IDIAP RESEARCH REPORT

Evidences of Equal Error Rate Reduction in Biometric Authentication Fusion

Norman Poh Hoon Thian [a] Samy Bengio [a]

IDIAP–RR 04-43

August 2004

submitted for publication

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

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

IDIAP Research Report 04-43

Evidences of Equal Error Rate Reduction in Biometric Authentication Fusion

Norman Poh Hoon Thian　　Samy Bengio

August 2004

submitted for publication

Abstract. Multimodal biometric authentication (BA) has shown perennial successes both in research and applications. This paper casts a light on why BA systems can be improved by fusing opinions of different experts, principally due to diversity of biometric modalities, features, classifiers and samples. These techniques are collectively called variance reduction (VR) techniques. A thorough survey was carried out and showed that these techniques have been employed in one way or another in the literature, but there was no systematic comparison of these techniques, as done here. Despite the architectural diversity, we show that the improved classification result is due to reduced (class-dependent) variance. The analysis does not assume that scores to be fused are uncorrelated. It does however assume that the class-dependent scores have Gaussian distributions. As many as 180 independent experiments from different sources show that such assumption is acceptable in practice. The theoretical explanation has its root in regression problems. Our contribution is to relate the reduced variance to a reduced classification error commonly used in BA, called Equal Error Rate. In addition to the theoretical evidence, we carried out as many as 104 fusion experiments using commonly used classifiers on the XM2VTS multimodal database to measure the gain due to fusion. This investigation leads to the conclusion that different ways of exploiting diversity incur different hardware and computation cost. In particular, higher diversity incurs higher computation and sometimes hardware cost and vice-versa. Therefore, this study can serve as an engineering guide to choosing a VR technique that will provide a good trade-off between the level of accuracy required and its associated cost.

1 Introduction

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 [27].

However, today, biometric-based security systems (devices, algorithms, architectures) still have room for improvement, particularly in their accuracy, tolerance to various noisy environments and scalability as the number of individuals increases. Biometric data is often noisy because of deformable nature of biometric traits, corruption by environmental noise, variability over time and occlusion by the user's accessories. The higher the noise, the less reliable the biometric system becomes.

Advancements in BA show several trends. The most prominent trend is the use of multiple biometric modalities (multi-modal biometrics), such as combining face and speech information [53], face and fingerprint [22], speech and ear [26], visible-light with infrared face information [11], palmprint and face [33], ear and face [9] audio and non-audio (using non-acoustic sensors) cues [7], among others. Three biometric modalities have also been proposed, such as frontal face, lips and speech cues [14], or hand, face and fingerprint [51]. However, in practical applications, where convenience, performance and robustness are concerned, simple and fast access techniques such as text-dependent speaker authentication (the user is required to speak specific text known to the system) coupled with fingerprint and keypad are actually preferred [41]. The second major trend is the exploration of various biometric traits, such as iris [13], palmprint [59, 60], gait [2, 20], on-line signature (using electronic pen [1] or mouse [17]) and key-stroke dynamics [17].

Meanwhile, traditional BA methods such as fingerprint, face or speaker authentication continue to evolve, particularly into dealing with robustness against different mismatched conditions, e.g., robust speaker authentication against noisy environments [47], robust face authentication against pose [54] or illumination variations [43].

Unfortunately, in the authors' opinion, there is still a lack of study on why multi-modal biometrics should work, even though many works have already been done in multimodal biometrics and most have shown improvement. Pankanti *et al.* [23] shaded some lights on this subject. They demonstrated that combining the expert opinions using AND and OR will result into improved performance. Unfortunately they assumed that the baseline expert opinions are not correlated. Sanchez *et al.* [31] showed both theoretically and empirically that fusing multiple instances of biometric traits can indeed reduce the system error by as much as 40%. The theoretical analysis, unfortunately, again did not deal with the case when the expert opinions are correlated. Since multiple instances of the same biometric traits are likely to be correlated, it is not clear how correlation in expert opinions can hamper the expected improvement, although they 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, Kitter *et al.* [32] examined intramodel 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 [34] and Behaviour Knowledge Space [24]) 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 output of system expert's opinion (often called a score or a probability), such gain is downplayed by the increased ambiguity due to the weak experts. For intramodal fusion, where the expert scores are highly correlated, increasing the number of experts improve monotonically with fusion results. Unfortunately, the issue of correlation is not examined in details. Vermuulen *et al.* [58] studied empirically the case of combining two systems' hypotheses. Specifically, they examined the combination of two systems with equal performance, with unequal performance and when one system outperforms the other under certain 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. Unfortunately, the correlation of these errors are not explored further.

Section 2 explores different techniques that can be applied in fusion. These techniques are collectively called variance-reduction techniques. Section 3 proposes a theoretical framework that deals specifically with combining expert opinions that are *possibly correlated* and study how this will affect the final combined hypothesis. Section 4 proposes the F-ratio measure and shows how it can be related to the Equal Error Rate (EER) measure, often used to assess BA systems. It extends the findings of reduced variance in Section 3 to reduced EER. Section 5 provides empirical evidences to support the theoretical evidences proposed earlier. This is followed by conclusions in Section 6.

2 Exploring Diversity in BA

2.1 An Architectural Review of BA

The fundamental problem of BA can be viewed as a classification task to decide if person $\mathbf{x}$ is a client or an impostor. In a statistical framework, the probability that $\mathbf{x}$ is a client after a classifier f_θ observes his/her scanned biometric trait can be written as:

$$y \equiv f_\theta(f_e(s(\mathbf{x}))), \tag{1}$$

where, s is a sensor, f_e is a feature extractor, θ is a set of classifier parameters associated to the classifier f_θ. If the classifier is associated to a unique client identity i, we can replace θ by $\theta(i)$. Note that there exists several types of classifiers in BA, all of which can be represented by Eqn. (1). 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).

A BA system can be viewed as a series of subsystems working together, i.e., a result of concatenation of several subsystems, namely sensors, extractors and classifiers. This is shown in Figure 1a. The circle, square and diamond represent sensor, feature extractor and classifier, respectively. $\mathbf{x}$ is the person requesting an access and y is the opinion of the system that the person is a client, as defined in Eqn. (1).

A user's biometric trait is captured using *sensors*. Examples of sensors are Charged Couple Device (CCD) cameras, Infrared-Red (IR) cameras, fingerprint scanners and microphones. Each sensor has its own standard data representation. A set of operations, often based on signal- and image-processing algorithms, constitute the building blocks of extractors. *Extractors* have two functions: to detect and to extract user-discriminant information. Each extractor produces its own type of vectors or feature set, also called templates in a more generic setting. *Experts* (or classifiers) are used to categorise these produced vectors. Classifiers are a set of pattern-matching algorithms, which may be learning-based (e.g. Multi-Layer Perceptron, Support Vector Machine, etc) or template-based (dynamic time warping, Euclidean distance, normalised correlation, etc). Classifiers' role is to map a vector to an associated identity. They do so with a certain degree of confidence commonly called a score or a confidence measure. It could be a scalar value or a vector when more information is supplied. In some cases, a score can be interpreted as the estimated *a posteriori* probability of the claimed class label given the feature. When there are several classifiers, a *combination mechanism* (COM) (also known as supervisor or fusion module) merges different scores to obtain the final decision. To make the final decision, a score is compared with a pre-defined threshold. If the final decision is a match, then the system accepts the identity claim. If the decision is a non-match, then the system rejects the identity claim. Finally and optionally, if the decision is inconclusive, a fall-back procedure should be activated.

The serial concatenation process of sensors, extractors, classifiers and supervisors shows that error may accumulate along the chain (see Figure 1a) because each module depends on its previous module. In a separate study done by Jain and Pankanti [28], they used the terms *information limited behaviour*, *representation limited behaviour* and *invariance limited behaviour* to describe the errors of the first three components (sensors, extractors and classifiers). To our opinion, the term "limitation" is easier to be understood as errors due to sensors, extractors and classifiers, respectively.

2.2 Architectural Diversity

This section explores parallel structures that can be applied to the BA problem. The classical model has already been defined in Eqn. (1). In multimodal BA, it can be written in two parts, as follow:

$$\overrightarrow{X_F} = [f_\theta(f_e(s(\mathbf{x}_i)))]_{i\in\mathcal{I}} \quad (2)$$
$$\overrightarrow{X_F} = [f_\theta(f_e(s_i(\mathbf{x})))]_{i\in\mathcal{M}} \quad (3)$$
$$\overrightarrow{X_F} = [f_\theta(f_{e,i}(s(\mathbf{x})))]_{i\in\mathcal{F}} \quad (4)$$
$$\overrightarrow{X_F} = [f_{\theta,i}(f_e(s(\mathbf{x})))]_{i\in\mathcal{T}} \quad (5)$$
$$(6)$$

and

$$y_{COM} = f_{\theta_{COM}}(\overrightarrow{X_F}), \quad (7)$$

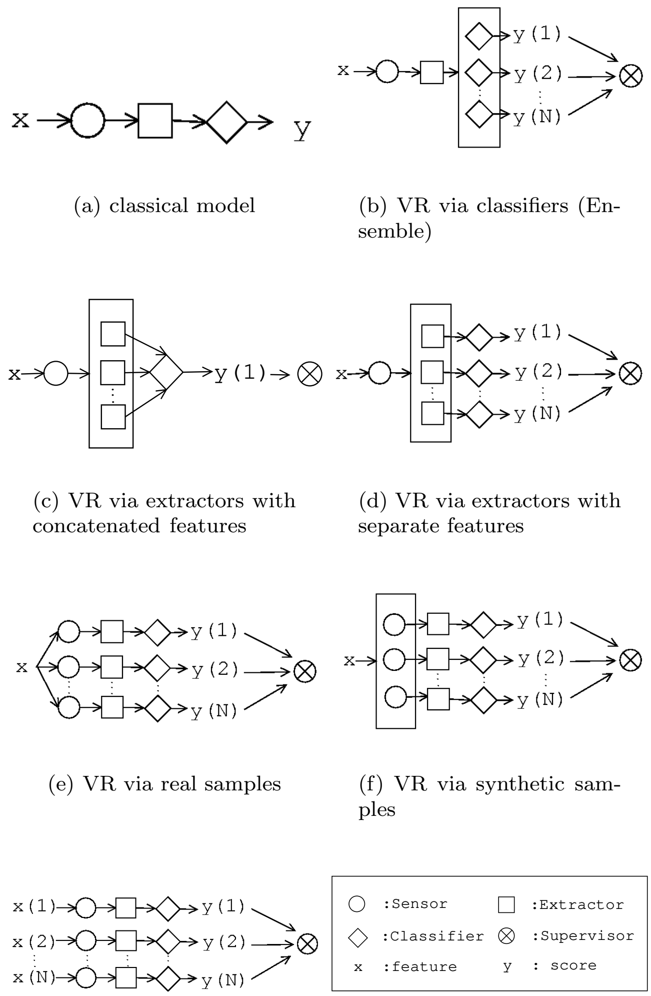
where $\mathcal{I}$, $\mathcal{M}$, $\mathcal{F}$ and $\mathcal{T}$ are sets of instances, modalities, feature extractors and classifier. The first part (Eqns. (2–5)) can be considered as the base-expert while the second part (Eqn. (7)) as the COM. Note that $\overrightarrow{X_F}$ is a vector consisting of y_i and $i = 1, \ldots, N$ when there are N responses. These approaches are shown graphically in Figure 1. Figure 1a is the usual mono-modal biometric approach. One can improve the system by using multiple classifiers (Figure 1b) which can also be called the ensemble method; multiple extractors with concatenated features (Figure 1c); multiple extractors with separate features (Figure 1d); multiple real samples (Figure 1e); multiple synthetic samples (Figure 1f); and multiple biometric modalities (Figure 1g). Each of these approaches is explained in more details in the next subsection. As will become clear in Section 3, exploiting such parallel structure can actually reduce variance of Eqn. (1) and thus can increase the accuracy of the overall system. For this reason, we called these techniques collectively as Variance Reduction (VR) techniques. They are detailed in the next subsection.

2.3 A Survey in the Literature

VR via classifiers is a kind of ensemble method. Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a (weighted) vote of each classifier prediction. The main idea is that ensembles are often much more accurate than the individual classifiers that make them up, provided that the individual classifiers are *accurate* and *diverse* (which includes complementary). These two aspects are examined in Section 3 for a special case of classification error called Equal Error Rate (EER).

Dietterich [15] groups ensemble methods into: (i) Bayesian voting, (ii) manipulation of the training examples (e.g. bagging [4, 37], cross-validated committee and boosting [18]), (iii) manipulation of the input features (e.g. sub-tasking), (iv) manipulation of the output targets (e.g. Error-Correcting Output-Coding (ECOC) [16]) and (v) injection of randomness, also known as learning with noise. Biometric features are very susceptible to noise and different deformation. Therefore, these techniques are important considerations in our framework. Kittler *et al.* [30] have convincingly shown that a modified version of ECOC called multi-seed ECOC improves face recognition on the XM2VTS database. These are all known methods to improve BA systems.

The main idea of **VR via extractors** is that given a raw biometric data, several features are extracted. For example, one can extract the following information from speech features: Linear Predictive Coding Coefficients or Mel Frequency Cepstrum Coefficients [50]. In face verification, common features are principal components, linear discriminant components [10] or more recently independent components [21]. Each feature is often classified by an associated classifier. Since these features are different, one can expect the corresponding trained classifiers to commit different errors. On the often false assumption that features are not correlated, the classifiers are therefore not correlated. In the hope that each classifier operating on different feature space makes different errors, the combined classifier should be able to reduce the errors.



(a) classical model
(b) VR via classifiers (Ensemble)
(c) VR via extractors with concatenated features
(d) VR via extractors with separate features
(e) VR via real samples
(f) VR via synthetic samples
(g) VR via modalities
(h) legend

Figure 1: Different possible VR techniques in BA

There are basically two variations of VR via extractors: with concatenated features (see Figure 1c) and separate features (1d). In the first case, the extracted features are normalised to the same range, concatenated and fed to a common classifier for training and matching. Often the curse of dimensionality [3] is an obstacle to this approach. In the second case, each feature set is treated separately by its own classifier. A decision fusion scheme is required to merge the scores coming from these classifiers. These techniques work because different features usually capture different or complementary information. Since they are extracted from the same sample, the scores from classifiers associated to these features are often correlated.

In the work of Brunelli and Falavigna [6], two speech experts (using respectively static and temporal derivative features) and three face experts (using respectively eye, nose and mouth areas of the face) are used for person verification. The weighted product approach was used to fuse the opinions, with the weights found automatically via a heuristic approach. The static and dynamic feature experts obtained an identification rate of 77% and 71%, respectively. Combining the two speech experts increased the identification rate to 88%. The eye, nose and mouth experts obtained an identification rate of 80%, 77% and 83% respectively. Combining the three facial experts increased the identification rate to 91%. The work shows that **VR via extractors with separate features** improve the accuracy of a BA system.

For the problem of face verification, Marcel and Bengio [39] have shown that instead of using just face images, one can use normalised face colour histogram as an additional feature to the existing normalised face image to train client specific classifiers. This yields an improved classification result.

Luettin [35] investigated the combination of speech and (visual) lip information using feature vector concatenation. In order to match the frame rates of both feature sets, speech information was extracted at 30 fps (frames per second) instead of the usual 100 fps. In text-dependent configuration, the fusion process resulted in a minor performance improvement. However, in text-independent configuration, the performance slightly decreased. This could probably be due to the curse of dimensionality explained earlier. These works [39, 35] showed that **VR via extractors with concatenated features** *may* also improve the accuracy of a BA system.

In this study, we decided to use VR via extractors with separate features for the following reasons: it is better understood; its use can be justified in Section 3; and it is often computationally less expensive compared to its counterpart with concatenated features.

VR via real samples has been demonstrated by Kittler *et al.* [31]. In their work, they combined multiple snapshots of a single biometric property using a Bayesian framework. It is observed that as more and more samples are used, the classification error decreases until a point where it is "saturated", i.e., further increase of samples will not decrease the classification error further.

We have shown that **VR via synthetic samples** [48] is also a viable solution when real samples are not available due to some reasons. For instance, the data transfer bandwidth is limited or taking several biometric samples are inconvenient. This approach works only if such transformation can be found. For face images, geometric transformations can be readily applied without the loss of information. Other image-to-image transformations, i.e., quotient image and methods based on the symmetric property of faces can also be used to normalise the face image against lighting variations. The only constraint is that such transformation itself must not require a lot of client-dependent training data.

Several studies have shown that **VR via different modalities** is superior, on average, to any single-modal biometrics. The following are some strategies proposed in the literature:

- Jain *et al.* [22] have proposed a multimodal biometric system design that integrates face and fingerprints to make a personal identification.

- Ross *et al.* [52] have used hand-scan, fingerprint and face-scan to improve the overall result.

- Poh [44] used eye features and voice features extracted via wavelets to verify a person's identity. Both the face and voice experts are combined using the AND operation. Experiments showed that combining both experts improved the accuracy of the system.

- Dieckmann *et al.* [14] used three experts (frontal face expert, dynamic lip image expert and text-dependent speech expert). A hybrid fusion scheme involving majority voting and opinion fusion was utilised. Two of the experts had to agree on the decision and the combined opinion had to exceed a pre-defined threshold. The hybrid fusion scheme provided better performance than using the underlying experts alone.

- Jourlin *et al.* [29] used a form of weighted summation fusion to combine the opinions of two experts: a text-dependent speech expert and a text-dependent lip expert. It was shown that fusion led to better performance than using the underlying experts alone.

- Sanderson [55] used face and noisy speech information and combined both modalities using adaptive weights and various fusion methods. The resultant system provides a good trade-off in both clean and noise conditions.

The above list is certainly not exhaustive. Most multi-modal approaches yield improvement in results. This is a common and promising approach to improve a BA system. The theoretical aspect of these approaches will be investigated in Sections 3 and 4 while the empirical evidences will be reported in Section 5.

3 Class-Dependent Variance Reduction

The analysis here, based on Equal Error Rate (EER), requires that the class label of the claimant be known in advance. EER is a commonly used performance evaluation criterion in BA and will be defined in Section 4. There are two sets of scores y as defined in Eqn. (1): those of clients and those of impostors, for a given i *response* (instance, modality, feature or classifier). Both sets are written as $y_i^{k=C}$ and $y_i^{k=I}$ here. We adopt the convention that the mean of $y_i^{k=C}$ is greater than that of $y_i^{k=I}$. To fix the idea, Figure 2 plots the client and impostor distributions when they are assumed to be Gaussian. In this case, the client distribution is a Gaussian with mean 1 and variance 0.9, denoted as $\mathcal{N}(1, 0.9)$ and the impostor distribution is $\mathcal{N}(-1, 0.6)$. These are typical plots of client and impostor distributions and are often very close to real score distributions (which are not necessarily Gaussian). Note that the analysis in this section *does not* require the Gaussian assumption but Section 4 will require this assumption.

Suppose $y_{i,j}^k$ is the j-th observed sample of the i-th response of class k, recalling that $i = 1, \ldots, N$ and $k = \{C, I\}$. We assume that this observed variable has a deterministic component and a noise component and that their relation is additive. The deterministic component is due to the fact that the class is discrete in nature, i.e., during authentication, we know that a user in *either* a client or an impostor. The noise component is due to some random processes during biometric acquisition (e.g. degraded situation due to light change, miss-alignment, etc) which in turns affect the quality of extracted features. Indeed, it has a distribution governed by the extracted feature set $f_e\left(s(\mathbf{x})\right)$ often in a non-linear way. By ignoring the source of distortion in extracted biometric features, we actually assume the noise component to be random (while in fact they may be not if we were able to systematically incorporate all possible variations into the base-expert model).

Let μ_i^k be the deterministic component. Note that its value is *only dependent on* the class $k = \{C, I\}$ and independent of j. We can now model $y_{i,j}^k$ as a sum of this deterministic value plus the noise term $w_{i,j}^k$, as follows:

$$y_{i,j}^k = \mu_i^k + w_{i,j}^k, \tag{8}$$

for $k \in \{C, I\}$ where $w_{i,j}^k$ follows an unknown distribution W_i^k with zero mean and $(\sigma_i^k)^2$ variance, i.e., $w_{i,j}^k \sim W_i^k\left(0, (\sigma_i^k)^2\right)$. By adopting such a simple model, from the fusion point of view, we effectively encode the i-th response of a biometric system as the sum of a deterministic value and another random variable, in a class-dependent way. Following Eqn. (8), we can deduce that $y_{i,j}^k \sim Y_i^k \equiv W_i^k\left(\mu_i^k, (\sigma_i^k)^2\right)$. Hence, the expectation of Y_i^k (over different j samples) is:

$$E[Y_i^k] = E[\mu_i^k] + E[W_i^k] = \mu_i^k. \tag{9}$$

Let us consider two cases here. In the first case, for *each access*, N responses are available and are used independently of each other. The *average of variance* of Y_i^k over all $i = 1, \ldots, N$, denoted as $(\sigma_{AV}^k)^2$ is:

$$\begin{aligned}\left(\sigma_{AV}^k\right)^2 &= \frac{1}{N}\sum_{i=1}^{N} Cov(Y_i^k, Y_i^k) \\ &= \frac{1}{N}\sum_{i=1}^{N} E[W_i^k W_i^k] \\ &\equiv \frac{1}{N}\sum_{i=1}^{N} \left(\sigma_i^k\right)^2,\end{aligned} \tag{10}$$

where we adopted the following notation: $Cov(Y_i^k, Y_j^k)$ is the covariance between Y_i^k and Y_j^k, for any $i, j \in \{1, \ldots, N\}$. By definition, $Cov(Y_i^k, Y_j^k) \equiv E[W_i^k W_j^k]$. When $i = j$, we obtain the variance of Y_i^k, which is denoted as $(\sigma_i^k)^2$.

In the second case, all N responses are used together and are combined using the mean operator; the resultant score can be written as:

$$Y_{COM}^k = \frac{1}{N} \sum_{i=1}^{N} Y_i^k, \tag{11}$$

for any $k \in \{C, I\}$. The expected value of Y_{COM}^k, for $k = \{C, I\}$, is:

$$\begin{aligned} \mu_{COM}^k &\equiv E[Y_{COM}^k] \\ &= \frac{1}{N} \sum_{i=1}^{N} E[Y_i^k] \\ &= \frac{1}{N} \sum_{i=1}^{N} \mu_i^k. \end{aligned} \tag{12}$$

The variance of Y_{COM}^k (over many accesses), denoted as $(\sigma_{COM}^k)^2$, is called the *variance of average*, and can be calculated as follows:

$$\begin{aligned} (\sigma_{COM}^k)^2 &= Cov(Y_{COM}^k, Y_{COM}^k) \\ &= E\left[\left(Y_{COM}^k - E[Y_{COM}^k]\right)^2\right] \\ &= E\left[\left(\frac{1}{N}\sum_{i=1}^{N} Y_i^k - \frac{1}{N}\sum_{i=1}^{N} \mu_i^k\right)^2\right] \\ &= E\left[\left(\frac{1}{N}\sum_{i=1}^{N} Y_i^k - \mu_i^k\right)^2\right] \\ &= E\left[\left(\frac{1}{N}\sum_{i=1}^{N} W_i^k\right)^2\right]. \end{aligned} \tag{13}$$

where Eqns. (9) and (11) are used.

To expand Eqn. (13), one should take care of possible correlation between different W_m^k and W_n^k, as follows:

$$\begin{aligned} (\sigma_{COM}^k)^2 &= E\left[\frac{1}{N^2}\left(\sum_{m=1}^{N}\sum_{n=1}^{N} W_m^k W_n^k\right)\right] \\ &= \begin{array}{l} \underbrace{\frac{1}{N^2}\sum_{i=1}^{N} E[W_i^k W_i^k]}\ + \\ \underbrace{\frac{2}{N^2}\sum_{m=1, m<n}^{N} E[W_m^k W_n^k]} \end{array} \\ &\equiv (V_{AV}^k)^2 + (V_{COV}^k)^2, \end{aligned} \tag{14}$$

where $(\sigma_{COM}^k)^2$ is split into the summation of $(V_{AV}^k)^2$ and $(V_{COV}^k)^2$. The first term corresponds to the sum of the diagonal of $E[W_m^k W_n^k]$ while the second term is the sum of all other elements, such that $m \neq n$. $(V_{AV}^k)^2$ measures the average spread (variance) of the base-experts Y_i^k. Using Eqn. (10), it can be further simplified to:

$$\begin{aligned}(V_{AV}^k)^2 &= \frac{1}{N^2}\sum_{i=1}^{N}(\sigma_i^k)^2 \\ &= \frac{1}{N}(\sigma_{AV}^k)^2. \end{aligned} \tag{15}$$

The second term, $(V_{COV}^k)^2$, measures the covariance[1] of error among different W_m^k and W_n^k, for $k = \{C, I\}$. It is related to correlation by:

$$(V_{COV}^k)^2 = \frac{2}{N^2}\sum_{m=1,m<n}^{N} \rho_{m,n}^k \sigma_m^k \sigma_n^k. \tag{16}$$

where $\rho_{m,n}^k$ is the correlation coefficient between Y_m^k and Y_n^k for $k \in \{C, I\}$. The correlation coefficient $\rho_{m,n}^k$ is defined as:

$$\rho_{m,n}^k = \frac{E[W_m^k W_n^k]}{\sigma_m^k \sigma_n^k}. \tag{17}$$

Note that $\rho_{n,n}^k = 1$ and $-1 \leq \rho_{m,n}^k \leq 1$ for all $m, n \in \{1, \dots, N\}$ and $k \in \{C, I\}$.

Now, we need to consider two cases: when W_m^k and W_n^k are uncorrelated (i.e., $\rho_{m,n}^k = 0$) and when they are (i.e., $\rho_{m,n}^k \neq 0$).

3.1 Uncorrelated Assumption: $\rho_{m,n}^k = 0$

In this case, $E[W_m^k W_n^k] = 0$, hence $\rho_{m,n}^k = 0$. As a consequence,

$$(V_{COV}^k)^2 = 0. \tag{18}$$

As a result, substituting Eqns. (15) and (18) into Eqn. (14) will give:

$$(\sigma_{COM}^k)^2 = \frac{1}{N}(\sigma_{AV}^k)^2, \tag{19}$$

which is true when W^m and W^n are not correlated. This is the lowest theoretical bound that $(\sigma_{COM}^k)^2$ can achieve. Basically, this shows that by averaging N scores, the *variance of average* $((\sigma_{COM}^k)^2)$ can be reduced by a factor of N with respect to the *average of variance* $((\sigma_{AV}^k)^2)$, when two instances of Y_m^k and Y_n^k are not correlated, for each class $k = \{C, I\}$.

3.2 Correlated Assumption: $\rho_{m,n}^k \neq 0$

The upper bound can be derived by assuming that W_m and W_n are correlated, i.e. $\rho_{m,n}^k \neq 0$. We will show that the worst-case bound is in fact equal to $(\sigma_{AV}^k)^2$, i.e., there is no gain. To be more explicit, we wish to test the hypothesis:

$$(\sigma_{COM}^k)^2 \leq (\sigma_{AV}^k)^2. \tag{20}$$

Using Eqns. (14), (15) and (16), $(\sigma_{COM}^k)^2$ can be written as follows:

$$\begin{aligned}(\sigma_{COM}^k)^2 = &\underbrace{\frac{1}{N^2}\sum_{i=1}^{N}\left(\sigma_i^k\right)^2}+ \\ &\underbrace{\frac{2}{N^2}\sum_{m=1,m<n}^{N} \rho_{m,n}^k \sigma_m^k \sigma_n^k}, \end{aligned} \tag{21}$$

[1]Note that $(V_{COV}^k)^2$ *could* be negative, hence its square-root, V_{COV}^k, does not represent any meaningful value. The square representation is used here to be consistent with $(V_{AV}^k)^2$ which is strictly positive.

Using Eqns. (21) and (10), the inequality of Eqn. (20) can be expressed as:

$$\frac{1}{N^2}\sum_{i=1}^{N}(\sigma_i^k)^2 + \frac{2}{N^2}\sum_{m=1,m<n}^{N}\rho_{m,n}^k\sigma_m^k\sigma_n^k \leq \frac{1}{N}\sum_{i=1}^{N}(\sigma_i^k)^2 \tag{22}$$

By multiplying both sides by N^2 and rearranging them, we obtain:

$$0 \leq (N-1)\sum_{j=1}^{N}\sigma_j^2 - 2\sum_{m=1,m<n}^{N}\rho_{m,n}\sigma_m\sigma_n. \tag{23}$$

Given that:

$$(N-1)\sum_{i=1}^{N}\sigma_i^2 = \sum_{i=1,i<j}^{N}(\sigma_i^2+\sigma_j^2) \tag{24}$$

(the proof can be found in the Appendix A) the inequality can further be simplified to:

$$\begin{aligned}
0 &\leq \sum_{m=1,m<n}^{N}\left((\sigma_m^k)^2+(\sigma_n^k)^2\right) - 2\sum_{m=1,m<n}^{N}\rho_{m,n}^k\sigma_m^k\sigma_n^k \\
0 &\leq \sum_{m=1,m<n}^{N}\left((\sigma_m^k)^2 - 2\rho_{m,n}^k\sigma_m^k\sigma_n^k + (\sigma_n^k)^2\right) \\
0 &\leq \sum_{m=1,m<n}^{N}\left((\sigma_m^k-\rho_{m,n}^k\sigma_n^k)^2 + (1-\rho_{m,n}^k)(\sigma_n^k)^2\right).
\end{aligned}$$

In other words, hypothesis in Eqn. (22) is always true, regardless of the value $\rho_{m,n}^k$. As a consequence, we have just validated the inequality of Eqn. (20). Taking this conclusion and that of Eqn. (19), one can conclude that:

$$\frac{1}{N}(\sigma_{AV}^k)^2 \leq (\sigma_{COM}^k)^2 \leq (\sigma_{AV}^k)^2. \tag{25}$$

Referring back to Eqn. (21) if $\rho_{m,n}^k < 0$ (i.e., negatively correlation), then $(V_{COV}^k)^2$ would be negative and consequently $(\sigma_{COM}^k)^2 \leq \frac{1}{N}(\sigma_{AV}^k)^2$! Obviously, negative correlation would help improve the results. However, and unfortunately, in reality, negative correlation will not happen if the underlying experts are trained separately, i.e., for a given instance i, y_i^k for $i = 1,\ldots,N$, will tend to agree with each other (hence positively correlated) most often than to disagree with each other (hence negatively correlated). One possible exception will be that the experts are specifically trained to be decorrelated or even negatively correlated *in a collaborative way*, e.g. [5]. By fusing scores obtained from experts that are trained independently (which is often so in multimodal fusion), one can almost be certain that $0 \leq \rho_{m,n}^k \leq 1$.

3.3 Introduction of α as a gain factor

To measure *explicitly* the factor of reduction, we introduce α, which can be defined as follows:

$$\alpha = \frac{(\sigma_{AV}^k)^2}{(\sigma_{COM}^k)^2}. \tag{26}$$

By dividing Eqn. (25) by $(\sigma_{COM}^k)^2$ and rearranging it, we can deduce that:

$$1 \leq \alpha \leq N. \tag{27}$$

One direct implication of variance reduction is that **the more hypotheses used** (increasing N), **the better the combined system**, even if the hypotheses of underlying experts are correlated. This will come at a cost of more computation proportional to N. Experiments in [57] (in speech recognition) and [31] (in face verification) provide strong empirical evidences to support this claim. The above analysis, however, was only possible when the class-labels are already known in advance. This analysis is relevant for regression problems, by simply removing the class-label k from the equations (since in regression problems there is no class label). However, it is not clear how (class-dependent) variance reduction can lead to better classification, often measured in terms of false rejection rate (FRR) and false acceptance rate (FAR) in BA tasks. These two error terms will be defined and treated in the next section.

4 EER Reduction

This section first presents, intuitively, how reduced variance will result in reduced classification error before working it out in a more detailed manner. We then show how F-ratio can be derived and how it is related to EER. Finally, how reduced class-dependent variance can lead to reduced EER will be explained.

Figure 2 illustrates the effect of averaging scores in a two-class problem, such as in BA where an identity claim could belong either to a client or an impostor. Let us assume that the genuine user scores in a situation where 3 samples are available but are used separately, follow a normal distribution of mean 1.0 and variance ($\sigma^2_{AV}(\mathbf{x})$ of genuine users) 0.9, i.e., $\mathcal{N}(1, 0.9)$, and that the impostor scores (in the mentioned situation) follow a normal distribution of $\mathcal{N}(-1, 0.6)$ (both graphs are plotted with "+"). If for each access, the 3 scores are used, according to Eqn. (27), the variance of the resulting distribution will be reduced by a factor of 3 or less. Both resulting distributions are plotted with "o". Note the area where both the distributions overlap before and after. The latter area is shaded in Figure 2. This area corresponds to the zone where minimum amount of mistakes will be committed given that the threshold is optimal[2]. Decreasing this area implies an improvement in the performance of the system.

Let the scores' probability density function (*pdf*) be $P(Y_i^{k=C})$ for the client scores and $P(Y_i^{k=I})$ for the impostor scores and i indicates that it belongs to the i-th base-expert[3]. Let us first assume that these *pdf*s are Gaussians and have mean μ_i^k and standard deviation σ_i^k, for $k = \{C, I\}$. FRR is a threshold-dependent measure and is defined as a ratio between the total number of wrongly rejected client accesses and the total number of client accesses. FAR is also a threshold-dependent measure and is defined as a ratio between the total number of wrongly accepted impostor accesses and the total number of impostor accesses. FRR and FAR can then be written as:

$$\begin{aligned}
\text{FRR}(\Delta) = & \int_{-\infty}^{\infty} P(Y_i^{k=C} = y) dy \\
= & \int_{-\infty}^{\infty} \frac{1}{\sigma_i^C \sqrt{2\pi}} \exp\left[\frac{-(y-\mu_i^C)^2}{2(\sigma_i^C)^2}\right] dy \\
= & \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{\Delta - \mu_i^C}{\sigma_i^C \sqrt{2}}\right), \text{ and}
\end{aligned} \tag{28}$$

[2]Optimal in the Bayes sense, when (1) the cost and (2) probability of both types of errors are equal.

[3]This analysis applies to any independent experiments and not necessary base-experts. For instance, by replacing i with COM, the findings can be applied to *fusion* experiments as well.

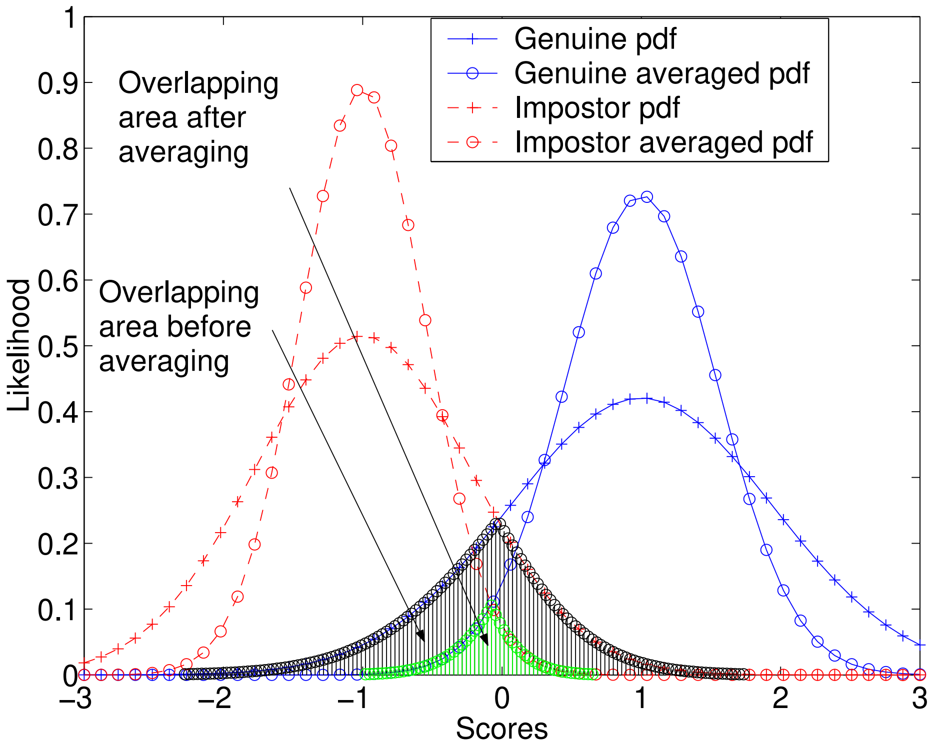


Figure 2: Averaging score distributions in a two-class problem

$$\begin{aligned}
\mathrm{FAR}(\Delta) &= \int_{\Delta}^{\infty} P(Y_i^{k=I} = y)dy \\
&= 1 - \int_{-\infty}^{\Delta} P(Y_i^{k=I} = y)dy \\
&= 1 - \left[\frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{\Delta - \mu_i^I}{\sigma_i^I \sqrt{2}}\right)\right] \\
&= \frac{1}{2} - \frac{1}{2}\mathrm{erf}\left(\frac{\Delta - \mu_i^I}{\sigma_i^I \sqrt{2}}\right),
\end{aligned} \tag{29}$$

where

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp\left[-t^2\right] dt, \tag{30}$$

which is the so-called error function. Note that the use of an error function for such analysis has been reported in [12], but with differences in the definition of the error function. In another similar work (but limited to the context of combining multiple samples) [31], the Equal Error Rate (EER) curve was not calculated explicitly and validated via experiments as done here. Furthermore, the issue on how the dependency among samples affects the resultant variance was not studied theoretically as done in Section 3.

The minimal error happens when:

$$\mathrm{FAR}(\Delta) = \mathrm{FRR}(\Delta) = \mathrm{EER}(\Delta),$$

i.e., the Equal Error Rate. Making these two terms equal (Eqns. (28) and (29)) and using the property

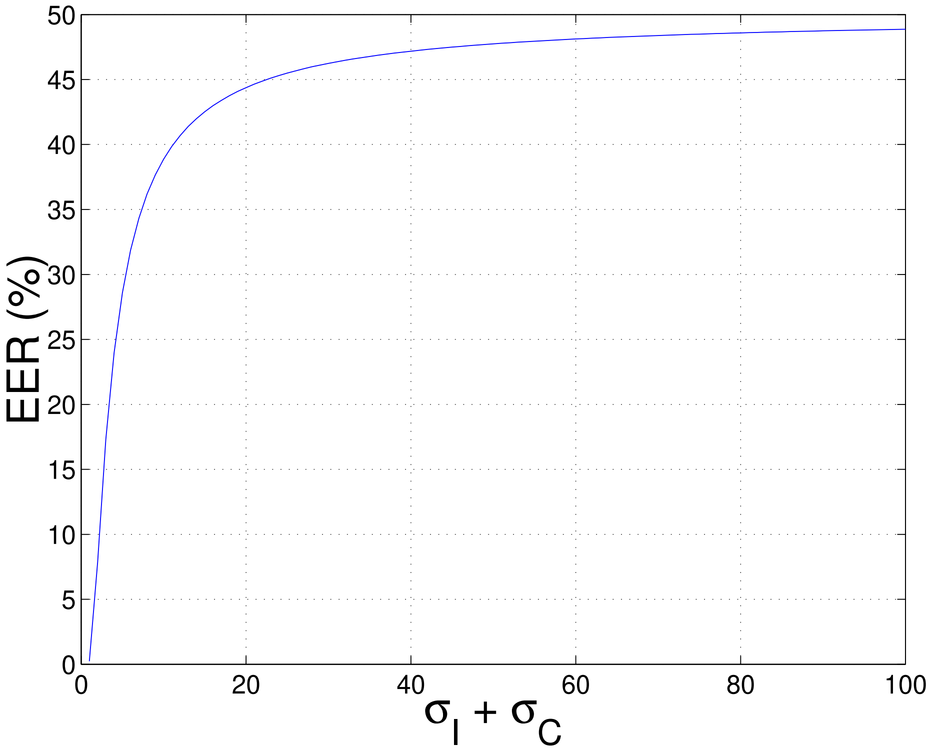


Figure 3: Equal error rate versus the sum of standard deviations of client and impostor scores

that $\text{erf}(-z) = -\text{erf}(z)$, we can deduce that:

$$\Delta = \frac{\mu_i^I \sigma_i^C + \mu_i^C \sigma_i^I}{\sigma_i^I + \sigma_i^C}. \tag{31}$$

By introducing Eqn. (31) into Eqn. (29) (or equivalently into Eqn. (28)), we obtain:

$$\text{EER} = \frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{\text{F-ratio}}{\sqrt{2}}\right), \tag{32}$$

where

$$\text{F-ratio} = \frac{\mu_i^C - \mu_i^I}{\sigma_i^C + \sigma_i^I} \tag{33}$$

F-ratio has a particular meaning. It is in fact similar to Fisher-ratio. For a two-class problem, using the same terms, the Fisher-ratio [3, pg. 107] is defined as:

$$\frac{\mu_i^{k=C} - \mu_i^{k=I}}{(\sigma_i^{k=C})^2 + (\sigma_i^{k=I})^2} \tag{34}$$

F-ratio or equally Fisher-ratio measures the degree of separability (ability to discriminate) between the client and impostor scores. The higher this ratio is, the further and better the scores are discriminated and hence the lower EER will be.

To give an idea how EER would look like, let us first suppose that $\mu_C = 1$ and $\mu_I = -1$. EER as a function of $\sigma_i^C + \sigma_i^I$ is then plotted, as shown in Figure 3. EER appears to be a monotonically increasing function as the sum of standard deviations of client and impostor distributions increases.

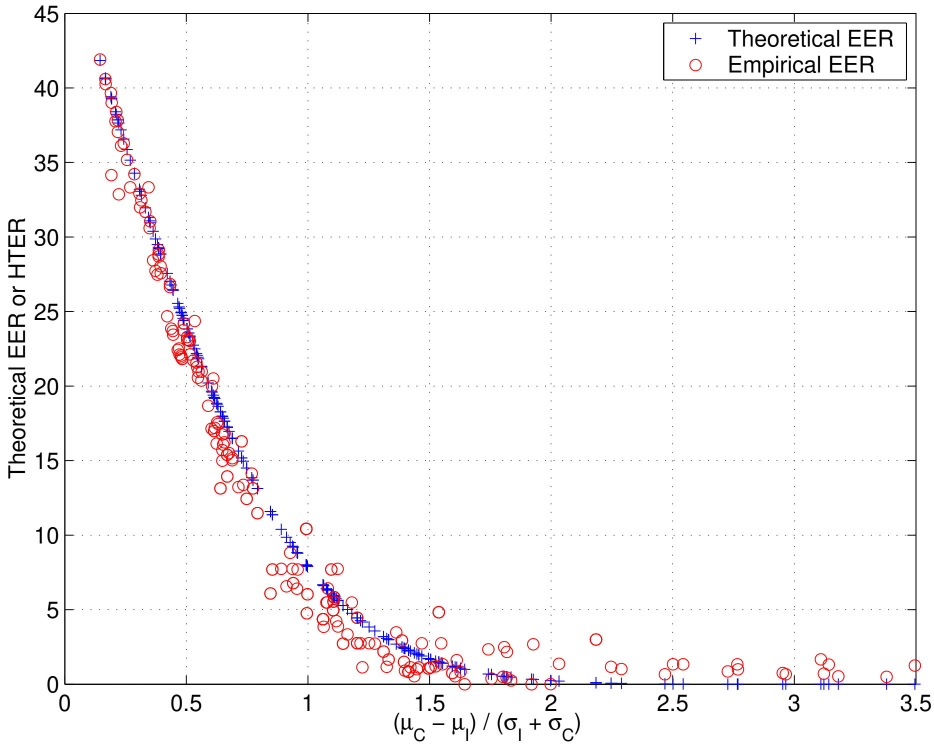


Figure 4: A Comparison between EER and HTER versus F-ratio, carried out on 180 independent experiments on XM2VTS (45 experiments), BANCA (63 experiments) and NIST2001 (72 experiments). Details can be found in [46].

Intuitively, this means that smaller class-dependent standard deviations are desirable to attain lower EER.

We call the EER based on Gaussian assumption, presented in the last section, the *theoretical EER*, to distinguish it from the *empirical EER*, which is calculated by direct minimization of threshold according to the following criterion:

$$\Delta^* = \arg\min_{\Delta} |\text{FAR}(\Delta) - \text{FRR}(\Delta)| \tag{35}$$

and approximated by the commonly used Half Total Error Rate:

$$\text{HTER} = \frac{\text{FAR}(\Delta^*) + \text{FRR}(\Delta^*)}{2}. \tag{36}$$

The so-called "empirical EER" in this case is HTER with threshold Δ^*. Empirical EER and HTER are used interchangeably in this paper.

To check how accurate the EER function (as in Eqn. (32)) is as compared to its empirical counterpart (as in Eqn. (36)), we conducted many experiments on XM2VTS, NIST2001 and BANCA score datasets taken from [45, 47, 38], respectively. The results are shown in Figure 4. Details on how these experiments were carried out were reported in [46]. By visual inspection, it can be seen that the theoretical EER as a function of F-ratio is quite accurate at the high ends of HTER and the accuracy decreases as HTER decreases. The degree of deviation is proportional to how well the real underlying distributions (of the client and the impostor scores) follow the Gaussian distribution at *a given EER point*. The reason for such a deviation is that due to finite number of data (client and impostor) accesses, empirical FAR and FRR are not smooth functions. As a result, small change of the threshold Δ will cause a big change in HTER. Nevertheless, the theoretical EER *does* reflect the actual HTER fairly accurately.

Using the notations in Section 3, let σ^I_{COM} and σ^C_{COM} (as in Eqn. (21)) be the standard deviations of the fused scores (using the mean operator) of both the impostor and client distributions, respectively. These definitions also apply to the average of the standard deviations σ^I_{AV} and σ^C_{AV} (as in Eqn. (10)).

From Eqn. (25), we can deduce that:

$$\sigma^I_{COM} \leq \sigma^I_{AV} \text{ and } \sigma^C_{COM} \leq \sigma^C_{AV}.$$

Since EER is a monotonically increasing function as shown in Figure 3, these inequalities imply that:

$$\mathrm{EER}(\sigma^I_{COM}, \sigma^C_{COM}) \leq \mathrm{EER}(\sigma^I_{AV}, \sigma^C_{AV}),$$

when both the μ_C and μ_I are *normalised* such that they are constant across different streams, bands and modalities[4].

In fact, without assuming the Gaussian distribution, as long as the EER function has a monotonically increasing behaviour with respect to $\sigma_I + \sigma_C$, the above conclusions remain valid.

To require that EER be a monotonically increasing function, the necessary condition is that the right tail of the impostor *pdf* is a decreasing function and the left tail of the client *pdf* is an increasing function. A Gaussian function exhibits such behaviour on its left and right tails. Unfortunately, in the case of non-Gaussian *pdf*s, the analytical analysis such as the one done here is more difficult.

To evaluate the improvement due to variance reduction, we can define a gain factor β, similar to α defined in Eqn. (26), as follows:

$$\beta_{mean} = \frac{\mathrm{mean}_i(\mathrm{EER}_i)}{\mathrm{EER}_{COM}} \tag{37}$$

where EER_{COM} is the EER of the combined system (with reduced variance) and EER_i is the EER of the i-th system. Indeed, *all experiments* reported in Section 5 verified that $\beta_{mean} \geq 1$, which is theoretically achievable. β_{mean} can only measure the relative improvement with respect to the average EER of the underlying expert. In practice, one wishes to know whether the resultant combined expert is better than the *best* underlying expert. This can be measured using:

$$\beta_{min} = \frac{\min_i(\mathrm{EER}_i)}{\mathrm{EER}_{COM}}, \tag{38}$$

which is defined very similarly to β_{mean}, except that the minimum EER of the underlying experts is used. $\beta_{min} \geq 1$ implies that the resultant expert is better than the best underlying expert.

All the VR techniques discussed can be evaluated using β_{mean} and β_{min}, except VR via synthetic samples. This is because, the new system using synthetic samples uses the real samples in the original system. As a result, β_{min} is not meaningful anymore. A meaningful comparison will be to use the following "gain" ratio:

$$\beta_{real} = \frac{\mathrm{EER}_{real}}{\mathrm{EER}_{COM}}, \tag{39}$$

In fact, for both β_{mean} and β_{min}, $(\beta^{-1} - 1) \times 100\%$ measures the relative change of the EER of the combined expert with respect to the average EER or the minimum EER of the underlying experts.

5 Empirical Results

5.1 XM2VTS Database

The XM2VTS database [40] contains synchronized video and speech data from 295 subjects, recorded during four sessions taken at one month intervals. On each session, two recordings were made, each

[4]This normalization will affect the corresponding variance [46, Sec. VI] but should not change the F-ratio.

consisting of a speech shot and a head shot. The speech shot consisted of frontal face and speech recordings of each subject during the recital of a sentence.

The database is divided into three sets: a training set, an evaluation set and a test set. The training set was used to build client models, while the evaluation set (Eval) was used to compute the decision thresholds (as well as other hyper-parameters) used by classifiers. Finally, the test set (Test) was used to estimate the performance.

The 295 subjects were divided into a set of 200 clients, 25 evaluation impostors and 70 test impostors. There exists two configurations or two different partitioning approaches of the training and evaluation sets. They are called Lausanne Protocol I and II, denoted as **LP1** and **LP2** in this paper. Thus, besides the data for training the model, the following four data sets are available for evaluating the performance: LP1 Eval, LP1 Test, LP2 Eval and LP2 Test. Note that LP1 Eval and LP2 Eval are used to calculate the optimal thresholds that will be used in LP1 Test and LP2 Test, respectively. Results are reported only for the test sets, in order to be as unbiased as possible (using an *a priori* selected threshold). Table 1 is the summary of how the data sets are partitioned. In both configurations, the test set remains the same. However, there are three training shots per client for LP1 and four training shots per client for LP2. More details can be found in [36].

5.2 Summary of Fusion Experiments

As many as 13 baseline experiments and 104 fusion experiments were carried out to evaluate various VR techniques on the common XM2VTS database. The baseline experiments have been reported elsewhere in the literature (see Appendix B) whereas the fusion experiments appeared in [45]. However, a very brief summary of these experiments were included in Appendix B[5]. The baseline system performances can be found in Table 2. The 104 fusion experiments are divided as follow: 63 (21 datasets $\times$ 3 fusion methods) multimodal fusion experiments (see Table 3), 27 (9 datasets $\times$ 3 methods) multi-feature fusion experiments, 6 (2 datasets $\times$ 3 methods) multi-classifier experiments (see Table 4) and 8 (2 datasets $\times$ 4 methods) synthetic multi-sample experiments (see Table 5). Each of these experiments corresponds to the following techniques: VR via modalities, extractors, classifiers and synthetic samples, as surveyed in Section 2.2. The experiments are summarised by evaluating their β_{mean} and β_{min}, as shown in Figures 5 and 6, respectively. Note that the β's for VR via synthetic samples are actually β_{real}, as clarified in Section 4 already.

In both figures, the VR techniques are ordered such that the one with the highest computation and/or physical cost is ranked first while the one with the lowest cost is ranked last. VR via modalities has the highest cost since adding a biometric modality requires additional hardware (sensor) and software (feature extractor, classifier module, etc). To implement VR via extractors, only additional

[5]The goal of this paper is to provide evidences that EER can be reduced due to VR techniques. This section essentially provide empirical evidences and it complements the theoretical evidences already examined in Section 3 and 4. Due to overwhelmingly many details involved, we cannot discuss how those baseline/fusion experiments are carried out in this paper. Interested readers are strongly encouraged to refer to [45].

Table 1: The Lausanne Protocols of XM2VTS database

Data sets	Lausanne Protocols	
	LP1	LP2
Training client accesses	3	4
Evaluation client accesses	600 (3×200)	400 (2×200)
Evaluation impostor accesses	40,000 ($25 \times 8 \times 200$)	
Test client accesses	400 (2×200)	
Test impostor accesses	112,000 ($70 \times 8 \times 200$)	

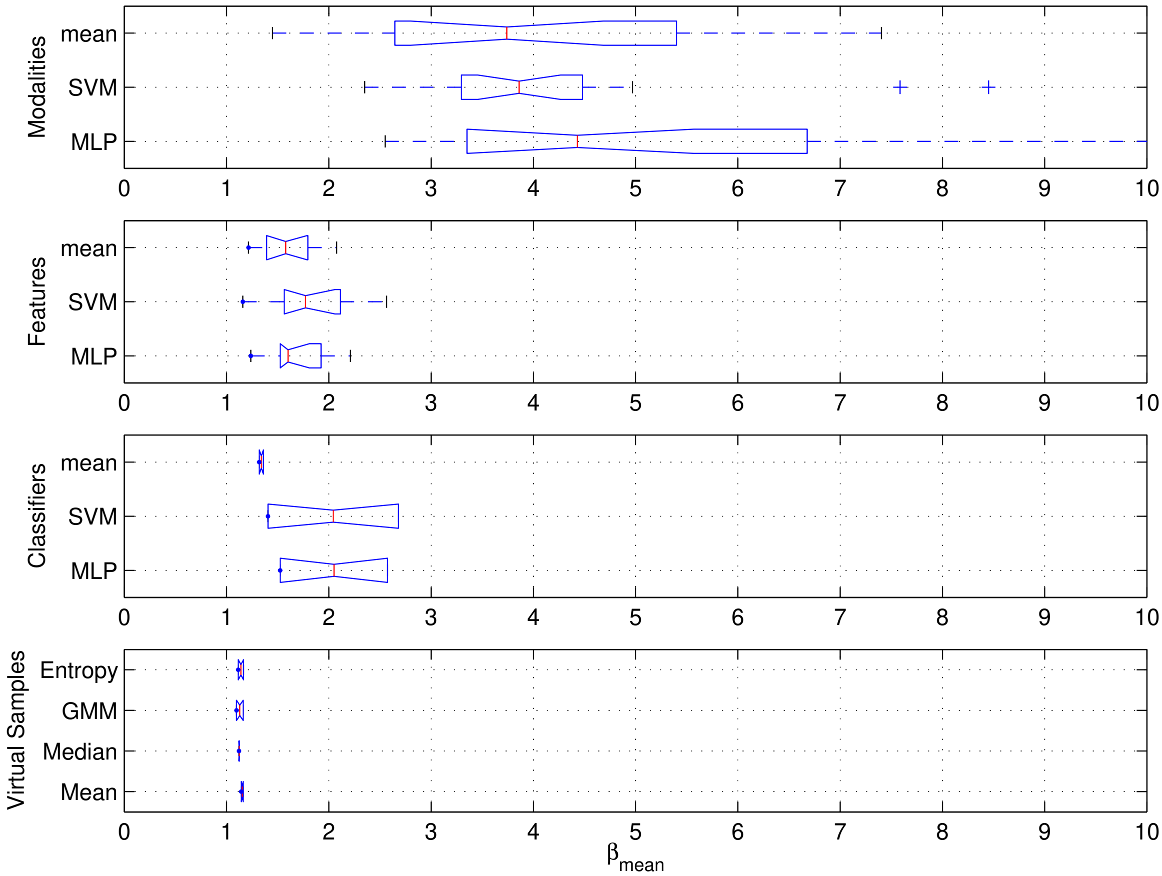


Figure 5: Boxplot of β_{mean}. Each bar shows the data within 95% of confidence. The vertical line around the middle of each bar is the median of β_{mean}. Dotted lines at each end of a bar are extreme values laying outside the 95% confidence interval. For VR via synthetic samples, β_{real} is used in place of β_{mean}. The x-axis of all the boxplots are aligned so that β_{mean} across different VR techniques are comparable.

feature extractors and classifiers are required. The implementation cost of VR via classifiers is less than that of VR via extractors because the former does not need feature extractors[6].

Recall that β_{mean} compares the average performance of baseline systems against the combined system whereas β_{min} compares the *best* baseline system against the combined system. For both cases, $\beta \geq 1$ implies that the combined system is better. Figure 5 shows that β_{mean} of all the VR techniques are greater than or equal to 1. This confirms the theoretical findings in Sections 3 and 4.

On the other hand, not all VR techniques can achieve $\beta_{min} \geq 1$. In both cases, VR via modalities outperform other VR techniques. It is not obvious that VR via extractors are better than VR via classifiers. It should be noted that there are only two experiments in VR via classifiers *for each method* compared to 9 experiments in VR via extractors. Hence, it is difficult to compare both sets of experiments. It is certain that, however, VR via synthetic/virtual samples brings the least gain.

One conclusion that can be drawn from here is that **higher diversity will incur higher com-**

[6]Feature extractors can be seen as a pre-processing to classifiers. In practice, this pre-processing *improves* class discriminability. In some situations, the boundary between input features and "extracted" features is not clear. As a result, the difference between diversity due to features and due to classifiers is subtle. The discussion here is principally motivated by architectural diversity.

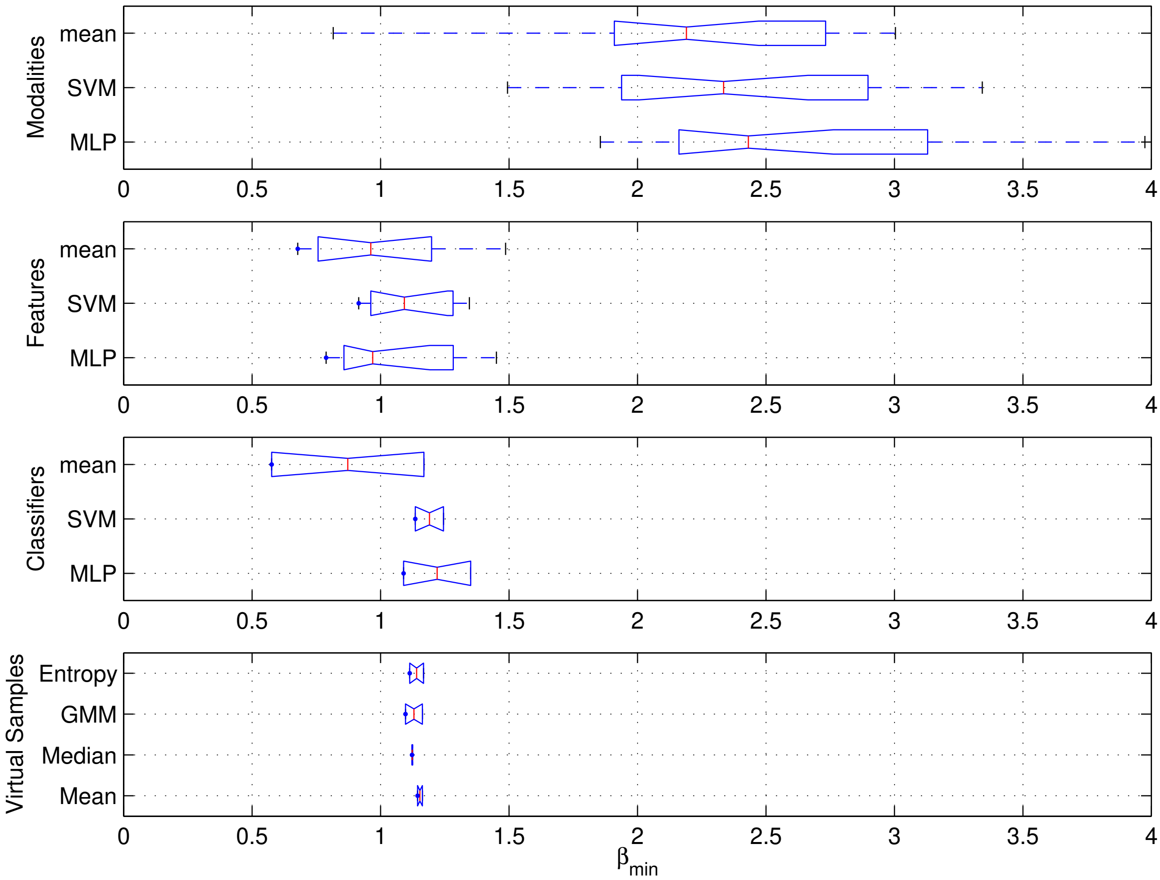


Figure 6: Boxplot of β_{min}. Each bar shows the data within 95% of confidence. The vertical line around the middle of each bar is the median of β_{min}. Dotted lines at each end of a bar are extreme values laying outside the 95% confidence interval. For VR via synthetic samples, β_{real} is used in place of β_{min}. The x-axis of all the boxplots are aligned so that β_{min} across different VR techniques are comparable.

putation/hardware cost, and vice versa. Apart from the implementation cost, there is also a matching *user convenience* cost. In general, by adding more biometric sensors, the system becomes less convenient to the user, if the user was obliged to provide all the biometric data. From the engineering point of view, one should achieve a balance between the level of accuracy that is required and the different "costs" associated to increasing the diversity via VR techniques.

In terms of the type of fusion methods to use, we could categorise the mean operator as non-trainable fusion whereas support vector machines (SVMs) and Multi-Layer Perceptrons (MLPs) as trainable fusion. For VR via modalities, extractors and classifiers, it seems that trainable fusion performs at least as good as non-trainable fusion, if not better. On the other hand, the choice of classifier is less obvious for VR via synthetic samples.

6 Conclusions

Multimodal biometric authentication has shown perennial successes both in research and applications. This paper casts a light on why BA systems can be improved by fusion (fusing opinions of different experts), principally due to diversity of biometric modalities, features, classifiers and samples. We call these techniques collectively as *Variance Reduction (VR) techniques* because fusion effectively reduces class-dependent variance. The mentioned diversity are hence called VR via modalities, extractors/features, classifiers and samples. A survey in the literature showed that these interrelated techniques have already been presented elsewhere. Unfortunately, until now, there is no systematic comparative study of these techniques yet. This somewhat motivated this paper.

We showed that by reducing class-dependent variance (of client and impostor scores), the Equal Error Rate is effectively reduced. By assuming that the client and impostor scores are generated by two Gaussians, we derived a measure of quality called "F-ratio", which happens to be very similar to the Fisher-ratio criterion. EER is shown to be a function of F-ratio. In fact, EER calculated from F-ratio can roughly approximate EER measured directly from the score data. Different from most analysis, our analysis does not make the assumption that the baseline expert opinions are uncorrelated. Indeed, the correlation is taken into account by the covariance matrix of different baseline experts.

Having acquired a better understanding of EER, we proceed to show that reduced class-dependent variance leads to reduced EER. In addition to the theoretical evidence, empirical evidences on 104 fusion experiments confirmed that the EER of the combined system is guaranteed to be lower than the average EER of all the baseline systems. Unfortunately, there is no evidence that the EER of the combined system is better than the minimum EER of all the baseline systems. The experiments *do* confirm that VR via modalities is the most effective way to reduce EER among the VR techniques while VR via samples is the least effective. Interestingly, the most effective technique incurs also the highest cost – in terms of additional computation, hardware and user inconvenience, and vice-versa for the least effective technique, i.e., VR via samples. Although the VR techniques were tested on XM2VTS database, we conjecture that the same trend will be observed if different databases have had been used.

Based on the 104 fusion experiments, trainable fusion classifiers (using Multi-Layer Perceptrons and Support Vector Machines) is at least as good as non-trainable fusion classifiers (provided that the capacity (degree of flexibility) of the trainable classifiers are tuned correctly) for all the VR techniques investigated, except for VR via synthetic samples. For this case, there is no evidence that a trainable classifier is better than a non-trainable one. The success of fusion hence is attributed to two aspects: the underlying diversity among different baseline systems; and the discriminative power of the combination mechanism (fusion). Although both aspects are important, we conjecture that the first aspect has a higher influence than the second aspect for all VR techniques except VR via synthetic samples; in other words, it is sufficient to merge scores obtained from VR via synthetic samples using the mean operator.

This paper could serve as a useful guide from engineering point of view, since choosing a particular VR technique will bring about a given level of accuracy but at the same time also incur a particular cost. As such, in biometric applications involving the highest level of security, VR via modalities should be used and with it, user's convenience may be tolerated.

A Proof of $(N-1)\sum_{i=1}^N \sigma_i^2 = \sum_{i=1,i<j}^N(\sigma_i^2+\sigma_j^2)$

Let σ_i be a random variable and $i = 1,\dots,N$. The term $\sum_{i=1,i<j}^N(\sigma_i^2+\sigma_j^2)$ can be interpreted as $\sum_{i=1}^N\sum_{j=i+1}^N(\sigma_i^2+\sigma_j^2)$. The problem now is to count how many σ_k^2 there are in the term, for any $k = 1,\dots,N$.

There are two cases here. The first case is when $i = k$, the term $\sum_{i=1}^N\sum_{j=i+1}^N(\sigma_i^2+\sigma_j^2)$ becomes: $\sum_{j=k+1}^N(\sigma_k^2+\sigma_j^2)$. There are $(N-k)$ terms of σ_k^2.

In the second case, when $j = k$, the term $\sum_{i=1}^N\sum_{j=i+1}^N(\sigma_i^2+\sigma_j^2)$ then becomes: $\sum_{i=1}^{k-1}(\sigma_i^2+\sigma_k^2)$. There are $(k-1)$ terms of σ_k^2.

The total number of σ_k^2 is just the sum of these two cases, which is $(N-k)+(k-1) = (N-1)$, for any k drawn from $1,\dots,N$. The sum of $(N-1)$ σ_k^2 over all possible $k = 1,\dots,N$ then gives

$(N-1)\sum_{k=1}^{N}\sigma_k^2$. Therefore, $(N-1)\sum_{i=1}^{N}\sigma_i^2 = \sum_{i=1,i<j}^{N}(\sigma_i^2+\sigma_j^2)$. $\square$

B Baseline Experts used on the XM2VTS Database

B.1 Face and Speech Features

The face baseline experts are based on the following features:

1. **FH**: normalised **f**ace image concatenated with its RGB **H**istogram (thus the abbreviation **FH**) [39].

2. **DCTs**: DCTmod2 features [56] extracted from face images with a size of 40×32 (rows $\times$ columns) pixels. The DCT coefficients are calculated from an 8×8 window with horizontal and vertical overlaps of 50%, i.e., 4 pixels in each direction. Neighbouring windows are used to calculate the "delta" features. The result is a set of 35 feature vectors, each having a dimensionality of 18. (**s** indicates the use of this small image compared to the bigger size image with the abbreviation **b**.)

3. **DCTb**: Similar to DCTs except that the input face image has 80×64 pixels. The result is a set of 221 feature vectors, each having a dimensionality of 18.

The speech baseline experts are based on the following features:

1. **LFCC**: The Linear Filter-bank Cepstral Coefficient (LFCC) [50] speech features were computed with 24 linearly-spaced filters on each frame of Fourier coefficients sampled with a window length of 20 milliseconds and each window moved at a rate of 10 milliseconds. 16 DCT coefficients are computed to decorrelate the 24 coefficients (log of power spectrum) obtained from the linear filter-bank. The first temporal derivatives are added to the feature set.

2. **PAC**: The PAC-MFCC features [25] are derived with a window length of 20 milliseconds and each window moves at a rate of 10 milliseconds. 20 DCT coefficients are computed to decorrelate the 30 coefficients obtained from the Mel-scale filter-bank. The first temporal derivatives are added to the feature set.

3. **SSC**: These features, originally proposed for speech recognition [42], were used for speaker authentication in [49]. It was found that mean-subtraction could improve these features significantly. The mean-subtracted SSCs are obtained from 16 coefficients. The γ parameter, which is a parameter that raises the power spectrum and controls how much influence the centroid, is set to 0.7. Also The first temporal derivatives are added to the feature set.

B.2 Classifiers

Two different types of classifiers were used for these experiments: Multi-Layer Perceptrons (MLPs) and a Bayes Classifier using Gaussian Mixture Models (GMMs) [3]. While in theory both classifiers could be trained using any of the previously defined feature sets, in practice MLPs are better at matching feature vectors of fixed-size while GMMs are better at matching sequences (feature vectors of unequal size). Whatever the classifier is, the hyper-parameters (e.g. the number of hidden units for MLPs or the number of Gaussian components for GMMs) are tuned on the evaluation set LP1 Eval. The same set of hyper-parameters are used in both LP1 and LP2 configurations of the XM2VTS database.

For each client-specific MLP, the feature vectors associated to the client are treated as positive patterns while all other feature vectors *not* associated to the client are treated as negative patterns. All MLPs reported here were trained using the stochastic version of the error-back-propagation training algorithm [3].

For the GMMs, two competing models are often needed: a world and a client-dependent model. Initially, a world model is first trained from an external database (or a sufficiently large data set) using the standard Expectation-Maximisation algorithm [3]. The world model is then adapted for each client to the corresponding client data using the Maximum-A-Posteriori adaptation [19] algorithm.

B.3 Baseline Systems

The baseline experiments based on DCTmod2 feature extraction were reported in [8] while those based on normalised face images and RGB histograms (FH features) were reported in [39]. Details of the experiments, coded in the pair **(feature, classifier)**, for the face experts, are as follows:

1. **(FH, MLP)** Features are normalised **F**ace concatenated with **H**istogram features. The client-dependent classifier used is an MLP with 20 hidden units. The MLP is trained with geometrically transformed images [39].

2. **(DCTs, GMM)** The face features are the DCTmod2 features calculated from an input face image of 40×32 pixels, hence, resulting in a sequence of 35 feature vectors each having 18 dimensions. There are 64 Gaussian components in the GMM. The world model is trained using *all the clients* in the training set [8].

3. **(DCTb, GMM)** Similar to (DCTs,GMM), except that the features used are DCTmod2 features calculated from an input face image of 80×64 pixels. This produces in a sequence of 221 feature vectors each having 18 dimensions. The corresponding GMM has 512 Gaussian components [8].

4. **(DCTs, MLP)** Features are the same as those in (DCTs,GMM) except that an MLP is used in place of a GMM. The MLP has 32 hidden units [8]. Note that in this case a training example consists of a *big single* feature vector with a dimensionality of 35×18. This is done by simply concatenating 35 feature vectors each having 18 dimensions[7].

5. **(DCTb, MLP)** The features are the same as those in (DCTb,GMM) except that an MLP with 128 hidden units is used. Note that in this case the MLP in trained on a *single* feature vector with a dimensionality of 221×18 [8].

and for the speech experts:

1. **(LFCC, GMM)** This is the Linear Filter-bank Cepstral Coefficients (LFCC) obtained from the speech data of the XM2VTS database. The GMM has 200 Gaussian components, with the minimum relative variance of each Gaussian fixed to 0.5, and the MAP adaptation weight equals 0.1. This is the best known model currently available [47] under clean conditions.

2. **(PAC, GMM)** The same GMM configuration as in LFCC is used. Note that in general, 200-300 Gaussian components would give about 1% of difference of HTER [47]. This system is particularly robust to very noisy conditions (more than 18 dBs, as tested on the NIST2001 one-speaker detection task).

3. **(SSC, GMM)** The same GMM configuration as in LFCC is used [49]. This system is known to provide an optimal performance under moderately noisy conditions (18-12 dBs, as tested on NIST2001 one-speaker detection task).

The baseline performances are shown in Table 2. The fusion experiments based on VR via modalities are shown in Table 3; VR via extractors in Table 4 (rows 1–9); VR via classifiers in Table 4 (rows 10–11); and VR via synthetic samples in Table 5.

[7]This may explain why MLP, an inherently discriminative classifier, has worse performance compared to GMM, a generative classifier. With high dimensionality yet having only a few training examples, the MLP cannot be trained optimally. This may affect its generalisation on unseen examples. By treating the features as a sequence, GMM was able to generalise better and hence is more adapted to this feature set.

Table 2: Baseline performance in HTER(%) of different modalities evaluated on XM2VTS based on *a priori* selected thresholds

Data sets	(Features, classifiers)	FAR	FRR	HTER
Face LP1 Test	(FH,MLP)	1.751	2.000	1.875
Face LP1 Test	(DCTs,GMM)	4.454	4.000	4.227
Face LP1 Test	(DCTb,GMM)	1.840	1.500	1.670
Face LP1 Test	(DCTs,MLP)	3.219	3.500	3.359
Face LP1 Test	(DCTb,MLP)	4.443	8.000	6.221
Speech LP1 Test	(LFCC,GMM)	1.029	1.250	1.139
Speech LP1 Test	(PAC,GMM)	4.608	8.000	6.304
Speech LP1 Test	(SSC,GMM)	2.374	2.500	2.437
Face LP2 Test	(FH,MLP)	1.469	2.250	1.860
Face LP2 Test	(DCTb,GMM)	1.039	0.250	0.644
Speech LP2 Test	(LFCC,GMM)	1.349	1.250	1.300
Speech LP2 Test	(PAC,GMM)	5.283	8.000	6.642
Speech LP2 Test	(SSC,GMM)	2.276	1.750	2.013

Acknowledgement

The authors wish to thank the Swiss National Science Foundation for supporting this work through the National Centre of Competence in Research (NCCR) on "Interactive Multimodal Information Management (IM2)".

References

[1] b. Herbst and D. Richards. On an Automated Signature Verification System. In *Proc. IEEE Int'l Symp. Industrial Electornics*, volume 2, pages 600–604, 1998.

[2] C. BenAbdelkader, R. Cutler, H. Nanda, and L. Davis. Eigengait: Motion-Based Recognition of People Using Image Self-Similarity. In *3rd Int'l Conf. on Audio- and Video-Based Person Authentication*, pages 284–293, June 2001.

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

[4] L. Breiman. Bagging Predictors. *Machine Learning*, 24(2):123–140, 1996.

[5] G. Brown. *Diversity in Neural Network Ensembles.* PhD thesis, School of Computer Science, Uni. of Birmingham, 2003.

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

[7] W. M. Campbell, J. P. Campbell T. Quatieri and, and C. J. Witnstein. Multimodal Speaker Authentication using Nonacoustic Sensors. In *Workshop on (Multimodal-User Authenticaion (MMUA)*, pages 215–230, 2003.

[8] F. Cardinaux, C. Sanderson, and S. Marcel. Comparison of MLP and GMM Classifiers for Face Verification on XM2VTS. In *4th Int'l Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA'03)*, pages 911–920, Guildford, 2003.

[9] K. Chang, K.W. Bowyer, S. Sarkar, and B. Victor. Comparison and Combination of Ear and Face Images in Apperance-Based Biometrics. *IEEE. Trans. Pattern Analysis and Machine Intelligence*, 25(9):1160–1165, 2003.

[10] R. Chellappa and S. Sirohey. Human and Machine Recognition of Faces: A Survey. *Proc. IEEE*, 83(9):705–740, May 1995.

[11] X. Chen, P. Flynn, and K. Bowyer. Visible-Light and Infrared Face Recognition. In *Workshop on (Multimodal-User Authenticaion (MMUA)*, pages 48–55, 2003.

[12] A. Cohen and Y. Zigel. On Feature Selection for Speaker Verification. In *Proc. COST 275 workshop on The Advent of Biometrics on the Internet*, pages 89–92, Rome, November 2002.

[13] J. Daugman. High confidence visual recognition of persons by a test of statistical independence. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(11):1148–1161, 1993.

[14] U. Dieckmann, P. Plankensteiner, and T. Wagner. SESAM: A Biometric Person Identification System Using Sensor Fusion. *Pattern Recognition Letters*, 18:827–833, 1997.

[15] T.G. Dietterich. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems*, pages 1–15, 2000.

[16] T.G. Dietterich and G. Bakiri. Solving Multiclass Learning Problems via Error-Correcting Output Codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.

[17] R.A.J. Everitt and P.W. McOwan. Java-Based Internet Biometric Authentication System. *IEEE. Trans. Pattern Analysis and Machine Intelligence*, 25(9):1166–1171, 2003.

[18] Y. Freund and R. Schapire. A Short Introduction to Boosting. *J. Japan. Soc. for Artificial Intelligence*, 14(5):771–780, 1999.

[19] J.L. Gauvain and C.-H. Lee. Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Obervation of Markov Chains. *IEEE Tran. Speech Audio Processing*, 2:290–298, 1994.

[20] J. Han and B. Bhanu. Individual Recognition Using Gait Energy Image. In *Workshop on Multimodal User Authentication (MMUA 2003)*, pages 181–188, Santa Barbara, 2003.

[21] C. Havran, L. Hupet, J. Czyz, J. Lee, L. Vandendorpe, and M. Verleysen. Independent Component Analysis for Face Authentication. In *Int'l Conf. Knowledge-Based Intelligent Information and Engineering*, pages 1207–1211, September 2002.

[22] L. Hong and A.K. Jain. Multi-Model Biometrics. In *Biometrics: Person Identification in Networked Society*, 1999.

[23] 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.

[24] 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.

[25] S. Ikbal, H. Misra, and H. Bourlard. Phase Auto-Correlation (PAC) derived Robust Speech Features. In *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP-03)*, Hong Kong, 2003.

[26] K. Iwano, T. Hirose, E. Kamibayashi, and S. Furui. Audio-Visual Person Authentication Using Speech and Ear Images. In *Workshop on (Multimodal-User Authenticaion (MMUA)*, pages 85–90, 2003.

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

[28] A.K. Jain and S. Pankanti. Biometrics Systems: Anatomy of Performance. Technical Report MSU-CSE-00-20, Department of Computer Science, Michigan State University, East Lansing, Michigan, September 2000.

[29] P. Jourlin, J. Luettin, D. Genoud, and H. Wassne. Acoustic-Labial Speaker Verification. *Pattern Recognition Letters*, 18(9):853–858, 1997.

[30] J. Kittler, R. Ghaderi, T. Windeatt, and J. Matas. Face Identification and Verification via ECOC. In *Proc. 3rd Int. Conf. Audio-Visual Biometric Person Authentication (AVBPA'01)*, pages 1–13, 2001.

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

[32] 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.

[33] A. Kumar and D. Zhang. Integrating Palmprint with Face for User Authentication. In *Workshop on (Multimodal-User Authenticaion (MMUA)*, pages 107–112, 2003.

[34] 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.

[35] J. Luettin. *Visual Speech and Speaker Recognition.* PhD thesis, Department of Computer Science, University of Sheffield, 1997.

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

[37] R. Maclin and D. Opitz. An Empirical Evaluation of Bagging and Boosting. In *National (American) Conf. on Artificial Intelligence AAAI/IAAI*, pages 546–551, 1997.

[38] Christine Marcel. Multimodal Identity Verification at IDIAP. Communication Report 03-04, IDIAP, Martigny, Switzerland, 2003.

[39] S. Marcel and S. Bengio. Improving Face Verification Using Skin Color Information. In *Proc. 16th Int. Conf. on Pattern Recognition*, Quebec, 2002.

[40] J. Matas, M. Hamouz, K. Jonsson, J. Kittler, Y. Li, C. Kotropoulos, A. Tefas, I. Pitas, T. Tan, H. Yan, F. Smeraldi, J. Begun, N. Capdevielle, W. Gerstner, S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz. Comparison of Face Verification Results on the XM2VTS Database. In *Proc. 15th Int'l Conf. Pattern Recognition*, volume 4, pages 858–863, Barcelona, 2000.

[41] P. R. Morin and J-C. Junqua. A Voice-Centric Multimodal User Authentication System for Fast and Convenient Physical Access Control. In *Workshop on (Multimodal-User Authenticaion (MMUA)*, pages 19–24, 2003.

[42] K. K. Paliwal. Spectral Subband Centroids Features for Speech Recognition. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 617–620, Seattle, 1998.

[43] F. Perronnin and J-L. Dugelay. A Model of Illumination Variation for Robust Face Recognition. In *Workshop on Multimodal User Authentication (MMUA 2003)*, pages 157–164, Santa Barbara, 2003.

[44] N. Poh. Biometric Authentication System. Master's thesis, USM, Penang, Malaysia, August 2001.

[45] N. Poh and S. Bengio. Non-Linear Variance Reduction Techniques in Biometric Authentication. In *Workshop on Multimodal User Authentication (MMUA 2003)*, pages 123–130, Santa Barbara, 2003.

[46] 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.

[47] 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, 2004.

[48] N. Poh, S. Marcel, and S. Bengio. Improving Face Authetication Using Virtual Samples. In *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, pages 233–236 (Vol. 3), Hong Kong, 2003.

[49] N. Poh, C. Sanderson, and S. Bengio. An Investigation of Spectral Subband Centroids For Speaker Authentication. In *Int'l Conf. on Biometric Authentication*, pages 631–639, Hong Kong, 2004.

[50] L. Rabiner and B-H Juang. *Fundamentals of Speech Recognition*. Oxford University Press, 1993.

[51] A. Ross, A. Jain, and J-Z. Qian. Information Fusion in Biometrics. Technical Report MSU-CSE-01-18, Department of Computer Science, Michigan State University, East Lansing, Michigan, May 2001.

[52] A. Ross, A. Jain, and J-Z. Qian. Information Fusion in Biometrics. In *3rd Int. Conf. Audio-Visual Biometric Person Authentication (AVBPA'01)*, pages 354–359, 2001.

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

[54] C. Sanderson and Samy Bengio. Augmenting frontal face models for non-frontal verification. In *Workshop on Multimodal User Authentication (MMUA 2003)*, pages 165–172, Santa Barbara, 2003.

[55] C. Sanderson and K. K. Paliwal. Information Fusion and Person Verification Using Speech & Face Information. Research Report 02-33, IDIAP, Martigny, 2002.

[56] C. Sanderson and K.K. Paliwal. Fast Features for Face Authentication Under Illumination Direction Changes. *Pattern Recognition Letters*, 24(14):2409–2419, 2003.

[57] S. Sharma, D. Ellis, S. Kajarekar, P. Jain, and H. Hermansky. Feature Extraction Using Non-Linear Transformation for Robust Speech Recognition on the Aurora Database. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'03)*, pages II:1117–1120, Hong Kong, 2003.

[58] 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.

[59] W. Shu, G. Rong, Z. Bian, and D. Zhang. Automatic Palmprint Verification. *International Journal of Image and Graphics*, 1(1):135–151, 2001.

[60] D. Zhang, W.-K. Kong, J. You, and M. Wong. Online Palmprint Identificaiton. *IEEE. Trans. Pattern Analysis and Machine Intelligence*, 25(9):1041–1050, 2003.

Table 3: Performance in (%) of HTER of VR via modalities on XM2VTS based on *a priori* selected thresholds. "Joint HTER" are the HTERs of the combined experts via the mean operator, MLP and SVM. The column "mean HTER" refers to the mean HTER of each of the two underlying experts. Similarly, the column "min HTER" refers to the minimum HTER of the two underlying experts. Numbers in bold are the best HTER among the three fusion methods.

(a) Face experts and (LFCC,GMM) expert

Data sets	Face, Experts	Joint HTER			mean	min
		mean	MLP	SVM	HTER	HTER
LP1 Test	(FH,MLP)	0.399	**0.366**	0.381	1.507	1.139
LP1 Test	(DCTs,GMM)	**0.537**	0.576	0.613	2.683	1.139
LP1 Test	(DCTb,GMM)	0.520	0.483	**0.475**	1.405	1.139
LP1 Test	(DCTs,MLP)	0.591	0.611	**0.587**	2.249	1.139
LP1 Test	(DCTb,MLP)	0.497	0.489	**0.485**	3.680	1.139
LP2 Test	(FH,MLP)	0.151	**0.150**	0.389	1.580	1.300
LP2 Test	(DCTb,GMM)	0.147	**0.130**	0.252	0.972	0.644

(b) Face experts and (PAC,GMM) expert

Data sets	Face, Experts	Joint HTER			mean	min
		mean	MLP	SVM	HTER	HTER
LP1 Test	(FH,MLP)	1.114	**0.856**	0.970	4.090	1.875
LP1 Test	(DCTs,GMM)	1.407	1.425	**1.402**	5.266	4.227
LP1 Test	(DCTb,GMM)	**0.899**	0.900	0.923	3.987	1.670
LP1 Test	(DCTs,MLP)	1.248	1.056	**1.009**	4.832	3.359
LP1 Test	(DCTb,MLP)	3.978	**2.455**	2.664	6.263	6.221
LP2 Test	(FH,MLP)	1.282	**0.765**	0.855	4.251	1.860
LP2 Test	(DCTb,GMM)	0.243	**0.222**	0.431	3.643	0.644

(c) Face experts and (SSC,GMM) expert

Data sets	Face, Experts	Joint HTER			mean	min
		mean	MLP	SVM	HTER	HTER
LP1 Test	(FH,MLP)	0.972	0.786	**0.742**	2.156	1.875
LP1 Test	(DCTs,GMM)	**1.028**	1.175	1.213	3.332	2.437
LP1 Test	(DCTb,GMM)	0.756	**0.704**	0.742	2.053	1.670
LP1 Test	(DCTs,MLP)	1.167	**0.829**	0.850	2.898	2.437
LP1 Test	(DCTb,MLP)	2.986	1.176	**1.121**	4.329	2.437
LP2 Test	(FH,MLP)	0.901	**0.302**	0.404	1.937	1.860
LP2 Test	(DCTb,GMM)	**0.049**	0.162	0.383	1.329	0.644

Table 4: Performance in (%) of HTER of VR via extractors and classifiers on XM2VTS based on *a priori* selected thresholds. Columns are explained in Table3.

Data sets	(Features, classifiers)	Joint HTER			mean	min
		mean	MLP	SVM	HTER	HTER
LP1 Test	(FH,MLP) (DCTs,GMM)	1.641	**1.379**	1.393	3.051	1.875
LP1 Test	(FH,MLP) (DCTb,GMM)	**1.123**	1.151	1.528	1.772	1.670
LP1 Test	(FH,MLP) (DCTs,MLP)	**1.475**	1.667	1.476	2.617	1.875
LP1 Test	(FH,MLP) (DCTb,MLP)	1.948	**1.933**	1.938	4.048	1.875
LP1 Test	(LFCC,GMM) (SSC,GMM)	1.296	1.444	**1.142**	1.788	1.139
LP1 Test	(PAC,GMM) (SSC,GMM)	3.594	2.954	**2.663**	4.370	2.437
LP2 Test	(FH,MLP) (DCTb,GMM)	0.896	0.670	**0.488**	1.252	0.644
LP2 Test	(LFCC,GMM) (SSC,GMM)	1.107	**1.034**	1.063	1.656	1.300
LP2 Test	(PAC,GMM) (SSC,GMM)	2.614	2.316	**2.125**	4.328	2.013
LP1 Test	(DCTs,GMM) (DCTs,MLP)	2.873	**2.486**	2.697	3.793	3.359
LP1 Test	(DCTb,GMM) (DCTb,MLP)	2.898	1.532	**1.471**	3.946	1.670

Table 5: Performance in (%) of HTER of different combination methods of synthetic scores. The original method use only real samples while the other methods used. real and synthetic samples, i.e., geometrically transformed face images.

Method	HTER	
	LP1	LP2
Original	1.875	1.737
Mean	1.612	1.518
Median	1.667	1.547
GMM†	1.709	**1.493**
Entropy ‡	**1.606**	1.559

- † "GMM" is fusion using log-likelihood ratio between a Gaussian Mixture Model (GMM) modeling the client scores and another modeling the impostor scores.
- ‡ "Entropy" is fusion of virtual scores, by evaluating the difference of two relative entropies: the relative entropy between distribution of the virtual scores and that of client scores; and the relative entropy between distribution of the virtual scores and that of impostor scores. Details can be found in [45].