EFFICIENT SPEAKER VERIFICATION SYSTEM USING SPEAKER MODEL CLUSTERING FOR T AND Z NORMALIZATIONS

Kiran Ravulakollu
New Mexico State University
Klipsch School of Elect. Eng.
Las Cruces, NM 88003
USA
kiranrs@nmsu.edu

Vijendra Raj Apsingekar
New Mexico State University
Klipsch School of Elect. Eng.
Las Cruces, NM 88003
USA
vijendra@nmsu.edu

Phillip L. De Leon
New Mexico State University
Klipsch School of Elect. Eng.
Las Cruces, NM 88003
USA
pdeleon@nmsu.edu

***Abstract*—**In speaker verification (SV) systems based on Gaussian Mixture Model-Universal Background Model (GMM-UBM), normalization is an important component in the decision stage. Many normalization methods including the T- and Z-norms, have been proposed and investigated and these have contributed to state-of-the-art SV systems which have extremely low equal-error rates (EERs). In this paper, we consider application of both T- and Z-norms to a carefully selected subset of speakers using a data driven approach which can significantly reduce computation resulting in faster SV decisions and lower EER. Unfortunately, selection of the subset is critical and must be representative of the entire speaker model space otherwise error rates will increase. In order to properly select the subset of speakers for the normalizations, we propose a novel method which first clusters the speaker models using the K-means algorithm and the Kullback-Leibler (KL) divergence and then selects a set of speakers within the cluster. We evaluate the approach using both the TIMIT, NTIMIT and NIST-2002 corpora and compare against standard T- and Z-normalizations.

Index Terms – Speaker recognition, Clustering methods

I. INTRODUCTION

The objective of speaker verification (SV) is to verify an identity claim of a voice sample [1]. SV is a two-stage procedure consisting of training and testing. In the training stage, speaker-dependent feature vectors are extracted from the training speech signals and a speaker model λ_s is built by MAP-adapting the training feature vectors to a Gaussian mixture model-universal background model (GMM-UBM) [2]. Normally, SV systems use mel-frequency cepstral coefficients (MFCCs) as a $L \times 1$ feature vector and the speaker model λ_s is parameterized by the set $\{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$ where w_i are the weights, $\boldsymbol{\mu}_i$ are the mean vectors, and $\boldsymbol{\Sigma}_i$ are the covariance matrices. In the testing stage, feature vectors $\mathbf{X}_m^{\text{test}}$ are extracted from a test signal. A log-likelihood ratio $\Lambda(\mathbf{X}_m^{\text{test}})$ is computed by scoring the test feature vectors against the claimant model and the UBM.

$$\Lambda(\mathbf{X}_m^{\text{test}}) = \log p(\mathbf{X}_m^{\text{test}}|\lambda_s) - \log p(\mathbf{X}_m^{\text{test}}|\lambda_{\text{UBM}}). \quad (1)$$

The claimant speaker is accepted if

$$\Lambda(\mathbf{X}_m^{\text{test}}) \geq \theta \quad (2)$$

or else rejected [3].

The log-likelihood ratio given in (1) essentially measures how well the claimant's model scores compared to a background model for a given test utterance [4]. Prior to the use of GMM-UBM techniques for SV systems, the second term on the right hand side of (1) is replaced by a function, such as average or maximum, operating on a set of speaker models other than the claimant model [2]. The set of other speaker models is called as cohort set and should be selected in such a way that it covers the expected impostors encountered during testing.

The important problem in SV is to find a decision threshold θ for the decision making [5], [6]. The uncertainty in θ is mainly due to score variability between the trials. Score variability is due to the nature of enrollment material: phonetic content, signal duration, environmental noise, intraspeaker variability, transmission channel and quality of the speaker model training [3]. In the literature, the following three ways have been used to deal with the score variability: client-specific threshold, client-specific fusion and client-specific score normalization [7]. In client-specific threshold, each user has a different threshold [8], which can be a function of a global threshold [9]. However, the decision threshold has to be tuned for each user separately making it difficult for large population applications.

In the literature, several client-specific fusion classifiers have been proposed [7], [6], [10]. In [6], a support vector machine (SVM) classifier was used which shares the user-specific and user-independent data. The SVM was trained using user-specific data scores and user-independent scores. The relative influence of both the scores are weighted by controlling the contribution of each set of scores in the SVM. In [7], a two level client-specific fusion strategy was developed. In the first stage, N base systems have to be developed for J users and thus $J \times N$ scores have to be trained. In the second stage, N normalized outputs are formed by using a global fusion classifier, which is common to all the users.

In client-specific score normalization, normalization parameters are estimated for each speaker and applied after (1) is calculated, such that only a global threshold is needed. A general block diagram of the SV system using score normalization is shown in Fig. 1. Several score normalization techniques have been proposed in the literature, such as Z-norm, H-norm, T-norm, HT-norm, C-norm, F-norm, and D-norm. The need for score normalization was first studied in [11].

In [11], researchers observed large variance from both distributions of claimant scores and impostor scores during SV tests. To reduce the overall score distribution variance, the authors in [11] proposed impostor score distribution normalization. The basic idea is to center the impostor score distribution by applying on each score generated by SV system the following normalization

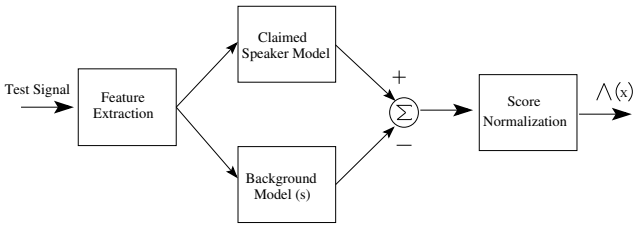

Fig. 1. Overview of speaker verification system using score normalization.

$$\tilde{\Lambda}(\mathbf{X}) = \frac{\Lambda(\mathbf{X}) - \mu_\lambda}{\sigma_\lambda} \tag{3}$$

where μ_λ and σ_λ are normalization parameters for speaker λ and $\tilde{\Lambda}(\mathbf{X})$ is the distribution of the normalized scores. The parameters μ_λ and σ_λ need to be estimated. Among the various normalization techniques Z-norm and T-norm are most widely used.

In Zeros normalization (Z-norm), a speaker model is tested against a set of example impostor utterances, resulting in an impostor similarity score distribution. Speaker dependent mean and variance are estimated from this distribution and are used in (3). The advantage of Z-norm is that the estimation of normalization parameters can be performed during the training stage of the system resulting in no additional testing stage computation [5].

In Test normalization (T-norm), during testing a set of impostor models are scored against the test utterance to yield an impostor score distribution. The advantage of T-norm over a cohort normalization is the use of the variance parameter which approximates the distribution of cohort population more accurately. The advantage of T-norm over Z-norm is that any acoustical mismatch between test utterance and impostor utterances are avoided. The disadvantage of T-norm is that it is performed during the testing stage resulting in an additional testing stage computation [5].

It is observed that, most of the client speaker models respond differently, for telephone speech, when the handset type used during training is different from that of testing. To deal with the handset mismatch between training and testing, H-norm, a variant of Z-norm, was proposed in [12]. Here, each speaker model is tested against different handset dependent speech signals produced by impostors to estimate the mean and variance of the handset dependent normalization parameters. The knowledge of the handset type used for the incoming test utterance determines the right set of the parameters to be used in score normalization.

As opposed to Z-norm, D-norm does not require any speech data to estimate the normalization parameters [13]. Here, pseudo-impostor data is generated using the UBM. A Monte-Carlo based symmetric Kullback-Leibler (KL) distance is used to obtain a set of client and impostor data using client model and UBM respectively. However, results presented in [13] show that Z-norm always outperformed D-norm, particularly at low miss-detection rates.

C-norm was introduced in [14] to deal with the cellular data when speech was recorded using several unidentified handsets. Here, the data required to estimate the normalization parameters is clustered followed by a H-norm like process, assuming each cluster consists of data generated from different handset types. However, authors claim that there is no performance gain when compared to T-norm and is more expensive to implement. HT-norm is based on same observation as H-norm but a variant of T-norm.

From the brief literature survey one can understand that Z-norm and T-norm are the most widely used normalization techniques and there exist many variants of these, specific to the database or application. Two important questions are 1) What is the proper amount of impostor utterances required to estimate Z-norm parameters and 2) How many cohort models are required for T-norm? Researchers in [15], [16] propose two different ways of selecting the cohort models for T-norm.

In [15], speaker adaptive cohort selection was proposed based on a city-block distance. Here, each speaker model is scored against N-impostor utterances to get a N-dimensional vector. Also, a pool of P T-norm models are scored against N-impostor utterances to get a $P \times N$ matrix. Using city-block vector distance measure, K-nearest T-norm models are chosen from the set of P models. This method was called AT-norm. The experiments were conducted on NIST-2004 corpus consisting of 435 male speakers and 550 female speakers, with $K = 55$. Results show AT-norm outperformed T-norm.

The procedure in [16] is much similar to AT-norm but instead of using city-block distance, here researchers used an approximation of KL divergence and called it KL-T-norm. The KL divergence between each speaker model and P T-norm models is computed and K nearest models are chosen. Experiments were performed on NIST-2005 corpus and KL-T-norm outperformed T-norm, with a cohort size of 75. However, no comparisons were made to AT-norm.

In this paper, the focus is on selecting the impostor utterances for estimating the Z-norm parameters and selecting the cohort models for T-norm We propose the use of training stage *speaker model clustering* (SMC) to guide us in selecting the impostor utterances and T-norm models. We are unaware of any paper dealing with selection of impostor utterances for Z-norm and papers dealing with T-norm selection are [15], [16]. Our work differs from [15], instead of finding the city-block distance between all the speaker model pairs and impostor utterances, we use KL distance between the speaker models in the selected clusters. Our work also differs from [16], in that we find a Monte-Carlo approximation of KL divergence and instead of finding the nearest speakers, we cluster the speaker models. The advantage of SMC is that we can select any number of cohorts without re-estimating the distances between all the speaker model pairs.

This paper is organized as follows. In Section II, we describe our method of speaker model clustering. In Section III, we describe the selection of cohort models for T-norm and impostor utterances for Z-norm using SMC. In Section IV, we describe the experimental evaluation and provide results using TIMIT, NTIMIT and NIST-2002 corpora; these corpora are among the most common, large population speech databases used in SV research. We conclude the article in Section V.

II. Speaker Model Clustering

The earlier work dealing with the cohort selection for T-norm used some broad client-specific information, such as matching the speaker's sex or enrollment handset type [5].

More client-specific, data driven approaches for selecting the T-norm models are required [15]. There has been little research in finding the number of impostor utterances and diversity in utterances required for estimating the Z-norm parameters.

In [17], we proposed speaker model clustering (SMC) for speeding-up the test stage computations in a speaker identification (SI) system. The objective of SI is to determine which voice sample from a set of known voice samples best matches the characteristics of an unknown input voice sample [18]. Similar to SV, in SI also speaker models are built during training. However, during testing, the log-likelihood score of unknown test utterance is computed against all the speaker models in the database and maximum scoring speaker is identified. In SI, likelihood computations between test feature vectors and all the speakers in the database can be time consuming and detrimental to applications where fast SI is required [17].

Thus to speed-up the test stage computations, we proposed to cluster the speaker models after the training models are built using simple k-means algorithm. During testing, we select only a subset of clusters containing speaker models which are likely to give large likelihood values for the given test utterance.

In order to develop clustering methods which are based on the k-means algorithm and can scale with population size, we begin by representing the speaker model simply as a point in L-dimensional space determined by the weighted mean vector (WMV) [17]

$$\bar{\boldsymbol{\mu}} \quad = \quad \sum_{i=1}^{W} w_i \boldsymbol{\mu}_i. \tag{4}$$

where W is the number of component densities in the GMM. From (4), one can conveniently define the centroid of a cluster of GMM speaker models as

$$\mathbf{r} \quad = \quad \frac{1}{K} \sum_{k=1}^{K} \bar{\boldsymbol{\mu}}_k \tag{5}$$

where $\bar{\boldsymbol{\mu}}_k$ is the WMV for λ_k and K is the number of speaker models in the cluster. Fig. 2 gives an illustration of the speaker model space. In order to select the cluster that will be searched in the test stage, the average of test feature vectors from the unknown speaker is computed. Next, Euclidean distance between this average and cluster centroids are computed and the nearest cluster is searched. Rather than selecting a single cluster to search, we selected a subset of clusters ranked according to the distance between the average of test feature vectors and cluster centroids. Using a subset of clusters allows a smooth trade-off between accuracy loss (due to searching too few clusters) and speed. Using this approach we could gain a speed-up of $3\times$ with little or no loss in accuracy on TIMIT and NTIMIT speech corpora [17].

Although our work in [17] shows that Euclidean distance based clustering works for SI systems, the SV decision in (1) is based on log-likelihood measure and not on an Euclidean distance. As the cluster centroid does not have the required GMM parameters $\{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$, many distance measures such as the Kullback-Leibler (KL) divergence cannot be directly used in conventional k-means clustering.

To avoid the centroid not having the sufficient GMM parameters and to facilitate the clustering based on KL divergence, we identify the speaker model, λ_n^{CR} which is nearest to each cluster centroid

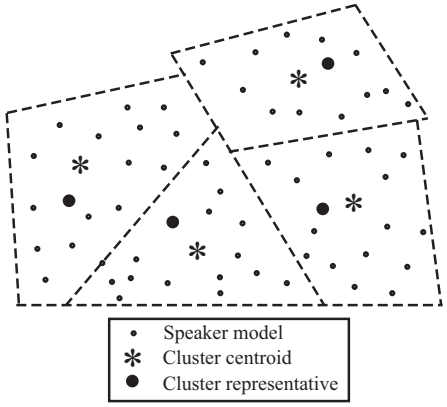

Fig. 2. Space of speaker models and clusters.

$$\lambda_n^{\mathrm{CR}} \quad = \quad \arg \min_{1 \le s \le S} d_1(\lambda_s, \mathbf{r}_n),\ 1 \le n \le N. \tag{6}$$

Where d_1 is defined as

$$d_1(\lambda_s, \mathbf{r}_n) \quad = \quad \left[(\bar{\boldsymbol{\mu}}_s - \mathbf{r}_n)^T (\bar{\boldsymbol{\mu}}_s - \mathbf{r}_n) \right]^{1/2}. \tag{7}$$

and N is the total number of clusters. This speaker model is called the "cluster representative" (CR) and is illustrated in Fig. 2. Below, we present a KL-divergence based distance measure which can be used in k-means for partitioning the speaker model space into N clusters.

A natural "distance" measure between two distributions f and g is KL divergence,

$$D(f||g) \quad = \quad \int f(x) \log \frac{f(x)}{g(x)} dx \tag{8}$$

however, there is currently no known closed-form expression for the KL divergence between GMMs which could be used for speaker model clustering [19]. Several distance measures between GMMs have alternatively been proposed and investigated for application to speaker recognition [19]. One of these methods uses actual acoustic data (feature vectors) from speakers to approximate the KL divergence between any two speaker models.

Following the approach in [19], we propose a distance measure which can be used in speaker model clustering by approximating the KL divergence from λ_s to λ_n^{CR} with

$$\begin{aligned} d_2(\lambda_s, \lambda_n^{\mathrm{CR}}) \quad \approx \quad & \frac{1}{M} \sum_{m=1}^{M} \log p(\mathbf{x}_{s,m}^{\mathrm{train}} | \lambda_s) - \\ & \frac{1}{M} \sum_{m=1}^{M} \log p(\mathbf{x}_{s,m}^{\mathrm{train}} | \lambda_n^{\mathrm{CR}}) \end{aligned} \tag{9}$$

where M is the number of test feature vectors. The algorithm for clustering using the above KL approximation is given in Algorithm 1.

III. Speaker Model Clustering for Normalization

A. Clusters for T-norm

Test Normalization, as the name suggests, is performed at the testing stage of a speaker verification system to reduce

Algorithm 1 Speaker model clustering using a log-likelihood distance

1: Initialize cluster representatives, λ_n^{CR}, $1 \leq n \leq N$ using randomly-chosen speaker models
2: Compute distance using (9) from λ_s to λ_n^{CR}, $1 \leq s \leq S$
3: Assign each λ_s to the cluster with the minimum distance
4: Compute new cluster centroids using (5) and determine λ_n^{CR} using (6)
5: Goto step 2 and terminate when cluster membership does not change.

inter-session variability. Fig. 3 shows the block diagram of clusters for selecting the cohort models for T-norm. Speaker models are built from the utterances and along with the claimant models are clustered to form N clusters according to the clustering technique mentioned in Section II. The main idea behind clustering being, we would rather consider a few cohort models to estimate the T-norm parameters instead of all the available models for our approach. P cohort models which are in the same cluster as the claimant are then selected and the claimant test utterance is scored against these models. The mean and variance of these scores are our normalization parameters. In the scenario where a cluster doesn't have P models, the clusters nearest to the present cluster according to (9), are merged until we get the required number of models for selection. In case the number of speakers in a cluster is more than P, we select P models out of the available models according to Algorithm 2 [4].

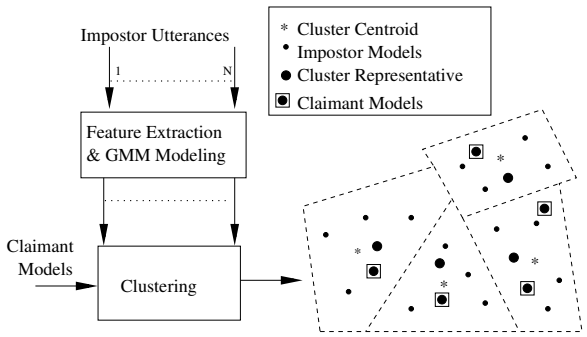

Fig. 3. T-norm selection procedure

Algorithm 2 Selecting Cohort models for T-norm

1: Let the available speakers be $Q(> P)$. Set of Q speakers be $\mathscr{Q}(i)$ and let required set of P speakers be $\mathscr{P}(i)$ for speaker i
2: Move the closest speaker according to (9) (between speaker i and all the speakers in $\mathscr{Q}(i)$) from $\mathscr{Q}(i)$ to $\mathscr{P}(i)$, $P' = 1$
3: Move speaker q from $\mathscr{Q}(i)$ to $\mathscr{P}(i)$, where q is found by

$$q = \arg \max_{q \in \mathscr{Q}(i)} \left\{ \frac{1}{P'} \sum_{p \in \mathscr{P}(i)} \frac{d_2(\lambda_p, \lambda_q)}{d_2(\lambda_i, \lambda_q)} \right\}, P' \leftarrow P' + 1$$

4: Repeat step (3) until $P' = P$

The cohorts selected according to the above algorithm are nearest to the claimant in that cluster and maximally spread from each other [4]. Researchers in [15], [20] suggested that cohort models closest to the claimant would yield lower EER than randomly selected ones. This process is repeated for all the speakers in the database.

B. Clusters for Z-norm

Similar to T-norm, speaker dependent mean and variance are estimated here again from impostor score distribution but this time instead of cohort models we have impostor utterances. Here again, we cluster the impostor models together with the speaker model into N clusters as we intend not to consider all the available impostor utterances to estimate the normalization parameters. Once the clustering is done, the selection of P near cohorts is same as that of T-norm. For the T-norm, the cohort models under consideration and the speaker should be from the same corpus unlike Z-norm where in there no such constraint on the selection of impostor utterances.

IV. Experiments and Results

Experiments have been performed on the TIMIT, NTIMIT and NIST 2002 corpora. To demonstrate the applicability of the methods proposed in Section II to a wide variety of GMM-UBM systems, we have added some additional elements such as delta MFCCs, cepstral mean subtraction (CMS) and RASTA processing depending on the corpus being used. Specifically, our baseline system uses an energy-based voice activity detector to remove silence; feature vectors composed of 29 MFCCs for TIMIT, 20 MFCCs for NTIMIT and 13 MFCCs + 13 delta MFCCs for NIST 2002 extracted every 10 ms using a 25 ms hamming window; CMS and RASTA processing are applied to NIST 2002. A 1024 component density UBM is built for each corpus by concatenating the training feature vectors of all the speakers within that corpus. Individual speaker models have then been built by MAP adaptation of parameters of the mean alone with a relevance factor of 16. For TIMIT/NTIMIT, we use approximately 24s training signals and 6s test signals and for NIST 2002 (one speaker detection cellular task) we use approximately 90s training signals and 30s test signals. Our system has baseline (no normalization) EERs of 0.11%, 3.64% for 630-speaker TIMIT,NTIMIT corpus respectively. For the 330-speaker NIST 2002 corpus our baseline EER of 12.25% agrees with the value published in [14]. In addition to the EER, another metric to measure the performance of SMC has been included called the minimum decision cost function (DCF), defined in [14] as

$$DCF = 0.1 \times Pr(\mathrm{miss}) + 0.99 \times Pr(\mathrm{falsealarm}) \qquad (10)$$

A. T-norm experiments

Our first set of experiments consists of observing the change in EER while varying the number of T-norm models used. A comparison between AT-norm and SMC for each of TIMIT, NTIMIT and NIST 2002 corpus is summarized in Tables I, II and III respectively. We observe that for TIMIT SMC performs better than AT-norm for cohort size of 20 and equals AT-norm's performance for 40 and 60 cohorts. We see that there is only a little improvement in the EER as we move from 20 cohorts to 40 cohorts in SMC. Thus we can achieve a similar EER using 20 cohorts with the SMC approach as opposed to 40 or more cohorts with the AT-norm. This is seen as an advantage

in computation as T-norm is performed online and we would be scoring against only half the number of models for each speaker. For NTIMIT, comparing the EER values we observe that SMC outperforms AT-norm for all the cohort sizes considered. Also for the NIST corpus, we see that SMC has done better than AT-norm for different cohort sizes in our experiments.

TABLE I
EER COMPARISON BETWEEN AT-NORM AND SMC T-NORM WITH VARYING COHORT SIZES FOR TIMIT DATABASE

Cohort Size	EER	
	AT-norm	SMC
20	0.31%	0.18%
40	0.16%	0.16%
60	0.16%	0.16%

TABLE II
EER COMPARISON BETWEEN AT-NORM AND SMC T-NORM WITH VARYING COHORT SIZES FOR NTIMIT DATABASE

Cohort Size	EER	
	AT-norm	SMC
20	3.33%	3.29%
40	3.35%	3.10%
60	3.17%	3.04%

TABLE III
EER COMPARISON BETWEEN AT-NORM AND SMC T-NORM WITH VARYING COHORT SIZES FOR NIST 2002 DATABASE

Cohort Size	EER	
	AT-norm	SMC
10	11.5%	8.5%
20	11.5%	8.0%
30	10.5%	7.5%

In Figs. 4, 5 and 6, we plot the detection error trade-off (DET) curves for TIMIT, NTIMIT and NIST 2002 corpora respectively. For TIMIT,the lowest EER values achieved by both SMC and AT-norm are identical. But for a relatively smaller cohort size of 20 itself SMC achieves an EER which is quite close to its lowest value and outperforms AT-norm at this level by around 0.12%. Also the minimum DCF values drops from 1.2×10^{-3} to 1.0×10^{-3} when SMC is used instead of AT-norm. In the case of NTIMIT, the lowest EER value achieved by SMC is better than that of AT-norm by 0.13% but the minimum DCF raises to 1.82×10^{-2} from 1.80×10^{-2}. Similarly for NIST, SMC surpasses the lowest EER value from AT-norm by a margin of 3% and the minimum DCF value drops from 9.40×10^{-2} in the case of AT-norm to 8.12×10^{-2} for SMC. We notice that for the TIMIT and NIST corpora, SMC has lower false alarm rate than AT-norm for the entire operating range . For NTIMIT, for operating conditions with a low miss detection rate SMC has a better false alarm rate compared to AT-norm.

B. Z-norm experiments

For Z-norm experiments we compared the performance of SMC Z-norm against the conventional Z-norm technique. By conventional Z-norm we mean that all available utterances are utilized for estimating the Z-norm parameters instead of only the P "closest" utterances based on SMCs. Our system has baseline (no normalization) EERs of 0.11%, 3.64% and 12.25%

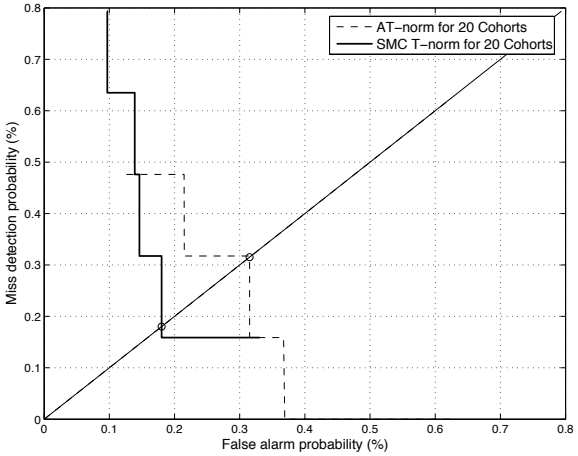

Fig. 4. T-norm DET curves for TIMIT

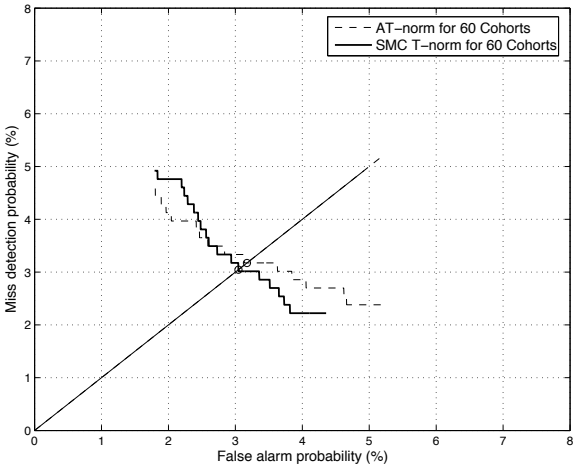

Fig. 5. T-norm DET curves for NTIMIT

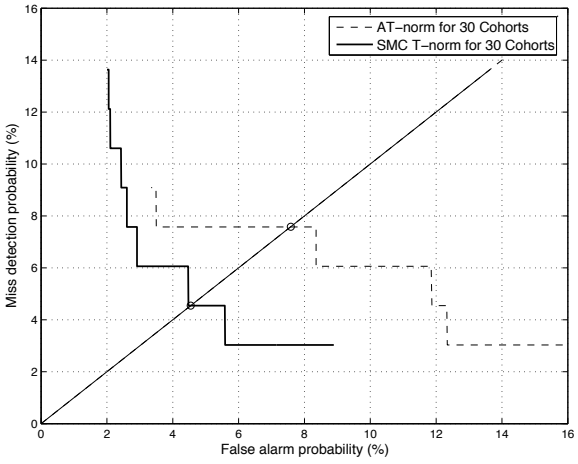

Fig. 6. T-norm DET curves for NIST

on TIMIT, NTIMIT and NIST-2002 corpora respectively. Using conventional Z-norm techniques our system has an EER of 0.12%, 3.64% and 12.12% on TIMIT, NITIMIT and NIST-2002 corpora respectively.

TABLE IV
EER VALUES FOR SMC Z-NORM WITH VARYING NUMBER OF IMPOSTOR UTTERANCES. EER FOR CONVENTIONAL Z-NORM IS SHOWN IN PARENTHESIS.

Number of Impostor Utterances	TIMIT (0.12%)	NTIMIT (3.64)
20	0.16%	3.51%
40	0.13%	3.55%
60	0.12%	3.65%

Similar to T-norm experiments, EER with varying number of impostor utterances is shown in Table IV for TIMIT and NTIMIT corpora and in Table V for NIST 2002 corpus. For TIMIT, SMC Z-norm equals the performance of the conventional Z-norm yielding a lowest EER of 0.12% which is higher than the baseline value of 0.11%. Our experiments reveal that because of the proximity of TIMIT speech to ideal conditions, score normalization did not improve the baseline EER value. For NTIMIT, SMC Z-norm achieved better EER values for different number of impostor utterances considered. The lowest being 3.51% using 20 impostor utterances as against an EER of 3.64% using conventional Z-norm. For the NIST corpus (Table V), using 10 impostor utterances, SMC Z-norm yields an EER which is little higher than the conventional Z-norm. It performs marginally better than the conventional technique for 20 impostor utterances and is as good as the conventional technique when 30 impostor utterances are considered for Z-norm parameter estimation.

TABLE V
EER VALUES FOR SMC Z-NORM WITH VARYING NUMBER OF IMPOSTOR UTTERANCES. EER FOR CONVENTIONAL Z-NORM IS SHOWN IN PARENTHESIS.

Number of Impostor Utterances	EER (12.12%)
10	13.60%
20	12.11%
30	12.12%

We observe that on TIMIT corpus, the performance of both SMC and conventional Z-norm is similar which is consistent with Fig. 7. The minimum DCF value drops off from 5.4405×10^{-4} to 5.0481×10^{-4} as we move from conventional Z-norm to SMC. For NTIMIT, from Fig. 8 we observe that the performance of conventional Z-norm technique is better than SMC. However, EER achieved by conventional technique is close to 3.64% and by applying SMC Z-norm we could reduce it to 3.51%. The minimum DCF value changed from 2.44×10^{-2} to 2.55×10^{-2} when SMC replaces conventional Z-norm. For the NIST corpus from Fig. 9, SMC and conventional Z-norm yield similar results with SMC proving to be 0.01% better in terms of EER. The minimum DCF value drops from 9.03×10^{-2} to 8.84×10^{-2} when SMC replaces the conventional technique. The False alarm rate for TIMIT using SMC is nearly equal to that of the conventional technique for operating conditions at low miss detection probabilities. On NIST-2002 corpus, only for operating conditions with a low miss detection rate does SMC have a better false alarm rate compared to conventional Z-norm.

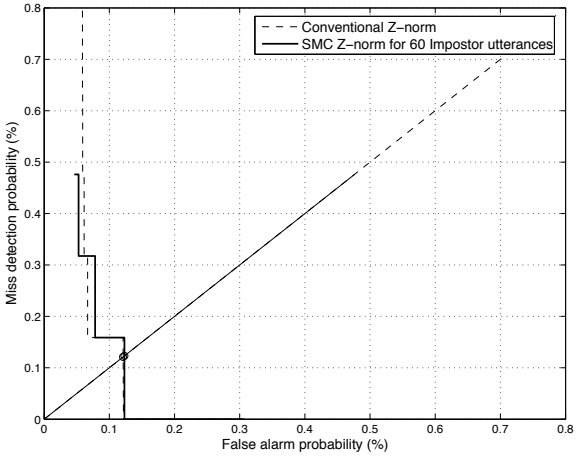

Fig. 7. Z-norm DET curves for TIMIT

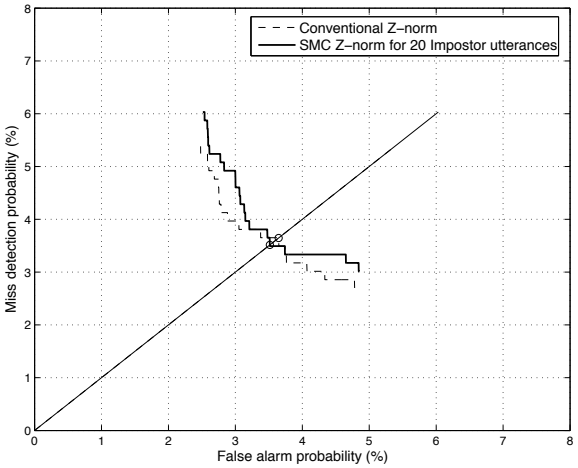

Fig. 8. Z-norm DET curves for NTIMIT

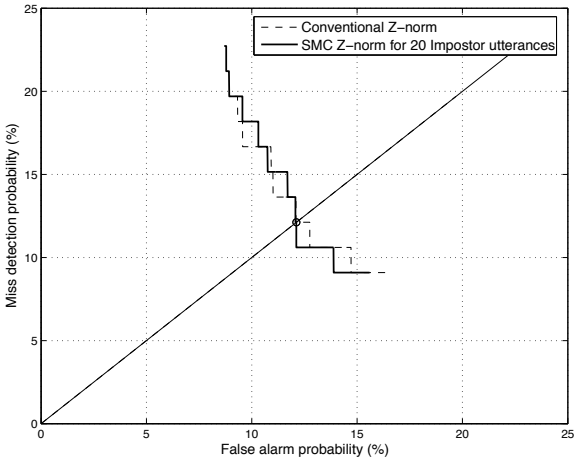

Fig. 9. Z-norm DET curves for NIST

V. CONCLUSIONS

Score normalization is an important phase in any speaker verification system which transforms the output scores to minimize score variability. It relies on having a global threshold for the entire database to accept or reject a claim. To estimate the normalization parameters, we need a set of impostor utterances or cohort models. In this paper, we proposed a new method called speaker model clustering for the clustering of speaker models before the selection of the cohort set. This method helps us select the utterances or models nearest to the speaker according to certain distance criteria. Clearly, we observe that the selection of impostor utterances for Z-norm on the basis of SMC either outperforms or is as good as the conventional Z-norm technique. Also the selection of cohort models for T-norm based on SMC has outperformed the AT-norm. In case of the T-norm, SMC did better than AT-norm for all the three corpora either in terms of computation or in achieving lower EER value. For Z-norm, SMC performed no better than the conventional method on all the three corpora. In general, the minimum DCF value was reduced using SMC as compared to AT-norm or conventional Z-norm. The SMC-based approach for cohort selection used in T- and Z- normalizations can be applied to other normalizations.

REFERENCES

[1] T. F. Quatieri, *Discrete-Time Speech Signal Processing Principles and Practice*. Prentice-Hall, Inc., 2002.

[2] 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.

[3] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleaua, S. Meignier, T. Merlin, J. Ortega-Garcia, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 430–451, 2004.

[4] D. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, 1995.

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

[6] J. Fierrez-Aguilar, D. Garcia-Romero, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Exploiting general knowledge in user-dependent fusion strategies for multimodal biometric verification," *Proc. Int. Conf. Acoust. Speech, Signal Process.*, vol. 5, pp. 617–620, 2004.

[7] N. Poh and J. Kittler, "Incorporating model-specific score distribution in speaker verification systems," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 16, no. 3, pp. 594–606, 2008.

[8] K. Chen, "Towards better making a decision in speaker verification," *Pattern Recognition*, vol. 36, no. 2, pp. 329–346, 2003.

[9] J. Lindberg, J. W. Koolwaaji, H. P. Hutter, D. Genoud, M. Blomberg, J. B. Pierrot, and F. Bimbot, "Techniques for a priori decision threshold estimation in speaker verification," *Proc. Workshop Reconnaissance du Locuteur et ses Applications Commerciales et Criminalistiques (RLA2C)*, pp. 89–92, 1998.

[10] J. Fierrez-Aguilar, D. Garcia-Romero, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Bayesian adaptation for user-dependent multimodal biometric authentication," *Pattern Recognition*, vol. 38, pp. 1317–1319, 2005.

[11] K. P. Li and J. E. Porter, "Normalizations and selection of speech segments for speaker recognition scoring," *Proc. IEEE. Int. Conf. Acoustics, Speech and Signal Processing*, vol. 1, pp. 595–598, April 1988.

[12] D. Reynolds, "The effect of handset variability on speaker recognition performance: experiments on the switchboard corpus," *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 113–116, May 1996.

[13] M. Ben, R. Blouet, and F. Bimbot, "A monte-carlo method for score normalization in automatic speaker verification using kullback-leibler distances," *Proc. Int. Conf. Acoust. Speech, Signal Process.*, vol. 1, pp. 689–692, 2002.

[14] D. Reynolds, "Channel robust speaker verification via feature mapping," *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 53–56, 2003.

[15] D. E. Sturim and D. A. Reynolds, "Speaker adaptive cohort selection for Tnorm in text-independent speaker verification," *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 741–744, 2005.

[16] D. Ramos-Castro, J. Fierrez-Aguilar, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "Speaker verification using speaker- and test-dependent fast score normalization," *Pattern Recognition Letters*, vol. 28, pp. 90–98, Jan. 2007.

[17] P. L. De Leon and V. R. Apsingekar, "Reducing speaker model search space in speaker identification," in *Proc. IEEE Biometrics Symposium*, 2007.

[18] D. Reynolds, "Large population speaker identification using clean and telephone speech," *IEEE Signal Process. Lett.*, vol. 2, no. 3, pp. 46–48, Mar. 1995.

[19] J. Goldberger and H. Aronowitz, "A distance measure between gmms based on the unscented transform and its application to speaker recognition," in *Proc. of Interspeech*, 2005, pp. 1985–1988.

[20] D. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," *Proc. Int. Conf. of Eurospeech*, pp. 963–966, 1997.

VI. VITA

K. Ravulakollu: Kiran Ravulakollu received his B.Tech. in electrical engineering from Jawaharlal Nehru Technological University, Hyderabad, India, in 2004. In 2004 he joined Satyam computers and was a embedded programmer. In Fall of 2006 he joined Klipsch School of Electrical and Computer Engineering, New Mexico State University, Las Cruces and presently pursuing his Masters degree.

V. R. Apsingekar: Vijendra Raj Apsingekar received his B.Tech. in electrical engineering from Jawaharlal Nehru Technological University, Hyderabad, India, in 2004. He received the M.S.E.E. degree in electrical engineering from Klipsch School of Electrical and Computer Engineering, New Mexico State University, Las Cruces in 2006. He is presently working on his Ph.D. degree in electrical engineering at New Mexico State University. His research interests include automatic speech recognition, speaker recognition and speech enhancement.

P. De Leon: Phillip De Leon received the B.S. Electrical Engineering and the B.A. in Mathematics from the University of Texas at Austin, in 1989 and 1990 respectively and the M.S. and Ph.D. degrees in Electrical Engineering from the University of Colorado at Boulder, in 1992 and 1995 respectively. Previously he worked at AT&T (and later Lucent Technologies) Bell Laboratories in Murray Hill, N.J. Currently, he serves as a Professor in the Klipsch School, Director of the Advanced Speech and Audio Processing Laboratory, and Associate Director of the Center for Space Telemetering and Telecommunications at NMSU. His research interests are in adaptive-, multirate-, real-time-, and speech-signal processing as well as wireless communications. Dr. De Leon is a senior member of IEEE.