Predicting Biometric Authentication System Performance Across Different Application Conditions: A Bootstrap Enhanced Parametric Approach

Norman Poh and Josef Kittler

CVSSP, University of Surrey, Guildford, GU2 7XH, Surrey, UK
normanpoh@ieee.org, j.kittler@surrey.ac.uk

Abstract. The performance of a biometric authentication system is dependent on the choice of users and the application scenario represented by the evaluation database. As a result, the system performance under different application scenarios, e.g., from cooperative user to non-cooperative scenario, from well controlled to uncontrolled one, etc, can be very different. The current solution is to build a database containing as many application scenarios as possible for the purpose of the evaluation. We propose an alternative evaluation methodology that can reuse existing databases, hence can potentially reduce the amount of data needed. This methodology relies on a novel technique that projects the distribution of scores from one operating condition to another. We argue that this can be accomplished efficiently only by modeling the genuine user and impostor score distributions for each user parametrically. The parameters of these model-specific class conditional (MSCC) distributions are found by maximum likelihood estimation. The projection from one operating condition to another is modelled by a regression function between the two conditions in the MSCC parameter space. The regression functions are trained from a small set of users and are then applied to a large database. The implication is that one only needs a small set of users with data reflecting both the reference and mismatched conditions. In both conditions, it is required that the two data sets be drawn from a population with similar demographic characteristics. The regression model is used to predict the performance for a large set of users under the mismatched condition.

1 Introduction

The performance of a biometric system involves a biometric database with multiple records of N users. In a typical experimental evaluation, the records of a reference user are compared with the remaining users, hence resulting in impostor match scores. The comparisons among records of the same reference user result in genuine user (or referred to as "client") match scores. By sweeping through all possible decision threshold values given these two sets of class conditional match scores, one obtains the system performance in terms of pairs of false acceptance (FAR) and false rejection (FRR) rates, respectively. Consequently, the measured performance is inevitably database- and protocol-dependent. For instance, running the same matching algorithm on a single database but with two different protocols (or ways of partitioning the training and test

S.-W. Lee and S.Z. Li (Eds.): ICB 2007, LNCS 4642, pp. 625–635, 2007.

data) may result in two correlated but nevertheless slightly different performance measures. A good example is the XM2VTS database and its two Lausanne protocols. As a result, if one algorithm outperforms another, one cannot be certain that the same relative performance is repeatable when a different set of users is involved. The same concern about the preservation of relative performance applies to changes in application scenarios, e.g., from a controlled to uncontrolled ones, and from cooperative to non-cooperative users. These changes will result in *variability* of the measured performance.

The goal of this paper is to propose a statistical model that can capture the above factors of variability such that given a small set of development (training) data where samples acquired during the reference and mismatched conditions are present, one can apply the model to predict the performance on an evaluation (test) data set composed of samples collected in the mismatched condition only. We use the term *mismatched* condition as the one that is significantly different *relative* to the reference one. In this work, we assume that the population of users in both the development and the evaluation sets are drawn from a *common* demographic population. This itself is a difficult problem because given this common demographic population, one still has to deal with the other three sources of variability.

The novelty of this study is to propose an evaluation methodology as well as a statistical model developed for that purpose to avoid collecting more data (although more data is always better) but instead to put more emphasis on *predicting the performance* under various operating conditions based only on the development data set. The proposed methodology can thus significantly reduce the cost of data collection.

This contrasts with the conventional evaluation methodology which is inherently demanding in terms of the amount of data required, e.g., the BANCA [1] and FERET [2] evaluations. Furthermore, given limited resources in data collection, one has to trade-off the amount of data (users) against the number of application scenarios available.

Preliminary experiments on 195 experiments carried out on the BANCA database show that predicting the performance from the reference to the mismatched condition is possible even when the performance associated with these two conditions were assessed on two *separate* populations of users.

2 Towards Generalising Biometric Performance

In order to model the change of score distribution, we will need a statistical model that can effectively model the score distribution *within* a given operating condition. We opt to build a model known as the *model-specific class conditional* (MSCC) distributions. If there are J users, there will be $2 \times J$ MSCC distributions since each user has two class conditional scores, one reflecting genuine match scores and the other impostor match scores. An MSCC distribution represents a statistical summary of class-conditional scores specific to each user. In essence, it captures the sample variability but conditioned on the class label. Then, in order to capture the change from a reference operating condition to a mismatched one, we need a *projection function* that can map a set of MSCC distributions to another set. One reasonable assumption that is used here is that the effect of the mismatch, as observed given only the scores, is the *same*

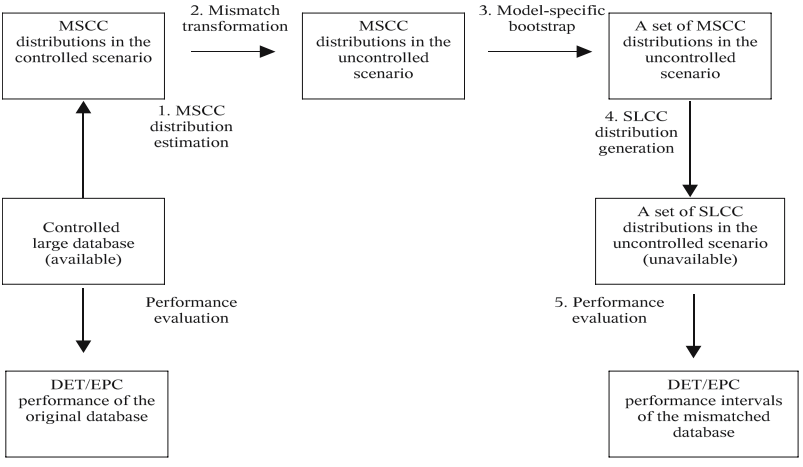


Fig. 1. A procedure to predict performance as well as its associated confidence intervals based on MFCC distributions under mismatched application scenario and user composition. Since the MSCC parameters under mismatched situation can only be estimated probabilistically, we propose a slightly different version of model-specific bootstrap which also referred to as the bootstrap subset technique [3].

across all the user models of the same group, i.e., the users are drawn from the common demographic population. (e.g., bankers or construction workers, but not both).

Figure 1 outlines an algorithm to train and infer the proposed statistical model. In Step 1, the parameters of the MSCC distributions are estimated via the maximum likelihood principle. In Step 2, the projection function between the set of MSCC distributions corresponding to the reference operating condition and the mismatched one is modeled via regression. The projection function has to be learnt from the smaller data set containing biometric samples recorded in both types of conditions. The transformation is then applied to a larger data set recorded in the reference operating condition. Thanks to this projection, the MSCC distributions of the mismatched condition becomes available, while, avoiding the need of actually collecting the same amount of data set for the mismatched condition.

In Step 3, we derive the confidence interval around the predicted performance. The state-of-the-art technique to do so in biometric experiments is known as the *bootstrap subset* technique [3]. This technique is different from the conventional bootstrap because it does not draw the score samples directly but draws the user models acquired for the database. In our context, this technique draws with replacement the user models associated with *a pair of* MSCC distributions in each round of the bootstrapping process. The bootstrap subset technique assumes that the parameters of the MSCC distributions are known. We propose a Bayesian approach which defines a distribution over each of the MSCC parameters. In order to estimate confidence intervals around the

predicted performance we propose to sample from this distribution J different sets of MSCC parameters (one for each user) in each round of bootstrap. As will be shown, the proposed Bayesian approach gives systematically better estimate of confidence interval than the bootstrap subset technique.

Step 4 attempts to derive the system-level class-conditional (SLCC) score distributions from the set of MSCC distributions. Finally, Step 5 visualises the performance in terms of conventional plots, e.g., receiver's operating characteristic (ROC) and detection error trade-off (DET).

The following sections present an algorithm to implement the proposed evaluation methodology shown in Figure 1.

2.1 Model Specific Class Conditional Score Distribution

Let the distribution of model-specific class-conditional (MSCC) scores be written as $p(y|k,j,m)$ where $y \in \mathbb{R}$ is the output of a biometric system, j is the user index and $j \in [1,\ldots,J]$ and m is the m-th model/template belonging to user j. Without loss of generality, $p(y|k)$ can be described by:

$$p(y|k) = \sum_j \underbrace{\sum_m p(y|k,j,m)P(m|j,k)}\, P(j|k) \tag{1}$$

$$p(y|k) = \sum_j p(y|k,j)P(j|k) \tag{2}$$

where $p(y|k,j,m)$ is the density of y conditioned on the user index j, true class label k and the m-th model (a discrete variable) specific to user j; $P(m|j,k)$ specifies how probable it is that m is used (when $k = C$) or abused (when $k = I$); and, $P(j|k)$ specifies how probable it is that user j uses the system or persons impersonating him/her abuses his identity. All the data sets we deal with have only one model/template per user, i.e., $m = 1$. As a result, (1) and (2) are identical. The false acceptance rate (FAR) and false rejection rate (FRR) are defined as a function of the global decision threshold $\Delta \in [-\infty,\infty]$ in the following ways:

$$\text{FAR}(\Delta) = 1 - \Psi_I(\Delta) \tag{3}$$

$$\text{FRR}(\Delta) = \Psi_C(\Delta), \tag{4}$$

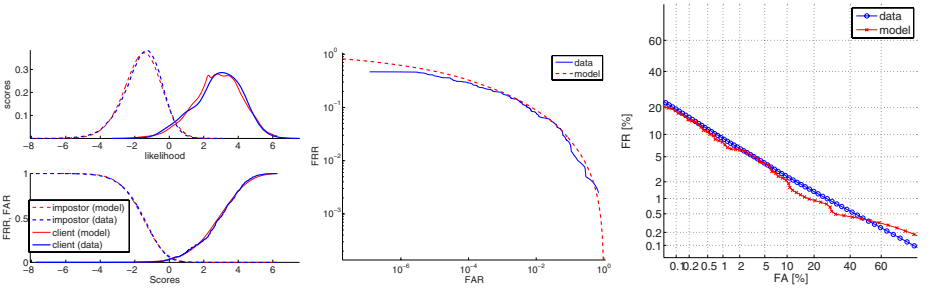
where Ψ_k is the cumulative density function (*cdf*) of $p(y|k)$, i.e.,

$$\Psi_k(\Delta) = p(y \le \Delta|k) = \int_{-\infty}^{\Delta} p(y|k)dy \tag{5}$$

In this study, we assume that $p(y|k,j)$ (or more precisely $p(y|k,j,m=1)$) is Gaussian, i.e.,

$$p(y|k,j) = \mathcal{N}\big(\mu_j^k, (\sigma_j^k)^2\big), \tag{6}$$

where, μ_j^k and $(\sigma_j^k)^2$ are respectively the mean and the variance of the underlying scores. When the true score distribution is known, it should be used. However, as is often the case, due to the paucity of user-specific data, especially the genuine user match



(a) *pdf* and *cdf* (b) ROC (c) DET

Fig. 2. (a) Comparison between the *pdf*'s (top) and *cdf*'s (bottom) estimated from the model and the data; the same comparison when visualised using (b) ROC in log-log scales and (c) DET (in normal deviate scales). In order to estimate the *pdf*'s in the top figure of (a), the kernel density (Parzen window) method was used with a Gaussian kernel; the *cdf*'s of the data used for visualising the ROC curve are based on (5).

Table 1. The partition of the data sets. The first row and first column should read "the score set $\mathcal{Y}^C_{small}$ is available".

Data set size	reference		degraded	
	Genuine (C)	impostor (I)	Genuine (C)	impostor (I)
small	available	available	available	available
large	available	available	not available	not available

scores, one may not be able to estimate the parameters of the distribution reliably. As a result, a practical solution may be to approximate the true distribution with a simpler one, e.g., using only the first two order of moments as represented by a Gaussian distribution.

Figure 2 compares the class-conditional score distribution estimated from the data with the one estimated from the model. Referring back to Figure 1, Step 1 is an application of (1); Step 4 of (5) and Step 5 of (3) and (4), respectively.

2.2 Estimation of Score Distribution Projection Function

In this section, we develop the projection function from the score distribution of the reference operating condition to a mismatched one. Suppose that for a small number of user models, we have access to both their reference and degraded class-conditional scores, i.e., $\{\mathcal{Y}^k_{small}|Q\}$ for $Q \in \{\text{ref}, \text{deg}\}$ (for reference and degraded, respectively) and $k \in \{C, I\}$. For another set with a much larger number of user models, we have only access to their data captured in the reference condition, $\{\mathcal{Y}^k_{large}|Q = \text{ref}, \forall_k\}$ and we wish to predict the density of the degraded scores $\{\mathcal{Y}^k_{large}|Q = \text{deg}, \forall_k\}$ which we do not have access to. Table 1 summarises the availability of scores data. In our experimental setting, $\{\mathcal{Y}^k_{large}|Q = \text{deg}, \forall_k\}$ serves as the ground-truth (or *test*) data whereas the other six data sets are considered as the *training* data in the usual sense.

Let y be the variable representing a score in $\mathcal{Y}_s^k$ where $k \in \{C, I\}$ and $s \in \{small, large\}$. The *pdf* $p(y|k)$ estimated from $\mathcal{Y}_s^k$ as given by (2) is a function of $p(y|j,k) \equiv \mathcal{N}\left(\mu_j^k, (\sigma_j^k)^2\right)$ for all j, k. Therefore, $p(y|k)$ (for a given data set s) can be fully represented by keeping the parameters $\{\mu_j^k, \sigma_j^k | \forall_j, s\}$ estimated from $\mathcal{Y}_s^k$.

Our goal is to learn the following projection

$$f : \{\mu_j^k, \sigma_j^k | \forall_j, Q = \text{ref, s}\} \rightarrow \{\mu_j^k, \sigma_j^k | \forall_j, Q = \text{deg}, s\} \tag{7}$$

separately for each k from the small data set ($s = small$) and apply it to the large data set ($s = large$). One could have also learnt the above transformation jointly for both $k \in \{C, I\}$ if one assumed that the noise source influenced both types of scores in a similar way. Our preliminary experiments show that this is, in general, not the case. For instance, degraded biometric features affect the magnitude of the genuine scores much more than that of the impostor scores. For this reason, it is sensible to model the transformation functions for each class of access claims separately. This is done in two steps for each of the two parameters separately:

$$f_\mu : \{\mu_j^k | \forall_j, Q = \text{ref}, s\} \rightarrow \{\mu_j^k | \forall_j, Q = \text{deg}, s\} \tag{8}$$

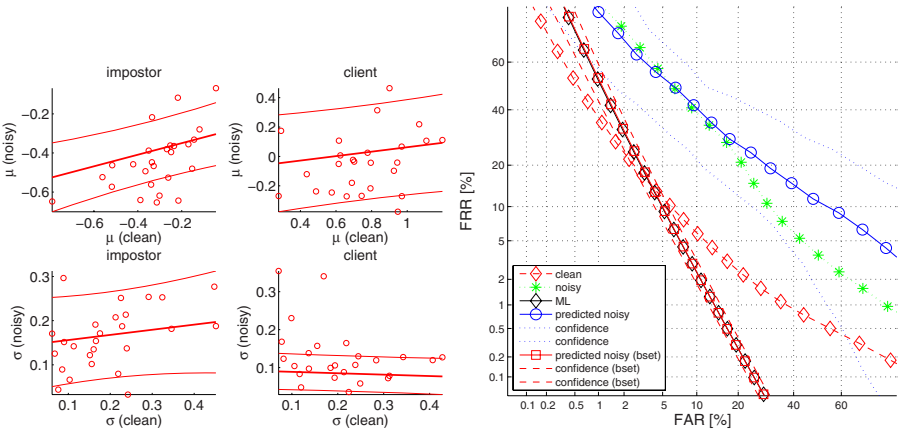
$$f_\sigma : \{\sigma_j^k | \forall_j, Q = \text{ref}, s\} \rightarrow \{\sigma_j^k | \forall_j, Q = \text{deg}, s\} \tag{9}$$

where f_{param} is a polynomial regression function for $param \in \{\mu, \sigma\}$. Note that in (9), the function is defined on the standard deviation and not on the variance because noise is likely to be amplified in the latter case. This observation was supported by our preliminary experiments (not shown here). We have also considered modeling the joint density of $\{\mu_j^k, \sigma_j^k | \forall_j\}$ and found that their correlation is extremely weak across the 195 experiments taken from the BANCA database (to be described in Section 3), i.e., on average -0.4 for the impostor class and 0 for the genuine user (client) class. This means that the two Gaussian parameters are unlikely to depend on each other and there is no additional advantage to model them jointly. These two regression functions give us the predicted degraded Gaussian parameters – $\hat{\mu}_j^k$ and $\hat{\sigma}_j^k$ – given the Gaussian parameters of the reference condition–μ_j^k and σ_j^k.

The polynomial coefficients are obtained by minimising the mean squared error between the predicted and the true values of the larger data set given the values of the smaller data set. We used Matlab's `polyfit` function for this purpose. As a by-product, the function also provides a 95% confidence around the predicted mean value, which corresponds to the variance of the prediction, i.e., $Var[f_\mu(\mu_j^k)]$ for the mean parameter and $Var[f_\sigma(\sigma_j^k)]$ for the standard deviation parameter. These two by-products will be used in Section 2.3.

In our preliminary experiments, the degree of polynomial was tuned using a two-fold cross validation procedure. However, due to the small number of samples (in fact the number of users) used, which is 26, using a quadratic function or even of higher order does not necessarily generalise better than a linear function. Following the Occam's Razor principle, we used only a linear function for (8) and (9), respectively.

Examples of fitted regression functions for each f_μ and f_σ conditioned on each class k are shown in Figure 3(a). The four regression functions aim collectively to predict a DET curve in degraded condition given the MSCC parameters of the reference condition. Figure 3(b) shows both the reference and predicted degraded DET curves. Two



(a) Regression on MSCC parameters (b) DET

Fig. 3. One of 195 BANCA experiments showing the regression fits to project the MSCC parameters under the reference condition to those under the degraded conditions. Note that there is also user composition mismatch because the reference and degraded data come from two disjoint sets of users (the g1 and g2 sets according to the BANCA protocols). Four regression fits are shown in (a) for each of the client and impostor classes and for each of the two Gaussian parameters (mean and standard deviation). The regression lines together with their respective 95% confidence intervals were estimated on the development set whereas the data points (each obtained from a user model) were obtained from the evaluation set. Figure (b) shows the DET curves (plotted with '$\circ$') along with its upper and lower confidence bounds (dotted lines) as compared to the original reference ('$\diamondsuit$') and degraded ('$*$') DET curves. The 95% confidence intervals around the predicted degraded DET curves were estimated by 100 bootstraps as described in Section 2.3 (not the bootstrap subset technique). In each bootstrap (which aims to produce a DET curve), we sampled the predicted MSCC parameters given the reference MSCC parameters from the four regression models assuming Gaussian distribution.

procedures were used to derive the predicted degraded DET curves, thus resulting in two predicted degraded curves as well as their corresponding confidence intervals. They will be described in Section 2.3.

2.3 Bootstraps Under Probabilistic MSCC Parameters: A Bayesian Approach

This section deals with the case where the MSCC parameters can only be estimated probabilistically, e.g., when applying (7) which attempts to project from the reference MSCC parameters to the degraded ones, the exact degraded MSCC parameters are unknown. This is in contrast to the bootstrap subset technique which assumes that the MSCC parameters can be estimated deterministically (which corresponds to the usual maximum likelihood solution).

Let $\hat{\mu}_j^k = f_\mu(\mu_j^k)$ be the predicted mean parameter given the reference mean parameter μ_j^k via the regression function f_μ. The predicted standard deviation is defined similarly, i.e., $\hat{\sigma}_j^k = f_\sigma(\sigma_j^k)$. We assume that the predicted mean value is normally distributed, i.e.,

$$p(\hat{\mu}_j^k) = \mathcal{N}(f_\mu(\mu_j^k), Var[f_\mu(\mu_j^k)]), \tag{10}$$

and so is the predicted standard deviation value, i.e.,

$$p(\hat{\sigma}_j^k) = \mathcal{N}(f_\sigma(\sigma_j^k), Var[f_\sigma(\sigma_j^k)]). \tag{11}$$

Both distributions $p(\hat{\mu}_j^k)$ and $p(\hat{\sigma}_j^k)$ effectively characterise the variability due to projecting the parameters of the reference MSCC distribution as represented by $\{\mu_j^k, \sigma_j^k\}$ to the degraded ones $\{\hat{\mu}_j^k, \hat{\sigma}_j^k\}$, which one cannot estimate accurately.

In order to take into account the uncertainty associated with the predicted Gaussian parameters, we propose to sample from the distributions $p(\hat{\mu}_j^k)$ and $p(\hat{\sigma}_j^k)$ for all j and both classes k, thus obtaining a set of MSCC parameters $\{{v_\mu}_j^k, {v_\sigma}_j^k | \forall_j, \forall_k\}$ from which we can evaluate a DET curve thanks to the application of (1), (3) and (5) in each round of bootstraps. By repeating this process U times, one obtains U bootstrapped DET curves.

We give a brief account here how to obtain the expected DET curve and its corresponding 95% confidence intervals given the U bootstrapped DET curves. This technique was described in [4]. First, we project a DET curve into polar coordinates (r, θ), i.e., radius and DET angle, such that $\theta = 0$ degree is parallel to FRR=0 in normal deviate scale and $\theta = 90$ degree is parallel to FAR=0. To obtain $\alpha \times 100\%$ confidence, given the set of bootstrapped DET curves in polar coordinates, we estimate the upper and lower bounds:

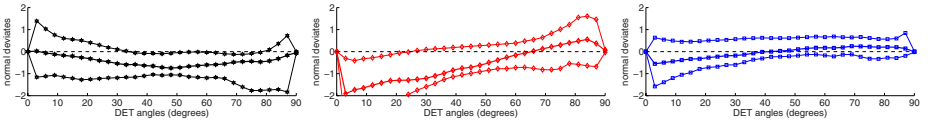
$$\frac{1-\alpha}{2} \leq \Psi_\theta(r) \leq \frac{1+\alpha}{2},$$

where $\Psi_\theta(r)$ is the empirical *cdf* of the radius r observed from the U bootstrapped curves for a given θ. Note that each bootstrapped curve cuts through θ exactly once. The lower, upper and median DET curves are given by setting r to be $\frac{1-\alpha}{2}$, $\frac{1+\alpha}{2}$, and $\frac{1}{2}$, respectively. By back-projecting these three curves from the polar coordinates to the DET plane, one obtains the expected DET curve as well as its associated confidence region at the desired $\alpha \times 100\%$ level of confidence. This technique was described in [4].

A large U is necessary in order to guarantee stable results. Our preliminary experiments show that $U > 40$ is fine. We used $U = 100$ throughout the experiments. Our initial experiments show that the most probable DET curve derived this way generalises much better compared to plotting a DET directly from the predicted MSCC parameters, i.e., $\{\hat{\mu}_j^k, \hat{\sigma}_j^k | \forall_{j,k}\}$ because by using these parameters, one does not take the uncertainty of the prediction into consideration. Our preliminary experiments, as shown for instance in Figure 3(b), suggest that in order to generalise to a different user population and under a mismatched situation, it is better to sample from (10) and (11) for each user and for each class k in each round of bootstraps than to use the bootstrap subset procedure with the predicted MSCC parameters.

3 Database, Performance Evaluation and Results

We chose the BANCA database for our experiments for the following reasons:



(a) Before correction (b) After correction, (bootstrap subset) (c) After correction, (Bayesian bootstrap)

Fig. 4. Comparison of DET bias due to noise and user composition mismatches on 195 BANCA systems under degraded (Ud) conditions as compared to the reference (Mc) conditions. Each of the three figures here presents the distribution of 195 DET radius bias for all $\theta \in [0, 90]$ degrees. The upper and lower confidence intervals represent 90% of the data. Prior to bias correction, in (a), the DET bias is large and is systematic (none zero bias estimate); in (b), after bias correction using the predicted MSCC parameters, i.e., following the bootstrap subset technique, the DET radius bias is still large and non-zero; in (c), after bias correction using the Bayesian bootstrap approach, the bias is significantly reduced and is close to zero – indicating the effectiveness of the regression functions in projecting the MSCC parameters from the reference to both the degraded (Ud) conditions. An example of the actual DET curve was shown in Figure 3(b). Similar results were obtained when the experiments were repeated to predict the performance on the adverse (Ua) conditions instead of the degraded (Ud) conditions (not shown here).

(a) **Availability of mismatched conditions:** It has three application scenarios: controlled, degraded and adverse operating conditions.
(b) **Different sets of user:** It comes with a defined set of protocols that has two partitions of gender-balanced users, called g1 and g2. This allows us to benchmark the quality of performance prediction by training the projection function on g1 and testing it on g2. In each data set, there are only 26 users. 3 genuine scores are available per user; and 4 for the impostor scores to estimate $p(y|j, k, m = 1)$.
(c) **Availability of many systems:** Being a benchmark database, 195 face and speech verification systems have been assessed on this database. These systems were obtained from
"ftp://ftp.idiap.ch/pub/bengio/banca/banca_scores" as well as from [5].

In order to assess the quality of predicted DET curve, we consider three DET curves: two derived experimentally using the reference and the degraded data, and one based on the MSCC models (i.e., the predicted degraded curve from the reference data). These three curves can be expressed by $r_u^{ref}(\theta)$, $r_u^{deg}(\theta)$ and $r_u^{pdeg}(\theta)$ ($pdeg$ for predicted degraded curve), respectively, in polar coordinates for convenience with $\theta \in [0, \frac{\pi}{2}]$. In order to quantify the merit of the proposed approach, we define the bias of the reference and the predicted degraded curves, with respect to the ground-truth degraded DET curve as follows:

$$\text{bias}_u^{ref}(\theta) = r_u^{ref}(\theta) - r_u^{deg}(\theta) \tag{12}$$

and

$$\text{bias}_u^{pdeg}(\theta) = r_u^{pdeg}(\theta) - r_u^{deg}(\theta), \tag{13}$$

where u is one of the 195 data sets. By performing $U = 195$ independent experiments, we can then estimate the density $p(\text{bias}_u^{data}(\theta))$ for each of the θ values for $data \in \{ref, pdeg\}$ condition) separately. We expect that the expected bias due to the

predicted degraded DET curve, $E_u[\text{bias}_u^{pdeg}(\theta))]$ be around zero and to have small confidence intervals whereas $E_u[\text{bias}_u^{ref}(\theta))]$ to be further away from zero and has comparatively larger confidence intervals. It is desirable to have positive bias for the predicted degraded DET curve, i.e., $E_u[\text{bias}_u^{pdeg}(\theta))]$, because systematically overestimating an error is better than underestimating it. The experimental results for predicting from the reference to the degraded operating condition, summarised over 195 BANCA systems, are shown in Figure 4.

4 Conclusions

While prior work has been reported on the effect of the sample and user variability, e.g., [6,3], to the best of our knowledge, none could be used to predict the biometric performance under mismatched conditions. As a result, the conventional methodology in biometric evaluation has always relied on collecting more data and one had to decide on a trade-off between the amount of data and the number of application scenarios available. We propose an evaluation methodology along with an algorithm that does not rely on a large quantity of data. Instead, it attempts to predict the performance under the mismatched condition by adequately modeling the score distribution and then projecting the distribution into one that matches the target mismatched condition for a given target population of users. Generalisation to different population of users (but of the *same* demographic characteristic) can be inferred from the resulting confidence interval. As a result, significantly fewer data is needed for biometric evaluation and existing databases can be reused to predict the performance behaviour of a system.

Our on-going work attempts to remove the Gaussian assumption made regarding the MSCC distribution. One important assumption about the current choice of regression algorithm (polyfit) used as the projection function is that the variance of the error terms is constant. A complete non-parametric modelling of the pair of MSCC parameters would have been more desirable. This issue is also being investigated. Finally, more comprehensive experiments, notably as to how well each predicted DET curve performs, will also be conducted.

Acknowledgment

This work was supported partially by the prospective researcher fellowship PBEL2-114330 of the Swiss National Science Foundation, by the BioSecure project (www.biosecure.info) and by Engineering and Physical Sciences Research Council (EPSRC) Research Grant GR/S46543. This publication only reflects the authors' view.

References

1. Bailly-Baillière, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariéthoz, J., Matas, J., Messer, K., Popovici, V., Porée, F., Ruiz, B., Thiran, J.-P.: The BANCA Database and Evaluation Protocol. In: Kittler, J., Nixon, M.S. (eds.) AVBPA 2003. LNCS, vol. 2688. Springer, Heidelberg (2003)

2. Phillips, P.J., Rauss, P.J., Moon, H., Rizvi, S.: The FERET Evaluation Methodology for Face Recognition Algorithms. IEEE Trans. Pattern Recognition and Machine Intelligence 22(10), 1090–1104 (2000)
3. Bolle, R.M., Ratha, N.K., Pankanti, S.: Error Analysis of Pattern Recognition Systems: the Subsets Bootstrap. Computer Vision and Image Understanding 93(1), 1–33 (2004)
4. Poh, N., Martin, A., Bengio, S.: Performance Generalization in Biometric Authentication Using Joint User-Specific and Sample Bootstraps, IDIAP-RR 60, IDIAP, Martigny. IEEE Trans. Pattern Analysis and Machine Intelligence 2005 (to appear)
5. Cardinaux, F., Sanderson, C., Bengio, S.: User Authentication via Adapted Statistical Models of Face Images. IEEE Trans. on Signal Processing 54(1), 361–373 (2006)
6. Doddington, G., Liggett, W., Martin, A., Przybocki, M., Reynolds, D.: Sheep, Goats, Lambs and Woves: A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation. In: Int'l Conf. Spoken Language Processing (ICSLP), Sydney (1998)