ESTIMATING THE CONFIDENCE INTERVAL OF EXPECTED PERFORMANCE CURVE IN BIOMETRIC AUTHENTICATION USING JOINT BOOTSTRAP

Norman Poh[a,b,c] *and Samy Bengio*[a]

[a] IDIAP Research Institute, CP 592, 1920 Martigny, Switzerland, and
[b] Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.
[c] CVSSP, University of Surrey, Guildford, GU2 7XH, Surrey, UK
{norman,bengio}@idiap.ch

ABSTRACT

Evaluating biometric authentication performance is a complex task because the performance depends on the user set size, composition and the choice of samples. We propose to reduce the performance dependency of these three factors by deriving appropriate confidence intervals. In this study, we focus on deriving a confidence region based on the recently proposed Expected Performance Curve (EPC). An EPC is different from the conventional DET or ROC curve because an EPC assumes that the test class-conditional (client and impostor) score distributions are unknown and this includes the choice of the decision threshold for various operating points. Instead, an EPC selects thresholds based on the training set and applies them on the test set. The proposed technique is useful, for example, to quote realistic upper and lower bounds of the decision cost function used in the NIST annual speaker evaluation. Our findings, based on the 24 systems submitted to the NIST2005 evaluation, show that the confidence region obtained from our proposed algorithm can correctly predict the performance of an unseen database with two times more users with an average coverage of 95% (over all the 24 systems). A coverage is the proportion of the unseen EPC covered by the derived confidence interval.

***Index Terms*—** Biometric authentication, pattern recognition, classification

1. INTRODUCTION

Biometric authentication is in general considered as a two-class classification problem that aims at accepting or rejecting the identity claim of a user based on some biometric sample (voice, face, etc). This is done by selecting an appropriate discriminant function for each user, as well as a corresponding threshold that better suit a given cost function. A common practice is to compare solutions with respect to the whole range of possible threshold values. The result is then visualized using a Receiver's Operating Characteristic (ROC) curve [1] or a Detection Error Trade-off (DET) curve [2] estimated on some test set.

In order to provide quantitative comparisons, one or several operating points of the DET/ROC are selected. For example, one commonly used operating point is called Equal Error Rate (EER) and is a special point where False Acceptance Rate (FAR) equals False Rejection Rate (FRR). Finally, some aggregate measure is chosen to represent the performance at these operating points. One such measure is the Half Total Error Rate, which is the average between FAR and FRR.

When the operating point is chosen according to the test set, we say the underlying performance is *a posteriori*, as it was obtained by looking at the test set in order to modify some parameter of the model (the threshold). On the other hand, when the operating point is chosen on some independent validation set, the underlying test set performance is called *a priori*, and is unbiased.

A priori evaluation is used in the NIST speaker evaluation [3, Chap. 8] and the BANCA face and speech [4] databases. In NIST, the operating performance is called the decision cost function (DCF; or the C_{DET} point). In BANCA, three cases of operating performance are quoted in terms of HTER.

The *a priori* evaluation is more realistic because in a real application, the *true* class-conditional distributions are unknown. In [5] as well as a follow-up study [6], it was argued that comparing two models based on several chosen points on a DET/ROC curve can be misleading. As a result, the authors proposed the Expected Performance Curve (EPC) [5]. An EPC is the operating performance due to a systematic evaluation of all possible *a priori* chosen thresholds.

The goal of this study is to provide a reasonable confidence interval around the *a priori* evaluation coming from an EPC. The confidence interval of *a posteriori* evaluation using DET/ROC curve has already been examined, in [7] for instance.

Considering the problem of biometric authentication as a two-class problem is only partly true. If there are J users, then, there are really J two-class problems because there are J models which are very often individually optimized for each user. A special property is that the decision scores of a particular user model – the intra-model scores (all decisions taken by a given client model) – are probably dependent with each other, whereas two sets of scores due to two different user models – the inter-model scores – are probably much less dependent. As a result, when deriving a confidence interval around a ROC/DET/EPC curve, one has to take into consideration the intra-model scores dependency: simple statistical tests such as the proportion test or a plain bootstrap based technique [8] are thus prone to optimistic biases. One technique that preserves this dependency is called the *bootstrap subset* technique [9]. The idea is to bootstrap the model (user) identities such that when a model identity is selected, its associated client and impostor scores are all selected. This is in contrast to the conventional bootstrap technique that bootstraps the scores directly without considering the model identity. In so doing, it destroys the intra-model scores dependency. As a result, the conventional bootstrap technique grossly underestimates the confidence interval compared to the bootstrap subset technique [9]. Similarly, instead of using the bootstrap approach, conventional parametric approaches (such as the proportion test) also grossly underestimates the true confidence interval because they do not preserve the intra-model scores dependency neither. The original contributions of this paper are as follows:

- **Proposal of joint user-specific and sample bootstrap:** A biometric authentication experiment is largely dependent on three factors: the composition of users, the number of users and the choice of samples for a given user. By preserving the intra-model scores dependency, the bootstrap subset technique effectively considers

This work was supported in part by the the Swiss NSF through the prospective researcher fellowship PBEL2-114330 and through NCCR on IM2; the IST Program of the European Community, under the PASCAL Network of Excellence, IST-2002-506778; and the BioSecure project. The authors also thank the Speech Group at NIST for providing the data. This publication only reflects the authors' view.

the composition of users. However, it does not consider the sample variability. For example, it is known that a biometric device cannot produce exactly the same score even if two biometric samples are acquired from the same person. This is because biometric traits are deformable over both short and long periods of times (seconds or years), susceptible to environmental noise, the state of the user, etc. We therefore propose to take into account both the user- and sample-variability using a two-level bootstrap approach.

- **A more realistic confidence interval estimation of EPC:** Prior works in this direction include [10] where a parametric approach was proposed to estimate the EPC confidence interval; and [6, 8] where a bootstrap approach was used. In both approaches, the intra-model scores dependency was not preserved. Thanks to our joint bootstrap approach, as will be shown later, the estimated confidence interval is much better in terms of *coverage* and is guaranteed to be as good as, if not better than, the bootstrap subset technique [9].

Coverage is a conventional way to quantify the quality of an estimated confidence interval, e.g., [11]. A coverage is the proportion of a future (test) EPC curve, in our case, that is completely covered by the confidence bound derived from a current (train) EPC curve. Both the future and current EPCs may be different in terms of user composition, number of users and choice of samples.

1.1. Organization

Section 2 gives an overview of EPC. Section 3 discusses four variants of bootstrap techniques that can be used to derive the confidence interval of an EPC. The database and experiments are reported in Sections 4 and 5, respectively. This is followed by conclusions in Section 6.

2. THE EPC PROCEDURE

The EPC procedure [6] requires two sets of labeled score data. Let us call them the development and the evaluation sets, i.e., $\mathcal{D} \in \{dev, eva\}$. The False Acceptance Rate (FAR) and False Rejection Rate (FRR) can be calculated from each set of labeled score data as:

$$\text{FAR}(\Delta|\mathcal{D}) = 1 - \Psi(\Delta|\mathcal{D}, I) \quad (1)$$

$$\text{FRR}(\Delta|\mathcal{D}) = \Psi(\Delta|\mathcal{D}, C). \quad (2)$$

where $\Psi(\Delta|\mathcal{D}, k)$ is the empirical cumulative density function (*cdf*) of the labeled score set $\mathcal{D}$ according to whether the data contains client or impostor attempts, i.e., $k \in \{C, I\}$, and up to the accept/reject decision threshold Δ.

The development score set is used to choose a threshold according to a cost to be minimized, i.e.,

$$\Delta_\beta = \arg\min_{\Delta} \text{cost}_\beta(\Delta|dev), \quad (3)$$

where $\beta \in [0, 1]$ parameterizes the cost function. Three types of parametric cost functions are commonly used, i.e., Weighted Error Rate, FAR and FRR. They are defined as:

$$\text{cost}^{wer}_\beta(\Delta|dev) = \beta\text{FAR}(\Delta|dev) + (1-\beta)\text{FRR}(\Delta|dev), \quad (4)$$

$$\text{cost}^{far}_\beta(\Delta|dev) = |\beta - \text{FAR}(\Delta|dev)|, \quad (5)$$

$$\text{cost}^{frr}_\beta(\Delta|dev) = |\beta - \text{FRR}(\Delta|dev)|. \quad (6)$$

For WER, the role of β can be seen as the relative cost of FAR with respect to FRR. Minimizing $\text{cost}^{far}_\beta(\Delta|dev)$ can be interpreted as finding the threshold where the empirically observed FAR is closest to β, and similarly for $\text{cost}^{frr}_\beta(\Delta|dev)$.

Once the optimal threshold that minimizes a chosen β is found, the performance of the three costs when calculated on the evaluation set are respectively:

$$\text{Perf}^{wer}_{\beta,\gamma}(\Delta|eva) = \gamma\text{FAR}(\Delta_\beta|eva) + (1-\gamma)\text{FRR}(\Delta_\beta|eva), \quad (7)$$

$$\text{Perf}^{frr}_\beta(\Delta|eva) = \text{FRR}(\Delta_\beta|eva), \quad (8)$$

$$\text{Perf}^{far}_\beta(\Delta|eva) = \text{FAR}(\Delta_\beta|eva). \quad (9)$$

For WER, two types of γ are commonly used. When $\gamma = 0.5$, $\text{Perf}^{wer}_{\beta,0.5}$ is called the Half Total Error Rate (HTER). When $\gamma = \beta$, $\text{Perf}^{wer}_{\beta,\beta}$ is called the Weighted Error Rate (WER).

Our discussion here generalizes to the NIST and BANCA evaluations. For NIST, the WER cost, $\text{cost}^{wer}_{0.91}$, along with $\text{Perf}^{wer}_{0.91,0.91}$ is used to rank the participating systems. For BANCA, cost^{wer}_β is used along with $\text{Perf}^{wer}_{\beta,0.5}$, for three particular values of β: $\{0.09, 0.5, 0.91\}$.

3. CONFIDENCE ESTIMATION USING BOOTSTRAP TECHNIQUES

This section explains how a set of bootstrapped EPC curves can be generated given the development and evaluation data sets, i.e., $\mathcal{D} \in \{dev, eva\}$. In each round of bootstraps, if a set of users is chosen, their associated scores will be chosen too. For some databases, e.g., the XM2VTS database [12], the same users are present in both dev and eva set. In this case, in each round of bootstraps, the same set of users must be chosen in both the dev and eva set. In the general case, e.g., the NIST database that we will be using, each round of bootstraps can be performed independently on the dev and eva sets. The following subsection describes this process more formally for a given data set (dev or eva) and identifies four variants of bootstrap to do so.

Let the labeled development set be $\{y^k_{j,m}|dev\}$, i.e., a set of scores due to a genuine (client) or impostors, i.e., $k \in \{C, I\}$ generated by the user-specific model $j \in \{1, \ldots, J\} \equiv \mathcal{J}$ and indexed by $m \in \{1, \ldots, M^k_j\} \equiv \mathcal{M}^k_j$. There are J users in $\mathcal{J}$ and M^k_j accesses in the set $\mathcal{M}^k_j$ for each user j and class k. The evaluation set is defined similarly.

Let Y^k_j be a random variable representing the scores $y^k_{j,m}$. Let $Y^{k,\prime}$ be another random variable representing the set of scores y^k_m, which is the union of all $y^k_{j,m}$ for all users j, and let M^k be the size of that set. Therefore, $M^k \gg M^k_j$. For example, in the NIST database that we will use (to be described in Section 4), the average value 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 $M^C = 1172$.

3.1. Four Bootstrap Techniques

We describe here four variants of bootstrap to generate a pool of *cdf*s that capture different sources of variability.

3.1.1. 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, $\mathcal{M}^k_s$. The *cdf* due to the bootstrap $\mathcal{M}^k_s$ is:

$$\Psi^{k,\prime}_s = P(Y^{k,\prime} < \Delta|\mathcal{M}^k_s). \quad (10)$$

3.1.2. Bootstrap Subset

The bootstrap subset approach [9] uses user-specific subset which, written in the form of Eqn. (10), is

$$\Psi^k_u = P(Y^k < \Delta|\mathcal{J}_u), \quad (11)$$

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.

3.1.3. User-Constrained Sample Bootstrap

Similarly, one can also consider the sample variability of a class-conditional *cdf* by using the following definition:

$$\Psi_s^k = P(Y_j^k < \Delta | j \in \mathcal{J}, \mathcal{M}_j^k(s)), \tag{12}$$

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

3.1.4. Joint/Two-Level Bootstrap

Combining both the variability due to Eqn. (11) and Eqn. (12), one obtains the following class conditional *cdf*:

$$\Psi_{u,s}^k = P(Y_j^k < \Delta | j \in \mathcal{J}_u, \mathcal{M}_j^k(s)). \tag{13}$$

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 $\Psi_{s,u}^k$ for both $k = \{C, I\}$

End, **End**

The complexity in this case is $O(U \times S)$. In Eqn. (10) or Eqn. (12), the complexity is $O(S)$ whereas in Eqn. (11), the complexity is $O(U)$. Therefore, Eqn. (13) has a slightly added overhead. However, we expect its confidence bound to have the highest coverage, which is a more important goal.

3.2. Defining Confidence Bounds

We will explain below the performance in terms of HTER, i.e., $\text{Perf}_{\beta,0.5}^{wer}$. Generalization to other performance measures such as WER ($\text{Perf}_{\beta,\beta}^{wer}$), $\text{Perf}_{\beta}^{far}$ and $\text{Perf}_{\beta}^{frr}$ is straightforward. Let the performance (HTER here) be v_β for a given operating point β. We will now describe how to derive a confidence region using the joint bootstrap approach. Generalizing the procedure to the three other variants of bootstrap is also straightforward.

Given $\Psi_{s,u}^k$ for $s = 1, \ldots, S$ and $u = \{1, \ldots, U\}$, there will be a total of $S \times U$ v_β values (for a fixed β).

Let Ψ be the empirical *cdf* of a set of v_β values. We can now estimate a range of values around v_β that can be expected in probability. Let us define a $(1-\alpha) \times 100\%$ confidence region as follows:

$$\frac{1-\alpha}{2} \le \Psi(v_\beta) \le \frac{1+\alpha}{2}$$

where α has been set to 0.95 in this paper. The upper and lower bounds of v_β are given by $v_\beta^{lower} = \Psi^{-1}(\frac{1-\alpha}{2})$ and $v_\beta^{upper} = \Psi^{-1}(\frac{1+\alpha}{2})$, respectively. The upper and lower confidence region of an EPC is then bounded by $[v_\beta^{lower}, v_\beta^{upper}]$ for all $\beta \in [0, 1]$. The *width* of the confidence interval is simply $v_\beta^{upper} - v_\beta^{lower}$ in the unit scale of v_β (percentage of HTER in our case). In order for the interval to be useful, its width must be as narrow as possible. We quantify the *expected confidence width* across all β by:

$$\text{confidence width} = E_\beta[v_\beta^{upper} - v_\beta^{lower}].$$

If "confidence $\text{width}_{S,U}$" is the confidence interval of the joint bootstrap, we conjecture that, for any S and U:

$$\text{conf. width}_{S,1} \le \text{conf. width}_{1,U} \le \text{conf. width}_{S,U}, \tag{14}$$

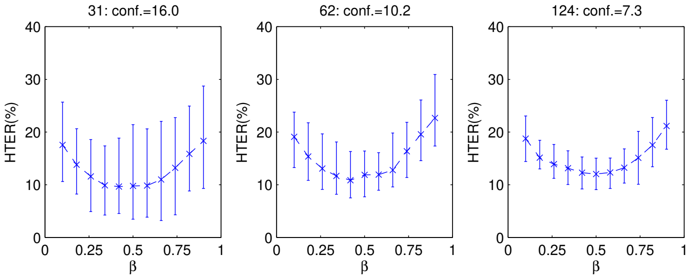

Fig. 1. EPC confidence regions derived using scores of 31, 62 and 124 users evaluated on one of the 24 systems that participated to the NIST2005 evaluation campaign. For each data set, the corresponding confidence interval (shown as "conf.") is quoted in HTER%.

i.e., the sample variability is smaller than the user-induced variability and the joint effect of the two has even larger variability.

Three examples of EPC confidence regions, derived using either 31, 62 or 124 users from one of the 24 participating systems of the NIST2005 evaluation campaign can be found in Figure 1. Note that as more and more data is available, the confidence width reduces from about 15% of HTER to about 8% of HTER. This figure is further explained in Section 5. Obviously, a large confidence interval such as 24% may produce a high coverage for any unseen EPC but its width makes it an imprecise (hence useless!) EPC predictor.

4. THE NIST2005 DATABASE

The NIST2005 score data set [13] 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[1]. Since the current study does not take into account such mismatched conditions, a subset of this data set 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 using different levels of speech information. Some systems combine different types of classifier but each classifier uses 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.

5. EMPIRICAL EVALUATION

The goal of this section is two-fold: to empirically verify the effectiveness of the joint bootstrap strategy as compared to user-specific and sample bootstrap approaches; 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 as follows:

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

In order to evaluate the coverage of a given bootstrap technique, for each of the 24 systems, two score data sets are needed: training and test sets. The training set is used to derive an EPC confidence region. The test

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

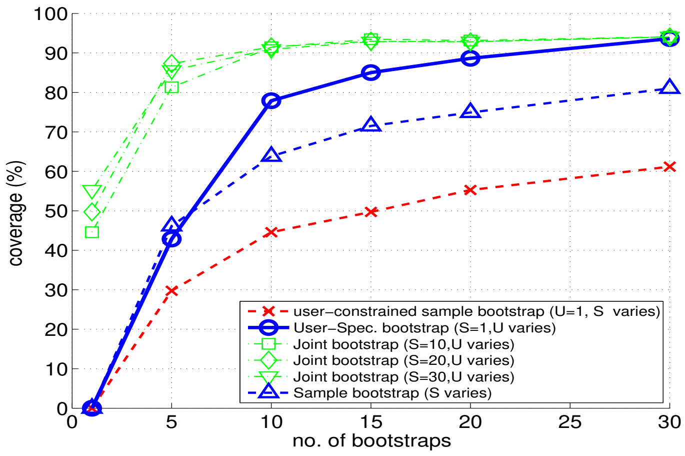

Fig. 2. Average coverage (in %) with respect to the number of bootstrap samples, over 24 experiments using different variants of bootstrap based confidence estimates. Each confidence interval is estimated using scores from 31 users and is tested using scores from another set of 62 users. Higher coverage implies more confidence on the interval estimate. 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. For this example, the maximal coverage obtained is 0.940 when α was set to 0.95. We then repeated another set of experiments but with more training data, i.e., with the scores of 62 users. The resulting maximal attainable coverage became 0.948. We conjecture that the discrepency between α and the maximal attainable coverage is possibly due to the measurement error introduced by the limited number of observed "training" users when constructing the confidence interval.

set enables us to plot an EPC. Thanks to this EPC, we can measure the quality of prediction due to the confidence region in terms of coverage. In this case, coverage is the proportion of the test EPC falling inside the confidence region. $(1-\alpha) \times 100\%$ coverage implies perfect prediction (since one should not expect a better coverage than what one asked for as reflected by α). For this experiment, the training set contains the (client and impostor) scores of 31 users (or more exactly user models), and the evaluation set contains the scores of 64 users. Note that the users in the test set are different from those in the training set. In this way, we estimate a confidence region based on the 31-user data set and evaluate its quality of prediction, in terms of coverage, based on the 64-user data set. The four variants of bootstrap algorithms are tested in this way.

Figure 2 shows the effects of varying one of the two free parameters S and U (on the X-axis). The Y-axis is the average coverage over the 24 systems.

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. (14). Observation three confirms the finding in [9]. Observation four implies that S and U superior to 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.

The average interval width, in terms of HTER units, across all β as well as *over all the 24 systems* are $\{15.15, 11.08, 8.28\}$ for the data sets with users $\{31, 62, 124\}$, respectively. Therefore, more data is needed in order to increase the precision of the estimate (decreasing the confidence interval). This trend can graphically be observed in Figure 1 as well. Although Figure 2 shows that one can predict an unseen EPC with two times more users at a seemingly impressive 95% coverage, this is achieved with an unsatisfactory large confidence interval of about 15% HTER. This suggests a need to devise novel algorithms that can narrow the confidence interval (hence increasing the precision) without demanding more data.

6. CONCLUSIONS

Interpreting a biometric authentication performance curve using a DET/ROC or EPC curve is problematic because the curve is dependent on the composition of users, the number of users and the choice of samples obtained from each user. We thus proposed a joint bootstrap approach that can put realistic upper and lower bounds on *a priori* performance evaluation based on EPC. The proposed joint bootstrap technique is shown to be always better than the bootstrap subset technique in terms of coverage.

7. REFERENCES

[1] D. Green and J. Swets, *Signal Detection Theory and Psychophysics*. John Wiley & Sons, 1964.

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

[4] E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Marithoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran, "The BANCA Database and Evaluation Protocol," in *LNCS 2688, 4th Int. Conf. Audio- and Video-Based Biometric Person Authentication, AVBPA 2003*. Springer-Verlag, 2003.

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

[6] S. Bengio, J. Marithoz, and M. Keller, "The Expected Performance Curve," in *Int'l Conf Machine Learning, ICML, Workshop on ROC Analysis in Machine Learning*, 2005.

[7] N. Poh, A. Martin, and S. Bengio, "Performance Generalization in Biometric Authentication Using Joint User-Specific and Sample Bootstraps," IDIAP, Martigny, IDIAP-RR 60, 2005, to appear in IEEE Trans. Pattern Analysis and Machine Intelligence, March 2007.

[8] M. Keller, S. Bengio, and S. Wong, "Benchmarking non-parametric statistical tests," in *Advances in Neural Information Processing Systems, NIPS 18(2005)*, 2005.

[9] R. Bolle, N. 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.

[10] S. Bengio and J. Marithoz, "A Statistical Significance Test for Person Authentication," in *The Speaker and Language Recognition Workshop (Odyssey)*, Toledo, 2004, pp. 237–244.

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

[12] J. Lüttin, "Evaluation Protocol for the XM2FDB Database (Lausanne Protocol)," IDIAP, Martigny, Switzerland, Communication 98-05, 1998.

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