IDIAP RESEARCH REPORT

Towards Explaining the Success (Or Failure) of Fusion in Biometric Authentication

Norman Poh [a] Samy Bengio [a]

IDIAP–RR 05-43

July 2005

submitted for publication

[a] IDIAP, CP 592, 1920 Martigny, Switzerland

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 05-43

Towards Explaining the Success (Or Failure) of Fusion in Biometric Authentication

Norman Poh Samy Bengio

July 2005

submitted for publication

Abstract. Combining multiple information sources, typically from several data streams is a very promising approach, both in experiments and to some extents in various real-life applications. A system that uses *more than one* behavioural and physiological characteristics to verify whether a person is who he/she claims to be is called a *multimodal* biometric authentication system. Due to lack of large true multimodal biometric datasets, the biometric trait of a user from a database is often combined with another different biometric trait of yet another user, thus creating a so-called a *chimeric user*. In the literature, this practice is justified based on the fact that the underlying biometric traits to be combined are assumed to be independent of each other given the user. To the best of our knowledge, there is no literature that approves or disapproves such practice. We study this topic from two aspects: 1) by clarifying the mentioned independence assumption and 2) by constructing a pool of chimeric users from a pool of *true* modality matched users (or simply "true users") taken from a bimodal database, such that the performance variability due to chimeric user can be compared with that due to true users. The experimental results suggest that for a large proportion of the experiments, such practice is indeed questionable. Biometric authentication is a process of verifying an identity claim using a person's behavioral and physiological characteristics. Due to vulnerability of the system to environmental noise and variation caused by the user, fusion of several biometric-enabled systems is identified as a promising solution. In the literature, various fixed rules (e.g. min, max, median, mean) and trainable classifiers (e.g. linear combination of scores or weighted sum) are used to combine the scores of several base-systems. Despite many empirical experiments being reported in the literature, few works are targeted at studying a wide range of factors that can affect the fusion performance. Some of these factors are: 1) dependency among features to be combined, 2) the choice of fusion classifier/operator, 3) the choice of decision threshold, 4) the relative base-system performance, 5) the presence of noise (or the degree of robustness of classifiers to noise), and 6) the type of classifier output. To understand these factors, we propose to model Equal Error Rate (EER), a commonly used performance measure in biometric authentication. Tackling factors 1–5 implies that the use of class conditional Gaussian distribution is imperative, at least to begin with. When the class conditional scores (client or impostor) to be combined are based on a multivariate Gaussian, factors 1, 3, 4 and 5 can be readily modeled. The challenge now lies in establishing the *missing* link between EER and the fusion classifier mentioned above. Based on the EER framework, we can even derive such missing link with *non-linear* fusion classifiers, a proposal that, to the best of our knowledge, has not been investigated before. The principal difference between the theoretical EER model proposed here and previous studies in this direction is that scores are considered log-likelihood ratios (of client versus impostor) and the decision threshold is considered a prior (or log-prior ratio). In the previous studies, scores are considered posterior probabilities whereby the role of adjustable threshold *as a prior adjustment parameter* is somewhat less emphasized. Several issues which are untreated in the EER models are also discussed and supported by some 1186 experiments. These issues are 1) what if the scores are known to be not approximately normally distributed (for instance those due to Multi-Layer Perceptron outputs); 2) what if scores among classifiers to be combined are not comparable in range (their distributions are different from each other); 3) how to evaluate the performance measure other than EER using the proposed EER models.

1 Introduction

1.1 Background

Biometric authentication (BA) is a process of verifying an identity claim using a person's behavioral and physiological characteristics. BA is becoming an important alternative to traditional authentication methods such as keys ("something one has", i.e., by possession) or PIN numbers ("something one knows", i.e., by knowledge) because it is essentially "who one is", i.e., by biometric information. Therefore, it is not susceptible to misplacement or forgetfulness. Examples of biometric modalities are fingerprint, face, voice, hand-geometry and retina scans [12]. Despite its great potential, *automatic* BA systems suffer a major drawback as compared to the traditional alternatives: its relatively low accuracy and reliability (as compared to manual process). For this reason, combining multiple BA systems is identified to be a promising solution. Fusion of several systems can be performed at abstract, rank or measurement levels [5]. In the first case, only the most probable class label is returned by the system. In the second case, a list of most probable class labels is returned. Finally, in the third case, the raw output scores are used for further combining. We will focus in the last case as most information is preserved. Fusion of several systems of different biometric modalities is called *multimodal fusion* whereas fusion involving the *same* modality but *different* architectural configurations (due to different features or classifiers) is called *intramodal fusion*. The latter includes multi-classifier and multi-feature fusion. Another mode of fusion is to combine multiple samples, also called multi-sample fusion. All these mode of fusions are the central subject of this paper.

1.2 Factors Influencing Fusion Performance

To understand the factors influencing a BA system's performance, one has to know that a BA system can commit two types of error upon making a decision after comparing a score with a pre-defined threshold: falsely accepting an impostor and falsely rejecting a genuine user, or client. They are respectively Type I and Type II errors in statistics. We list here a possibly non-exhaustive unordered list of factors that could affect a combined system's performance:

- **the dependency among features of base-systems.** If features of two BA systems are dependent on the same sample (in intramodal fusion) or in time (e.g., multi-sample fusion), then, the observed scores will be likely to be dependent as well. On the other hand, the system scores derived from multimodal biometric features will be less likely to be dependent. The multi-sample fusion was examined in [14] whereby it was shown both theoretically and empirically that such approach can indeed reduce the system error by as much as 40%. It was observed that "saturation" may happen, i.e., using more instances of the same biometric trait cannot help improve the performance further. Using the XM2VTS database, Kittler *et al.* [15] examined intramodal and multimodal expert fusion. According to this empirical study, for multimodal fusion, there is no strong evidence that trainable fusion strategies (based on Decision Template [17] and Behavior Knowledge Space [10]) offer better performance than simple rules (based on sum and vote). They remarked that although adding more experts can reduce (class-dependent) variance of the combined system output, such gain is downplayed by the increased ambiguity due to the weak systems. For intramodal fusion, where the system scores are highly correlated, increasing the number of experts improve monotonically with fusion results. This empirical work complements well with our previous study [22] which, using the mean operator, can explain this phenomenon.

- **the type of output of classifier of the base-systems.** A classifier is constructed often based on a set of assumptions, more or less corresponding to the problem at hand. There are in general two broad categories of classifiers, as long as BA applications are concerned: template-based or model-based. A template-based system compares (extracted) features known to belong to an identity with features representing an access request. A model-based system derives a set of parameters to represent the identity of a person. Often, a template-based system outputs a distance measure or a correlation coefficient. A model-based system can output a posterior probability (of being a client), a log-likelihood ratio (LLR) or a similarity score. All these outputs are generally called scores. Combining heterogeneous scores of different systems has been treated in [11] in BA applications or [19] in other domains, for instance.

- **the choice of fusion operator.** There are two broad categories of fusion classifiers, namely fixed rules and trainable classifiers. As their names imply, fixed rules do not have free parameters to adjust while trainable classifiers do. Examples of fixed rules are AND, OR, mean and ordered statistics (OS) classifiers such as minimum, maximum and median. Examples of trainable classifiers are weighted sum (or linear opinion pool), weighted product (or log-opinion pool) and standard machine-learning algorithms such as Multi-Layer Perceptrons (MLPs) and Support Vector Machines (SVMs).

- **the choice of decision threshold.** Often, a reported performance due to a particular combined system's setting could be *over-confidently* stated just because the choice of threshold favors this particular setting. Changing this threshold would favor another particular setting. This is a well-known dilemma in ROC-based analysis.

- **the relative performance of base-systems.** It was reported in [7], that "a better biometric system is better used alone than combining it with a weaker one". Taking this remark out of its context would have been a grave mistake as the author was referring to the use of AND and OR operators, with a particular strategy of choosing the threshold of each base-system. However, to what extent this statement is true in a more general context (say substituting the AND and OR operators with other fixed rule operators)? Vermuulen *et al.* [32] studied empirically the case of combining two systems with equal performance, with unequal performance and when one system outperforms the other under some specific conditions. They observed that fusing two systems is advantageous when the errors committed by both systems are not correlated, i.e., the combined system may benefit from the case where, for the same access, one system commits an error and the other makes the right decision and vice-versa.

- **the presence of noise**. Some BA systems are more vulnerable to noise than other systems. With the presence of noise, how will the combined system behave? This problem was treated in [31] in multimodal fusion and in [24] in intramodal fusion.

The citations above are but a small yet representative sampling of literature that treat the listed factors. Due to the empirical approach adopted, most of these factors are studied in isolation. Our aim here is precisely to propose a theoretical model that can address jointly most if not all the factors mentioned.

1.3 Literature on Theoretical Aspects of Fusion

While many fusion experiments have been reported in the literature over the past few years, the theoretical models to explaining why fusion works (or fails) also follow closely. We review several of them here. In [9], it was demonstrated that combining several multimodal system scores using AND and OR will result in improved performance. The underlying assumption is that multimodal system scores are independent. As we understood, the issue of relative performance among systems and the strategy of choosing the decision threshold *prior to fusion* were not thoroughly considered.

In [18], the theoretical classification error of six classifiers are thoroughly studied for a two-class problem. This study assumes that the base classifier scores are probabilities $\in [0, 1]$. Hence probability of one class is one minus the probability of the other class and the optimal threshold is always set to 0.5. It also assumes that all base classifier scores are drawn from a common distribution. Gaussian and uniform distributions are studied. The first assumption is not always applicable to biometric authentication. This is because the output of a biometric system is often not necessarily a probability but a distance measure, a similarity or a log-likelihood ratio. Moreover, decisions are often taken by comparing a classifier score with a threshold. The second assumption, in practice, is also unrealistic in most situations, particularly in multimodal fusion. This is because the (class-dependent) score distributions are often *different* across different classifiers.

In [34], order statistics (OS) combiners, i.e., min, max and median, are examined both theoretically and empirically. Two concepts are introduced: biased and unbiased. Unbiased classifiers have optimal decision boundary whereas biased classifiers have a systematic offset from the optimal decision boundary. The unbiased classifier has an *inherent error* (or Bayes error) due to the (limited training) data itself, whereas the biased classifier has an *added error* due to the boundary offset, in addition to the same inherent error. It was shown in [34] that the added error due to the OS combiners can be reduced with respect to the average of any single classifier's added error. Although it was claimed that OS combiners are good alternatives to taking the average of scores (the mean operator), empirical evidences only show that the performance of the OS combiners are comparable to that of the mean operator. While the analysis in [34] is certainly interesting, there is no direct way of inferring the overall classification performance given a data set. It is also unclear how *correlation affects the OS combiners*. This is partly due to the fact that incorporating correlation into the OS combiners is analytically *intractable*. In intramodal fusion of biometric authentication tasks, scores are often correlated as they are derived from the *same* biometric sample, i.e, the same source of information.

In [13], sum and product rules are discussed in a Bayesian framework. According to this study, several fixed rules such as min, max, median and majority vote can be seen as approximations to the aforementioned rules. In particular, it was shown that the sum rule (or mean in our context) outperforms the rest of the fixed rules and even better than the single best underlying system. A further investigation showed that the sum rule is most resilient to estimation error of individual classifier than the product rule. This study, too, uses the probability framework similar to [18], i.e., the decision threshold is always 0.5. Another similarity between [13] and [18] is the *common probability distribution* assumption across scores of different classifiers. Hence, if scores are not probabilities, they have to be transformed into the range $[0, 1]$ and treated as probabilities. Such is the approach adopted in [11] or [19].

1.4 Contribution and Organization

In this paper, we adopt a rather different approach from those proposed in [18, 34, 13]. Firstly, we remark that in BA systems, scores are not always probabilities. For instance, most fingerprint-based systems are template-based and output a distance measure, the state-of-the-art speaker verification system, based on Gaussian Mixture Model [29], outputs LLR, popular face verification systems based on Principal Component Analysis or Linear Discriminant Analysis output a correlation measure (but not *necessarily* corresponding to a probability) or a distance measure [35], the commonly used Support Vector Machine in general pattern recognition task outputs also a score proportional to the distance (in the feature space defined by its kernel) of a sample from the decision hyperplane [36], etc. Secondly, the threshold often used in the decision making should reflect a system's *prior*. Hence, the performance/error measure should be a function of this threshold. To the best of our knowledge, there is no theoretical model that can *adequately* address this issue. Our preliminary work [22] addressed this subject to a limited extent. In order that the error measure be a function of the decision threshold, one cannot avoid making some assumptions about the conditional score distributions. In [22] and in this study as well, we assume that class conditional scores (client or impostor) are normally distributed. Although this seems to be rather *naive* at first, some 1186 experiments showed that the error measure estimated using the Gaussian assumption is *fairly* robust to deviation from this Gaussian. In particular, the error measure used is called Equal Error Rate (EER). This is the error rate where the probability of false acceptance equals that of false rejection. By proposing this model, our ultimate goal is to address *all of the factors* listed in Section 1.2. Without an error model, it is almost impossible to generalize empirical results due to a particular experimental setting. On the other hand, realizing that a theoretical model cannot answer all possible settings, simulations such as [] can be a handy tool to study the error behaviour.

This work can be seen as an extension to [22] in the following ways: 1) proposing many more fusion classifiers to the EER model (weighted sum, weighted product, OS classifiers, linear and quadratic classifiers) and 2) studying the joint effects of base-system performance imbalance and correlation on fusion classifiers. Realizing that the greatest weakness about the proposed EER model lies in its Gaussian assumption, we propose a *probabilistic inversion* procedure such that some BA systems with probabilistic output[1] can be fitted well in the picture. Finally, as a possible outlook, we show that removing the Gaussian assumption is actually possible.

This paper is organized as follows: Section 2 introduces the notations and presents the theoretical EER models of commonly used fusion classifiers/rules as well as a novel non-linear (quadratic) classifier. Section 3 presents the databases used. Several issues not directly treated by the EER models are discussed in Section 4. This is followed by the conclusions in Section 5.

2 Theoretical EER

We first present notations and fusion classifiers in Section 2.1. The first group of fusion classifiers based on sum and product rules are discussed in Section 2.2 whereas the second group of fusion classifiers based on OS are discussed in Section 2.3. The EER models for a family of polynomial classifiers, representing one possible but important representation of non-linear classifiers, are discussed in Section 2.4. Finally, a comparison with existing fusion theories is made in Section 2.5.

2.1 Preliminary

Let us denote C and I as the one of the two class labels k can take, client and impostor classes, respectively, i.e., $k \in \{C, I\}$. To decide whether to accept or reject an access requested by a person, denoted as "p", one can evaluate the *posterior probability ratio* in logarithmic domain (called log-posterior ratio, LPR):

$$\begin{aligned} \text{LPR} &\equiv \log\left(\frac{P(C|\mathrm{p})}{P(I|\mathrm{p})}\right) = \log\left(\frac{p(\mathrm{p}|C)P(C)}{p(\mathrm{p}|I)P(I)}\right) \\ &= \underbrace{\log\frac{p(\mathrm{p}|C)}{p(\mathrm{p}|I)}} + \underbrace{\log\frac{P(C)}{P(I)}} \equiv y - \Delta, \end{aligned} \quad (1)$$

where we introduced the term y (also called a Log-Likelihood Ratio, LLR) and a threshold Δ to handle the case of different priors. This constant also reflects the different *costs* of false acceptance and false rejection. In both cases, the threshold Δ has to be fixed *a priori*. The decision of accepting or rejecting an access is then:

$$\text{decision(LPR)} = \begin{cases} \text{accept} & \text{if LPR} > 0 \\ \text{reject} & \text{otherwise.} \end{cases} \quad (2)$$

or

$$\text{decision}_\Delta(y) = \begin{cases} \text{accept} & \text{if } y > \Delta \\ \text{reject} & \text{otherwise.} \end{cases} \quad (3)$$

Although both forms are equivalent, the explicit presence of a threshold in the second decision function shows that the log-prior ratio can be adjusted *separately* from the score y. Note that y is a direct function of the person and the *whole system*. This relationship can be explicitly written as:

$$y \equiv f_\theta(f_e(s(\mathrm{p}))), \quad (4)$$

[1]Probabilistic scores are *not* normally distributed.

where, s is a sensor, f_e is a feature extractor, θ is a set of classifier parameters associated to the classifier f_θ.

Note that there exists several types of classifiers in BA, all of which can be represented by Eqn. (4). They can be categorized by their output y, i.e., probability (within the range $[0,1]$), distance metric (more than or equal to zero), or log-likelihood ratio (a real number). In the context of multimodal BA, y is associated to the subscript i, which takes on different meanings in different context of fusion, as follows:

$$y_i(\mathrm{p}) = \begin{cases} f_\theta(f_e(s(\mathrm{p}_i))) & \text{if multi-sample} \\ f_\theta(f_e(s_i(\mathrm{p}))) & \text{if multimodal} \\ f_\theta(f_{e,i}(s(\mathrm{p}))) & \text{if multi-feature} \\ f_{\theta,i}(f_e(s(\mathrm{p}))) & \text{if multi-classifier.} \end{cases} \tag{5}$$

Note that i is the index of the i-th sample in the context of multi-sample fusion. i can also mean the i-th biometric modality in multimodal fusion, etc. In a general context, we refer to $y_i(\mathrm{p})$ as the i-th *response* and there are altogether N responses ($i = 1, \ldots, N$), but all *from the same person* and *from the same access*. We write y_i instead of $y_i(\mathrm{p})$ for simplicity, while bearing in mind that y_i is always dependent on the person.

To decide if an access should be granted or not, all $y_i|\forall_i$ have to be combined to form a single output. This can be expressed as:

$$y_{COM} = f_{COM}(y_1, \ldots, y_N) \tag{6}$$

Several types of combination strategies are used in the literature, e.g., min, max, median, mean (or sum), weighted sum, product and weighted product. They are defined as follow:

$$y_{min} = \min_i(y_i), \tag{7}$$

$$y_{max} = \max_i(y_i), \tag{8}$$

$$y_{med} = \mathrm{median}_i(y_i), \tag{9}$$

$$y_{wsum} = \sum_{i=1}^{N} w_i y_i, \tag{10}$$

$$y_{wprod} = \prod_{i=1}^{N} y_i^{w_i}, \tag{11}$$

where $w_i|\forall_i$ are parameters that need to be estimated. The mean operator is a special case of weighted sum with $w_i = \frac{1}{N}$. Similarly, the product operator is a special case of weighted product with $w_i = 1$. Kittler *et al.* [13] provides an explanation on how these fusion rules can arise as approximation to the product and sum rules in a Bayesian framework.

The family of classifiers studied as shown in Eqn. (6) here should be contrasted with the one that makes use of decision as input, i.e., $\mathrm{decision}_{\Delta_1}(y_1), \ldots, \mathrm{decision}_{\Delta_N}(y_N)$, where Δ_i is the threshold for system i. Note that in this case, to fuse N systems, N thresholds will be needed. Examples of the latter family are Behavioral Knowledge Space [10], AND and OR rules. It has several weaknesses as compared to the former family of classifiers. Firstly, due to the threshold, precious information is lost. Secondly, due to change of any threshold, the former may have to be retrained. Thirdly, to employ the latter, N thresholds will have to be fixed in advance whereas the former will only have one threshold to fix. Hence, the former family of classifiers is of our primary concern.

Because of the binary nature of decision, the system commits two types of error called False Acceptance (FA) and False Rejection (FR) errors, as a function of the threshold Δ. FA is committed when the access claims is from an impostor and is wrongly accepted by the system (as a client) whereas FR is committed when the access claims is from a client and is wrongly accepted by the system. They can be quantified by False Acceptance Rate (FAR) and False Rejection Rate (FRR) as follow:

$$\mathrm{FAR}(\Delta) = \frac{\mathrm{FA}(\Delta)}{NI} \tag{12}$$

$$\mathrm{FRR}(\Delta) = \frac{\mathrm{FR}(\Delta)}{NC}, \tag{13}$$

where $\mathrm{FA}(\Delta)$ counts the number of FA, $\mathrm{FR}(\Delta)$ counts the number of FR, NI is the total number of impostor accesses and NC is the total number of client accesses.

At this point, it is convenient to introduce two conditional variables, $Y^k \equiv Y|k$, for each k being client or impostor, respectively i.e., $k \in \{C, I\}$. Hence, $y^k \sim Y^k$ is the score y when person p is $k \in \{C, I\}$. Let $p(y^k)$ be the probability density function (*pdf*) of y^k. Eqns. (12) and (13) can then be re-expressed by:

$$\mathrm{FAR}(\Delta) = 1 - p(Y^I < \Delta), \tag{14}$$

$$\mathrm{FRR}(\Delta) = p(Y^C < \Delta), \tag{15}$$

respectively. Because of Eqn. (3), it is implicitly assumed that $E[Y^C] > E[Y^I]$, where $E[z]$ is the expectation of z. When $p(Y^k)$ for both $k \in \{C, I\}$ are assumed to be Gaussian (normally distributed), they take on the following parametric forms (see [26]):

$$\mathrm{FAR}(\Delta) = \frac{1}{2} - \frac{1}{2}\mathrm{erf}\left(\frac{\Delta - \mu^I}{\sigma^I\sqrt{2}}\right), \tag{16}$$

$$\mathrm{FRR}(\Delta) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{\Delta - \mu^C}{\sigma^C\sqrt{2}}\right) \tag{17}$$

where μ^k and σ^k are mean and standard deviation of $p(Y^k)$, and the erf function is defined as follows:

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp\left[-t^2\right] dt. \tag{18}$$

At Equal Error Rate (EER), FAR=FRR. Solving this constraint yields (see [26]):

$$\mathrm{EER} = \frac{1}{2} - \frac{1}{2}\mathrm{erf}\left(\frac{\text{F-ratio}}{\sqrt{2}}\right) \equiv \mathrm{eer}(\text{F-ratio}), \tag{19}$$

where,

$$\text{F-ratio} = \frac{\mu^C - \mu^I}{\sigma^C + \sigma^I}. \tag{20}$$

The function eer is introduced here to simplify the EER expression as a function of F-ratio because eer will be used frequently in this paper. Note that the threshold Δ is omitted since there is only one unique point that satisfies the EER criterion.

A more general performance evaluation measure is called Half Total Error Rate (HTER) and is defined as:

$$\mathrm{HTER}(\Delta) = \frac{1}{2}(\mathrm{FAR}(\Delta) + \mathrm{FAR}(\Delta)), \tag{21}$$

for any arbitrary threshold Δ.

When the cost of false acceptance is not the same as false rejection, we can also associate a pair of FAR and FRR with a particular cost. This function is sometimes called Decision Cost Function (DCF) in the NIST evaluation [20]. It is defined as:

$$\mathrm{DCF}(\Delta) = P(I)\mathrm{Cost}(I)\mathrm{FAR}(\Delta) + P(C)\mathrm{Cost}(C)\mathrm{FRR}(\Delta), \tag{22}$$

Table 1: Summary of theoretical EER based on the assumption that class-independent scores are normally distributed.

Fusion methods	EER
average baseline[1]	$\mathrm{EER}_{AV} = \mathrm{eer}\left(\frac{\mu^C_{AV}-\mu^I_{AV}}{\sigma^C_{AV}+\sigma^I_{AV}}\right)$ $\mu^k_{AV} = \frac{1}{N}\sum_i \mu^k_i$ $\left(\sigma^k_{AV}\right)^2 = \frac{1}{N}\sum_i \left(\sigma^k_i\right)^2$
single-best classifier	$\mathrm{EER}_{best} = \mathrm{eer}\left(\max_i \left(\frac{\mu^C_i-\mu^I_i}{\sigma^C_i+\sigma^I_i}\right)\right)$
mean rule	$\mathrm{EER}_{mean} = \mathrm{eer}\left(\frac{\mu^C_{mean}-\mu^I_{mean}}{\sigma^C_{mean}+\sigma^I_{mean}}\right)$ $\mu^k_{mean} = \frac{1}{N}\sum_i \mu^k_i$ $\left(\sigma^k_{mean}\right)^2 = \frac{1}{N^2}\sum_{i,j} \mathbf{\Sigma}^k_{i,j}$
weighted sum, linear classifier[2]	$\mathrm{EER}_{wsum} = \mathrm{eer}\left(\frac{\mu^C_{wsum}-\mu^I_{wsum}}{\sigma^C_{wsum}+\sigma^I_{wsum}}\right)$ $\mu^k_{wsum} = \sum_i \omega_i \mu^k_i$ $\left(\sigma^k_{wsum}\right)^2 = \sum_{i,j} \omega_i\omega_j \mathbf{\Sigma}^k_{i,j}$
OS combiners[3]	$\mathrm{EER}_{OS} = \mathrm{eer}\left(\frac{\mu^C_{OS}-\mu^I_{OS}}{\sigma^C_{OS}+\sigma^I_{OS}}\right)$ $\mu^k_{OS} = \mu^k + \gamma_1\sigma^k$ $\left(\sigma^k_{OS}\right)^2 = \gamma_2\left(\sigma^k\right)^2$

Remark 1: This is not a classifier but the average performance of baselines when used independently of each other. By its defintion, scores are assumed independent as classifiers function independently of each other. **Remark** 2: The weighted product takes the same form as weighted sum, except that log-normal distribution is assumed instead. **Remark** 3: OS classifiers assume that scores *across classifiers* are i.i.d. The reduction factor γ is listed in Table 2. The mean and weighted sum classifiers *do not* assume that scores are i.i.d.

and it requires the costs to be specified in advance so as the prior probabilities. We will use a similar but simplified function, called Weighted Error Rate (WER), which has only one parameter to be specified in advance. It is defined as:

$$\mathrm{WER}(\Delta) = \alpha\mathrm{FAR}(\Delta) + (1-\alpha)\mathrm{FRR}(\Delta), \tag{23}$$

where α balances between the contribution of FA and FR. As can be seen, WER and DCF is equivalent when $\alpha = P(I)\mathrm{Cost}(I)$ and $\mathrm{Cost}(I) = \mathrm{Cost}(C) = 1$. HTER is also a special case of WER when $\alpha = 0.5$.

2.2 Theoretical EER of Fusion Classifiers

We now derive several parametric forms of fused scores using different types of classifiers, namely the single-best classifier, mean, weighted sum, product rule and Order Statistics (OS)-combiners such as min, max and median. The OS-combiners are further discussed in Section 2.3.

The analysis in this section is possible due to the simple expression of F-ratio, which is a function of four parameters: $\{\mu^k, \sigma^k|\forall_{k=\{C,I\}}\}$ as shown in Eqn. (19). Suppose that the i-th response is y^k_i sampled from $p(Y^k_i)$ and there are N classifiers, i.e., $i = 1, \ldots, N$. The *average baseline* performance of classifiers, considering that each of them works independently of the other, is shown in the first row of Table 1. The (class-dependent) average variance, σ^k_{AV}, is defined as the average over all the variances of classifier. This is in fact not a fusion classifier but the *average performance* of classifiers measured in EER. The single-best classifier in the second row chooses the baseline classifier that maximizes the F-ratio. This is the same as choosing the one with minimum EER because F-ratio is inversely proportional to EER, as implied by the left part of Eqn. (19).

The derivation of EER of weighted sum (as well as mean) fusion can be found in [25]. The central idea consists of projecting the N dimensional score onto a one dimensional score via Eqn. (11). Suppose that the class conditional scores (prior to fusion) are modeled by a multivariate Gaussian with mean $(\boldsymbol{\mu}^k)^T = \mu_1^k, \ldots, \mu_N^k$ and covariance $\boldsymbol{\Sigma}^k$ of N-by-N dimensions. Let $\boldsymbol{\Sigma}_{i,j}^k$ be the i-th row and j-th column of covariance matrix $\boldsymbol{\Sigma}^k$ for $k = \{C, I\}$. $E[\cdot]$ is the expectation operator (over samples) and $\omega_i \sim W_i^k$ is an instance of a while noise associated to classifier i for all k. The linear score projection from N dimensions to one dimension has the same effect on the Gaussian distribution: from N multivariate Gaussian distribution to a single Gaussian distribution with mean μ_{wsum}^k and variance $(\sigma_{wsum})^2$ defined in the fourth row of Table 1 for each class k. The mean operator is derived similarly with $w_i = \frac{1}{N} \forall i$. Note that the weight w_i affects both the mean and variance of fused scores. In [22], it was shown mathematically that the EER of mean, EER_{mean}, is always smaller than or equal to the EER of the average baseline performance (EER_{AV}). This is closely related to the ambiguity decomposition [16] often used in the regression context (as opposed to classification as done in [22]). However, there is no evidence that $\text{EER}_{mean} \leq \text{EER}_{best}$, i.e., the EER of the best-classifier. In [4], it was shown that $\sigma_{wsum}^k \leq \sigma_{mean}^k$, supposing that the $w_i \forall_i$ are optimal. In [8], when the correlation among classifiers is assumed to be zero, $w_i \propto (\text{EER}_i)^{-1}$. As a result, this implies that $\text{EER}_{wsum} \leq \text{EER}_{mean}$. The finding in [4] is more general than that of [8] because the underlying correlation among baseline classifiers is captured by the covariance matrix. Hence, fusion using weighted sum can, in theory, have better performance than the mean rule, assuming that the weights are tuned optimally. Although there exists several methods to tune the weights in the literature, to the best of our knowledge, no standard algorithm *directly* optimizes EER (hence requiring further investigation which cannot be dealt here). For the product operator, it is necessary to bound Y to be within the range $[0, 1]$, otherwise the multiplication is not applicable. Consider the following case: two instances of classifier score can take on any real value. The decision function as in Eqn. (3) is used with optimal threshold being zero. With an impostor access, both classifier scores will be negative if correctly classified. Their product, on the other hand, will be positive. This is clearly undesirable.

The weighted product (and hence product) at first seems slightly cumbersome to obtain. However, one can apply the following logarithmic transform instead: $\log(Y_{wprod}^k) = \sum_i w_i \log(Y_i^k)$, for any y_i^k sampled from $p(Y_i^k)$. This turns out to take the same form as weighted sum. Assuming that Y_i^k is log-normally distributed, we can proceed the analysis in a similar way as the weighted sum case (and hence the mean rule).

2.3 Theoretical EER of Order Statistics Combiners

Attempting to analyze analytically the EER values of as fixed rule *order statistics* (OS) such as the maximum, minimum and median combiners, as done in the previous section, is difficult without making (very) constraining assumptions. The first assumption is that the instance of scores must be *comparable*. If scores of various types of classifiers are involved for fusion, their range may not be comparable. Hence, score normalization is imperative while this pre-processing step is *unnecessary* in the previous section. The second assumption assumes that scores are i.i.d. In this case, there exists a very simple analytical model[2]. Although this model seems too constraining, it is at least applicable to fusion with multiple samples which satisfies some of the assumptions stated here: scores are comparable; and they are *identically distributed* but unfortunately not necessarily *independently* sampled.

[2]This assumption will be *removed* during experimentation with synthetic data.

Table 2: Reduction factors γ_1 and γ_2 versus different fixed rules assuming i.i.d samples (zero correlation)

N	γ_2 values			γ_1 values	
	OS combiners		mean	OS combiners	
	min, max,	median	$(\frac{1}{N})$	min	max
1	1.000	1.000	1.000	0.00	0.00
2	0.682	0.682	0.500	-0.56	0.56
3	0.560	0.449	0.333	-0.85	0.85
4	0.492	0.361	0.250	-1.03	1.03
5	0.448	0.287	0.200	-1.16	1.16

Reduction factor γ_2 (2 for the second moment) with respect to the standard normal distribution due to fusion with min, max (the second column) and median (third column) OS combiners for the first five samples (indicated by N) according to [1]. The fourth column is the *maximum* reduction factor due to mean (at zero correlation), with minimum reduction factor being 1 (at perfect correlation). The fifth and sixth columns show the shift factor γ_1 (for the first moment) as a result of applying min and max for the first five samples. These values also exist in tabulated forms but here they are obtained by simulation. For median, γ_1 is relatively small (in the order of 10^{-4}) beyond 2 samples and hence not shown here. γ_1 approaches zero as N is large and γ_2 approaches $1/N$.

All OS combiners will be collectively studied. The subscript OS can be replaced by min, max and median. Suppose that $y_i^k \sim Y_i^k$ is an instance of i-th response knowing that the associated access claim belongs to class k. y_i^k has the following model:

$$y_i^k = \mu_i^k + \omega_i^k, \tag{24}$$

where μ_i^k is a deterministic component and ω_i^k is a noise component. Note that in the previous section ω_i^k is assumed to be normally distributed with zero mean. The fused scores by OS can be written as: $y_{OS}^k = \mathrm{OS}(y_i^k) = \mu^k + \mathrm{OS}(\omega_i^k)$, where i denotes the i-th sample (and not the i-th classifier output as done in the previous section). Note that μ^k is constant across i and it is *not affected* by the OS combiner. The expectation of y_{OS}^k as well as its variance are shown in the last row of Table 1, where γ_2 is a reduction factor and γ_1 is a shift factor, such that $\gamma_2(\sigma^k)^2$ is the variance of $\mathrm{OS}(\omega_i^k)$ and $\gamma_1\sigma^k$ is the expected value of $\mathrm{OS}(\omega_i^k)$. Both γ's can be found in tabulated form for various noise distributions [1]. A similar line of analysis can be found in [34] except that class-independent noise is assumed. The reduction factors of combining the first five samples, assuming Gaussian distribution, are shown in Table 2. The smaller γ_2 is, the smaller the associated EER. The fourth column of Table 2 shows the reduction factor due to mean (as compared to the second and third columns). It can be seen that mean is overall superior under the i.i.d. assumption and common class-conditional variance (but different class-conditional mean, of course)

2.4 Beyond Linear Classifiers

The EER models discussed in 2.2 are limited to linear classifiers so far. The next more flexible (with higher capacity) than linear is perhaps a quadratic classifier of the form:

$$y_{quad} = \sum_{i=1,j=1}^{N} w_{i,j} y_i y_j \tag{25}$$

where $w_{i,j}$ is the i-th row and j-th column of the weight matrix $\boldsymbol{w}$. Note that the threshold constant is absent because this threshold adjusts for the prior in Eqn. (1). Another possible form would be:

$$y_{quad,2} = \sum_{i=1,j=1}^{N} w_{i,j}^{(2)} y_i y_j + \sum_{i=1}^{N} w_i^{(1)} y_i \tag{26}$$

which includes a linear interaction term found in the weighted sum classifier. We will first study the first form and generalize to the second form. In order to estimate the F-ratio of the quadratic classifier, we need to estimate the conditional mean and standard deviation of the resultant combined scores. Let us denote these two parameters by μ^k_{quad} and σ^k_{quad}. Let the class conditional variable of the combined score be Y^k_{quad}. To be consistent with the previous models, the same noise model as in Eqn. (24) is assumed here. The noise model due to $y_i y_j$ can thus be expressed by[3]:

$$y_i y_j = \mu_i \mu_j + \mu_i \omega_j + \mu_j \omega_i + \omega_i \omega_j \tag{27}$$

Using Eqn. (27), the mean is:

$$\begin{aligned}
\mu^k_{quad} &= E[\sum_{i,j} w_{i,j} y_i y_j] \\
&= E\left[\sum_{i,j} w_{i,j}(\mu_i \mu_j + \mu_i \omega_j + \mu_j \omega_i + \omega_i \omega_j)\right] \\
&= \sum_{i,j} w_{i,j} \mu_i \mu_j,
\end{aligned}$$

since $E[z\omega] = 0$ for any constant z. The variance is:

$$\begin{aligned}
\left(\sigma^k_{quad}\right)^2 &= E\left[\left\{\sum_{i,j} w_{i,j} y_i y_j - E[Y^k_{quad}]\right\}^2\right] \\
&= E\left[\left\{\sum_{i,j} w_{i,j} y_i y_j - \sum_{i,j} w_{i,j} \mu_i \mu_j\right\}^2\right] \\
&= E\left[\left\{\sum_{i,j} w_{i,j}(\mu_i \omega_j + \mu_j \omega_i + \omega_i \omega_j\right\}^2\right].
\end{aligned}$$

This result can be further expanded by considering the product between $\mu_i \omega_j + \mu_j \omega_i + \omega_i \omega_j$ and $\mu_m \omega_n + \mu_n \omega_m + \omega_m \omega_n$ for all possible integer m, n, i, j in $1, \ldots, N$. The result can be summarized by introducing the class-conditional matrix $\boldsymbol{z}$, whose element is:

$$z^k_{m,n,i,j} = \begin{bmatrix} \mu_m \omega_n \mu_j \omega_i & \mu_m \mu_n \mu_i \omega_j & \mu_m \mu_n \omega_i \omega_j \\ \mu_n \omega_m \mu_j \omega_i & \mu_n \mu_m \mu_i \omega_j & \mu_n \mu_m \omega_i \omega_j \\ \omega_m \omega_n \mu_j \omega_i & \omega_m \omega_n \mu_i \omega_j & \omega_m \omega_n \omega_i \omega_j \end{bmatrix}$$

where $\omega_i = \sqrt{\Sigma^k_{i,i}}$ and $\omega_{i,j} = \Sigma^k_{i,j}$ (the i-th and j-th element of the class k covariance matrix for $k = \{C, I\}$). Hence, the variance of the combined scores is:

$$(\sigma^k_{quad})^2 = \sum_{m,n,i,j} w_{m,n} w_{i,j} z^k_{m,n,i,j}.$$

[3] *All* these terms are conditioned on the class label k. They are deliberately removed here for simplicity.

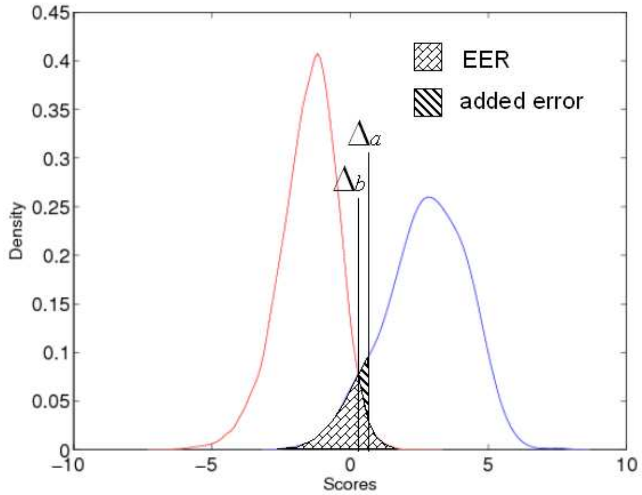


Figure 1: Probabilistic Density Functions of client (right) and impostor (left) classes fitted on a speech experiment. The threshold Δ_b is due to an unbiased classifier who commits the Bayes error (also known as EER in our context). This is the smallest error the system can commit. The threshold Δ_a is due to a biased classifier and hence it commits an added error apart from the Bayes error.

The second form can be easily generalized to incorporate the additional term $\sum_i w_i y_i$. In this case, the resultant mean and standard deviation will be:

$$\begin{aligned}
(\mu^k_{quad,2})^2 &= \mu^k_{quad} + \mu^k_{wsum} \\
(\sigma^k_{quad,2})^2 &= (\sigma^k_{quad})^2 + \sum_{i,j} w_{i,j}\omega_i\omega_j,
\end{aligned}$$

respectively. Polynomial of degree three and higher can be derived in the same way as the steps presented here, i.e., first write down the noise model $y_p y_q y_r \ldots$ and then its two Gaussian parameters, with the help of $z^k_{s,t,u,\ldots,p,q,r\ldots}$, etc. Hence, the EER model presented here is not limited to linear classifier models.

2.5 Compatibility and Comparison with Existing Theory

The theoretical analysis proposed here is different from [34, 8] in terms of application, context and methodology. In [34, 8], two types of errors are introduced, namely Bayes (inherent) error and added error. The former is due to an unbiased classifier whose class posterior estimates correspond to the true posteriors. The latter is due to a biased classifier which results in wrongly estimated class posteriors. In the context of a binary classification problem, both Bayes and added errors can be illustrated by Figure 1. From the application point of view, the EER used here is commonly found in binary classification problems while the error (sum of Bayes error and added error) applies to any number of classes. It is tempting to conclude that EER is equivalent to the Bayes error for a two-class problem. However, in the context of [34, 8], the Bayes error is due to additive error in the feature space near the decision boundary. This additive error causes a slight change in the outputs (posteriors) of the respective classifier (one output for each class). This change between input features and output posteriors is linearly approximated around features *near the decision boundary*, although the classifier model may be non-linear. In EER, the input measurement is not a set of features but a set of scores of one or more base-classifiers. The output posteriors between the two classes are assumed to be (integral of) Gaussian. The local continuity at the boundary is implicitly assumed. Since the Bayes error cannot be reduced, it is understandable that the analyzes in [34, 8] focus on reducing the locally defined added error (on features near the decision boundary between a pair of class posteriors). To the best of our understanding, this error cannot be (or is not useful to be) calculated explicitly due to the local definition. Only the final effect of classification error can be measured. EER, on the other hand, is a performance measure (hence *explicitly* calculated) and can only be measured *globally* over the whole data set. As a result, it is defined everywhere in the score space (not just near the decision boundary) contrary to the error terms in [34, 8] which are local. It is hence unavoidable in EER to make assumption about the two class posteriors (known as False Acceptance Rate and False Rejection Rate). Since Bayes error cannot be reduced, the focus in [34, 8] is to reduce the added error. This is understandable because features are susceptible to noise. On the contrary, our focus is to reduce EER (the analogous term of Bayes error) at the decision boundary and other errors (e.g., Half Total Error Rate, Weighted Error Rate) not *necessarily* near the decision boundary. Perhaps the most remarkable difference is that Bayes error *cannot* be reduced due to is definition, whereas EER can [22]. Putting our work in the context of [34, 8], in this study, we focus only on unbiased classifiers. The biased case has been studied for the mean rule in [22].

In terms of methodology, for simplicity, additional assumptions are made at some points of discussion in [34, 8]. For instance, class-independent variance and correlation is assumed in [30, Sec. III-C] (extension from [8]) when studying unbiased and correlated estimation errors; and in [34, Secs. 3,4], the indeterministic noise component (error) is assumed to be identical for different classifiers when discussing combination of biased and unbiased classifiers through OS and linear combiners. While such simplifications are necessary to allow studying of multiple classes, they are simply too unrealistic to be useful in biometric authentication since only two classes are involved (we know that in practice, the class-dependent variances are *not the same* even for Gaussian-distributed scores). Finally, despite the differences, both theoretical analyzes lead to rather similar forms. In fact, it is possible to write the added error in the terms used in EER.

Before doing so, it is useful to define the total error, which is a sum of Bayes error and added error. In terms of FAR and FRR, this is:

$$E_{tot}(\Delta) = \mathrm{FAR}(\Delta) + \mathrm{FRR}(\Delta), \tag{28}$$

where Δ is a threshold. Figure 1 shows two instances of the threshold, namely Δ_b (for Bayes) and Δ_a (for added). By definition, at Δ_b,

$$E_{tot}(\Delta_b) = \mathrm{FAR}(\Delta_b) + \mathrm{FRR}(\Delta_b) = 2\mathrm{EER}$$

since $\mathrm{FAR}(\Delta_b) = \mathrm{FRR}(\Delta_b) = \mathrm{EER}$.

The added error can be written as:

$$\begin{aligned} E_{added} &= E_{tot}(\Delta_a) - E_{tot}(\Delta_b) \\ &= \mathrm{FAR}(\Delta_a) - \mathrm{FAR}(\Delta_b) + \\ &\quad \mathrm{FRR}(\Delta_a) - \mathrm{FRR}(\Delta_b) \end{aligned} \tag{29}$$

The added error can also be written in terms of client and impostor distributions, as follows:

$$E_{added} = \begin{cases} P(\Delta_b \le Y^C \Delta_a) - P(\Delta_b \le Y^I \Delta_a) & if \Delta_b < \Delta_a \\ P(\Delta_a \le Y^I \Delta_b) - P(\Delta_a \le Y^C \Delta_b) & if \Delta_a < \Delta_b \\ 0 & if \Delta_a = \Delta_b \end{cases} \tag{30}$$

3 Database

Before dealing with several critical issues to be discussed in Section 4, we present the two datasets used: BANCA [2] and XM2VTS [3, 21]. The experimental scores of these two databases are publicly available[4]. Both databases contains face and speech modalities. The BANCA score dataset contains 1186 experimental outcomes and they cover a large range of HTER values. They are used principally to test the proposed theoretical EER model. The XM2VTS score dataset contains 13 baseline experiments and several fusion protocols are defined [21]. One fusion protocol contains 32 intramodal and multimodal fusion experiments and this will be used in a user-independent manner.

4 Discussion

The fundamental assumption about the EER proposed in the previous section is that both classes of conditional scores are each normally distributed. What if this assumption is not true? This is addressed in Section 4.1. For the OS classifiers to work, it is assumed that the scores to be combined are drawn from the same distribution. What if they are not? This is typically the case when combining heterogeneous classifiers (with different output type and value range). This is addressed in Section 4.2. Section 4.3 treats two important issues not addressed so far: performance measure other than EER and relaxing the Gaussian assumption.

4.1 Handling Class-Dependent Gaussian Assumption

It should be warned that the analysis here is founded on the assumption that class-dependent *pdf* is normally distributed. Consider the case where the base-classifier output is an MLP which outputs posterior probability (within the range $[0, 1]$) typically due to using a logistic activation function or outputs scores within the range $[-1, 1]$ typically due to using hyperbolic tangent activation function. Then, one knows that the scores *cannot be adequately* modeled by Gaussian distributions because simply they are indeed *not* normally distributed. Ideally, one should *always* use the output just before applying any one of the two squashing functions mentioned. However, when this is not possible, (for instance, due to using a commercial-off-the-shelf product), reversing this process is possible, to a certain extent. The usual definition of sigmoid and tangent hyperbolic are:

$$
\begin{aligned}
\mathrm{sigmoid}(z) &= \frac{1}{1+\exp(-z)} \\
\tanh(z) &= \frac{\sinh(z)}{\cosh(z)}
\end{aligned}
$$

respectively. If y is an output of a sigmoid or a hyperbolic tangent function, its inverse is:

$$
\mathrm{sigmoid}^{-1}(y) = -\log\left(\frac{1}{y}-1\right) \tag{31}
$$

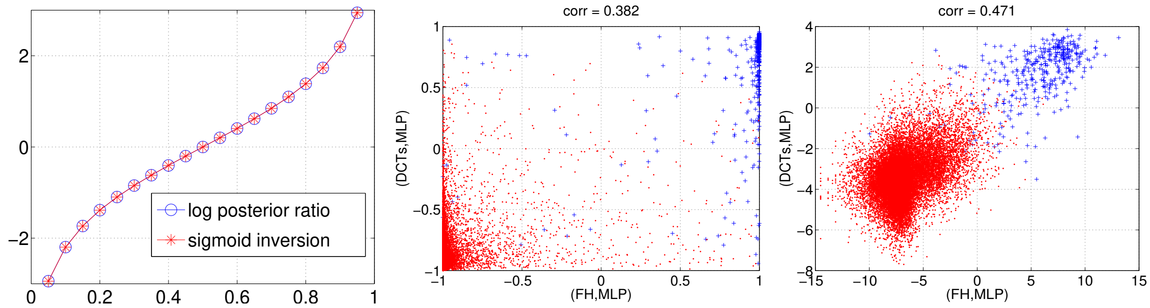
$$
\tanh^{-1}(y) = \frac{1}{2}\log\left(\frac{1+y}{1-y}\right), \tag{32}
$$

respectively.

Another more principled way of handling posterior is to turn them into LPR (log-posterior ratio). Consider y to be the posterior probability of being client given the person p, i.e., $P(C|\mathrm{p})$. Then the posterior of being the impostor, $P(I|\mathrm{p})$, is simply $1 - P(C|\mathrm{p})$. By definition, LPR is

$$
\mathrm{LPR} = \log\frac{P(C|\mathrm{p})}{1-P(C|\mathrm{p})}. \tag{33}
$$

[4]BANCA: "ftp://ftp.idiap.ch/pub/bengio/banca/banca_scores" and XM2VTS: "http://www.idiap.ch/∼norman/fusion"



(a) probabilistic inversion (b) original (c) prob. inverse

Figure 2: (a): The probabilistic inversion based on LPR and sigmoid functions. For both curves, when $x = 0$, $y = -\infty$ and when $x = 1$, $y = \infty$. (b) and (c): Scatter plots of one of the 32 fusion data sets using (b) the original score prior to fusion and (c) probabilistic inversed scores. The two base-classifiers use the same modality but different feature set. The X-axis is a face expert based on histogram features and an MLP classifier, labeled as (FH,MLP). The Y-axis is also a face expert based on DCTMod2 features and an MLP classifier, labeled as (DCTs,MLP). Hence, they are expected to be somewhat correlated. Their corresponding correlations are measured to be 0.382 for (a) and 0.471 for (b). In all 32 datasets involving MLPs, the correlation among classifiers are systematically under-estimated in the original score space with respect to the probabilistic inverse space.

Under such system, the decision function as in Eqn. (2) is effective. This is equivalent to making the accept decision when $P(C|\mathrm{p}) > P(I|\mathrm{p})$ or $P(C|\mathrm{p}) > 0.5$ and vice-versa for making the reject decision. Figure 2(a) shows transformation using Eqn. (33) and compares it with that of Eqn. (31). As can be seen, both approaches are equivalent (proof omitted); these transformations are linear around the mid-range and non-linear towards the range limit. As a result, when y is near the range limits, its inversion will be limited to the machine precision represented by the value y. Hence, such remedial procedure cannot "reverse" perfectly the process. Fortunately, we now know that data points near the range limits cannot influence the decision function. On the other hand, only those data points near the decision boundary, which is found to be in the mid-range, typically those found within the margin, can influence the decision function [36].

Figures 2(b) and (c) show the effect before and after applying probabilistic inversion on one of the 32 XM2VTS fusion dataset. This example is chosen such that the two underlying face systems are based on MLP architecture. Stronger correlation is expected since both systems operate on the same biometric sample *for each access request*. The supposedly observed correlation is squashed by each of the MLP output activation. The probabilistic inversion reveals that the correlation is indeed much stronger than before applying the procedure. Furthermore, the scores are also much more normally distributed than before. To evaluate more objectively the probabilistic inversion, we use the 1186 experiments in BANCA. The goal is to measure how "far" the score distributions are before and after applying the probabilistic inversion. For this purpose, we applied the Lillie-test [6]. It evaluates the hypothesis that a set of (client or impostor) scores has a normal distribution with unspecified mean and variance against the alternative that the set of scores does not have a normal distribution. This test is similar to Kolmogorov-Smirnov (KS) test, but it adjusts for the fact that the parameters of the normal distribution are estimated from the set of scores rather than specified in advance. In the BANCA score database, there are 474 experiments based on MLP, 514 on GMM and 182 on SVM. For the MLP classifiers, we measured the KS-statistics twice, once before the probabilistic inversion and once after. The results are shown in Figure 3. As can be seen, the output of MLPs (trained using sigmoid output function) gives high KS-statistics whereas the outputs of SVMs and GMMs

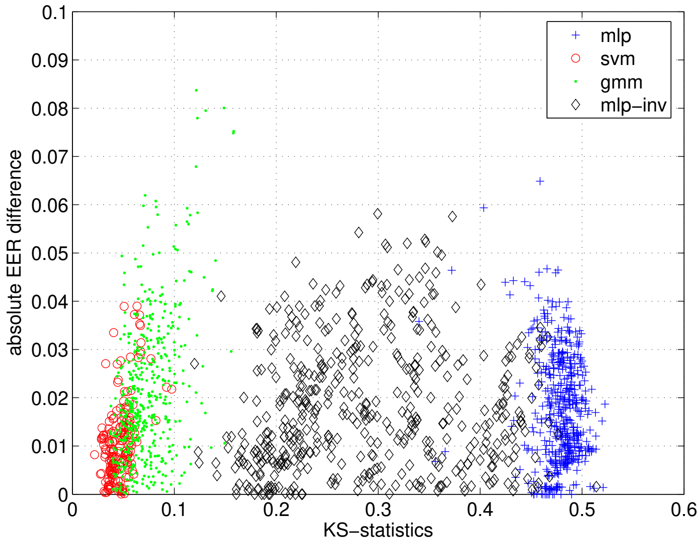


Figure 3: Absolute EER difference between theoretical EER and empirical EER versus the average KS-statistics between the corresponding client and impostor distributions. Each data point represents the pair of values (absolute EER difference, average KS) over 1186 experiments of the BANCA score database. Three sets of classifiers are distinguished here: MLP, SVM and GMM. The fourth type of classifier shown here (denoted as "mlp-inv") is the MLP classifier output preprocessed with the probabilistic inversion procedure before calculating its theoretical EER and KS-statistics. KS-statistics measures the degree of deviation from Gaussian assumption.

conform better to the Gaussian assumption. Furthermore, the scores after probabilistic inversion are more conforming to the Gaussian distribution than those before, although they are still far (in KS-statistics sense) from those of GMMs and SVMs. Prior to this experiment, we thought that deviation from Gaussian would mean large absolute EER difference. If this were to be the case, absolute EER difference would have been increasing proportionally with respect to the KS-statistics. It turns out that this is not the case. Hence, despite the use of Gaussian assumption, the empirical EER is still somewhat predictable.

4.2 Handling Score-Comparability Assumption

Score-comparability is somewhat crucial to the success of fixed rule classifiers (OS classifiers and mean). As long as application in BA is concerned, a pioneer study in [11] shows that score-normalization is important especially when combining heterogeneous classifier output. Another study in this direction can be found in [19] which deals with fusion of posterior outputs. For the purpose of this study, we will begin with an illustration in Figure 4. Suppose that both the client and the impostor scores of two systems to combine are normally distributed, but the distribution of these two systems are not well aligned (see Figure 4, left). One well-known technique to mitigate this miss-alignment is called Z-score normalization. This normalization attempts to center the scores such that the impostor score distribution is aligned. Such alignment, unfortunately can cause the client distribution to be miss-aligned (see Figure 4, middle). Another normalization that does not suffer from this drawback is called F-score normalization [28]. This procedure aims to align both the client and impostor score distribution *simultaneously* (see Figure 4, right). The above illustration motivates the need to normalize scores obtained from heterogeneous systems, such that the order relationship "$\leq$" is meaningful. Although the EER model of OS classifier assumes common score distribution *for convenience*, in practice, one

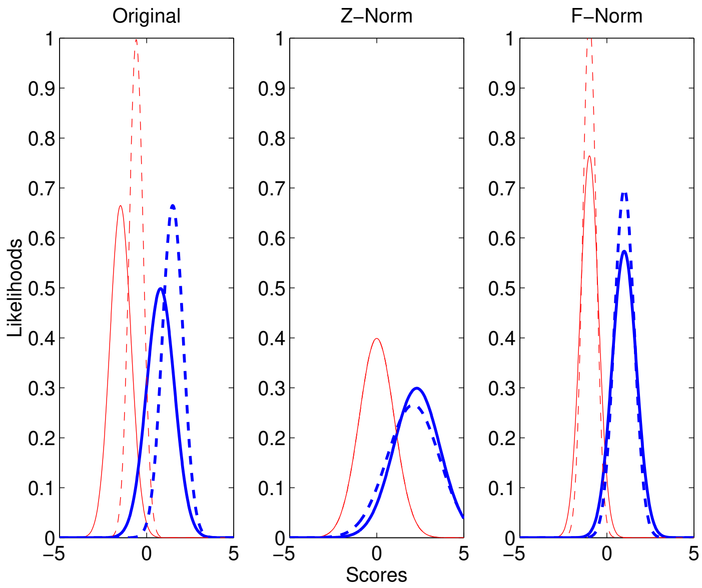


Figure 4: The *pdf* of a client score distribution (thick line) and that of an impostor score distribution (thin line) of two systems to combine (dotted and continuous lines) before score normalization (left), after applying Z-score normalization (middle) and after applying F-score normalization.

only needs to make sure that the order relationship is respected. Z-score normalization is often used in speaker verification tasks whereby the score variation due to different client models can be high. It is used in a user-specific manner (one per client model). This same technique is useful in combining heterogeneous system outputs as well. For both techniques, the major assumption is that the class conditional scores are normally distributed. All linear normalization techniques have the following form:

$$y^{lin} = A(y - B) \tag{34}$$

where B is a bias and A is a scaling factor. Suppose the mean and standard deviation of the class conditional distribution is μ^k and σ^k, respectively, for $k \in \{C, I\}$, such that $p(y) = \mathcal{N}\left(y^k|\mu^k, (\sigma^k)^2\right)$. Z–norm can be defined by $B = \mu^I$ and $A = (\sigma^I)^{-1}$. To define F-norm, one needs to define two anchor points, each defining the "desired" class center c^k.

$$y^F = \frac{c^C - c^I}{\mu^C - \mu^I}(y - \mu^I) + c^I. \tag{35}$$

When $c^I = 0$ and $c^C = 1$, one can write this formulation using Eqn. (34), with $B = \mu^I$ and $A = (\mu^C - \mu^I)^{-1}$. Other linear normalization techniques can be found in [11], such as min-max normalization:

$$B = \min(y) \qquad \text{and} \qquad A = (\max(y) - \min(y))^{-1} \ , \tag{36}$$

decimal scaling normalization:

$$B = 0 \qquad \text{and} \qquad A = \left(10^{\log_{10} \max y}\right)^{-1} \ , \tag{37}$$

and median normalizations:

$$\begin{aligned} B &= \text{median}(y) \text{ and} \\ A &= \text{median}(|y - \text{median}(y)|)^{-1}. \end{aligned} \tag{38}$$

These linear transformations "preserve" the distribution since they only shift and scale the original distribution. Non-linear normalization techniques can change the distribution drastically and can force the scores to be aligned. Two such techniques are proposed in [11]. They are the double sigmoid function:

$$y_i^{dsig} = \begin{cases} \frac{1}{1+\exp\left(-2\left(\frac{y_i - m}{s_1}\right)\right)} & \text{if } y_i < m, \\ \frac{1}{1+\exp\left(-2\left(\frac{y_i - m}{s_2}\right)\right)} & \text{otherwise }, \end{cases} \tag{39}$$

and the tanh-estimator

$$y^{tanh} = \frac{1}{2}\left\{\tanh\left(0.01\left(\frac{y-\mu^{k=C}}{\sigma^{k=C}}\right)\right)\right\}, \tag{40}$$

respectively. Note that the double sigmoid function has free parameters such as $m, s1$ and $s2$. There is no clear way to estimate these parameters and it is done by cross-validation [11] (as we understood). Motivated by robust statistics, $\mu^{k=C}$ and $\sigma^{k=C}$ in the tanh-estimator are estimated using the Hampel estimator. Here, we introduce a margin-based normalization. It is defined as (see [23]):

$$y^{margin} = \text{FRR}(y) - \text{FAR}(y),$$

where FRR and FAR are *empirically estimated* from the (training) data set using Eqns. (14 and 15). It has the property that y^{margin} is confined in $[-1, 1]$, making it attractive to be interpreted as a confidence measure. Furthermore $y^{margin}/2 + 0.5$ can be *naively* interpreted as a probability (having the range $[0, 1]$) (although they do not necessarily correspond to the *true class posterior*). The advantage of margin-derived score is that it has no free parameters to estimate and is entirely dependent on the training data (to estimate FAR and FRR curves). Figure 5 compares the resultant scatter plot of applying these three normalization techniques with the original one.

4.3 Beyond EER and Gaussian Assumption

Lastly, although only the EER value is studied here, one can extend the present finding to a more general case, whereby the EER constraint by its definition, i.e, $\text{EER}(\Delta) = \text{FAR}(\Delta) = \text{FRR}(\Delta)$, does not hold anymore.

Suppose μ^k_{COM} and σ^k_{COM} are mean and standard deviation of any of the fusion techniques presented in Table 1. Then all performance measures not satisfying EER such as DCF Eqn. (22), WER Eqn. (23), and HTER Eqn. (21) can still be evaluated using Eqns. (16) and (17).

It is also possible to replace the class conditional Gaussian assumption with a mixture of Gaussians or Gaussian Mixture Model (GMM) such that:

$$p(y_{com}|k) = \sum_{c=1}^{C} q_c^k \mathcal{N}\left(y_{com}|\mu^k_{com,c}, (\sigma^k_{com,c})^2\right) \tag{41}$$

where q_c^k is the Gaussian prior, μ_c^k the mean and σ_c^k the standard deviation for the c-th component and for class $k = \{C, I\}$. The FAR and FRR can be *numerically* evaluated (by integration over a range of Δ values) using Eqns. (14 and 15), respectively.

To illustrate this idea, we calculated WER with $\alpha = \{0.1, 0.5, 0.9\}$ over 1186 BANCA datasets using three approaches, namely Mixture of Gaussians as in Eqn. (41), Gaussian without probabilistic inversion preprocessing for the MLP output and Gaussian with preprocessing. The results are shown in Figure 6 for Gaussian without probabilistic inversion preprocessing and GMM. The number of Gaussian components in GMM is selected by validation on a subset of training scores not used by the Expectation-Maximization (EM) algorithm to tune the GMM parameters. As can be roughly observed, the WER estimated by GMM matches better the true WER (that is calculated from the scores directly), as compared to the WER estimated using the Gaussian assumption (with probabilistic inversion), across different costs. The WER estimated using the Gaussian assumption without pre-processing is not shown. They are similar to its Gaussian counterpart (with pre-processing). To measure objectively the estimated WER, we calculated the WER difference between a WER estimate and the true WER. The distribution of WER difference for the three methods are shown in Figure 7 for the three different costs of α. Narrower WER difference distribution implies better performance. The GMM approach provides the best estimates of WER compared to those using Gaussian assumption. Hence, as can be seen, extending the finding from EER to the more general case is possible.

When is Eqn. (41) applicable? Or, how did one obtain the multiple-component Gaussian score distribution in the first place? Suppose we model the class-conditional (multi-dimensional) *input* scores with a GMM of the form:

$$p(\boldsymbol{y}|k) = \sum_{c=1}^{C} q_c^k \mathcal{N}\left(\boldsymbol{y}|\boldsymbol{\mu}_c^k, \boldsymbol{\sigma}^k\right) \tag{42}$$

Let us suppose further that a linear decision hyperplane is used and whose weight is w_i for each i-th system. Then, for each component c, the mean and variance will be:

$$\begin{aligned} \mu_{com,c}^k &= \sum_i w_i \mu_i^k \\ \left(\sigma_{com,c}^k\right)^2 &= \sum_i \sum_j w_i w_j \Sigma_{i,j,c}^k, \end{aligned}$$

respectively (based on Table 1), where $\Sigma_{i,j,c}^k$ is the i-th and j-th element of k-th class covariance of the $c-th$ component. The Gaussian priors q_c^k for all c and k remain the same in Eqns. (41 and 42). These Gaussian priors are found by (typically) the EM algorithm trained on the N dimensional score space. Although a linear decision function is used here for illustrative purpose, a possible extension is to use a non-linear decision function.

5 Conclusions

Several important factors that can influence the performance of multimodal and intramodal BA systems are: 1) the dependency of features of base-systems, 2) the type of output of classifier of the base-systems, 3) the choice of fusion operator, 4) the choice of decision threshold, 5) the relative performance of base-systems and 6) the presence of noise. The above aspects have been studied in isolation by different authors, for instance, [15, 11, 7, 31] However, trying to study simultaneously all these factors is almost impossible by experiments. This empirically driven approach is difficult to carry out due to the combinatorial factors of correlation, classifier performance imbalance, fusion classifiers used, among others. The first and foremost obstacle is that the datasets simply do not permit one to test the joint effects. In this paper, we propose to study the above issues, especially 1–5 (see [22] for the 6-th factor) by proposing theoretical EER models. The fundamental and seemingly the most *naive* assumption is that class conditional scores (for client and impostor classes) are normally distributed. The EER of several classifiers based on this assumption is then proposed. These classifiers can be divided into two categories: sum and product rule-motivated classifiers (e.g., mean, weighted sum, product, weighted product) and Order Statistics (OS) classifiers (e.g., min, max, median). We even further propose to model EER of a family of polynomial classifiers with the quadratic classifier as a concrete example. To the best of our knowledge, as long as fusion is concerned, a theoretical analysis on non-linear classifier has not been found elsewhere in the literature. This can be a good starting point to answer the question: "Could a non-linear fusion classifier be better than a linear fusion classifier?". Some empirical studies such as [33] based on a reduced polynomial expansion discriminant function *do* support this hypothesis.

One justification that we put forward is to use the Log-Posterior Ratio (LPR). It can be decomposed into Log-Likelihood Ratio (LLR) and a log-prior ratio probability. We interpret a system's output score as a LLR and the threshold used to make the accept/reject decision to be the log-prior ratio probability. This is to be contrasted with the previous proposed Bayesian framework by [18, 13], whereby a BA system output is interpreted as a probability. Our justification is that the threshold often used in a practical BA system corresponds to one's prior belief and it is tailored to a specific application. Hence, LPR should be more appropriate in reflecting BA applications. We identify our proposed work with that of [34] but with significant differences. The most important differences can be summarized as follows: 1) EER is reducible and Bayes error in [34] is not; 2) EER is defined globally (after observing all scores); Bayes error is defined locally around a given feature sample near the decision boundary; 3) EER is measurable; Bayes is not (since it is defined locally); and, 4) the local continuity of FAR and FRR around EER is enforced by the Gaussian assumption; the added error is linear around a local feature by approximation using Taylor expansion.

In the case where a BA system output is posterior probability (typically due to Multi-Layer Perceptrons trained with a logistic/sigmoid function), whereby scores are no longer approximately normally distributed, we propose to use the scores just before passing through the non-linear activation function of the classifier. When this is not possible due to some reasons (e.g., using an off-the-shelf system), we propose to invert this process using probabilistic inversion procedures. Using 1186 XM2VTS experimental scores, we showed our proposed EER model can estimate EER and other performance errors not satisfying EER *fairly robustly*, despite the fact that the scores are not exactly normally distributed. As a possible outlook, we also proposed to *remove* the Gaussian assumption central to this work. As a replacement, we propose a semi-parametric approach, realized using a mixture of Gaussian. Although experimentally shown to be very powerful, by so doing, one looses the interpretability offered by those models based on Gaussian assumption. There are basically two contradictory goals here: to understand the influencing factors or to model/predict the performance. If the goal is the former, then the EER models proposed here are useful and relevant. If the goal is the latter, then one can simply carry out a fusion experiment and then empirically measure the performance. The proposed semi-parametric approach is particularly useful when empirically carrying out the experiments is prohibitive. This is the case, for instance, in the problem of choosing a subset of candidates for fusion. This is a combinatorial problem with a complexity of $2^N - 1$ with N classifiers to choose from. In this case, the semi-parametric approach will solve this problem by inferring the performance directly after modeling the data with conditional mixture of Gaussians, without actually carrying out experiments. Finally, although no actual empirical fusion experiments are performed here, interested readers can refer to a preliminary version of this paper in [27] which contains some 256 fusion experiments. It also contains some simulation of performance of several fixed rules and the weighted sum classifier. These simulations are carried out to study the joint effects of correlation and the relative strengh between classifiers to be combined on the performance of a particular fusion classifier. Under some restricted assumptions, weighted sum can be showned to theoretically outperform other fixed rule classifiers. The mean rule turns out to be very competitive.

Acknowledgement

The authors thank the anonymous reviewers of [22], for giving constructive questions which we thoroughly addressed in the paper. This work was supported in part by the IST Program of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, funded in part by the Swiss Federal Office for Education and Science (OFES) and the Swiss NSF through the NCCR on IM2. This publication only reflects the authors' view.

References

[1] B.C. Arnold, N. Balakrishnan, and H.N. Nagaraja. *A First Course in Order Statistics.* Wiley, New York, 1992.

[2] E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran. The BANCA Database and Evaluation Protocol. In *Springer LNCS-2688, 4th Int. Conf. Audio- and Video-Based Biometric Person Authentication, AVBPA'03*. Springer-Verlag, 2003.

[3] Samy Bengio, Johnny Mariethoz, and Sebastien Marcel. Evaluation of Biometric Technology on XM2VTS. IDIAP-RR 21, IDIAP, 2001.

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

[5] R. Brunelli and D. Falavigna. Personal Identification Using Multiple Cues. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17(10):955–966, 1995.

[6] W. J. Conover. *Practical Nonparametric Statistics.* Wiley, 1980.

[7] J. Daugman. Biometric decision landscapes. Technical Report TR482, University of Cambridge Computer Laboratory, 2000.

[8] G. Fumera and F. Roli. Analysis of Linear and Order Statistics Combiners for Fusion of Imbalanced Classifiers. In *LNCS 2364, Proc. 3rd Int'l Workshop on Multiple Classifier Systems (MCS 2002)*, pages 252–261, Cagliari, 2002.

[9] L. Hong, A.K. Jain, and S. Pankanti. Can Multibiometrics Improve Performance? Technical Report MSU-CSE-99-39, Computer Science and Engineering, Michigan State University, East Lansing, Michigan, 1999.

[10] Y. Huang and C. Suen. A Method of Combining Multiple Experts for the Recognition of Unconstrained Handwritten Numerals. *IEEE Trans. Pattern Recognition and Machine Intelligence*, 17(1):1, 1995.

[11] A. Jain, K. Nandakumar, and A. Ross. Score Normalisation in Multimodal Biometric Systems. *Pattern Recognition (to appear)*, 2005.

[12] A.K. Jain, R. Bolle, and S. Pankanti. *Biometrics: Person Identification in a Networked Society.* Kluwer Publications, 1999.

[13] J. Kittler, M. Hatef, R. P.W. Duin, and J. Matas. On Combining Classifiers. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[14] J. Kittler, G. Matas, K. Jonsson, and M. Sanchez. Combining Evidence in Personal Identity Verification Systems. *Pattern Recognition Letters*, 18(9):845–852, 1997.

[15] J. Kittler, K. Messer, and J. Czyz. Fusion of Intramodal and Multimodal Experts in Personal Identity Authentication Systems. In *Proc. Cost 275 Workshop*, pages 17–24, Rome, 2002.

[16] A. Krogh and J. Vedelsby. Neural Network Ensembles, Cross-Validation and Active-Learning. *Advances in Neural Information Processing Systems*, 7, 1995.

[17] L. Kuncheva., J.C. Bezdek, and R.P.W. Duin. Decision Template for Multiple Classifer Fusion: An Experimental Comparison. *Pattern Recognition Letters*, 34:228–237, 2001.

[18] L.I. Kuncheva. A Theoretical Study on Six Classifier Fusion Strategies. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(2):281–286, February 2002.

[19] C-L. Liu. Classifier Combination Based on Confidence Information. *Pattern Recognition*, (38):11–28, 2004.

[20] A. Martin. NIST Year 2001 Speaker Recognition Evaluation Plan, 2001.

[21] N. Poh and S. Bengio. Database, Protocol and Tools for Evaluating Score-Level Fusion Algorithms in Biometric Authentication. Research Report 04-44, IDIAP, Martigny, Switzerland, 2004. Accepted for publication in *AVBPA 2005*.

[22] N. Poh and S. Bengio. How Do Correlation and Variance of Base Classifiers Affect Fusion in Biometric Authentication Tasks? Research Report 04-18, IDIAP, Martigny, Switzerland, 2004. accepted for publication in *IEEE Trans. Signal Processing*, 2005.

[23] N. Poh and S. Bengio. Improving Fusion with Margin-Derived Confidence in Biometric Authentication Tasks. Research Report 04-63, IDIAP, Martigny, Switzerland, 2004. Accepted for publication in *AVBPA 2005*.

[24] N. Poh and S. Bengio. Noise-Robust Multi-Stream Fusion for Text-Independent Speaker Authentication. In *The Speaker and Language Recognition Workshop (Odyssey)*, pages 199–206, Toledo, 2004.

[25] N. Poh and S. Bengio. Towards Predicting Optimal Subsets of Base-Experts in Biometric Authentication Task. In *IDIAP Research Report 04-17, Martigny, Switzerland*, Accepted for publication in *Joint AMI/PASCAL/IM2/M4 Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, 2004.

[26] N. Poh and S. Bengio. Why Do Multi-Stream, Multi-Band and Multi-Modal Approaches Work on Biometric User Authentication Tasks? In *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pages vol. V, 893–896, Montreal, 2004.

[27] N. Poh and S. Bengio. EER of Fixed and Trainable Classifiers: A Theoretical Study with Application to Biometric Authentication Tasks. In *LNCS 3541, Multiple Classifiers System (MCS)*, pages 74–85, Monterey Bay, 2005.

[28] N. Poh and S. Bengio. F-ratio Client-Dependent Normalisation on Biometric Authentication Tasks. In *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pages 721–724, Philadelphia, 2005.

[29] D. A. Reynolds, T. Quatieri, and R. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1–3):19–41, 2000.

[30] F. Roli and G. Fumera. A Theoretical and Experimental Analysis of Linear Combiners for Multiple Classifier Systems. *IEEE Trans. Pattern Analysis and Machine Intelligence*, (in press).

[31] C. Sanderson. *Automatic Person Verification Using Speech and Face Information*. PhD thesis, Griffith University, Queensland, Australia, 2002.

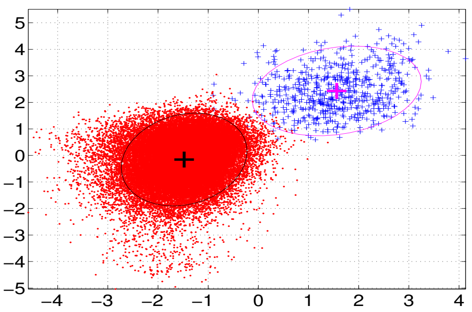
[32] S. Sharma, H. Hermansky, and P. Vermuulen. Combining Information from Multiple Classifiers for Speaker Verification. In *Proc. Speaker Recogniton and Its Commercial and Forestic Applications Workshop (RLA2C)*, pages 115–119, Avignon, 1998.

[33] K.-A. Toh, W.-Y. Yau, and X. Jiang. A Reduced Multivariate Polynomial Model For Multimodal Biometrics And Classifiers Fusion. *IEEE Trans. Circuits and Systems for Video Technology (Special Issue on Image- and Video-Based Biometrics)*, 14(2):224–233, 2004.

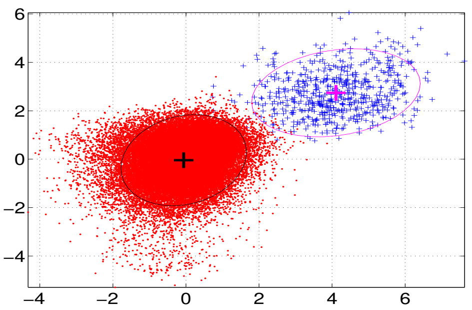
[34] K. Tumer and J. Ghosh. Robust Combining of Disparate Classifiers through Order Statistics. *Pattern Analysis and Applications*, 5:189–200, 2002.

[35] M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscienc*, 3(1):71–86, 1991.

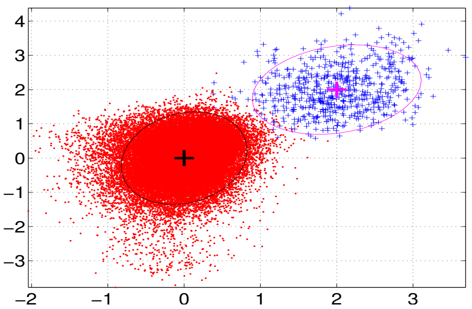
[36] V. N. Vapnik. *Statistical Learning Theory*. Springer, 1998.



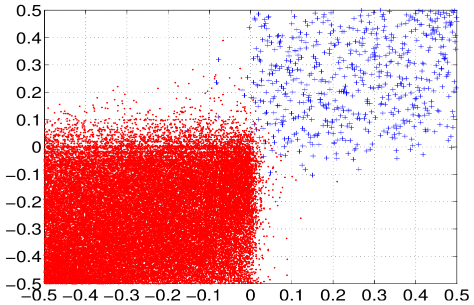
(a) original



(b) Z-norm

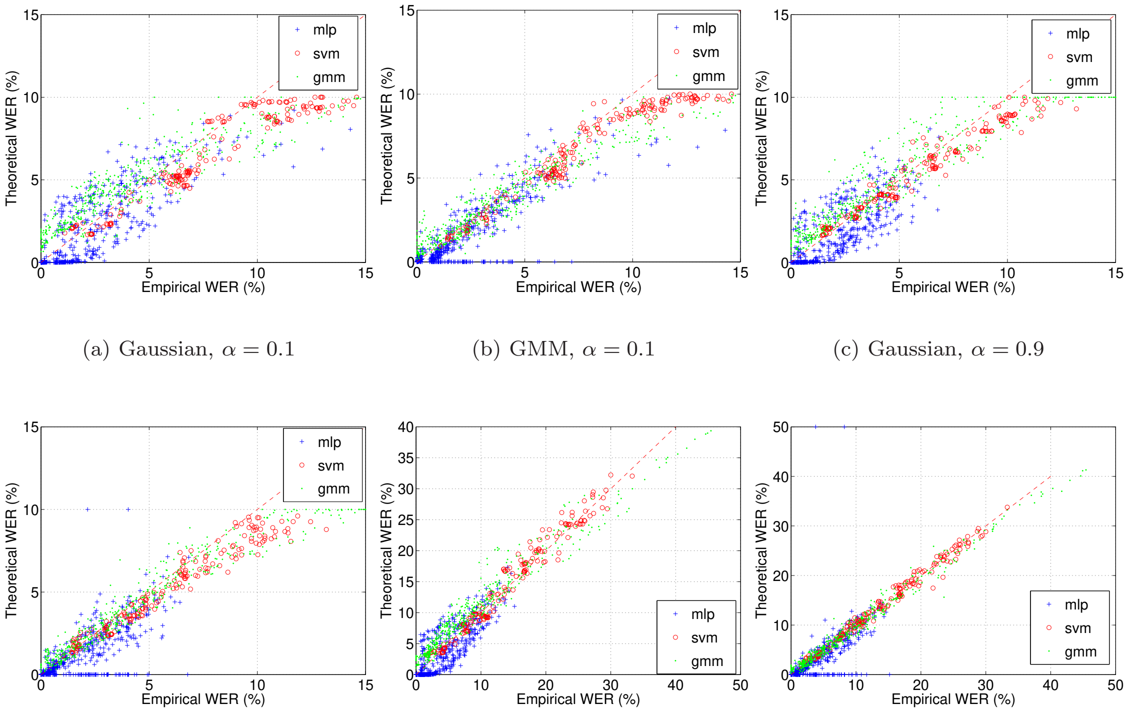


(c) F-norm



(d) margin

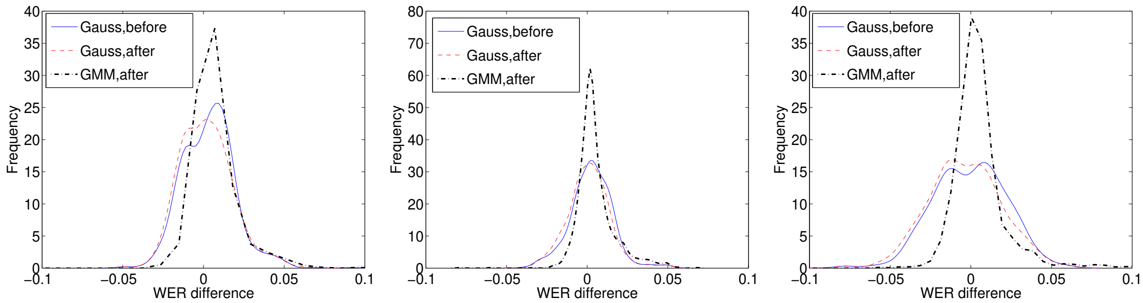
Figure 5: Scatter plots of one of the fusion data sets using (a) the original score, (b) Z-norm, (c) F-norm and (d) margin-transformed scores scaled into the range $[-0.5, 0.5]$. The X- and Y-axes are the output of two BA systems. For (a)–(c), a bi-variate Gaussian fit is also shown on each class of scores with mean marked as a big plus sign and width displayed as an oval. The client cluster of scores (small plus signs) are on the upper right corner and the those of impostor (small dots) are on the lower left corner. Note that for (b), the impostor centers are always zero for the two systems whereas the client centers could take on any values. In (c), not only the impostor centers are always zero, the client centers are also fixed to 2 in this case (or any number desired). Due to being linear transformations, both Z- and F-norms *preserve* the score distribution linearly whereas margin-score transformation *changes* the original score distribution.



(a) Gaussian, $\alpha = 0.1$ (b) GMM, $\alpha = 0.1$ (c) Gaussian, $\alpha = 0.9$

(d) GMM, $\alpha = 0.9$ (e) Gaussian, $\alpha = 0.5$ (f) GMM, $\alpha = 0.5$

Figure 6: Empirical WERs (a, c and e) versus approximated WERs (b, d and f) using class conditional Gaussian assumption with probabilistic inversion pre-processing and GMM with three different α values, namely 0.1 (a vs. b), 0.9 (c vs. d) and 0.5 (e vs. f), evaluated over 1186 BANCA datasets. The 1186 WERs approximated using class conditional Gaussian assumption *without* probabilistic inversion pre-processing are not shown as they are very similar to a, c and e.



(a) $\alpha = 0.1$ (b) $\alpha = 0.9$ (c) $\alpha = 0.5$

Figure 7: The distribution of absolute WER difference due to Gaussian assumption before probabilistic inversion, after probabilistic inversion and GMM. Narrower distribution implies better estimation of WER with respect to the true WER that is calculated from scores directly.