Score Distribution Scaling for Speaker Recognition

Vinod Prakash, John H.L. Hansen

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas
Richardson, Texas 75083-0688, U.S.A.
{vinod.prakash, john.hansen}@utdallas.edu

10.21437/Interspeech.2007-167

Abstract

In this study, we transform the verification scores of a speaker recognition system in order to standardize the imposter score distribution, this facilitates setting of a speaker-independent threshold at desired False Alarm (FA) rates. Impostor score distributions are estimated using GMMs, and a univariate Gaussianization [1] transform (which is a monotonically increasing mapping) is applied on the scores. It is shown that if a monotonically increasing mapping is used, the Probability of correct detection for a given setting of the FA is maintained as before. Hence, the proposed technique performs distribution scaling without affecting the False Alarm to False Reject relationship of the original test statistic. The maximum (relative) mismatch between the obtained and desired False Alarm rates is less than 10% for a wide range of False Alarm rates. When compared to modeling the imposter score distributions using a single Gaussian (Z-norm case), the overall relative mismatch is reduced by an average of 30%. While the application focus is on speaker recognition, the proposed technique can be used for other binary speech classification tasks as well.

Index Terms: Speaker Verification, Speaker Normalization, Score Normalization, Binary Classification

1. Introduction

Speaker Recognition is focused on characterizing subjects on the basis of their voice [2]. In Speaker Verification, enrollment data is provided for a speaker, and when given a test utterance, the recognizer is required to produce a binary decision as to whether the test utterance was spoken by the enrolled speaker or not (i.e., the test utterance was from an "imposter"). (All experiments in this study, and the terminology for the remainder of the paper addresses the verification problem.) Speaker verification systems typically construct statistical models for the enrolled speaker and potential imposters. Given a verification utterance, a test statistic(score) is calculated using these models and the observations. This score is then compared against a threshold, and a decision is made to accept or reject the speaker. The threshold determines the tradeoff between the two types of error every non-perfect binary classification system can make: *False Acceptance* (FA) of an imposter or *False Rejection* (FR) of an enrolled speaker.

Depending on the application the threshold is set so as to satisfy a pre-determined FA or FR rate or is based on the minimization of a cost function involving the FA and FR rates (as in the NIST-SRE's[3]). In practice, depending on the actual technique used, speaker-specific calibration of the threshold is problematic due to biases that are either speaker dependent and/or dependent on the validation set used. This results in poor generalizations on unseen data[3, 4, 5]. In evaluations like the NIST-SRE's, the verification scores for all Hypothesized speakers should be on a common scale since a single threshold is used for performance evaluation. For this purpose, most commonly imposter score distributions are modeled since there would not be enough data to accurately model the score distribution of enrolled speakers.

In this paper we perform speaker dependent transformations of the scores so that a robust speaker independent threshold, at a desired setting of the FA rate, can be used for verification. Also the proposed transformation does not affect the performance of the original verification system, in the sense that after the transformation, the Probability of correct detection (P_D) for a given False Alarm rate P_{FA} is maintained as before. The transformation we use attempts to map the distribution of the imposter scores to a standard normal distribution. A widely used imposter-centric distribution scaling strategy is the Z-norm [6, 7], where imposter score distributions are estimated using a single Gaussian. We generalize that by modeling the distribution using a mixture of Gaussians (GMMs).

The remainder of this paper is structured as follows, the next section covers the objective formulation of the problem and obtains a functional form for the transformations. Sec. 3 contains details about the proposed algorithm. Experimental results and analysis are provided in Sec. 4., with overall conclusions in Sec. 5.

2. Objective Formulation

Cast in terms of binary hypothesis testing, the speaker verification problem can be expressed as, H_0: The verification utterance is *not* from the claimed speaker. vs. H_1: The verification utterance is from the claimed speaker. Since the focus here is on transformations to the *final* score returned by a speaker verification system. The decision rule for a verification utterance $\vec{O}$ can be written as:

$$X \begin{cases} \geq \tau & : \text{accept } H_1, \\ < \tau & : \text{reject } H_1 \text{ (accept } H_0). \end{cases} \quad (1)$$

where τ is the decision threshold, and $X = S(\vec{O})$ is a decision/test statistic calculated from the observations and speaker (imposter) models. For example, in a GMM-UBM based verification system [8], X would be the average log-likelihood ratio of the speaker model against the UBM for the test utterance. The following proposition specifies a condition under which the transformed scores, $Y = g(X)$, maintain the relationship between P_{FA} and P_D as before. (In what follows, it is assumed

This work was supported by RADC under contract FA8750-05-C-0029 and by University of Texas at Dallas under project EMMITT.

that the random variables used have continuous and strictly increasing Cumulative Distribution Function's (CDF)).

Proposition 1. *for a given P_{FA} the test statistic's X and Y have the same P_D if $g(.)$ is a strictly increasing function*

Proof. For a given threshold τ,

$$\begin{aligned} P_{FA} &= \int_{\tau}^{\infty} f_{X|0}(x|0)dx \\ &= 1 - F_{X|0}(\tau) \\ \therefore \tau &= F_{X|0}^{-1}(1 - P_{FA}) \\ P_D(X) &= \int_{\tau}^{\infty} f_{X|1}(x|1)dx \\ &= 1 - F_{X|1}(\tau) \\ &= 1 - F_{X|1}(F_{X|0}^{-1}(1 - P_{FA})) \end{aligned}$$

Similarly for the statistic Y,

$$P_D(Y) = 1 - F_{Y|1}(F_{Y|0}^{-1}(1 - P_{FA})) \quad (2)$$

Since $g(.)$ is monotonically increasing (and hence invertible), for some $\beta \in (-\infty, \infty)$ and $\gamma \in [0, 1]$,

$$\begin{aligned} Pr\{X \leq \beta\} &= Pr\{g(X) \leq g(\beta)\} \\ \Rightarrow F_X(\beta) &= F_Y(g(\beta)) = \gamma \quad (3) \\ \therefore \beta = F_X^{-1}(\gamma) &= g^{-1}(F_Y^{-1}(\gamma)) \quad (4) \end{aligned}$$

From (2), (3) and (4),

$$\begin{aligned} P_D(X) &= 1 - F_{X|1}(F_{X|0}^{-1}(1 - P_{FA})) \\ &= 1 - F_{Y|1}(g(g^{-1}(F_{Y|0}^{-1}(1 - P_{FA})))) \\ &= 1 - F_{Y|1}(F_{Y|0}^{-1}(1 - P_{FA})) \\ &= P_D(Y) \end{aligned}$$

□

Equipped with the result of Proposition 1, the task is to determine a monotonically increasing transformation such that the distribution of Y given imposter utterances matches that of a standard normal. For this we use the technique of univariate Guassianization [1]. Denoting the CDF of the standard normal density by $G(.)$ and using (3),

$$\begin{aligned} F_{X|0}(x) &= F_{Y|0}(g(x)) \\ \text{and it is desired that, } F_{Y|0}(y) &= G(y) \\ \therefore y = g(x) &= G^{-1}(F_{X|0}(x)) \quad (5) \end{aligned}$$

Since the CDF's are monotone rising, utilizing the result of proposition 1 the transformation given by (5) does not affect the P_{FA} to P_D relationship of the untransformed scores. In practice $F_{X|0}(.)$ has to be estimated using verification scores obtained for pseudo-imposters. The next section addresses details on the estimation procedures used in this study.

3. Proposed Method

3.1. GMM based Modeling

In our experiments, we modeled $F_{X|0}(.)$ using GMMs,

$$\begin{aligned} f_{X|0}(x) &= \sum_{i=1}^{N} \omega_i \mathcal{N}(x; \mu_i, \sigma_i^2) \\ \Rightarrow F_{X|0}(x) &= \sum_{i=1}^{N} \omega_i G\left(\frac{x - \mu_i}{\sigma_i}\right) \\ \therefore y &= g(x) \\ &= G^{-1}\left(\sum_{i=1}^{N} \omega_i G\left(\frac{x - \mu_i}{\sigma_i}\right)\right) \quad (6) \end{aligned}$$

Experiments were conducted on model orders N ranging from 2 to 32. The following two methods were used to estimate the speaker-dependent GMM parameters:

3.1.1. ML Estimation

For an enrolled speaker a fixed number of pseudo-imposter utterances are scored. These scores are used to obtain a Maximum Likelihood estimate of the GMM parameters via the EM algorithm.

3.1.2. MAP Estimation

Using development data, a speaker verification system is established. The imposter verification scores of all enrolled speakers for this system are pooled together and a speaker independent (SI) GMM is constructed. Pseudo-imposter scores for a target speaker are used to MAP adapt[9] the mixture weights, means and variances of the SI GMM to obtain the speaker dependent GMM parameters.

4. Experiments

4.1. Experimental setup

The speech analysis frame rate is set to 20 ms with a 10 ms skip rate. Speech is pre-emphasized with the filter $(1 - 0.95z^{-1})$. Nineteen-dimensional Mel-Frequency Cepstral Coefficients (MFCC's) are used as features. Low-energy speech frames are removed using an adaptive TEO based frame selection technique[10]. A standard UBM-GMM[8] setup (means only adaptation, relevance factor of 16) was used for speaker modeling. The average of the log-likelihood ratio values was used as the output statistic. No additional normalization (e.g., H-norm, T-norm) was performed.

Evaluations were carried out using the 10sec train/ 10sec test NIST-SRE 2006 data (A total of 2942 true and 29608 imposter trials).We operated under the low enrollment/verification mode in order to have low turnaround times for our experiments. SRE 2004 data was used to build a 64 mixture UBM. SRE 2005 data was used for score distribution scaling as follows:1000 SRE 2005 test files from each gender were used as (gender-specific) pseudo-imposters for both the MAP and ML methods. All the scores from the SRE 2005 imposter trials were pooled together, (in a gender-specific way) to construct an SI base model for MAP estimation.

4.2. Evaluations

The verification scores were transformed according to (6). Threshold values were set for a range of P_{FA}'s using CDF equa-

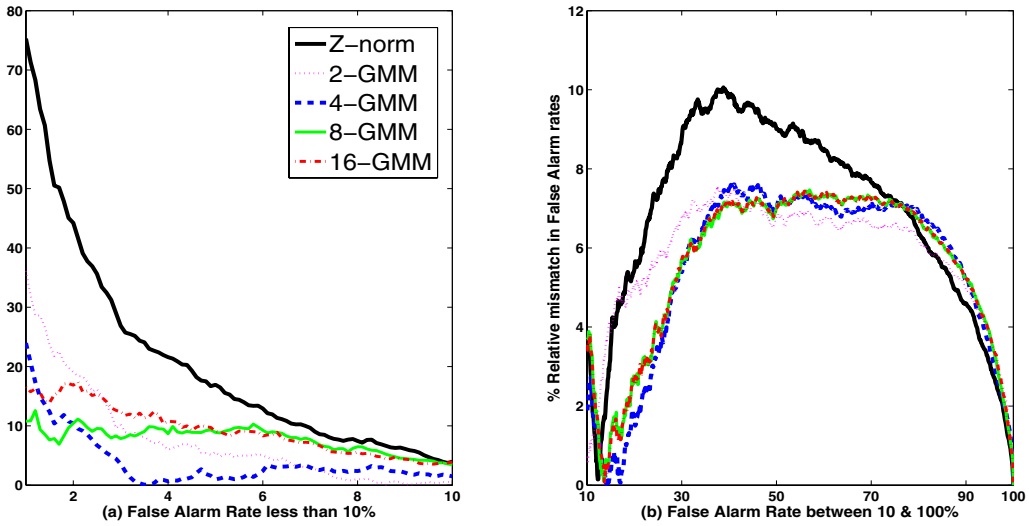

Figure 1: *% Relative mismatch, after Gaussian CDF based threshold setting, between ideal and empirical false alarm rates for (a) low FA range (0 - 10%) and (b) wider FA range (10 - 100%). Note: Z-norm and GMM transform based result key is the same for both (a) & (b). Also note the difference in the Y axis values across (a) & (b)*

tions for the standard Gaussian, and the empirical false alarm rates at these threshold settings was obtained. Fig 1 shows the relative mismatch between the obtained and desired false alarm rates for different model orders. The area beneath each of the curves in Fig 1 provides an overall measure of the average mismatch in expected False Alarm rates. This average mismatch is used as the main performance metric and is labeled "avg. mismatch" in the last column of Table 1. There the value for 1-GMM (i.e. Z-norm) corresponds to the total area under the "Z-norm" plots from Fig 1 (a) and 1 (b). Likewise the avg. mismatch of 5.29% corresponds to the area under the 4-GMM plots in Fig 1 (a) and 1 (b). Some additional performance metrics (EER, area under ROC) obtained for the different model orders are also given in Table 1. Fig 2 shows the DET curves for un-normalized, Z-norm and a 4 mixture GMM (4-GMM). The actual decision points at a desired false alarm setting of 5% are also labeled. (Note: The reason for the performance difference between the raw and transformed scores is because distribution scaling compensates for the mismatch due to pooling together of scores of all enrolled speakers.)

The results shown are for GMMs trained using the ML method (Sec. 3.1.1). We also ran experiments using GMMs trained using the MAP method (Sec. 3.1.2) and found some improvement in performance but in general the method was sensitive to the relevance factor setting.

4.3. Analysis and Discussion

From Fig. 1(a) (low P_{FA} range) it can be seen that the observed False Alarm rate deviates from the desired significantly when imposter scores are normalized using a single Gaussian (Z-norm case). Whereas for higher GMM orders the deviation is significantly reduced. When viewed along with Fig. 1(b) (higher P_{FA}), where the relative mismatch is lower, it can be seen that the variation of the mismatch is smaller over the entire P_{FA} range. This implies that the threshold can be estimated fairly robustly. A quantitative measure of the benefit can be seen in the last column of Table1 where the areas under the curves in Fig. 1 are calculated. Relatively, a 30 - 35% improvement is

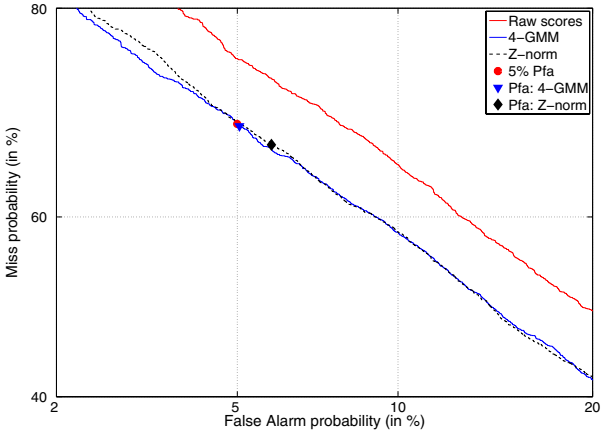

Figure 2: *DET curve showing performance of un-normalized, Z-norm(1-GMM) and 4-GMM modeling of imposter score distributions. Also shown are the observed and obtained decision thresholds at a Pfa of 5%.*

System	EER(%)	Area under ROC(%)	Avg. mismatch (%)
0-GMM	34.84	29.20	-
1-GMM	30.29	23.59	**8.09**
2-GMM	30.42	23.61	5.93
4-GMM	30.42	23.63	**5.29**
8-GMM	30.35	23.63	5.74
16-GMM	30.39	23.63	5.88

Table 1: *Summary statistics for score distribution scaling schemes, 0-GMM denotes raw scores, 1-GMM is equivalent to Z-norm. As far as ROC based metrics go, all score distribution scaling schemes perform comparably.*

seen over the Z-norm(1-GMM case).

Also, while there is little separation between the *a posteriori* performance of the distribution scaled scores for the single Gaussian and the GMM case, as can be seen from the DET curve (Fig 2) and the second and third columns of 1. The difference between the two methods is apparent when actual decisions need to be made based on *a priori* threshold settings. One instance of this is seen in the deviation in the False Alarm rate (marked with a diamond in Fig. 2), when the desired setting was 5%.

5. Conclusion

In this study, we determined transformations for imposter distribution scaling to enable a more robust setting of the decision thresholds given a desired False Alarm Rate. Having formulated a scheme to standardize imposter distributions without affecting the P_{FA} to P_D relationship between the original scores , we evaluated the method on the SRE corpus. Results showed that the GMM based score transformation method outperformed Z-norm by 35%. In terms of the average mismatch error between desired and obtained False Alarm rates.

6. References

[1] S.Chen, R.A. Gopinath, "Gaussianization", Proc. NIPS 2000, Denver, Colorado.

[2] J.P. Campbell, Jr., "Speaker recognition: a tutorial", Proceedings of the IEEE , vol.85, no.9. Sep 1997.

[3] G.R. Doddington, M.A. Przybocki, A.F.Martin, D.A.Reynolds, "The NIST speaker recognition evaluation - Overview methodology, systems, results, perspective", Speech Comm. 31 (2000).

[4] F. Bimbot, M. Blomberg, L. Boves, D. Genoud, H.P. Hutter, C. Jaboulet, J. Koolwaaij, J. Lindberg, J.B. Pierrot, "An overview of the CAVE project research activities in speaker verification", Speech Comm. 31 (2000).

[5] K. Chen, "Towards better making a decision in speaker verification", Pattern Recognition 36 (2003) .

[6] R. Auckenthaler, M. Carey, H. Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification Systems", Digital Signal Processing, vol. 10, 2000.

[7] D.A. Reynolds, "Comparison of Background Normalization methods for text-independent speaker verification", Proc. Eurospeech'97, 1997.

[8] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, "Speaker verification using adapted Gaussian Mixture Models", Digital Signal Processing, vol. 10, 2000.

[9] J.-L. Gauvain, C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", IEEE. Trans. on SAP, vol 2., Apr 1994.

[10] X. Zhang, J.H.L. Hansen, "In-set/Out-of-set speaker identification based on discriminative speech frame selection", Interspeech'05, 2005.