A STATISTICAL SIGNIFICANCE TEST FOR PERSON AUTHENTICATION

Samy Bengio *Johnny Mariéthoz*

IDIAP
CP 592, rue du Simplon 4
1920 Martigny, Switzerland
{bengio,marietho}@idiap.ch

ABSTRACT

Assessing whether two models are *statistically significantly* different from each other is a very important step in research, although it has unfortunately not received enough attention in the field of person authentication. Several performance measures are often used to compare models, such as *half total error rates* (HTERs) and *equal error rates* (EERs), but most being aggregates of two measures (such as the *false acceptance rate* and the *false rejection rate*), simple statistical tests cannot be used as is. We show in this paper how to adapt one of these tests in order to compute a confidence interval around one HTER measure or to assess the statistical significantness of the difference between two HTER measures. We also compare our technique with other solutions that are sometimes used in the literature and show why they yield often too optimistic results (resulting in false statements about statistical significantness).

1. INTRODUCTION

The general field of biometric person authentication is concerned with the use of several biometric traits such as the voice, the face, the signature, or the fingerprints of persons in order to assess their identity [1]. In all these cases, researchers tend to use the same performance measures to estimate and compare their models. Most of them, such as the *half total error rate* (HTER) or the *detection cost function* (DCF) are in fact aggregates of other measures such as *false acceptance rates* (FARs) and *false rejection rates* (FRRs). However, when it is time to compare a novel model to existing solutions on the same problem, a quick review of the current literature in person authentication shows that either no statistical test is used to assess the difference between models, or, worse, statistical tests are wrongly used, which often ends up in over-optimistic results, tending to show, for instance, that the new model is indeed statistically significantly better than the state-of-the-art while it might not be the case in fact.

In this paper, we present a proper method to compute a simple statistical test, known as the *test of two proportions*, or *z-test*, adapted to the problem of aggregate measures such as HTER and DCF.

In section 2, we first review the main performance measures used in verification tasks, then in section 3 we recall the purpose of the *z-test*, based on the Binomial distribution, and some of its variants. In section 4, we extend this test to the case of aggregate measures such as HTER, while in section 5, we present other possible solutions, which, as explained, can lead to improper results. In fact, section 6 compares our solution to these other methods and show why they yield over-optimistic results. Section 7 concludes this paper with some proposed future work.

2. PERSON AUTHENTICATION MEASURES

A verification system has to deal with two kinds of events: either the person claiming a given identity is the one who he claims to be (in which case, he is called a *client*), or he is not (in which case, he is called an *impostor*). Moreover, the system may generally take two decisions: either *accept* the *client* or *reject* him and decide he is an *impostor*.

Thus, the system may make two types of errors: a *false acceptance*, when the system accepts an *impostor*, and a *false rejection*, when the system rejects a *client*.

Let FA be the total number of *false acceptances* made by the system, FR be the total number of *false rejections*, NC be the number of client accesses, and NI be the number of impostor accesses.

In order to be independent on the specific dataset distribution, the performance of the system is often measured in terms of rates of these two different errors, as follows:

$$\text{FAR} = \frac{\text{FA}}{\text{NI}}\,, \quad \text{FRR} = \frac{\text{FR}}{\text{NC}}\,. \tag{1}$$

A unique measure often used combines these two ratios into the so-called *detection cost function* (DCF) [2] as follows:

$$\text{DCF} = \left\{ \begin{array}{l} \text{Cost(FR)} \cdot P(\text{client}) \cdot \text{FRR} + \\ \text{Cost(FA)} \cdot P(\text{impostor}) \cdot \text{FAR} \end{array} \right. \tag{2}$$

where $P(\text{client})$ is the prior probability that a client will use the system, $P(\text{impostor})$ is the prior probability that an impostor will use the system, Cost(FR) is the cost of a false rejection, and Cost(FA) is the cost of a false acceptance.

A particular case of the DCF is known as the *half total error rate* (HTER) where the costs are equal to 1 and the probabilities are 0.5 each:

$$\text{HTER} = \frac{\text{FAR} + \text{FRR}}{2} \ . \tag{3}$$

Most authentication systems are measured and compared using HTERs or variations of it. The main question we address in this paper is thus: *how can we compute a confidence interval around an HTER or assess the difference between two systems yielding different HTERs*.

Note that in most benchmark databases used in the authentication literature, there is a significant unbalance between the number of client accesses and the number of impostor accesses. This is probably due to the relatively higher cost of obtaining the former with respect to the latter. This unbalance is the main reason why people use HTER to compare models and not the usual classification error used in the machine learning literature.

3. THE Z-TEST ON PROPORTIONS

Several statistical tests are available in the literature. For standard classification tasks, a simple yet often used test is known as the *z-test*, or *test between two proportions*. The rationale of this test is the following: given a set of n examples, each drawn independently and identically distributed (i.i.d.) from an unknown distribution, our system is going to take a decision for each example, and this decision will be correct or not. Let us now look at the distribution of the number of errors that will be made by our classification system. Since each decision is independent from the others and is binary, it is reasonable to assume that the random variable $\mathbf{X}$ representing[1] the number of errors should follow a *Binomial* distribution $\mathcal{B}(n,p)$ where n is the number of examples and p is the percentage of errors.

Moreover, it is known that a Binomial $\mathcal{B}(n,p)$ can be approximated by a Normal distribution $\mathcal{N}(\mu,\sigma^2)$ with

$$\mu = np \quad \text{and} \quad \sigma^2 = np(1-p)$$

when n is large enough[2].

Finally, if $\mathbf{X} \sim \mathcal{N}(np, np(1-p))$, then the distribution of the proportion of errors $\mathbf{Y} = \frac{\mathbf{X}}{n} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$.

[1]In this paper we use the following notation: bold letters such as **FA** represent random variables, while normal letters such as FA represent a particular value of the underlying random variable.

[2]A rule of thumb often used is to have $np(1-p)$ larger than 10.

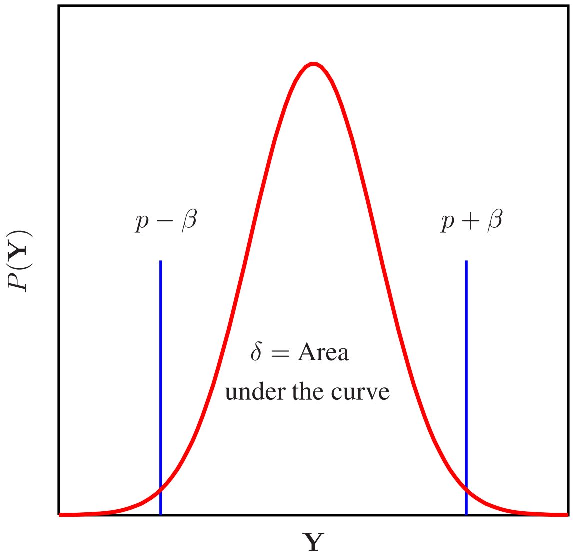

Fig. 1. Confidence intervals are computed using the area under the Normal curve.

3.1. Confidence Intervals

In order to compute a confidence interval around p, we can search for bounds $\{p-\beta, p+\beta\}$ such that

$$P\left(p-\beta < \mathbf{Y} < p+\beta\right) = \delta \tag{4}$$

where δ represents our confidence. This is called a *two-sided* test since we are searching for two bounds around p. Fortunately, finding β in (4) for a given δ can be done efficiently for the Normal distribution. Figure 1 illustrates graphically the problem and Figure 2 summarizes the procedure to obtain the confidence interval.

3.2. Difference Between Proportions

Alternatively, if one wants to verify whether a given proportion of errors p_A is statistically significantly different from another proportion p_B, a similar test can be performed. In the case where we already know that p_A cannot be lower than p_B, a *one-sided* test is used, otherwise we use a *two-sided* test. Noting respectively $\mathbf{Y}_A$ and $\mathbf{Y}_B$ the random variables representing the distribution of p_A and p_B, the *one-sided* test is based on

$$P(\mathbf{Y}_A - \mathbf{Y}_B < p_A - p_B) = \delta \tag{5}$$

while the *two-sided* test is based on

$$P(|\mathbf{Y}_A - \mathbf{Y}_B| < |p_A - p_B|) = \delta \tag{6}$$

which can be solved using the fact that the difference between two independent Normal distributions is a Normal distribution where the mean is the difference between the two Normal means and the variance is the sum of the two Normal variances, hence, if $\mathbf{Y}_A$ is not statistically different from $\mathbf{Y}_B$, then

$$\mathbf{Y}_A - \mathbf{Y}_B \sim \mathcal{N}\left(0, \frac{p_A(1-p_A) + p_B(1-p_B)}{n}\right) \quad (7)$$

and if δ is higher than a predefined value (such as 95%), then one can state that p_A is significantly different from p_B. Note that a better estimate of the variance of (7) can be obtained when assuming $p_A = p_B$ (which should be the case if they are not significantly different). In that case, equation (7) becomes

$$\mathbf{Y}_A - \mathbf{Y}_B \sim \mathcal{N}\left(0, \frac{2p(1-p)}{n}\right) \quad (8)$$

with

$$p = \frac{p_A + p_B}{2} .$$

Note however that using this test to verify whether two models give statistically significantly different results on the same test database makes a wrong hypothesis, since $\mathbf{Y}_A$ and $\mathbf{Y}_B$ are not really independent as they correspond to decisions taken on *the same test set*.

3.3. Dependent Case

One possible solution proposed in [3] is to only take into account the examples for which the two models disagree. Let p_{AB} be the proportion of examples correctly classified by model A and incorrectly classified by model B, and similarly p_{BA} be the proportion of examples correctly classified by model B and incorrectly classified by model A. In that case, the distribution $\mathbf{Y}_{AB}$ of the difference between the proportions of errors committed by each model is still Normal distributed and, assuming the two models are not different from each other, should follow

$$\mathbf{Y}_{AB} \sim \mathcal{N}\left(0, \frac{p_{AB} + p_{BA}}{n}\right) \quad (9)$$

with the corresponding two-sided test

$$P(\mathbf{Y}_{AB} < |p_{AB} - p_{BA}|) = \delta . \quad (10)$$

This test is in fact very similar to the well-known McNemar test, based on a χ^2 distribution.

In the literature, most people adopt equation (8) and some adopt equation (9); remember that in order to use equation (9), one needs to have access to all the scores of both models, and not just the numbers of errors. When possible, we will look at both solutions here, for the case of person authentication.

4. A STATISTICAL TEST FOR HTERS

HTERs are not proportions, but they are an average of two well-defined proportions (FAR and FRR). Given this, and assuming some hypotheses regarding FAR and FRR[3], *we propose here to extend the test between two proportions for the case of HTERs* as follows.

4.1. Confidence Intervals

Let the random variable **FA** represent the number of false acceptances. We can model it by a Binomial, and hence by a Normal, as follows:

$$\begin{aligned} \mathbf{FA} &\sim \mathcal{B}\left(\mathrm{NI}, \frac{\mathrm{FA}}{\mathrm{NI}}\right) \\ &\sim \mathcal{N}\left(\mathrm{NI}\cdot\frac{\mathrm{FA}}{\mathrm{NI}}, \mathrm{NI}\cdot\frac{\mathrm{FA}}{\mathrm{NI}}\cdot\left(1-\frac{\mathrm{FA}}{\mathrm{NI}}\right)\right) \\ &\sim \mathcal{N}\left(\mathrm{FA}, \mathrm{FA}\cdot(1-\mathrm{FAR})\right) . \end{aligned} \quad (11)$$

The random variable **FR** can be modeled accordingly. We can now write the distribution of the random variable **FAR** representing the ratio of false acceptances:

$$\begin{aligned} \mathbf{FAR} &\sim \mathcal{N}\left(\frac{\mathrm{FA}}{\mathrm{NI}}, \frac{\mathrm{FA}\,(1-\mathrm{FAR})}{\mathrm{NI}\cdot\mathrm{NI}}\right) \\ &\sim \mathcal{N}\left(\mathrm{FAR}, \frac{\mathrm{FAR}\,(1-\mathrm{FAR})}{\mathrm{NI}}\right) \end{aligned} \quad (12)$$

and similarly for the random variable **FRR**. Given the distribution of **FAR** and **FRR**, we can estimate the distribution of the random variable **HTER** as follows:

$$\mathbf{FAR{+}FRR} \sim \mathcal{N}\left(\mathrm{FAR{+}FRR}, \frac{\mathrm{FAR}\,(1-\mathrm{FAR})}{\mathrm{NI}} + \frac{\mathrm{FRR}\,(1-\mathrm{FRR})}{\mathrm{NC}}\right)$$

$$\frac{\mathbf{FAR{+}FRR}}{2} \sim \mathcal{N}\left(\frac{\mathrm{FAR{+}FRR}}{2}, \frac{\mathrm{FAR}\,(1-\mathrm{FAR})}{4\cdot\mathrm{NI}} + \frac{\mathrm{FRR}\,(1-\mathrm{FRR})}{4\cdot\mathrm{NC}}\right)$$

$$\mathbf{HTER} \sim \mathcal{N}\left(\mathrm{HTER}, \frac{\mathrm{FAR}(1-\mathrm{FAR})}{4\cdot\mathrm{NI}} + \frac{\mathrm{FRR}(1-\mathrm{FRR})}{4\cdot\mathrm{NC}}\right) \quad (13)$$

Using this last definition, we can now compute easily confidence intervals around HTERs using the methodology presented in section 3 and summarized in Figure 2 for classical confidence values used in the scientific literature,

Moreover, the test can be easily extended to variations of HTER, such as the DCF. For instance, in the case of

[3] such that the distributions of FAR and FRR should be independent, which may look false since they are both linked by the same model and threshold, but in fact, *given a model and associated threshold* these two quantities are indeed independent since they are computed on separate data (the client accesses and the impostor accesses), assuming the model was estimated on a separate training set, as it should be.

the well-known NIST evaluations performed yearly to compare speaker verification systems, and which use the DCF measure described by equation (2) with Cost(FR) $=10$, P(client) $=0.01$, Cost(FA) $=1$ and P(impostor) $=0.99$, the underlying Normal becomes:

$$\mathbf{DCF} \sim \mathcal{N}\left(\text{DCF}, \frac{\text{FAR}\,(1-\text{FAR})}{0.99^{-2}\cdot\text{NI}} + \frac{\text{FRR}\,(1-\text{FRR})}{100\cdot\text{NC}}\right). \quad (14)$$

4.2. Difference Between HTERs

The distribution of the difference between two HTERs assuming *independence* between the two underlying distributions is

$$\mathbf{HTER}_A - \mathbf{HTER}_B \sim \mathcal{N}\left(0, \sigma^2_{\mathbf{INDEP}}\right) \quad (15)$$

with

$$\sigma^2_{\mathbf{INDEP}} = \left\{ \begin{array}{l} \dfrac{\text{FAR}_A\,(1-\text{FAR}_A) + \text{FAR}_B\,(1-\text{FAR}_B)}{4\cdot\text{NI}} + \\ \dfrac{\text{FRR}_A\,(1-\text{FRR}_A) + \text{FRR}_B\,(1-\text{FRR}_B)}{4\cdot\text{NC}} \end{array} \right.$$

while the distribution of the difference between two HTERs assuming *dependence* between the two underlying distributions becomes

$$\mathbf{HTER}_A - \mathbf{HTER}_B \sim \mathcal{N}\left(0, \sigma^2_{\mathbf{DEP}}\right) \quad (16)$$

with

$$\sigma^2_{\mathbf{DEP}} = \frac{\text{FAR}_{AB} + \text{FAR}_{BA}}{4\cdot\text{NI}} + \frac{\text{FRR}_{AB} + \text{FRR}_{BA}}{4\cdot\text{NC}}$$

where $\text{FAR}_{AB} = \frac{\text{NI}_{AB}}{\text{NI}}$ and NI_{AB} is the number of impostor accesses correctly rejected by model A and incorrectly accepted by model B, with similar definitions for FAR_{BA}, FRR_{AB}, and FRR_{BA}.

Hence, in summary, and using the standard confidence values used in the scientific literature, we obtain the simple methodology described in Figure 2 in order to compute statistical tests for person authentication tasks[4].

5. OTHER STATISTICAL TESTS

While several researchers have pointed out the use of the *z-test* to compute statistical tests around values such as FAR or FRR (see for instance [4]), we are not aware, to the best of our knowledge, of any similar attempt for aggregate measures such as HTERs (or EER, or DCF). However, most people publishing results in verification use HTERs or DCF to assess the quality of their methods.

[4]While this summary concerns HTERs, it should now be obvious to extend it to the general DCF function.

The confidence interval (CI) around an HTER is

$\text{HTER} \pm \sigma\cdot Z_{\alpha/2}$ with

$$\sigma = \sqrt{\frac{\text{FAR}(1-\text{FAR})}{4\cdot\text{NI}} + \frac{\text{FRR}(1-\text{FRR})}{4\cdot\text{NC}}}$$

$$Z_{\alpha/2} = \begin{cases} 1.645 & \text{for a } 90\% \text{ CI} \\ 1.960 & \text{for a } 95\% \text{ CI} \\ 2.576 & \text{for a } 99\% \text{ CI} \end{cases}$$

and similarly, HTER_A and HTER_B are statistically significantly different if $z > Z_{\alpha/2}$ with

$$z = \frac{|\text{HTER}_A - \text{HTER}_B|}{\sqrt{\begin{array}{l} \dfrac{\text{FAR}_A\,(1-\text{FAR}_A) + \text{FAR}_B\,(1-\text{FAR}_B)}{4\cdot\text{NI}} + \\ \dfrac{\text{FRR}_A\,(1-\text{FRR}_A) + \text{FRR}_B\,(1-\text{FRR}_B)}{4\cdot\text{NC}} \end{array}}}$$

in the independent case, and

$$z = \frac{|\text{FAR}_{AB} - \text{FAR}_{BA} + \text{FRR}_{AB} - \text{FRR}_{BA}|}{\sqrt{\dfrac{\text{FAR}_{AB} + \text{FAR}_{BA}}{4\cdot\text{NI}} + \dfrac{\text{FRR}_{AB} + \text{FRR}_{BA}}{4\cdot\text{NC}}}}$$

in the dependent case.

Fig. 2. Methodology for statistical tests around HTERs.

One simple solution could be to consider the classification error instead of the HTER and compute statistical tests around it. Since the classification error is a well-defined proportion, we can apply the *z-test* as well; Let **CLASS** be defined as the following random variable:

$$\mathbf{CLASS} = \frac{\mathbf{FA+FR}}{\text{NC+NI}}$$

then, the corresponding underlying Normal becomes:

$$\mathbf{CLASS} \sim \mathcal{N}\left(\frac{\text{FA+FR}}{\text{NC+NI}}, \frac{\text{FA+FR}}{(\text{NC+NI})^2}\left(1 - \frac{\text{FA+FR}}{\text{NC+NI}}\right)\right) \quad (17)$$

but remember that while this test is correct to assess models according to their respective classification error, it does not say anything on the confidence one has over the corresponding HTER, which is the measure of interest in person authentication. In fact, we will show in section 6.1 that, under reasonable assumptions, the variance of **CLASS** in equation (17) is always smaller than the variance of **HTER** in equation (13), hence confidence tests using (17) will always result in over-confident statistical significance (or smaller confidence intervals). This will be explored further in the following section.

Another possible solution is to consider the HTER itself as a proportion (which it is not directly) and compute the statistical test on it. Let **NAIVE** be the random variable of this value; the underlying Normal becomes:

$$\mathbf{NAIVE} \sim \mathcal{N}\left(\text{HTER}, \frac{\text{HTER}(1-\text{HTER})}{\text{NC+NI}}\right) \tag{18}$$

Again, we will show in section 6.1 that under reasonable assumptions, the variance of **NAIVE** in equation (18) is always smaller than the variance of **HTER** in equation (13), hence confidence tests using (17) should always result in over-confident statistical significance (or smaller confidence intervals).

Yet another solution that has been proposed by some researchers (see for instance [5]) is to compute a statistical test for FAR and FRR separately and then combine the results[5]. For instance, in order to compute a confidence interval for HTER, one would average both upper bounds and both lower bounds found separately by the FAR and FRR tests. On top of the fact that there is no theoretical ground to justify such an approach, there is an evident problem with all approaches that consider separately FARs and FRRs. Two models could yield very similar HTERs but for some reason (in general linked to the choice of the threshold, which is done in general on a separate data set) one could be slightly biased toward FRRs and the other one slightly biased toward FARs. In such a case, these tests would consider them statistically significantly different while they would not be when considering globally their respective HTER instead. For this reason, we will not consider this solution further here.

6. ANALYSIS

We would like to compare in this section the use of the proposed statistical test for HTERs, with respect to the two other tests presented in section 5. We will first show that under some reasonable conditions, increasing the ratio between NI and NC will increase the difference between the variance of the Normal of the proposed test and the variance of the Normal of the other tests. Afterward, we present two real case studies where the use of the proposed statistical test would have yielded a different conclusion with regard to the confidence intervals and the difference between the compared models.

[5]The well-known NIST evaluation campaigns have also apparently recently investigated the use of the McNemar test to assess speaker verification methods, but have considered separately FARs and FRRs [6].

6.1. Theoretical Analysis

Let us first look in which conditions $\sigma^2(13)$, the variance of **HTER** as written in equation (13) is higher than $\sigma^2(18)$, the variance of **NAIVE** as written in equation (18):

$$\sigma^2(13) > \sigma^2(18) \tag{19}$$

implies that

$$\frac{\text{FAR}\,(1-\text{FAR})}{4\cdot\text{NI}} + \frac{\text{FRR}\,(1-\text{FRR})}{4\cdot\text{NC}} > \frac{\text{HTER}(1-\text{HTER})}{\text{NC+NI}}$$

which can be simplified and yields

$$0 > (\text{FAR}\cdot\text{NC} - \text{FRR}\cdot\text{NI})(\text{NI}(1-\text{FRR}) - \text{NC}(1-\text{FAR}))$$

which means that inequation (19) will be true when either NC is much less or much higher than NI (which is in general the case), and FAR is similar to FRR (again, when the threshold is chosen such that we have *equal error rate* (EER) on a separate validation set, as it is often done, this is reasonable).

Let us now look in which conditions $\sigma^2(13)$ is higher than $\sigma^2(17)$, the variance of **CLASS**, representing the classification error:

$$\sigma^2(13) > \sigma^2(17) \tag{20}$$

implies that

$$\frac{\text{FAR}(1-\text{FAR})}{4\cdot\text{NI}} + \frac{\text{FRR}(1-\text{FRR})}{4\cdot\text{NC}} > \frac{\text{FA+FR}}{(\text{NC+NI})^2} \cdot \frac{(\text{FA+FR})^2}{(\text{NC+NI})^3}$$

which can be re-written as

$$(1-\text{FRR})\text{NI}(3\text{NC}+\text{NI}) > (1-\text{FAR})\text{NC}(3\text{NI}+\text{NC})$$

and assuming FAR is similar to FRR, it can be simplified into

$$\text{NI}^2 > \text{NC}^2 \tag{21}$$

which is true as long as NI is higher than NC, which is in general the case, again.

In order to verify these relations graphically, we have fixed some variables to reasonable values (FAR = 0.1, FRR = 0.2, NC = 100) and have varied NI, the number of impostor accesses. Figure 3 shows the relation between the standard deviation of the underlying Normal distributions and the ratio between NI and NC. As expected, the higher the ratio $\frac{\text{NI}}{\text{NC}}$, the bigger the difference between the standard deviation of the Normal distributions related to the three statistical tests. Moreover, we see that the standard deviation of the proposed **HTER** distribution stays close to the one of the **FRR** distribution, which is mostly influenced by NC, the number of client accesses, and does not decrease with the increase of NI, contrary to the two other solutions. Since the size of the confidence interval is directly related to the standard deviation, this Figure essentially shows that

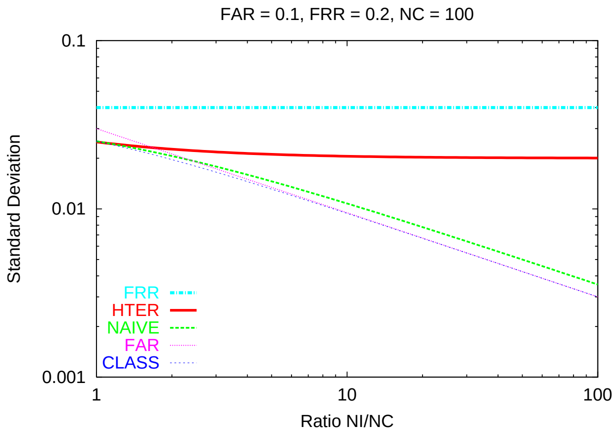

Fig. 3. Standard deviation of the Normal distributions underlying the three different choices of distributions for a statistical test on HTERs. Also shown: standard deviations of both the **FAR** and **FRR** distributions. All curves are in log-log scale. The order in the legend corresponds to the order of the curves at the right of the figure.

the confidence interval computed using the proposed technique will always be larger than that of the two other techniques. Hence two verification methods yielding two different HTERs could easily be considered statistically significantly different using one of the methods described in section 5, while they would not be considered statistically significantly different using the proposed technique. In fact, the Figure shows that the confidence interval is directly influenced by the minimum of NC and NI and not their sum.

In the next two subsections, we present two real case studies where the use of the proposed statistical test would have yielded a different conclusion.

6.2. Empirical Analysis on XM2VTS

In the first case, the well-known text-independent audio-visual verification database XM2VTS [7] was used. In this database, the test set consists of up to 112000 impostor accesses and only 400 client accesses, for a total of 112400 accesses. In a recent competition [8], several models were compared[6] on a face verification task and we will look here at the results of the best model, hereafter called *model A*, and the third best model, hereafter called *model B*, apparently significantly worse. Table 1 shows the difference of performance in terms of HTER between models A and B. Having up to 112400 examples, one could indeed expect the difference between the two models to be statistically significant.

Method	FAR (%)	FRR (%)	HTER (%)
Model A	1.15	2.50	1.82
Model B	1.95	2.75	2.35

Table 1. HTER Performance comparison on the test set between models A and B when the threshold was selected according to the Equal Error Rate criterion (EER) on a separate validation set.

δ	HTER eq (13)	NAIVE eq (18)	CLASS eq (17)
90%	1.285%	0.131%	0.105%
95%	1.531%	0.156%	0.125%
99%	2.013%	0.206%	0.164%

Table 2. Confidence intervals around results of model A, computed using three different hypotheses (and their respective equation).

Table 2 shows the size of the confidence intervals computed around the result (using HTER or the classification error) obtained by model A for the three methods for three different values of δ (90%, 95% and 99%). As we can see, for all values of δ, the size of the interval is about one order of magnitude larger for the proposed method than for the two other methods.

	HTER DEP, eq (16)	HTER INDEP, eq (15)	NAIVE eq (18)	CLASS eq (17)
δ	69.2%	64.7%	100.0%	100.0%
σ	0.0052	0.0057	0.0006	0.0005

Table 3. Confidence value δ on the fact that model A is statistically significantly different from model B, according to their respective performance (HTER or classification error), and computed using four different hypotheses (and their respective equation). For each method, we also give σ, the standard deviation of the corresponding statistical test.

Table 3 verifies whether the HTER obtained by model A gives statistically significantly different results than the one obtained by model B, using the *two-sided* test of equation (6) for the independent cases and (10) for the dependent case. According to both proposed HTER method (independent and dependent cases), both models are equivalent (the confidence on their difference is much less than, say, 90%),

[6] While this is not the topic of this paper (since it should apply to any data/model), people interested in knowing more about the problem tackled in this case study are referred to [8]; we used results of the models of IDIAP and UniS-NC on the automatic registration task, using Lausanne Protocol I. Furthermore, note that the results of UniS-NC are slightly different from those published in [8], but correspond to the list of scores provided by one of the authors of the method.

while according to both other methods, the models would be different (with 100% confidence!). Remember that there was only 400 client accesses during the test, hence it is reasonable that only one error on these accesses makes a visible difference in HTER while it cannot seriously be considered statistically significant. This is well captured by our technique, but not by the other ones. Moreover, in this case, the dependence/independence assumption did not have any impact on the final decision.

6.3. Empirical Analysis on NIST'2000

In the second case, the well-known text-independent speaker verification benchmark database NIST'2000 was used. Here, the test set consists of 57748 impostor accesses and 5825 client accesses, for a total of 63573 accesses. We compared the performance of two models[7] hereafter called *models C* and *D*. Note that, while on XM2VTS the ratio between the number of impostor and client accesses was very high (280 times more), for the NIST database, the ratio is more reasonable, but still high (around 10).

Method	FAR (%)	FRR (%)	HTER (%)
Model C	13.1	9.6	11.4
Model D	15.8	7.8	11.8

Table 4. HTER Performance comparison on the test set between models C and D when the threshold was selected according to the Equal Error Rate criterion (EER) on a separate validation set.

δ	HTER eq (13)	NAIVE eq (18)	CLASS eq (17)
90%	0.676%	0.414%	0.436%
95%	0.805%	0.493%	0.519%
99%	1.058%	0.648%	0.682%

Table 5. Confidence intervals around results of model C, computed using three different hypotheses (and their respective equation).

We now present the same kinds of results as for the XM2VTS case. Table 4 shows the difference of performance in terms of HTER between models C and D; Table 5 shows the size of the confidence intervals computed around the result obtained by model C; as we can see, given a ratio of impostor and client accesses around 10 instead of 280, the difference between all the confidence intervals is less drastic but still exists; Table 6 verifies whether the HTER obtained by model C gives statistically significantly different results than the one obtained by model D. For each test, we show both the confidence value δ and the standard deviation σ of the corresponding statistical test.

	HTER DEP, eq (16)	HTER INDEP, eq (15)	NAIVE eq (18)	CLASS eq (17)
δ	98.8%	89.1%	98.9%	100.0%
σ^2	0.0016	0.0028	0.0018	0.0019

Table 6. Confidence value δ on the fact that model C is statistically significantly different from model D, according to their respective performance (HTER or classification error), and computed using four different hypotheses (and their respective equation). For each method, we also give σ, the standard deviation of the corresponding statistical test.

As it can be seen, in the DEP case, σ is very small, even smaller than the NAIVE and CLASS solutions, hence obtaining a very high confidence that the two models are different. In order to explain this unexpected result, note than none of the tests take into account the possible dependence existing between the compared *models*. Indeed, if the two models are based on the same technique (which is often the case; for instance, in speaker verification, most systems are often based on Gaussian Mixture Models, but trained with slightly different assumptions), then both systems will have a natural tendency to answer very correlated scores on the same example. In the case of the two models trained on the XM2VTS database, they were very different (one was based on a Gaussian Mixture Model, while the other one was based on Linear Discriminant Analysis and Normalized Correlation); while for the models trained on the NIST database, both were in fact variations of Gaussian Mixture Models, hence are probably very correlated. Unfortunately, there exist no test that take this dependency into account. Hence, for instance, the variance $\frac{p_{AB}+p_{BA}}{n}$ of equation (9) will be quickly very small simply because the models are correlated (and not just because the examples are the same). Using this equation will thus result in an underestimate of the true variance when models are very correlated, as empirically shown in Table 6.

On the other hand, the INDEP case does not take into account the dependency between the data, but somehow it is reasonable to expect that the effect of this error may be balanced by the fact that it does not take into account the dependency between the models neither. The correct solution probably lies somewhere between these two solutions, hence, one should probably favor the most difficult test so as to only assess statistical differences when both tests agree on this fact (hence, here, with only 89.1% confidence).

[7] Once again, while this is not the topic of this paper, people interested in knowing more about the problem tackled in this case study are referred to [9].

7. CONCLUSION

In this paper, we have proposed a proper method to compute statistical tests on aggregate measures such as HTER or DCF often used in person authentication. We have also shown why using other approximations such as tests on the classification error instead would result in over-optimistic decisions. We have given some empirical evidence using two benchmark databases. It is important to note that the test of two proportions is not the ultimate statistical test and there exist other tests that are known to be sometimes more appropriate for classification tasks (such as complex cross-validation techniques for instance [10]). However, none of these tests have so far addressed the problem of dependence between the tested models. Nevertheless, an important finding of this paper is that when people design new databases for person authentication, they should keep in mind that it is probably not worth having a huge unbalance between client and impostor access numbers, since the statistical significantness of the results will mainly depend on the smallest of these two numbers (providing equal costs for false acceptances and false rejections).

8. ACKNOWLEDGMENTS

This research has been carried out in the framework of the Swiss NCCR project (IM)2. The authors would like to thank Ivan Magrin-Chagnolleau for suggesting the problem, and Mohammed Sadeghi for providing the scores of Model A.

9. REFERENCES

[1] P. Verlinde, G. Chollet, and M. Acheroy, "Multimodal identity verification using expert fusion," *Information Fusion*, vol. 1, pp. 17–33, 2000.

[2] A. Martin and M. Przybocki, "The NIST 1999 speaker recognition evaluation - an overview," *Digital Signal Processing*, vol. 10, pp. 1–18, 2000.

[3] G. W. Snedecor and W. G. Cochran, *Statistical Methods*, Iowa State University Press, 1989.

[4] J.L. Wayman, "Confidence interval and test size estimation for biometric data," in *Proceedings of the IEEE AutoID Conference*, 1999.

[5] J. Koolwaaij, *Automatic Speaker Verification in Telephony: a probabilitic approach*, PrintPartners Ipskamp B.V., Enschede, 2000.

[6] A Martin, "Personal communication," 2004.

[7] J. Lüttin, "Evaluation protocol for the the XM2FDB database (lausanne protocol)," Tech. Rep. COM-05, IDIAP, 1998.

[8] K. Messer, J. Kittler, M. Sadeghi, S. Marcel, C. Marcel, S. Bengio, F. Cardinaux, C. Sanderson, J. Czyz, L. Vandendorpe, S. Srisuk, M. Petrou, W. Kurutach, A. Kadyrov, R. Paredes, B. Kepenekci, F. B. Tek, G. B. Akar, F. Deravi, and N. Mavity, "Face verification competition on the XM2VTS database," in *4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA*. 2003, Springer-Verlag.

[9] J. Mariéthoz and S. Bengio, "An alternative to silence removal for text-independent speaker verification," Technical Report IDIAP-RR 03-51, IDIAP, Martigny, Switzerland, 2003.

[10] T.G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1924, 1998.