EER of Fixed and Trainable Fusion Classifiers: A Theoretical Study with Application to Biometric Authentication Tasks

Norman Poh and Samy Bengio

IDIAP Research Institute, Rue du Simplon 4, CH-1920 Martigny, Switzerland
norman@idiap.ch, bengio@idiap.ch

Abstract. Biometric authentication is a process of verifying an identity claim using a person's behavioural and physiological characteristics. Due to the vulnerability of the system to environmental noise and variation caused by the user, fusion of several biometric-enabled systems is identified as a promising solution. In the literature, various fixed rules (e.g. $\min$, $\max$, median, mean) and trainable classifiers (e.g. linear combination of scores or weighted sum) are used to combine the scores of several base-systems. How *exactly* do correlation and imbalance nature of base-system performance affect the fixed rules and trainable classifiers? We study these *joint* aspects using the commonly used error measurement in biometric authentication, namely Equal Error Rate (EER). Similar to several previous studies in the literature, the central assumption used here is that the class-dependent scores of a biometric system are approximately normally distributed. However, different from them, the novelty of this study is to make a *direct link* between the EER measure and the fusion schemes mentioned. Both synthetic and real experiments (with as many as 256 fusion experiments carried out on the XM2VTS benchmark score-level fusion data sets) verify our *proposed theoretical modeling of EER* of the two families of combination scheme. In particular, it is found that weighted sum can provide the best generalisation performance when its weights are estimated correctly. It also has the additional advantage that score normalisation prior to fusion is not needed, contrary to the rest of fixed fusion rules.

1 Introduction

There exists a vast literature study that proposes to model theoretical classification errors for fusion, e.g., [1–3]. However, to the best of our knowledge, a direct modeling of Equal Error Rate (EER), i.e., an evaluation error commonly used in biometric authentication tasks, has not been attempted. This is partly because of the unknown decision threshold which prevents further analysis. Analysis of EER is cumbersome without making any assumption about the distribution of the classifier scores, e.g., using a non-parametric approach. We tackle this problem by assuming that the class-dependent scores are normally distributed. With a very large number of independent experiments, our previous work [4] shows that although the class-dependent scores are often not normally distributed, the estimated EER is *fairly robust* to deviation from such assumption.

In [1], the theoretical classification error of six classifiers are thoroughly studied for a two-class problem. This study assumes that the base classifier scores are probabilities $\in [0, 1]$. Hence probability of one class is one minus the probability of the other class and the optimal threshold is always set to 0.5. It also assumes that all baseline classifier scores are drawn from a common distribution. Gaussian and uniform distributions are studied. The first assumption is not always applicable to biometric authentication. This is because the output of a biometric system is often not necessarily a probability but a distance measure, a similarity or a log-likelihood ratio. Moreover, decisions are often taken by comparing a classifier score with a threshold. The second assumption, in practice, is also unrealistic in most situations, particularly in multimodal fusion. This is because the (class-dependent) score distributions are often *different* across different classifiers. The proposed EER model is also different from the one presented in [2, 3] in terms of application, assumption and methodology (see Section 3).

The goal of this paper is thus to study the EER of fixed and trainable fusion classifiers with respect to the correlation and the imbalance performance nature of baseline systems. Section 2 briefly discusses the general theoretical EER framework and how it can be applied to study several commonly used fusion classifiers. Section 3 discusses the important assumptions made and draws differences between EER and current theoretical model to explaining why fusion works. Sections 4 and 5 present experimental results on synthetic and real data. These are followed by conclusions in Section 6.

2 Theoretical EER

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

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

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

Note that there exists several types of classifiers in biometric authentication, 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). the context of multimodal BA, y is associated to the subscript i, which takes on different meanings in different context of fusion, as follows:

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

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 to $y_i(x)$ as the i-th *response* and there are altogether N responses ($i = 1, \ldots, N$). It is important to note that all $y_i(x)$ belong to the *same* access. We write y_i instead of $y_i(x)$ for simplicity, while bearing in mind that y_i is always dependent on x.

To decide if an access should be granted or not, all $y_i|\forall_i$ have to be combined to form a single output. This can be expressed as: $y_{COM} = f_{COM}(y_1, \ldots, y_N)$. Several types of combination strategies are used in the literature, e.g., min, max, median, mean (or sum), weighted sum, product and weighted product. They are defined as follow:

$$\begin{array}{ccc} y_{min} = \min_i(y_i), & y_{max} = \max_i(y_i) & y_{med} = \mathrm{median}_i(y_i), \\ y_{wsum} = \sum_{i=1}^{N} w_i y_i, & \text{and} & y_{wprod} = \prod_{i=1}^{N} y_i^{w_i}, \end{array} \tag{3}$$

where $w_i|\forall_i$ are parameters that need to be estimated. The mean operator is a special case of weighted sum with $w_i = \frac{1}{N}$. Similarly, the product operator is a special case of weighted product with $w_i = 1$.

The decision function based on the score y (for any y after fusion $\{y_{COM}|COM \in \{\min, \max,$ $\text{mean}, \text{median}, wsum, prod, wprod\}$ or any y_i prior to fusion; both cases are refered to simply as y) is defined as:

$$\text{decision} = \begin{cases} \text{accept} & \text{if } y > \Delta \\ \text{reject} & \text{otherwise.} \end{cases} \quad (4)$$

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

$$\text{FAR}(\Delta) = \frac{\text{FA}(\Delta)}{NI} \qquad \text{and} \qquad \text{FRR}(\Delta) = \frac{\text{FR}(\Delta)}{NC}, \quad (5)$$

where FA(Δ) counts the number of FA, FR(Δ) counts the number of FR, NI is the total number of impostor accesses and NC is the total number of client accesses.

At this point, it is convenient to introduce two conditional variables, $Y^k \equiv Y|k$, for each k being client or impostor, respectively i.e., $k \in \{C, I\}$. Hence, $y^k \sim Y^k$ is the score y when person x is $k \in \{C, I\}$. Let $p(Y^k)$ be the probabilistic density function (*pdf*) of Y^k. Eqns. (5) can then be re-expressed by:

$$\text{FAR}(\Delta) = 1 - p(Y^I > \Delta) \qquad \text{and} \qquad \text{FRR}(\Delta) = p(Y^C > \Delta). \quad (6)$$

Because of Eqn. (4), it is implicitly assumed that $E[Y^C] > E[Y^I]$, where $E[z]$ is the expectation of z. When $p(Y^k)$ for both $k \in \{C, I\}$ are assumed to be Gaussian (normally distributed), they take on the following parametric forms (see [4]):

$$\text{FAR}(\Delta) = \frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{\Delta - \mu^I}{\sigma^I\sqrt{2}}\right) \qquad \text{and} \qquad \text{FRR}(\Delta) = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{\Delta - \mu^C}{\sigma^C\sqrt{2}}\right) \quad (7)$$

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

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp\left[-t^2\right] dt. \quad (8)$$

At Equal Error Rate (EER), FAR=FRR. Solving this constraint yields (see [4]):

$$\text{EER} = \frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{\text{F-ratio}}{\sqrt{2}}\right) \equiv \text{eer}(\text{F-ratio}) \qquad \text{where} \qquad \text{F-ratio} = \frac{\mu^C - \mu^I}{\sigma^C + \sigma^I}. \quad (9)$$

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

2.1 Theoretical EER of Fusion Classifier

We now derive several parametric forms of fused scores using different types of classifiers, namely the single-best classifier, mean, weighted sum, product rule and Order Statistics (OS)-combiners such as $\min$, $\max$ and median. The OS-combiners are further discussed in Section 2.2.

The analysis in this section is possible due to the simple expression of F-ratio, which is a function of four parameters: $\{\mu^k, \sigma^k|\forall_{k=\{C,I\}}\}$ as shown in Eqn. (9). Suppose that the i-th response is y_i^k sampled from $p(Y_i^k)$ and there are N classifiers, i.e., $i = 1, \ldots, N$. The *average baseline* performance of classifiers, considering that each of them works independently of the other, is shown in the first row of Table 1. The (class-dependent) average variance, σ_{AV}^k, is defined as the average over all the variances of classifier. This is in fact not a fusion classifier but the *average performance* of classifiers measured in EER. The single-best classifier in the second

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

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

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

row chooses the baseline classifier that maximises the F-ratio. This is the same as choosing the one with minimum EER because F-ratio is inversely proportional to EER, as implied by the left part of Eqn. (9).

The derivation of EER of weighted sum (as well as mean) fusion can be found in [5]. The central idea consists of projecting the N dimensional score onto a one dimensional score via the fourth equation in Eqns. (3). Suppose that the class conditional scores (prior to fusion) are modeled by a multivariate Gaussian with mean $(\boldsymbol{\mu}^k)^T = \mu^k_1, \ldots, \mu^k_N$ and covariance $\boldsymbol{\Sigma}^k$ of N-by-N dimensions. Let $\boldsymbol{\Sigma}^k_{i,j}$ be the i-th row and j-th column of covariance matrix $\boldsymbol{\Sigma}^k$ for $k = \{C, I\}$. $E[\cdot]$ is the expectation operator (over samples) and W^k_i is the noise variable associated to classifier i for all k. The linear projection from N dimensions of score to one dimension of score has the same effect on the Gaussian distribution: from N multivariate Gaussian distribution to a single Gaussian distribution with mean μ^k_{wsum} and variance $(\sigma_{wsum})^2$ defined in the fourth row of Table 1 for each class k. The mean operator is derived similarly with $w_i = \frac{1}{N}\forall i$. Note that the weight w_i affects both the mean and variance of fused scores. In [4], it was shown mathematically that the EER of mean, EER_{mean}, is always smaller than or equal to the EER of the average baseline performance (EER_{AV}). This is closely related to the ambiguity decomposition [6] often used in the regression context (as opposed to classification as done in [4]). However, there is no evidence that $\mathrm{EER}_{mean} \leq \mathrm{EER}_{best}$, i.e., the EER of the best-classifier. In [7], it was shown that $\sigma^k_{wsum} \leq \sigma^k_{mean}$, supposing that the $w_i \forall_i$ are optimal. In [3], when the correlation among classifiers is assumed to be zero, $w_i \propto (\mathrm{EER}_i)^{-1}$. As a result, this implies that $\mathrm{EER}_{wsum} \leq \mathrm{EER}_{mean}$. The finding in [7] is more general than that of [3] because the underlying correlation among baseline classifiers is captured by the covariance matrix. Hence, fusion using weighted sum can, in theory, have better performance than the mean rule, assuming that the weights are tuned optimally. A brief discussion of weight-tuning procedures are discussed in Section 5.2. Although there exists several methods to tune the weights in the literature, to the best of our knowledge, no standard algorithm *directly* optimises EER (hence requiring further investigation which cannnot be dealt here).

For the product operator, it is necessary to bound Y to be within the range $[0, 1]$, otherwise the multiplication is not applicable. Consider the following case: two instances of classifier score can take on any real value. The decision function Eqn. (4) is used with optimal threshold being zero. With an impostor access, both classifier scores will be negative if correctly classified. Their product, on the other hand, will be positive. This is clearly undesirable.

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

2.2 Theoretical EER of Order Statistics Combiners

To implement fixed rule *order statistics* (OS) such as the maximum, minimum and median combiners, scores must be comparable. Unfortunately, attempting to analyse analytically the EER values as done in the previous section is difficult without making (very) constraining assumptions.

The first assumption is that the instance of scores must be *comparable*. If scores of various types of classifiers are involved for fusion, their range may not be comparable. Hence, score normalisation is imperative while this pre-processing step is *unnecessary* in the previous section. The second assumption assumes that scores are i.i.d. In this case, there exists a very simple analytical model[1]. Although this model seems too constraining, it is at least applicable to fusion with multiple samples which satisfies some of the assumptions stated here: scores are comparable; and they are *identically distributed* but unfortunately not necessarily *independently* sampled.

All OS combiners will be collectively studied. The subscript OS can be replaced by min, max and median. Supposing that $y^k_i \sim Y^k_i$ is an instance of i-th response knowing that the associated access claim belongs to class k. y^i has the following model: $y^k_i = \mu^k_i + \omega^k_i$, where μ^k_i is a deterministic component and ω^k_i is a noise component. Note that in the previous section ω^k_i is assumed to be normally distributed with zero mean. The fused scores by OS can be written as: $y^k_{OS} = \mathrm{OS}(y^k_i) = \mu^k + \mathrm{OS}(\omega^k_i)$, where i denotes the i-th sample (and not the i-th classifier output as done in the previous section). Note that μ^k is constant across i and it is *not affected* by the OS combiner. The expectation of y^k_{OS} as well as its variance are shown in the last row of Table 1, where γ_2 is a reduction factor and γ_1 is a shift factor, such that $\gamma_2(\sigma^k)^2$ is the variance of $\mathrm{OS}(\omega^k_i)$ and $\gamma_1\sigma^k$ is the expected value of $\mathrm{OS}(\omega^k_i)$. Both γ's can be found in tabulated form for various noise distributions [8]. A similar line of analysis can be found in [2] except that class-independent noise is assumed. The reduction factors of combining the first five samples, assuming Gaussian distribution, are shown in Table 2. The smaller γ_2 is, the smaller the associated EER. The fourth column of Table 2 shows the reduction factor due to mean (as compared to the second and third columns). It can be seen that mean is overall superior.

3 General Discussion

We gather here a list of assumptions made that will be used in simulating a theoretical comparison of fixed and trainable fusion classifiers listed in Table 1. For each assumption, we discuss its relevance and acceptability in practice.

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

Table 2. Reduction factor γ_2 of variance (2 for the second moment) with respect to the standard normal distribution due to fusion with $\min$, $\max$ (the second column) and median (third column) OS combiners for the first five samples according to [8]. The fourth column is the *maximum* reduction factor due to mean (at zero correlation), with minimum reduction factor being 1 (at perfect correlation). The fifth and sixth columns show the shift factor γ_1 (for the first moment) as a result of applying $\min$ and $\max$ for the first five samples. These values also exist in tabulated forms but here they are obtained by simulation. For median, γ_1 is relatively small (in the order of 10^{-4}) beyond 2 samples and hence not shown here. It approaches zero as N is large.

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

1. **Class-dependent gaussianity assumption.** Perhaps this is the most severe assumption as this does not necessarily hold in reality. In [4], 1186 data sets of scores were used to verify this assumption using the Kolmogorov-Smirnov statistics. Only about a quarter of the data sets supported the gaussianity assumption. However, to much surprise, the theoretical EER (estimated using the Gaussian assumption) matches closely its empirical counterpart (obtained by directly estimating the EER from scores). Hence, the theoretical EER employed here is somewhat robust to deviation from such assumption. This in part may be due to the fact that the classifier scores are unimodal but not necessarily Gaussian. The Gaussianity assumption is used mainly because of its easy interpretation. A mixture of Gaussian components could have been used in place of a single Gaussian. However, this subject requires a dedicated study which cannot be adequately dealt in the present context.
2. **Score comparability assumption.** This assumption is *only necessary* for OS combiners because of their nature that requires comparison relation "$\geq$". Scores can be made comparable by using score normalisation techniques. We use here the zero-mean unit-variance normalisation (or z-score), where a score is subtracted from its global mean and divided by its standard deviation, both of which are calculated from a training set. For the product rule which naturally assumes classifier outputs are probabilistic (in the range $[0, 1]$), the min-max normalisation is used. This is done by subtracting the score from its smallest value and divided by its range (maximum minus minimum value), all of which calculated from a training set
3. **Class-dependent correlation assumption.** Under such assumption, one assumes that the correlation of client and impostor distributions are correlated, i.e., $\rho^I \propto \rho^C$. This means that knowing the covariance of impostor joint distribution, one can actually estimate the covariance of the client joint distribution. A series of 70 intramodal and multimodal fusion experiments taken from the BANCA database were analysed in [4] and it was shown that the correlation between ρ^I and ρ^C is rather strong, i.e., 0.8.

Different from studies in [1, 2], we do not assume identical distribution across different classifiers. In fact, for OS combiners, the analytical EER expression that does not commit such assumption is cumbersome to be evaluated. Hence, we propose to resolve to simulations, which are relatively easier to carry out and reflect *better* the fusion tasks in biometric authentication.

Note that we do not make the independence assumption in the sense that correlation across different classifiers is non-zero. In fact, the correlation among classifier scores is captured by the covariance matrix via the definition of correlation, as follows: $\rho^k_{i,j} \equiv \frac{\boldsymbol{\Sigma}^k_{i,j}}{\sigma^k_i \sigma^k_j}$. This indicates that if one uses a multivariate Gaussian, the correlation is automatically taken care of by the model.

Our theoretical analysis is different from [2, 3] in several aspects. In [2, 3], two types of errors are introduced, namely Bayes (inherent) error and added error. The former is due to unbiased classifier whose class posterior estimates correspond to the true posteriors. The latter is due to biased classifiers which result in wrongly estimated class posteriors. The EER used here is commonly found in binary classification problems while the error (sum of bayes error and added error) applies to any number of classes. It is tempting to conclude that EER is equivalent to the Bayes error for a two-class problem. There are, however, important differences. In [2, 3] (the former), the bayes error is due to additive error in the feature space near the decision boundary. In EER (the latter), the input measurement is not a set of features but a set of scores of one or more base-classifiers. The output posteriors between the two classes in the former are enforced by linear approximation, whereas in the latter, they are assumed to be (integral of) Gaussian. The local continuity at the boundary is hence implicitly assumed. Furthermore, the Bayes error cannot be reduced (the added error can) but EER can [4].

4 Experiments with Synthetic Data

We designed a series of 110 synthetic experiment settings. Each experiment setting consists of a fusion task of *two* classifier outputs. All three assumptions mentioned in Section 3 are used here, i.e., (1) the 2D scores will be sampled from a multivariate Gaussian distribution for each class (client and impostor); (2) scores are comparable, i.e., the mean of client and impostor distributions are fixed to 0 and 1, respectively. However, for the product rule, scores are further normalised into the range $[0, 1]$ by the min-max normalisation; and finally, (3) the *same* covariance matrix is used for the client and impostor distributions.

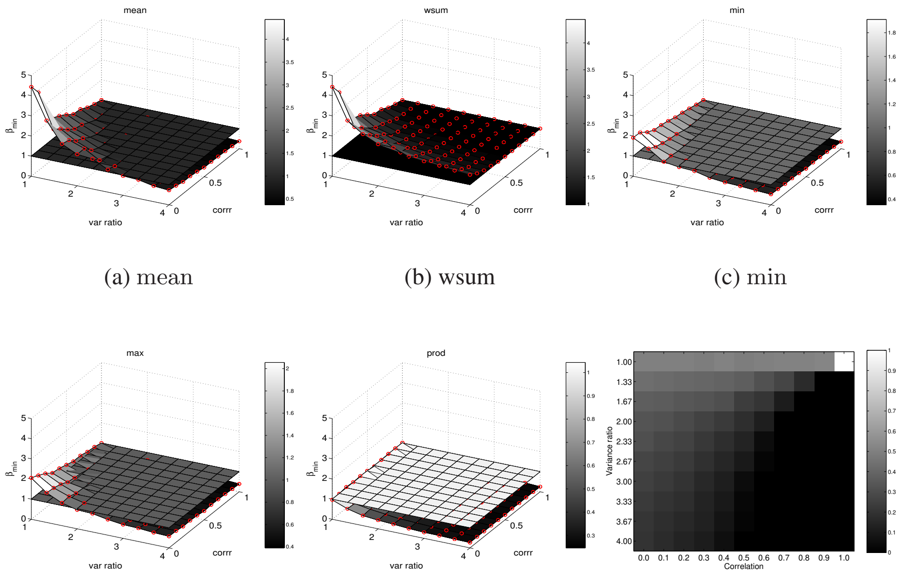
In order to evaluate classifier performance, Half Total Error Rate (HTER) is commonly used for biometric authentication. It is defined as: $\text{HTER} = \frac{1}{2}(\text{FAR}(\varDelta) + \text{FRR}(\varDelta))$, where the threshold $\varDelta$ is chosen to minimise the Weighted Error Rate (WER) at a given pre-defined $\alpha \in [0, 1]$ which balances between FAR and FRR. WER is defined as:

$$\text{WER}(\alpha) = \alpha \text{FAR}(\varDelta) + (1 - \alpha)\text{FRR}(\varDelta). \tag{10}$$

To optimise the EER criterion, instead of WER, $\alpha = 0.5$ is used. We further define a performance gain variable called β_{min}, as follows: $\beta_{min} = \frac{\text{HTER}_{best}}{\text{HTER}_{COM}}$ where COM is any one of the fusion classifiers/rules under study. When $\beta_{min} > 1$, the particular fusion classifier is better than the best underlying system.

The first classifier, designed as the *better* classifier of the two, has a (class-dependent) variance of 0.5 and is kept constant across all synthetic data sets, whereas the second classifier has a variance that varies with a ratio between 1 to 4 (or absolute variance value between 0.5 to 2). This causes the first expert to have a HTER between 5.3% and 6.2%, with a mean of 5.8% and the second expert between 5.4% and 22% of HTER with a mean of 15% at the EER point. Furthermore, the correlation value is varied between 0 and 1, at a step of 0.1 increment.

The simulation results are shown in Figure 1. For figures (a)-(e), the plane with $\beta_{min} = 1$ indicates the best single classifier, i.e., the baseline performance. As can be seen, the weighted sum classifier achieves the best overall gain. In fact, its $\beta_{min} > 1$ across all variance ratios and



(a) mean (b) wsum (c) min

(d) max (e) product (f) weights

Fig. 1. Performance gain of HTER, at EER criterion, with respect to the best underlying classifier, β_{min}, (the Z-axis) across different variance ratios (of two experts) from 1 to 4 (the X-axis) and different correlation values from 0 to 1 (the Y-axis), as a result of fusing synthetic scores of two expert systems (classifiers) assuming class-dependent scores are normally distributed. The scores are combined using (a) mean, (b) weighted sum, (c) min, (d) max and (e) product fusion classifiers. (f): the weight of the *weaker* expert found in the weighted sum after training. This can be thought of as the degree of "reliance on the weaker expert".

across all correlation values. The mean rule shows that the performance gain is more than 1 only when the variance ratio is 3 at correlation=0. As correlation increases, to maintain a positive gain, the variance ratio has to be decreased. This behaviour has been theoretically verified in [4]. The min and max rules follow the same trend as mean and weighted sum except that their gain is much smaller. There is no significant difference between the min and max rules. This is somewhat expected following their theoretical EER models presented in Table 1.

We further examined the weight attributed to the second (weaker) classifier by the weighted sum classifier to see how the weights evolve with various variance ratios and correlations. This weight can be interpreted as "reliance on the weaker system". This is shown in Figure 1(f). On this Cartesian coordinate system (X is correlation and Y is variance ratio), the point (1,1) implies that the two classifiers have exactly the same performance. Hence, the weight attributed to classifier 1 or 2 makes no difference. However, at exact correlation (=1), the weight attributed to classifier 2 (the weaker one) immediately becomes zero as the variance ratio increases. Furthermore, there is absolutely no improvement for this case (see in Figure 1(b)). On the other hand, at zero-correlation, the weaker classifier *contributes* to fusion (i.e., the weights are not zero). The corresponding performance gain *always increases* with decreasing variance ratio (increasingly stronger weak classifier). The product rule only has performance as good as the single-best classifier at variance ratio=1 while does not match the rest of the fusion classifiers. Its performance does not evolve with the correlation. One plausible explanation of such suboptimal performance comes from [9], stating that the the product rule is more sensitive to error as compared to the sum (or mean) rule. Despite their difference, all fusion classifiers except the product rule show that low correlation and low variance ratio increase the fusion performance. Note that no generalisation performance is involved here. In real applications, where there is a mismatch between training and test data sets, generalisation performance becomes an important concern. This is treated in the next section with real data.

5 Experiments with Real Data

5.1 Database Settings and Evaluation

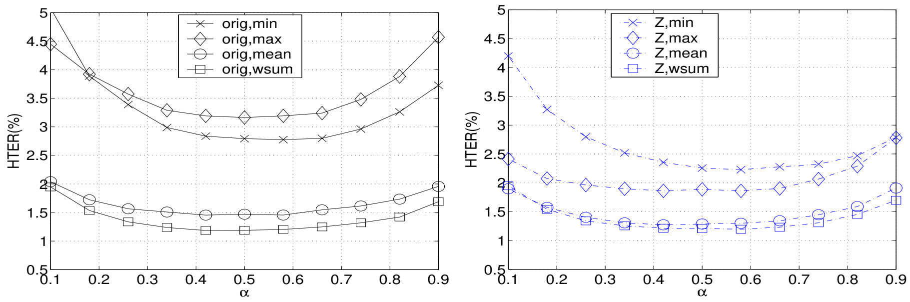
The publicly available[2] XM2VTS benchmark database for score-level fusion [10] is used. There are altogether 32 fusion data sets and each data set contains a fusion task of two experts. These fusion tasks contain multimodal and intramodal fusion based on face and speaker authentication tasks. For each data set, there are two sets of scores, from the *development* and the *evaluation* sets. The development set is used *uniquely* to train the fusion classifier parameters, including the threshold (bias) parameter, whereas the evaluation set is used uniquely to evaluate the generalisation performance. They are in accordance to the two originally defined Lausanne Protocols [11]. The 32 fusion experiments have 400 (client accesses) $\times$ 32 (data sets)= 12,800 client accesses and 111,800 (impostor accesses) $\times$ 32 (data sets) = 3,577,600 impostor accesses.

The most commonly used performance visualising tool in the literature is the Decision Error Trade-off (DET) curve. It has been pointed out [12] that two DET curves resulting from two systems are not comparable because such comparison does not take into account how the thresholds are selected. It was argued [12] that such threshold should be chosen *a priori* as well, based on a given criterion. This is because when a biometric system is operational, the threshold parameter has to be fixed *a priori*. As a result, the Expected Performance Curve (EPC) [12] was proposed. This curve is constructed as follows: for various values of α in Eqn. (10) between 0 and 1, select the optimal threshold Δ on a development (training) set, apply it on the evaluation (test) set and compute the HTER on the evaluation set. This HTER is then plotted with respect to α. The EPC curve can be interpreted similarly to the DET curve, i.e., the lower the curve, the better the generalisation performance. In this study, the *pooled* version of EPC is used to visualise the performance. The idea is to plot a single EPC curve instead of 32 EPC curves for each of the 32 fusion experiments. This is done by calculating the *global* false acceptance and false rejection errors over the 32 experiments for *each* of the α values. The pooled EPC curve and its implementation can be found in [10].

5.2 Experimental Results and Discussion

Figure 2 shows the pooled EPC curves of several fusion classifiers/rules under study, each over the 32 XM2VTS fusion data sets. As can be observed, the weighted sum gives the best generalisation performance. The mean rule follows closely. As expected, both $\min$ and $\max$ rules have improved generalisation performance *after* score-normalisation. For the normalised case (see figure (b)), $\max$ turns out to outperform $\min$ significantly for a large of α, according to HTER significance test at 90% of confidence [13].

[2] Accessible at http://www.idiap.ch/∼norman/fusion



(a) original (b) Z-Norm

Fig. 2. Pooled EPC curves, each derived from 32 fusion data sets, as a result of applying $\min$, $\max$, mean and weighted sum fusion, with (a) unnormalised orignal scores, (b) margin-transformed scores, (c) z-scores and (d) F-ratio transformed scores.

The weight parameters in the weighted sum are optimised using a 1D search procedure with a constant step-size of 0.05 within the bound $[0, 1]$ since only two classifier outputs are involved. This strategy has been employed by [14] for user-specific weighting. The advantage of this technique over the technique assuming zero-correlation, such as [3] or Fisher-ratio [7, Sec. 3.6] is that no assumption is made about the underlying class-dependent distribution. Support Vector machines with linear kernel could also have been used instead since it too does not make this assumption. We actually carried out the two control experiments using the two techniques mentioned and found that their generalisation performance are significantly inferior to our line search or SVM approach (not shown here). This is a probable reason why the empirical study conducted here is somewhat different from [15], where the authors did not find weighted sum significantly outperforms the mean rule, although the *same* database was used.

6 Conclusions

In this study, the theoretical and empirical aspects of fixed and trainable fusion classifiers are studied using the EER. Although this subject is well studied [1–3], the effects of correlation on Order Statistics (OS) combiners, e.g, $\min$, $\max$, and median, are largely unknown or rarely discussed due to intractable analysis. We studied the *joint effect* of correlation and base-classifier imbalance performance on EER by simulation. This simulation is based on three major assumptions: class-dependent Gaussianity assumption, score comparability assumption and class-dependent correlation assumption. Each assumption are adequately addressed (see Section 3). In particular, for the second assumption, several score normalisation techniques are discussed. Based on 4 fusion classifiers $\times$ 2 normalisation techniques (and) $\times$ 32 data sets = 256 fusion experiments, we show that weighted sum, when weights are tuned correctly, can achieve the best generalisation performance, with the additional advantage that no score normalisation is needed.

Acknowledgment

This work was supported in part by the IST Program of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, funded in part by the Swiss Federal Office for Education and Science (OFES) and the Swiss NSF through the NCCR on IM2. This publication only reflects the authors' view.

References

1. L.I. Kuncheva, "A Theoretical Study on Six Classifier Fusion Strategies," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24(2), pp. 281–286, February 2002.
2. K. Tumer and J. Ghosh, "Robust Combining of Disparate Classifiers through Order Statistics," *Pattern Analysis and Applications*, vol. 5, pp. 189–200, 2002.
3. G. Fumera and F. Roli, "Analysis of Linear and Order Statistics Combiners for Fusion of Imbalanced Classifiers," in *LNCS 2364, Proc. 3rd Int'l Workshop on Multiple Classifier Systems (MCS 2002)*, Cagliari, 2002, pp. 252–261.
4. N. Poh and S. Bengio, "How Do Correlation and Variance of Base Classifiers Affect Fusion in Biometric Authentication Tasks?," Research Report 04-18, IDIAP, Martigny, Switzerland, 2004, accepted for publication in *IEEE Trans. Signal Processing*, 2005.
5. 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.
6. A. Krogh and J. Vedelsby, "Neural Network Ensembles, Cross-Validation and Active-Learning," *Advances in Neural Information Processing Systems*, vol. 7, 1995.
7. C. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1999.
8. B.C. Arnold, N. Balakrishnan, and H.N. Nagaraja, *A First Course in Order Statistics*, Wiley, New York, 1992.
9. J. Kittler, M. Hatef, R. P.W. Duin, and J. Matas, "On Combining Classifiers," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.
10. N. Poh and S. Bengio, "Database, Protocol and Tools for Evaluating Score-Level Fusion Algorithms in Biometric Authentication," Research Report 04-44, IDIAP, Martigny, Switzerland, 2004, Accepted for publication in *AVBPA 2005*.
11. J. Matas, M. Hamouz, K. Jonsson, J. Kittler, Y. Li, C. Kotropoulos, A. Tefas, I. Pitas, T. Tan, H. Yan, F. Smeraldi, J. Begun, N. Capdevielle, W. Gerstner, S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Comparison of Face Verification Results on the XM2VTS Database," in *Proc. 15th Int'l Conf. Pattern Recognition*, Barcelona, 2000, vol. 4, pp. 858–863.
12. S. Bengio and J. Mariéthoz, "The Expected Performance Curve: a New Assessment Measure for Person Authentication," in *The Speaker and Language Recognition Workshop (Odyssey)*, Toledo, 2004, pp. 279–284.
13. S. Bengio and J. Mariéthoz, "A Statistical Significance Test for Person Authentication," in *The Speaker and Language Recognition Workshop (Odyssey)*, Toledo, 2004, pp. 237–244.
14. A. Jain, K. Nandakumar, and A. Ross, "Score Normalisation in Multimodal Biometric Systems," *Pattern Recognition (to appear)*, 2005.
15. F. Roli, G. Fumera, and J. Kittler, "Fixed and Trained Combiners for Fusion of Imbalanced Pattern Classifiers," in *Proc. 5th Int'l Conf. on Information Fusion*, 2002, pp. 278–284.