



HOW DO CORRELATION AND VARIANCE OF BASE-EXPERTS AFFECT FUSION IN BIOMETRIC AUTHENTICATION TASKS?

Norman Poh Hoon Thian ^a Samy Bengio ^a
IDIAP-RR 04-18

APRIL 2004

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

HOW DO CORRELATION AND VARIANCE OF BASE-EXPERTS AFFECT FUSION IN BIOMETRIC AUTHENTICATION TASKS?

Norman Poh Hoon Thian

Samy Bengio

APRIL 2004

Abstract. 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 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 and XM2VTS multi-modal databases and NIST 2001 speaker verification database confirm our findings using VR-EER analysis. Furthermore, F-ratio is shown to be a valid criterion in place of EER as an evaluation criterion. 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.

1 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 [8]. 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. Pankanti et al [6] shed some lights 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 [9] 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 [10] 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 [12] and Behaviour Knowledge Space [7]) 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 [20] 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 in [17] 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 “factorise” 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 [11] and bias-variance-covariance decomposition [22] 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 [3].

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 2 presents briefly the BANCA experiment setup whereby 70 fusion experiments will be conducted. Section 3 discusses the preliminary findings of the VR-EER analysis and notations used. Section 4 presents what happens to fusion when scores are normalised. The effects of variance and correlation are verified in Section 5. Using these findings, we analysed four commonly encountered scenarios of fusion in Section 6. Two important analysis tools and concepts that are well-studied in regression problems are linked to a specific classification problem in Sections 7 and 8. This is followed by conclusions in Section 9.

2 Experiment Setup

The BANCA database [1] 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. 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 26 males and 26 females in the database. A set of face and speaker authentication experiments were carried out by University of Surrey (face experiments), IDIAP (speaker), UC3M (speaker) and UCL (face)¹. Details of these experiments can be found in [14]. For each protocol, we used the following score files:

- 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

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. Similarly by combining each time three baseline experts, we obtain a total of $7 \times {}^5C_3 = 70$ experiments.

Furthermore, we also needed two additional databases, namely NIST2001 and XM2VTS, to verify the relationship between theoretical and empirical ways of deriving Equal Error Rate (EER) function explained in Section 3.2. The NIST2001 database for speaker authentication task [15] that we used is taken from experiments carried out in [18]. Four types of speech features which exhibit different degrees of noise-robustness were used. For each feature set, 18 speech experiments are carried out using different noise types and signal-to-noise ratios, hence resulting in a total of 72 experiments. XM2VTS [13] is a multimodal (face and speech) database. The scores are taken from experiments carried out in [16]. For both Lausanne Protocols I and II defined in XM2VTS, there are 7 face baseline experts, 6 speech baseline experts and 32 fusion experiments.

¹Available at “ftp://ftp.idiap.ch/pub/bengio/banca/banca_scores”

3 Preliminary and Recent Findings on VR-EER

Our proposed theoretical model [17] 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).

3.1 Variance Reduction

Let $f_i(\mathbf{x})$ be the i -th instance of a feature set representing the biometric trait of person \mathbf{x} . Note that f consists of a biometric scanning device and feature extraction algorithm. For instance, $f(\mathbf{x})$ could be a set of Principal Component projected features of the digitised face image of person \mathbf{x} . The system has to decide if person \mathbf{x} is a client or impostor, i.e., the class label $k \in \{C, I\}$. The probability that \mathbf{x} belongs to class k given by an expert system with parameter θ can be written as:

$$y_i^k \equiv P(k|f_i(\mathbf{x}), \theta_i) \quad (1)$$

For example, i could denote the i -th subband of a spectrogram representing the speech of a person, the i -th stream or type of feature (e.g. Mel-scale Frequency Cepstrum Coefficients), the i -th biometric modality (e.g., speech, face or fingerprint), the i -th sample, the i -th classifier (but for the *same* access). In this context, y_i^k is referred to as an instance of the i -th *response* of person \mathbf{x} 's biometrics given by an expert system (often called a score in the literature). Typically, this output (e.g. score) is compared to a predefined threshold to make the accept/reject decision. Let μ_i^k be the mean score of the i -th response associated to class k estimated from a training set. Then we can write the mapping function of each response as the summation between the desired function and an error w_i^k :

$$y_i^k = \mu_i^k + w_i^k, \quad (2)$$

for $k \in \{C, I\}$. Note that the error term w_i^k follows an unknown distribution W_i^k with zero mean. It's distribution is governed by $f_i(\mathbf{x})$. Also, different from [17], 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 of discriminative training (i.e., using Multi-Layer Perceptrons (MLPs) or Support Vector Machines (SVMs)). Here, we would like to keep the discussion general where the output scores need not be constrained to be within $[-1, 1]$. Since w_i^k is dependent on \mathbf{x} , it is obvious that y_i^k , which follows the distribution Y_i^k , is also dependent on \mathbf{x} . We can write the expectation of Y_i^k , $E[Y_i^k]$, as:

$$E[Y_i^k] = E[\mu_i^k] + E[W_i^k] = \mu_i^k. \quad (3)$$

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 $i = 1, \dots, N$, denoted as $(\sigma_{AV}^k)^2$ is, according to [17]:

$$\begin{aligned} (\sigma_{AV}^k)^2 &= \frac{1}{N} \sum_{i=1}^N Cov(Y_i^k, Y_i^k) \\ &= \frac{1}{N} \sum_{i=1}^N E[W_i^k W_i^k] \\ &\equiv \frac{1}{N} \sum_{i=1}^N (\sigma_i^k)^2, \end{aligned} \quad (4)$$

where we defined the variance of Y_i^k to be $(\sigma_i^k)^2 \equiv E[W_i^k W_i^k]$ by definition. In the second case, all N responses are used together and are combined using the mean operator; the resultant score can be written as:

$$Y_{COM}^k = \frac{1}{N} \sum_{i=1}^N Y_i^k, \quad (5)$$

for any $k \in \{C, I\}$. The variance of Y_{COM} (over many accesses), denoted as σ_{COM}^2 , is called the *variance of average*, and can be calculated as follows (see [17] for details of this derivation):

$$\begin{aligned} (\sigma_{COM}^k)^2 &= \text{Cov}(Y_{COM}^k, Y_{COM}^k) \\ &= \frac{1}{N^2} \sum_{j=1}^N (\sigma_j^k)^2 + \frac{2}{N^2} \sum_{m=1, m < n}^N \rho_{m,n}^k \sigma_m^k \sigma_n^k, \\ &= \underbrace{\frac{1}{N} (\sigma_{AV}^k)^2}_{\text{average variance}} + \underbrace{\frac{2}{N^2} \sum_{m=1, m < n}^N \rho_{m,n}^k \sigma_m^k \sigma_n^k}_{\text{covariance}} \end{aligned}$$

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

$$\rho_{m,n}^k \sigma_m^k \sigma_n^k = E[W_m^k W_n^k], \quad (6)$$

by definition of correlation. Note that $\rho_{n,n}^k = 1$ for $k \in \{C, I\}$. The VR analysis shows that [17]:

$$(\sigma_{COM}^k)^2 \leq (\sigma_{AV}^k)^2. \quad (7)$$

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

$$\frac{1}{N} (\sigma_{AV}^k)^2 \leq (\sigma_{COM}^k)^2. \quad (8)$$

Hence, by combining N responses using the mean operator, the resultant variance is assured to be smaller than the average (not the minimum) variance.

3.2 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). $\text{FAR}(\Delta)$ is calculated by integrating the impostor score distribution from a given threshold Δ in the score space to $+\infty$ while $\text{FRR}(\Delta)$ 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 [17] for details of this derivation) :

$$\text{EER}_p = \frac{1}{2} - \frac{1}{2} \text{erf} \left(\frac{\text{F-ratio}_p}{\sqrt{2}} \right), \quad (9)$$

where

$$\text{F-ratio}_p = \frac{\mu_p^{k=C} - \mu_p^{k=I}}{\sigma_p^{k=C} + \sigma_p^{k=I}}, \quad (10)$$

and

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp[-t^2] dt. \quad (11)$$

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 is defined as

$$\frac{\mu_p^{k=C} - \mu_p^{k=I}}{(\sigma_p^{k=C})^2 + (\sigma_p^{k=I})^2} \quad (12)$$

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 [4]. It measures how separable the client distribution is from its impostor counterpart. It is defined as:

$$d' = \frac{|\mu_p^{k=C} - \mu_p^{k=I}|}{\sqrt{\frac{1}{2}(\sigma_p^{k=C})^2 + \frac{1}{2}(\sigma_p^{k=I})^2}} \quad (13)$$

To our opinion, F-ratio should be used instead since it is directly related to EER by Eqn. (9)).

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

$$\Delta^* = \arg \min_{\Delta} |\text{FAR}(\Delta) - \text{FRR}(\Delta)|, \quad (14)$$

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

$$\text{HTER} = \frac{\text{FAR}(\Delta^*) + \text{FRR}(\Delta^*)}{2}. \quad (15)$$

Empirical EER and HTER are used interchangeably in this paper.

Let μ_{COM}^k be the mean of fused scores and μ_{AV}^k be that of the average of N responses discussed earlier, for both $k = \{C, I\}$. The first term is:

$$\begin{aligned} E[Y_{COM}^k] &= \frac{1}{N} \sum_{i=1}^N E[Y_i^k] \\ &= \frac{1}{N} \sum_{i=1}^N \mu_i^k \equiv \mu_{COM}^k, \end{aligned} \quad (16)$$

while the second term is by definition:

$$\mu_{AV}^k = \frac{1}{N} \sum_{i=1}^N \mu_i^k. \quad (17)$$

Hence, $\mu_{COM}^k = \mu_{AV}^k$. Let EER_{COM} be the EER of the combined scores and EER_{AV} be the EER of the average of scores of N responses. It can be shown that: $\text{EER}_{COM} \leq \text{EER}_{AV}$ since we know that $\sigma_{COM}^k \leq \sigma_{AV}^k$ for both $k \in \{I, C\}$. Therefore, fusing scores can reduce variance which results in reduction of EER. This formed the argument in [17] for why fusion using multiple modalities, features, and classifiers works for BA tasks.

To check how accurate the EER function is as compared to its empirical counterpart, we conducted many experiments on XM2VTS, NIST2001 and BANCA score databases as described in Section 2. The results are shown in Figure 1. The NIST2001 experiments (after some additive noise degradation) cover from 10% to 45% EER space while XM2VTS experiments cover from near 0% to 6% EER space. There is a gap between 6% and 10% and this is nicely covered by the BANCA experiments which cover from 0% to 34% EER space. Details of these fusion experiments can be found in [17, 18, 14].

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. On the other hand, despite the finite data, the over estimation of the theoretical EER reflects the actual HTER.

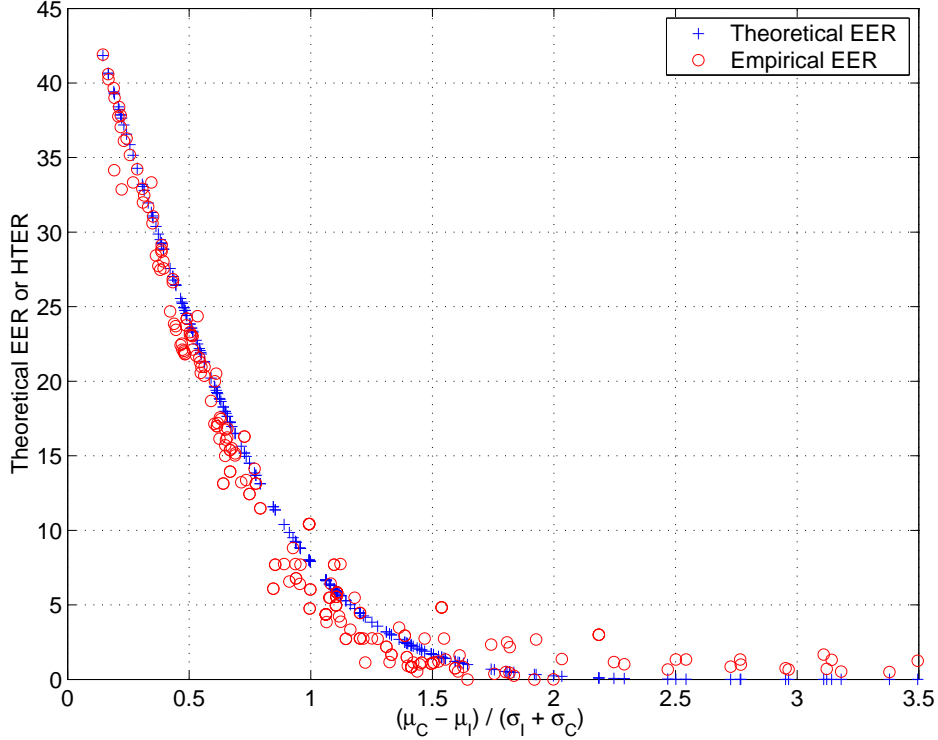


Figure 1: A Comparison between EER and HTER versus F-ratio, carried out on 180 independent experiments on XM2VTS (45 experiments), BANCA (63 experiments) and NIST2001 (72 experiments).

The second observation is that F-ratio is inversely proportional to EER, i.e., the larger F-ratio is, the better the degree of separability between the two classes and hence the lower the EER is. Hence, instead of using HTER as an empirical evaluation criterion, one alternative is actually to use F-ratio, bearing in mind that by so doing, one assumes that the underlying client and impostor distributions are Gaussian.

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

4 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 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:

$$\begin{aligned}
 Y_i^{norm,k} &= \frac{Y_i^k - E[Y_i^{all}]}{\sqrt{Cov(Y_i^k, Y_i^k)}} \\
 &\equiv \frac{Y_i^k - \mu_i^{all}}{\sigma_i^{all}},
 \end{aligned}
 \tag{18}$$

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

$$Y_{COM}^{norm,k} = \frac{1}{N} \sum_{i=1}^N Y_i^{norm,k}. \quad (19)$$

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

$$\begin{aligned} \mu_{COM}^{norm,k} &\equiv E[Y_{COM}^{norm,k}] \\ &= \frac{1}{N} \sum_{i=1}^N E[Y_i^{norm,k}] \\ &= \frac{1}{N} \sum_{i=1}^N \frac{E[Y_i^k] - \mu_i^{all}}{\sigma_i^{all}} \\ &= \frac{1}{N} \sum_{i=1}^N \frac{\mu_i^k - \mu_i^{all}}{\sigma_i^{all}}. \end{aligned} \quad (20)$$

Using Eqns. (3), (18) and (19), the variance of $Y_{COM}^{norm,k}$ is:

$$\begin{aligned} (\sigma_{COM}^{norm,k})^2 &= Cov(Y_{COM}^{norm,k}, Y_{COM}^{norm,k}) \\ &= E \left[\left(Y_{COM}^{norm,k} - E[Y_{COM}^{norm,k}] \right)^2 \right] \\ &= E \left[\left(\frac{1}{N} \sum_{i=1}^N \frac{Y_i^k - \mu_i^{all}}{\sigma_i^{all}} - \frac{1}{N} \sum_{m=1}^N \frac{\mu_m^k - \mu_m^{all}}{\sigma_m^{all}} \right)^2 \right] \\ &= E \left[\left(\frac{1}{N} \sum_{i=1}^N \frac{Y_i^k - \mu_i^k}{\sigma_i^{all}} \right)^2 \right] \\ &= E \left[\left(\frac{1}{N} \sum_{i=1}^N \frac{W_i^k}{\sigma_i^{all}} \right)^2 \right]. \end{aligned} \quad (21)$$

To expand Eqn. (21), one should take care of possible correlation between different W_m^k and W_n^k , similar to Eqn. (6), as follows:

$$\begin{aligned} (\sigma_{COM}^{norm,k})^2 &= E \left[\frac{1}{N^2} \left(\sum_{m=1}^N \sum_{n=1}^N \frac{W_m^k W_n^k}{\sigma_m^{all} \sigma_n^{all}} \right) \right] \\ &= \frac{1}{N^2} \sum_{j=1}^N \frac{E[W_j^k W_j^k]}{\sigma_j^{all}} \\ &\quad + \frac{2}{N^2} \sum_{m=1, m < n}^N \frac{E[W_m^k W_n^k]}{\sigma_m^{all} \sigma_n^{all}}. \\ &\equiv (V_{AV}^k)^2 + (V_{COV}^k)^2, \end{aligned} \quad (22)$$

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:

$$F\text{-ratio}_{COM}^{norm} = \frac{\mu_{COM}^{norm,k=C} - \mu_{COM}^{norm,k=I}}{\sigma_{COM}^{norm,k=C} + \sigma_{COM}^{norm,k=I}}. \quad (23)$$

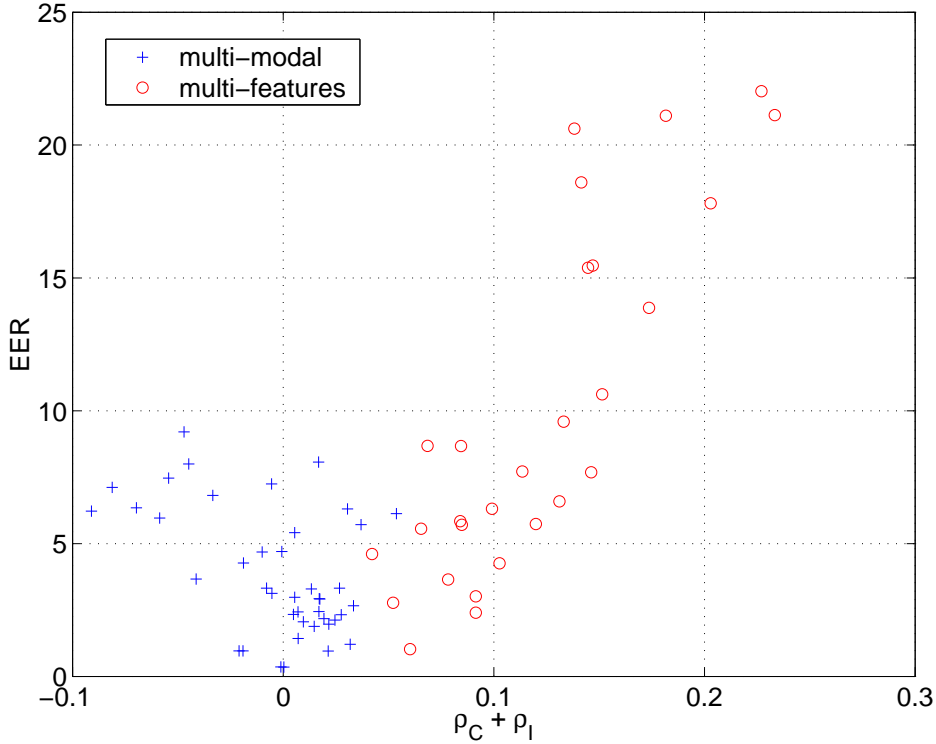


Figure 2: Empirical EER of combining 2 baseline experts versus $\rho_C + \rho_I$ using the BANCA database. The crosses represent experiments combining 2 modalities while the circles represent those combining 2 features of the *same* modality.

5 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 $F\text{-ratio}_{COM}^{norm}$, we carried out experiments on the BANCA fusion database, each time combining only two experts. These experiments can be divided into two types: multi-modal fusion (fusion of two different modalities, i.e, face and speech experts) and feature expert fusion (of two face experts *or* two speech experts). We expect the multi-modal fusion to be less correlated while the feature expert fusion to be more correlated. This is an important aspect so that both sets of experiments will cover a large range of correlation values.

The goal is to find empirically the relationship between HTER and the sum of correlation of client and impostor distributions. Let the client and impostor-dependent correlations between them be the scalars ρ_C and ρ_I , respectively². The results are shown in Figure. 2. From this figure, it can be observed that multi-modal fusion experiments have less correlated scores while multi-feature fusion experiments have high correlated scores. One would have expected that HTER (or empirical EER) 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

²In 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 $\rho_{1,2}$.

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. The evaluation procedures are as follows:

1. Verify the relationship between theoretical and empirical $V_\mu \equiv \mu_{COM}^{norm,k=C} - \mu_{COM}^{norm,k=I}$, i.e, the numerator of Eqn. (23).
2. Verify the relationship between theoretical and empirical $V_\sigma \equiv \sigma_{COM}^{norm,k=C} + \sigma_{COM}^{norm,k=I}$ i.e, the denominator of Eqn. (23).
3. Verify the relationship between theoretical and empirical F-ratio, i.e., Eqn. (23) itself.

The results of steps 1, 2 and 3 are shown in Figures 3(a), (b) and (c), respectively. From (a) and (b), it is observed that both V_μ and V_σ are actually predictable. As a result, in (c), the theoretical F-ratios match exactly its empirical counterparts.

We now know *exactly* the composition of F-ratio of the fused (using mean) and normalised (using zero-mean unit-variance) scores. It is now possible to decompose the F-ratio into its three important components, namely

1. the mean difference V_μ ,
2. the sum of standard deviations of base-experts $V_{AV}^{k=C} + V_{AV}^{k=I}$, and
3. the sum of square-root of covariance of base-experts $V_{COV}^{k=C} + V_{COV}^{k=I}$.

These three components are factorised according to Eqn. (23), with variance obtained from Eqn. (22). By such factorisation, we consider one of the three factors at a time (the rest of the factors are thus considered absent).

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 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 of covariance matrix of the fused scores (for both client and impostor scores). It measures how correlated the base-experts are. For the last two components, the higher they are, the lower the F-ratio will be. Hence, for good fusion, one should maximise the first component and minimise the second and third components. Because of these three interrelated factors, analysing any one of them alone, as done in Figure 2, 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 3 and 4 experts and were able to predict the F-ratio accurately. The results are similar to those presented in Figure. 3 (not shown here). This is somewhat expected because the VR-EER analysis is not limited to 2 experts.

A follow-up study using weighted sum [19] 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 their theoretical counterparts (see Figure 4). The weights are estimated using Fisher-ratio criterion [2, pg. 110]. However, F-ratios of fusion experiments *between* the development and the evaluation sets *do not match* exactly, but they are correlated (see Figure 5). This is due to the inherent mismatch between the development and evaluation sets. These two experiments were performed using the *same data sets* as fusion experiments presented in Figure 3 (which used the mean operator instead of weighted sum). More details on how to derive the F-ratio of weighted-sum fusion can be found in [19].

6 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 multi-modal 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 [20] 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 $\sigma_i^{k'}$ by their 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:

$$\text{F-ratio} = \frac{\mu_i^{k=C} - \mu_i^{k=I}}{\sigma_i^{k=C} + \sigma_i^{k=I}} = \frac{1 - (-1)}{\sigma_i^{k=C'} + \sigma_i^{k=I'}}. \quad (24)$$

The solution is:

$$\sigma_i^{k'} = \alpha'_i \sigma_i^k, \quad (25)$$

where,

$$\alpha'_i = \frac{2}{\mu_i^{k=C} - \mu_i^{k=I}},$$

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

By taking the square of Eqn. (25) and applying the definition of variance of y_i , we obtain

$$\begin{aligned} (\sigma_i^{k'})^2 &= (\alpha'_i)^2 E[(y_i - E[y_i])^2] \\ &= E[(\alpha'_i(y_i - E[y_i]))^2] \end{aligned} \quad (26)$$

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 $V_{AV,i}^{k=C} \propto V_{AV,i}^{k=I}$ for system $i \in \{1, 2\}$ and $V_{COV}^{k=C} \propto V_{COV}^{k=I}$ (covariance between the two systems). 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. (23).

For simplicity,

- $\sigma_{i=1}^2 = (V_{AV,i=1}^{k=C})^2 = a$,
- $\sigma_{i=2}^2 = (V_{AV,i=2}^{k=C})^2 = b$, and
- $(V_{COV}^k)^2 = c$.

For the first case, without loss of generality, we have $a > b$ and $c \simeq 0$. Hence, for the combination to be *better than the best system*, i.e., system a , it is required that:

$$\begin{aligned} \sigma_{COM}^2 &< \sigma_{i=1}^2 \\ \frac{a+b+2c}{4} &< a \end{aligned} \quad (27)$$

σ_{COM}^2 is calculated using Eqn. (6) with $N=2$.

We see that:

$$b < 3a - 2c.$$

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

Hence, the combined system will benefit from the fusion when $\sigma_{i=2}^2$ is *at most* less than 3 times of $\sigma_{i=1}^2$ since $c \simeq 0$.

Furthermore, correlation (or equivalently covariance; one is proportional to the other; See Eqn. (6)) between the two systems penalises this margin of $3\sigma_{i=1}^2$. This is particularly true for the second case since $c > 0$. It can be seen that the fusion will loose to the best system when the covariance $c > 3a/2$. Also, it should be noted that $c \leq 0$ (which implies negative correlation) could allow for larger b . 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 [3].

For the third and fourth cases, we have $a \simeq b$. Hence, Eqn. (27) becomes

$$c < a.$$

Note that for the third case, $c \simeq 0$ which satisfies the condition $c < a$. Therefore, fusion will *definitely* lead to better performance. On the other hand, for the fourth case, fusion will lead to better performance only when the covariance between the two systems is less than the variance of the best system ($\sigma_{i=1}^2$).

7 Relation to Ambiguity Decomposition

We would like to link our findings with those of Krogh and Vedelsby [11], who showed that, in our context:

$$\begin{aligned} E[Y_{COM}^k - \mu_{COM}^k]^2 &= \sum_i \alpha_i E(Y_i^k - \mu_{COM}^k)^2 \\ &\quad - \sum_i \alpha_i E(Y_i^k - Y_{COM}^k)^2 \\ (\sigma_{COM}^k)^2 &\equiv \text{acc}^k - \text{div}^k, \end{aligned} \quad (28)$$

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. (28), the accuracy of $Y_{COM}^{norm,k}$ as discussed in section 4 is:

$$\text{acc}^k = \frac{1}{N} \sum_i E[Y_{COM}^{norm,k} - \mu_{COM}^{norm,k}]^2$$

$$\begin{aligned}
 &= \frac{1}{N} \sum_i E \left[\frac{Y_i^k - \mu_i^{all}}{\sigma_i^{all}} - \frac{1}{N} \sum_j \frac{\mu_j^k - \mu_j^{all}}{\sigma_j^{all}} \right]^2 \\
 &= \frac{1}{N} \sum_i \frac{E[W_i^k W_i^k]}{(\sigma_i^{all})^2} = (V_{AV}^k)^2.
 \end{aligned} \tag{29}$$

From Eqns. (28) and (22), it is obvious that divergence is simply:

$$\text{div}^k = -(V_{COV}^k)^2. \tag{30}$$

The negative sign in this term shows that the divergence is indeed negatively proportional to the covariance component. Hence, conclusions drawn in Section 5 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 [21], 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 3.2 and [17]. 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.

8 Relation to Bias-Variance-Covariance Decomposition

Ueda and Nakano [22] presented the bias-variance-covariance decomposition while Brown [3] 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 [3, 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.

Let us introduce a bias term h_i^k into the assumption presented in Eqn. (2). This term is due to mismatch during testing as oppose to μ_i^k and w_i^k , both of which are mean and noise that are associated to a training set. This new noise model can now be represented as follows:

$$y_i^k = \mu_i^k + w_i^k + h_i^k. \tag{31}$$

Note that the above equation 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:

$$y_{COM}^{norm,k} = \mu_{COM}^{norm,k} + w_{COM}^{norm,k} + h_{COM}^{norm,k}, \tag{32}$$

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

$$h_{COM}^{norm,k} = \frac{1}{N} \sum_{i=1}^N \frac{h_i^k}{\sigma_i}, \tag{33}$$

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

According to the bias-variance decomposition ([5, 11], also see [2, Sec. 9.1]), for any Y^k (i.e., Y_i^k or Y_{COM}^k , and their normalised counterparts: $Y_i^{k,norm}$ and $Y_{COM}^{k,norm}$), we have:

$$\begin{aligned}
 E[(Y^k - h^k)^2] &= (E(Y^k - h^k))^2 \\
 &\quad + E((Y^k - (\mu^k))^2).
 \end{aligned} \tag{34}$$

The first part corresponds to (bias)² while the second part corresponds to variance. The bias-variance decomposition of $Y_{COM}^{norm,k}$ can be found using this framework. Let the variance of $Y_{COM}^{norm,k}$ be:

$$(Q_{COM}^{norm,k})^2 \equiv E \left[(Y_{COM}^{norm,k} - h_{COM}^{norm,k})^2 \right]. \quad (35)$$

The resultant variance can be written as:

$$(Q_{COM}^{norm,k})^2 = \underbrace{(B^k)^2} + \underbrace{(V_{AV}^k)^2 + (V_{COV}^k)^2}. \quad (36)$$

where the first term corresponds to (bias)² and the second term is variance, according to Eqn. (34). The bias term (without the square) is:

$$\begin{aligned} B^k &= E \left[Y_{COM}^{norm,k} - h_{COM}^{norm,k} \right] \\ &= \mu_{COM}^{norm,k} - h_{COM}^{norm,k} \\ &= \frac{1}{N} \sum_{i=1}^N (\mu_i^k - h_i^k), \end{aligned} \quad (37)$$

The second term in Eqn. (36) is the original variance of the noise model (see Eqn. (2)). With the assumption in Eqn. (32), the mean of $Y_{COM}^{norm,k}$ is:

$$\begin{aligned} P_{COM}^{norm,k} &= \frac{1}{N} \sum_{i=1}^N \frac{\mu_i^k + h_i^k - \mu_i^{all}}{\sigma_i^{all}} \\ &= \mu_{COM}^{norm,k} + \frac{1}{N} \sum_i \frac{h_i^k}{\sigma_i^{all}} \\ &\equiv \mu_{COM}^{norm,k} + h_{COM}^{norm,k}. \end{aligned} \quad (38)$$

Finally, the F-ratio of this system is:

$$\text{F-ratio}_{biasCOM}^{norm} = \frac{P_{COM}^{norm,k=C} - P_{COM}^{norm,k=I}}{Q_{COM}^{norm,k=C} + Q_{COM}^{norm,k=I}}. \quad (39)$$

By writing Eqn. (39) in the terms of Eqn. (23), F-ratio_{biasCOM}^{norm} becomes:

$$\frac{\mu_{COM}^{norm,k=C} - \mu_{COM}^{norm,k=I} + h_{COM}^{norm,k=C} - h_{COM}^{norm,k=I}}{\sqrt{(B^{k=C})^2 + (\sigma_{COM}^{norm,k=C})^2} + \sqrt{(B^{k=I})^2 + (\sigma_{COM}^{norm,k=I})^2}}. \quad (40)$$

We see that $h_{COM}^{norm,k}$ for $k = \{C, I\}$ are bias correction terms and $(B^k)^2$ are variance correction terms to the resultant F-ratio. These correction terms can be very useful for correcting mismatch between training and test sets if they can be learned from the data. F-ratio can be translated into EER by using Eqn (9). With this, we provided a link between techniques well-studied in regression problems to a specific classification problem using EER, which is relevant to biometric authentication tasks.

9 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. 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 consider this factor and verify the proposed theoretical model using the BANCA multi-modal database. Experimental 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 measure 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; 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.

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. The theoretical analysis shows that if the weaker base-expert has (class-dependent) variance three times larger than 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 6). This somewhat confirms the findings in [3]. 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 scores of the combined system. 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.

In this work, we also showed that F-ratio is a valid evaluation criterion in place of the commonly used EER. One potential use of F-ratio is to select an optimal subset of base-experts for fusion. This is because adding a base-expert always incurs either additional physical hardware or computational cost. Hence, using F-ratio, one can design a systematic algorithm, and derive statistics according to Eqn. (23) analytically, without actually carrying out experiments, as often done in the literature in a black-box fashion. Preliminary experiments in [19] showed that this is indeed a feasible approach, with much lower computation than carrying out experiments by brute-force testing. This study also includes how F-ratios of fused scores *using weighted-sum* can be estimated.

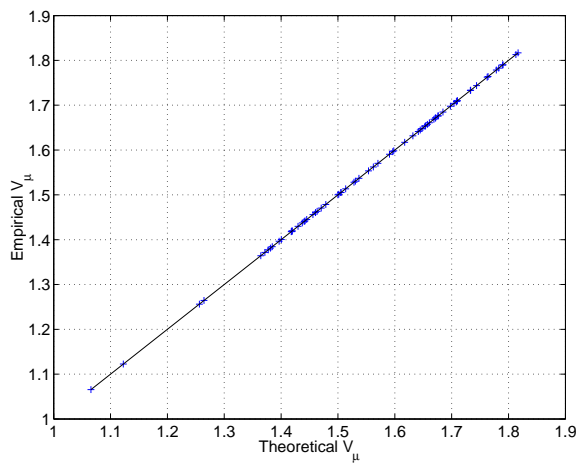
Acknowledgement

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)”.

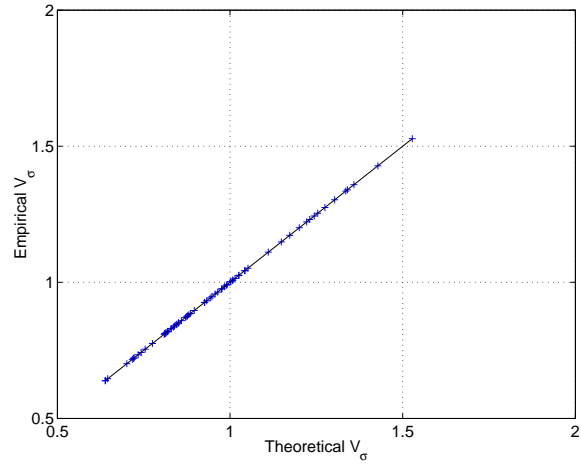
References

- [1] 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 *4th Int. Conf. Audio- and Video-Based Biometric Person Authentication, AVBPA'03*. Springer-Verlag, 2003.
- [2] C. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1999.
- [3] G. Brown. *Diversity in Neural Network Ensembles*. PhD thesis, School of Computer Science, Uni. of Birmingham, 2003.
- [4] J. Daugman. Biometric decision landscapes. Technical Report TR482, University of Cambridge Computer Laboratory, 2000.
- [5] S. Geman, E. Bienenstock, , and R. Doursat. Neural networks and the bias/variance dilemma. In *Neural Computation*, volume 4, pages 1–52, 1992.
- [6] 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.
- [7] 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.
- [8] A.K. Jain, R. Bolle, and S. Pankanti. *Biometrics: Person Identification in a Networked Society*. Kluwer Publications, 1999.
- [9] J. Kittler, G. Matas, K. Jonsson, and M. Sanchez. Combining Evidence in Personal Identity Verification Systems. *Pattern Recognition Letters*, 18(9):845–852, 1997.
- [10] 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.
- [11] A. Krogh and J. Vedelsby. Neural network ensembles, cross-validation and active-learning. *Advances in Neural Information Processing Systems*, 7, 1995.
- [12] L. Kuncheva., J.C. Bezdek, and R.P.W. Duin. Decision template for multiple classifier fusion: An experimental comparison. *Pattern Recognition Letters*, 34:228–237, 2001.
- [13] J. Lüttin. Evaluation Protocol for the XM2FDB Database (Lausanne Protocol). Communication 98-05, IDIAP, Martigny, Switzerland, 1998.
- [14] Christine Marcel. Multimodal Identity Verification at IDIAP. Communication Report 03-04, IDIAP, Martigny, Switzerland, 2003.
- [15] A. Martin. NIST Year 2001 Speaker Recognition Evaluation Plan, 2001.
- [16] 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.
- [17] N. Poh and S. Bengio. Why Do Multi-Stream, Multi-Band and Multi-Modal Approaches Work on Biometric User Authentication Tasks? Research Report 03-59, IDIAP, Martigny, Switzerland, 2003. to appear in IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP), 2004.

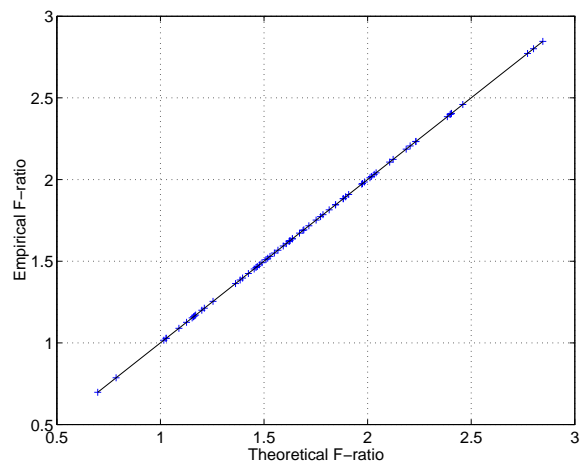
- [18] N. Poh and S. Bengio. Noise-robust multi-stream fusion for text-independent speaker authentication. Research Report 04-01, IDIAP, Martigny, Switzerland, 2004. to appear in The Speaker and Language Recognition Workshop (Odyssey), Toledo, 2004.
- [19] Norman Poh and Samy Bengio. Towards Predicting Optimal Subsets of Base-Experts in Biometric Authentication Task. Research Report 04-17, IDIAP, Martigny, Switzerland, 2004.
- [20] S. Sharma, H. Hermansky, and P. Vermuulen. Combining Information from Multiple Classifiers for Speaker Verification. In *Proc. Speaker Recognition and Its Commercial and Forensic Applications Workshop (RLA2C)*, pages 115–119, Avignon, 1998.
- [21] C.A. Shipp and L.I. Kuncheva. Relationships between combination methods and measures of diversity in combining classifiers. *Information Fusion*, 3:135–148, 2002.
- [22] N. Ueda and R. Nakano. Generalisation Error of Ensemble Estimators. In *Proc. Int'l conf. on Neural Networks*, pages 90–95, 1990.



(a) Numerator of F-ratio



(b) Denominator of F-ratio



(c) F-ratio

Figure 3: Empirical versus theoretical V_μ , V_σ and F-ratio, by fusing 2 experts. Experiments were carried out on the development set of the BANCA multi-modal database.

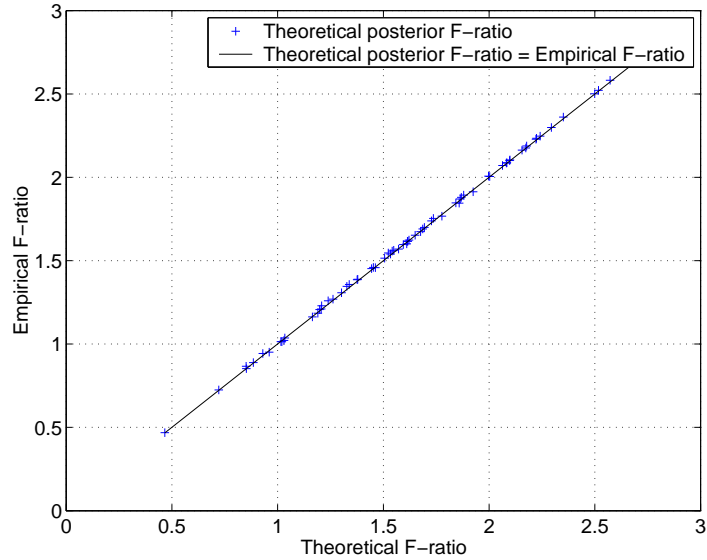


Figure 4: Comparison of a theoretical prior F-ratio an empirical F-ratio, using weighted sum, based on BANCA development set, over all possible combinations ($2^5 - 1 = 31$ of them) on all 7 different protocols, giving a total of 217 experiments.

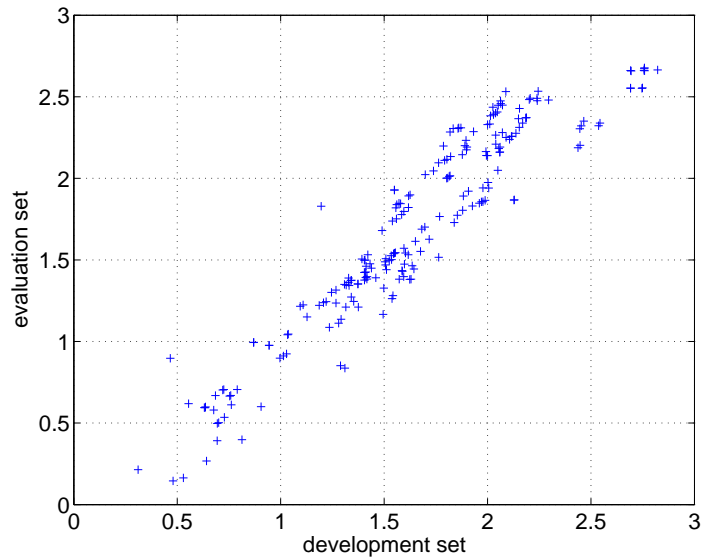


Figure 5: F-ratios of combined scores, using weighted sum, on the *development* set versus those of the *evaluation* set, over all possible 31 combinations and all 7 protocols, giving a total of 217 experiments.