How Do Correlation and Variance of Base-Experts Affect Fusion in Biometric Authentication Tasks?

Norman Poh and Samy Bengio

Abstract*— Combining multiple information sources such as subbands, streams (with different features) and multi modal data has been shown to be a very promising trend, both in experiments and to some extents in real-life biometric authentication applications. Despite considerable efforts in fusions, there is a lack of understanding on the roles and effects of correlation and variance (of both the client and impostor scores of base-classifiers/experts). Often, scores are assumed to be independent. In this paper, we *explicitly* consider this factor using a theoretical model, called Variance Reduction-Equal Error Rate (VR-EER) analysis. Assuming that client and impostor scores are approximately Gaussian distributed, we showed that Equal Error Rate (EER) can be modeled as a function of *F-ratio*, which itself is a function of 1) correlation, 2) variance of base-experts and 3) difference of client and impostor means. To achieve lower EER, smaller correlation and average variance of base-experts, and larger mean difference are desirable. Furthermore, analysing any of these factors independently, e.g. focusing on correlation alone, could be miss-leading. Experimental results on the BANCA multimodal database confirm our findings using VR-EER analysis. We analysed four commonly encountered scenarios in biometric authentication which include fusing correlated/uncorrelated base-experts of similar/different performances. The analysis explains and shows that fusing systems of different performances is *not always* beneficial. One of the most important findings is that positive correlation "hurts" fusion while negative correlation (greater "diversity", which measures the spread of prediction score with respect to the fused score), improves fusion. However, by linking the concept of ambiguity decomposition to classification problem, it is found that diversity is not sufficient to be an evaluation criterion (to compare several fusion systems), unless measures are taken to normalise the (class-dependent) variance. Moreover, by linking the concept of bias-variance-covariance decomposition to classification using EER, it is found that if the inherent mismatch (between training and test sessions) can be learned from the data, such mismatch can be incorporated into the fusion system as a part of training parameters.

***Index Terms*— Biometric authentication, variance reduction, bias-variance-covariance decomposition, ambiguity decomposition, pattern recognition, classification**

I. INTRODUCTION

BIOMETRIC authentication (BA) is the problem of verifying an identity claim using a person's behavioural and physiological characteristics. Examples of biometric modalities are fingerprint, face, voice, hand-geometry and retina scans [1]. 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. This affects the accuracy and the reliability of a BA system. One popular trend to improve accuracy is to use multiple modalities of biometric traits, or multiple features (of the same biometric traits), multiple classifiers or multiple samples. Scores are then fused using a COmbination Mechanism (COM, also called a supervisor, a fusion expert/classifier).

Although fusion in the context of BA has been discussed elsewhere, in the authors' opinion, there is still a lack of theoretical analysis and understanding, particularly with respect to correlation. Hong et al [2] shed some light on this subject by demonstrating that combining the expert opinions using AND and OR will result in improved performance. Unfortunately they assumed that the baseline expert opinions are not correlated. Sanchez et al [3] showed both theoretically and empirically that fusing multiple instances of biometric trait can indeed reduce the system error by as much as 40%. The theoretical analysis, unfortunately, again did not deal with the case where 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, Kittler et al [4] examined *intramodal* (i.e., different base-experts of the *same* biometric trait) and *multimodal* (i.e., base-experts of different biometric traits) expert fusion. According to this empirical study, for multimodal fusion, there is no strong evidence that trainable fusion strategies (based on Decision Template [5] and Behaviour Knowledge Space [6]) offer better performance than simple rules (based on sum and vote). They remarked that although adding more experts can reduce variance, 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 [7] studied empirically the case of combining two systems' hypotheses. Specifically, they examined the combination of two systems with equal performance, with unequal performance and with one system outperforming 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. Again, the correlation of these errors are not explored further.

The goal of this study is to apply the VR-EER analysis (the first part is Variance Reduction (VR) and the second part is Equal Error Rate (EER) analysis) that we have proposed

N. Poh and S. Bengio are with IDIAP Research Institute, Rue du Simplon 4, 1920 Martigny, Switzerland. E-mail: {norman,bengio}@idiap.ch.

in [8] on the fusion using a non-trainable COM, namely the mean operator. Different from our previous work, this study takes into account the effect of score normalisation such that the resultant scores have zero-mean and unit-variance. The VR-EER analysis provides a very simple framework to analyse what happens when the scores are correlated, or when the variances of the base-expert are high/low. Since these factors are actually inter-related, attempts to analyse one or the other often fail. Using the proposed framework tested on the BANCA database, we were able to identify different contributing factors that determine the success and failure of fusion, in the context of BA. Based on the VR-EER analysis, four commonly encountered scenarios of fusion in biometric authentication are discussed and analysed.

In this paper, we also linked the concepts of ambiguity decomposition [9] and bias-variance-covariance decomposition [10] that are important analysis tools in regression problems to specific classification problems (using Equal Error Rate evaluation criterion). To the best of our knowledge, the link between these concepts and classification problems have not been shown elsewhere in the literature, as also pointed out by Brown [11].

In the literature, fusion in BA often relies on one or two reported experiments. It should be stressed that our approach to fusion is different in that, we tried to conduct as many experiments as available to us, such that some meaningful statistics can be derived and generalised to other fusion using the same technique.

Section II presents briefly the BANCA experiment setup whereby 1186 experiments were used to study EER and another 70 experiments were used to study fusion. Section III discusses the preliminary findings of the VR-EER analysis and notations used. Section IV presents what happens to fusion when scores are normalised. The effects of variance and correlation are verified in Section V. Using these findings, we analysed four commonly encountered scenarios of fusion in Section VI. Two important analysis tools and concepts that are well-studied in regression problems are linked to a specific classification problem in Sections VII and VIII. Future extensions and limitations of the study are discussed in Section IX. This is followed by conclusions in Section X.

II. Experiment Setup

The BANCA database [12] is the principal database used in this paper. It has a collection of face and voice prints of up to 260 persons in 5 different languages. In this paper, we only used the English subset, containing only a total of 52 persons; 26 females and 26 males. There are altogether 7 protocols, namely, Mc, Ma, Md, Ua, Ud, P and G, each simulating matched control, matched adverse, matched degraded, uncontrolled adverse, uncontrolled degraded, pooled and grant test, respectively. For protocols P and G, there are 312 client accesses and 234 impostor accesses. For all other protocols, there are 78 client accesses and 104 impostor accesses. There are altogether 1186 score files containing single modality experiments or fusion experiments, thanks to a study conducted in [13][1]. The classifiers involved are Gaussian Mixture Models (GMMs), Multi-Layer Perceptrons (MLPs) and Support Vector Machines (SVMs). *All* the score files are used to test the Gaussian hypothesis as reported in Section III-C.

A subset of single modality experiments were selected to study fusion as reported in Section V. These experiments were carried out by University of Surrey (2 face experiments), IDIAP (1 speaker experiment), UC3M (1 speaker experiment) and UCL (1 face experiment). The specific score files used are as follow:

- `IDIAP_voice_gmm_auto_scale_33_200`
- `SURREY_face_svm_auto`
- `SURREY_face_svm_man`
- `UC3M_voice_gmm_auto_scale_34_500`
- `UCL_face_lda_man`

for each of the 7 protocols. Each of these files contains the following columns of data: the true identity, the claimed identity, a unique access tag and the associated expert score for the access. Moreover, for each protocol, there are two subgroups, called g1 and g2. In this paper, g1 is used as a development set (called **dev**) while g2 is used as an evaluation set (called **eva**). By combining each time two baseline experts of a protocol, one can obtain 10 fusion experiments, given by ${}^5C_2$ (5 "choose" 2). This results in a total of 70 experiments for all 7 protocols.

III. Preliminary and Recent Findings on VR-EER

Our proposed theoretical model [8] has two parts. The first one deals with Variance Reduction (VR) and the second relates F-ratio (which involves variance discussed in the first part) to Equal Error Rate (EER).

A. Variance Reduction

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 j, we can replace θ by $\theta(j)$. 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). 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:

$$y_i = \begin{cases} f_\theta(f_e(s(\mathbf{x}_i))) & \text{if multi-sample} \\ f_\theta(f_e(s_i(\mathbf{x}))) & \text{if multimodal} \\ f_\theta(f_{e,i}(s(\mathbf{x}))) & \text{if multi-feature} \\ f_{\theta,i}(f_e(s(\mathbf{x}))) & \text{if multi-classifier} \end{cases} \tag{2}$$

[1] Available at "ftp://ftp.idiap.ch/pub/bengio/banca/banca_scores"

Note that i is the index to 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 y_i as the i-th *response* and there are altogether N responses ($i = 1, \ldots, N$).

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 III-B. Suppose that y_i^k is the i-th response (sample, modality, feature or classifier) belonging to class $k = \{C, I\}$, i.e., either client or impostor. We adopt the convention that the mean of $y_i^{k=C}$ is greater than that of $y_i^{k=I}$.

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 turn affects the quality of extracted features. Indeed, it has a distribution governed by the extracted feature set $f_e(s(\mathbf{x}))$ 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:

$$y_{i,j}^k = \mu_i^k + w_{i,j}^k, \tag{3}$$

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. (3), 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:

$$E[Y_i^k] = E[\mu_i^k] + E[W_i^k] = \mu_i^k. \tag{4}$$

Note that different from [8], here we do not require μ_i^k to take on specific values such as -1 for $k = I$ and 1 for $k = C$. This assumption is true for discriminative training (i.e., using Multi-Layer Perceptrons (MLPs) or Support Vector Machines (SVMs)), but not applicable for distance-based scores or log-likelihood ratios. Hence, removal of such assumption makes the analysis applicable to a wider context.

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, according to [8]:

$$\begin{aligned}\left(\sigma_{AV}^k\right)^2 &= \frac{1}{N}\sum_{i=1}^N Cov(Y_i^k, Y_i^k) \\ &= \frac{1}{N}\sum_{i=1}^N E[W_i^k W_i^k] \\ &\equiv \frac{1}{N}\sum_{i=1}^N \left(\sigma_i^k\right)^2,\end{aligned} \tag{5}$$

where we adopted the following notation: $Cov(Y_i^k, Y_j^k)$ as 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:

$$Y_{COM}^k = \frac{1}{N}\sum_{i=1}^N Y_i^k, \tag{6}$$

for any $k \in \{C, I\}$. 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 (see [8] for details of this derivation):

$$\begin{aligned}\left(\sigma_{COM}^k\right)^2 &= Cov(Y_{COM}^k, Y_{COM}^k) \\ &= \frac{1}{N^2}\sum_{j=1}^N \left(\sigma_j^k\right)^2 + \\ &\quad \frac{2}{N^2}\sum_{m=1, m<n}^N \rho_{m,n}^k \sigma_m^k \sigma_n^k, \\ &= \underbrace{\frac{1}{N}\left(\sigma_{AV}^k\right)^2}_{} + \\ &\quad \underbrace{\frac{2}{N^2}\sum_{m=1, m<n}^N \rho_{m,n}^k \sigma_m^k \sigma_n^k}_{},\end{aligned} \tag{7}$$

where $\rho_{m,n}^k$ is the correlation coefficient between Y_m^k and Y_n^k for $k \in \{C, I\}$. The first underbrace term is the *average variance* of the base-experts while the second underbrace term is the *covariance* between Y_m^k and Y_n^k for $m \neq n$. This is because the term $\rho_{m,n}^k \sigma_m^k \sigma_n^k$ is by definition equivalent to correlation, i.e.,

$$\rho_{m,n}^k \sigma_m^k \sigma_n^k = E[W_m^k W_n^k], \tag{8}$$

Note that $\rho_{n,n}^k = 1$ for $k \in \{C, I\}$. The VR analysis shows that [8]:

$$\left(\sigma_{COM}^k\right)^2 \leq \left(\sigma_{AV}^k\right)^2. \tag{9}$$

When $0 \leq \rho_{m,n}^k \leq 1$, it can be shown that:

$$\frac{1}{N}\left(\sigma_{AV}^k\right)^2 \leq \left(\sigma_{COM}^k\right)^2. \tag{10}$$

In other words, the upper bound of $\left(\sigma_{COM}^k\right)^2$ is shown in Eqn. (9) and its lower bound is shown in Eqn. (10). They are attained in perfect correlation in the former case and uncorrelated case in the latter case. Any other positive correlation values will cause $\left(\sigma^k_{COM}\right)^2$ to take on values between these bounds. Hence, by combining N responses using the mean operator, the resultant variance is assured to be smaller than the average (not the minimum) variance.

B. Equal Error Rate Analysis

Let $\mu_p^{k=C}$ and $\mu_p^{k=I}$ be the means of client and impostor access scores of a given experiment p. Without loss of generality, we assume that $\mu_p^{k=C} > \mu_p^{k=I}$. Let $\sigma_p^{k=C}$ and $\sigma_p^{k=I}$ be the standard deviation of the client and impostor scores. In BA, there are two types of errors committed by the system, often measured by False Acceptance Rates (FARs) and False Rejection Rates (FRRs). FAR(Δ) is calculated by integrating the impostor score distribution from a given threshold Δ in the score space to $+\infty$ while FRR(Δ) is calculated by integrating the client distribution from $-\infty$ to Δ. Equal Error Rate (EER) is a unique point where FAR equals FRR. By assuming that the client and impostor scores follow Gaussian distributions, one can derive the EER of a given experiment p as (see [8] for details of this derivation) :

$$\text{EER}_p = \frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{\text{F-ratio}_p}{\sqrt{2}}\right), \tag{11}$$

where

$$\text{F-ratio}_p = \frac{\mu_p^{k=C} - \mu_p^{k=I}}{\sigma_p^{k=C} + \sigma_p^{k=I}}, \tag{12}$$

and

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp\left[-t^2\right] dt. \tag{13}$$

It should be noted that the term F-ratio is used here because this value is somewhat *similar to* the standard Fisher ratio, but not defined exactly in the same way. In a two-class problem, the Fisher ratio [14, pg. 107] is defined as

$$\frac{\mu_p^{k=C} - \mu_p^{k=I}}{(\sigma_p^{k=C})^2 + (\sigma_p^{k=I})^2} \tag{14}$$

F-ratio is used here just to underpin the idea that the degree of separability of the class distribution affects the authentication performance measured by EER. There exists similar measures such as the *d-prime* metric proposed by Daugman [15]. It measures how separable the client distribution is from its impostor counterpart. It is defined as:

$$d' = \frac{|\mu_p^{k=C} - \mu_p^{k=I}|}{\sqrt{\frac{1}{2}(\sigma_p^{k=C})^2 + \frac{1}{2}(\sigma_p^{k=I})^2}}. \tag{15}$$

In our opinion, F-ratio should be used instead since it is directly related to EER by Eqn. (11)).

We call the EER based on Gaussian assumption the *theoretical EER*, to distinguish it from the *empirical EER*, which is calculated by direct minimisation of the following criterion:

$$\Delta^* = \arg\min_{\Delta} |\text{FAR}(\Delta) - \text{FRR}(\Delta)|, \tag{16}$$

where Δ is the threshold, and approximated by the commonly used Half Total Error Rate:

$$\text{HTER}_{\min} = \frac{\text{FAR}(\Delta^*) + \text{FRR}(\Delta^*)}{2}. \tag{17}$$

$\text{HTER}_{\min}$ is minimum at Δ^*. This is because FRR is an increasing function (a cumulative density function; *cdf*) and FAR is a decreasing function (one minus *cdf*). Δ^* is the point where these two functions intersect. Let EER_{COM} be the EER of the combined scores and EER_{AV} be the EER of the average of scores of N responses. Using Eqns. (11 and 12), their corresponding F-ratios can be defined as follow:

$$\text{F-ratio}_{COM} = \frac{\mu_{COM}^{k=C} - \mu_{COM}^{k=I}}{\sigma_{COM}^{k=C} + \sigma_{COM}^{k=I}} \tag{18}$$

$$\text{F-ratio}_{AV} = \frac{\mu_{AV}^{k=C} - \mu_{AV}^{k=I}}{\sigma_{AV}^{k=C} + \sigma_{AV}^{k=I}}. \tag{19}$$

In order to show the hypothesis that the EER of the combined scores is less than the EER of the the average of N responses, i.e.,

$$\text{EER}_{COM} \leq \text{EER}_{AV}, \tag{20}$$

we first need to calculate μ_p^k and σ_p^k for $k = \{C, I\}$ and $p = \{COM, AV\}$. $\sigma_p^k | p = \{COM, AV\}$ were defined by Eqns. (5 and 7), respectively. μ_{AV}^k is the average of N responses when used separately. It is defined as:

$$\mu_{AV}^k \equiv \frac{1}{N}\sum_{i=1}^{N} \mu_i^k, \tag{21}$$

μ_{COM}^k is the mean of the combined scores of N responses (used simultaneously). It is defined as:

$$\begin{aligned} E[Y_{COM}^k] \quad &\equiv \quad \mu_{COM}^k = \frac{1}{N}\sum_{i=1}^{N} E[Y_i^k] \\ &= \quad \frac{1}{N}\sum_{i=1}^{N} \mu_i^k = \mu_{AV}^k. \end{aligned} \tag{22}$$

Hence, $\mu_{COM}^k = \mu_{AV}^k$. Since F-ratio is non-linearly and inversely proportional to EER as shown in Eqn. (11), the inequality of Eqn. (20) can be rewritten as:

$$\text{F-ratio}_{COM} \geq \text{F-ratio}_{AV}, \tag{23}$$

Replacing the two F-ratio terms using Eqns. (22 and 21) into Eqn. (23) and using the relation $\mu_{COM}^k = \mu_{AV}^k$, we obtain:

$$\begin{gathered} \frac{\mu_{COM}^{k=C} - \mu_{COM}^{k=I}}{\sigma_{COM}^{k=C} + \sigma_{COM}^{k=I}} \geq \frac{\mu_{AV}^{k=C} - \mu_{AV}^{k=I}}{\sigma_{AV}^{k=C} + \sigma_{AV}^{k=I}} \\ \sigma_{COM}^{k=C} + \sigma_{COM}^{k=I} \leq \sigma_{AV}^{k=C} + \sigma_{AV}^{k=I} \\ \sum_{k=\{C,I\}} \sigma_{COM}^k \leq \sum_{k=\{C,I\}} \sigma_{AV}^k \end{gathered} \tag{24}$$

This inequality confirms the upper bound already found in Eqn. (9). Hence, the inequality of Eqn. (20) is true, i.e., fusing scores can reduce variance which results in reduction of EER (with respect to the case where scores are used separately). This formed the argument in [8] for why fusion using multiple modalities, features, and classifiers works for BA tasks.

C. Validity of the Gaussian Assumption

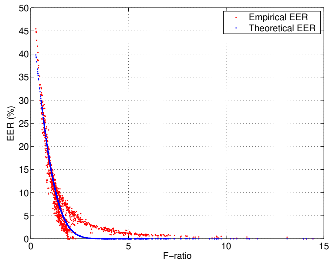
To check how accurate the EER function is as compared to its empirical counterpart, we conducted as many as 1186 experiments on the BANCA database as described in Section II. There are 490 experiments from the output of Multi-Layer Perceptrons (MLPs), 182 experiments from the output of Support Vector machines (SVMs) and 514 experiments from the output of Gaussian Mixture Models (GMMs). Two approaches are adopted here. The first approach is to test whether for each of the 1186 experiments, the respective client and impostor scores are normally distributed or not. The second approach is to directly compare the empirical EER against its theoretical counterpart (assuming that client and impostor distributions are normally distributed).

For the first approach, we applied the Lillie-test [16]. 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. Using this test, we found that 22.85% of impostor scores and 25.89% of client scores (out of 1186 experiments) supported the hypothesis that they are Gaussian distributed. Hence, only approximately a quarter of the distributions are Gaussian according to the Lillie-test.

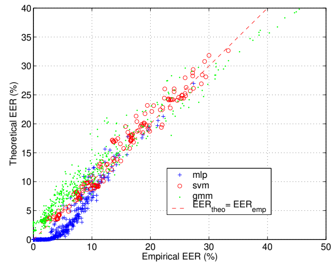
The results of the second approach are shown in Figure 1. From Figure 1(a), it can be seen that both the theoretical and empirical EERs are non-linearly and inversely proportional to their F-ratio. Removing the F-ratio, we compared the theoretical EER directly with its empirical counterpart in Figure 1(b). Here the output of different classifiers are plotted with different symbols. If the theoretical EER matches exactly its empirical EER, the points (each one corresponding to a single experiment) should be on the diagonal line. One measure of agreement is to use correlation. Its value is evaluated to be 0.9573, indicating the the variables are *strongly correlated*. In other words, knowing theoretical EER, one can use the correlation to *approximately* estimate the empirical EER.

Figure 1(c) plots the absolute EER difference (between theoretical EER and empirical EER) versus the average KS-statistics of their respective client and impostor distributions (note that from each experiment, we will have two KS-statistics values, one for each distribution). The KS-statistics measures the degree of divergence from normal distribution. As can be seen, the output of MLPs (trained using sigmoid output function) gives high KS-statistics whereas the outputs of SVMs and GMMs conform better to the Gaussian assumption.

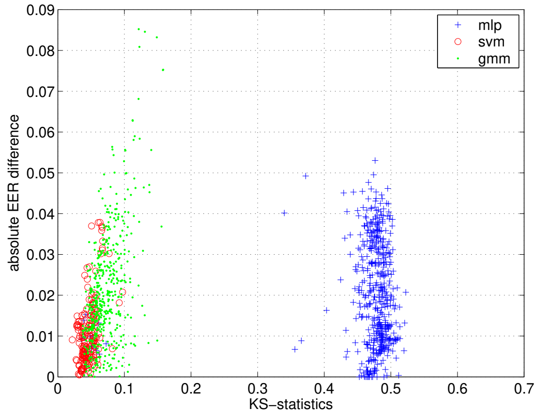
Prior to this experiment, we thought that deviation from Gaussian would mean large absolute EER difference. If this was the case, absolute EER difference would have been increasing proportionally with respect to the KS-statistics. It turns out that this is not the case. In Figure 1(c), despite high KS-statistics of MLP outputs, their corresponding absolute EER differences are spread below 0.06; some are even near 0! Hence, deviation from Gaussian does not mean large absolute EER difference. In other words, **the theoretical EER is fairly robust to deviation from the Gaussian assumption**.



(a) Empirical and theoretical EER vs. F-ratio



(b) Empirical vs. theoretical EER



(c) Absolute EER difference vs. KS-statistics

Fig. 1. Results of experiments carried out using all the available 1186 experiments on the BANCA score database: (a) Theoretical EER and empirical EER (HTER) versus their common F-ratio (b) Theoretical EER versus empirical EER (HTER) using output of different classifiers – 490 MLPs, 182 SVMs and 514 GMMs; the correlation coefficient between the two variables is 0.9573. (c) Absolute EER difference between theoretical EER and empirical EER versus the average KS-statistics between the corresponding client and impostor distributions. KS-statistics measures the degree of deviation from Gaussian assumption.

It should be noted that a more interesting issue to investigate is the *relative* values of EER, i.e., if the empirical EER of experiment a is more than the empirical EER of experiment b, does the theoretical EER of these experiments also follow the same trend? Using the data at hand, we calculated all the possible combinations of two EER experiments. This turns out to be ${}^{1186}C_2 = 702,705$ combinations. The number of "disagreement", d, can be calculated as follows:

$$d = \left|(\mathrm{EER}_a^{emp} > \mathrm{EER}_b^{emp}) - (\mathrm{EER}_a^{theo} > \mathrm{EER}_b^{theo})\right| \quad (25)$$

for $(a,b) \in \{(1,2),(1,3),\ldots,(1185,1186)\}$ and

$$(z_1 > z_2) = \begin{cases} 1 & \text{if true} \\ 0 & \text{otherwise.} \end{cases} \quad (26)$$

The percentage of disagreement turns out to be 11%. If the 1186 experiments are representative of biometric authentication tasks, we can conclude that to compare any two experiments, 89% of the time, the theoretical EER (calculated from the F-ratio) can give a correct answer as compared to using the empirical EER as the ground-truth. Of course, a mixture of Gaussians or non-parametric Parzen window with Gaussian kernel could have been used to accurately model the underlying client and impostor distributions. In so doing, we may not be able to further perform the analysis in sections that follow.

The VR-EER analysis presented here is not simply theoretical. In the following section, we propose to put this analysis to test.

IV. Score Normalisation

To begin with, we would like to fuse the scores of two systems using the simple mean operator (trainable weighted sum and non-linear functions could be included in this analysis in the future).

Before fusing the scores, it is necessary to normalise them so that the scores of a given base-expert with high variance will not dominate the fused decision. We used the *zero-mean unit-variance* approach. This is done by subtracting an input score from its *global mean* (estimated from a training set) and divide it by its standard deviation. Let y_i^k be a raw output score which follows the distribution Y_i^k. The normalised score distribution, $Y_i^{norm,k}$, can be written as follows:

$$\begin{aligned} Y_i^{norm,k} &= \frac{Y_i^k - E[Y_i^{all}]}{\sqrt{Cov(Y_i^{all}, Y_i^{all})}} \\ &\equiv \frac{Y_i^k - \mu_i^{all}}{\sigma_i^{all}}, \end{aligned} \tag{27}$$

for $k \in \{C, I\}$ and $Y_i^{all} = 1/2(Y_i^{k=C} + Y_i^{k=I})$, i.e, the union of the two distributions. When combining the scores using mean, we obtain:

$$Y_{COM}^{norm,k} = \frac{1}{N}\sum_{i=1}^{N} Y_i^{norm,k}. \tag{28}$$

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

$$\begin{aligned} \mu_{COM}^{norm,k} &\equiv E[Y_{COM}^{norm,k}] \\ &= \frac{1}{N}\sum_{i=1}^{N} E[Y_i^{norm,k}] \\ &= \frac{1}{N}\sum_{i=1}^{N} \frac{E[Y_i^k] - \mu_i^{all}}{\sigma_i^{all}} \\ &= \frac{1}{N}\sum_{i=1}^{N} \frac{\mu_i^k - \mu_i^{all}}{\sigma_i^{all}}. \end{aligned} \tag{29}$$

Using Eqns. (4), (27) and (28), the variance of $Y_{COM}^{norm,k}$ is:

$$\begin{aligned} (\sigma_{COM}^{norm,k})^2 &= Cov(Y_{COM}^{norm,k}, Y_{COM}^{norm,k}) \\ &= E\left[\left(Y_{COM}^{norm,k} - E[Y_{COM}^{norm,k}]\right)^2\right] \\ &= E\left[\left(\frac{1}{N}\sum_{i=1}^{N}\frac{Y_i^k - \mu_i^{all}}{\sigma_i^{all}} \frac{1}{N}\sum_{m=1}^{N}\frac{\mu_i^k - \mu_i^{all}}{\sigma_i^{all}}\right)^2\right] \\ &= E\left[\left(\frac{1}{N}\sum_{i=1}^{N}\frac{Y_i^k - \mu_i^k}{\sigma_i^{all}}\right)^2\right] \\ &= E\left[\left(\frac{1}{N}\sum_{i=1}^{N}\frac{W_i^k}{\sigma_i^{all}}\right)^2\right]. \end{aligned} \tag{30}$$

To expand Eqn. (30), one should take care of possible correlation between different W_m^k and W_n^k, similar to Eqn. (7), as follows:

$$\begin{aligned} (\sigma_{COM}^{norm,k})^2 &= E\left[\frac{1}{N^2}\left(\sum_{m=1}^{N}\sum_{n=1}^{N}\frac{W_m^k W_n^k}{\sigma_m^{all}\sigma_n^{all}}\right)\right] \\ &= \frac{1}{N^2}\sum_{j=1}^{N}\frac{E[W_j^k W_j^k]}{\sigma_j^{all}} \\ &\quad + \frac{2}{N^2}\sum_{m=1,m<n}^{N}\frac{E[W_m^k W_n^k]}{\sigma_m^{all}\sigma_n^{all}}. \\ &\equiv (V_{AV}^k)^2 + (V_{COV}^k)^2, \end{aligned} \tag{31}$$

for any $k \in \{C, I\}$. The term $(V_{AV}^k)^2$ is the average *normalised* variance of the base-expert scores while the second term $(V_{COV}^k)^2$ is the *normalised* covariance between $Y_m^{norm,k}$ and $Y_n^{norm,k}$ for $m \neq n$.

The F-ratio is:

$$\text{F-ratio}_{COM}^{norm} = \frac{\mu_{COM}^{norm,k=C} - \mu_{COM}^{norm,k=I}}{\sigma_{COM}^{norm,k=C} + \sigma_{COM}^{norm,k=I}}. \tag{32}$$

V. Effects of Correlation, Variance of Base Expert Scores on Fusion

Having derived all the parameters in the VR-EER analysis in the previous section, namely, $\mu_{COM}^{norm,k}$, $\sigma_{COM}^{norm,k}$, and F-ratio$_{COM}^{norm}$, we carried out experiments on the BANCA fusion database, each time *combining only two* experts. These experiments can be divided into two types: multimodal fusion (fusion of two different modalities, i.e, face and speech experts) and intramodal expert fusion (of two face experts *or* two speech experts). We expect multimodal fusion to be less correlated while intramodal fusion to be more correlated. This is an important aspect so that both sets of experiments will cover a large range of correlation values.

A naive approach to analyse fusion is to find empirically the relationship between minimum *a posteriori* HTER and the sum of correlation of client and impostor distributions. Let the client and impostor-dependent correlations between two baseline systems (to be fused) be the scalars ρ_C and ρ_I, respectively[2]. The results are shown in Figure. 2. From this

[2]In general, the correlation of scores of N responses are a matrix of N by N with elements $\rho_{m,n}$. It has the property that $\rho_{m,m} = 1$ and $\rho_{m,n} = \rho_{n,m}$. In the case of two responses, we simply write ρ in place of $\rho_{1,2}$.

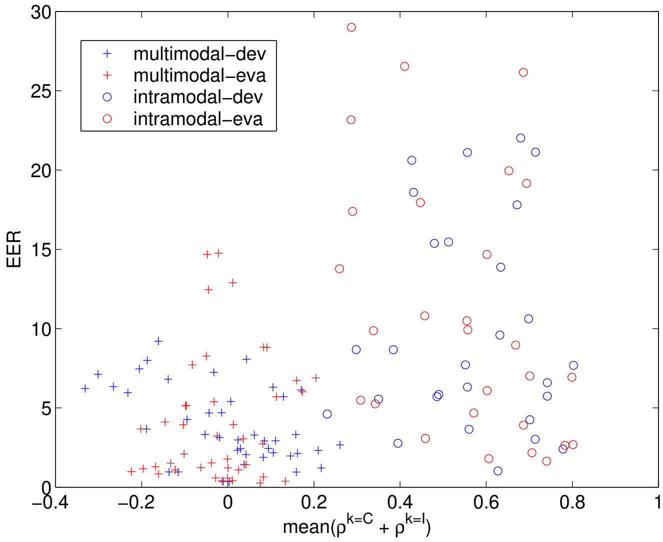
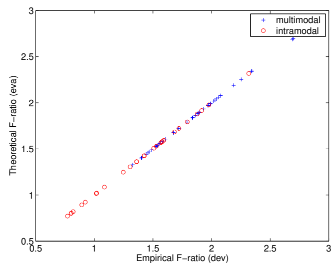
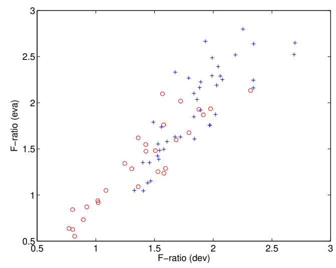


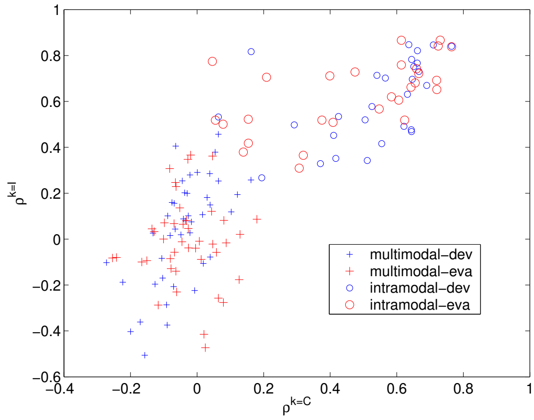
Fig. 2. Empirical EER of combining 2 baseline experts versus $\rho_C + \rho_I$ using the BANCA database. The crosses represent experiments combining 2 modalities while the circles represent those combining 2 features of the *same* modality. The correlation between the two variables is 0.38.



(a) Empirical F-ratio vs. theoretical F-ratio



(b) F-ratio (dev) vs F-ratio (eva)



(c) Correlation of client vs impostor scores

Fig. 3. Experiments carried out on fusion of ${}^5C_2 \times 7 = 70$ experiments, i.e., combining 2 expert systems each time out of five available systems, for all the 7 BANCA protocols: (a) Empirical F-ratio versus theoretical F-ratio on the development set. (b) F-ratio of development set versus its evaluation set counterpart. The correlation between the two variables is 0.90. (c) Correlation of client scores versus correlation of impostor scores. The correlations between the two variables (class-dependent correlations) on the development and evaluation sets are 0.85 and 0.80, respectively.

figure, it can be observed that multimodal fusion experiments have less correlated scores while multi-feature fusion experiments have high correlated scores. One would have expected that the minimum *a posteriori* HTER is somewhat proportional to $\rho_C + \rho_I$. This is actually partially true because the variance of base-experts are not taken into account. As a result, there is no clear trend in this graph and one cannot conclude that HTER is proportional to correlation.

By making use of the enhanced VR-EER analysis with zero-mean unit-variance normalisation, we propose to evaluate the *theoretical* versus *empirical* parameters in the VR-EER analysis. For each of the parameters tested here, *theoretical* means that the respective parameter is directly estimated using the unnormalised input score set. This score set is of dimension two, since only two expert scores are fused at a time. *Empirical* means that the respective parameter is estimated using the resultant fused score.

Figure 3(a) shows empirical F-ratio versus its theoretical counterpart calculated uniquely on the development set. As can be seen both empirical and theoretical F-ratio are exactly the same. Their equivalence can be shown mathematically (see Appendix I). Hence, the performance of fusion is determined by F-ratio, assuming that the scores are normally distributed, as defined in Eqn. (32). The fused and normalised mean and variance is defined in Eqn. (29) and Eqn. (31), respectively. From Eqn. (31), we know that the fused and normalised variance has two components: variance and covariance. Based on the Gaussian assumption, fusion performance consists of three factors: (1) mean difference (the nominator of Eqn. (32)), (2) variance and (3) covariance of baseline experts. The first component measures how far the client mean of the fused score is from its impostor counterpart. The second component corresponds to the sum of square-root of the diagonal terms of covariance matrix of the fused scores (for both client and impostor scores). This term measures, in average, how good the base-experts are, when acting alone. The last component corresponds to the sum of square-root of the non-diagonal terms of covariance matrix of the fused scores (for both client and impostor scores). Note that the last two factors *cannot* be separated due to the square-root of variance in the denominator of F-ratio. Understanding how these three factors are related (by F-ratio) can guide us to understand how EER should be minimised or how F-ratio should be maximised, i.e., maximising the mean difference and minimising the variance and covariance components. Because of these three interrelated factors, analysing any one of them alone, as done in Figure 2 or in Figure 3(c), does not lead to any convincing conclusion.

The above analysis was performed by combining two baseline experts. It is natural to ask if the analysis would work by combining more than two experts. We repeated the above experiments for combining three and four experts and were able to predict the F-ratio accurately. The results are similar to those presented in Figure 3(a) (not shown here). This is somewhat expected because the VR-EER analysis is not limited to two experts. Similarly, the analytical proof showing the equivalence between empirical F-ratio and theoretical F-ratio in Appendix I is a general framework that, of course, includes fusion of N experts.

Figure 3(b) plots the F-ratio found on the development set versus the F-ratio found on the evaluation set. They are not exactly the same this time because there is a mismatch between these two data sets. Nevertheless, their correlation is 0.90, indicating that knowing F-ratio from the development set, it is possible to predict reasonably F-ratio of the evaluation set. A follow-up study using weighted sum [17], instead of mean as done here, also showed that using weighted sum operator, where weights are found on a *development set*, empirical F-ratios of fusion experiments (using all possible combination of base-experts) match *approximately* the F-ratio on the evaluation set. The plot of F-ratio between the development and the evaluation sets is similar to Figure 3(b) (not shown here), i.e., strong correlation is also observed. More details on how to derive the F-ratio of weighted-sum fusion (instead of mean as done here) can be found in [17].

Figure 3(c) shows a 2D plot of the following two variables: correlation of client and that of impostor scores. The overall correlation between these two variables is 0.83. This indicates that knowing the covariance (or correlation; since one is proportional to the other as shown in Eqn. (8)) of the impostor scores, one can approximate the covariance of the client scores. Note that all intramodal fusion experiments have high correlation values. Figure 3(c) thus has two clusters. The cluster in the upper right corner belongs to intramodal fusion experiments whereas the cluster in the lower left corner belongs to multimodal fusion experiments.

VI. Analysis of Commonly Encountered Scenarios in Biometric Authentication

Suppose we have the following scenarios:

1) Combining 2 uncorrelated experts with very different performances
2) Combining 2 highly correlated experts with very different performances
3) Combining 2 uncorrelated experts with very similar performances
4) Combining 2 highly correlated experts with very similar performances

The first and third cases are often encountered in multimodal fusions while the second and fourth cases are encountered in intra-modal (multi-feature) fusions. Fusing experts of similar and different performances are encountered in almost all biometric authentication problems. It should be noted that empirical evidences of these scenarios have been examined in [7] but unfortunately there was a lack of theoretical explanation.

To make analysis simple, let us assume that (i) the two base-experts have the same numerator of F-ratio and that (ii) for each base-expert, the variance and covariance of client and impostor distributions are proportional. The first assumption is actually reasonable because scores can be normalised to have canonical client and impostor means. For instance, we can map σ_i^k to $\sigma_i^{k,\prime}$ by constraining that they have the same F-ratio while assuming that μ_i^k of the resultant conversion takes on -1 for impostor distribution and 1 for client distribution, as follows:

$$\text{F-ratio} = \frac{\mu_i^{k=C} - \mu_i^{k=I}}{\sigma_i^{k=C} + \sigma_i^{k=I}} = \frac{1-(-1)}{\sigma_i^{k=C,\prime} + \sigma_i^{k=I,\prime}}. \tag{33}$$

The solution is:

$$\sigma_i^{k,\prime} = \alpha_i \sigma_i^k, \tag{34}$$

where,

$$\alpha_i = \frac{2}{\mu_i^{k=C} - \mu_i^{k=I}},$$

for $k = \{C, I\}$.

By taking the square of Eqn. (34) and applying the definition of variance of y_i, we obtain

$$\begin{aligned}(\sigma_i^{k,\prime})^2 &= (\alpha_i)^2 E\left[(y_i^k - E[y_i^k])^2\right] \\ &= E\left[\left(\alpha_i(y_i^k - E[y_i^k])\right)^2\right]\end{aligned} \tag{35}$$

Since α_i is not dependent on the class label k, Eqn. (35) is also valid when applying to y_i, instead of y_i^k. Therefore, to map the client and impostor means to canonical values, one needs to modify the variance *without affecting* the F-ratio and the corresponding EER. This simply translates into multiplying score y_i with α_i.

The second assumption implies that $\sigma_i^{k=C} \propto \sigma_i^{k=I}$ for system $i \in \{1, 2\}$ as well as their covariance

$$\rho^{k=I} \sigma_1^{k=C} \sigma_2^{k=C} \propto \rho^{k=C} \sigma_1^{k=I} \sigma_2^{k=I}.$$

It is rather intuitive and actually not necessary. It just simplifies the analysis so that one considers only one class at a time. The variance of the two classes can be merged by using the relation found in the denominator of Eqn. (32). Hence, the class label k can be dropped.

For the first case, without loss of generality, we assume $\sigma_1 \leq \sigma_2$ (i.e., system 1 is better than system 2) and $\rho \simeq 0$. Hence, for the combination to be *better than the best system*, i.e., system 1, it is required that:

$$\begin{aligned}\sigma_{COM}^2 &< \sigma_1^2 \\ \frac{\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2}{4} &< \sigma_1^2\end{aligned} \tag{36}$$

σ_{COM}^2 is calculated using Eqn. (7) with N=2.

We see that:

$$\sigma_2^2 < 3\sigma_1^2 - 2\rho\sigma_1\sigma_2.$$

Note that in general, the covariance $\rho \geq 0$. For instance, in multimodal fusion, c is around zero while in multi-feature fusion, ρ is positive.

Hence, the combined system will benefit from the fusion when σ_2^2 is *at most* less than 3 times of σ_1^2 since $\rho \simeq 0$.

Furthermore, correlation (or equivalently covariance; one is proportional to the other; See Eqn. (8)) between the two systems penalises this margin of $3\sigma_1^2$. This is particularly true for the second case since $\rho > 0$. Also, it should be noted that $\rho \leq 0$ (which implies negative correlation) could allow for larger σ_2^2. As a result, adding another system that is negatively correlated, but with large variance (hence large EER) *will* improve fusion. Unfortunately, in biometric authentication, 2 systems are either positively correlated or not correlated, unless these systems are *jointly trained* together by algorithms such as negative correlation learning [11].

For the third and fourth cases, we have $\sigma_1^2 \simeq \sigma_2^2$. Hence, Eqn. (36) becomes

$$\rho\sigma_2^2 < \sigma_1^2. \tag{37}$$

Note that for the third case, $\rho \simeq 0$ which will satisfy the constraint of Eqn. (37). Therefore, fusion will *definitely* lead to better performance. On the other hand, for the fourth case where $\rho \simeq 1$, according to Eqn. (37), fusion may not necessarily lead to better performance.

VII. Relation to Ambiguity Decomposition

We would like to link our findings with those of Krogh and Vedelsby [9] (see also [14, pages 368]), who showed that, in our context:

$$\begin{aligned} E[Y_{COM}^k - \mu_{COM}^k]^2 &= \sum_i \alpha_i E\left(Y_i^k - \mu_{COM}^k\right)^2 \\ &\quad - \sum_i \alpha_i E\left(Y_i^k - Y_{COM}^k\right)^2 \\ (\sigma_{COM}^k)^2 &\equiv \text{acc}^k - \text{div}^k, \end{aligned} \tag{38}$$

where α_i are the weights in weighted sum combination. This equation is also true for the normalised version of Y_{COM}^k, i.e., $Y_{COM}^{norm,k}$. Note that $\alpha_i = 1/N$ because we are using the mean operator instead of weighted sum. The first term, denoted as acc (or "accuracy"), measures how accurate each base-expert is with respect to the mean score of the combined mechanism. It depends only on the individual base-experts. The second term, denoted as div (or "divergence"), measures the spread of prediction of the base-experts relative to the score of combined mechanism.

Based on the definition of accuracy in Eqn. (38), the accuracy of $Y_{COM}^{norm,k}$ as discussed in section IV is:

$$\begin{aligned} \text{acc}^k &= \frac{1}{N}\sum_i E[Y_{COM}^{norm,k} - \mu_{COM}^{norm,k}] \\ &= \frac{1}{N}\sum_i E\left[\frac{Y_i^k - \mu_i^{all}}{\sigma_i^{all}} - \frac{1}{N}\sum_j \frac{\mu_j^k - \mu_i^{all}}{\sigma_j^{all}}\right]^2 \\ &= \frac{1}{N}\sum_i \frac{E[W_i^k W_i^k]}{(\sigma_i^{all})^2} = (V_{AV}^k)^2. \end{aligned} \tag{39}$$

From Eqns. (38) and (31), it is obvious that divergence is simply:

$$\text{div}^k = -(V_{COV}^k)^2. \tag{40}$$

The negative sign in this term shows that the divergence is indeed negatively proportional to the covariance component. Hence, conclusions drawn in Section V also apply here: divergence (negative covariance) is not a sufficient metric for measuring classification error diversity. This explains why a number of heuristics to define classification error diversity have been proposed in the literature [18], all based on zero-one loss function where a threshold has already been applied. What we really want to do is in fact measuring the diversity *without fixing the threshold* in advance. For a specific case in biometric authentication, this can be done via EER as proposed in Section III-B and [8]. By so doing, one assumes that the client and impostor scores can be modeled by Gaussian distributions, and that the prior class distributions and cost of two types of errors are equal.

VIII. Relation to Bias-Variance-Covariance Decomposition

Ueda and Nakano [10] presented the bias-variance-covariance decomposition while Brown [11] provided the link between this concept and the ambiguity decomposition. However, both discussions were limited to the context of regression, as clearly pointed out by Brown [11, Sec. 3.1.2]. So far, we have not discussed about mismatch between training and test conditions. The introduction of bias in classification can actually be very useful for countering such a problem, as will become clear later.

Suppose that the noise model in Eqn. (3) can only be calculated from a training set. During testing, the noise model deviates from the one observed during training, i.e., there is a *mismatch* between training and testing. Suppose that the new noise model now is:

$$y_i^{k,\prime} = \mu_i^{k,\prime} + w_i^k, \tag{41}$$

where

$$\mu_i^{k,\prime} = \mu_i^k + h_i^k. \tag{42}$$

In other words, h_i^k is a bias. By using the new noise model, we also assume that the noise term $w_i^k|\forall_i$ *do not change* in both training and test sessions. By rewriting Eqn. (41) using Eqn. (42), we obtain:

$$y_i^{k,\prime} = \mu_i^k + h_i^k + w_i^k. \tag{43}$$

Note that Eqn. (43) is also true for Y_{COM}^k and their normalised counterparts (i.e., $Y_{COM}^{k,norm}$ and $Y_i^{k,norm}$). Therefore, it is also valid to write:

$$y_{COM}^{norm,k,\prime} = \mu_{COM}^{norm,k} + h_{COM}^{norm,k} + w_{COM}^{norm,k}, \tag{44}$$

By definition of $y_{COM}^{norm,k}$, it follows that:

$$h_{COM}^{norm,k} = \frac{1}{N}\sum_{i=1}^{N}\frac{h_i^k}{\sigma_i}, \tag{45}$$

and $w_{COM}^{norm,k}$ is defined similarly.

With the noise model in Eqn. (44), the mean of $Y_{COM}^{norm,k,\prime}$ is:

$$\begin{aligned} \mu_{COM}^{norm,k,\prime} &= E\left[Y_{COM}^{norm,k,\prime}\right] = \frac{1}{N}\sum_{i=1}^{N} E\left[\mu_i^{norm,k,\prime}\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{\mu_i^k + h_i^k - \mu_i^{all}}{\sigma_i^{all}} \\ &= \mu_{COM}^{norm,k} + \frac{1}{N}\sum_i \frac{h_i^k}{\sigma_i^{all}} \\ &\equiv \mu_{COM}^{norm,k} + h_{COM}^{norm,k}. \end{aligned} \tag{46}$$

Using Eqn. (46), the class-dependent Mean-Squared Error (MSE) due to this mismatch can be calculated as follows:

$$
\begin{aligned}
& E\left[\left(Y_{COM}^{norm,k}-\mu_{COM}^{norm,k,\prime}]\right)^2\right] \\
&= E\left[\left(Y_{COM}^{norm,k}-\mu_{COM}^{norm,k}-h_{COM}^{norm,k}\right)^2\right] \\
&= \left(h_{COM}^{norm,k}\right)^2 + E\left[\left(Y_{COM}^{norm,k}-\mu_{COM}^{norm,k}\right)^2\right] \\
&= \underbrace{\left(h_{COM}^{norm,k}\right)^2} + \underbrace{(V_{AV}^{norm,k})^2 + (V_{COV}^{norm,k})^2}. \qquad (47)
\end{aligned}
$$

where the first underbraced term is bias^2 and the second underbraced term is variance of the fused score (found in the training set). As defined in Eqn. (31), the second term can be further decomposed into $(V_{AV}^{norm,k})^2$ (i.e., the average variance of all experts when used separately) and $(V_{COV}^{norm,k})^2$ (i.e., the spread of prediction; negative divergence as found in Eqn. (40)). Eqn. (47) is the so-called *bias-variance-covariance* decomposition. Note that this is a decomposition of MSE. In the context of classification, MSE is not relevant; Half Total Error Rate (HTER) is. HTER has already been defined in Eqn. (17) as a function of threshold Δ (note that HTER_{min} is only a *specific* case where the threshold gives minimum EER).

The variance of $Y_{COM}^{norm,k,\prime}$ is:

$$
\begin{aligned}
(\sigma_{COM}^{norm,k,\prime})^2 &\equiv E\left[\left(Y_{COM}^{norm,k,\prime} - E\left[Y_{COM}^{norm,k,\prime}\right]\right)^2\right] \\
&= E\left[\left(\begin{array}{c}\left(Y_{COM}^{norm,k}+h_{COM}^{norm,k}\right)- \\ \left(\mu_{COM}^{norm,k}+h_{COM}^{norm,k}\right)\end{array}\right)^2\right] \\
&= E\left[\left(Y_{COM}^{norm,k}-\mu_{COM}^{norm,k}\right)^2\right] \\
&= (\sigma_{COM}^{norm,k})^2 \qquad (48)
\end{aligned}
$$

Under the new noise model, it is interesting to note that the class-dependent variance of the fused score is indeed not affected by the bias, whereas the MSE is. However, the paragraphs that follow will show that the presence of bias can adversely affect the classification error measured by HTER.

When one knows the amount of mismatch (i.e, one has access to the test data), the *a posteriori* F-ratio is:

$$
\begin{aligned}
& \text{F-ratio}_{COM,apost}^{norm} \\
&= \frac{\mu_{COM}^{norm,k=C,\prime}-\mu_{COM}^{norm,k=I,\prime}}{\sigma_{COM}^{norm,k=C,\prime}+\sigma_{COM}^{norm,k=I,\prime}} \\
&= \frac{\left(\begin{array}{c}(\mu_{COM}^{norm,k=C}+h_{COM}^{norm,k=C})- \\ (\mu_{COM}^{norm,k=I,\prime}+h_{COM}^{norm,k=I})\end{array}\right)}{\sigma_{COM}^{norm,k=C}+\sigma_{COM}^{norm,k=I}}. \qquad (49)
\end{aligned}
$$

Note that at the *a posteriori* F-ratio and its corresponding *a posteriori* EER, their corresponding threshold is at:

$$
\Delta_{apost} = \frac{\left(\begin{array}{c}(\mu_{COM}^{k=I,norm}+h_{COM}^{k=I,norm})\sigma_i^C+ \\ (\mu_{COM}^{k=C,norm}+h_{COM}^{k=C,norm})\sigma_i^I\end{array}\right)}{\sigma_i^I+\sigma_i^C}. \qquad (50)
$$

The corresponding Half Total Error Rate (HTER) will be:

$$
\begin{aligned}
\text{HTER}_{COM,apost}^{norm} &\equiv \text{EER}_{COM,apost}^{norm} \\
&= \frac{1}{2}-\frac{1}{2}\text{erf}\left(\frac{\text{F-ratio}_{COM,apost}^{norm}}{\sqrt{2}}\right) \qquad (51)
\end{aligned}
$$

When one does not know the amount of mismatch, the *a priori* threshold that will be used is the one that is estimated from the training set, i.e.,

$$
\Delta_{apri} = \frac{\mu_{COM}^{k=I,norm}\sigma_i^C+\mu_{COM}^{k=C,norm}\sigma_i^I}{\sigma_i^I+\sigma_i^C}. \qquad (52)
$$

This threshold is then applied to the mismatched test set. As a result, the *a priori* HTER (on the test set) will be:

$$
\text{HTER}_{COM,apost}^{norm} \equiv \text{HTER}_{COM}^{norm}(\Delta_{apost}) \qquad (53)
$$

where, in a general context, for any given Δ, the corresponding HTER is:

$$
\text{HTER}_{COM}^{norm}(\Delta) = \frac{1}{2}\left(\text{FAR}_{COM}^{norm}(\Delta)+\text{FRR}_{COM}^{norm}(\Delta)\right), \qquad (54)
$$

where

$$
\text{FAR}_{COM}^{norm}(\Delta) = \frac{1}{2}-\frac{1}{2}\text{erf}\left(\frac{\Delta-\mu_{COM}^{k=I,norm}-h_{COM}^{k=I,norm}}{\sigma_{COM}^{k=I,norm}\sqrt{2}}\right), \qquad (55)
$$

and

$$
\text{FRR}_{COM}^{norm}(\Delta) = \frac{1}{2}+\frac{1}{2}\text{erf}\left(\frac{\Delta-\mu_{COM}^{k=C,norm}-h_{COM}^{k=C,norm}}{\sigma_{COM}^{k=C,norm}\sqrt{2}}\right). \qquad (56)
$$

It is possible to show that

$$
\text{HTER}_{COM,apost}^{norm} < \text{HTER}_{COM,apri}^{norm}.
$$

This can be done by showing that $\text{HTER}_{COM,apost}^{norm}$ is the *global* minimum, i.e.,

$$
\Delta_{apost} = \arg\min_{\Delta} \text{HTER}_{COM}^{norm}(\Delta). \qquad (57)
$$

Hence any $\Delta \neq \Delta_{apost}$ will *not be optimal*, including Δ_{apri}. In fact this global minimum happens at EER where FAR=FRR because FRR is an increasing function of the threshold and FAR is a decreasing function of the threshold.

In summary, this section shows that the bias-variance-covariance decomposition (of MSE) is not relevant for classification problems. Specifically, in a two-class problem such as biometric authentication, the concepts of *a priori* and *a posteriori* threshold play an important role in decision-making because these thresholds directly affect a specific classification error called EER[3]. Under mismatch between training and test data sets (due to the bias parameter h_i), the optimal threshold will not be the one found on the training data. This section actually provides a realistic theoretical framework that can be used to measure the amount of error due to this mismatch (in terms of HTER) by using Eqns. (54–56). Under such mismatch, the best decision one can make, supposing that one has access to the test data, is to use the *a posteriori* threshold.

[3]This idea can be extended to the general classification error but this is not the focus of this paper. The last paragraph of Section IX drafts a procedure to do so.

Of course in reality, the mismatch is unknown in advance. One possible solution will be to *estimate* the probable mismatch and then to pick the corresponding bias h_i^k. This bias can then be used to calculate the new threshold using Eqn. (50).

IX. Limitations and Future Extensions

It should be noted that the present theoretical study is limited to the *mean* operator. An extension to the weighted sum was investigated in [17]. To decide whether to perform fusion or not, the best way is perhaps to directly perform the fusion experiments. However, a more efficient way to achieve the same goal is to *predict* the fusion performance based on F-ratio (without explicitly running the experiments) as done in [17]. Performance estimation is particularly useful to select a subset of experts that will provide the best fusion performance in a more efficient way as compared to direct experimentation. This is our current research direction.

This study is also limited to the assumption that the client and impostor score distributions are *Gaussian*. However, as shown in this study, deviation from such assumption *did not severely* impact the predicted EER performance. This may be due to the unimodal nature of the class-dependent distribution. To make the analysis more general, one obvious way is to *avoid* the Gaussian assumption. One possible research direction is to use different distributions to fit the nature of the scores (e.g., Poisson distribution for distance scores, $\mathcal{X}^2$ distribution for log-likelihood ratios, etc) or by an emprical procedure that *finds the most suitable* and existing distribution using the KS-statistics. Another direction is to use a mixture of Gaussians. In this case, an analytical study such as the one done here may not be possible. Yet another direction is to make the client-dependent scores more normally distributed. This will maintain the analytical solution while tolerating for imprecision due to non-confirmity of the Gaussian assumption.

Although this study considers only zero-mean unit-variance normalisation (or z-score), one can replace the global mean and standard deviation parameters to any bias and scaling values estimated by different procedure. Suppose that these two parameters are called B and A. Any linear transformation will necessarily make use of these two parameters. As a result, our analysis can be *directly* extended to any *linear score* transformation (or normalisation), such as those proposed in [19]. However, the same analysis cannot be proceeded for non-linear score transformation. One way to get around this issue is to estimate the VR-EER analysis parameters on the *transformed* (or normalised) scores by the respective non-linear function instead of working on the unnnormalised score space.

Lastly, although only the EER value is studied here (as in Section III-B), one can extend the present finding to a more general case, whereby the EER constraint by its definition, i.e, $\mathrm{EER}(\Delta) = \mathrm{FAR}(\Delta) = \mathrm{FRR}(\Delta)$, does not hold anymore. Given the mean and variance of client and impostor distributions, the following procedure can be used to find FAR and FRR for an arbitrary threshold Δ:

$$\mathrm{FAR}(\Delta) = 1 - \int_{-\infty}^{\Delta} Y^{k=I} dy \tag{58}$$

and

$$\mathrm{FRR}(\Delta) = \int_{\Delta}^{\infty} Y^{k=C} dy \tag{59}$$

When they are assumed to be Gaussian distributed, they have the following parametric forms:

$$\mathrm{FAR}(\Delta) = \frac{1}{2} - \frac{1}{2}\mathrm{erf}\left(\frac{\Delta - \mu^{k=I}}{\sigma^{k=I}\sqrt{2}}\right), \tag{60}$$

and

$$\mathrm{FRR}(\Delta) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{\Delta - \mu^{k=C}}{\sigma^{k=C}\sqrt{2}}\right) \tag{61}$$

where the erf function is defined in Eqn. (13). When applying the constraint $\mathrm{FAR}(\Delta) = \mathrm{FRR}(\Delta)$, one obtains Eqn. (11). For any other values of Δ, there is unfortunately no simple expression, without further introducing simplication to the erf function. However, they can be computed *numerically* for each of the Δ values, using Eqns. (58) and (59), or Eqns. (60) and (61) when assuming that the class-dependent scores are normally distributed. Because of the need of using numerical calculation, analysis from Section VI onwards is cumbersome, although not impossible. Note that the plot of ($\mathrm{FAR}(\Delta)$, $\mathrm{FRR}(\Delta)$) for all Δ values is called the Receiver's Operating Curve (ROC) or the DET curve when plotted on a non-linear scale. When derived using Eqns. (60) and (61), they are truely Gaussians and will show up as straight lines on the Decision Errror Trade-off (DET) curve. Hence, as can be seen, extending the finding from EER to the more general case is possible. However, numerical simulation is unavoidable[4].

X. Conclusion

Combining multiple information sources such as subbands, streams (with different features) and multi modal data has shown to be a very promising trend, both in experiments and to some extent in real-life biometric authentication applications. Despite considerable efforts in fusion, there is a lack of understanding on the roles and effects of correlation and variance (of both the client and impostor scores of base-experts). In this paper, we proposed a theoretical model of Equal Error Rate as a function of F-ratio, which itself is a function of correlation, variance of base-expert and the difference of mean of both client and impostor distributions. The fundamental assumption is that the underlying client and impostor scores distributions are Gaussian. Although this assumption is not always true, based on 1186 experiments taken from the BANCA multimodal database, it was found that the predicted theoretical EER points based on the Gaussian assumption match approximately the EER points computed empirically from the scores directly. This is a strong indication the Gaussian assumption is acceptable in practice.

This analysis takes into account the effect of score normalisation. While there exists a lot of literature on fusion, scores are often assumed to be independent. Here, we explicitly considered this factor and verified the proposed theoretical model using the BANCA multimodal database. Experimental

[4]Due to limited space, we only describe the procedure to do so here, even though the DET curve could have been plotted as an illustration

results show that the higher the variance of base-experts and its covariance counterpart, the lower the F-ratio will be and consequently the higher Equal Error Rate (EER) will be. This is because F-ratio is inversely proportional to EER. Variance of base-experts determines how good their average quality is when each base-expert acts individually. Lower variance means better performance. Covariance among base-experts measures how dependent they are (note that by definition, correlation is a"normalised covariance", hence correlation is proportional to covariance). The more dependent they are, the lesser the gain one can benefit out of fusion.

Furthermore, through the VR-EER analysis, it is discovered that variance and covariance of base-experts are not the only criterion that determine fusion performance, the mean difference between fused client and impostor scores is another. The bigger it is, the better F-ratio and hence the lower EER will be.

Using a mean operator as fusion, we analysed four commonly encountered scenarios in biometric authentication which include fusing correlated/uncorrelated base-experts of similar/different performances. The analysis explains and shows that fusing two systems of different performances is *not always* beneficial. The theoretical analysis shows that if the weaker base-expert has (class-dependent) variance three times larger that of the best base-expert, the gain due to fusion breaks down. This is even more true for correlated base-experts as correlation penalises this limit further. We also showed that fusing two uncorrelated base-experts of similar performance *always* result in improved performance. Finally, fusing two correlated base-experts of similar performance will be beneficial only when the covariance of the two base-experts are less than the variance of the best expert. In any case, positive correlation "hurts" fusion.

We also linked the concepts of ambiguity decomposition and bias-variance-covariance decomposition to classification problems using EER evaluation criterion. The result of analysis shows that "diversity", which measures the spread of prediction score with respect to the fused score, is actually negative of covariance. As a result, analysing diversity alone is necessary *but not sufficient* to estimate good fusion, unless measures are taken to normalise the variance against a "canonical" mean (Section VI). This somewhat confirms the findings in [11]. By linking bias-variance-covariance decomposition to classification problems, we showed that bias or mismatch between training and test sets of scores of the base-experts can affect the mean and variance components of the fused scores. It is found that if the bias of base-experts can be learned from the data, such bias can be incorporated into the fusion system.

Finally, several limitations of our analysis were presented in Section IX. Future research directions will concentrate on removing the Gaussian assumption, extending the analysis to the more general linear combination of scores (instead of mean). Although zero-mean unit-variance normalisation was used here, we also showed that the analysis can be generalised easily to *any linear score transformation* or normalisation (on a per base-expert basis). Although EER is studied here, the more general case of error whereby the constraint FAR=FRR does not hold any more can be extended easily. Unfortunately, in this case, numerical analysis (instead of analytical analysis as done here) would have been required.

Acknowledgments

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. The authors also thank the anonymous reviewers for their excellent work in pointing out errors and offering suggestions for correction. This publication only reflects the authors' view.

Appendix I
Proof of Equivalence between Empirical F-ratio and Theoretical F-ratio

The estimated theoretical and empirical parameters can be shown to be exactly the same mathematically. Suppose there are M^k accesses, where $M^{k=C}$ are the number of client accesses and $M^{k=I}$ are the number of impostor accesses. Suppose also that $Y^k_{i,u}$ is the output of the i-expert and u-th access given that the class label is $k = \{C, I\}$, and $i = 1, \ldots, N$ and $u = 1, \ldots, M^k$. μ^k_i can be estimated by:

$$\hat{\mu}^k_i \equiv \frac{1}{M}^k \sum_{u=1}^{M^k} Y^k_{i,u} \equiv \bar{Y}^k_{i,\cdot}. \tag{62}$$

For the u-th access, the combined score is:

$$\frac{1}{N}\sum_{i=1}^{N} Y^k_{\cdot,u} \equiv \bar{Y}^k_{\cdot,u}. \tag{63}$$

The empirical estimate of μ^k_{COM}, $\hat{\mu}^k_{COM,emp}$ is given by:

$$\frac{1}{M}\sum_{u=1}^{M^k} \bar{Y}^k_{\cdot,u} \equiv \bar{Y}^k_{\cdot,\cdot}. \tag{64}$$

Note that:

$$\begin{aligned}
\hat{\mu}^k_{COM,emp} &= \frac{1}{M}\sum_{u=1}^{M^k} \bar{Y}_{\cdot,u} \\
&= \frac{1}{N}\sum_{i=1}^{N} \bar{Y}_{i,\cdot} \quad \left(\begin{array}{c}\text{interchange the} \\ i \text{ and } u \text{ summations}\end{array}\right) \\
&= \frac{1}{N}\sum_{i=1}^{N} \hat{\mu}^k_i \\
&= \hat{\mu}^k_{COM,theo}.
\end{aligned} \tag{65}$$

Hence, they are the same. The empirical variance can be calculated as follows:

$$\left(\hat{\sigma}^k_{COM,emp}\right)^2 = \frac{1}{M}\sum_{u=1}^{M} \left(\bar{Y}_{\cdot,u} - \bar{Y}_{\cdot,\cdot}\right) \tag{66}$$

The theoretical variance is obtained by estimating the terms $(\sigma^k_i)^2$ and $\rho^k_{i,j}\sigma^k_i\sigma^k_j$ in the expression of $(\sigma^k_{COM})^2$, as shown in Eqn. (7). The estimate of $(\sigma^k_i)^2$ is given by:

$$\frac{1}{M}\sum_{u=1}^{M} \left(Y^k_{i,u} - \bar{Y}^k_{i,\cdot}\right)^2. \tag{67}$$

The estimate of $\rho^k_{i,j}\sigma^k_i\sigma^k_j$ is given by:

$$\frac{1}{M}\sum_{u=1}^{M}\left(Y^k_{i,u}-\bar{Y}^k_{i,\cdot}\right)\left(Y^k_{j,u}-\bar{Y}^k_{j,\cdot}\right)^2. \quad (68)$$

Plugging in these two estimates into the expression for $(\sigma^k_{COM})^2$, we get the theoretical estimate of the variance of the fused scores as:

$$\begin{aligned}
&\left(\hat{\sigma}^k_{COM,theo}\right)^2\\
&=\frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{M}\sum_{u=1}^{M}\left(Y^k_{i,u}-\bar{Y}^k_{i,\cdot}\right)\right]\\
&\quad+\frac{2}{N}\sum_{i=1,j>i}^{N}\left[\left(Y^k_{i,u}-\bar{Y}^k_{i,\cdot}\right)\left(Y^k_{j,u}-\bar{Y}^k_{j,\cdot}\right)\right]\\
&=\frac{1}{M}\sum_{u=1}^{M}\left[\frac{1}{N^2}\sum_{i,j=1}^{N}\left(Y^k_{i,u}-\bar{Y}^k_{i,\cdot}\right)\left(Y^k_{j,u}-\bar{Y}^k_{j,\cdot}\right)\right]\\
&=\frac{1}{M}\sum_{u=1}^{M}\left(\bar{Y}_{\cdot,u}-\bar{Y}_{\cdot,\cdot}\right)\\
&=\left(\hat{\sigma}^k_{COM,emp}\right)^2. \quad (69)
\end{aligned}$$

Because the empirical and theoretical μ^k_{COM} and σ^k_{COM} are the *same*, the empirical and theoretical F-ratios will be exactly the same. Using the definition of F-ratio in Eqn. (12), the theoretical F-ratio of the combined score can be defined as:

$$\text{F-ratio}_{COM,theo} \equiv \frac{\hat{\mu}^{k=C}_{COM,theo}+\hat{\mu}^{k=I}_{COM,theo}}{\hat{\sigma}^{k=C}_{COM,theo}+\hat{\sigma}^{k=I}_{COM,theo}}. \quad (70)$$

The empirical F-ratio is:

$$\begin{aligned}
\text{F-ratio}_{COM,emp} &\equiv \frac{\hat{\mu}^{k=C}_{COM,emp}+\hat{\mu}^{k=I}_{COM,emp}}{\hat{\sigma}^{k=C}_{COM,emp}+\hat{\sigma}^{k=I}_{COM,emp}}\\
&= \frac{\hat{\mu}^{k=C}_{COM,theo}+\hat{\mu}^{k=I}_{COM,theo}}{\hat{\sigma}^{k=C}_{COM,theo}+\hat{\sigma}^{k=I}_{COM,theo}}\\
&= \text{F-ratio}_{COM,theo} \quad (71)
\end{aligned}$$

Hence, the theoretical F-ratio is exactly the same as the empirical F-ratio. This applies also for normalised version of Y. □

REFERENCES

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

[2] L. Hong, A. Jain, and S. Pankanti, "Can Multibiometrics Improve Performance?" Computer Science and Engineering, Michigan State University, East Lansing, Michigan, Tech. Rep. MSU-CSE-99-39, 1999.

[3] J. Kittler, G. Matas, K. Jonsson, and M. Sanchez, "Combining Evidence in Personal Identity Verification Systems," *Pattern Recognition Letters*, vol. 18, no. 9, pp. 845–852, 1997.

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

[5] L. Kuncheva., J. Bezdek, and R. Duin, "Decision Template for Multiple Classifer Fusion: An Experimental Comparison," *Pattern Recognition Letters*, vol. 34, pp. 228–237, 2001.

[6] 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*, vol. 17, no. 1, p. 1, 1995.

[7] 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)*, Avignon, 1998, pp. 115–119.

[8] 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)*, Montreal, 2004, pp. vol. V, 893–896.

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

[10] N. Ueda and R. Nakano, "Generalisation Error of Ensemble Estimators," in *Proc. Int'l conf. on Neural Networks*, 1990, pp. 90–95.

[11] G. Brown, "Diversity in neural network ensembles," Ph.D. dissertation, School of Computer Science, Uni. of Birmingham, 2003.

[12] 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.

[13] C. Marcel, "Multimodal Identity Verification at IDIAP," IDIAP, Martigny, Switzerland, Communication Report 03-04, 2003.

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

[15] J. Daugman, "Biometric decision landscapes," University of Cambridge Computer Laboratory, Tech. Rep. TR482, 2000.

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

[17] 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.

[18] C. Shipp and L. Kuncheva, "Realtionships Between Combination Methods and Measures of Diversity in Combining Classifiers," *Information Fusion*, vol. 3, pp. 135–148, 2002.

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

Norman Poh Hoon Thian obtained his M.Sc. degree from the University of Science, Penang, Malaysia and obtained DEA (Diplome d'Étude Approfondie) from Université Louis Pasteur, Strasbourg, France. He is currently a Ph.D. student and research assistant at the IDIAP Research Institute, Switzerland. His research interest is in biometric authentication and various pattern recognition-related algorithms, particularly in fusion based on classifiers and different features.

Samy Bengio is a senior researcher in the machine learning group at the IDIAP Research Institute since 1999, where he supervises PhD students and post-doctoral fellows working on many areas of machine learning such as support vector machines, time series prediction, mixture models, large-scale problems, speech and speaker recognition, multimodal problems, brain computer interfaces, and many more. He has obtained his PhD in computer science from Université de Montreal (1993), and spent three post-doctoral years at CNET, the research center of France Telecom, and INRS-Telecommunications (Montreal). He then worked as a researcher for CIRANO, an economic and financial academic research center, applying learning algorithms to finance. Before joining IDIAP, he was also research director at Microcell Labs, a private research center in mobile telecommunications. His current interests include all theoretical and applied aspects of learning algorithms.