Speaker Verification Score Normalization Using Speaker Model Clusters

Vijendra Raj Apsingekar and Phillip L. De Leon

Klipsch School of Electrical and Computer Engineering,
New Mexico State University,
Las Cruces, NM 88003 USA.
+1 (575) 646-3771 Tel, +1 (575) 646-1435 Fax,
{vijendra, pdeleon}@nmsu.edu

Abstract

Among the various proposed score normalizations, T- and Z-norm are most widely used in speaker verification systems. The main idea in these normalizations is to reduce the variations in impostor scores in order to improve accuracy. These normalizations require selection of a set of cohort models or utterances in order to estimate the impostor score distribution. In this paper we investigate basing this selection on recently-proposed speaker model clusters (SMCs). We evaluate this approach using the NTIMIT and NIST-2002 corpora and compare against T- and Z-norm which use other cohort selection methods. We also propose three new normalization techniques, Δ-, ΔT- and TC-norm, which also use SMCs to estimate the normalization parameters. Our results show that we can lower the equal error rate and minimum decision cost function with fewer cohort models using SMC-based score normalization approaches.

Key words: Speaker verification, score normalization

1. Introduction

The objective in speaker verification (SV) is to accept or reject a claim of identity based on a voice sample (Reynolds, 1995a). During the training stage, speaker-dependent feature vectors are extracted from the training speech signal and used to build a statistical model λ_s through MAP-adaptation of a Gaussian mixture model-universal background model (GMM-UBM) (Reynolds et al., 2000). Each feature vector consists of mel-frequency cepstral coefficients (MFCCs). 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 (diagonal) covariance matrices of the GMM. During the test stage, the sequence of feature vectors $\mathbf{X}$ is extracted from a test signal and a log-likelihood ratio $\Lambda(\mathbf{X})$ is computed by scoring the test feature vectors against the claimant model, λ_c and the UBM, λ_{UBM}

$$\Lambda(\mathbf{X}) = \log p(\mathbf{X}|\lambda_c) - \log p(\mathbf{X}|\lambda_{\text{UBM}}). \tag{1}$$

The claimant speaker is accepted if

$$\Lambda(\mathbf{X}) \geq \theta \tag{2}$$

or else rejected, where θ is decision threshold (Bimbot et al., 2004).

Large variance in the distributions of both claimant and impostor scores has been observed (Li and Porter, 1988). To reduce this variance, Li and Porter (1988) proposed impostor score distribution normalization. The basic idea is to use a normalized version of (2) where the normalization is

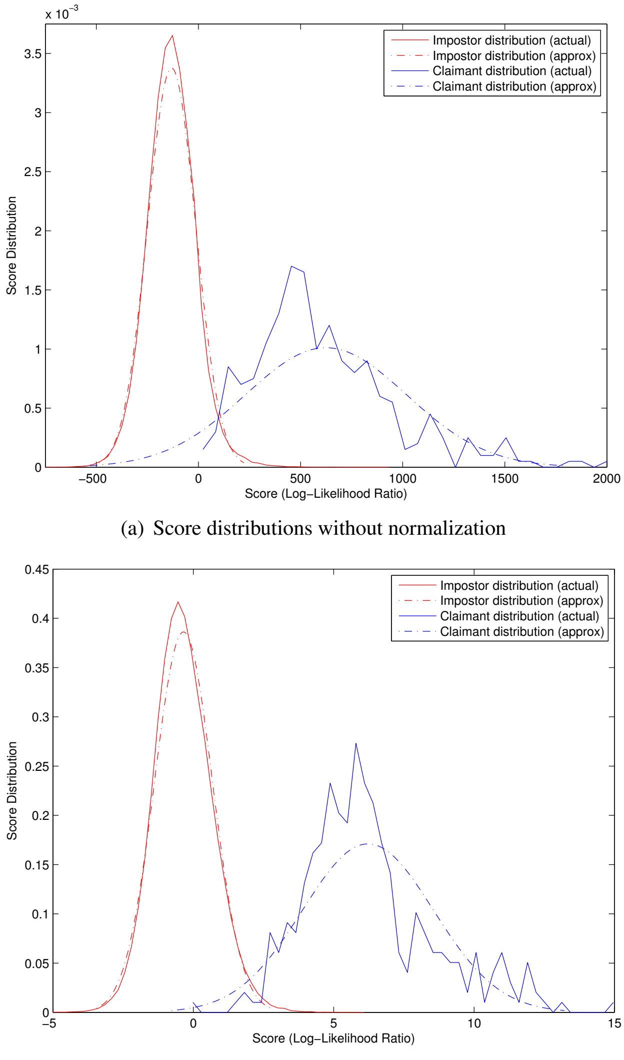
$$\tilde{\Lambda}(\mathbf{X}) = \frac{\Lambda(\mathbf{X}) - \alpha_c}{\beta_c} \tag{3}$$

and α_c, β_c are the estimated mean, standard deviation respectively, of the distribution of impostor log-likelihood scores for λ_c. Among the various normalization techniques, Zero-normalization (Z-norm) and Test-normalization (T-norm) are the most widely used methods to estimate the normalization parameters, α_c and β_c.

In Z-norm, during the training stage a set of impostor utterances is scored against each potential claimant model. The resulting impostor score distribution is used to estimate the normalization parameters in (3). Since the estimation is done at the training stage, there is no additional test-stage computation aside from (3), which is seen as an advantage for Z-norm (Auckenthaler et al., 2000). Generally all the available impostor utterances are used in estimating the Z-norm parameters (Auckenthaler et al., 2000).

In T-norm, during the test stage the test utterance is scored against a pre-selected set of cohort models (pre-selection is based on the claimant model). The resulting score distribution is then used to estimate the normalization parameters in (3). The advantage of T-norm over Z-norm is that any acoustic or session mismatch between test and impostor utterances is reduced. However, the disadvantage of T-norm is the additional test-stage computation in scoring the cohort models (Auckenthaler et al., 2000).

As shown in Fig. 1(a) for the NIST-2002 corpus, we observe considerable overlap among both the impostor and claimant score distributions thus resulting in verification errors and higher EER (Auckenthaler et al., 2000). Using score normalization methods, the impostor score distribution can normalized to zero mean and unit variance. As shown in Fig. 1(b), we observe that the T-norm reduces the overlap among the distributions resulting in fewer verification errors and lower EER.



(a) Score distributions without normalization

(b) T-norm

Figure 1: Approximate impostor and claimant score distributions from NIST-2002 speech corpus. With no score normalization (a), distributions overlap resulting in verification errors. With T-normalization (b), the overlap is reduced leading to fewer verification errors.

In (Sturim and Reynolds, 2005), speaker adaptive cohort selection for T-norm (known as AT-norm) was proposed based on a city-block distance. Given a claimant model λ_c, R impostor utterances, and Q T-norm models, impostor utterances are scored against each claimant model in the system and also against the Q T-norm models. Then city-block distance is used to select the L nearest T-norm models as the cohorts for the given claimant model. Prior to AT-norm (Sturim and Reynolds, 2005), cohorts for T-norm were selected based on some broad speaker-specific information, such as the speaker's gender or handset used (Auckenthaler et al., 2000).

Similar to AT-norm, Ramos-Castro et al. (2007) used an approximation of Kullback-Leibler (KL) divergence for the distance measure and called it KL-T-norm. The KL divergence between each speaker model and Q T-norm models is computed and the L nearest models are chosen. Experiments were performed on NIST-2005 corpus and showed that KL-T-norm outperformed T-norm, with a cohort size of 75. However, no comparisons were made to AT-norm.

From the literature review, there appears to be only limited research in selection of impostor utterances for Z-norm and cohort models for T-norm. In this paper, we propose the use of speaker model clusters (SMCs) as a general technique to assist with these selection problems in score normalization (Apsingekar and DeLeon, 2009), (Ravulakollu et al., 2008). The main contributions of this paper are two-fold. First, we present a unified approach which uses SMCs in selecting a subset of impostor utterances for Z-norm and selecting cohort models for T-norm. This approach can be extended to other similar normalization techniques such as H-norm, C-norm, and D-norm (Bimbot et al., 2004). Furthermore, with SMC-based T-norm, we can reduce computation resulting in a speed-up of the verification. Second, we propose three new normalization techniques, Δ-, ΔT- and test-cluster (TC) norm, all of which use SMCs for estimating the normalization parameters. When compared with existing methods, the proposed SMC-based normalizations can lower the equal error rate (EER) and minimum decision cost function (DCF).

This paper is organized as follows. In Section 2, we describe speaker model clustering and in Section 3, we describe the selection of cohort models for T-norm and impostor utterances for Z-norm using SMCs. In Sections 4 and 5, we describe the three new normalization techniques. In Section 6, we describe the experimental evaluation and provide results using both NTIMIT and NIST-2002 corpora. In Section 7, we provide analysis and discussion of the results and conclude the article in Section 8.

2. Speaker Model Clustering

In prior research, we proposed SMCs in order to speed-up the test stage in speaker identification (SI). In this section, we first summarize the work in (Apsingekar and DeLeon, 2009) before discussing application to SV score normalization.

We begin by representing the speaker model simply as a single point determined by the weighted mean vector (WMV)

$$\bar{\boldsymbol{\mu}} = \sum_{i=1}^{W} w_i \boldsymbol{\mu}_i. \qquad (4)$$

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

$$\mathbf{r} = \frac{1}{K} \sum_{k=1}^{K} \bar{\boldsymbol{\mu}}_k \qquad (5)$$

where $\bar{\boldsymbol{\mu}}_k$ is the WMV for λ_k and K is the number of speaker models in the cluster.

Next, we identify the speaker model λ_n^{CR}, which is nearest to each cluster centroid

$$\lambda_n^{\mathrm{CR}} = \arg \min_{1 \leq s \leq S} \left[\left(\bar{\boldsymbol{\mu}}_s - \mathbf{r}_n\right)^T \left(\bar{\boldsymbol{\mu}}_s - \mathbf{r}_n\right) \right]^{1/2}, \; 1 \leq n \leq N \quad (6)$$

where N is the total number of clusters and S is the number of speakers. This speaker model is called the "cluster representative" (CR) and the space of SMCs, speaker models, centroids, and CRs is illustrated in Fig. 2.

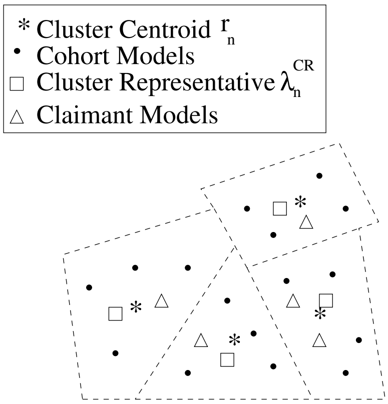


Figure 2: Space of speaker model clusters, cohort (speaker) models, cluster centroids and representatives. The SMCs facilitate a grouping of speaker models which can aid in efficient speaker identification and speaker verification score normalization.

Finally, we cluster speaker models using the k-means algorithm where the distance measure is based on an approximation to KL divergence

$$d(\lambda_s, \lambda_n^{\mathrm{CR}}) \approx \frac{1}{M}\sum_{m=1}^{M} \log p(\mathbf{x}_{s,m}^{\mathrm{train}}|\lambda_s) - \frac{1}{M}\sum_{m=1}^{M} \log p(\mathbf{x}_{s,m}^{\mathrm{train}}|\lambda_n^{\mathrm{CR}}) \tag{7}$$

where M is the number of training feature vectors and $\mathbf{x}_{s,m}^{\mathrm{train}}$ are the feature vectors from speaker s. The algorithm for speaker model clustering is given in Algorithm 1 (Apsingekar and DeLeon, 2009).

Algorithm 1 Speaker model clustering using KL divergence

```
1: Initialize cluster representatives, $\lambda_n^{\mathrm{CR}}$, $1 \leq n \leq N$ using randomly-chosen speaker models where $N$ is the desired number of clusters
2: Compute distance using (7) from $\lambda_s$ to $\lambda_n^{\mathrm{CR}}$, $1 \leq s \leq S$
3: Assign each $\lambda_s$ to the cluster with the minimum distance
4: Compute new cluster centroids using (5) and determine $\lambda_n^{\mathrm{CR}}$ using (6)
5: Goto step 2 and terminate when cluster membership does not change.
```

In the original application of SMCs for speeding-up the test stage in SI, we first select the cluster whose log-likelihood, measured against λ_n^{CR}, is highest

$$C_n = \arg\max_{1\leq n\leq N} \left[\sum_{m=1}^{M'} \log p(\mathbf{x}_{s,m}|\lambda_n^{\mathrm{CR}})\right], \tag{8}$$

where $\mathbf{x}_{s,m}$ is the test feature vector of speaker s and M' is the number of test feature vectors. Then the test utterance is scored against speaker models belonging to the selected cluster in order to make the identification. To increase SI accuracy, rather than selecting a single cluster, we use a subset of available clusters ranked according to (8). By using 20% of the SMCs, we were able to speed-up the SI test-stage using the TIMIT, NTIMIT, and NIST-2002 corpora by a factor of 5× with no loss in accuracy compared to a full search (Apsingekar and DeLeon, 2009).

3. Score Normalization using Speaker Model Clusters

We require impostor utterances to estimate the Z-norm parameters and cohort models to estimate the T-norm parameters. In both normalizations, cohort models are first built for each available impostor utterance and clustered along with the claimant models to form SMCs using Algorithm 1.

3.1. SMC-Based T-norm

In T-norm, the task is to select L cohort models which are nearest to the given claimant model yet diverse from each other (Reynolds, 1995b). By selecting cohort models from the same SMC as that of the claimant we are guaranteed to select models nearest to the claimant model [according to (7)]. There are three possibilities in the selection process: the number of speaker models in the claimant's cluster (excluding the claimant model) is equal to, less than, or greater than L. If the number of available models in the selected cluster is equal to L, then all intra-cluster models are used as the cohort models.

If the number of speaker models in the selected cluster is less than L, then the nearest clusters according to (7) (KL divergence between the CRs) are merged until the number of available models, Q is greater than or equal to L. If the number of speaker models, Q is greater than L, then the cohort models are selected according to Algorithm 2 which is similar to the algorithm used for selecting background cohort set in (Reynolds, 1995b). Cohort models selected according Algorithm 2 are nearest to the claimant in that cluster and maximally spread from each other (Reynolds, 1995b). Sturim and Reynolds (2005) suggested that cohort models closest to the claimant would yield lower EER than randomly selected ones. In our research, we used cohort models consisting of speaker models other than the claimant to estimate the T-norm parameters.

3.2. SMC-Based Z-norm

In Z-norm, we normally use all the available impostor utterances in estimating the normalization parameters (Auckenthaler et al., 2000). Using SMCs, however, we may estimate the parameters using a significantly smaller subset of impostor utterances. In our proposed SMC-based Z-norm, the impostor utterances associated with cohort models selected using SMCs (Algorithm 2) are used for parameter estimation. We note that the cohort models used in obtaining impostor utterances for Z-norm could be different than the cohort models for T-norm. In our research, we used the training utterances of speakers other than the claimant as impostor utterances to estimate the Z-norm parameters.

Algorithm 2 T-norm cohort model selection for claimant

```
1: Initialize the set of required cohort models $\mathscr{L}$ to null set
   and the set of speakers chosen from merged SMCs to be $\mathscr{Q}$
2: Move the closest speaker model according to (7) (between
   the claimant model and all the speakers in $\mathscr{Q}$) from $\mathscr{Q}$ to
   $\mathscr{L}$, $L' = 1$
3: Move speaker model $\lambda_q$ from $\mathscr{Q}$ to $\mathscr{L}$, where $\lambda_q$ is found
   by
   $$\lambda_q = \arg\max_{\lambda_q \in \mathscr{Q}} \left\{ \frac{1}{L'} \sum_{l \in \mathscr{L}} \frac{d(\lambda_l, \lambda_q)}{d(\lambda_c, \lambda_q)} \right\}, L' \leftarrow L' + 1$$
   (where $\lambda_c$ is the claimant model)
4: Repeat step (3) until $L' = L$
```

4. Normalization through SMC-Based Score Offset

In the previous section, we proposed how to use SMCs to select impostor utterances (Z-norm) and cohort models (T-norm) in order to determine the normalization parameters in (3). We can also use SMCs to offset or bias the log-likelihood score in order to further separate overlapping claimant and impostor score distributions.

4.1. Δ-Normalization

We begin by clustering all speaker models in the system as described in Section 2. Once the test utterance is acquired, clusters are scored and ranked according to (8). If the claimant speaker model is a member of the highest scoring cluster C [according to (8)], we add an offset $\Delta > 0$ to the log-likelihood score in (1) otherwise we subtract an offset

$$\tilde{\Lambda}_{\Delta}(\mathbf{X}) = \begin{cases} \Lambda(\mathbf{X}) + \Delta, & \lambda_c \in C \\ \Lambda(\mathbf{X}) - \Delta, & \lambda_c \notin C. \end{cases} \tag{9}$$

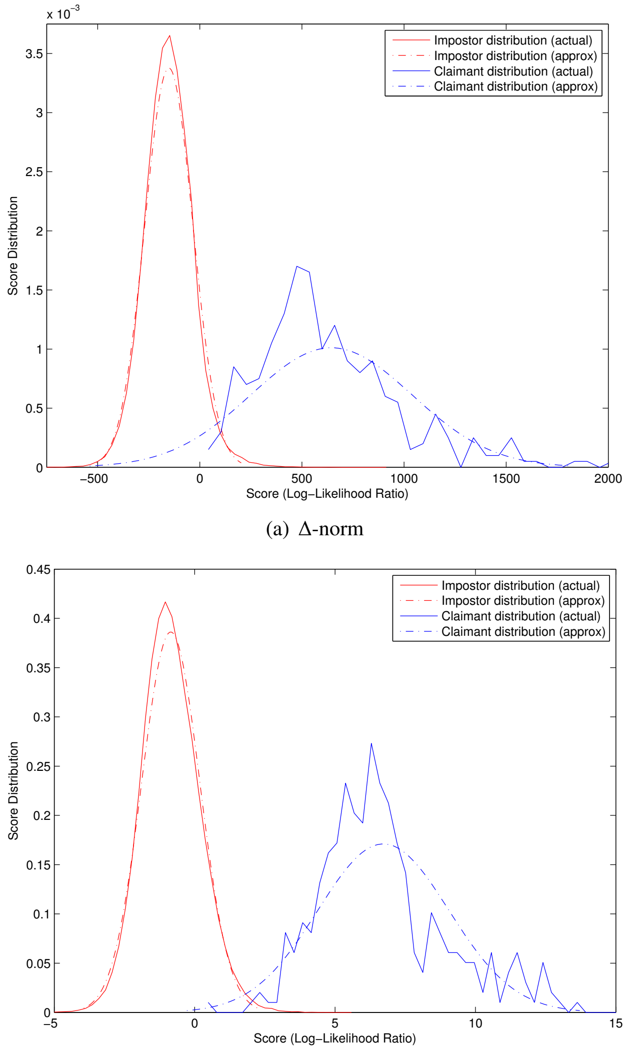
The addition or subtraction of the offset serves to further separate overlapping claimant and impostor score distributions as shown in Fig. 3(a). We can extend this idea to include not just the highest scoring cluster but rather a set of the highest-scoring clusters. Since there is the possibility that the set of highest-scoring clusters does not contain the true claimant or does contain the impostor model and $\mathbf{X}$ is from an impostor, the value of Δ is limited and found through experiment.

4.2. ΔT-Normalization

We can combine the proposed Δ-norm with SMC-based T-norm, resulting in ΔT-norm

$$\tilde{\Lambda}_{\Delta\mathrm{T}}(\mathbf{X}) = \begin{cases} \dfrac{\Lambda(\mathbf{X}) + \Delta - \alpha_c}{\beta_c}, & \lambda_c \in C \\ \dfrac{\Lambda(\mathbf{X}) - \Delta - \alpha_c}{\beta_c}, & \lambda_c \notin C. \end{cases} \tag{10}$$

where the definitions of α_c and β_c are same as that for conventional T-norm, i.e. using all available speakers as cohorts. Δ is not used in calculating the normalization parameters α_c and β_c. ΔT-norm can mitigate the effects of impostor and true claimant being in the same cluster since the SMC-based T-norm parameters are estimated from the cohort models chosen within the cluster. If the impostor and claimant are in the same cluster, then this impostor would also be a member of the cohort models for the claimant, thus reducing the final score in (10). Thus ΔT-norm can further reduce the overlap between the score distributions as shown in Fig. 3(b) and reduce the EER.



(a) Δ-norm

(b) ΔT-norm

Figure 3: Approximate impostor and claimant score distributions from NIST-2002 speech corpus.With (a) Δ- and (b) ΔT score normalizations, the overlap is further reduced over no score normalization and T-norm [see Fig. 1] leading to fewer verification errors.

5. Test-Cluster Normalization

The last proposed score normalization technique utilizing SMCs is called "Test Cluster (TC)" normalization. We begin by clustering all speaker models in the system as described in Section 2. Once the test utterance is acquired, clusters are scored and ranked according to (8). The speaker models within the highest scoring clusters serve as cohort models for estimating the TC-norm parameters α_c and β_c. Should the claimant speaker model may be a member of the highest scoring cluster, it is not used as a cohort model. In TC-norm, cohort models are chosen in the test-stage using the test utterance. This is different than the T-norms (AT-norm, KLT-norm, and SMC-based T-norm) where cohort models are chosen in the training stage.

As discussed earlier, the advantage of T-norm over Z-norm is that acoustical mismatch between the testing and training utterances can be eliminated (Auckenthaler et al., 2000). However, this acoustical mismatch is not completely eliminated in T-norm because the cohort models are chosen during training. However, with TC-norm, any acoustical mismatch between cohort models and test utterance can be further reduced as compared to T-norm.

6. Experiments and Results

Speaker verification experiments using the proposed SMC-based score normalization techniques have been performed using the NTIMIT and NIST-2002 corpora. Our baseline SV system uses an energy-based voice activity detector to first remove silence. Feature vectors composed of 20 MFCCs for NTIMIT and 15 MFCCs, 15 delta MFCCs, log-energy, and delta log-energy for NIST-2002 are extracted every 10 ms using a 25 ms hamming window. Cepstral mean subtraction (CMS) and relative spectra (RASTA) processing are applied. For experiments with the NIST-2002 corpus, we also apply feature warping.

For the NTIMIT corpus, a 1024 component UBM was built using all the available training feature vectors. The individual speaker models are then MAP-adapted (only the mean vectors) from the GMM-UBM. The NIST-2002 corpus consists of 330 speakers (139 male and 191 female) in the single speaker detection cellular task. We have separated the NIST-2002 corpus into two groups consisting of 2/3 and 1/3 of the speakers in the corpus [1]. The second group (last 1/3rd speakers of NIST-2002 corpus, consisting of 45 male and 65 female) are used to build a gender dependent, 1024 component GMM-UBM and these separated speakers are used as the cohort speakers in all the experiments. The speakers in the first group (first 2/3 speakers of NIST-2002 corpus, consisting of 94 male and 126 female speakers), individual speaker models are MAP-adapted (only the mean vectors) from the gender specific GMM-UBMs. The test utterances and claimant models from the second group, with which the UBM was built, are not used in calculating EER. Finally, we use 100 speaker model clusters with NTIMIT and 10 speaker model clusters with NIST-2002 corpus (Apsingekar and DeLeon, 2009). In experiments where we select a subset of clusters, we specify this as a percentage of the total number of clusters.

For the NTIMIT corpus, our baseline EER (no score normalization) is 3.64% and for the NIST-2002 corpus, our baseline EER (no score normalization) is 11.02%. Our NIST-2002 baseline EER is lower than the 12.10% published in (Longworth and Gales, 2009) due to fewer test speakers used in system evaluation, however, our system design closely matches that described in (Longworth and Gales, 2009). We note that other researchers (Ramaswamy et al., 2003), (Zhang and Mak, 2009) and (David and Leeuwen, 2005) have reported lower EERs with NIST-2002 with different variations on the GMM-UBM system in (Longworth and Gales, 2009). For this work, we use the system and baseline results as recently reported in (Longworth and Gales, 2009) as our benchmark.

In addition to the EER, the minimum DCF is also calculated to evaluate the performance of the proposed system. DCF is defined in (Reynolds, 2003) as

$$\text{DCF} = 0.1 \times \Pr(\text{Miss}) + 0.99 \times \Pr(\text{False alarm}). \tag{11}$$

For the NTIMIT corpus, our baseline minimum DCF (no score normalization) is 1.82×10^{-2} and for the NIST-2002 corpus, our baseline minimum DCF (no score normalization) is 9.82×10^{-2}.

6.1. Z-norm

For Z-norm experiments, we compared the performance of SMC-based Z-norm from Section 3.2 against the conventional Z-norm. By conventional Z-norm we mean that all available impostor utterances (629 on NTIMIT and 45 male/65 female for NIST-2002) are utilized for estimating the Z-norm parameters. Using the conventional Z-norm, for the NTIMIT, NIST-2002 corpus our system has an EER of 3.54%, 10.92% respectively. For the NTIMIT, NIST-2002 corpus, the minimum DCF is 2.55×10^{-2}, 9.03×10^{-2} respectively.

Our first set of experiments measure the EERs while varying the number of impostor utterances selected using SMCs. The results are shown in Tables 1 and 2 and we find that with as few as 20 impostor utterances, SMC-based Z-norm nearly matches the performance of conventional Z-norm which used **all** available impostor utterances. In addition, for the NTIMIT, NIST-2002 corpus the minimum DCF for SMC-based Z-norm (20, 20 impostor utterances) is 2.44×10^{-2}, 8.84×10^{-2} respectively.

Table 1: Speaker model cluster based Z-norm using the NTIMIT corpus. EER for conventional Z-norm (using all 629 impostor utterances) is 3.54% while SMC-based Z-norm can use as few as 20 impostor utterances with equivalent performance.

Number of Impostor Utterances	NTIMIT (3.54%)
20	3.51%
40	3.55%
60	3.55%

6.2. T-norm

For T-norm experiments, we compared the performance of SMC-based T-norm from Section 3.1 against AT-norm while varying the cohort size. First we note that when using conventional T-norm (all the available speakers as cohorts to estimate the T-normalization parameters), the EERs are 2.96%,

[1] By splitting NIST-2002 into two sets, we avoid using impostor data to train background models. The correct protocol is to use a different corpus to train the background models and T-norm models. Thus our results with NIST-2002 cannot be directly compared with other research which follows the protocol.

Table 2: Speaker model cluster based Z-norm using the NIST-2002 corpus. EER for conventional Z-norm (using all 45 male/65 female impostor utterances) is 10.92% while SMC-based Z-norm can use as few as 20 impostor utterances with equivalent performance.

Number Impostor	EER
Utterances	(10.92%)
10	11.02%
20	10.98%
30	11.02%

8.10% for NTIMIT, NIST-2002 respectively. This is not generally used due to the significant computation required at the test stage, hence the proposed variations on T-norm to select a subset for cohorts, e.g. AT-norm, KL-T-norm. In Tables 3 and 4, columns 2 and 3 provide the EER results for various cohort sizes using the NTIMIT and NIST-2002 corpora. We find that SMC-based T-norm outperforms AT-norm for all fixed cohort sizes.

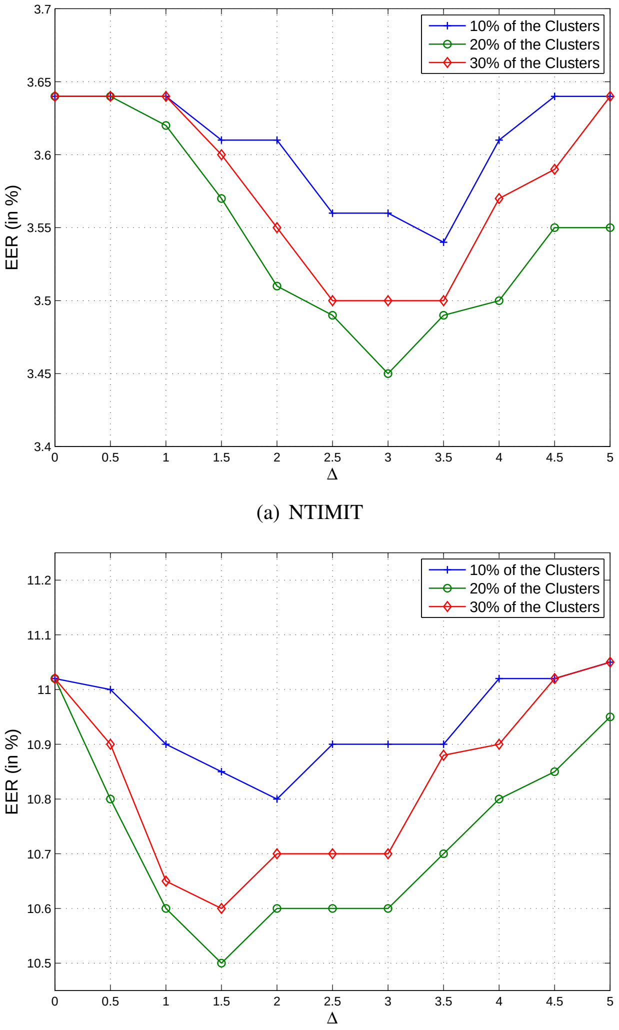
Table 3: Comparison of AT-norm, SMC-based T-norm, and ΔT-norm using the NTIMIT. For a fixed cohort size, SMC-based T-norm has lower EER than AT-norm which in turn has lower EER than the baseline system (no score normalization) of 3.64%. ΔT-norm improves upon SMC-based T-norm resulting in even lower EER for fixed cohort sizes.

Cohort Size	EER	EER	EER ΔT-Norm
Size	AT-norm	SMC T-norm	(Δ=3.0, 20% of Clusters)
20	3.33%	3.29%	3.28%
40	3.35%	3.10%	3.00%
60	3.17%	3.04%	2.97%

Table 4: Comparison of AT-norm, SMC-based T-norm, and ΔT-norm using the NIST-2002. For a fixed cohort size, SMC-based T-norm has lower EER than AT-norm which in turn has lower EER than the baseline system (no score normalization) of 11.02%. ΔT-norm improves upon SMC-based T-norm resulting in even lower EER for fixed cohort sizes.

Cohort Size	EER	EER	EER ΔT-norm
	AT-norm	SMC T-norm	(Δ=1.5, 20% of Clusters)
10	10.98%	9.81%	9.50%
20	10.05%	9.45%	9.13%
30	9.95%	8.99%	8.50%

For the NTIMIT corpus, we measure EER of 3.17% for AT-norm and 3.04% for SMC-based T-norm with 60 cohorts. This is an improvement over the baseline (no score normalization) result of 3.64%. For the NIST-2002 corpus, we measure EER of 9.95% for AT-norm and 8.99% for SMC-based T-norm with 30 cohorts. This is an improvement over the baseline (no score normalization) result of 11.02%. Using SMC-based cohort selection allows fewer cohorts for fixed EER than AT-norm. For fixed cohort size SMC-based score normalization produced lower EER than AT-norm. As T-norm is performed during test stage, scoring with fewer cohort models for each speaker translates into a computational advantage. Increasing the cohort size beyond 80 speakers on NTIMIT and 40 speakers on NIST-2002, SMC-based and AT-norm based score normalization technique produce similar limiting results. In addition, for the NTIMIT corpus the minimum DCF for AT-norm (60 cohorts) is 1.80×10^{-2} while SMC-based T-norm (60 cohorts) is 1.75×10^{-2}; for NIST-2002 the minimum DCF for AT-norm (30 cohorts) is 9.40×10^{-2} while SMC-based T-norm (30 cohorts) is 8.72×10^{-2}.



(a) NTIMIT

(b) NIST-2002

Figure 4: Δ-normalized EER with varying offset, Δ for different percentages of highest-scoring speaker model clusters used in the set C in (9). For the NTIMIT corpus (a), the baseline (no score normalization) EER is 3.64% and can be lowered with Δ-norm to 3.45% while for the NIST-2002 corpus (b), the baseline EER is 11.02% and can be lowered with Δ-norm to 10.50%. The baseline results are equivalent to Δ = 0.

6.3. Δ- and ΔT-norm

We evaluated the performance of the proposed Δ-norm from Section 4.1. The effect of Δ-norm on EER with varying Δ and percentage of highest-scoring clusters used in the selection process is shown in Fig. 4. We see that the choice of 20% of the SMCs and Δ = 3.0, 1.5 for NTIMIT, NIST-2002 respectively achieved the best results. With these choices we have an EER of 3.45%, 10.45% for NTIMIT, NIST-2002 respectively for Δ-norm. This is only slightly better than the baseline (no score normalization) EERs of 3.64% (NTIMIT) and 11.02% (NIST-2002) or equivalently, with $\Delta = 0$. In addition, the minimum DCF for Δ-norm is 1.82×10^{-2}, 9.82×10^{-2} for NTIMIT, NIST-2002 respectively which is the same as the baseline.

From SMC-based SI results on NTIMIT and NIST-2002 (Apsingekar and DeLeon, 2009), it was found that when searching 20% of clusters, there was no loss in accuracy compared to a full search. In other words, the true speaker of the test utterance is one among the speakers from highest scoring 20% of the clusters. In SV, the only change to that of SI is the knowledge of the claimant. If the claimed identity of the test utterance is a true claimant, then speakers within the subset of highest scoring clusters are likely to contain the claimant. This is utilized by adding Δ, assuming the test utterance is truly coming from the claimant. If the claimant is not contained in the selected SMCs, then the test utterance might have been produced by an impostor and so subtracting Δ would result in lower EER.

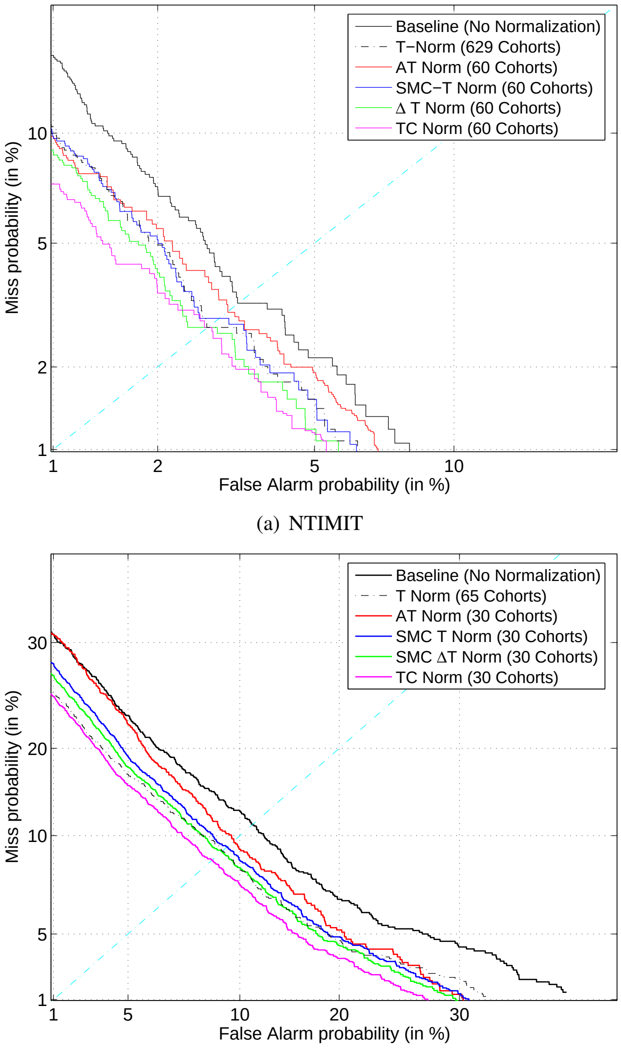
On the other hand, the test utterance might be from an impostor present in the selected SMCs or the true claimant speaker might not be contained in selected SMCs (equivalent to miss-ID in SI). Thus the percentage of clusters to search and the value of Δ are set experimentally. However, our research has found that selecting 20%–30% of the SMCs and $1 \leq \Delta \leq 4$ produce results better than baseline (see Fig. 4).

We also evaluated the performance of the proposed ΔT-norm from Section 4.2. Table 3, column 4 provides the NTIMIT results. For a cohort size of 60, ΔT-norm has an EER of 2.97% which is lower than the baseline (no score normalization) EER of 3.64% and AT-normalized EER of 3.17%. Table 4, column 4 provides the NIST-2002 results. For a cohort size of 30, ΔT-norm has an EER of 8.50% which is lower than the baseline (no score normalization) EER of 11.02% and AT-normalized EER of 9.95%. We find that of the family of T-norms, ΔT-norm has the lowest EER for fixed cohort size. In addition, the minimum DCF for ΔT-norm (60, 30 cohorts) is 1.75×10^{-2}, 8.59×10^{-2} for NTIMIT, NIST-2002 corpus respectively.

6.4. TC-norm

For the last proposed score normalization, we evaluated the performance of the TC-norm from Section 5 while varying the percentage of highest-scoring SMCs used in the selection process. Table 5, column 2 provides the results for the NTIMIT corpus where we see that using 30% of the clusters results in EER of 2.94% which is lower than the baseline EER of 3.64%. In Table 5, column 3 are the results for the NIST-2002 corpus where we also see that using 30% of the clusters results in an EER of 7.99% which is significantly lower than the baseline (no score normalization) EER of 11.02%. The TC-norm EER results are, in fact, better than all other score-normalized EERs.

We note that the way the cohort models are determined in T-norm and TC-norm are very different possibly explaining better performance for 30% of clusters than 20%. In T-norm cohort models are selected during the training stage, however the test utterance is scored against the cohort models during testing. In TC-norm, cohort models are selected during the test stage, thus T-norm is claimant model dependent and TC-norm is test utterance dependent. Thus the percentage of clusters to search may



(a) NTIMIT

(b) NIST-2002

Figure 5: DET plots for (a) NTIMIT and (b) NIST-2002 corpus using all the proposed normalization techniques. SMC-based T-norm outperforms AT-norm, while both techniques perform better than the baseline (no normalization). Among the family of T-norms proposed, SMC-based ΔT-norm has the lowest EER compared to SMC-based T-norm and AT-norm. TC-norm has the lowest EER among all the normalization techniques proposed.

Table 5: Performance of proposed TC-norm on NTIMIT and NIST-2002 corpus. Using 30% of the clusters on NTIMIT and NIST-2002 corpus, results in an EER of 2.94%, 7.99%, which is significantly lower than the baseline (no score normalization) EER of 3.64 and 11.02% respectively.

% of Selected Clusters	EER NTIMIT	EER NIST-2002
10	3.30%	9.13%
20	3.10%	8.40%
30	**2.94%**	**7.99%**
40	2.98%	8.25%
50	2.96%	8.25%

be different from ΔT-norm and TC-norm. Finally, the minimum DCF for TC-norm is 1.75×10^{-2}, 8.32×10^{-2} for NTIMIT, NIST-2002 corpus respectively.

The DET curves for all the proposed normalizations are shown in Fig. 5. On both the NTIMIT and NIST-2002 corpora, AT-norm performs better than the baseline (no nomalization). SMC-based T-norm generally performed better than the AT-norm, however at low false alarm probabilities on NTIMIT, AT-norm is slightly better than the SMC-based T-norm. Among the family of T-norms proposed, ΔT-norm has the better performance on the entire operating range. The performance of TC-norm is the best among all the proposed techniques with EERs as low as 2.92% and 7.99% on NTIMIT and NIST-2002 corpora respectively.

7. Analysis and Discussion

Score normalization techniques improve performance of SV systems by transforming claimant and imposter score distributions in order to better match the assumed normal score distributions. Distributions which are more normal lead to better estimates of the normalization parameters in (3). The Jarque-Bera (JB) goodness-of-fit test, measures departure from the normal distribution and is based on sample kurtosis and skewness (Judge et al., 1988). Table 6 gives the JB test statistics for null hypothesis rejection (scores are normally-distributed) at a 5% significance for the various score normalizations presented in this research. We see that all score normalization methods improve the goodness-of-fit to the normal distribution as compared to when no score normalization method is used. The proposed SMC-based normalizations result in score distributions with the higher JB test statistics indicating a better match the assumed normality of the distributions.

Table 6: Jarque-Bera test statistics for normality of score distributions at 5% significance. Results were generated using 629 NTIMIT, 45/65 Male/Female NIST-2002 cohorts for conventional T-norm and 60 NTIMIT, 30 NIST-2002 cohorts for AT-norm and SMC-based norms.

	Conventional T-Norm	AT-norm	SMC T-norm	SMC TC-norm
NTIMIT	92.89%	92.3%	94.81%	93.37%
NIST-2002	90.54%	95.55%	95.64%	96.91%

8. Conclusions

Score normalization transforms the log-likelihood ratio score in order to minimize score variability and is an important component in any SV system. In our previous work, speaker model clusters (SMCs) were used in order to speed-up the SI test stage. In this paper, we extend the use of SMCs for score normalization. SMCs allow us to select fewer impostor utterances for Z-norm and fewer cohort models for T-norm as compared to standard versions of these normalization techniques while simultaneously having lower EERs and minimum DCFs. In addition, we also introduced three new score normalizations—Δ, ΔT-norm and Test-Cluster (TC) normalization which also utilize SMCs. With the TC-norm, we were able to reduce the baseline (no score normalization) EER from 3.64% to 2.94% for the NTIMIT corpus and from 11.02% to 7.99% for the NIST-2002 corpus.

References

Apsingekar, V. R., DeLeon, P., May 2009. Speaker Model Clustering for Efficient Speaker Identification in Large Population Applications. IEEE Trans. Audio, Speech, and Language Process. 17 (4), 848–853.

Auckenthaler, R., Carey, M., Lloyd-Thomas, H., 2000. Score normalization for test-independent speaker verification system. Digital Signal Processing 10 (1), 42–54.

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

David, A., Leeuwen, V., 2005. Speaker adaptation in the nist speaker recognition evaluation 2004. Interspeech.

Judge, G. G., Hill, R. C., Griffiths, W. E., Lutkepohl, H., Lee, T. C., 1988. The theory and practice of Econometrics, 2nd Edition. John Wiley and Sons, Inc, Hoboken, NJ.

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

Longworth, C., Gales, M. J. F., May 2009. Combining derivative and parametric kernels for speaker verification. IEEE Trans. on Audio, Speech, and Language Process. 17 (4), 748–757.

Ramaswamy, G. N., Navratil, A., Chaudhari, U. V., Zilca, R. D., 2003. The IBM system for the NIST-2002 cellular speaker verification evaluation. Proc. Int. Conf. on Acoustics, Speech, and Signal Processing.

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

Ravulakollu, K., Apsingekar, V. R., P. L. De Leon, 2008. Efficient speaker verification system using speaker model clustering for T- and Z-normalizations. In: Proc. Int. Carnahan Conf. on Security Technology (ICCST).

Reynolds, D., 1995a. Automatic speaker recognition using gaussian mixture speaker models. The Linciln Laboratory Journal 8 (2), 173–191.

Reynolds, D., 1995b. Speaker identification and verification using gaussian mixture speaker models. Speech Communication 17, 91–108.

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

Reynolds, D. A., Quatieri, T. F., Dunn, R. B., 2000. Speaker verification using adapted gaussian mixture models. Digital Signal Processing 10, 19–41.

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

Zhang, S. X., Mak, M. W., 2009. Optimization of discriminative kernels in svm speaker verification. Interspeech.