Enrolled Template Specific Decisions and Combinations in Verification Systems

Sergey Tulyakov, Jiang Li and Venu Govindaraju

***Abstract*— The matching scores in biometric systems are usually calculated using one enrolled (gallery) template and one test (probe, user) template. In this paper we investigate the dependencies existing between scores related to the same enrolled biometric template or to the same user biometric template. We discuss linear score dependency models which are handled by the Z- or T-normalization, and sample statistics based models. We show that different models might better account for score dependence in different matches. The dependency models might also be different for enrollee or for user specific score sets. Finally, we investigate the application of two such models, Z-normalization and second best score model, to construct enrollee specific verification system decision and combination algorithms. The experiments are performed on NIST BSSR1 biometric score dataset.**

I. Introduction

The biometric matching score s is usually calculated as a distance between two templates, one enrolled in the biometric database and the other originating from the person currently being authenticated. The former might be called a gallery template, T_g , and the latter a test or probe template, T_p: , $s_{g,p} = s(T_g, T_p)$. The matching score is subsequently used in accept/reject decisions, in the combination of biometric matchers, or in some other decision process of biometric system.

Though the matching score s can be used in its original form, one can notice that during the biometric system operation some additional information about the strength of matching score $s_{g,p}$ might be available. For example, the previous authentication attempts might have produced a set of k matching scores $s_{g,p_1}, \dots, s_{g,p_k}$ between same gallery template T_g and some set of probes $T_{p_1}, \dots, T_{p_k}$. We can compare the current score $s_{g,p}$ with this set of previous scores and be more certain of whether the current score is genuine or impostor.

In this paper we explore the different approaches to using such background information. In particular, we are interested in utilizing the sets of matching scores related to specific enrolled template. In contrast to most works considering enrolled template specific decisions and combinations we only assume the availability of impostor scores related to enrolled template, and not genuine scores.

The content of the paper is the following. First, we introduce our dataset and the terminology. This would further clarify our goals before any other discussions. Then we present the discussion of relevant previous work. In section IV we analyze the possible dependences between matching scores and apply this analysis to our dataset. In the section V we present the results of experiments on accepting verification results using enrollee-specific methods, and in the section VI we present the results of enrollee-specific combinations.

The authors are with the Center for Unified Biometrics and Sensors (CUBS), University at Buffalo, USA `tulyakov@cubs.buffalo.edu`

II. Definitions and Experimental Setup

A. Dataset

In this work we used the biometric matching score set BSSR1 distributed by NIST[1]. This set contains matching scores for a fingerprint matcher and two face matchers ’C’ and ’G’. Fingerprint matching scores are given for left index ’li’ finger matches and right index ’ri’ finger matches. For experiments in acceptance decisions in verification system each of the four matchers was considered separately. For experiments in combinations, we considered all 6 possible 2-matcher combinations.

We used the bigger subsets of the database involving 6000 users (identification trials). Since the scores in these subsets originate from different persons, we assumed the independence of fingerprint and face matching scores, and considered randomly paired set of scores corresponding to 6000 identification trials of 3000 enrolled persons. Note that correspondence of scores to the same physical person was retained when combining scores of the same modality. Also note, that some enrollee and user scores had to be discarded due to apparent template acquisition errors, resulting in 5982 identification trials and 2991 enrollees. This data can be considered consisting of two 2991x2991 parts as shown in Figure 1.

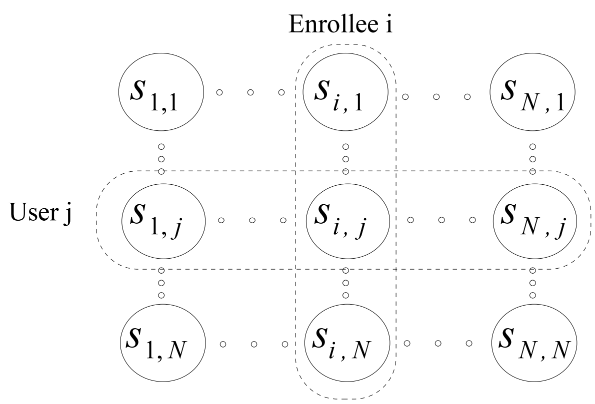


Fig. 1. The structure of matching scores in BSSR1 dataset.

We follow the BSSR1 score set definitions, calling the *enrollee* a person who is enrolled in the database, and *user* is a person being authenticated. The enrolled template is equivalent to gallery template, and user template is equivalent to probe or test template. Each score $s_{i,j}$ in the database is produced by matching the template of enrollee i with the template of user j. When the identities of an enrollee and the user coincide, $i = j$, then the matching score s_{ii} is the genuine matching score, and all other scores are impostor scores. The important characteristics of the BSSR1 set is that for each enrollee there is only one genuine score (actually, for face matchers there two genuine scores, but since we pair them with fingerprint matchers, we assume the same structure as in Figure 1), and that for each user there is only one genuine score.

B. Enrollee and User Specific Methods

Two major types of score dependencies can be discerned general structure of the matching score sets of Figure 1. First, the scores obtained using the same enrolled template i, $s_{i,1}, \ldots, s_{i,N}$. Since the same template participates in their calculation, it is expected that there might be some dependence between them. Second, the scores obtained using the same test template of user j, $s_{1,j}, \ldots, s_{N,j}$; again we expect some dependence between them. The two types of dependencies are not necessarily the same. In section IV-C we present some analysis of these dependencies in BSSR1 set.

The decision making and combination algorithms utilizing the matching scores can make special considerations for the existence of these dependencies. In this paper, we will call the methods accounting for dependence in scores related to a particular enrollee as *enrollee-specific* methods, and the methods accounting for dependence in scores related to a particular user as *user-specific* methods. Some methods might account for both dependencies and be both enrollee and user specific.

In order to utilize any of the score dependencies, a corresponding set of matching scores should be obtained first. For enrollee-specific methods we have to match the given enrolled template with few other templates (of the same person or of other persons). This matching, as well as the training of enrollee specific method, can be performed off-line and thus does not have any performance hits. For user-specific methods we have to do the matching and any user-specific adjustments of the method in real time. Thus we might anticipate additional processing time in this case.

There are two general approaches to utilizing score dependencies. In one case, a predetermined transform is performed on the matching score $s_{i,j}$ based on obtained enrollee-specific score set $s_{i,1}, \ldots, s_{i,N}$ or user-specific score set $s_{1,j}, \ldots, s_{N,j}$. The examples of such transformations are Z-normalization and T-normalization. In the other case, some statistics of score sets can be extracted, $\boldsymbol{\theta}_{enr,i}$ or $\boldsymbol{\theta}_{user,j}$, and used along with score $s_{i,j}$ to construct an enrollee- or user-specific method. We will explore the use of second best score statistics in our experiments.

C. Training and Testing Procedures

We use the bootstrap sample testing technique [2] in our experiments. For each bootstrap test, we chose randomly testing, training and validation sets, each of size 2x997x997 (two separate square matrices). Such specific size was prompted by the maximum number of scores related to a single user, 2991. Note, that a bootstrap sample does not contain all 5982x2991 scores of original set. Such configuration implies that for each user we have 997 enrolled templates, and one of them genuine, and for each enrollee we have 997 users which are tested against this enrollee. Thus, the effective size of matrices of Figure 1 is N=997. 100 bootstrap tests were performed for each experiment.

Most of our algorithms relied on likelihood ratio decision making and likelihood ratio combinations and the densities of genuine or impostor scores have been estimated by Parzen kernel method. Since each enrollee- or user-specific score set has 1 genuine and 996 impostor scores we chose to use only a single random impostor score for each genuine score from corresponding set for training and testing the score densities. The validation sets were used to estimate the kernel sizes for score density approximations.

III. PREVIOUS WORK

Our work is most closely connected with two general research directions - the user-specific decision making and combinations in arbitrary biometric systems, and the score normalization techniques for speaker authentication.

Working in the first direction, Toh et al. [3] learn combination functions (multivariate polynomials) for each enrolled person separately (local learning), or use different decision thresholds for different enrollees (local decision). Since the number of genuine templates for each enrollee available for training was small (5 templates, 5*4/2 genuine scores), the authors chose to randomly generated additional pseudo-genuine scores by adding noise to original genuine scores.

Fierrez-Aguilar et al. [4] modeled the densities of genuine and impostor scores by normal distributions. The parameters of these distributions were adjusted for each enrollee using a set of training scores. Though such modeling delivers a robustness in the parameter estimates for specific enrollees, the assumption of score distribution normality might not hold and can lead to incorrect error rate estimations [5].

The enrollee-specific weighted sum combinations were explored in [6] and [7]. Jain and Ross [6] searched the weights for each enrollee by the exhaustive search with the criteria of minimizing EER. Snelick et al. [7] assigned the weights heuristically using d-prime metric.

All these presented methods are enrollee-specific, rather than user-specific. In all cases a significant number of genuine scores for each enrollee was available, which allowed proper training. In the case of our work, BSSR1 database has only single genuine score for each enrollee, so these methods might not be applicable.

The second direction, the score normalization methods in speaker authentication, considered some user- or enrollee-specific techniques which might be applicable in the case of genuine score non-availability. The T('Test')-normalization and Z('Zero')-normalization are two such methods [8], [9]. The normalization is done using the same formula:

$$s \to \frac{s-\mu}{\sigma} \quad (1)$$

but the parameters μ and σ are calculated differently. For T-normalization, μ and σ are the sample mean and variance of the set of user-specific impostor scores. These are the scores produced during a single *test* or authentication attempt. Z-normalization, on the other hand, uses a set of enrollee-specific scores. Note, that sometimes Z-normalization is defined as normalization with parameters μ and σ calculated using all available training scores of all enrollees; in these paper we will assume that only scores of particular enrollee are used. The use of normalization techniques seems to be present in many algorithms evaluated by NIST, for example, in speaker recognition [10] and face recognition [11]. Note, that though [11] uses term Z-normalization, the actual method seems to be based on user-specific impostor score sets, and following definitions of this paper it should be rather called T-normalization.

Instead of using all available impostor scores related to a particular enrollee or user, we might consider some selection of them. The *cohort methods* [12], [13] introduced in speaker verification find a cohort - a subset of enrolled templates close to the particular enrolled template. During matching test template to the enrolled, the templates belonging to the cohort of enrolled templates can be used for normalizations. Auckenthaler et al. [8] separated cohort normalization methods into cohorts found during training (constrained) and cohorts dynamically formed during testing (unconstrained cohorts). As we have presented in [14], different use of cohorts can result in different types of normalizations (user-specific, enrollee-specific or both). The example, adaptive cohort model T-normalization, combining both enrollee- and user-specific parameters for normalization is given in [15].

The T- and Z-normalizations given by Equation 1 are able to deal with the linear score dependencies (section IV-A). In order to deal with possible more complex dependencies between matching scores, we introduced second best score model [16], which considers a pair of original score and second best score from a particular set of impostor scores instead of a single original matching score. In our previous research we investigated the use of second best model construction of user-specific methods for verification decision [16] and for combination of biometric matchers [14]. In those works we used the term 'identification model' to describe user-specific method, since the set of user-specific scores usually is obtained in a single identification trial. Second best score model was also utilized in cohort normalization for fingerprint verification [17].

In this work we consider the verification decision and combination algorithms based on enrollee-specific models: Z-normalization and second best impostor model. We discuss the difference between enrollee-specific and user-specific algorithm modification, and present their performance for both tasks, acceptance decision and combination.

IV. Dependencies between Matching Scores

A. Linear Score Dependence

Let $s_{i,1}, \ldots, s_{i,N}$ denote a set of matching scores related to one particular enrollee i. Let $\hat{\mu_i}$ and $\hat{\sigma_i}$ be the sample mean and sample variance of this score set. The Z-normalization performs the following transformation of matching scores:

$$s_{i,j} \to \frac{s_{i,j} - \hat{\mu_i}}{\hat{\sigma_i}} \quad (2)$$

The following discussion holds relevant for T-normalization as well, since it has same normalization formula with μ and σ derived from scores related to a particular user.

Suppose we have some score density $p(x)$ with mean of 0 and the variance of 1. Also, suppose that for each enrollee i we are given two random parameters μ_i and σ_i, and the scores in the enrollee-specific set are independently sampled according to

$$p_i(s) = p_{\mu_i,\sigma_i}(s) = \frac{1}{\sigma_i} p(\frac{s-\mu_i}{\sigma_i}) \quad (3)$$

It is easy to show that in this case the mean of scores in the enrollee-specific set is μ_i and the variance is σ_i. By calculating sample mean and variance estimates, $\hat{\mu_i}$ and $\hat{\sigma_i}$, and by applying Z-normalization (2) to the enrollee i related scores, the transformed scores will be approximately (due to approximations $\mu_i \approx \hat{\mu_i}$ and $\sigma_i \approx \hat{\sigma_i}$) distributed according to $p(x)$.

Equation (3) represents a possible model of how the dependencies between matching scores in enrollee- or user-specific sets originate. We can call it the *linear score dependency model*. Previously, Navratil and Ramaswamy [18] described the T-normalization using the property of *local gaussianity*, which assumes that function $p_i(x)$ is close to normal density with mean μ_i and variance σ_i. In our description we are not making any assumptions on the form of $p_i(x)$ except that it is generated for each specific enrollee or user by Eq. (3) using some common density p. There is also no assumptions on distributions of μ_i and σ_i (which are randomly chosen for each particular enrollee or user).

According to linear score dependency model the range of scores for each enrollee or user is shifted by μ_i and stretched by σ_i. Note, that there are two types of scores in enrollee- or user-specific sets - genuine and impostors, and it is quite possible that they might have different dependence models. But the number of genuine scores per enrollee or per user is usually limited (only one genuine score in our dataset BSSR1), and it might not be possible to learn the dependency model for genuine scores. Therefore, we will assume that the same model is applied for both types of scores; the sample estimates $\hat{\mu_i}$ and $\hat{\sigma_i}$ can be computed using both genuine and impostor samples, but in this work we use only impostor score samples.

If the genuine and impostor scores for each enrollee or user are the result of linear score dependency model and

Matchers	$\hat{\mu}$	$\hat{\sigma}$	OS_1	$OS_{N/4}$	$OS_{N/2}$
C	0.2300	0.1798	0.2169	0.2310	0.2210
G	0.2860	0.2983	0.2453	0.2335	0.2373
li	0.2917	0.3303	0.2904	0.2915	0.2496
ri	0.3423	0.3639	0.3349	0.3378	0.3100

TABLE I
CORRELATIONS BETWEEN GENUINE SCORE AND THE STATISTICS OF ENROLLEE-SPECIFIC IMPOSTOR SETS.

Matchers	$\hat{\mu}$	$\hat{\sigma}$	OS_1	$OS_{N/4}$	$OS_{N/2}$
C	-	-	0.0427	-0.0556	0.0223
G	-	-	0.4949	-0.2352	0.2727
li	-	-	0.0336	-0.0121	-0.0330
ri	-	-	0.0458	-0.0107	-0.0547

TABLE II
CORRELATIONS BETWEEN GENUINE SCORE AND THE STATISTICS OF ENROLLEE-SPECIFIC IMPOSTOR SETS AFTER Z-NORMALIZATION WAS PERFORMED.

Matchers	$\hat{\mu}$	$\hat{\sigma}$	OS_1	$OS_{N/4}$	$OS_{N/2}$
C	0.1439	0.1263	0.1485	0.1457	0.1352
G	0.1587	-0.0352	0.1569	0.1565	0.1496
li	0.2857	0.3227	0.2901	0.2747	0.2506
ri	0.3541	0.3691	0.3269	0.3487	0.3191

TABLE III
CORRELATIONS BETWEEN GENUINE SCORE AND THE STATISTICS OF USER-SPECIFIC IMPOSTOR SETS.

Matchers	$\hat{\mu}$	$\hat{\sigma}$	OS_1	$OS_{N/4}$	$OS_{N/2}$
C	-	-	0.0417	-0.0369	0.0228
G	-	-	0.0805	0.1664	0.2052
li	-	-	0.0315	-0.0524	0.0004
ri	-	-	0.0104	-0.0105	-0.0310

TABLE IV
CORRELATIONS BETWEEN GENUINE SCORE AND THE STATISTICS OF USER-SPECIFIC IMPOSTOR SETS AFTER T-NORMALIZATION WAS PERFORMED.

have distributions

$$\begin{aligned} p_{gen,i}(s) = p_{gen,\mu_i,\sigma_i}(s) = \frac{1}{\sigma_i} p_{gen}(\frac{s-\mu_i}{\sigma_i}) \\ p_{imp,i}(s) = p_{imp,\mu_i,\sigma_i}(s) = \frac{1}{\sigma_i} p_{imp}(\frac{s-\mu_i}{\sigma_i}) \end{aligned} \quad (4)$$

then after the corresponding Z- or T-normalization genuine and impostor scores will be independently and identically distributed according to $p_{gen}(x)$ and $p_{imp}(x)$. It is not necessary that linear score dependence model exactly describes the dependencies of scores in identification trials and the actual dependencies might be more complex. In order to check whether the linear dependence models adequately describe existing dependences in matching scores, we can perform statistical tests on score dependence after Z- or T-normalization.

B. Score Dependencies in BSSR1 Set

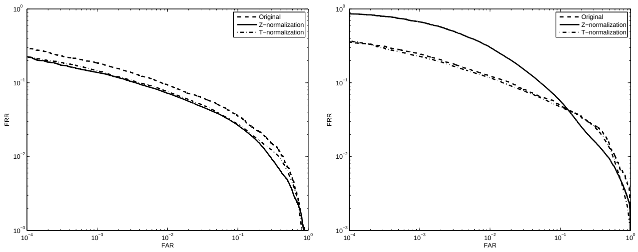
We conducted a series of experiments trying to describe the existing score dependencies in BSSR1 set. The utilized technique was to measure the correlations between some variables in either enrollee-specific or user-specific score sets. In particular, we were interested in the correlations between genuine and impostor scores. If we knew how the genuine score relates to impostor scores for the same enrollee or the same user, we would know how to properly assign the confidence of a match.

Instead of using a genuine and a single random impostor score for correlations, we used a genuine score and some statistics of the corresponding impostor score sets for calculating correlations. Tables I- IV show the results of these experiments. The particular statistics employed in our experiments are: $\hat{\mu}$ and $\hat{\sigma}$ - sample mean and variance, and order statistics - best impostor score OS_1 , N/4-th best impostor score $OS_{N/4}$ and score median $OS_{N/2}$.

Tables I and III show correlations for original matching scores. We can notice that generally all matchers exhibit relatively high dependencies. This implies that accounting for these dependencies might benefit any system utilizing these matchers. The interesting feature about these tables is the difference in the correlation coefficients for face matchers. Whereas statistics of impostor scores give almost the same correlations with genuine scores for matchers 'li' and 'ri' irrespective of whether enrollee or user related scores were used, the correlations for face matchers are higher when the statistics are calculated from enrollee-specific scores than the statistics from user-specific sets. This implies that either the calculation of matching scores for face matchers 'C' and 'G' is non-symmetric with regards to enrolled and test templates, or that some type of enrollee- or user-specific score normalization was already performed.

In order to see whether the linear score dependence model (Equation 3) holds for considered biometric matchers, we looked at the same correlations after corresponding Z- or T-normalization has been performed. Tables II and IV show the results of these experiments. If linear score dependence model was true, then correspondingly normalized scores would have been independent, and the correlations between genuine and the statistics of the impostor sets would have been near 0. The correlations are indeed near 0 for all matchers except 'G' for enrollee-specific score sets. This means that linear score dependence model is not sufficient in this case, and in order to properly construct enrollee-specific methods for matcher 'G' we need to use something different from Z-normalization.

In order to investigate this further, we compared the effects of Z-normalization and T-normalization on all our matchers. Both normalization algorithm resulted in identical performance improvement for both fingerprint match score sets 'li' and 'ri' (we omitted the performance graph). This again confirms that fingerprint matcher used in producing these sets gives symmetric recognition results with respect to enrolled and test fingerprints. But face matchers showed



(a) C (b) G

Fig. 2. ROC curves for verification decisions of single matchers utilizing enrollee-specific models.

different effects from Z- and T-normalizations, and Figure 2 presents the results. For 'C' matcher both normalizations were successful with Z-normalization showing little better results. For 'G' matcher T-normalization gave slightly better results than the use of original scores, and Z-normalization was showing mostly the decrease in performance.

C. Statistics of Impostor Score Sets

Instead of applying pre-defined normalization formula in Z- and T-normalizations, we can try to simply pass the involved sample statistics $\boldsymbol{\theta}_{enr,i}$ or $\boldsymbol{\theta}_{user,j}$ to the algorithm utilizing the matching score ($\boldsymbol{\theta}$ will consist of $\hat{\mu}$ and $\hat{\sigma}$ for these normalizations). In this case the enrollee- or user-specific normalization is done implicitly by the external algorithm. Generally, this will make the task of training external algorithm more difficult, and our preliminary results showed that increased training complexity almost negates the positive effect from adding statistics in case of Z- and T-normalizations.

But by using adequate statistics we might be able to get the good results. In [14] we investigated the use of second best score for matcher combinations in verification systems, where the second best score is defined as the best score among all matching scores related to a particular user (the scores the same identification trial) besides currently considered one. Most experiments showed that using second best score resulted in better combination than combination operated on original scores. Moreover, the performance improvements from T-normalization and second best model complemented each other, and using both led to superior performance. In this paper, we are trying to perform similar experiments with Z-normalization and second best score statistics derived from the set of enrollee-specific scores.

Note, that when we obtain a set of user-specific matching scores, we do not have knowledge whether some particular scores are genuine or impostor. But we have a definite knowledge (assuming closed set identification system) that there is only one genuine score and all other scores are impostors. The second best score model effectively utilizes this knowledge: for a currently considered score if second best score (best score besides current) is very high, this high score is likely to be genuine and currently considered score is therefore impostor. Reverse is also true: if second best score is low relative to current, then this score is very likely to be genuine. The second best score statistics works for user-specific models even if the scores in user-specific sets are independent, see [16] for detailed analysis.

The situation is different for second best score statistics computed in enrollee-specific sets. Though we might have a knowledge on whether the previously collected matching scores for a particular enrollee i are impostor or genuine, this does not give us any information about whether the current score is genuine or impostor. For example, the previously collected set of scores might already include some genuine score, but we can not use this information and say that it is definite that current score is impostor; in real system any enrolled person might be authenticated arbitrary number of times (including zero). In order to simulate the work of a real system, we actually disregard the only genuine score available for each enrollee from calculating the statistics of enrollee specific scores. Thus, instead of defining second best score statistics as the best score besides current one which we used in user-specific score sets, for enrollee specific score sets we define second best score statistics $sbs_{i,j}$ as the *best impostor* besides current score $s_{i,j}$. Note, that calculation of $sbs_{i,j}$ in user specific score sets did include genuine score, and this is a major difference between two statistics.

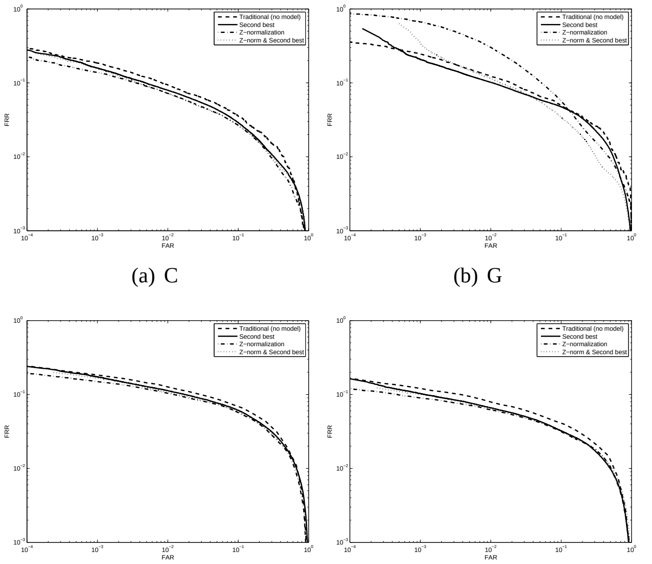
V. EXPERIMENTS IN ACCEPTANCE DECISIONS OF VERIFICATION SYSTEMS

In the experiments of these section we used enrollee-specific normalization and score statistics for enhancing the algorithm of accepting decisions in verification systems. The results of the experiments are shown in Figure 3. The traditional approach is to simply compare matching score to the threshold to accept or reject the verification attempt. The Z-normalization approach also use thresholds the scores, but the scores are Z-normalized. In order to use $sbs_{i,j}$ model we perform the following score transformation [16]:

$$s_{i,j} \rightarrow \frac{p_{gen}(s_{i,j}, sbs_{i,j})}{p_{imp}(s_{i,j}, sbs_{i,j})} \tag{5}$$

As we explained in [16], by treating the pair $(s_{i,j}, sbs_{i,j})$ as an input to the classification algorithm, Equation 5 delivers optimal Bayesian classification decision. Therefore, if $sbs_{i,j}$ carries any useful information complementary to $s_{i,j}$ and the learning of densities p_{gen} and p_{imp} has sufficiently small errors, we can expect the improvement to the decision algorithm relying exclusively on $s_{i,j}$. As Figure 3 shows, the performance improves from using $sbs_{i,j}$ compared to original scores in all cases.

We have also tried to use both models: first perform Z-normalization, and then use second best score model of Equation 5. The observed performance was generally worse than the performance of a single Z-normalization except for matcher 'G' where Z-normalization fails. The combination of both models somewhat improves the performance of original scores of matcher 'G', but it is not as good as using second best score model alone for smaller FAR values. The absence of additional performance improvement is confirmed by the score dependency analysis of section IV-B: for all



(a) C (b) G

(c) li (d) ri

Fig. 3. ROC curves for verification decisions of single matchers utilizing enrollee-specific models.

matchers except 'G' the linear model adequately describes score dependence and Z-normalization might be sufficient to significantly reduce it.

VI. Experiments in the Combination of Matchers in Verification Systems

In the experiments of this section we tried to utilize enrollee-specific normalization and score statistics in the combinations of biometric matchers in verification systems. The results of the experiments are shown in Figure 4. The presented experiments resemble those presented in [14], only instead of user-specific combinations we considered enrollee specific combinations here.

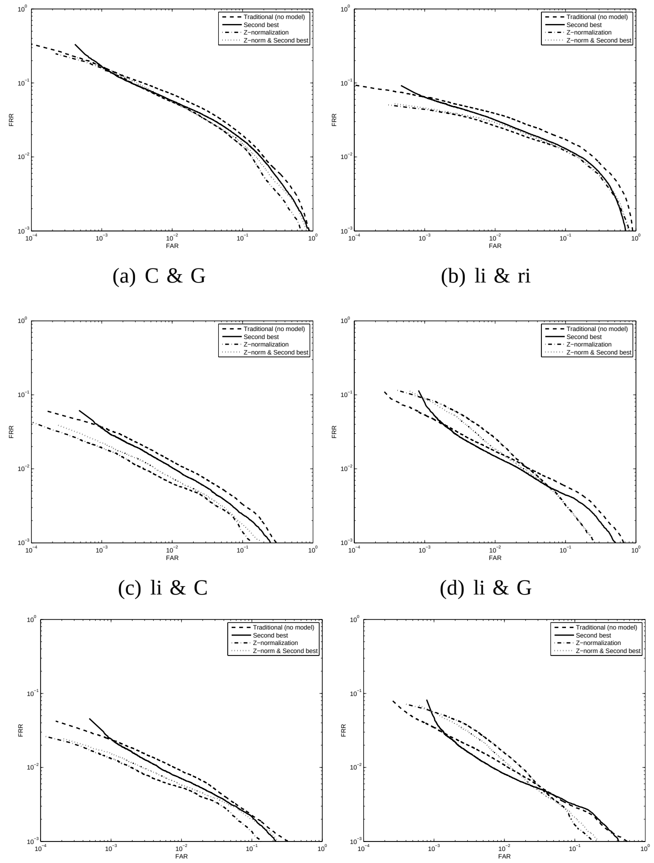
We used the likelihood ratio combination method in order to compare the usefulness of enrollee-specific information. This is theoretically optimal combination method for verification system [5] and consists in assigning a combined score a value of the ratio between genuine and impostor score densities:

$$S = \frac{p_{gen}(s^1, s^2)}{p_{imp}(s^1, s^2)} \tag{6}$$

where s^m is the verification matching score assigned by the matcher m (we omitted indexes i and j from score notation). The likelihood ratio with Z-normalization will operate by the same formula, only using Z-normalized scores s^m. The likelihood ratio method using second best score model will consider the joint densities of scores and second best score statistics:

$$S = \frac{p_{gen}(s^1, sbs(s^1), s^2, sbs(s^2))}{p_{imp}(s^1, sbs(s^1), s^2, sbs(s^2))} \tag{7}$$

The use of Z-normalization and second best score model at the same time implies first Z-normalization of combined



(a) C & G (b) li & ri

(c) li & C (d) li & G

(e) ri & C (f) ri & G

Fig. 4. ROC curves for likelihood ratio combinations utilizing enrollee-specific models.

scores, and then using second best score model likelihood ratio combination of above formula.

The results of combinations experiments somewhat reflect the results we obtained for decision thresholding of single matchers: 1. both Z-normalization and second best score model deliver better performance than original systems with the exception of combinations involving matcher 'G', 2. for matcher 'G' the use of second best score model is beneficial, 3. for all matchers except 'G' the Z-normalization has better results than second best score model. These results are different than the results we obtained in [14], where we considered user-specific combinations - the second best score model is less beneficial here. The difference seems to be due to the different knowledge these statistics use as we explained in section IV-C.

VII. Conclusion

In this paper we investigated the use of enrollee-specific Z-normalization and enrollee-specific second best score statistics in the acceptance decision thresholding and matcher combination in biometric verification systems. We analyzed the difference of these enrollee-specific methods with previously investigated user-specific methods. The major conclusion derived in these paper is that the dependencies of matching scores in enrollee specific score sets and user specific score sets might be different. As a result, the optimal methods utilizing either enrollee-specific information or user-specific information might also differ, even if the same matcher is considered.

In terms of the complexity types of combinations [19] the combinations considered here belong to medium I type. Most of our previous research with regards to matching score dependencies addressed combinations of medium II type. Though some of the previously developed ideas, like using score set statistics, are directly applicable to medium I type combinations, current results shows that the choice of particular statistics might be different in two cases.

The enrollee-specific decision and combination algorithms we presented here are somewhat new for general biometric systems. Though the speaker recognition community extensively uses some techniques (like Z-normalization), most enrollee-specific algorithms developed for general biometric systems require significant number of genuine scores available for training for each enrollee. In our case, we construct enrollee-specific algorithms using only impostor scores, and thus we are able to perform our experiments on NIST BSSR1 dataset.

REFERENCES

[1] "Nist biometric scores set. http://www.nist.gov/biometricscores/."

[2] R. M. Bolle, N. K. Ratha, and S. Pankanti, "Error analysis of pattern recognition systems–the subsets bootstrap," *Computer Vision and Image Understanding*, vol. 93, no. 1, pp. 1–33, 2004.

[3] K.-A. Toh, X. Jiang, and W.-Y. Yau, "Exploiting global and local decisions for multimodal biometrics verification," *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, vol. 52, no. 10, pp. 3059–3072, 2004.

[4] J. Fierrez-Aguilar, D. Garcia-Romero, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Bayesian adaptation for user-dependent multimodal biometric authentication," *Pattern Recognition*, vol. 38, no. 8, pp. 1317–1319, 2005.

[5] S. Prabhakar and A. K. Jain, "Decision-level fusion in fingerprint verification," *Pattern Recognition*, vol. 35, no. 4, pp. 861–874, 2002.

[6] A. Jain and A. Ross, "Learning user-specific parameters in a multibiometric system," in *International Conference on Image Processing. 2002*, vol. 1, 2002, pp. I–57–I–60 vol.1.

[7] R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. Jain, "Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 3, pp. 450–455, 2005.

[8] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42–54, 2000.

[9] J. Mariethoz and S. Bengio, "A unified framework for score normalization techniques applied to text independent speaker verification," *IEEE Signal Processing Letters*, vol. 12, 2005.

[10] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The nist speaker recognition evaluation - overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, no. 2-3, pp. 225–254, 2000.

[11] P. Grother, "Face recognition vendor test 2002 supplemental report, nistir 7083," Tech. Rep., 2004.

[12] A. Rosenberg and S. Parthasarathy, "Speaker background models for connected digit password speaker verification," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1, 1996, pp. 81–84 vol. 1.

[13] J. Colombi, J. Reider, and J. Campbell, "Allowing good impostors to test," in *Signals, Systems & Computers, 1997. Conference Record of the Thirty-First Asilomar Conference on*, vol. 1, 1997, pp. 296–300 vol.1.

[14] S. Tulyakov, Z. Zhang, and V. Govindaraju, "Comparison of combination methods utilizing t-normalization and second best score model," in *CVPR 2008 Workshop on Biometrics*, 2008.

[15] D. Sturim and D. Reynolds, "Speaker adaptive cohort selection for tnorm in text-independent speaker verification," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, vol. 1, 2005, pp. 741–744.

[16] S. Tulyakov and V. Govindaraju, "Combining matching scores in identification model," in *8th International Conference on Document Analysis and Recognition (ICDAR 2005)*, Seoul, Korea, 2005.

[17] G. Aggarwal, N. Ratha, and R. Bolle, "Biometric verification: Looking beyond raw similarity scores," in *Computer Vision and Pattern Recognition Workshop, 2006 Conference on*, 2006, p. 31.

[18] J. Navratil and G. N. Ramaswamy, "The awe and mystery of t-norm," in *8th European Conference on Speech Communication and Technology (EUROSPEECH-2003)*, Geneva, Switzerland, 2003, pp. 2009–2012.

[19] S. Tulyakov and V. Govindaraju, "Classifier combination types for biometric applications," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), Workshop on Biometrics*, New York, USA, 2006.