# SPEAKER ADAPTIVE COHORT SELECTION FOR TNORM IN TEXT-INDEPENDENT SPEAKER VERIFICATION •

*D. E. Sturim and D. A. Reynolds*
{sturim,dar}@ll.mit.edu

MIT Lincoln Laboratory, Lexington, MA USA

## ABSTRACT

In this paper we discuss an extension to the widely used score normalization technique of test normalization (Tnorm) for text-independent speaker verification. A new method of speaker Adaptive-Tnorm that offers advantages over the standard Tnorm by adjusting the speaker set to the target model is presented. Examples of this improvement using the 2004 NIST SRE data are also presented.

## 1. INTRODUCTION

Score normalization is the transformation of speaker verification output scores to enhance the effectiveness of the detection threshold by aligning the score distributions of individual speaker models, Score normalization can be used to reduce the effects of both speaker-dependent and speaker-independent modifications on the signal. For example, Znorm attempts to align between-speaker differences of imposter scores distributions, while Hnorm attempts to remove speaker-dependent scale and bias effects from different channels and microphones. In both Znorm and Hnorm, score normalization parameters are estimated from scores derived by scoring a set of imposter utterances through each speaker model.

In the popular score normalization method of Tnorm [1], the normalization parameters are estimated using scores derived at test time from a set of imposter speaker models. As shown in **Error! Reference source not found.**, a fixed set of imposter or Tnorm speaker models are scored in parallel with the target speaker model. The mean and standard deviation of the imposter scores are then used to adjust the target speaker score as

$$S_{tgt|tnorm}(O) = \frac{S_{tgt}(O) - \mu_{tnorm}}{\sigma_{tnorm}} \qquad (1)$$

where $S_{tgt}(O)$ is the target speaker score for observations $O$. Tnorm is particularly efficient in an adapted UBM system since the adapted universal background model (UBM) provides fast scoring [2] allowing for the scoring of a large set of Tnorm speaker models with little additional computation.

The Tnorm method is very similar to earlier methods employing cohort, likelihood ratio or background model sets (e.g., [3] and [4]) used before UBM approaches were widely adopted. When using cohort sets for likelihood ratio computations, it was observed that better performance could be obtained by using speaker-specific sets of cohorts, selected using speaker characteristics (e.g., sex) or via data driven approaches (e.g., based on distance measures [4]). While Tnorm sets use some broad speaker-specific information, such as matching the speaker's sex or enrollment handset type [1], there has been little research in more speaker-specific, data driven Tnorm selection approaches. In this paper we present an approach for speaker dependent Tnorm selection to help improve verification performance.

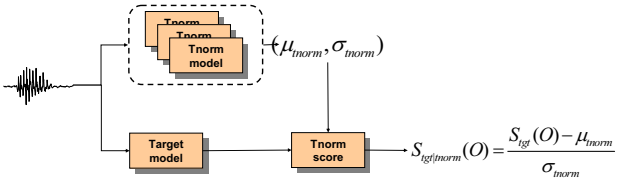


Figure 1 System for normalizing score distributions. The scaling parameters are derived from scores of the test message as scored against the set of Tnorm models.

The remainder of this paper is organized as follows. Section 2 describes the adaptive cohort model selection Tnorm system. Section 3 describes the NIST-04 extended data corpus and the experiment paradigm. Section 4 presents experiment results and analysis using the adaptive cohort model selection Tnorm system trained with various numbers of training utterances. In Section 5, we discuss the results and suggest possible future directions.

## 2. ADAPTIVE COHORT MODEL SELECTION FOR TEST NORMALIZATION

In this paper, we use a Gaussian Mixture Model-Universal Background Model (GMM-UBM) speaker-verification system [5]. The GMM-UBM system is a likelihood-ratio detector in which we compute the likelihood ratio for an unknown test utterance between a speaker-independent acoustic distribution (UBM) and a speaker-dependent acoustic distribution (see [5] for details). Typically 2048 mixtures are used for the UBM.

A data-driven approach is used for Tnorm selection where we attempt to find a set of Tnorm models that produce scores to imposter utterances similar to those from the target model.

*Figure* 2 System for selection of the K-nearest Atnorm cohort models. N-impostor test messages are scored against all cohort Atnorm models and the target model. Through a vector distance comparison (city-block distance) the K-nearest models are chosen.

• This work is sponsored by the Department of Defense under Air Force Contract F19628-00-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government

presents the system used to select the set of Tnorm models for a target speaker. A set of *N*-impostor test messages is scored against a large pool *P* of potential Tnorm models. The pool contains models whose training utterances comprise varying handset types, durations and number of sessions. The *N*-impostor messages are scored against all Tnorm models, forming *P* score vectors for each of *N*-dimensions. The *N*-impostor test messages are also scored against the target model forming a single N-dimensional vector of the raw output likelihood scores. Using city-block vector distance comparison, the K-nearest Tnorm models are chosen for the set of Tnorm models for the target model. K is set experimentally using a development corpus.

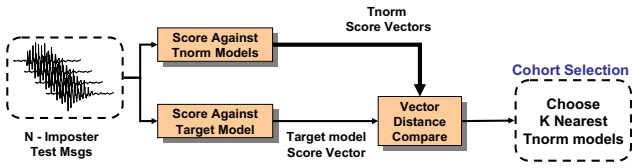


Figure 2 System for selection of the K-nearest Atnorm cohort models. N-impostor test messages are scored against all cohort Atnorm models and the target model. Through a vector distance comparison (city-block distance) the K-nearest models are chosen.

After the selection, each target model will have its own set of Tnorm models leading to a unique set of scaling parameters per target model. Ideally, the pool of cohort models *P* should be large enough to provide a representative coverage of Tnorm models from which to draw.

## 3. THE EXTENDED DATA TASK

For the 2004 NIST speaker recognition evaluation (NIST SRE), the extended data task was a continuation of a task that first debuted in 2002. The aim is to encourage the use of techniques and approaches precluded by limited training data. The evaluation data was the newly recorded Mixer corpora [6,7], which was exposed to the community for the first time this year. New this year, multi-lingual speakers were included (Arabic, Mandarin, Russian, and Spanish along with English). Experiments in this paper are for pooling results over all conditions (which included cross-language trials). Speaker models could be trained with up to 16 conversations of training speech (~40 minutes). Training conditions for model training and could use 1, 3, 8, and 16 conversation sides. (A conversation side consists of nominally 1–2.5 minutes of speech.) In this paper we will use the notation 1c, 3c and 8c to denote the conditions of 1, 3 and 8 conversation sides. The trials consisted of mismatched handset trials. For the evaluation, NIST supplied speaker-model training lists and index files indicating which models were to be scored against which conversation sides. A full description of the 2004 SRE can be found at http://www.nist.gov/speech/tests/spk/2004/.

The development corpus used for Tnorm model training was Switchboard II Phases 1, 2, 3, 4, and 5[i]. The Tnorm system used models consisting of a pool of $P_{male}$ = 435 male models if the target model is male or $P_{female}$ = 550 female models if the target model is female (the maximum that could be selected). The Atnorm system used the same gender model pools (male = 435; female = 550) for model selection. The number of impostor test messages was $N$ = 800, and they were also drawn from speech in Switchboard II Phases 1, 2, 4, and 5 that was not used in background model training or Tnorm model training. Cohort model selection used $K$ = 55 of the closest models determined from the procedure described in section 2. The parameter K was chosen from experiments run on the development corpus over different training/testing conditions. K proved to be a relatively stable parameter over the range [50-70].

[i] The Tnorm speakers all spoke English.

## 4. EXPERIMENTS

Experiments were conducted using the extended data setup. Results reported below are for the 2004 NIST speaker-recognition evaluation with UBMs trained using data from Switchboard II Phases 1–5. The experiments examined the performance of the baseline GMM-UBM compared against Tnorm and Atnorm. The training conditions varied from a single conversation of 10 seconds to 16 sides of a conversation. The test speech messages ranged from single conversation sides of 10 and 30 seconds to a whole conversation side averaging 2.5 minutes.

### 4.1 Detection Error Trade-Off Performance

Figure 3 is a detection error trade-off (DET) plot for the NIST speaker recognition evaluations for the 8-conversation-side training condition with 1 conversation side test. The ATnorm system used the same sex model pools (male = 435; female = 550) for model selection. The parameter K was chosen to be K=55. The Tnorm system displays the characteristic improvement in the low false-alarm region and nominal gain in the EER region. In contrast, the Atnorm system shows improvement across the entire DET curve.

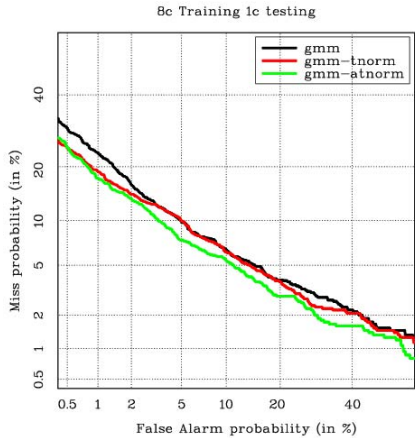


Figure 3 DET plot for the NIST speaker-recognition evaluations for the 8-conversation-side training condition with 1 conversation side test.

Figure 4 displays a bar chart of the EER versus the number of training/test conversation sides. The general trend is that the Atnorm system offers similar or better performance over the baseline GMM and the Tnorm systems.


Figure 4 Bar chart of the EER versus number of target model training conversation sides for matched and mismatched cohort model training conditions.

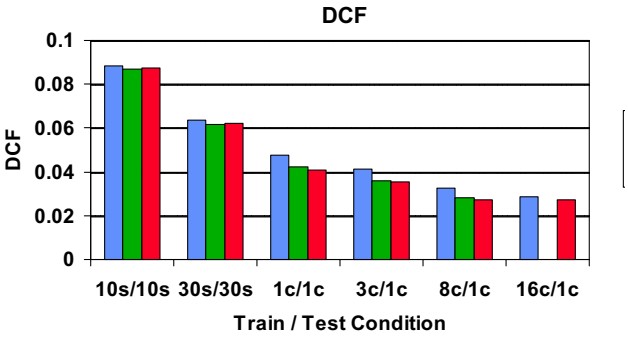


Figure 5 Bar chart of the DCF versus number of target model training conversation sides for matched and mismatched cohort model training conditions.

Figure 5 shows a similar bar chart plotting minimum decision cost function (DCF) versus the number of training/test conversation sides. The decision cost function is the relative cost of detection errors; the ratio of the cost of miss to cost of false alarm is 10 to 1 [8]. In this figure the Atnorm system offers a real improvement over the baseline system and a moderate improvement over the Tnorm system.

## 4.2 Atnorm Target Model Training Mismatch

We can pose the question of how critical it is to match the Atnorm training condition with the training condition of the target model. Training conditions are the amount of training data (in number of conversation sides) used to generate the particular target model. The following experiments compare the performance of matched and mismatched Atnorm training/target model training conditions.

Figure 6 and Figure 7 show bar charts for EER/DCF versus a varied number of target model training conditions. The black arrows point to the cases in which the cohort model training condition matches the target training. In both figures it can be seen that performance drops off when the target model training condition does not match that of the Atnorm cohort model. The disparity in performance becomes most severe when the target training has a single-conversation training side (1c).

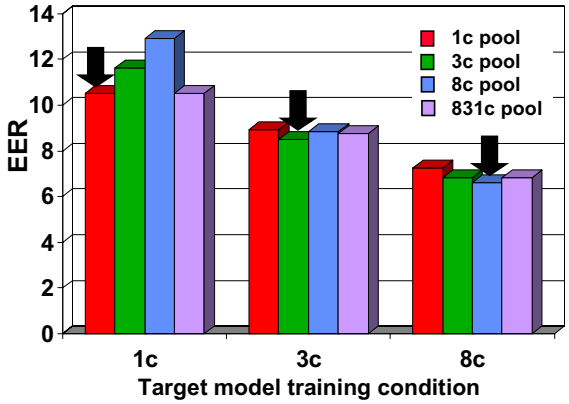


Figure 6 Bar chart of the EER versus number of training conversation sides over various Tnorm model training conditions.

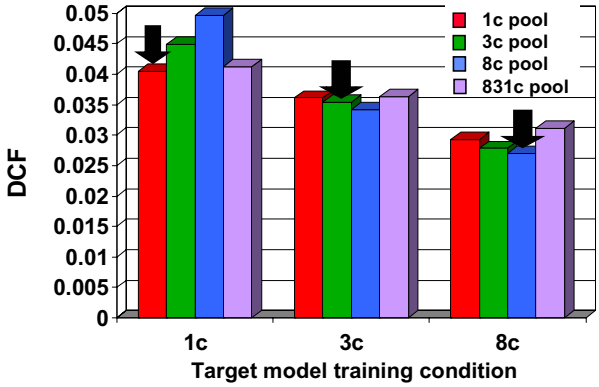


Figure 7 Bar chart of the DCF versus number of training conversation sides over various Tnorm model training conditions.

We notice two points from Figures 6 and 7; (1) Atnorm functions best when the target training conditions match the Atnorm cohort model training, and (2) The all-pooling training condition (831c) displays reasonable performance.

We can further analyze which cohorts are being selected in the Atnorm system when cohorts are selected from a combined pool of 1, 3 and 8 training conversations (831c). Table 1 presents a percentage break down of which cohorts are selected under different target model training conditions (1c or 8c).

Table 1 shows that when the target training is a single side the system is selects virtually all of its Atnorm cohorts from the 1c models. When the target model is trained from 8 conversation sides the Atnorm system selects cohorts from all conditions, but surprisingly most from the 3c set. This results in a performance hit at DCF when compare to the cohort matched system (Table 2 and Figure 7). However the EER performance is better than the performance in the mismatched Atnorm training/target model training.

Table 2 displays the EER and DCF performance matrix for the various Atnorm systems presented in this section versus the 1c

and 8c target training conditions. The bolded entries are the matched Atnorm training/target model training conditions. For both targets training conditions, the matched Atnorm systems perform best. However the Atnorm system with the 831c cohort pooling offers stable performance.

Table 1 *P*ercentage break down of which Atnorm cohorts are selected under different target model training conditions (1c or 8c).

| | Atnorm Selection | | |
|---|---|---|---|
| **Tgt Train** | **1c** | **3c** | **8c** |
| 1c | 93.9% | 6.0% | 0.2% |
| 8c | 19.2% | 47.8% | 33.0% |

Table 2 Equal error rate (EER in percent) and decision cost function (DCF x 10^-3) performance for pooling combinations versus target training conditions.

| **Tgt Train** | Atnorm pool 1c 3c 8c | Atnorm 1c | Atnorm 8c |
|---|---|---|---|
| 1c | 10.51 / 41.2 | **10.52 / 40.5** | 12.91 / 49.7 |
| 8c | 6.62 / 31.2 | 7.25 / 29.3 | **6.84 / 27.0** |

## 5. CONTRASTING SYSTEMS

It should be noted that there were two other systems tried in the formulation of the Atnorm technique presented in Section 2. The first contrastive system tried to utilize the target training utterances to generate speaker-centric normalization parameters directly. This system could use up to the 16 training conversation sides in the extended data task of the NIST evaluation. It yielded no performance gain and was not pursued. We conclude that, even given 16 training sides, there still were not enough utterances to provide meaningful normalization statistics.

The second contrastive system is structured in the same manner as the Atnorm presented in Section 2 with the exception that the cohort models are selected with direct model comparison using a Bhattacharya comparison technique. This second system offered similar performance when compared to the technique of Section 2; however, it proved to be very compute-intensive and was abandoned because of efficiency considerations.

## 6. CONCLUSIONS AND FUTURE WORK

In this paper we have presented the results of using a speaker adaptive cohort selection for Tnorm to improve speaker-verification performance for a text-independent task. On the NIST 2004 extended-data task, the work demonstrated that choosing Atnorm cohort models based on the target speaker could produce low error rates when compared to traditional Tnorm.

We also investigated how best performance is achieved when the Atnorm cohort model training duration matches that of the target model. A back-off system was shown to be the all-pooling Atnorm system. This system combines all available cohort model types into a large population. The all-pooling system provides slightly worse performance compared to the cohort-matched Atnorm system, but outperforms the cohort-mismatched Atnorm system.

## REFERENCES


[1] C. Auckenthaler, Lloyd-Thomas, "Score Normalization for Text-independent Speaker Verification Systems," *Digital Signal Processing*, vol. 10 No 1-3, 2000.

[2] D. A. Reynolds, "Comparison of Background Normalization Methods for Text-Independent Speaker Verification," in Proceedings of EUROSPEECH '97, Rhodes, Greece, 1997, pp. 963-966.

[3] A. Rosenberg, J. Delong, C. Lee, B. Juang, and F. Soong, "The use of cohort normalized scores for speaker recognition," *In Proc. ICSLP*, pp. 599–602, 1992.

[4] D. A. Reynolds, "Robust text-independent speaker identification using Gaussian mixture speaker models," *Speech Communications*, vol. 17, pp. 91-108, 1995.

[5] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10(1–3), pp. 19-41, 2000.

[6] A. Martin, et al., "Conversational Telephone Speech Corpus Collection for the NIST Speaker Recognition Evaluation 2004," *LREC 2004*, 2004.

[7] J. P. Campbell, et al., "The MMSR Bilingual and Cross-Channel Corpora for Speaker Recognition Research and Evaluation," *Odyssey 2004*, 2004.

[8] G. R. Doddington, et al., "The NIST Speaker Recognition Evaluation-Overview, Methodology, Systems, Results, Perspective," *Speech Communication*, vol. 31 (2-3), pp. pp. 225-254, 2000.