Mono and Multi-Modal Biometric Systems Assessment by a Common Black Box Testing Framework

Antoine Cabana, Christophe Charrier

Mono and Multi-Modal Biometric Systems Assessment by a Common Black Box Testing Framework

Antoine Cabana and Christophe Charrier
Normandie Univ, ENSICAEN, UNICAEN, CNRS, GREYC, 14000 Caen, France
Email: {antoine.cabana, christophe.charrier}@unicaen.fr

Abstract

As a trending method for the authentication, biometrics tends to be integrated in various devices, and in particular in smartphones. If the evaluation is performed on operational device, the biometric sample and algorithm are not reachable by the assessors. So, these latter have to perform an evaluation on a system considered as a black box. This kind of evaluation implies numerous manual comparison.

This paper proposes a methodology to perform an evaluation of biometric black boxes. In order to obtain this methodology, two experiments were performed in order to determine an optimized conduct. Nevertheless, these experiments were realized with small test population, in order to assess the methodology a full scale evaluation was performed. This paper describes the used methodology to perform evaluation on black boxes systems, and the results obtained on the systems under test.

Keywords

Biometrics; performance evaluation; black box evaluation.

I. Introduction

The ubiquity of connected devices in every economic sector, and the broad range of applications that rely on it, such as smart speaker, smartphones, smart cars, online banking applications, to name a few, necessitates means to evaluate the security of such devices [1], [2], [3], since they transform the real world objects into intelligent virtual objects. Usually, to access embedded services, many of them are based on the use of at least one biometric modality, such as voice, fingerprint, face, etc. Once one has been authenticated, amongst the numerous available smart applications, some of them allow access to critical data or functionalities, such as e-payment, online blanking, professional data, privacy-relative informations, and so on. Consequently, the application providers may seek to ensure the suitability of the biometric authentication methods with these critical purposes [4], [5].

The main objective of the evaluation is to determine the performances of the considered system. Subsequently, these performances permit to ensure the suitability of the system with the envisioned uses during a certification process. As part of a certification scheme, the evaluation and the certification processes are generally performed by third parties, in order to ensure the authenticity of the results and decisions.

Three different methods exist to evaluate the performance of any biometric system, depending on the internal data access a tester has: 1) White Box testing, 2) Gray Box testing and 3) Black Box testing.

White Box testing provides a full access to intermediate data (*e.g.* captured samples, templates, matching scores, references) and permit to capture or inject some data (in particular samples captured beforehand). White box systems allow to perform the evaluation according to favorable methodology; *i.e.* the processes of capturing samples and the cross-comparison may be separated. This evaluation conduct permits to capture numerous biometric samples and to perform these captures through several sessions. Nevertheless, the methodologies proposed in the literature and standards are generally developed by academics or manufacturers.

Grey Box testing is testing technique performed with with the limited or some knowledge and understanding of the internal features, functionality or details of the embedded software.

Black box testing assesses a system solely from the outside, without the operator or tester knowing what is happening within the system to generate responses to test actions. A black box refers to a system whose behavior has to be observed entirely by inputs and outputs. If the evaluation is performed by a third party (as a part of a certification process, for instance), manufacturers may not provide a white box system, or the certification test plan may specify the system under test is an operational device. In that case biometric samples cannot be captured prior to the evaluation which does not allow to inject these samples as system's entries. In order to process the cross-comparison, each comparison and enrollment have to be performed manually; the biometric samples are directly provided by the test population. These limitations make the evaluation process particularly costly and long.

The purpose of this paper is to investigate how it is possible to evaluate the performance of any biometric systems when one has no access to internal data, that is the worst case. The idea is to formulate recommandations (in terms of population test size for example) for labs during certification processes. That way, we investigate how usual black box testing protocol can be adapted and we determine the minimum size of the population test needed to ensure a robust evaluation of biometric systems.

The aim of this paper is to present a methodology to achieve the evaluation of a black box biometric system. This paper is structured as follows. In Section II, we present existing methodologies published in the literature and standards. Section III summarizes some preliminary experiments. In section IV, we present and discuss the proposed protocol. The apparatus is also presented. In sections V and VI, the extension of the proposed protocol for multimodal biometric systems is detailed and discussed. This is followed by a concluding section.

II. Evaluating biometric systems

As alluded above, in order to ensure the suitability of a system with a given use case, the system under test has to be evaluated. This problematic has already been addressed in the literature where several solutions to evaluate a biometric system are described. Phillips *et. al.* [6] and Jain & Prabhakar [7] introduce the biometric evaluation and addresses its problematics.

A. Performance metrics

In order to qualify the inherent quality of a biometric system, different performance metrics have been defined. The definition of such metrics have been addressed in the literature and even have been subject to a standardization. Phillips *et. al.* [6] present two types of error admitted by the biometric authentication system : false reject and false alarm. These errors respectively corresponds to the wrongful reject of a genuine user, and to the abnormal acceptation of an unenrolled person. They characterize the failures admitted by the whole system and are based on the final decision (*i.e.* the binary choice to accept or reject the identity claim).

Mansfield *et. al.* present other metrics in [8], in order to investigate the performances of the sensor and algorithm outputs. The FTA and FTE are sensor specific error rates, which respectively address the system issues on the biometric data acquisition failure (Failure To Acquire) and biometric enrollment (Failure To Enroll). The FTE may denote a user's inability to be enrolled in the system, which may be due to a (temporary or permanent) incapacity to present the biometric modality, a biometric feature unable to produce samples of sufficient quality, and so on.

The algorithm performances are also measured by the FMR (False Matching Rate) and the FNMR (False Non Matching Rate). The FMR addresses the wrongful matching of a reference and a sample from different users, and is considered as the most serious of biometric security error. Whereas the

FNMR represents the algorithm tendencies to wrongly non-match a reference and a sample from the same genuine user. This error does not necessarily indicate a flaw in the biometric system

B. Evaluation protocol

Protocols to evaluate biometric systems described in the literature are numerous and cover different biometric features [9], [10], [11], [12], [13]. An evaluation protocol may be divided into two parts: 1) sample collection, 2) cross-comparison (this latter addresses the problematic of computing the matching scores and final decision given a set of samples).

The description of samples acquisition generally covers the description of the used sensors and of the apparatus to measure or control the environmental conditions (depending on the captured biometric feature, *e.g.* illumination for the face, noise for speaker recognition, temperature humidity and even illumination for fingerprints . . .). The population is also described, mainly in terms of age and gender.

In order to perform cross-comparison, some platforms have been developed such as EVABIO by Mahier *et. al.* [14], or the Finger Verification Competition [15], [16], [17], [18]. An algorithm is uploaded on these platforms, which then perform the matching using biometric databases. The obtained matching scores are used to compute the error rates, and the result are usually presented in the form of ROC or DET curves.

Given the size constraints of embedded systems, these fingerprint systems use small size sensors. Evaluation and study of partial fingerprints systems drawbacks have been address by Fernandez-Saavedra et. al. [19], where the impact of sensor size on the systems performances have been investigated. Roy et. al. [20] addressed this problem too, studying the vulnerabilities of partial fingerprint authentication systems and investigating the possibility of generating a ”master print” (*i.e.*, a synthetic or natural partial print able to match one or several templates).

However in the case of operational devices, these two alluded above processes cannot be separately applied. In response to the need of evaluating these black box systems (*e.g.*, biometric authentication process on smartphones), two experiments and one evaluation have been performed.

In our case of study, three assumptions have been formulated:

- We investigate how it is feasible to evaluate an operational device without any access to internal data where only final results (granted or refused) are available,
- The population test is not necessarily familiar with biometrics,
- The evaluation protocol has been designed to be consistent with the recommendations provided by the ISO 19795-1 and 19795-2 [21], [22], Mansfield and Wayman [23].

III. Preliminary experiments

Prior to the protocol definition, two experiments are conducted in order to ensure the feasibility of black box evaluation. These experiments also permit to determine the difficulties and constraints of such processes, and do not aim to obtain exploitable results. Consequently, the test crew involved in these pilot evaluations are minimal (a dozen people constitutes the test population in each experiment). However in each experimentation, as mentioned before, the test crew is not a familiar with biometrics systems evaluation; the main drawback is to conciliate participants' schedule and the conduct of the test evaluations.

A. Common evaluation conduct

These two experiments implement a common conduct in order to perform the evaluation. Indeed these test evaluations were performed on the same operational device : a biometric system embedded into a smartphone. Given the system under test was an operational device, access to the biometric authentication process was not granted. This biometric device can be considered as a black box which does not allow to separate the capture and comparison parts.

In order to address this issue, a generic evaluation conduct is designed. This conduct is split in as many round as people in the test crew. Each round consists of enrolling one person, and then the remaining test crew attempts to perform impostures. Then, after a while, the enrolled person tries to perform genuine transactions. Obviously, the enrolled person changes at each round, and consequently this specific round is performed for each person in the test crew.

The obtained results for the impostor and the genuine attempts are recorded separately. This separated storage ensures to avoid enumeration error while computing the system performances.

As the trial system does not permit to access to the internal data of the biometric recognition process, it is not possible to store the samples in order to replay the comparisons, or to access to the matching scores to plot ROC or DET curves. Since the only exploitable information is the final decision provided by the system and the ground truth known by the evaluator, the performances of a biometric black box are finally defined as an operating point.

Population description: The collection protocol to design the population test crew is the following: 1) the collection has been performed at different times, 2) in different environments, 3) with different devices and 4) controls have been done to control data integrity. The protocol yields us to estimate the minimum number of involved subjects, as well as the distribution of age and gender.

To achieve the alluded above evaluation experiments, two similar population have been designed. Their similarity is mainly due to the important overlapping. This overlapping comes from people's availability to perform this experiment. The two populations involved in these experiments are detailed in table I.

Even if the test crew is quite small (a dozen persons) to ensure the obtained performances on the biometric devices, those tests evaluations permit to estimate the required time to perform a complete biometric evaluation.

TABLE I

POPULATION DESCRIPTION INVOLVED INTO THE TWO PRELIMINARY EXPERIMENTS.

Test crew	First experiment	Second experiment
Total number of testers :	12	13
Age range :	25-60	20-60
Age range representation :		
20-30	1	2
31-40	2	2
41-50	7	7
51-60	2	2

B. First experiment: roaming protocol

The first experiment aims 1) to determine the feasibility of the black box evaluation process, and 2) to estimate the required time to complete the evaluation process. Given the compromise between the evaluation and people's schedule we need to address, the chosen trade-off was to avoid the testers moving. Consequently, the system under test has been brought to each people office. This specific conduct leads to the following two constraints:

1) Since the evaluation has to be performed in different rooms, the environmental conditions may vary between each of them. And given the latency of the used thermometer, and in order to permit a quick result collection, the recording of the environmental information has been withdrawn. That way, we assume that the environmental conditions do not significantly influence results.
2) The result recording is performed using fillable forms in order to reduce the time needed to perform the test. The supervisor is in charge of filling the form while the current tester is running the

biometric presentations. However, the main noticeable drawback is that the supervisor may not efficiently observe the interaction of the tester with the system under test.

This first experiment allows to determine the constraints and process's improvements. We deduce that the main possible improvement is the replacement of the fillable forms by a dedicated software to record the results of biometric comparisons.

C. Second experiment: assisted evaluation

This second experiment implements a different test protocol. Given the difficulties to collect users' feedback as well as observations of the interaction between the tester of the biometric system, the implemented protocol aims to collect the comparison results as well as observations. These comments consists of users' feedback, error messages and users' behavior.

In order to perform this experiment including the proposed improvement, result recording processes are achieved by a dedicated software. This dedicated software permits to record results (granted transaction, rejected transaction, or acquisition failure) into a database which is designed to store the relevant information of the evaluation. In this database, each result is linked to the testers (both those who perform the enrollment and those who perform the genuine or impostor attempt) and the involved biometric feature is also registered.

Given the objective to improve the collection of relevant observation, the recording software was handled by the test crew under the supervision of the person in charge of the evaluation. Consequently, the supervisor was able to record comments on the user behavior, on the user interaction with the system, and so on, while the testers record theirs own results. As a consequence and on the contrary of the first experiment, this evaluation is performed in a dedicated room.

Yet, performing the evaluation in a single room presents pros and cons. People have to move to the dedicated room for the evaluation. They need to finish or hold their current task in abeyance, which may imply people can be distracted. But on the other hand, the environmental conditions have been recorded easily due to the lack of changes.

This new environment setup permits to perform the evaluation while collecting information on the interactions of the testers with the system under test. The collected information brings additional feedback on the system used. Past the first rounds, the collected feedback is quite redundant and its collection is mostly useless. The comments address the interaction/positioning of the tester modality on the sensor, the modality condition (*i.e.* if a scar or a wound affects the modality, or the wetness/dryness . . .) as well as difficulties encountered during the experiment (*e.g.* problematic fingers, explanations to numerous FTA-case error, . . .).

IV. Protocol presentation

Given the black box systems constraints, the proposed methodology is subdivided in two sub-methodologies in order to respectively estimate the FAR and the FRR values.

Similarly to the previous experiments, this protocol uses as guideline the recommendations and requirements presented in the ISO 19795-1, 19795-2 and Mansfiel and Wayman [21], [22], [23].

A. Systems under test

In order to perform this evaluation, several biometric recognition systems have been selected. The first column of Table II presents the five trial biometric systems under test.

These systems are embedded in smartphones as authentication solution, none of them allow to handle internal data. This section proposes to describe each of the five systems[1] used in FAR evaluation part. Only the first three have been used in the FRR part.

All the trial systems allow five presentation before requiring a password to be unlocked.

Some technical characteristics of the trail systems are listed hereafter:

- **System A:** the fingerprint sensor is embedded into the front facing home button. The sensor dimensions are 10 mm in length and 4 mm wide, for a sensing surface of 40 mm^2
- **System B:** the fingerprint sensor is embedded in a back facing position. The sensor dimensions are 8 mm side length, for a sensing surface of 64 mm^2
- **System C:** the fingerprint sensor is embedded into the right side facing sleep button. The button dimensions are 14 mm in length and 3.5 mm wide, for a maximal sending area of 49 mm^2. This system implements a different handling of acquisition failure.

  This system implement a different management of failures acquisition compared to the System A and B. Indeed, when the other systems consider FTA case as a valid biometric transaction, this system does not provide any feedback, and considers any FTA case as a valid biometric transaction.
- **System D:** the fingerprint sensor is embedded into a back facing button. This device is only used to estimate the FAR value.

[1] The systems under test have been anonymized, the evaluation has been performed without the manufacturer's consent or participation

- **System E:** the fingerprint sensor is embedded into the front facing home button This device is only used to estimate the FAR value.

The second column of Table II shows the eligible fingers for the collection of genuine attempts for each set of system depending of the position of their sensor. Considering the different sensor positions, we define the eligible fingers for the collection of the genuine as well as impostor attempts in order to avoid inconsistent results:

- Front facing sensors (System A and E): usually, this placement implies that sensor is embedded in the home button, at the bottom of the device. The thumb and index finger of both hands are especially suitable for the presentation kinematic.
- Back facing sensors (System B and D): usually, the sensor is placed at the center in width and at higher part of the length. In this case, the index and middle finger of both hands are suitable for the presentation kinematic.
- Right side facing sensors (System C): the sensor is placed at the center in length and width of the side. Given the sensor lateral position, the eligible fingers of both hands are different. For the right hand, the thumb and index finger are the more suitable. For the left handed persons, the index and middle finger are the more suitable.

B. Objectives of the proposed protocol

The proposed protocol mainly aims to estimate a biometric system's performances resumed here as an operating point calculated both FAR and FRR values.

1) Estimation of the false acceptance rate: Given the difficulty to perform as many comparisons as white box evaluation and the possible low probability of imposture error on tested systems, the main objective of this protocol is to ensure a superior threshold of this error rate. According to [21], the "*Rule of 3*" addresses the problem of determining the lowest error rate which can be estimated given N independent identically distributed comparisons. The lowest error rate is expressed, for a 95% confidence level, as :

$$p \approx \frac{3}{N} \tag{1}$$

where N is the number of performed comparison in the evaluation.

The protocol is designed in order to ensure the imposture probability is not over a defined threshold, with a 95% confidence interval. in this study, $p = 10^{-4}$ in order to determine if the system may be

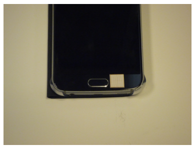
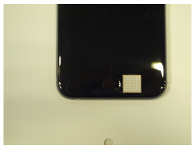
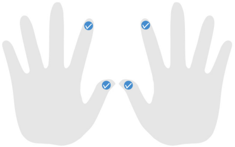
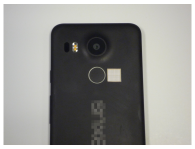
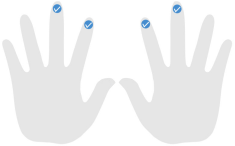
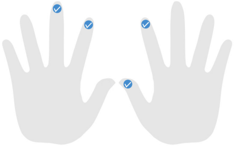
Systems under test	Eligible fingers
 System A  System E	
 System B System D	
System C	

TABLE II

The five trial systems under tests and their corresponding fingers for the collection of the genuine attempt . System A and E have their fingerprint sensors embedded into a face button, System B and D have their fingerprint sensors embedded into a back button, and System C has its fingerprint sensor embedded into the side button.

considered as secure as a 4-digit PIN-based system. This detection limit as been chosen in order to ensure the discriminating capacity of the system is at least equivalent to the probability of brute forcing a 4-digit PIN. The motivation of this choice is to ensure the suitability of a biometric system with payment use cases.

From Eq. 1, $N = 30,000$. This means that at least $30,000$ tests have to be performed to ensure an upper bound of the False Acceptance Rate lower than 10^{-4}.

2) Estimation of the false reject rate: The method used to estimate the False Reject Rate and the correspondent confidence interval is presented in the ISO 19795-1 [21]. The confidence interval is formulate as :

$$CI = \hat{p} \pm z(1 - \frac{\alpha}{2})\sqrt{\hat{V}(\hat{p})} \tag{2}$$

where $z()$ is the inverse of the standard normal cumulative distribution, α is the probability that the confidence interval does not contain the real value of the FRR, $\hat{p}$ is the observed error rate and $\hat{V}(\hat{p})$ is the estimated variance of the observed error rate.

$$\hat{p} = \frac{1}{m \times n} \sum_{i=1}^{n} a_i \tag{3}$$

$$\hat{V}(\hat{p}) = \frac{1}{(n-1)} \left(\frac{1}{m^2 \times n} \sum_{i=1}^{n} {a_i}^2 - \hat{p}^2 \right) \tag{4}$$

where a_i is the number of observed errors for the i^{th} tester, n the number of testers, m the number of attempts for each tester.

C. Envisioned evaluation conduct

The proposed protocol is adapted from the common protocol detailed in the experiment part III-A. Nevertheless due to the more ambitious objectives, some modifications have been done.

First, the implemented conduct splits the process to determine the impostor and the process related to genuine attempts. Each of them implements a different testing conduct, however they rely on a common part of the protocol.

1) Common protocol: As the tested devices permit to enroll four or five references at the same time and in order to optimize the evaluation process, the references are grouped by three, each reference is provided by different users. Thus when a presentation is performed, three transactions are computed by the system under test. Indeed, this conduct assumes the system performs three authentication attempts ($1:1$ comparison) and not a single identification attempt among three references ($1:N$ comparison with $N=3$), when an authentication attempt is performed on a three references group.

According to Sanchez-Reillo *et. al.* [24], the biometric systems integrated in smartphones may be used under different stances. In this evaluation, the accepted stances are:

- tester is standing holding the device
- tester is standing, while the device is placed on a platform (for instance a standing table)
- tester is sitting holding the device
- tester is sitting, while the device is placed on a platform (for instance a table or a desk)

A dedicated software has been developed to be embedded in a tablet in order to perform the result recording. The evaluation is performed accordingly to the roaming protocol of the first evaluation III-B, and the evaluator records the result using the tablet.

2) Genuine attempts' conduct: This process is divided into several sessions, each corresponding to a given group of person. Each session is divided into four rounds, each corresponding to a different enrolled finger. For each round, references are created enrolling three distinct testers, each enrolls a finger coherent with the presentation kinematic (Cf. section IV-A).

At this stage, the testers perform genuine verifications with their enrolled finger. The resulting decisions are recorded into the database. Yet, performing the verification just after the enrollment may induce a bias. Indeed, the enrollment process on such biometric systems requires an important number of presentation (the average number is 20 presentations per tester). The enrollment process reconstructs the reference thanks to the extraction of the fingerprint. However the resulting model may contain un-enrolled area of the fingerprint. A bias may be induced by the tester acknowledgment of the non enrolled areas directly after the enrollment. To avoid this bias, a latency period is introduced between the enrollment task and the verification task. Given this conduct and due to the limitations of the system under test, the latency period is defined as a trade-off between the logistical constraints (evaluation time and people availability for evaluation) and the necessary latency period. The latency period is fixed to a two-hour delay, which also models a frequent use of the biometric authentication.

Figure 1 illustrates the used protocol to perform genuine verifications in order to compute the FRR value, and Fig. 2 presents the used protocol to avoid training effects from tester when they enroll their fingerprints, as alluded above.

3) Impostor attempts' conduct: This conduct is divided in multiple rounds, each round corresponding to a group of references. As explained above, the references are grouped by three and the testers are enrolled once and only once in this conduct.

Once the references have been created, non-enrolled testers proceed with impostor attempts. In order to close a round, a given (and sufficient) number of impostor attempts has to be performed, in order to achieve the objectives of the evaluation to perform at least $30,000$ biometric transactions (Cf. section IV-B1). Fig 3 shows the used protocol.

D. Test crew composition and collection policies

Given the protocol objectives, a suitable population has been constituted. Unlike the previous experiments described in section III, here the objectives are to provide relevant estimation of the FAR and FRR. Given constraints determined during the experiments, a trade-off has been done between the population size and the time and logistical constraints of the evaluation conduct.

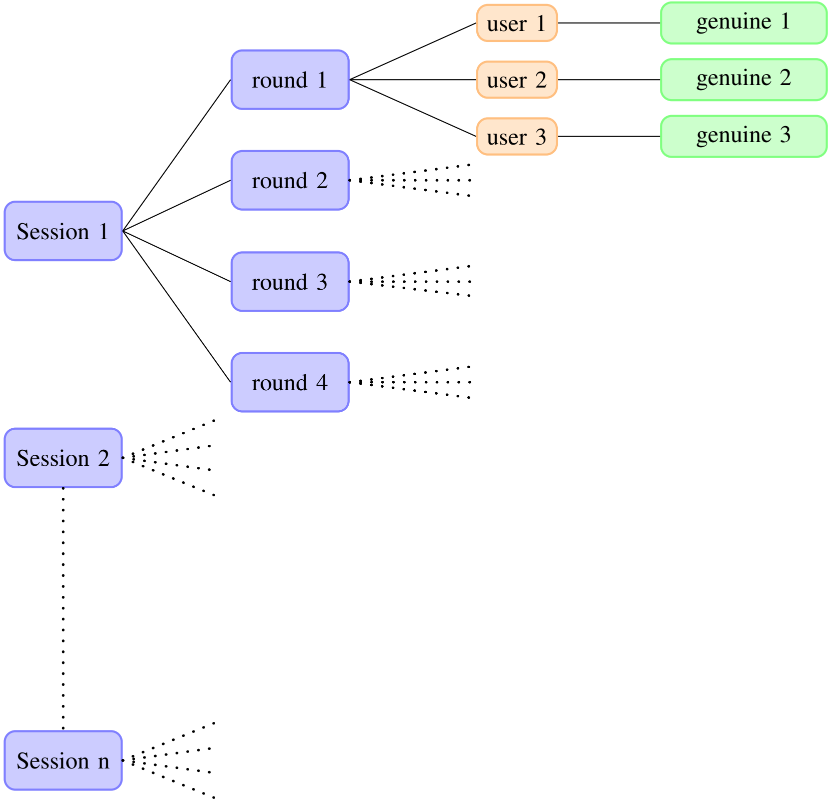


Fig. 1. Synopsis of the used protocol to perform genuine verifications in order to compute the FRR value.

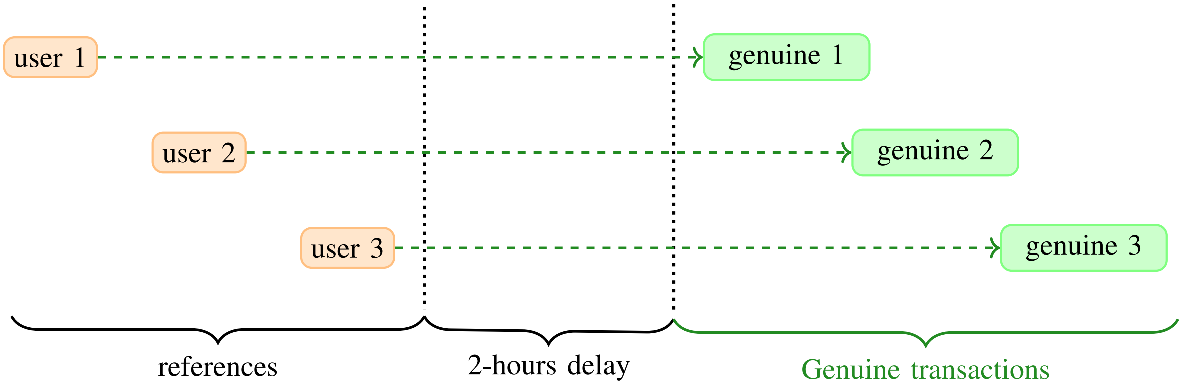


Fig. 2. Illustration of the applied latency period to avoid training of testers due to its acknowledgment of the non enrolled areas of the fingerprint directly after the enrollment.

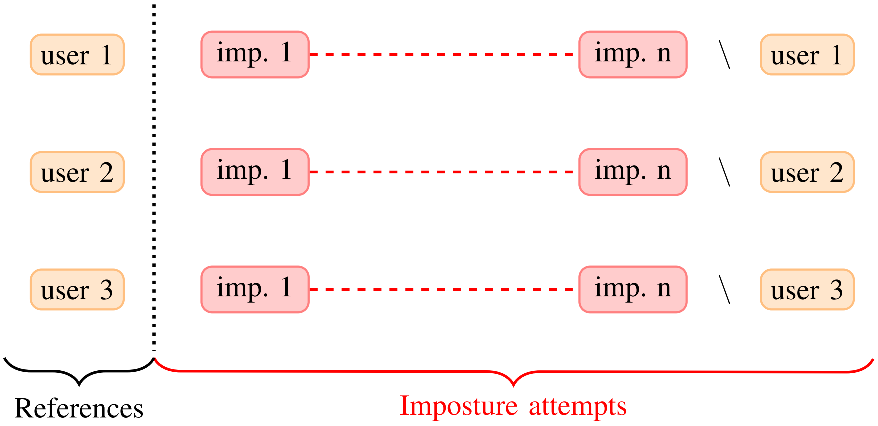


Fig. 3. Illustration of the Impostor attempts' conduct. $\forall i \in \{1, 2, 3\}$, User$_i$ do not participate to the corresponding imposture attempts.

1) Test crew planned and result collection policy: In order to ensure the aimed upper bound, at least $30,000$ impostures have to be performed during the evaluation (Cf. IV-B1). So the population size and the result collection policy have been studied in order to comply with the evaluation constraints and the relevancy of the results.

The main aim of a sample size calculation is to determine the number of participants needed to measure the performance of biometric systems. Usually, the number of people in a study is restricted due to cost and time considerations. However, if the sample size is too small, one may not be able to efficiently measure the performance of biometric systems under consideration, whereas samples that are too large may waste time, resources and money. It is therefore important to optimize the sample size. Moreover, calculating the sample size in the design stage of the study is increasingly becoming a requirement for many certification labs.

In order to calculate the sample size, it is required to have some idea of the results expected in a study. In general, the greater the variability in the outcome variable, the larger the sample size required to assess the performance of biometric systems. On the other hand, the more efficient the measure is, the smaller the sample size needed.

In this study, the primary outcome variable can be considered as binary variable (accepted or rejected) and the total number of subjects required (N) is defined as [25]:

$$N = 4z_\alpha^2 P(1 - P)/W^2 \tag{5}$$

where P is the expected proportion who have the characteristic of interest, W is the width of the confidence interval, and z_α is a value from the normal distribution related to and representing the confidence level (equal to 1.96 for 95% confidence, in our case).

In our study, $W = 0.14$ and $P = 0.96$. That way, from (5), $N = 30$ represents the minimum sample size required to assess the performance of the trial biometric systems.

a) Genuine attempts: As mentioned in section IV-C2, the biometric reference policy consists in enrolling four instances of fingers as references in the tested systems. Depending on the tested system, the fingers chosen as references have to be coherent with the defined eligible fingers (Cf. IV-A).

Consequently, the testers authenticate with the current enrolled finger. For each enrolled finger, the testers perform five attempts, each result is recorded in the result database.

The total number of performed genuine comparison is 600, for each system under test.

b) Impostor attempts: The policy for the collection of impostor attempts is to enroll a unique reference per user for each system under test. Testers are free to choose their reference as long as the chosen finger remains compatible with the eligible fingers presented in IV-A.

To comply with the evaluation objectives, at least $30,000$ tests have to be performed during the evaluation. The number of fingers used to perform the imposture attempts and the number of presentations have been adequately set. The impostors perform presentations with eight fingers: thumbs, index, middle and ring fingers of both hands. For each finger, the impostors have five attempts which are recorded in the result database.

The expected number of collected attempts is $32,400$ which allows a slight margin of error in the evaluation conduct.

2) Test population description: The result collection process is divided in two parts : 1) the collection of the results of impostor attempts and 2) of genuine attempts. Given their different conducts, the test population involved in each process slightly differs (overlapping each other).

As described in section IV-D1, the minimum size of the population test is 30. In this study, at least 33 persons participated to the evaluation. The structure of population is presented in table III.

E. Results

During the evaluation, the decisions provided by the trial systems were recorded into a database. Thanks to the recorded results, estimated error rate have been computed for each tested system.

1) Evaluation application: The evaluation has been performed along two months by one supervisor, with a 10-day stop due to a scheduled decrease attendance. The FAR estimation part has been achieved

TABLE III

EVALUATION'S POPULATION DESCRIPTION

Test crew	Genuine attempts collection	Impostor attempts collection
Total number of testers :	33	38
Age range :	20-60	20-60
Age range representation :		
18-34	9	12
35-49	17	19
50+	7	7

in the first part of the evaluation and then the FRR part has been processed in the second part. Prior to the evaluation, each tester has been trained in order to avoid misuse of the tested devices.

The training consists in an enrollment and some verification attempts immediately after. This short training permits to ensure that the testers became familiar with the use of the trial systems, and correctly perform the imposture attempts.

The FAR part has been achieved in 20 days, and 32,040 impostures attempts were performed and subsequently recorded. Due to some issues along the evaluation (mainly unscheduled decreased attendance), the expected number of impostor attempts have been barely unreached, however remaining in the margins of error allowed during the evaluation scaling. Indeed, the number of attempts is sufficient enough to achieve the objective of the evaluation (*i.e.* the number of collected transactions is over 30,000).

The FAR part achievement implies the population becomes familiar with the embedded biometric recognition.

2) FRR estimation: The FRR values of the three system under test have been determined during this evaluation. These three systems were selected due to their sensor positions: each system implements a different sensor positioning. These different sensor positioning allow to experiment the evaluation process on various interaction kinematics. The FRR result collection has been performed in a dozen days.

The estimated FRR error rates for each system under test are presented in the table IV. However, this performance metrics does not precisely highlight the overall population easiness to use the systems. In order to enhance the results comprehensibility and to balance the lack of performance curves (ROC or DET), two bar charts have been plotted.

TABLE IV

ESTIMATED FRR VALUES FOR EACH TESTED SYSTEM.

System under test	Estimated FRR	Confidence interval magins
System A	11.2%	±2.9%
System B	33.7%	±3.0%
System C	7.2%	±6.6%

Figure 4 presents the population distribution over different FRR ranges, and permits to highlight the easiness to use the system. On the basis of Doddington *et. al.*'s [26] definition of different category of users (in particular *goats*: which represents users difficulty to be recognized by their biometric feature), the user distribution has been studied across their personal FRR rate. In a similar way to CMC curves, the cumulated proportion of accepted user depending on their personal FRR value have been plotted in figure 5.

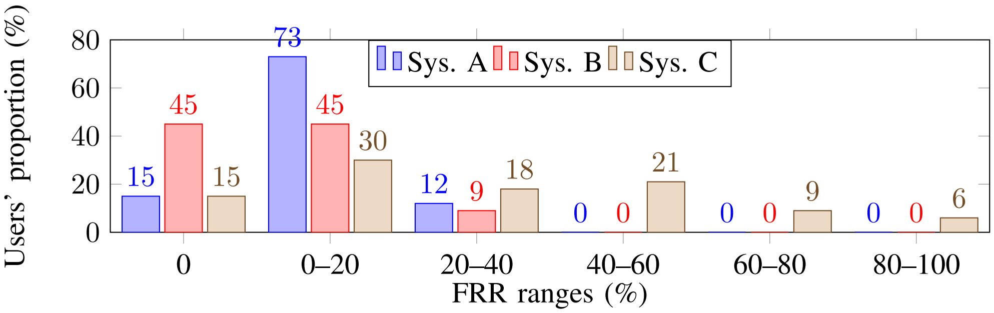


Fig. 4. Representation of the user proportion for different FRR ranges. This figure highlights the proportion of the population admitting a specific FRR range.

According to figure 5, the system A recognizes 88% of the population, while testers do not experience a personal false rejection rate (FRR) higher than 20%. Furthermore the subsequent 12% of the population admits a personal FRR between 20% and 40%.

According to figure 5, the system B is particularly efficient to recognize users as 91% of the population admits a personal FRR of 20% or less. As seen on figure 4, for 45% of the population, not reject has been observed across their presentations sessions, and another 45% of this population admits only an

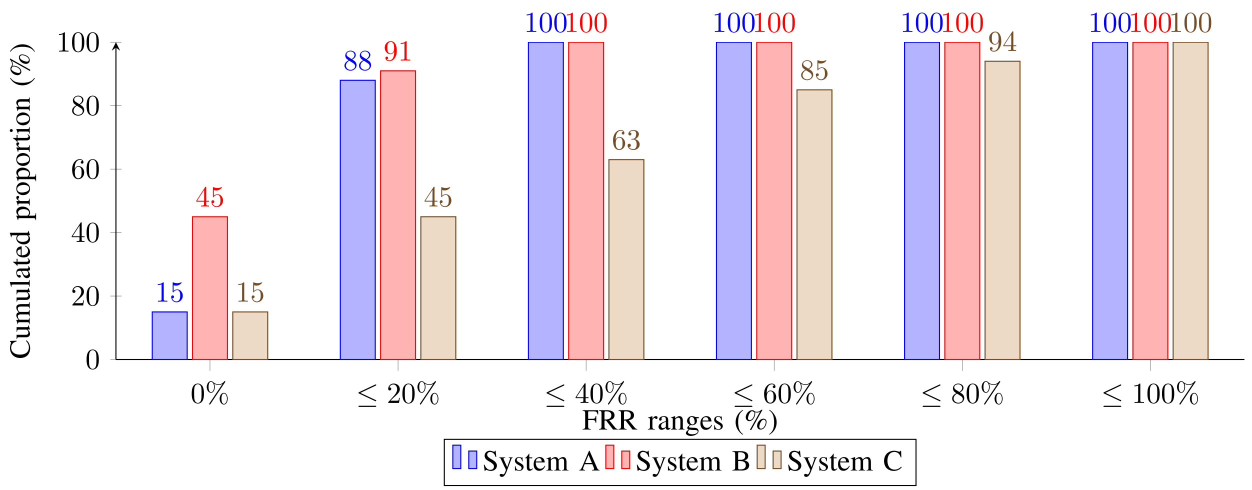


Fig. 5. This figure illustrates the cumulative proportion of accepted users depending on their observed personal FRR.

observed personal FRR of 20% or less[2]. This system only admits 9% of the population for a personal FRR included in $[20\%, 40\%]$. Consequently, this system appears as relatively easy to use.

According to the figure 5, the system C displays heterogeneous results. This system admits about 37% of the test population with a personal FRR higher than 40%, while the other tested systems do not.

According to figure 5, the systems A and B accept 100% of the genuine test population with a personal FRR lower than 40%. According to figure 4, for the system C, 37% of the test population admits a personal FRR higher than 40%.

3) FAR estimation: Along the whole evaluation, none successful impostor attempt was observed. In order to estimate the upper bound of the error rate, the "*rule of 3*" has been used. Given the number of tests performed, the estimated upper bound is $9.3633 * 10^{-5}$ with a 95% confidence level for all the five tested systems.

F. Remarks

This evaluation permits to estimate admitted error rates on several embedded biometric systems (five for the estimation of the FAR, and three for the FRR). Nevertheless, even if the evaluation process has been performed on a relatively small population (maximum 38 persons involved in the evaluation), the evaluation takes roughly two months (with a single full-time supervisor). An evaluation implementing

[2] Which roughly means this type of user would be accepted by the system at the first or second presentation.

a protocol separating the sample collection and the cross-comparison process is more time-efficient and achieves greater precision in error estimation. However, this type of methodology requires access to the algorithm and sensor output. The main drawback is that these methodologies are not applicable for operational devices.

The main improvement in term of processing speed up is to dedicate more than one supervisor to the evaluation (which implies the use of as many tested devices as full time supervisors). This conduct modification may be achieved by enrolling different testers as reference in the different devices, the evaluation is consequently conducted with multiple parallel rounds. However, in these experimentations, the population was not a dedicated population, which implies a more important availability for the evaluation process.

Since this evaluation was performed using a non-dedicated population, the main drawback is the introduction of a lag time due to the testers occupation. As things stand, the constitution of a dedicated test population seems not to be worthwhile. However the aforementioned improvement may render the recruitment of a test crew viable if sufficient devices are used in parallel (*i.e.* each tester in the population is regularly solicited throughout each working day).

The proposed evaluation does not offer any feedback on the population given the lack of captured samples. Consequently, the constituted population may bias the evaluation by its difficulty. In order to permit the bias estimation, the population has to be studied (*e.g.* with a reference systems which performances are known).

V. Extension to a multimodal system

The previously detailed black box protocol has been deployed to evaluate embedded biometric systems in wearable devices, such as smartphones. Yet, biometric systems under test were mono-modal system, based on the use of fingerprint.

To be able to evaluate multimodal biometric systems, the proposed protocol needs some improvements.

A. System under test

The trial biometric system is a prototype namely Multi-Biometric Authentication Device (MBAD). Figure 6 contains an image of the trial MBAD prototype. This wearable device allows a strong authentication of the holder. This authentication process is based on two biometric modalities: 1) fingerprint and 2) voice. The speaker recognition system is mainly based on the extraction of cepstral features from the voice signal, whereas the authentication process is based on the extraction of fingerprint features

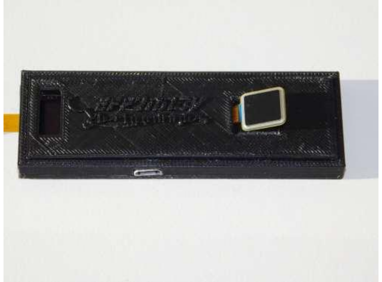


Fig. 6. The tested Multi-Biometric Authentication Device (MBAD) prototype

performed from a commercial application running like a black box. This application only allows users to get informations such as the choice of an operating point (each point is defined by its associated FAR value ranging from 1/50000 to 1/10000). In such a way, comparisons scores are not available. The final decision strategy is performed applying a cascade approach: the decision obtained from the fingerprint biometric module modulates the threshold level of the speaker recognition module. All functionalities were not embedded into the device, and is driven from an external PC. The advantage is that many comparisons with the same (voice sample, fingerprint sample) couple can be performed. The main advantage of such a process is to save evaluation time for people of the test population.

B. The used protocol

To respect the constraints of both the black box evaluation and the minimization down time of the test population, the detection limit of the FAR has been corrected. The final decision threshold has been fixed to 10^{-3}. From Eq. (1), the number of needed biometric transactions is $3,105$.

The acquisition process took place in a dedicated room, in which many environmental conditions were measured (temperature, noise level, humidity, etc.). Indeed, each condition can influence the acquisition performance of both the fingerprint and the voice.

C. Results

1) FAR estimation: During the evaluation process, no FAR case has been observed during the 3,105 imposture attempts.

The system can claim a FAR value lesser or equal to 10^{-3} with a 95% confidence interval. This result is not in contradiction with the white box evaluation performed at a larger scale, which allows the system to claim a FAR value lower than 0.0015%.

FRR value	lower bound	upper bound
45.39 ± 15.46%	29.93%	60.85%

TABLE V

COMPUTED FRR VALUE WITH ITS ASSOCIATED CONFIDENCE INTERVAL.

2) FRR estimation: The obtained results are shown in Table V. The white box evaluation of the system, reveals a FRR value equal to 8.28% for the system running with identical parameters. Those two evaluations do no allow to confirm or invalidate the performances in terms of false reject. Indeed, the black box evaluation tends to invalidate the performance observed when the system is evaluated using a white box test. Nevertheless, some remarks can be formulated, taking into account existing differences between the white box testing and the black box protocol:

1) For the white box testing, the samples capture is performed in one-shot. Thus, it is possible that due to a training effect, the obtained samples are of higher quality over the time. Considering the black box testing protocol, a delay between the creation of the reference samples and the genuine attempts is introduce (delay that does not exist for the white box testing). This can explain the observed difference between the results.
2) Each performance evaluation has been performed on two different prototypes of the MBDA. It has been observed that the prototype used during the black box evaluation suffered from a major drawback. A noise were added to the vocal sample by the sensor and could impact the performance of the speaker recognition algorithm.

VI. DISCUSSION

The proposed protocol yields to estimate biometric system performance with black box approach. During all experiments, the feasibility of this protocol has been highlighted. Yet, it is necessary to to implement a protocol validation process. Indeed, the obtained results may not be consistent with the results obtained under white box testing. Those bias may be induced by capture conditions (noise, temperature, humidity, etc.), test population (non representative population, heterogeneity of people, etc.), and have to be measured before the implementation of the protocol using a certification process. To do so, the proposed protocol has to be used to assess any biometric system for which the performances are known, in order to determine if the proposed protocol induced a biais or not. This experimentation allows to study the potential biais introduced during the samples capture and by the demography of the test population. Those potential biais will let us investigate recommendations for the implementation of such an evaluation.

The proposed protocol can be used for an information technology security evaluation for which users can specify security requirements for a particular security device in order to protect their personal data. Laboratories can evaluate such products and report results of conformance tests applying the proposed protocol. This will help to analyze and highlight possible threats to biometric verification systems introducing some vulnerability tests.

In addition, the proposed evaluation protocol will help certification lab to evaluate the performance of any biometric systems embedded in a smartphone, for example, for e-payment purpose, such a online purchase, or e-banking. In order to reduce the number of authentication based on the use of a PIN code, such applications would be unlocked using the embedded biometric sensor. Yet, major bank consortia, such as VISA or CB, require to assess to performance of the used biometric system guaranteed by a certificate issued by a trusted third party. As it is difficult and expensive to enroll more than one hundred people, the proposed protocol can help certification process with a minimum size of the population test, in a black box way.

Additionally to this protocol confirmation, the implementation of the white box testing will allow to study the test population, estimating and determining its potential bias. The white bow testing would concern a set of biometric sensors as well as a set of biometric systems for which the associated performances are given and known. Captured samples under this white box testing will help to analyze the consistency of the results obtained for each person of the test population. In addition, it should be possible to measure the imposture difficulty, identifying pairs of users with close modalities for which the biometric system (that serves as white box reference) yields to observe impostures. The obtained results based on the use of such a white box will permit to identify and understand obtained results using the proposed back box protocol.

VII. Conclusion

In this paper, a methodology dedicated to evaluate biometric systems under black box testing and two experiments have been investigated. The experiments allow to determine the issues pertaining to this kind of evaluation. The first experiment is a genuine implementation of a black box evaluation, aiming to simply obtain the error rates. The advantages of this method is its easiness to implement, and to be quickly executed. Nevertheless, the collection of additional information is hard, indeed the supervisor is already in charge of the result recording.

Whereas the second evaluation objectives were to provide both the error rates and additional information. However, this method requires to develop a dedicated test-tool, and increases the required time to perform. The development cost of the test tool counterbalances the efficient recording of results.The constituted database permits to store the results, and to efficiently ensure the result traceability.

The gained experience on these two experiments permits to performed a full scale evaluation in a reasonable time. This evaluation permits to estimate both false acceptance rate (FAR) and false rejection rate (FRR). The result database ensures the evaluation traceability, and repeatability to a lesser extent. In case of unusual results, the evaluator is able to replay the interaction, however the observed result at the first comparison occurrence may be different due to variation in the tester behavior or modality condition.

An extension of the proposed protocol has been investigated to a bi-modal biometric system based on the use of fingerprint and voice.

The protocol implemented in this evaluation allows to estimate the operating point of a biometric system functioning as a black box. Since the cost of such method is important both in term of time consumption and of human logistics, the sample size has been computed in order to determine the minimum number of people needed to ensure the robustness of results.

References

[1] C. Stergiou, K. E. Psannis, B.-G. Kim, and B. Gupta, "Secure integration of iot and cloud computing," *Future Generation Computer Systems*, vol. 78, pp. 964 – 975, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X1630694X

[2] M. Alhaidary, S. M. M. Rahman, M. Zakariah, M. S. Hossain, A. Alamri, M. S. M. Haque, and B. B. Gupta, "Vulnerability analysis for the authentication protocols in trusted computing platforms and a proposed enhancement of the offpad protocol," *IEEE Access*, vol. 6, pp. 6071–6081, 2018.

[3] A. Tewari and B. B. Gupta, "Cryptanalysis of a novel ultra-lightweight mutual authentication protocol for iot devices using rfid tags," *The Journal of Supercomputing*, vol. 73, no. 3, pp. 1085–1102, Mar 2017. [Online]. Available: https://doi.org/10.1007/s11227-016-1849-x

[4] B. Gupta, D. P. Agrawal, and S. Yamaguchi, *Handbook of Research on Modern Cryptographic Solutions for Computer and Cyber Security*, 1st ed. Hershey, PA, USA: IGI Global, 2016.

[5] B. B. Gupta, Ed., *Computer and Cyber Security: Principles, Algorithm, Applications, and Perspectives*, 1st ed. CRC Press, Taylor & Francis, nov. 2018.

[6] P. J. Phillips, A. Martin, C. L. Wilson, and M. Przybocki, "An introduction evaluating biometric systems," *Computer*, vol. 33, no. 2, pp. 56–63, 2000.

[7] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on circuits and systems for video technology*, vol. 14, no. 1, pp. 4–20, 2004.

[8] T. Mansfield, G. Kelly, D. Chandler, and J. Kane, “Biometric product testing final report,” *Computing, National Physical Laboratory, Crown Copyright, UK*, 2001.

[9] E. Bailly-Bailliére, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée *et al.*, “The banca database and evaluation protocol,” in *International conference on Audio-and video-based biometric person authentication*. Springer, 2003, pp. 625–638.

[10] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, “The feret database and evaluation procedure for face-recognition algorithms,” *Image and vision computing*, vol. 16, no. 5, pp. 295–306, 1998.

[11] D. Blackburn, J. Bone, and P. Phillips, “Frvt 2000 evaluation report, 2001,” 2001.

[12] P. J. Phillips, P. Grother, R. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone, “Face recognition vendor test 2002,” in *Analysis and Modeling of Faces and Gestures, 2003. AMFG 2003. IEEE International Workshop on*. IEEE, 2003, p. 44.

[13] P. J. Phillips, W. T. Scruggs, A. J. O’Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe, “Frvt 2006 and ice 2006 large-scale results,” *National Institute of Standards and Technology, NISTIR*, vol. 7408, no. 1, 2007.

[14] J. Mahier, B. Hemery, M. El-Abed, M. El-Allam, M. Bouhaddaoui, and C. Rosenberger, “Computation evabio: A tool for performance evaluation in biometrics,” *International Journal of Automated Identification Technology (IJAIT)*, p. 24, 2011.

[15] D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, and A. K. Jain, “Fvc2000: Fingerprint verification competition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 402–412, 2002.

[16] ——, “Fvc2002: Second fingerprint verification competition,” in *Pattern recognition, 2002. Proceedings. 16th international conference on*, vol. 3. IEEE, 2002, pp. 811–814.

[17] ——, “Fvc2004: Third fingerprint verification competition,” in *Biometric Authentication*. Springer, 2004, pp. 1–7.

[18] R. Cappelli, M. Ferrara, A. Franco, and D. Maltoni, “Fingerprint verification competition 2006,” *Biometric Technology Today*, vol. 15, no. 7, pp. 7–9, 2007.

[19] B. Fernandez-Saavedra, R. Sanchez-Reillo, R. Ros-Gomez, and J. Liu-Jimenez, “Small fingerprint scanners used in mobile devices: the impact on biometric performance,” *IET Biometrics*, vol. 5, no. 1, pp. 28–36, 2016.

[20] A. Roy, N. Memon, and A. Ross, “Masterprint: Exploring the vulnerability of partial fingerprint-based authentication systems,” *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 9, pp. 2013–2025, 2017.

[21] ISO, “Iec 19795-1: Information technology-biometric performance testing and reporting-part 1: Principles and framework,” *ISO/IEC, Editor*, 2006.

[22] ——, “Iso/iec 19795-2 biometric performance testing and reporting: Scenario testing,” 2007.

[23] T. Mansfield and J. L. Wayman, “Best practices in testing and reporting performance of biometric devices. on the web at www. cesg. gov. uk/site/ast/biometrics/media,” *BestPractice. pdf*, 2002.

[24] R. Sanchez-Reillo, D. Sierra-Ramos, R. Estrada-Casarrubios, and J. A. Amores-Duran, “Strengths, weaknesses and recommendations in implementing biometrics in mobile devices,” in *Security Technology (ICCST), 2014 International Carnahan Conference on*. IEEE, 2014, pp. 1–6.

[25] P. Mathews, *Sample Size Calculations: Practical Methods for Engineers and Scientists*. Mathews Malnar and Bailey, Inc., 2010.

[26] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, “Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation,” DTIC Document, Tech. Rep., 1998.