]$ containing top-E eigen vectors. D denotes the dimension of Mel-Frequency Cepstral Coefficient (MFCC) feature vectors. M^k_{sup} is the MLLR super-vector estimated with respect to GMM-UBM using a particular speaker’s training data. Note that the elements of the characterization vector are projections onto the orthogonal eigen-vector space instead of likelihoods calculated with respect to anchor models as in [1, 2, 3, 4].

Similarly, in test phase, first the characterization vector of test speech segment, y^t is estimated using MLLR super-vector of test data and projection matrix, B. Then, the best matched speaker in the database is found using cosine-similarity measure between y^k and y^t.

The proposed method is very similar to the recently proposed i-vector [6] based speaker-verification technique, where the speakers are represented by a characterization vector which is estimated in Joint Factor Analysis (JFA) frame-work. A similar approach is also used for rapid speaker adaptation [7]

This work is supported by SERC project funding SR/S3/EECE/058/2008 from Ministry of Science & Technology, India.

in speech recognition. The adapted model parameters of the speakers are formed by a linear combination of a number of reference eigen vectors [7] or speaker model parameters [8]. The eigen-vector space in [7] is built by eigenvector decomposition of the super vectors corresponding to a number of speaker dependent models. The speaker dependent models are adapted from the Speaker Independent (SI) Hidden Markov Model (HMM).

The eigen voice analysis of the proposed method is similar to [7, 9]. Unlike our proposed method of using MLLR supervectors, in [9] the MFCC feature vectors are projected onto the eigen space. Further, the projected features are used to train the GMM for individual speakers. In [7] the speaker dependent model parameters are used to find the eigenvectors, where as, in our case we use the MLLR super vectors to find the eigen vectors. We will refer to our proposed method as *Eigen voice anchor system* in this paper.

We compare the performance of our proposed method with conventional anchor modeling systems on a closed-set speaker identification task. The experimental results using speaker from NIST 2004 SRE core condition show that proposed method significantly performs better than conventional anchor systems.

The paper is organized as follows: In the next section, we describe our proposed method. The conventional anchor-based and Fast-MLLR systems are described in Section 3. Details of the experimental setup are provided in Section 4. Results and discussion are described in Section 6 and 7. Finally, the conclusion of the paper are presented in Section 8.

2. Eigen-voice Anchor System

In this section, we describe our proposed Eigen-voice based anchor modeling system. An important step in this approach is to identify the eigen-voice space and project the MLLR super-vectors onto this estimated low-dimensional space. We now provide details of the three steps of the proposed system: (i) estimation of eigen-voice space (ii) obtaining the speaker characterization vector, and (iii) finding the optimal speaker using a similarity measure.

2.1. Estimation of eigen-voice space

MLLR [10] is a speaker adaptation technique that is commonly used in Automatic Speech Recognition (ASR) to adapt the Speaker Independent (SI) model towards the speaker in a Maximum Likelihood (ML) sense using speaker's training (adaptation) data. It can be expressed as,

$$\hat{\mu} = W\mu + b, \quad \hat{\Sigma} = \Sigma \tag{5}$$

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

In our speaker-identification frame-work, MLLR transformations are estimated for a number of non-target speakers with respect to the speaker independent GMM-UBM (which is analogous to SI system in speech recognition). The coefficients of the MLLR transformation are stacked one-by-one to form the MLLR super vector [5] of a particular speaker. Bias parameter b is not considered in our experiment. For D dimension of MFCC feature vector, the MLLR super-vector will be $[DD \times 1]$ vector. Fig.1 shows the steps for MLLR super-vector estimation for speaker, k.

Let us define a matrix, M whose columns contain the MLLR super-vector of all the p speakers, i.e.

$$M_{[DD\times p]} = \left[M^{sup^1} \; M^{sup^2} \; \ldots \; M^{sup^p} \right] \tag{6}$$

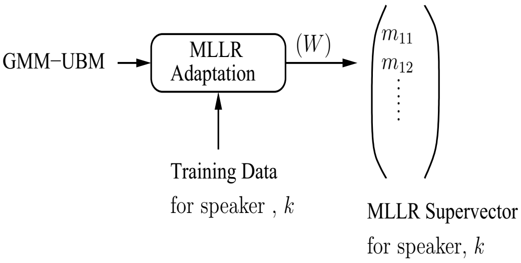

GMM−UBM
MLLR Adaptation
(W)
m_{11}
m_{12}
Training Data for speaker , k
MLLR Supervector for speaker, k

Figure 1: *MLLR super-vector.*

where, M^{sup^p} indicates the MLLR super-vector of non-target speaker, p. $M_{[DD\times p]}$ is a rectangular matrix. We apply Singular Value Decomposition (SVD) on M itself, i.e.

$$M_{[DD\times p]} = U_{[DD\times DD]} Q_{[DD\times p]} V^{*}_{[p\times p]} \tag{7}$$

where, U and V^* are unitary matrices and Q is a diagonal matrix. The E singular vectors of U corresponding to the E largest singular values point to the E most dominant directions in MLLR super-vector space. Let columns of B correspond to the singular vectors corresponding to largest E eigen values of U, i.e. $B_{[DD\times E]}$. The number of columns of the B is equivalent to number of anchor speakers in conventional anchor system, and the columns of B span the eigen-voice space.

2.2. Obtaining the Characterization Vector

During training, evaluation speakers are represented by Characterization Vectors (CVs) which are obtained by projecting their MLLR super-vector on space spanned by eigen vectors of B. *Algorithm* 1 describes the steps involved in estimating the speaker characterization vector for given speech segment of speaker, k.

Algorithm 1: Characterization Vector Computation

Step1 : Align the feature vectors of speaker, k with respect to GMM-UBM and obtained the MLLR matrix, W_k using Eqn.(5).

Step2 : Form MLLR super-vector of the k^{th} speaker, M^{sup^k} as shown in Fig.1.

Step3 : Obtained the characterization vector, y^k of k^{th} speaker for E eigen vectors of B using,

$$y^k_{[E\times 1]} = B^{+}_{[E\times DD]} M^{sup^k}_{[DD\times 1]} = (B^H B)^{-1} B^H M^{sup^k} \tag{8}$$

where, B^+ is the pseudo-inverse of B, which needs to be computed *only once* and stored – resulting in computational efficiency.

2.3. Finding Optimal Speaker

In test phase, the characterization vector, y^t of test utterance is calculated using *Algorithm-1*. Then the speaker is identified by finding the cosine angle similarity measure between CV of register speakers (obtained during training) in the database and y^t. The best matched speaker is recognized as the identified speaker, $\hat{k}$ of the test utterance, i.e.

$$\hat{k} = \arg \min_{1\leq k\leq L} arc\ cosine(y^t, y^k) \tag{9}$$

where, L is the number of speakers in database.

3. Baseline system

To compare the performance of our proposed method, we consider conventional GMM-UBM [1, 2] based anchor modeling system as well as recently proposed Fast MLLR Anchor system [3] as a baseline. In GMM-UBM based anchor technique, the anchor speakers are adapted from the GMM-UBM using Maximum a Posteriori (MAP) adaptation technique. Only mean parameters of GMM-UBM are adapted during MAP adaptation. The value of relevance factor is fixed at 16. The training/test characterization vector of the speech segment is calculated with top-C=15 fast scoring method [11] with respect to anchor models as described in Eqn.(1). The speaker of the test utterance is identified using cosine-similarity, i.e. Eqn.(9) between speaker specific-CV (obtained during training) and CV of test data.

In Fast MLLR anchor system, anchor speakers are represented by MLLR matrices instead of GMMs. The training/test characterization vector of the speech segment is estimated using anchor specific MLLR matrix and sufficient statistics accumulated from speech segment with respect to GMM-UBM. Similar to conventional anchor system, speaker of the test utterance is identified using cosine-similarity measure between CV of speakers and test segment. More details of Fast-MLLR method can be found in [3].

4. Experimental setup

The speaker-identification experiments are performed using speakers from NIST 2004 SRE core condition. The database contains 310 speakers including 124 male and 186 female speakers. There are only 306 speaker having both training and test utterances. Therefore, for a closed-set speaker identification task, we considered only these 306 speakers who have both train and test utterances. 1346 speakers (655 male and 691 female) are selected from NIST-1999, 2001 and speakers in training data of GMM-UBM for building anchor models and for finding the eigen-voice space. The different number of anchor speakers are selected from the above number of speakers.

For cepstral analysis, 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 using 20 ms Hamming window over the frequency band 300-3400 Hz. To remove the silence or less energy frames, two different frame removal techniques are followed [12]. Bi Gaussian modeling of energy components of the frames is applied for NIST 1999, 2001, 2002 SRE and Switchboard-1 Release-2. The tri Gaussian modeling of normalized energy components of the frames for NIST 2004 SRE. Silence-removed feature vectors are normalized to fit zero-mean and unit-variance at utterance level.

The GMM-UBM with 2048 mixture components is trained using data from NIST 2002 and Switchboard-1 Release-2 database. The covariance matrices of the Gaussian components are considered diagonal.

5. Selection of anchor speaker models

In this section, we describe how to select the desired number of anchor speakers from among the 1346 speakers in the database for anchor modeling systems. There are several algorithms available in literature to choose the anchor sets [2]. We consider the random clustering [2, 3] approach to select the anchor speakers. *Algorithm 2* describes the steps involved in the random selection of desired anchor speakers (see [3]).

Algorithm 2: Random selection of anchor speakers

Step 1: Let E_{anch} be the number of anchor models (e.g. 46, 146 etc.)

Step 2: Randomly select 10 different set of anchor speakers, with each anchor set having E_{anch} speakers from the entire speaker set (i.e. 1346 speakers).

Step 3: Perform speaker identification for each anchor set of E_{anch} models.

Step 4: Calculate average accuracy over the 10 sets of anchor speaker.

Step 5: Repeat Step 1 to 4 for each value of E_{anch} anchors

The value of E_{anch} is varied from 46 to 446 in steps of 100.

6. Results and discussion

Fig.2 shows the comparison of the speaker Identification Error Rate (IER) of proposed method with the other systems for different number of anchor speakers. In our system, we choose the corresponding number of top E singular vectors to match the number of anchor speakers used. It can be observed that proposed method performs significantly better than the two other methods. The proposed method shows *absolute* reduction of IER (%) by 7.01, 9.33, 10.33 and 11.2 compared to GMM-UBM anchor system for 46, 146, 246 and 346 anchor models respectively. Similarly for Fast MLLR anchor system, it shows 11.45, 15, 17.51 and 18.6 IER (%) reduction.

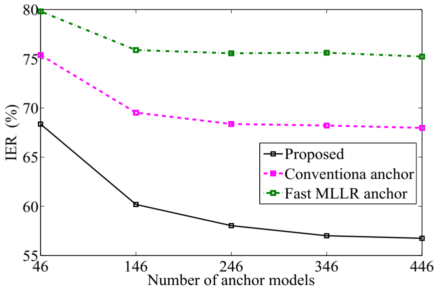
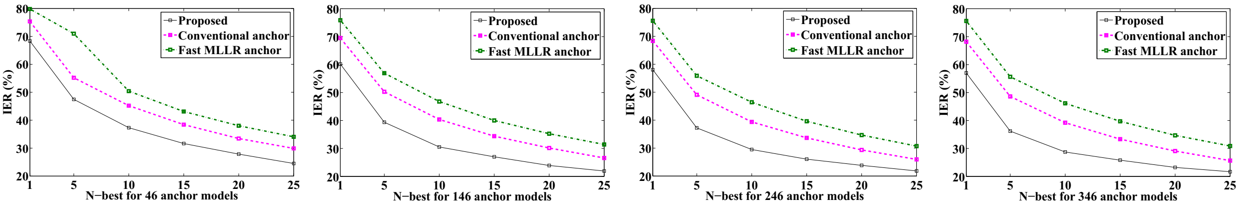


Figure 2: *Comparing speaker identification error rate of proposed method with baseline anchor systems for varying number of anchor models.*

Generally, anchor modeling technique is used as a front-end system to select a number of most probable speakers from a larger database. The optimal speaker is then found from the reduced set of the speakers using GMM-UBM based back-end system. Therefore, we investigate the N-best speaker identification performance of the different methods. Fig.3(a)-3(d) shows the N-best speaker identification performance of proposed method against the two other systems for 46, 146, 246 and 346 anchor speakers respectively. We observe from Fig.3(a)-3(d), that the proposed method consistently shows better performance for a wide range of N-best values when compared to the conventional and Fast-MLLR systems. This is because in the proposed method, the characterization vectors are calculated with respect to the orthogonal singular vectors which capture the most dominant directions in the speaker-space in the first few significant singular values.

7. Comparison of performance in cascade system framework

In this section, we compare the performance of the proposed method with baseline anchor systems in cascade mode. Fig. 4



(a) (b) (c) (d)

Figure 3: *Comparison of N-best speaker identification error rate between proposed method and baseline systems for varying anchor speaker models* **(a)** 46 . **(b)** 146. **(c)** 246. *and* **(d)** 346.

shows the block diagram of cascade anchor system in speaker identification task. The front-end systems selects the N-most

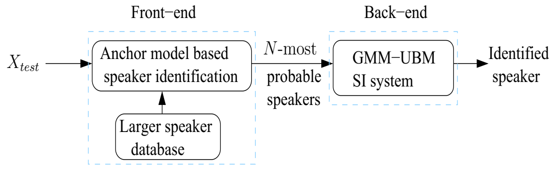


Figure 4: *Illustrates cascade anchor modeling system for speaker identification.*

probable speakers from the larger database and then the back-end GMM-UBM system finds the best speaker from the *reduced set*. For comparison of the different systems, we consider the 346 anchor models case. Fig.5 shows the speaker identification error rate of the cascade anchor system for different values of N. Table 1 shows the IER (%) of the cascade anchor systems.

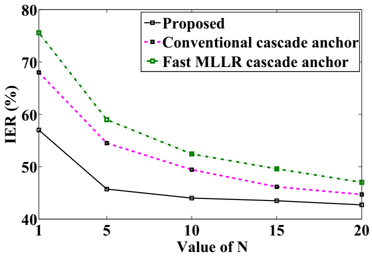


Figure 5: *Comparing speaker identification error rate of proposed method with baseline anchor systems in cascade mode.*

Table 1: *Comparing IER (%) of different systems in cascade mode for* 346 *anchor models.*

Cascade system	N=5		N=10		N=15	
	IER	Reduce by **(a)**	IER	Reduc. by **(a)**	IER	Reduc. by **(a)**
Proposed **(a)**	45.74	-	44.02	-	43.51	-
Conventional	54.51	8.77	49.44	5.42	46.17	2.66
Fast MLLR	58.99	13.25	52.45	8.43	49.61	6.10

From Fig.5 & Table, following observations can be drawn: Proposed anchor method shows significant reduction of IER compared to conventional and Fast MLLR anchor systems in cascade mode over a large range of N. The performance of the proposed method converges to conventional and Fast MLLR system as N increases, because performance of all systems approach that of the back-end GMM-UBM system. For case of N=5, proposed method gives *absolute* reduction of IER (%) by 8.77 and 13.25 over the conventional and Fast MLLR system respectively. Similarly, 5.42% and 8.43% in case of N=10, 2.66% and 6.10% in case of N=15 over the respective systems. The computational complexity of proposed method is similar to Fast-MLLR system, since both involve one alignment of data and a matrix multiplication.

8. Conclusion

In this paper, we propose a new anchor modeling system, where speakers are characterized by projecting their MLLR super-vector onto a eigen-voice space. The eigen voice space is built using MLLR super vectors of the non-target speaker. The proposed method shows significant improvement in speaker identification performance compared to anchor modeling techniques available in literature. Further, its computational complexity is similar to Fast-MLLR system. It gives *absolute* reduction of Identification Error Rate (IER) (%) by 7.01, 9.33, 10.33 and 11.2 against conventional anchor system for 46, 146, 246 and 346 anchor models respectively. Similarly, 11.45%, 15%, 17.51% and 18.6% compared to Fast MLLR anchor system. In cascade mode, proposed method also shows significant reduction of IER over the conventional and Fast MLLR anchor system with *absolute* reduction of IER (%) by 5.42 and 8.43 respectively in the case of N=10.

9. References

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

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

[3] A. K. Sarkar and S. Umesh, “Fast Computation of Speaker Characterization Vector using MLLR and Sufficient Statistics in Anchor Model Framework,” in *Interspeech*, 2010, pp. 2738–2741.

[4] E. Noor and H. Aronowitz, “Efficient Language Identification using Anchor Models and Support Vector Machines,” in *Proc. of Odyssey*, 2006.

[5] A. Stolcke *et al.*, “MLLR Transforms as Features in Speaker Recognition,” in *Proc. of Eurospeech*, 2005, pp. 2425–2428.

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

[7] R. Kuhn *et al.*, “Rapid Speaker Adaptation in Eigenvoice Space,” *IEEE Trans. Speech Audio Process*, vol. 8 (6), pp. 695–707, 2000.

[8] T. J. Hazen and J. R. Glass, “A Comparison of Novel Techniques for Instantaneous Speaker Adaptation,” in *Proc. of Eurospeech*, 1997, pp. 2047–2050.

[9] J. C. N. Wang, W. H. Tsai, and L. S. Lee, “Eigen-MLLR Coefficients as New Feature Parameters for Speaker Identification,” in *Proc. of Eurospeech*, 2001, pp. 1385–1388.

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

[11] 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.

[12] 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.