TOWARDS REDUCED FALSE-ALARMS USING COHORTS

Zahi N. Karam[1,2], William M. Campbell[2], Najim Dehak[3]

[1]DSPG, RLE at MIT, Cambridge, MA , [2]MIT Lincoln Laboratory, Lexington, MA , [3]CSAIL at MIT, Cambridge, MA

zahi@mit.edu, wcampbell@ll.mit.edu, najim@csail.mit.edu

ABSTRACT

The focus of the 2010 NIST Speaker Recognition Evaluation (SRE) [1] was the low false alarm regime of the detection error trade-off (DET) curve. This paper presents several approaches that specifically target this issue. It begins by highlighting the main problem with operating in the low-false alarm regime. Two sets of methods to tackle this issue are presented that require a large and diverse impostor set: the first set penalizes trials whose enrollment and test utterances are not nearest neighbors of each other while the second takes an adaptive score normalization approach similar to TopNorm [2] and ATNorm [3].

Index Terms— Speaker Recognition, False Alarms, Score Normalization, Adaptive Normalization

1. INTRODUCTION

The 2010 NIST Speaker Recognition Evaluation (SRE) [1] introduced a new detection cost function (DCF) that highly penalizes false alarms (FA): a typical system yields approximately 0.01% false alarms at the minimum DCF operating point. At that operating point the detection threshold falls in the tail of the non-target score distribution which is not a regime that typical speaker verification and normalization algorithms optimize for. Typical algorithms focus on ensuring a large degree of separability between target and non-target score distributions and typical score normalization schemes attempt to reduce score distribution variability over different target models and test utterances.

This work examines the low false alarm regime and proposes algorithms that attempt to tackle it directly. The approaches leverage a large set of unlabeled impostor utterances to identify suspect false alarm trials whose match score can then be penalized.

The paper begins by briefly introducing the baseline system used in this work which consists of the total variability (TV) speaker comparison system followed by symmetric score normalization (SNorm) and highlighting the difficulty encountered by this system in the low-FA regime. The proposed methods to tackle this difficulty are then presented and evaluated on an extended English telephony development set from the 2008 NIST SRE with promising outcomes. The methods are then applied to the telephony condition of the 2010 NIST SRE with less favorable results. This unexpected discrepancy between the 2008 and 2010 evaluations is explored and the likely reason identified and fixed resulting in improved performance on the 2010 SRE.

2. BASELINE SYSTEM AND THE PROBLEM

2.1. Total Variability (TV)

The baseline system used in this work is the total variability (TV) system proposed in [4]. A brief description is provided in this section, extended details may be found in the original work. The TV system is based on the TV space which is a linear subspace of the GMM supervector space that includes all variability, speaker and nuisance, observed in the GMM supervectors of the training dataset. The subspace is chosen such that for a given speech utterance its corresponding GMM supervector (m_{utt}), adapted from a universal background model (UBM), can be represented as

$$m_{utt} = m_{UBM} + T_{tel} w_{utt} \tag{1}$$

where m_{UBM} is the UBM mean, T_{tel} is the low-rank matrix defining the TV subspace, and w_{utt} is the corresponding factor of the utterance in the space. In this work T_{tel} is trained on telephony data such that the ws are Normally distributed with zero mean and unit variance.

In the TV space, the match score ($s(utt_a, utt_b)$) between two utterances utt_a and utt_b is computed as a weighted inner-product where the weighting effectively performs channel compensation:

$$s(utt_a, utt_b) = \frac{w^t_{utt_a} AW^{-1}A^t w_{utt_b}}{\sqrt{w^t_{utt_a} AW^{-1}A^t w_{utt_a}}\sqrt{w^t_{utt_b} AW^{-1}A^t w_{utt_b}}}.$$

A corresponds to a linear discriminant analysis (LDA) projection matrix, trained to project into a space that captures inter-speaker variability while avoiding within speaker variability, and W is the within speaker covariance matrix computed in the LDA space. It is of importance to note the symmetry in the scoring function.

The system operates on cepstral features, extracted using a 25 ms Hamming window. 19 Mel frequency cepstral coefficients together with log energy are calculated every 10 ms. Delta and double delta coefficients were then calculated using a 5 frame window to produce 60-dimensional feature vectors. This 60-dimensional feature vector was subjected to feature warping using a 3 s sliding window. The UBMs used are gender dependent Gaussian mixture models containing 2048 Gaussians. The UBM and the LDA projection are trained on data from the Switchboard II, Switchboard cellular, and telephone utterances from the 2004/05/06 NIST SRE. The TV subspace is trained on the these corpora as well as the Fisher English corpus. The WCCN matrix is computed using only the telephone utterances from the 2004/05/06 NIST SRE data sets. The focus of this work will be on the one conversation enroll one conversation test scenario and the development set is an extended trial set drawn from the 2008 NIST SRE English telephony data. Final performance will be measured on the extended condition 5 of the 2010 NIST SRE which consists of normal vocal effort English telephony speech.

2.2. Symmetric Score Normalization (SNorm)

It is common for speaker verification systems to be followed by a score normalization technique, the goal of which is to reduce within trial variability leading to improved performance, better calibration, and more reliable threshold setting. In this work symmetric score normalization (SNorm) [5] is used as the baseline. For every score

This work was sponsored by the Federal Bureau of Investigation under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions,and recommendations are those of the authors and are not necessarily endorsed by the United States Government. This work was also supported in part by the Texas Instruments Leadership University Program.

$s(utt_a, utt_b)$ between two utterances, the corresponding SNorm score $\hat{s}(utt_a, utt_b)$ is

$$\hat{s}(utt_a, utt_b) = \frac{s(utt_a, utt_b) - \mu_a}{\sigma_a} + \frac{s(utt_a, utt_b) - \mu_b}{\sigma_b} \quad (2)$$

where μ_i and σ_i are the mean and standard deviation of the scores of utt_i scored against an impostor list. Gender dependent impostor lists are used that consist of 614 female and 406 male English telephone utterances drawn from the 2005/06 NIST SRE data-sets.

2.3. The Problem

The 2010 NIST SRE set a very low prior of 0.001 on target trials in the detection cost function (DCF) which results in false alarms (FAs) costing significantly more than misses. The minimum DCF threshold, therefore, falls in the tail of the non-target trial scores as can be seen in Figure 1. For the TV baseline with and without SNorm the figure shows the minimum DCF threshold and the overlap of the histograms of the target and non-target trial scores of the development set used. The low overlap between target and non-target trials in both plots and the reduced variance of the scores for the SNormed system highlight the efficacy of the TV system for speaker verification and SNorm for score normalization. However, TV and SNorm, though effective, do not specifically tackle the tails of the score distributions in the overlap region, which we will attempt to do in this work.

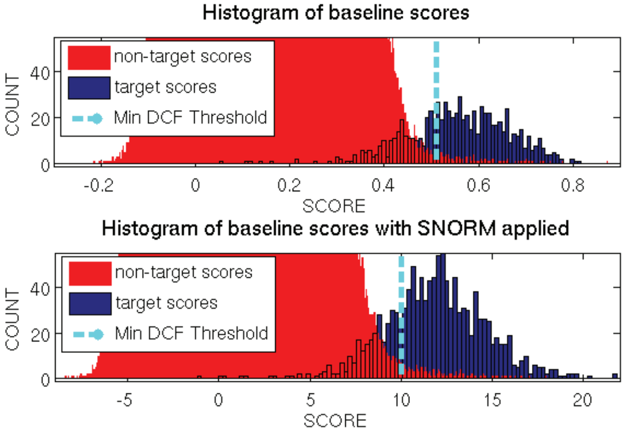

Fig. 1. The Problem

3. PROPOSED SYSTEMS

We tackle the problem by trying to identify the high scoring non-target trials, i.e. the trials in the tail of the non-target trial distribution. This is done by leveraging a wealth of available data as an impostor set, a set of utterances that do not share common speakers with the development or test set, and asking the question: "are the two utterances in the trial more similar to one another or to utterances in the impostor set?" Gender dependent impostor sets are used consisting of 9281 female and 6932 male telephony utterances from the 2004/05/06 NIST SREs excluding those used to perform SNorm. All match scores, between the trial utterances or a trial utterance and an impostor utterance, are computed using the symmetric equation (2).

In the proposed methods, one is not constrained to using a specific system to score trials. However, inner product scoring based systems, such as TV and inner product decision functions [6], are especially well suited because they allow for fast and efficient comparison of a large number of utterances, as is needed when scoring each trial utterance against the thousands of impostor utterances.

Table 1. % of trials flagged on the development set

Strategy	% target flagged	% non-target flagged
NN-OR	18.7	99.87
NN-AND	25.2	99.96

3.1. False Alarm Detectors

3.1.1. Nearest Neighbor AND/OR (NN-AND/NN-OR)

We begin with two strategies to detect whether a trial is likely a non-target trial, i.e. one that would contribute to false alarms. These strategies were motivated by previous work [7] that used NN-graphs and approximate geodesic distances to compare two utterances. The first proposed strategy, called NN-OR, flags a trial as a non-target if *either* of the trial utterances, enrollment or test, are closer, indicated by a higher match score, to utterances in the impostor set than to the other trial utterance. The second, called NN-AND, flags a trial as non-target if *both* trial utterances are closer to utterances in the impostor set.

We evaluate the two strategies on the development data-set by examining the percentage of target and non-target trials that get detected and labeled as non-target trials, a perfect detector would have detected and flagged 100% of the non-target and 0% of the target trials. Table 1 shows that while the majority of the non-target trials were detected correctly, a significant number of target trials were falsely detected.

This observation suggests a strategy that, rather than making a hard decision to label all utterances flagged by these detectors as non-targets, penalizes those trials by subtracting an offset from the trial score. Figure 2 shows the minDCF and EER values on the development set as a function of the offset, and shows that both strategies perform better than the baseline SNorm system and that NN-AND with an offset of 2 yields the best performance.

3.1.2. Nearest Neighbor Difference (NN-DIFF)

In both NN-AND and NN-OR each trial is either flagged as a non-target or not. We now propose to instead assign a confidence score $c_D(enr, tst)$, where enr is the enrollment utterance and tst is the test utterance, to each trial based on how suspect a trial is by:

$$\begin{aligned} c_D(enr, tst) &= \frac{1}{2}\{\hat{s}(enr, tst) - \hat{s}(enr, NN_1(enr))\} \\ &+ \frac{1}{2}\{\hat{s}(enr, tst) - \hat{s}(tst, NN_1(tst))\}. \quad (3) \end{aligned}$$

where $\hat{s}(.,.)$ is the SNormed TV match score (2), and $NN_1(utt)$ is the utterance in the impostor set that is nearest, has highest match score, to utt. c_D will therefore take on a large negative value when we are highly confident that a trial is a non-target and a large positive value when we are highly confident it is a target trial. The confidence score is then fused with the baseline SNorm score to obtain the final trial score

$$s_D(enr, tst) = (1 - f)\hat{s}(enr, tst) - f * c_D(enr, tst), \quad (4)$$

where $f \in [0, 1]$. Figure 2 shows the minDCF and EER values on the development set as a function of the fusion parameter, with $f = 0$ being the baseline SNorm system and $f = 1$ using the confidence score as the trial score. The parameter sweep suggests that a good choice of f is in the range of .3 to .6. Also, setting the trial score to be the confidence score, i.e. $f = 1$, performs well at the minDCF point yet poorly at the EER.

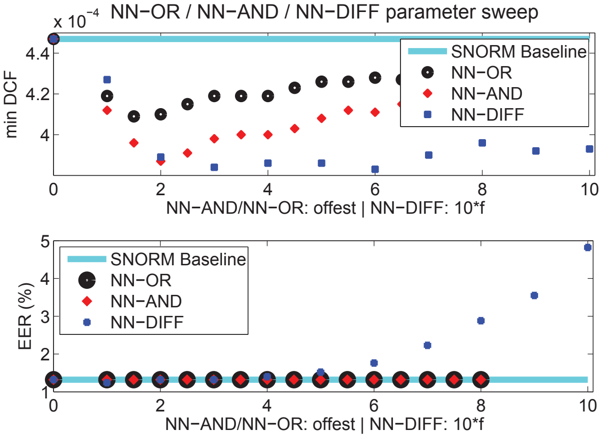

Fig. 2. Offset penalty sweep

3.2. K Nearest Neighbor Difference (KNN-DIFF) and Adaptive Symmetric Normalization (ASNorm)

The first set of proposed methods share a common shortcoming: they heavily rely on a single nearest neighbor from the impostor set. We therefore extend the NN-DIFF idea in an attempt to reduce this reliance by averaging the scores of the top K NNs rather than just the first, and call it KNN-DIFF. The confidence score is now

$$\begin{aligned} c_{KD}(enr, tst) &= \tfrac{1}{2}\{\hat{s}(enr, tst) - \mu(\hat{s}(enr, NN_K(enr)))\} \\ &+ \tfrac{1}{2}\{\hat{s}(enr, tst) - \mu(\hat{s}(tst, NN_K(tst)))\}, \quad (5) \end{aligned}$$

where $\mu(.)$ is the mean and $NN_K(.)$ is the set of the K NNs. As K gets large we can further divide out the standard deviation in the confidence score resulting in an adaptive symmetric normalization (ASNorm), similar to TopNorm [2] and ATNorm [3]:

$$\begin{aligned} c_{ASN}(enr, tst) &= \frac{\hat{s}(enr, tst) - \mu(\hat{s}(enr, NN_K(enr)))}{\sigma(\hat{s}(enr, NN_K(enr)))} \\ &+ \frac{\hat{s}(enr, tst) - \mu(\hat{s}(tst, NN_K(tst)))}{\sigma(\hat{s}(tst, NN_K(tst)))}, \quad (6) \end{aligned}$$

where $\sigma(.)$ is the standard deviation.

Figure 3 shows how increasing K affects each of the strategies. Notice that a lower number of cohorts, $K = 50$, is needed in KNN-DIFF, while $K = 1500$ is best for ASN.

We now choose the best performing confidence scores $c_{KD,K=50}$ and $c_{ASN,K=1500}$ and fuse them with the baseline SNorm scores,

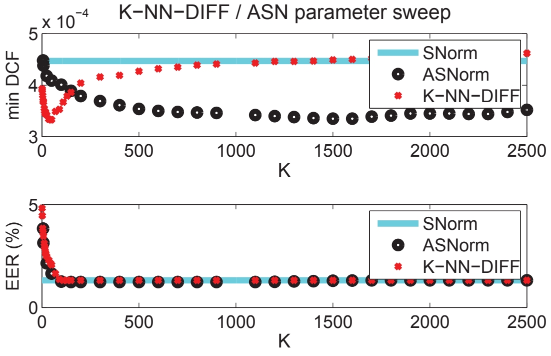

Fig. 3. Offset penalty sweep

$$\begin{aligned} s_{KD}(enr, tst) &= (1 - f)\hat{s}(enr, tst) - f c_{KD,K=50}(enr, tst) \\ s_{ASN}(enr, tst) &= (1 - f)\hat{s}(enr, tst) - f c_{ASN,K=1500}(enr, tst), \end{aligned}$$

and show the sweep of the fusion parameter f in Figure 4. The fusion shows that to optimize for minDCF f should be set to 0, meaning that the confidence score c_{KD} or c_{ASN} should be used rather than fusing with SNorm. However, the fusion does benefit EER, specifically in the KNN-DIFF case, where $f = .7$ seems to be a reasonable trade-off between DCF and EER.

Table 2. % of trials flagged on the development set

Strategy	DCF*1e4	EER (%)
Baseline: TV no SNorm	5.32	1.73
Baseline: TV with SNorm	4.47	1.32
NN-OR offset=1.5	4.09	1.32
NN-AND offset =2	3.87	1.32
NN-DIFF	3.93	4.82
NN-DIFF fused f=.5	3.86	1.52
KNN-DIFF K=50	3.33	2.07
KNN-DIFF K=50 fused f=.7	3.58	1.32
ASNorm K=1500	3.35	1.30
ASNorm K=1500 fused f=.7	3.46	1.24

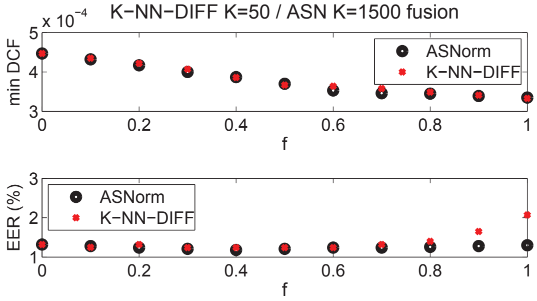

Fig. 4. Fusion of KNN-DIFF and ASNorm with SNorm

3.3. Analysis

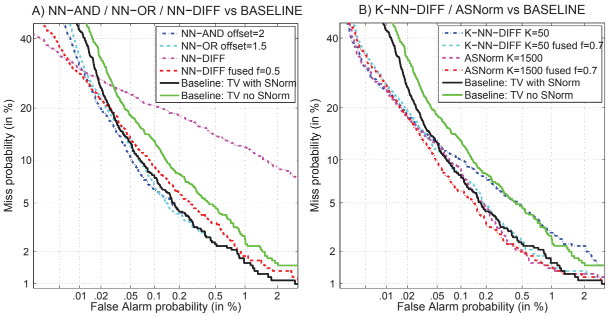

Fig. 5. DET plots of the different systems on the development set.

We first examine Table 2 and Figure 5 (A) and notice that even the simplest of the proposed strategies, that rely only on the first NN and make hard decisions to flag a trial as non-target, can yield overall improvement over SNorm and specifically a 13% relative improvement at minDCF. Using the confidence score in NN-DIFF as the trial score, however, aggressively targets the low-FA regime of the DET curve at the expense of the rest. Fusing the confidence score with SNorm provides a less aggressive system that improves in the regime of interest while performing reasonably elsewhere.

The results of KNN-DIFF and ASNorm shown in Table 2 and Figure 5 (B) show that utilizing more than one NN in the confidence score further improves performance at minDCF, yielding a 25% relative improvement over SNorm. However, the two methods differ greatly in performance over the rest of the DET curve: KNN-DIFF only shows improvement in the low-FA regime while ASNorm improves overall. Fusing the confidence score with the SNorm trial score trades off performance at the low-FA range for overall performance.

4. NIST SRE 2010 RESULTS

We now present in the first columns of Table 4 and Figure 6 the results of the proposed methods on the test set, condition 5 of the 2010 NIST SRE, versus the baselines. It is apparent from the DET plot that the improvement observed on the development data-set is not seen on the test set, specifically at the minDCF operating point.

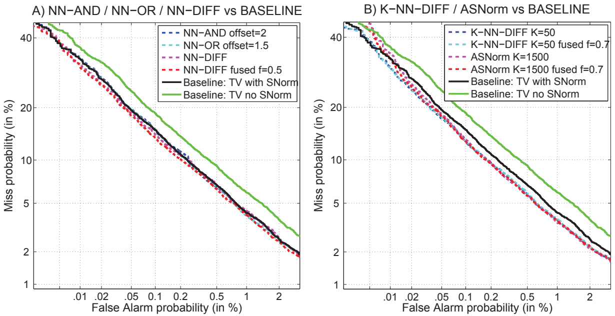

Fig. 6. DET plots of the different systems on 2010 NIST SRE.

In an attempt to resolve this discrepancy we examine the percentage of trials being flagged as non-targets in the simple NN-AND and NN-OR algorithms, this is shown in the first two columns of Table 3. Comparing these percentages to those in Table 1 it is apparent that the test data-set is interacting with the impostor set in a different manner than the development set: specifically a significantly smaller percentage of trials were being flagged as non-targets. This could be for one of two reasons: either the within set variability is lower for the test set than the development set, or the impostor set is better matched to the development data.

Changing the within set variability would require changing the system we are using to drive the experiments, we therefore attempt to better match the impostor set to the test set by including the 2008 NIST SRE English telephony utterances in the impostor set. The last two columns of Table 3 show that there is about a two-fold increase in the number of flagged utterances, indicating that the 2008 data is better matched to the 2010 data. The last two columns of Table 4 and Figure 7 show that augmenting the impostor set to better match the test data does improve performance over the original impostor set.

Table 3. % of trials flagged on the test set

$Strategy$	% target flagged	% non-tar flagged	% target flagged+08	% non-tar flagged+08
NN-OR	5.7	99.32	8.38	99.71
NN-AND	10.7	99.76	16	99.92

To provide a fair comparison between our proposed systems and the SNorm baseline we augment the SNorm set with a uniformly selected subset of utterances from the 2008 data-set. The comparison with the baseline is presented in Table 4 and Figure 7 and, even though the improvement is not as dramatic as was seen on the development data, there is a consistent improvement in performance over the DET range between the minDCF point and the EER point. Specifically, a $5-10\%$ and $8-10\%$ relative improvement at the minDCF and EER points respectively for the KNN-DIFF and ASNorm systems. However, even though the performance did improve it still falls short of expected. This may be because the percentage flagged in the last two columns of Table 3 are still lower than those in Table 1 indicating a likely persistent miss-match not addressed by augmenting the impostor set.

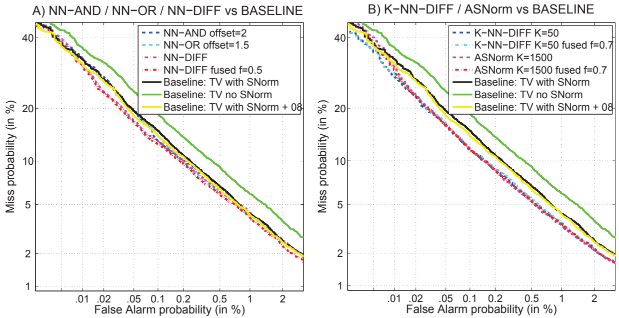

Fig. 7. DET plots of the different systems with the augmented impostor set on 2010 NIST SRE.

Table 4. minDCF and EER breakdown on test set

$Strategy$	DCF *1e4	EER (%)	DCFe4 with 08	EER (%) with 08
Baseline: TV no SNorm	4.62	2.82	4.62	2.82
Baseline: TV with SNorm	4.21	2.32	4.13	2.29
NN-OR offset=1.5	4.21	2.30	4.21	2.32
NN-AND offset =2	4.23	2.32	4.28	2.32
NN-DIFF	4.07	2.30	4.11	2.32
NN-DIFF fused f=.5	4.07	2.22	4.05	2.16
KNN-DIFF K=50	4.00	2.11	3.70	2.06
KNN-DIFF K=50 fused f=.7	4.01	2.13	3.80	2.09
ASNorm K=1500	4.33	2.09	4.02	2.08
ASNorm K=1500 fused f=.7	4.17	2.11	3.92	2.11

5. CONCLUSION AND FUTURE WORK

The goal of this work was to attempt to directly tackle the newly proposed DCF with systems that leverage a large impostor set. Our results on the development set were very promising with even the simplest algorithms outperforming the baseline, however, performance on the test set was on-par with the baseline. Upon exploring this discrepancy, it became apparent that an impostor set that is well matched to the data of interest is crucial to the proposed algorithms. Augmenting the impostor to better satisfy this criterion led to better performance. However, performance still fell short of what was observed on the development set, most likely due to not addressing all of the miss-match. Future work will focus on identifying well matched impostor sets, as well as further exploring this apparent miss-match between the 2010 NIST SRE data-set and the NIST SRE data from previous years.

6. REFERENCES

[1] "The NIST year 2010 speaker recognition evaluation plan," http://www.itl.nist.gov/iad/mig/tests/sre/2010/index.html, 2010.

[2] Y. Zigel and M. Wasserblat, "How to deal with multiple-targets in speaker identification systems," in *Proc. Odyssey*, 2006.

[3] D. E. Sturim and D. A. Reynolds, "Speaker adaptive cohort selection for tnorm in text-independent speaker verification," in *Proc. ICASSP*, 2005, vol. I, pp. 741–744.

[4] N. Dehak, P. Kenny, R. Dehak, P. Ouellet, and P. Dumouchel, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, To Appear.

[5] S. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification," in *Proc. Odyssey*, 2010.

[6] W. M. Campbell, Z. Karam, and D. E. Sturim, "Speaker comparison with inner product discriminant functions," in *Advances in NIPS 22*, 2009, pp. 207–215.

[7] Z. N. Karam and W. M. Campbell, "Graph-embedding for speaker recognition," in *Proc. Interspeech*, 2010.