Performance Generalization in Biometric Authentication Using Joint User-Specific and Sample Bootstraps

Norman Poh, Alvin Martin and Samy Bengio

Abstract

Biometric authentication performance is often depicted by a decision error trade-off (DET) curve. We show that this curve is dependent on the choice of samples available, the demographic composition and the number of users specific to a database. We propose a two-step bootstrap procedure to take into account of the three mentioned sources of variability. This is an extension to the Bolle *et al.*'s bootstrap subset technique. Preliminary experiments on the NIST2005 and XM2VTS benchmark databases are encouraging, e.g., the average result across all 24 systems evaluated on NIST2005 indicates that one can predict, with more than 75% of DET coverage, an unseen DET curve with 8 times more users. Furthermore, our finding suggests that with more data available, the confidence intervals become smaller and hence more useful.

Index Terms

Biometric authentication assessment, DET, ROC, bootstrap subset

N. Poh and S. Bengio are with IDIAP Research Institute, Rue du Simplon 4, 1920 Martigny, Switzerland. E-mail: {norman, bengio}@idiap.ch Phone: +41.27.721.{7753,7739} Fax: +41.27.721.77.12; and A. Martin is with NIST, 100 Bureau Drive, Gaithersburg, MD 20899 USA. E-mail: alvin.martin@nist.gov Phone: +1.301.975.3169 Fax: +1.301.670.0939

I. Introduction

Biometric authentication is a process of verifying an identity claim using a person's behavioral and physiological characteristics. There are several factors that can affect a biometric system's performance. Some of these factors are the deformable nature of biometric traits, corruption by environmental noise, variability of biometric traits over time, the state of users (especially behavioral biometrics) and occlusion by the user's accessories. As a consequence, even if two biometric samples are acquired from the same user, the system cannot produce *exactly* the same output score. Therefore, when assessing the performance, the uncertainty introduced by these numerous and often uncontrolled distortions has to be taken into account.

A biometric authentication system can make two types of error, i.e., falsely rejecting a genuine user (client) or falsely accepting an impostor. The respective error rates are called false acceptance rate (FAR) and false rejection rate (FRR)[1]. These two measures are fundamental building blocks to many visualizing tools. The most commonly used ones are receivers' operating cost (ROC) and decision error trade-off (DET) curves [2].

The goal of this paper is to establish a confidence interval around a DET curve by explicitly considering the *correlation structure* of match scores, i.e., the fact that match scores resulting from multiple attempts of a person making the same identity claim are correlated, regardless of whether the person is a client or an impostor. Confidence interval estimation techniques developed in the medical field, e..g, [3] and in machine learning, e.g., [4], [5], cannot be used in biometric authentication because the correlation structure is person-dependent. In [6], a bootstrap algorithm that exploits this person-dependent correlation structure was proposed to estimate the confidence of FAR given an FRR of interest, or the confidence of

[1] In the fingerprint and face communities, FAR is known as *false match rate* whereas FRR is *false non-match rate* [1, Chap. 6 pg. 50]. Furthermore, client accesses are considered *match* (or mated-pair) accesses and impostor accesses are *non-match* (or non-mated pair) accesses. In the speaker verification community (most represented in the NIST evaluation), FAR is known as *false alarm rate* and FRR is *miss detection rate* [1, Chap. 8 pg. 259]. Furthermore, clients are called *target* whereas impostors are sometimes referred to as *non-targets*. There exists two other error types where a system fails to give any output. They are failure-to-capture (FTC) and failure-to-enroll (FTE). These errors are not considered in this paper because we are concerned with algorithmic evaluation and not *operational* evaluation. Our choice is not a weakness because it is possible to modify FRR to take into account of FTC and FTE.

FRR given an FAR of interest. This algorithm was called “bootstrap subset” because it considers only a subset of scores associated to a claimed identity. For clarity, we also refer to this bootstrap as a *user-specific* bootstrap to distinguish it from the conventional *sample* bootstrap which does not take the claimed identity associated to each score into consideration. The bootstrap subset algorithm is better because it does not systematically *underestimate* the confidence interval as would any conventional parametric or non-parametric algorithm.

In this paper, we propose another bootstrap-based algorithm that can be seen as an improvement of the bootstrap subset in the following way:

- **Joint FAR-FRR estimate of confidence interval:** Instead of the *point-wise* estimation of confidence interval, i.e., fix FAR and then estimate the confidence interval of FRR, and vice-versa, we jointly estimate the FAR-FRR confidence interval of the *whole* DET curve.
- **Consideration of the effect of sample variability:** While the person-dependent variability was considered in [6], the sample variability, i.e., the choice of samples (given that the population of users are fixed), was not considered.

Our goal of estimating the confidence of the whole DET curve is so that one can predict a future DET curve that is different from the one available in the following ways:

1) A different sample set
2) A larger population size
3) A *completely different* population of users

Ideally, this future DET curve should be completely *covered* by the estimated confidence bound and the confidence bound should be as *narrow* as possible to be useful. It is conventional to measure the quality of prediction using “coverage”, e.g., [5]. Coverage is the proportion of a future DET curve that is completely covered by the confidence bound estimated from a present DET curve with variation due to one or more of the three factors just mentioned above. Note that in [6], the quality of confidence bound as a performance predictor was not the principal subject of investigation while it is our focus here. In [7], a semi-parametric approach that considers only the first two factors was proposed. The third factor is extremely important following the study in [8], which shows that users in a database exhibit very different behavior with respect to a biometric system. For instance, adding a vulnerable user, also known as a lamb [8], will quickly increase the error rate of a system. Similarly, adding a strong impostor (a wolf) will degrade the system performance. This issue is not the utmost concern when comparing two systems evaluated on the same sets of users, i.e., from the *same* database. It becomes a concern when different users are involved. The latter subject is our focus.

The original contribution of this paper is to propose a two-level bootstrap: a user-specific bootstrap followed by a user-constrained sample bootstrap. We show that the proposed algorithm can predict a *future* DET with reasonable accuracy, i.e., two thirds of coverage in the worst case scenario.

This paper is organized as follows: Section II presents the score data set to be used. Section III describes four available choices of bootstrap algorithm for generating a pool of DET curves. Section IV addresses the issue of defining a confidence region given a pool of DET curves. Section V presents some experimental results. Finally, Section VI contains some conclusions and future works.

II. Datasets

The NIST2005 data set [9] contains 24 verification systems which are all evaluated on a common database with a common protocol. This database contains mismatched training and test conditions. In this study, we only have access to the match scores, the true identity, the claimed identity, the hypothesized type of handset and the hypothesized gender information[2]. Since the current study does not take account of such mismatched conditions, a subset of this data sets was used such that it contains only females using land line handsets. This results in a subset of 124 user models, 11992 impostor accesses and 1172 genuine accesses. Therefore, on average, there are 96 impostor attempts and 9 genuine attempts per user in the evaluation. The 24 verification systems are based on Gaussian mixture models (GMMs), neural network-based classifiers and support vector machines. A few systems are actually combined systems

[2] The last two pieces of information are not available in the database so they are estimated using a gender and handset classifier.

using different levels of speech information. Some systems combine different types of classifier but each classifier use the same feature sets. In accordance with the NIST evaluation plan, the 24 systems are enumerated from 1 to 24 instead of using the actual system name.

III. Towards Establishing Confidence Bound via Bootstrap

In most biometric authentication systems, decisions are made by comparing a score y against a threshold Δ. The decision function is defined as:

$$\text{decision}_\Delta(y) = \begin{cases} \text{accept} & \text{if } y > \Delta \\ \text{reject} & \text{otherwise}. \end{cases} \quad (1)$$

A useful notation is to introduce the score y given the true class label k to which the biometric feature vector belongs, i.e., $y^k \equiv y|k$. Hence, a false acceptance is characterized by "accept $=$ decision$_\Delta(y^I)$" whereas a false rejection is characterized by "reject $=$ decision$_\Delta(y^C)$".

Let $Y^k_{j,m}$ be the variable of the match score y due to the m-th access claim of identity $j \in \{1,\ldots,J\} \equiv \mathcal{J}$ given that the match is due to the class label k (client or impostor) and $m \in \{1,\ldots,M^k_j\} \equiv \mathcal{M}^k_j$, i.e., there are M^k_j accesses in the set $\mathcal{M}^k_j$.

We also introduce another variable, $Y^{k,\prime}_m$ where $m \in \{1,\ldots,M^k\}$. While both $Y^k_{j,m}$ and $Y^{k,\prime}_m$ are two ways of specifying the same score data, the difference between them is that the former takes into consideration of the user index whereas the latter does not. Therefore, $M^k \gg M^k_j$. For example, in the NIST database that we are using, the average values of M^C_j and M^I_j (across all j) are respectively $E_j[M^C_j] = 9$ and $E_j[M^I_j] = 96$ and the number of users, J, is 124. Furthermore, $M^I = 11992$ and $M^C = 1172$.

The FAR and FRR given an *a priori* chosen threshold Δ, are defined as follow:

$$\begin{aligned} \text{FAR}(\Delta) &= 1 - \Psi^I(\Delta) && (2) \\ \text{FRR}(\Delta) &= \Psi^C(\Delta) && (3) \end{aligned}$$

where

$$\Psi^k = P(Y^{k,\prime}_{\cdot} < \Delta),$$

and $Y_{\cdot}^{k,\prime} \equiv Y_m^{k,\prime} | m \in \mathcal{M}^k$ is a class-conditional variable (dependent on k) that does not consider the user index j. Note that, in theory, the class-conditional cumulative density function (*cdf*) is a smooth function. In practice, however, it may be a stair-case like function if there are too few score samples. A DET curve [2] is plotted by tracing $\Delta \in [-\infty, \infty]$ across the following coordinate:

$$\mathbf{v} \equiv (v_{FAR}(\Delta), v_{FRR}(\Delta)) \equiv (\Psi^{-1}(\text{FAR}(\Delta)), \Phi^{-1}(\text{FRR}(\Delta))),$$

where Ψ^{-1} is the inverse of a normal *cdf*.

We will describe below four variants of bootstrap to generate a pool of *cdf*s that captures different sources of variability.

1) Conventional Bootstrap: In order to generate a confidence bound using the conventional sample bootstrap approach, one draws M^k samples *with replacement* from the sample index set $\mathcal{M}^k$ to create the s-th bootstrap, $\mathcal{M}_s^k$. The *cdf* due to the bootstrap $\mathcal{M}_s^k$ is:

$$\Psi_s^{k,\prime} = P(Y_m^{k,\prime} < \Delta | m \in \mathcal{M}_s^k). \tag{4}$$

2) Bootstrap Subset: The bootstrap subset approach [6] uses a user-specific subset which, written in the form of Eqn. (4), is

$$\Psi_u^k = P(Y_{j,\cdot}^k < \Delta | j \in \mathcal{J}_u), \tag{5}$$

where $\mathcal{J}_u$ is the u-th bootstrap of users drawn J times with replacement from the pool of all possible users $\mathcal{J}$. In this way, all the samples according to the selected users are drawn at the same time.

3) User-Constrained Sample Bootstrap: Similarly, one can also consider the sample variability of a class-conditional *cdf* by using the following definition:

$$\Psi_s^k = P(Y_{j,m}^k < \Delta | j \in \mathcal{J}, m \in \mathcal{M}_j^k(s)), \tag{6}$$

where, the set $\mathcal{J}$ is fixed but only their corresponding sample varies. Note that $\mathcal{M}_j^k(s)$ deontes the s-th bootstrap with replacement of the original user-specific index set $\mathcal{M}_j^k$. The DET curves due to Ψ_s^C and Ψ_s^I for different s bootstraps will reflect how the sample varies given the population.

4) Joint/Two-Level Bootstrap: Combining both the variability due to Eqn. (5) and Eqn. (6), one obtains the following class conditional *cdf*:

$$\Psi^k_{u,s} = P(Y^k_{j,m} < \Delta | j \in \mathcal{J}_u, m \in \mathcal{M}^k_j(s)). \tag{7}$$

Note that in this case, the user-specific bootstrap has to be performed before the sample bootstrap, i.e., an algorithm to do so will perform the following two loops:

```
For $u \in \{1, \ldots, U\}$ For $s \in \{1, \ldots, S\}$,
    Calculate $\Psi^k_{s,u}$ for both $k = \{C, I\}$
End, End
```

The complexity in this case is $O(U \times S)$. In Eqn. (4) or Eqn. (6), the complexity is $O(S)$ whereas in Eqn. (5), the complexity is $O(U)$. Therefore, Eqn. (7) has a slightly added overhead. However, we expect its confidence bound to have the highest coverage, which is a more important goal.

IV. Establishing Confidence region

This section deals with an algorithm to estimate a confidence region on a *two-dimensional* DET plan (spanned by FAR and FRR) given a set of class-conditional *cdf*s generated by any of the four bootstrap methods mentioned in Section III, i.e., Eqn. (4)–Eqn. (7). According to [5], there are several ways to construct a confidence region, called “sweeping methods”, given a set of class-conditional *cdf*s. These sweeping methods, in our context, are:

- **Vertical Averaging:** It works by fixing FAR and calculating the intervals of the corresponding FRR. A variant of this procedure is to fix FRR and calculate the corresponding FAR confidence bounds. A connected DET *region* can be constructed by joining all the neighboring vertical confidence bounds. A similar approach, termed *horizontal averaging*, fixes FRR instead FAR in order to estimate the corresponding FAR confidence bounds.
- **Threshold Averaging:** It works by averaging FAR and FRR values of different DET curves based on a *common* threshold. Bolle’s technique [6] that uses Eqn. (5) falls into this category.

- **Simultaneous Joint Confidence Regions:** This technique does not fix any threshold nor any axes on the DET plan but instead estimates a confidence region based on a set of paired (FAR,FRR) data points directly. Two variants were reported in [6], i.e., fixed-width band [10] and working-hotelling band [11]. The fixed-width band method, in our context, obtains a confidence region that is defined by two parallel DETs[3] with a fixed width distance such that the original observed DET is fully contained inside the region. The working-hotelling band fits the best regression line in the DET plan. Therefore, it assumes that the class-conditional scores follow a Gaussian distribution.

We propose here another method that also belongs to the third category. This method directly estimates the two-dimensional density of the bootstrapped DET curves spanned by all possible pairs of FAR and FRR values. In comparison to [10], the upper and lower DET curves do not have to be parallel or of fixed width because in our case, the bounded region is completely defined by the observed bootstrapped DET curves. In comparison to [11], our method is advantageous because one does not make the class-conditional Gaussian assumption.

The two-level bootstrap as in Eqn. (7) will be used since it generalizes Eqn. (5) and Eqn. (6). The generalization of the implementation to the conventional sample bootstrap, as in Eqn. (4), is straightforward. Our task here is to estimate $p(\mathbf{v})$, the likelihood of an arbitrary location in a DET plan – denoted by $\mathbf{v}$ – from the $U \times S$ bootstrapped DET curves. While several density estimation methods can be used, e.g., a mixture of Gaussian components and Parzen windows [12], one requirement of $p(\mathbf{v})$ is that the density must be defined everywhere in the DET plan and sum to one. Two approaches are proposed here:

- **Direct estimation using GMM:**

$$p(\mathbf{v}) = \sum_{c=1}^{C} w_c \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_c) \tag{8}$$

  where $\mathcal{N}$ is a bivariate Gaussian with mean $\boldsymbol{\mu}_c$ and covariance $\boldsymbol{\Sigma}_c$ for the c-th component, each weighted by w_c such that $\sum_c w_c = 1$. These parameters can be optimized using the Expectation Maximization algorithm and the number of components C can be optimized using cross-validation

[3] The original method applies to the ROC plan.

or some other criteria [13] (e.g., minimum description length). Unfortunately, very soon, we found that this method is not appropriate due to the GMM assumption that every $\mathbf{v}$ is independently and identically sampled. This assumption is violated since the $\mathbf{v}$ values that belong to the same curve (of a particular s-th and u-th bootstrap) are *not independent* on each other. Despite this weakness, the estimated $p(\mathbf{v})$ is still useful for characterizing the flatness of the distribution in terms of entropy (to be described further).

- **Estimation via a DET angle:** One way to overcome the mentioned weakness is to work on polar coordinates. By taking $\mathbf{v}$ in Cartesian coordinates, we define its corresponding polar coordinate to be (θ, r) where

$$\theta = \tan^{-1}\left(\frac{v_{FRR}(\Delta) - v_{FRR}(-\infty)}{v_{FAR}(\Delta) - v_{FAR}(-\infty)}\right),$$

and

$$r = \sqrt{(v_{FRR}(\Delta) - v_{FRR}(-\infty))^2 + (v_{FAR}(\Delta) - v_{FAR}(-\infty))^2},$$

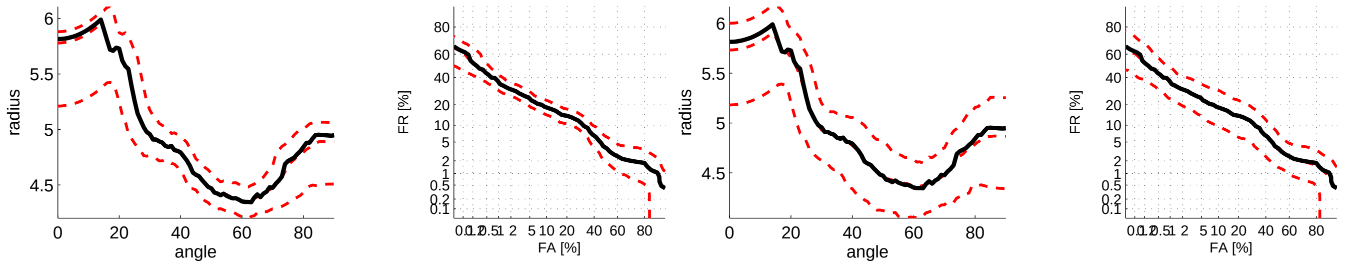
for $\theta \in [0, \pi/2]$, $r \in [-\infty, \infty]$ and $(v_{FAR}(-\infty), v_{FRR}(-\infty))$ is the origin[4]. To obtain $\alpha \times 100\%$ confidence given the set of bootstrapped DET curves in polar coordinates, we estimate the upper and lower bounds:

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

where $\Psi_\theta(r)$ is the empirical *cdf* of the radius r observed from the $U \times S$ bootstrapped curves for a given θ since each bootstrapped curve cuts through θ exactly once. The lower and upper r will be given by $r_{lower} = \Psi_\theta^{-1}(\frac{1-\alpha}{2})$ and $r_{upper} = \Psi_\theta^{-1}(\frac{1+\alpha}{2})$, respectively. Note that the inverse of Ψ_θ, i.e., Ψ_θ^{-1}, requires linear interpolation[5]. The corresponding lower (more optimistic) DET curve is

[4] Since $\Psi^{-1}(-\infty) = -\infty$, in practice, we replace the origin with the point $(\Psi^{-1}(1/N), \Psi^{-1}(1/N))$ where N is the total number of impostor attempts rounded to the nearest (and larger) power of 10. For example, if the number of impostor attempts is 3,800, then 10,000 can be used.

[5] In our implementation, we verified that by projecting a DET curve into polar coordinates and then reversing the process, one obtains *exactly* the same DET curve. Therefore, there is no loss of generality by working on polar coordinate as long as the *same* origin (according to footnote 4) is used.



(a) sample bootstrap

(b) user-specific bootstrap

Fig. 1. The 95% confidence of (a) conventional bootstrap ($U = 1, S = 400$) and (b) user-specific bootstrap ($U = 400, S = 1$), of one of the 24 systems in NIST2005, calculated from 80 users, are shown here in the (θ, r) polar coordinate (left) as well as its corresponding Cartesian coordinate or DET plan (right). The actual DET, plotted with a dark line, is always included in the upper and lower DET bounds.

given by $(r_{lower}\cos(\theta), r_{lower}\sin(\theta))$ across all $\theta \in [0, \pi/2]$. The upper (less optimistic) DET curve is defined similarly. By convention, the significance threshold α is set to 0.05 so that one obtains a 95% level of confidence. Note that DET angle was reported in [14] to combine several DET curves into a single one. Although DET angle seems to be an uncommon choice, three θ values are extremely commonly used: $\{0, \frac{\pi}{4}, \frac{\pi}{2}\}$. They correspond respectively to the estimates of confidence interval of FAR at FRR=0, EER and that of FRR at FAR=0. Therefore, the procedure described here can be seen as a generalization to this practice.

In this paper, we mainly use the DET angle approach to derive a confidence region from a set of bootstrapped DET curves. The GMM approach is used merely to quantify the flatness of the distribution since it is not directly obvious how this can be done with the DET angle approach. Note that deriving a confidence region around a DET curve is still an open question. Our experiments show that both approaches lead to very similar results in terms of coverage (to be reported in Section V) and conclusions (Section VI).

Two examples of the 95% confidence of a DET curve generated using the conventional and Bolle's bootstrap subset technique are shown in Figure 1. Note that like any density estimation algorithm, too few DET curves (due to too small a number of sample bootstraps S or user-specific U bootstraps,) will result in poor estimation of DET confidence.

Since we have an estimate of $p(\mathbf{v})$, we can characterize the flatness of the distribution $p(\mathbf{v})$ using entropy, i.e.,

$$\text{entropy}(p) = \int_{\mathbf{v}} -p(\mathbf{v}) \log p(\mathbf{v}).$$

We expect the following property to hold for biometric authentication tasks:

$$\text{entropy}(p|\Psi^k_{s,1}, s \in \mathcal{S}, \forall_k) \leq \text{entropy}(p|\Psi^k_{1,u}, u \in \mathcal{U}, \forall_k) \leq \text{entropy}(p|\Psi^k_{s,u}, u \in \mathcal{U}, s \in \mathcal{S}, \forall_k), \qquad (9)$$

for some fixed S and U. The first term is the entropy of the user-constrained sample bootstrap, the second term is the entropy of the bootstrap subset technique and the third term is entropy of our proposed two-level bootstrap. Smaller entropy implies a shaper distribution. The rationale of the above relationship is that the sample variability is lower than the user-induced variability and that the joint effect of the two sources of variability is larger than using either one. Eqn. (9) will be experimentally verified in Section V in terms of coverage. In [6], it was shown that the conventional sample bootstrap underestimates the confidence bounds compared to the bootstrap subset technique. This indicates that the entropy of the sample bootstrap is lower than that of the bootstrap subset technique. We verify this finding using an example already shown in Figure 1.

V. Empirical Evaluations

A. Effects of S and U Parameters

The goal of this section is two-fold: to empirically verify Eqn. (9) in terms of coverage and to determine the number of sample and user-specific bootstraps, S and U, that are needed in practice. Three variants of the same algorithm can be obtained by setting S and U

1) Sample bootstrap when S varies and $U = 1$
2) User-specific bootstrap when $S = 1$ and U varies
3) Joint user-specific and sample bootstraps when both S and U vary

An experiment is carried out for each of the 24 systems in the following ways: Two data sets are obtained such that the development set contains the (client and impostor) scores of 20 users, the evaluation set contains the scores 80 users and the development set is a subset of the evaluation set. In this way, we estimate a confidence region based on the 20-user data set and evaluate the prediction performance of the future DET, in terms of coverage, based on the 80-user data set. We then apply all the four bootstrap algorithms whose confidence region is estimated using the density-based approach as described in Section IV. Figure 2 shows the effects of varying one of the two free parameters S and U. The Y-axis is the average coverage over the 24 systems. A coverage is calculated as the proportion of the 80-user DET curve that is included in the confidence region obtained from the 20-user DET curve.

We can make the following observations

1) The user-constrained sample bootstrap technique has the lowest coverage.
2) The coverage of the joint bootstrap technique is never lower than that of the bootstrap subset technique in terms of coverage.
3) The conventional sample bootstrap technique has coverage lower than the user-specific bootstrap given asymptotically large number of bootstraps.
4) $S, U > 30$ are suitable.
5) The joint bootstrap technique and the user-specific bootstrap technique converge for large U.

Observations one and two confirm our conjecture in Eqn. (9). Observation three confirms the finding in [6]. Observation four implies that S and U above 30 is sufficient. The last observation implies that for biometric authentication tasks, the influence of U is more important than that of S. In other words, as long as U is large (30 or more), the joint bootstrap procedure is insensitive to different S values.

B. Assessment w.r.t. Larger Population

This section evaluates the quality of DET prediction with respect to the population size. We expect that a larger population of users should give a more accurate prediction – hence producing higher coverage and lower entropy (sharper distribution). We design a *progressive prediction* experiment described as follows. First, we divide the original data set (of 124 users) into 10, 20, 40 and 80 users such that the smaller data

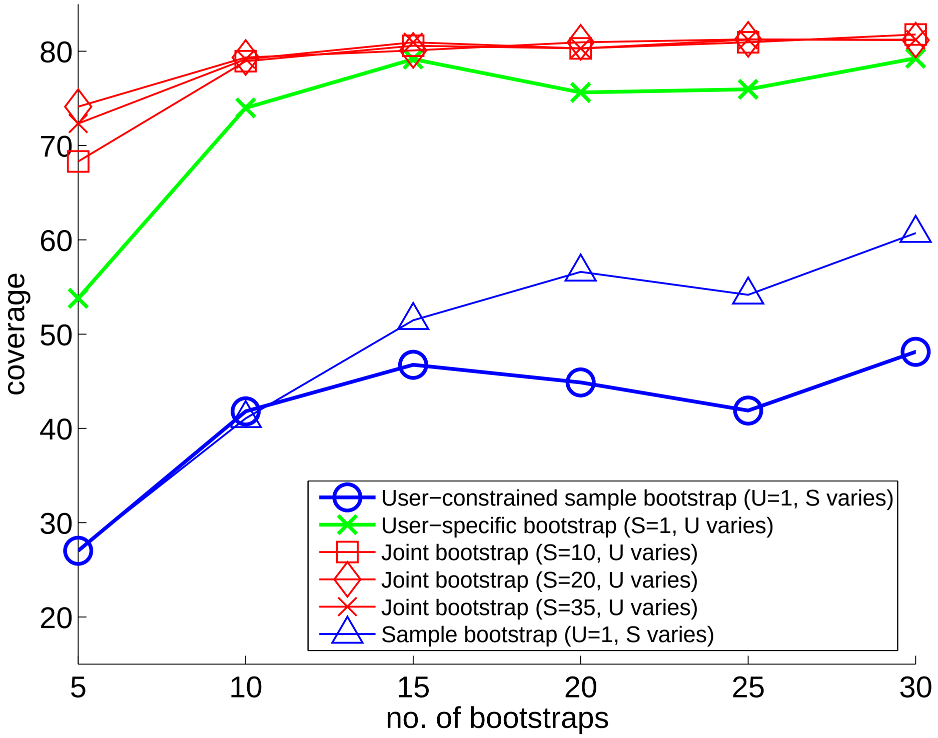
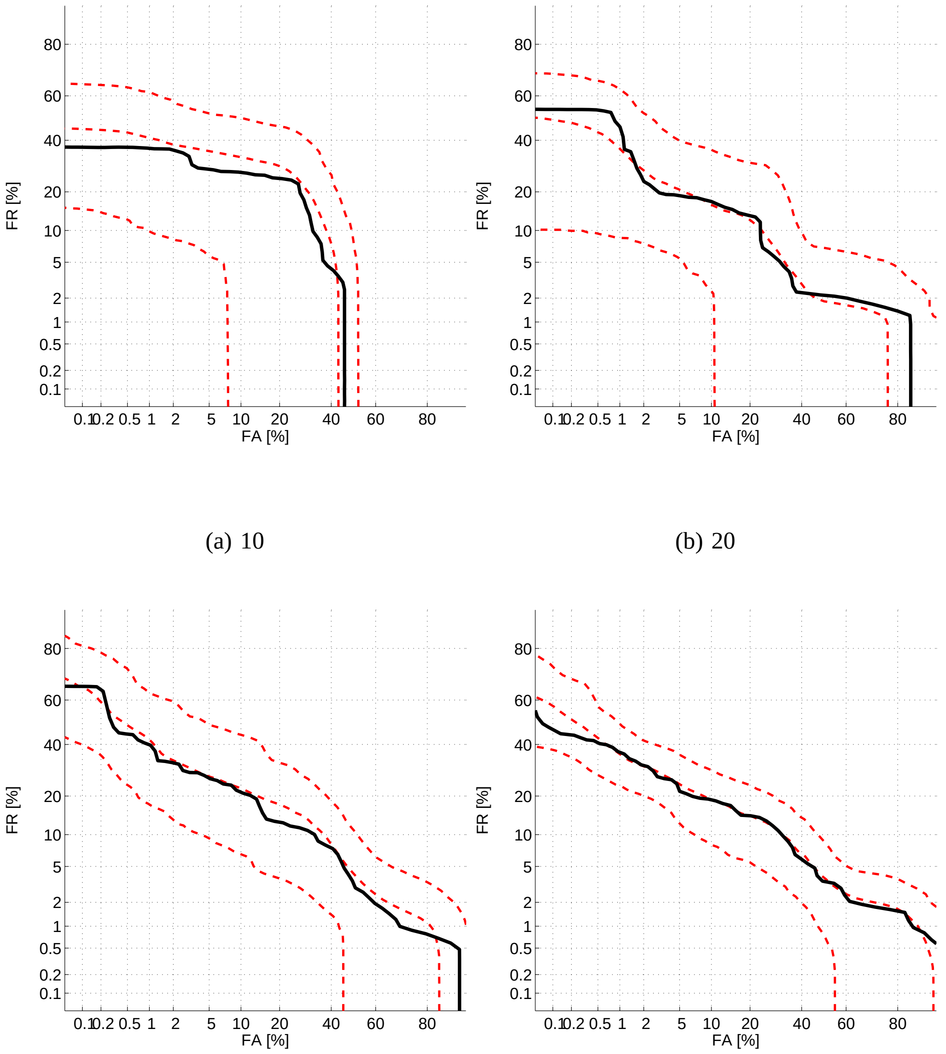


Fig. 2. Average coverage (in %) over 24 experiments using different variants of bootstraps. Each bootstrap is trained on a DET derived from 20 users and is tested on a DET derived on 80 users. Higher coverage implies better generalization. S is the number of sample bootstraps and U is the number of user-specific bootstraps. The X-axis reflects the change due to varying one of these two parameters. The user-specific bootstrap attained a maximum coverage of 81.7% while the joint-bootstrap attained 82.2%. The joint-bootstrap procedure almost always has higher coverage for any given S values.

set is always a subset of the larger one. Then, we apply the joint bootstrap technique on 10-user data set and measure the coverage of the 20-, 40- and 80-user DETs. The experiment is repeated but this time we apply the joint bootstrap procedure on the 20-user data set and test it on the 40- and 80-user data sets. Finally, the experiment is repeated with training on the 40-user data set and testing on the 80-user data set. The above procedure is tested using all the 24 systems available in NIST2005. A graphical output of this procedure for one of the 24 systems is shown in Figure 3.

The average entropy and coverage over all the 24 systems are shown in Table I and Table II, respectively. As can be observed and expected, the entropy of the bootstrapped DET mass decreases as more data is available. This trend can graphically be observed in Figure 3 as well. Coverage generally increases as more and more data is available.



(a) 10

(b) 20

(c) 40

(d) 80

Fig. 3. 95% DET mass of one of the 24 systems calculated using the DET angle approach for 10, 20, 40 and 80 users. Their corresponding entropies are 7.2116, 7.1109, 6.8357 and 6.7438, respectively (calculated using the GMM approach). For each figure, the DET confidence region is bounded by an upper and a lower DET plotted in dashed lines. The median of the region is also plotted by a dashed line. The actual observed DET from which the confidence region is derived is plotted in a bold continuous line.

C. Assessment w.r.t. User Composition Variation

This section assesses the robustness of the joint bootstrap technique to user composition. Using the subset of NIST2005 which contains 124 users, we randomly divided the data into four equal partitions, each containing 31 users. Data set 1 is used to estimate the confidence region while the rest of the data sets are used *separately* to evaluate the goodness of prediction in terms of coverage. The same procedure is repeated with data set 2, 3 and 4. The results are shown in Table III. Another experiment is repeated

TABLE I

ENTROPY OF DET MASS ESTIMATED USING DATA SETS OF DIFFERENT USER SIZE, AVERAGED OVER 24 SYSTEMS IN NIST2005.

User size	Entropy
10	7.094
20	6.975
40	6.841
80	6.658

Lower entropy implies sharper distribution.

TABLE II

AVERAGE COVERAGE OVER 24 SYSTEMS IN NIST2005 WITH 95% OF CONFIDENCE.

Actual	Coverage for predicted user size			
user size	10	20	40	80
10	$*$ 1.000	0.879	0.773	0.781
20	–	$*$ 1.000	0.891	0.859
40	–	–	$*$ 1.000	0.886
80	–	–	–	$*$ 1.000

Note: Numbers marked with “*” do not involve prediction.

but with only two partitions where each partition contains 62 users. The resultant matrix is similar to Table III except that it is 2-by-2 in dimension. Its values are:

$$\begin{bmatrix} *1.000 & 0.872 \\ 0.893 & *1.000 \end{bmatrix},$$

where the same conclusion as in Table III applies. The overall average coverage is evaluated to be 0.883 instead of 0.825. The increased coverage is due to the fact that the DET curves are estimated from an increased number of users.

TABLE III

AVERAGE COVERAGE OVER 24 SYSTEMS IN NIST2005

Data	Coverage due to data sets			
sets	1	2	3	4
1	1.000	0.802	0.859	0.874
2	0.794	1.000	0.798	0.862
3	0.866	0.743	1.000	0.846
4	0.846	0.794	0.819	1.000

Note: The diagonal of the table is one because the DET is tested on the data set from which it is derived. The average coverage across the non-diagonal elements is 0.8252.

D. Validation on the XM2VTS Database

In order to verify the repeatability of experimental results on other databases, we used the XM2VTS score-level fusion benchmark database [15][6]. The first Lausanne protocols was chosen and eight verification systems – three speech and five face systems – are available. This database contains 200 users and each user has two genuine samples and 600 impostor samples. We repeated a similar experimental setting as in Section V-C using two configurations:

1) Four partitions – hence 50 users per partition; test on same impostor set
2) Four partitions – hence 50 users per partition; test on *different* impostor set

Note that the most important difference between XM2VTS and NIST2005 is that in XM2VTS, the impostor population can be the same (using the fusion development set) or different (using the fusion evaluation set). By using $S = 5$ and $U = 100$, the coverage averaged over the eight systems is 70.5% for the first configuration and 67.2% for the second configuration. Therefore, by using a *different* impostor population set, the coverage is reduced. This indicates that by *re-using* the same impostors to generate impostor scores, one will obtain an overly optimistic bias of coverage.

[6] Available at http://www.idiap.ch/∼norman/fusion

VI. Conclusions

Generalizing performance of a system from a particular database to another is an important task. The current visualization tool via a DET curve does not guarantee such performance generalization. On the contrary, it is very sensitive the following three factors: the number of users, the choice of users and the choice of samples. Using a two-level bootstrap and a post-processing on the density of resultant DETs, we propose to establish a contour capturing $(1 - \alpha) \times 100\%$ of probability mass. The proposed two-level bootstrap approach generalizes the bootstrap subset technique as proposed by Bolle *et al.* [6] because our proposal takes into consideration the sample variability in addition to the user-induced variability. Both theoretical and empirical findings suggest that the two-level bootstrap approach has a systematically higher coverage than Bolle *et al.*'s bootstrap subset. Although the experimental settings can be different, the established confidence region from a small database of users can cover more than 75% of an actual unseen DET with 8 times the number of users. Ideally, a good indicator should score 95% of the actual DET. While a DET is inherently sensitive to the three aspects of variability mentioned, the proposed bootstrap procedure can mitigate such sensitivity to some extent but cannot totally remove it.

The following are some possible extensions to the current study:

- **Confidence interval estimation for threshold-dependent analysis:** A DET curve is a threshold-independent analysis. However, by using a DET curve, one assumes that the FAR and FRR distributions of the test data set are completely known. An recently proposed alternative is to use a threshold-dependent assessment whereby the system performance is calculated with thresholds fixed *a priori* [16] on a development set.
- **Mismatch between training and test sets:** The current study does not handle the case of mismatch between training and test sets. Research in this direction will require that some representative test samples to be available so that the confidence region of a DET derived from some training conditions can be transformed into that of the target test conditions.

ACKNOWLEDGMENTS

The authors thank Mark Przybocki for making the NIST2005 data sets available. Part of this work was carried out during the principle author's one month visit at NIST. This work was supported in part by the IST Program of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, funded in part by the Swiss Federal Office for Education and Science (OFES) and the Swiss NSF through the NCCR on IM2. This publication only reflects the authors' view.

REFERENCES

[1] J. Wayman, A. Jain, D. Maltoni, and D. Maio, *Biometric Systems: Technology, Design and Performance Evaluation*, Springer, 2005.

[2] A. Martin, G. Doddington, T. Kamm, M. Ordowsk, and M. Przybocki, "The DET Curve in Assessment of Detection Task Performance," in *Proc. Eurospeech'97*, Rhodes, 1997, pp. 1895–1898.

[3] G. Ma and Z.J. Hall, "Confidence Bands for ROC curves," , no. 13, pp. 191–197, 1993.

[4] S. Macskassy and F. Provost, "Confidence bands for ROC curves: Methods and an empirical study," in *Proc. 1st Workshop ROC Analysis in AI: ROCAI*, 2004, pp. 61–70.

[5] S. Macskassy, F. Provost, and S. Rosset, "ROC Confidence Bands: An Empirical Evaluation," in *Proc. 22nd Int'l. Conf. Machine Learning (ICML)*, Bonn, 2005.

[6] 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.

[7] S. Dass and A. Jain, "Effects of User Correlation on Sample Size Requirements," in *Proc. SPIE Vol. 5779, Biometric Technology for Human Identification II*, 2005, pp. 226–231.

[8] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "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.

[9] NIST, "The 2005 NIST Speaker Recognition Evaluation," 2005, [Available] http://www.itl.nist.gov/iad/894.01/tests/spk/2005/.

[10] G. Campbell, "Advances in Statistical Methodology for the Evaluation of Diagnostics and Laboratory Tests," *Statistics in Medicine*, vol. 13, 1994.

[11] H. Working and H. Hotelling, "Application of the Theory of Error to the Interpretation of Trends," *Statistics in Medicine*, vol. 24, 1929.

[12] C. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1999.

[13] M.A.T. Figueiredo and A.K. Jain, "Unsupervised learning on finite mixture models," *Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, March 2002.

[14] A. Adler and M. E. Schuckers, “Calculation of a Composite DET Curve,” in *LNCS 3546, 5th Int’l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2005)*, New York, 2005, pp. 860–868.

[15] N. Poh and S. Bengio, “Database, Protocol and Tools for Evaluating Score-Level Fusion Algorithms in Biometric Authentication,” *Pattern Recognition*, vol. 39, no. 2, pp. 223–233, February 2005.

[16] S. Bengio and J. Mariéthoz, “The Expected Performance Curve: a New Assessment Measure for Person Authentication,” in *The Speaker and Language Recognition Workshop (Odyssey)*, Toledo, 2004, pp. 279–284.

PLACE
PHOTO
HERE

Norman Poh is a visiting research fellow at the Centre for Vision, Speech and Signal Processing laboratory of University of Surrey. He recieved a Ph.D. in computer science from the Swiss Federal Institute of Technology in Lausanne or Ecole Polytechnique Fdrale de Lausanne (2006). His research interest is in biometric recognition and machine learning, in general, and multiple classifiers fusion, in particular. He won the best student poster award in the Audio Visual Biometric Person Authentication Conference (AVBPA) for his contribution on biometric fusion, 2005.

PLACE
PHOTO
HERE

Alvin Martin has served as a mathematician in the Speech Group at the National Institute of Standards and Technology since 1991. He has coordinated NIST’s evaluations over the past decade in the areas of speaker recognition and of language and dialect recognition, and has contributed to its evaluations of large vocabulary continuous speech recognition. This work has involved the collection, selection, and pre-processing of appropriate speech data, the writing of evaluation plans, the specification of metrics and charts for the scoring, presentation, and analysis of results, the implementation of statistical tests for determining the significance of performance differences, and the organization of workshops to review evaluation results. He received a Ph.D. in mathematics from Yale University (1977), has taught mathematics and computer science at the college level, and worked on the development of automatic speech recognition and speech processing systems before coming to NIST.

PLACE
PHOTO
HERE

Samy Bengio is a senior researcher and the machine learning group leader at the IDIAP Research Institute since 1999, where he supervises PhD students and postdoctoral fellows working on many areas of machine learning such as support vector machines, time series prediction, mixture models, large-scale problems, speech and speaker recognition, multi-modal problems, brain computer interfaces, and many more. He has obtained his PhD in computer science from Universite de Montreal (1993), and spent three post-doctoral years at CNET, the research center of France Telecom, and INRS-Telecommunications (Montreal). He then worked as a researcher for CIRANO, an economic and financial academic research center, applying learning algorithms to finance. Before joining IDIAP, he was a research director at Microcell Labs, a private research center in mobile telecommunications. His current interests include all theoretical and applied aspects of learning algorithms.