BEAT

Biometrics Evaluation and Testing

http://www.beat-eu.org/

Funded under the 7th FP (Seventh Framework Programme)

Theme SEC-2011.5.1-1
[Evaluation of identification technologies, including Biometrics]

D3.3: Description of Metrics For the Evaluation of Biometric Performance

Due date: 31/08/2012 **Submission date:** August 31, 2012
Project start date: 01/03/2012 **Duration:** 48 months
WP Manager: Julian Fierrez **Revision:** 1

Author(s): N. Poh (UNIS), C-H. Chan (UNIS), J. Kittler (UNIS), Julian Fierrez (UAM), and Javier Galbally (UAM)

Project funded by the European Commission in the 7th Framework Programme (2008-2010)		
Dissemination Level		
PU	Public	Yes
RE	Restricted to a group specified by the consortium (includes Commission Services)	No
CO	Confidential, only for members of the consortium (includes Commission Services)	No

D3.3: Description of Metrics For the Evaluation of Biometric Performance

Abstract:

The BEAT project aims to build a biometry-independent platform for Biometrics research, development and certification. By using this system, a users can easily compare results generated by several algorithms or different choices of parametrisation of the same algorithm with minimal interaction. The BEAT platform supports the evaluation of biometric systems with one or more biometric traits, including their fusion.

This document contributes to the above goal by reviewing the state-of-the-art methods in performance assessment and visualisation. Three common biometric tasks are described, i.e., verification, open-set identification (watchlist) and closed-set identification. For each task, two variants are possible, depending on whether the performance is assessed *a posteriori* or *a priori*. We have also considered the possibility of post-processing the system scores using cohort-based score normalisation techniques.

Contents

1 Introduction

Assessing the performance of a biometric system is a fundamental task in both research and application of the technology. This document discusses methodologies for testing a biometric system and reviews various performance statistics and charts that are involved in visualising . It is intended to allow the BEAT platform and its potential users to choose the most appropriate assessment methodology for their own application.

1.1 Evaluation Types

There are three main types of biometric performance evaluation with an increasing complexity as measured in terms of the number of uncontrolled variables, namely, technology, scenario, and operational. A thorough evaluation of a system for a specific purpose starts with a technology evaluation, followed by a scenario evaluation, and finally, an operational evaluation.

Technology evaluation aims at measuring the performance of a biometric system, typically, focusing on a particular system component. The evaluation is repeatable and usually short in duration compared to the other two evaluation types. In general technology evaluation is used to identify specific areas that require future research and to provide performance data that is useful when selecting algorithms for scenario evaluations.

The purpose of a *scenario evaluation* is to measure the performance of a biometric system operating in a particular application. The system to be tested includes the acquisition component (sensors). As a consequence, the test results are not always repeatable. Results from a typical scenario evaluation show areas that require additional system integration.

An *operational evaluation* is similar to a scenario evaluation except that the test is conducted at the actual site using actual end users, a subset of end users, or a representative set of subjects. Rather than testing for performance, which is often difficult, if not impossible, an operational evaluation aims to determine the impact of workflow due to the addition of a biometric system.

An ideal evaluation process consists of a technology evaluation that compares all applicable technologies that could conceivably meet requirements. While researchers use the performance data so obtained to identify areas for improvement, potential users use the results to determine the best system for their specific application. The performance data, together with the workflow impact data obtained from subsequent operational evaluations, will enable decision makers to develop a solid business case for potential installations.

As long as the BEAT project is concerned, we will focus on the technology evaluation, although the same platform can be used for the remaining two types of evaluation.

1.2 Types of Biometric Application

There are, in general, two types of biometric applications, namely verification and identification. It is useful to distinguish them here as they will have impact on the choice of performance evaluation.

Verification is the process of affirming that a claimed identity is correct by comparing the offered claims of identity with one or more previously enrolled templates. A synonym for verification is *authentication*.

Identification occurs when a biometric system attempts to determine the identity of an individual. A biometric sample is collected and compared to all the templates in a database. Identification is *close-set* if the person is assumed to exist in the database. In this case, the system must determine if the person is in the database. A *watchlist* task is an example of *open-set* identification whereby the person may not have been enrolled in the database. Another common task, namely, *negative identification* attempts to check that the person has not been enrolled in the database before. This task is useful to avoid duplicate identities in the same database.

Finally, *recognition* is a generic term that could imply either or both verification and identification. This term is generally avoided unless a broad coverage of both biometric applications is intended.

2 Performance Charts

2.1 Categorisation of charts

Several types of chart are commonly used in biometric performance analysis and evaluation. They can be categorised according to the type of application, namely verification or identification (the latter of which includes watch-list as a special case).

For biometric verification, two types of chart can be further distinguished, depending on whether the performance is assessed *a posteriori* or *a priori*. The main difference between these two types of performance evaluation methodology is that the former relies on a single matching score data set whereas the latter uses two score data sets: a *development set* that is used to establish a decision threshold based on a target criterion that one wishes to achieve, and an *evaluation set* that is used *uniquely* to assess the system performance using the established threshold. Figure 1 illustrates this idea.

For biometric identification, two types of chart are distinguished depending on whether or not the application is open-set or closed-set. Figure 2 summarises the different charts, as well as their corresponding metrics.

2.1.1 Verification

For verification, the metrics are:

- False Match Rate (FMR): an empirical estimate of the probability (the percentage of times) at which the system incorrectly declares that a biometric sample belongs to the claimed identity when the sample actually belongs to a different subject (impostor).
- False Non-Match Rate (FNMR): an empirical estimate of the probability at which the system incorrectly rejects a claimed identity when the sample actually belongs to the subject (genuine user).

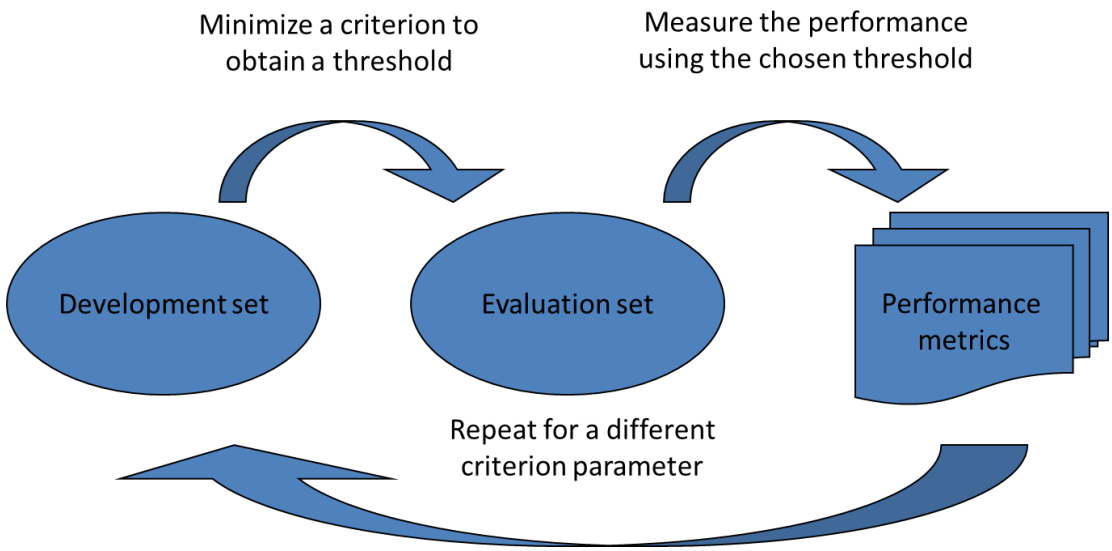


Figure 1: The two data sets used in an *a priori* performance evaluation methodology.

- Equal Error Rate (EER): The rate at which FMR is equal to FNMR.
- False Acceptance Rate (FAR) and False Rejection Rate (FRR): FAR and FMR are often used interchangeably in the literature, so as FNMR and FRR. However, their subtle difference is that FAR and FRR are system-level errors which include samples failed to be acquired or compared.
- True Acceptance Rate (TAR): It is defined as 1-FRR.
- Weighted Error Rate (WER): It is defined as the weighted sum between FNMR (FRR) and FMR (FAR).

The charts that are applicable to biometric verification are:

- Receiver Operating Characteristic (ROC): An ROC curve plots FNMR (in the Y-axis) versus FMR (in the X-axis), or FRR versus FAR. Alternatively, a ROC curve also plots TAR versus FAR.
- Detection Error Trade-off (DET) Curve: A DET curve is similar to ROC curve except that the axes are often scaled non-linearly to highlight the region of error rates of interest. Commonly used scales include normal deviate scale and logarithmic scale.
- Expected Performance Curve (EPC): While ROC and DET curves are *a posteriori* evaluation charts, an EPC is an *a priori* chart. It has three variants: (1) Use FMR as a performance criterion and report performance in FNMR; (2) Use FNMR as a performance criterion and report performance in FMR; (3) Use WER parametrised by β (see Section 1) as a performance criterion and report performance in HTER. In

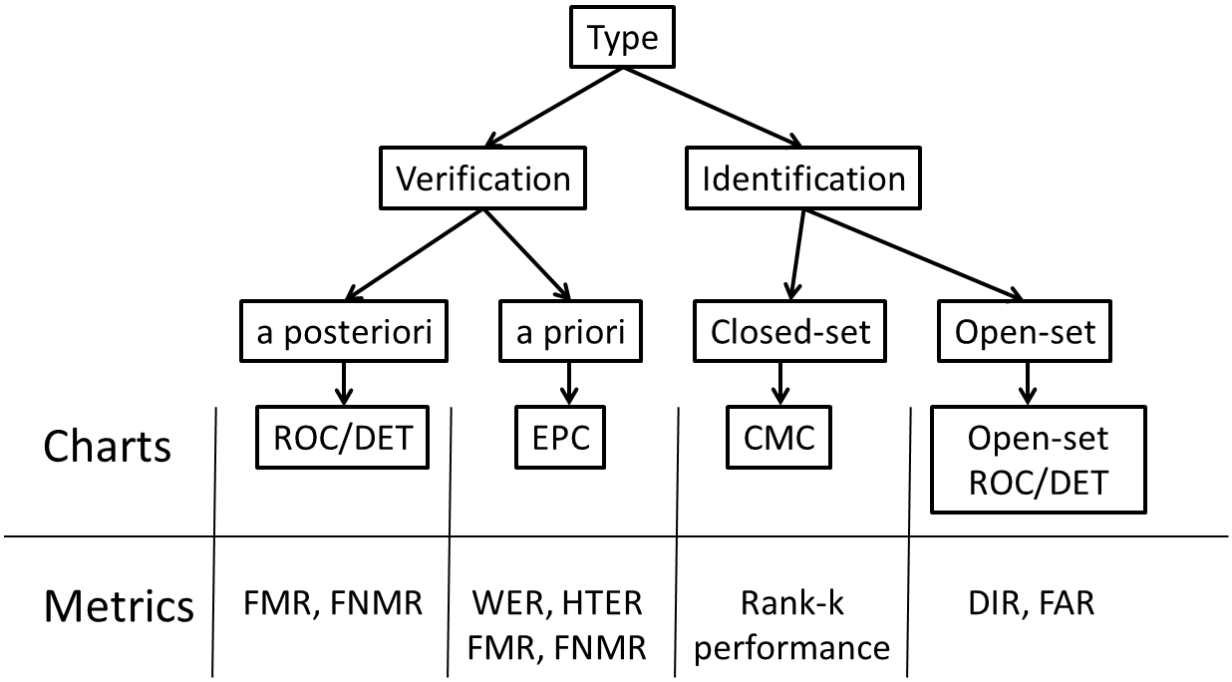
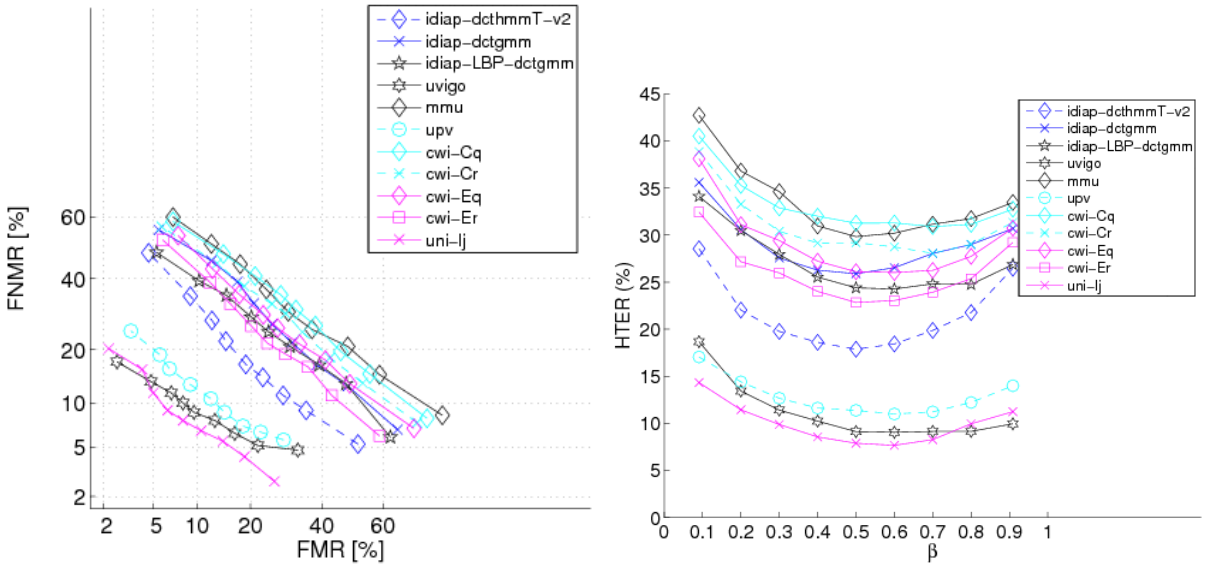


Figure 2: Classification of different biometrics performance charts.

usage 1, one plots FNMR as a dependent variable versus FMR which is an independent variable. In usage 2, the role of these two variables is interchanged so one plots FMR versus FNMR. Finally, in usage 3 which is the most common usage among all threes, one plots HTER versus the parameter α which is an adjustable variable that balances the cost or penalty incurred between false matches and false non-matches. The latter is, therefore, an independent variable.

Figure 3 shows some DET and EPC curves taken from the Face Video Competition presented in ICB2009 [1]. Each curve represents the performance of a face recognition system submitted to the Competition. The DET curves have been plotted using inverse normal deviate scale where small FMR and FNMR are emphasised. If both the match and non-match score distributions are Gaussian, the corresponding curve will appear as a straight line in DET. For DET, the best system is one that is closest to the lower left corner where FMR and FNMR are as small as possible. In this case, the system submitted by `uni-lj` is clearly the best system for most operating points being assessed. It is, however, beaten only by `uvigo`'s system at low FMR points. For EPC (panel (b)), a good system is one whose EPC is as low as possible. Again, we observe that the EPC of `uvigo`'s system is lower when β is closer to one. As will be explained, a higher β value penalises FMR higher. Since `uvigo`'s system is better when operating at low FMR values, its resultant EPC curve is lower than that of `uni-lj`'s. In this example, it can be clearly observed that both DET and EPC are consistent in reporting performance despite the fact that both charts use slightly different metrics: DET uses FNMR versus FMR whereas EPC uses



(a) DET (b) EPC

Figure 3: DET and EPC curves taken from the Face Video Competition presented in ICB2009 [1].

HTER versus β. However, this is not surprising considering that HTER is derived from FNMR and FMR.

2.1.2 Closed-set Identification

For closed-set identification, the Cumulative Match Characteristic (CMC) curve is commonly used. It is a plot of the *identification rate* at rank-k. A probe (or test sample) is given rank-k when the actual subject is ranked in position k by an identification system. Thus, a rank-1 outcome is considered a correct identification. The identification rate is an estimate of the probability that a subject is identified correctly at least at rank-k. Hence, identification rate is necessarily an increasing function of k. An example of CMC is shown in Figure 4.

2.1.3 Open-set Identification

There are two broad categories of architecture for open-set identification systems: exhaustive comparison and retrieval-based method. In the former architecture, the system compares a probe with all the all the gallery in the database. Face and iris identification systems are typically based on this approach because the computation involved in comparing a pair of sample is relatively small, or can be accelerated via parallel processes. In the latter approach, the system employs two or more cascaded subsystems, each of which acts as a filter, and typically each subsystem is more accurate but also computationally more costly than its precedent subsystem. Classical minutiae-based fingerprint systems are

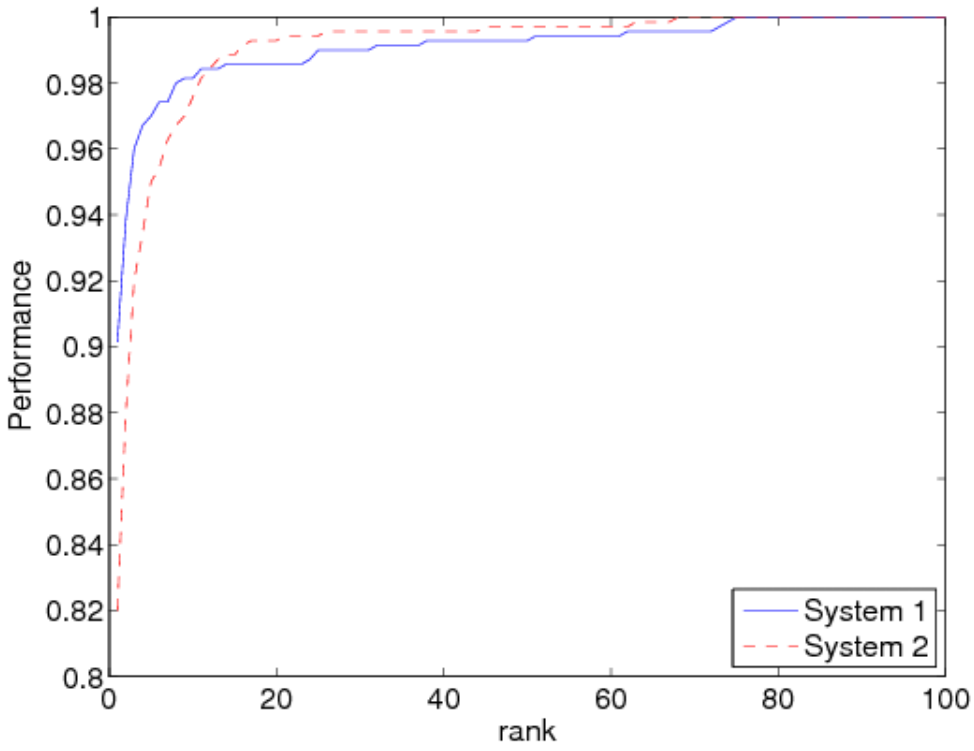


Figure 4: Identification performance of two systems reported on the AR face database using CMC. The gallery consisted of one image of 100 individuals. The probe set consisted of seven images for each of the 100 individuals. A better system is one that is closest to the top left corner of CMC. In this example, system 1 has a higher identification rate up to rank 12. Beyond this rank, system 2 is has a better identification rate. Hence, in this case, none of the two systems being compared is systematically better than another across all ranks. When the number of subjects is large, it is common to plot the X-axis in logarithmic scale.

based on this approach. The latter architecture is a generic methodology can potentially scale to extremely large gallery of subjects. In the subsequent sections, we will first discuss the performance evaluation methodology for a biometric identification with a retrieval-based architecture. However, the discussion is also applicable to systems that are based on exhaustive comparison.

In order to evaluate a retrieval-based open-set identification or watchlist biometric system, one has to consider the fact that the system operates with a threshold such that all samples with their corresponding similarity scores lower than the threshold will not be processed. A correct *detection and identification* implies that a probe must be detected and be subsequently processed. Although one can lower this threshold to allow more samples to be processed, doing so will increase the candidate list size returned by the identification system, hence, increasing the *false alarms*, i.e., the system alarms on the subjects in the watchlist when the probe does not belong to any of them. Two metrics are

used in evaluating an open-set identification system:

- Detection and Identification Rate (DIR): an estimate of the probability that a subject in the watchlist is detected.
- False Alarm Rate (FAR): an estimate of the probability an alarm is incorrectly sounded on an individual who is not in the database of a biometric system (watchlist).

Plotting DIR versus FAR produces a chart known as open-set ROC or DET. This chart is similar to a DET curve as shown in Figure 3(a).

In the discussions that follow, we will first introduce the notation and explain how different performance criteria and metrics are calculated.

2.2 Terminology and Notation

2.2.1 A Common Matrix Representation for Biometric Evaluation

In order to underpin the key variables, we shall first use a hypothetical example of a score matrix, as shown in Figure 5. This figure shows that there are two matrices, namely development and evaluation score matrices. Each matrix is further divided into two sub-matrices, namely, a left-hand side sub-matrix and a right-hand side sub-matrix. Both these sub-matrices are distinguished by its subject type, that is, whether the constituent subjects are in the gallery or the cohort set.

We shall use this matrix to store data required to compute metrics for verification, open-set identification, and closed-set identification (watchlist), each with the variant of carrying out (1) cohort-based score normalisation, and (2) *a priori* performance evaluation.

The left-hand side of the matrix contains a typical matrix of matching scores in which each column corresponds to the templates in the gallery and each row corresponds to a probe. Hence, a row of this matrix is a result of matching a probe with *all* the templates in the gallery.

In the right-hand side of the matrix, cohort subjects are added. Hence, a row of this right-hand side matrix corresponds to the scores of a given probe when compared to all the templates of these cohort subjects. The cohort scores are used for cohort-based score normalisation such as T-norm [2]. This technique has been shown to significantly improve the system performance for a number of biometric modalities, especially under strong mismatched conditions. If the cohort-based score normalisation is not required, then the matrix is simply empty.

Consistent with the *a priori* evaluation methodology, the development set is used to establish a threshold required in optimising a target performance metric whereas the evaluation set is used uniquely to estimate the system's generalisation performance (on unseen data).

In the *a posteriori* evaluation methodology, there would be only a single large matrix with $g_1, \ldots g_8$ as gallery subjects and $h_1, \ldots, h_{10}$ as cohort subjects, for this example. The single large matrix would have 12 rows indexed by the probes $p_1, \ldots p_{14}$. For example,

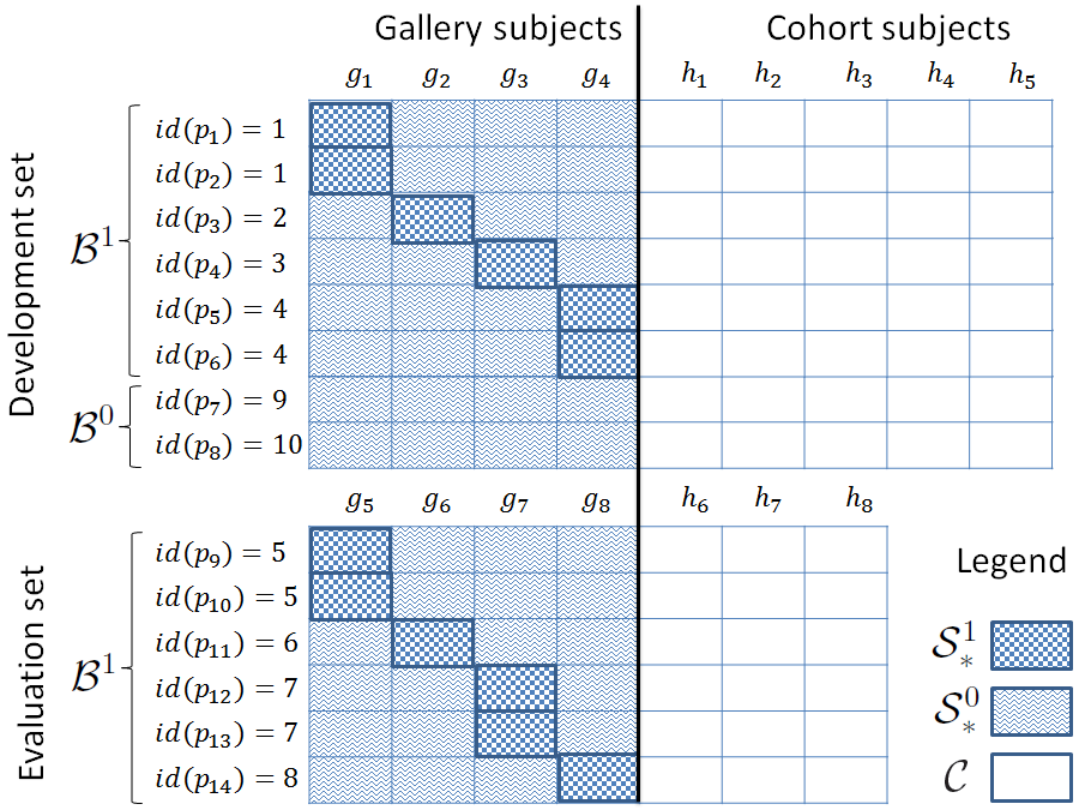


Figure 5: Categorisation of different visualisation methods and their performance metrics.

NIST challenge campaigns such as FRGC and MRGC use the *a posteriori* evaluation methodology.

The advantage of using the *a priori* evaluation methodology (with a development and an evaluation score matrix) is that there is the decision threshold is treated as a free parameter that needs to be optimised. Therefore, this is closer to real world applications. On the other hand, its disadvantage is that the data used for the development matrix do not contribute to the estimation of the performance. One way to mitigate this problem is to use a full matrix that is used in the *a posteriori* evaluation methodology and then bootstrap it to form a number of instances of a pair of development and evaluation matrices. For each bootstrap, one can then plot the performance. The ensemble of bootstrapped performance curves can be used to derive the expected or mean performance curve with confidence intervals. This forms the basis of deriving performance confidence intervals using the bootstrap approach. The confidence intervals around a DET curve has been reported in [3] and references therein.

2.2.2 Notation

We shall adopt the following terminology and notation:

- **Gallery, cohort, and probe sets**: Let $\mathcal{G}$ denote a *gallery* set; $\mathcal{C}$, a *cohort subject* set; and, $\mathcal{B}$, a *probe* set. The gallery consists of a set of enrolled users in a biometric database, contrary to the cohort set which contains all other subjects not enrolled in the database. The probe set may contain users who may or may not be present in the database. An element of the above sets is denoted by $g \in \mathcal{G}$, $b \in \mathcal{B}$, and $e \in \mathcal{E}$, respectively. g, b, and e are essentially biometric samples or features that are compatible with each other, that is, $g, b, e \in \mathcal{X}$, where $\mathcal{X}$ is a feature domain.

- **The identity function**: In order to refer to the identities, we use the function $id : \mathcal{X} \rightarrow \mathcal{U}$ where $\mathcal{U}$ is a set of unique identities. For example, $id(g)$ returns the identity (an integer) associated with the gallery. We use $\mathcal{U}_\mathcal{G}$ to denote the set of identities in the gallery; $\mathcal{U}_\mathcal{C}$, those in the cohort set; and, $\mathcal{U}_\mathcal{B}$, those in the probe set. For experiments in which cohort-based score normalisation is allowed, both the constituent identity sets are recommended to be disjoint, that is $\mathcal{U}_\mathcal{G} \cap \mathcal{U}_\mathcal{E} = \emptyset$. This is because the cohort subjects are often part of a design data set, whereas the probe set is an operational (test) data set.

- **Similarity score**: By comparing a gallery biometric sample g with a probe b, one obtains a similarity score $s = s(g, b)$. We shall use the convention that a larger similarity score indicates that the two biometric samples being compared are more similar. If a matcher produces a distance measure, one can invert the score, or multiply it by the minus sign.

- **User-specific class-conditional score set**: Let t be an identity in the gallery (t for template) and q be an identity in the probe set (q for query). A similarity score is a match score if $t = q$; otherwise, it is a nonmatch score. We represent the score set by

$$\mathcal{S}_{tq} \equiv \{s(g, b) | id(g) = t, id(b) = q, \forall_{g \in \mathcal{G}}, \forall_{b \in \mathcal{B}}\}.$$

  We use $\mathcal{S}_t^1 \equiv \mathcal{S}_{tt}$ (noting the superscript 1 that has just been introduced) to represent all the match scores due to the user $t \in \mathcal{U}_\mathcal{G}$, whereas

$$\mathcal{S}_t^0 \equiv \cup_{q \neq t, q \in \mathcal{U}_\mathcal{G}} \mathcal{S}_{tq}$$

  is his/her corresponding non-match scores. Using Figure 5 as an example, in which case each user has a single template in the gallery (so that $|\mathcal{G}| = |\mathcal{U}|$), $\mathcal{S}_3^1$ indicates that user 3 who is in the development set has one *match* score. The subject has seven nonmatch scores (the third column of data excluding its fourth row of data) as indicated by $\mathcal{S}_3^0$.

- **Aggregated class-conditional score set**: Having defined the user-specific score, we can define a gallery-aggregated score which is a union of all user-specific score sets, that is, $\mathcal{S}_*^0 \equiv \cup_{t \in \mathcal{U}} \mathcal{S}_t^0$ for the nonmatch score set and $\mathcal{S}_*^1 \equiv \cup_{t \in \mathcal{U}} \mathcal{S}_t^1$ for the match score set.

- **Gallery and non-gallery subjects in probe**: For open-set identification, it is necessary to distinguish two disjoint probe subsets $\mathcal{B}^\omega \subset \mathcal{B}$ for $\omega \in \{1, 0\}$, the samples of those who belong to the gallery subjects and those who do not:

$$\mathcal{B}^1 \equiv \{b \in \mathcal{B} | \exists_{g \in \mathcal{G}}, id(b) = id(g)\}$$

$$\mathcal{B}^0 \equiv \{b \in \mathcal{B} | \nexists_{g \in \mathcal{G}}, id(b) = id(g)\}$$

  Note that $\mathcal{B}^0$ is empty for closed-set identification.

2.3 Verification

In the following section, we will describe key performance indicators/metrics for biometric evaluation, and then present a metric that is commonly used as a performance criterion, that is, one that is used to find an operating threshold.

2.3.1 Performance Metric

In the verification task, a user must first claim that he/she is someone who has been enrolled into the system, and the biometric system then determines if the user's claim is true or false.

Two types of error can occur: (1) *false rejection* or *false nonmatch*, that is, falsely rejecting a genuine user's claim, or (2) *false acceptance* or *false match*, that is, falsely accepting the claim to be from a genuine user when the actual person is an impostor. Recall that false rejection and false acceptance are system-level events whereas false nonmatch and false match are events that happen at the algorithmic level, which is applicable when the database is stored and all biometric samples have already been successfully acquired.

One can estimate the error rates of the corresponding events in the following ways:

$$\text{FNMR}(\Delta) = \frac{|\{s | s \in \mathcal{S}_*^1, s < \Delta\}|}{|\mathcal{S}_*^1|}, \text{ and}$$

$$\text{FMR}(\Delta) = \frac{|\{s | s \in \mathcal{S}_*^0, s \geq \Delta\}|}{|\mathcal{S}_*^0|},$$

recalling that $\mathcal{S}_*^\omega$ denotes the population-aggregated score set with $\omega = 1$ indicating the scores due to match (genuine) comparisons; and $\omega = 0$, those with nonmatch comparisons.

We note that FNMR is a monotonically increasing function of the decision threshold, Δ, whereas FMR is a monotonically decreasing function of Δ. It is, therefore, not possible to minimise the two error rates simultaneously.

Two other commonly used quantities are EER and HTER. Although both quantities are derived from FNMR and FMR, EER is threshold-independent whereas HTER is threshold-dependent. HTER is defined as the average of FNMR and FMR, that is:

$$\text{HTER}(\Delta) = \frac{1}{2}(\text{FNMR}(\Delta) + \text{FMR}(\Delta))$$

EER is defined as a unique point where FNMR is equal to FMR.

$$\text{EER} = \text{FMR}(\Delta_*) = \text{FNMR}(\Delta_*)$$

where

$$\Delta_* = \arg\min_{\Delta}(|\text{FMR}(\Delta) - \text{FNMR}(\Delta)|)$$

ensures that the threshold found will satisfy the equality condition between FNMR and FMR as closely as possible. In practice, $\min_{\Delta}(|\text{FMR}(\Delta) - \text{FNMR}(\Delta)|)$ is non-zero but has a small ϵ value that is negligible. As a result, it is sensible to use the average of FNMR and FMR with the threshold so obtained, leading to:

$$\text{EER}_{emp} = \text{HTER}(\Delta_*)$$

2.3.2 Performance Criteria

A metric becomes a performance criterion when it is used to obtain a threshold but not for performance reporting. In biometrics evaluation, it is common to use FMR as a performance criterion and report the system performance in FNMR, and vice versa. For instance, one reports FMR when FNMR is at 0%, 0.001%, 0.01%, and so on. It is also common to report FNMR and FMR is equal to 0%. In this light, EER is a performance metric with FNMR and FMR used as performance criteria simultaneously, since EER is defined by both metrics with the constraint that they are equal.

Contrary to the above metrics, weighted Error Rate (WER) is commonly used as a performance criterion and not as a performance metric. It is defined as:

$$\text{WER}_{\beta}(\Delta) = (1-\beta)\text{FNMR}(\Delta) + (\beta)\text{FMR}(\Delta) \quad (1)$$

where $\beta \in [0, 1]$ adjusts the contribution between FMR and FNMR.

WER is a generalisation of the C_{DET} criterion used in the NIST speaker evaluation campaigns [4, Chap. 8]. It also generalises the criterion used in the BANCA protocols [5]. In the appendix, we show that a specific set of β values leads to the C_{DET} and the WER criterion established in the BANCA protocols.

The parameter β plays two roles simultaneously: (1) adjusting for the prior probability of the comparison and (2) determining the cost of false matches and false nonmatches for a given application. While the C_{DET} criterion does this explicitly via the use of cost of false matches, the cost of false non-matches, and the prior probabilities, hence three parameters in total (the two prior probabilities can be specified with a single parameter); WER does this implicitly. The use of β is convenient because it reduces the parameters that need to be taken into account from three to one, without losing any generality. Therefore, WER is an elegant criterion.

2.4 Closed-set Identification

In closed-set identification, the question one poses is: "Is the correct answer in the top k comparisons?". The first step in computing closed-set performance is to sort the similarity

scores, in decreasing order, between a given probe and the templates of all the users in the gallery, and then compute the rank of the sorted list, $rank(b)$ for $b \in \mathcal{B}$. The identification rate for rank k, IR(k), is the fraction of probes at rank k or lower, that is:

$$\text{IR}(k) = \frac{|\{b|rank(b) \leq k, \forall_{b\in\mathcal{B}}\}|}{|\mathcal{U}_{\mathcal{B}}|}$$

where the numerator is the size of the probe set. Since $rank(b)$ must lie in the range of $[1, |\mathcal{U}_{\mathcal{G}}|]$, identification rate must be in the range of $[0, 1]$.

A CMC chart is a plot $IR(k)$ versus k. $IR(k)$ is a non-decreasing function of k because of the property $|\{b|rank(b) \leq k_1\}| < |\{b|rank(b) \leq k_2\}|$ for any rank $k_1 < k_2$. For instance, if there are 100 probes and a system has 50 outputs with 50 rank-1 outcomes, 40 rank-2 outcomes, 5 rank-3 outcomes, 3 rank-4 outcomes, and 2 rank-5 outcomes, then, the number of elements with rank k or less is $\{50, 90, 95, 98, 100\}$ for ranks $k \in \{1, 2, 3, 4, 5\}$, respectively. Hence, the identification rate is 50% for rank-1 performance, 90% for rank-2 performance, and so on. As k increases, the identification rate increases and eventually attains 100%.

Closed-set identification performance is most often summarised with rank one performance, the other points such as rank 5, 10, or 20 are commonly used. The strength and weakness of the CMC is its implicit dependence on gallery size, $|\mathcal{U}_{\mathcal{G}}|$. An identification problem with one million subjects is harder than the same problem with one thousand subjects, because it is easier to wrongly identify one subject for another in the former. In order to remove the effect of gallery size, one can plot identification performance as a percentage of rank. Typically, the performance in the top 10% is plotted.

2.5 Open-set Identification

In the open-set identification task, a system determines if a probe corresponds to a person in a gallery. If the probe is determined to be in the gallery, then the algorithm identifies the person in the probe. The first task is detection whereas the second task is an identification. The Detection and Identification Rate (DIR) quantifies the *sensitivity* of the system in correctly identifying the subject at rank k. DIR is calculated as follow:

$$\text{DIR}(\Delta, k) = \frac{|\{s(g,b)|s(g,b), \forall_{g\in\mathcal{G}}, \forall_{b\in\mathcal{B}^1}, s(g,b) \geq \Delta, rank(b) \leq k\}|}{|\mathcal{B}^1|}$$

Effectively, the metric considers both the matching score $s(g, b)$ and its $rank(b)$. If the ranks were not considered, one effectively would calculate $1 - \text{FNMR}(\Delta)$. If only the ranks were considered, and not the scores, one would calculate the rank-k performance of a closed-set performance, since the non-gallery subjects in $\mathcal{B}^0$ are not considered. Therefore, the DIR metric is a hybrid of verification and identification.

A competing metric, namely, False Alarm Rate (FAR) is defined as:

$$\text{FAR}(\Delta, k) = \frac{|\{s(g,b)|s(g,b), \forall_{g\in\mathcal{G}}, \forall_{b\in\mathcal{B}^0}, s(g,b) \geq \Delta\}|}{|\mathcal{B}^0|},$$

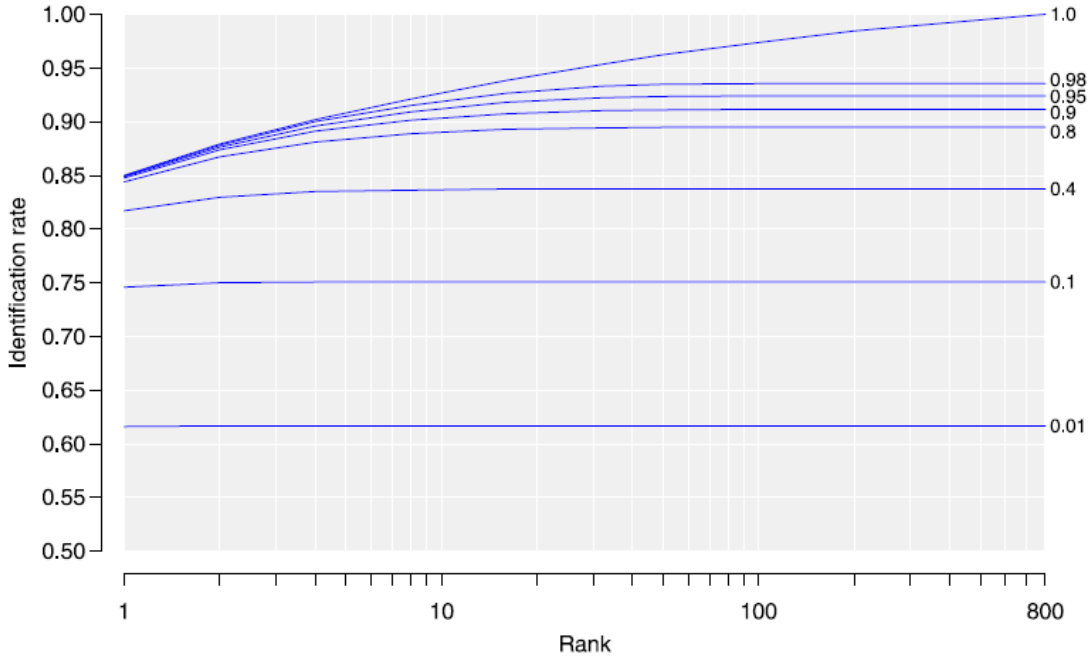


Figure 6: An example of open-set CMC chart. Each curve is a CMC with different FAR. The top line corresponds to FAR=1. The figure was taken from [6].

noting that only the non-gallery subjects are considered in the calculation.

It is apparent from the discussion that the performance metrics (DIR,FAR) is a function of Δ and k. This gives rise to at least two ways of visualising the performance: (1) an open-set ROC which plots (DIR,FAR) whilst fixing the rank k, hence, sweeping through all possible Δ; and (2) a set of CMC curves each of which vary by using a threshold that is optimised for a given FAR. The second chart hence plots DIR as a performance metric and uses FAR as a performance criterion. This chart is shown in Figure 6.

3 Impact of Score Normalisation

Biometric scores are influenced by many factors, such as the subject, the quality of template and query samples, and other unknown sources of variation that are difficult to quantify, such as ageing, environmental/ambient noise, and degradation in sensors.

A number of score normalisation schemes exist. They are cohort-based score normalisation, quality-based score normalisation, and user-specific score normalisation. Each scheme varies in the information sources they exploit to normalise a (raw) matching score. The cohort-based scheme relies on a set of cohort subjects. The quality-based scheme relies on biometric sample quality or quality measures. Finally, the user-specific scheme exploits the score statistics that are specific to each user or template, effectively reducing the effect now known as Doddington's Zoo [7] or Biometric Menagerie [8].

All these schemes can be specified by a function $f : \mathcal{S}, \mathcal{P} \rightarrow \mathbb{R}$ where $s \in \mathcal{S}$ is a raw matching score, $p \in \mathcal{P}$ represents some parameters which can be any of the three sources of information, and the output in $\mathbb{R}$ is a normalised matching score.

The score matrix we have introduced earlier has already considered the use of cohort information via the $\mathcal{C}$. For a given probe $b \in \mathcal{B}$ and a gallery sample $g \in \mathcal{G}$, a cohort-based score normalisation has the following form:

$$s'(b, g) = f_{norm}(s(b, g), \mathcal{S}(b))$$

where

$$\mathcal{S}(b) = \{s(b, t) | \forall_{t \in \mathcal{C}}\}$$

which returns a set of cohort scores due to the probe b.

Let $\mu(b)$ be the mean of the score set $\mathcal{S}(b)$ and $\sigma(b)$ be the corresponding standard deviation. T-norm is defined as:

$$s'(b, g) = \frac{s(b, g) - \mu(b)}{\sigma(b)}$$

The normalisation, known as T-norm, is carried out for all $g \in \mathcal{G}$. This technique has been reported to lower FMR without increasing FNMR [2]. Other forms of cohort-based normalisation normalisation exist (e.g., [9]) and can be implemented using the same score matrix representation.

4 Conclusions

In this deliverable, a number of commonly used performance metrics and charts have been discussed. Two variants are further considered, namely, with or without score normalisation, and the choice between *a priori* and *a posteriori* evaluation methodologies. A key distinction between the two methodologies is that the former has two score matrices, one for development, i.e., parameter tuning, including the decision threshold, and another for evaluation, i.e., it is used uniquely for measuring performance. In order to support different biometric applications, from verification to open/close-set identification, we have established a score matrix representation. This representation should be cater to all the need of the BEAT platform for the purpose of performance assessment and visualisation using commonly used charts such as DET, EPC, and CMC.

Appendix

A C_{DET} and Other Performance Criteria

The WER criterion of (1) is similar to the criterion used in the yearly NIST evaluation plans [4, Chap. 8] and the WER criterion used in the BANCA protocols [5].

The NIST evaluation plans use the C_{DET} point which is defined as:

$$C_{DET}(C_{FR}, C_{FA}) = \underbrace{C_{FR} \times P(C)}_{} \times \text{FRR}(\Delta) + \underbrace{C_{FA} \times P(I)}_{} \times \text{FAR}(\Delta), \tag{2}$$

where C_{FA} and C_{FR} are respectively the costs of false acceptance and false rejection, and $P(\omega)$ is the prior probability of comparison, which is either nonmatch or match, that is, $\omega \in \{0, 1\}$.

The BANCA protocols use a criterion also called "the WER criterion" but it is different from (1). It is defined as:

$$\text{WER}_{banca}(R, \Delta) = \frac{\text{FRR} + R\ \text{FAR}}{1 + R}, \tag{3}$$

where $R \geq 0$ balances the costs of FAR and FRR.

The two underbraced terms in C_{DET} as well as R of WER_{banca} play the same role as β in (1): they adjust for the different costs between FA and FR. Note that this adjustment parameter is not normalised for C_{DET}.

Let us explicitly write the grouped underbraced terms in C_{DET} as

$$C_{DET} = \beta_{FRR}\text{FRR}(\Delta) + \beta_{FAR}\text{FAR}(\Delta).$$

Since $\min_\Delta C_{DET}$ is equivalent to $\min_\Delta \frac{C_{DET}}{\beta_{FRR}+\beta_{FAR}}$, the normalised and non-normalised versions of C_{DET} are equivalent. As a result, (1) generalises to both the NIST and BANCA criteria.

In the NIST evaluation, the following constants are used:

$$C_{FR} = 10\ ,\ C_{FA} = 1\ ,\ P(C) = 0.01 \text{ and } P(I) = 0.99.$$

As a result, $C_{DET} = 0.1 \times \text{FRR} + 0.99 \times \text{FAR}$. By enforcing that the two costs sum to one, it can be observed that $\beta = 0.91$. For the BANCA protocols, three R values are used, namely 0.1, 1 and 10. They correspond to β values of 0.09, 0.5 and 0.91, respectively.

References

[1] N. Poh, J Kittler, C. H. Chan, S. Marcel, C. Mc Cool, E.A.Rua L. A. Castro, M. Villegas, R. Paredes, V. Struc, N. Pavesic, A.A. Salah, H. Fang, and N. Costen, "An evaluation of video-to-video face verification," *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 5, no. 4, pp. 781–801, 2011.

[2] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score Normalization for Text-Independant Speaker Verification Systems," *Digital Signal Processing (DSP) Journal*, vol. 10, pp. 42–54, 2000.

[3] N. Poh, A. Martin, and S. Bengio, "Performance generalization in biometric authentication using joint user-specific and sample bootstraps," *IEEE Transactions on Pattern Analysis and Machine (TPAMI)*, vol. 29, no. 3, pp. 492–498, March 2007.

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

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

[6] Stan Z. Li, Anil K. Jain, P. Jonathon Phillips, Patrick Grother, and Ross Micheals, "Evaluation methods in face recognition," in *Handbook of Face Recognition*, pp. 329–348. Springer New York, 2005, 10.1007/0-387-27257-7_15.

[7] 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," in *Int'l Conf. Spoken Language Processing (ICSLP)*, Sydney, 1998.

[8] N. Yager and T. Dunstone, "Worms, chameleons, phantoms and doves: New additions to the biometric menagerie," *Automatic Identification Advanced Technologies, 2007 IEEE Workshop on*, pp. 1–6, June 2007.

[9] A. Merati, N. Poh, , and J. Kitter, "Extracting discriminative information from cohort models," in *IEEE Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2010, pp. 1–6.