IDIAP RESEARCH REPORT

Why Do Multi-Stream, Multi-Band and Multi-Modal Approaches Work on Biometric User Authentication Tasks?

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

IDIAP–RR 03-59

November 2003

PUBLISHED IN
2004 Internation Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004), Montreal, vol. 5, pages 893-896, May 17-21, 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

IDIAP Research Report 03-59

Why Do Multi-Stream, Multi-Band and Multi-Modal Approaches Work on Biometric User Authentication Tasks?

Norman Poh Hoon Thian Samy Bengio

November 2003

published in
2004 Internation Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004), Montreal, vol. 5, pages 893-896, May 17-21, 2004.

Abstract. Multi-band, multi-stream and multi-modal approaches have proven to be very successful both in experiments and in real-life applications, among which speech recognition and biometric authentication are of particular interest here. However, there is a lack of a theoretical study to justify why and how they work, when one combines the streams at the feature or classifier score levels. In this paper, we attempt to cast a light onto the latter subject. Our findings suggest that combining several experts using the mean operator, Multi-Layer-Perceptrons and Support Vector Machines *always* perform better than the *average performance* of the underlying experts. Furthermore, in practice, *most* combined experts using the methods mentioned above perform better than *the best underlying expert*.

1 Introduction

Multi-band is a technique often used in speech recognition or speaker authentication that splits frequency into several subbands so that each subband will be processed separately by its corresponding classifier. The classifier scores are then merged by some combination mechanisms [1]. Multi-stream is a similar technique except that each stream uses different features. Multi-modal is yet another technique that is applied in Biometric Authentication, where each modality is a biometric trait associated to a person, such as face and speech. These approaches have proven to be very successful both in experiments and in real-life applications, e.g, [2, 1] for speech recognition and [3, 4, 5] for face and speaker authentication.

Unfortunately, there is a lack of a theoretical study to justify why and how they work, when one combines the streams at the feature or classifier score levels. The former is called feature combination while the latter is called posterior combination in [6]. In a separate study in biometric authentication (BA) [7], these two approaches are called Variance Reduction (VR) via extractors and VR via classifiers. The term variance reduction is originated from [8, Chap. 9], from the observation that when two classifier scores are merged by a simple mean operator, the *resultant variance* of the final score will be reduced with respect to the *average variance* of the two original scores.

To the authors opinion, theoretical justifications of these approaches have not been thoroughly investigated. In particular, (i) how does correlation in the classifier scores affect the combination mechanism, and (ii) how does this correlation affect the classification accuracy in terms of Equal Error Rate? These issues are the focus of this paper. In this study, the mean operator is used as a case study for studying these issues. In practice, non-linear trainable functions such as Multi-Layer Peceptrons and Support Vector Machines can also be used. Our findings suggest that the combined experts using the mean operator *always* perform better than the average of their participating experts. Furthermore, in practice, *most* combined experts, particulary those using non-linear trainable classifiers, perform better than *any* of their participating experts.

The rest of this paper is organised as follows: Section 2 studies variance reduction due to the mean operator and Section 3 shows its relation with classification error reduction. Section 4 discusses how non-linear combination mechanisms can be useful. Conclusions are in Section 5.

2 Variance Reduction

Let $\mathbf{x}$ be a *biometric measurement* that represents a person and $y_i(\mathbf{x})$ be the i-th measured relationship between the biometric trait $\mathbf{x}$ and the person. For example, i could denote the i-th subband of a spectrogram representing the speech of a person. i could also be the i-th stream for a given type of feature (e.g. Mel-scale Frequency Cepstrum Coefficients and Linear Predictive Cepstrum Coefficients). In the context of multi-modal biometrics, i could be the i-th biometric measurement (e.g., speech, face or fingerprint). In this context, i could be the i-th sample, i-th feature and even i-th classifier. $y_i(\mathbf{x})$ can be thought as the i-th response of the biometric measurement $\mathbf{x}$ given by an expert system. Typically, this output (e.g. score) is used to make the accept/reject decision. It can be defined as:

$$y_i(\mathbf{x}) = h(\mathbf{x}) + \eta_i(\mathbf{x}), \tag{1}$$

where $h(\mathbf{x})$ is the "target" function that one wishes to estimate and $\eta_i(\mathbf{x})$ is a random additive noise with zero mean, also dependent on $\mathbf{x}$. $h(\mathbf{x})$ can be viewed as the ideal function that consistently gives 1 when $\mathbf{x}$ corresponds to the client and -1 when corresponds to the impostor. The mean of y is

$$\bar{y}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} y_i(\mathbf{x}), \tag{2}$$

where there are N responses of streams, subbands or biometric modalities. The expected value of $y_i(\mathbf{x})$, denoted as $E[y_i(\mathbf{x})]$ is

$$E[y_i(\mathbf{x})] = E[h(\mathbf{x})] + E[\eta_i(\mathbf{x})] = h(\mathbf{x})$$

By using one hypothesis of y per access, the variance, by definition, is:

$$\begin{aligned} \mathrm{VAR}[y_i(\mathbf{x})] &= E[(y_i(\mathbf{x}) - E[y_i(\mathbf{x})])^2] \\ &= E[(y_i(\mathbf{x}) - h(\mathbf{x})])^2] = E[\eta_i(\mathbf{x})^2], \end{aligned} \tag{3}$$

When N hypotheses are available but are used separately, the *average of variance* of each hypothesis is:

$$\begin{aligned} \mathrm{VAR}_{AV}(\mathbf{x}) &= \frac{1}{N}\sum_{i=1}^{N} \mathrm{VAR}[y_i(\mathbf{x})] \\ &= \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}[\eta_i(\mathbf{x})^2], \end{aligned} \tag{4}$$

where Eqn. (3) is used. However, by combining N hypotheses per access via averaging, the *variance of average* is:

$$\begin{aligned} \mathrm{VAR}_{COM}(\mathbf{x}) &= E[(\bar{y}(\mathbf{x}) - h(\mathbf{x})])^2] \\ &= E\left[\left(\frac{1}{N}\sum_{i=1}^{N} y_i(\mathbf{x}) - h(\mathbf{x})\right)^2\right] \\ &= E\left[\left(\frac{1}{N}\sum_{i=1}^{N} \eta_i(\mathbf{x})\right)^2\right]. \end{aligned} \tag{5}$$

To expand Eqn. (5), one has to take into account the possible correlation among different $\eta_i(\mathbf{x})$ values which can be defined by:

$$\rho = \frac{E[\eta_i \eta_j]}{\sigma_i \sigma_j},$$

where σ_i and σ_j are the standard deviations of η_i and η_j. Note that with the introduction of ρ, Eqn. (3) can be written as: $\mathrm{VAR}[y_i(\mathbf{x})] = \rho E[\eta_i(\mathbf{x})^2] = E[\eta_i(\mathbf{x})^2] = \sigma_i^2$, since $\eta_i = \eta_j$ for $\forall_{i,j}$, and consequently $\rho = 1$ in this case. For clarity purposes, $\mathbf{x}$ will be dropped in all equations hereinafter. Note that this correlation equation has the property that $-1 \leq \rho \leq +1$. By taking into account the possible correlation in η_i, Eqn. (5) can be written as:

$$\begin{aligned} \mathrm{VAR}_{COM} &= E\left[\frac{1}{N^2}\left(\sum_{i=1}^{N}\sum_{j=1}^{N} \eta_i \eta_j\right)\right] \\ &= E\left[\frac{1}{N^2}\left(\sum_{i=1}^{N} \eta_i^2 + 2\sum_{i=1,i<j}^{N} \eta_i \eta_j\right)\right] \\ &= \frac{1}{N^2}\sum_{i=1}^{N} E[\eta_i^2] + \frac{2}{N^2}\sum_{i=1,i<j}^{N} E[\eta_i \eta_j] \\ &= \frac{1}{N^2}\sum_{i=1}^{N} \sigma_i^2 + \frac{2}{N^2}\sum_{i=1,i<j}^{N} \rho \sigma_i \sigma_j, \end{aligned} \tag{6}$$

since $E[\eta_i^2] = \sigma_i^2$. Now, we need to consider two cases: when η_i are independent from each other (i.e., $\rho = 0$) and when they are not (i.e., $\rho \neq 0$).

2.1 Assuming independence in η_i: $\rho = 0$

In this case, the right term in Eqn. (6) will be zero. In the same notation, Eqn. (4) can be rewritten as:

$$\mathrm{VAR}_{AV} = \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2, \tag{7}$$

Comparing Eqns. (6) and (7), it can be easily seen that:

$$\mathrm{VAR}_{COM} = \frac{1}{N}\mathrm{VAR}_{AV}, \tag{8}$$

which is true when η_i is uncorrelated. This is the lowest theoretical bound that VAR_{COM} can achieve. Basically, this shows that by averaging N scores, the *variance of average* (VAR_{COM}) can be reduced by a factor of N with respect to the *average of variance* (VAR_{AV}), when η_i is not correlated.

2.2 Assuming dependencies in η_i: $\rho \neq 0$

The upper bound can be derived from the second assumption that η_i is correlated, i.e. $\rho \neq 0$. This worst-case bound is in fact equal to VAR_{AV}, i.e., there is no gain. To be more explicit, we wish to test the hypothesis that $\mathrm{VAR}_{COM} \leq \mathrm{VAR}_{AV}$. By using Eqns. 6 and 7, this can be shown as follows:

$$\mathrm{VAR}_{COM} \leq \mathrm{VAR}_{AV}$$

$$\frac{1}{N^2}\sum_{i=1}^{N}\sigma_i^2 + \frac{2}{N^2}\sum_{i=1,i<j}^{N}\rho\sigma_i\sigma_j \leq \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2 \tag{9}$$

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

$$0 \leq (N-1)\sum_{i=1}^{N}\sigma_i^2 - 2\sum_{i=1,i<j}^{N}\rho\sigma_i\sigma_j.$$

Given that $(N-1)\sum_{i=1}^{N}\sigma_i^2 = \sum_{i=1,i<j}^{N}(\sigma_i^2+\sigma_j^2)$ (the proof can be found in the appendix), this inequality can further be simplified to:

$$\begin{aligned}
0 &\leq \sum_{i=1,i<j}^{N}(\sigma_i^2+\sigma_j^2) - 2\sum_{i=1,i<j}^{N}\rho\sigma_i\sigma_j \\
0 &\leq \sum_{i=1,i<j}^{N}\left(\sigma_i^2 - 2\rho\sigma_i\sigma_j + \sigma_j^2\right) \\
0 &\leq \sum_{i=1,i<j}^{N}\left((\sigma_i^2 - 2\rho\sigma_i\sigma_j + \rho^2\sigma_j^2 + (1-\rho^2)\sigma_j^2)\right) \\
0 &\leq \sum_{i=1,i<j}^{N}\left((\sigma_i - \rho\sigma_j)^2 + (1-\rho^2)\sigma_j^2\right).
\end{aligned} \tag{10}$$

In other words, hypothesis in Eqn. (9) is always true, whether η_i is correlated or not. As a consequence, we have just shown that $\mathrm{VAR}_{COM} \leq \mathrm{VAR}_{AV}$. Taking this conclusion and that of Eqn. (8), one can conclude that:

$$\frac{1}{N}\mathrm{VAR}_{AV} \leq \mathrm{VAR}_{COM} \leq \mathrm{VAR}_{AV}. \tag{11}$$

Referring back to Eqn. (6), if $\rho < 0$, i.e., η_i is negatively correlated, then the right hand term in this equation would be negative and consequently $\text{VAR}_{COM} \leq \frac{1}{N}\text{VAR}_{AV}$! Obviously, negative correlation would help improve the results. However, and unfortunately, in reality, negative correlation will not happen if the underlying experts are well-trained, i.e., for a given instant i, y_i for $i = 1, \ldots, N$, will tend to agree with each other (hence positively correlated) most often than to disagree with each other (hence negatively correlated).

2.3 Introduction of α as a gain factor

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

$$\alpha = \frac{\text{VAR}_{AV}(\mathbf{x})}{\text{VAR}_{COM}(\mathbf{x})}. \tag{12}$$

By dividing Eqn. (11) by VAR_{COM} and rearranging it, we can deduce that

$$1 \leq \alpha \leq N. \tag{13}$$

One direct implication of variance reduction is that **the more hypotheses used** (increasing N), **the better the combined system**, even if the hypotheses of underlying experts are correlated. This will come at a cost of more computation proportional to N. Experiments in [1] (in speech recognition) and [3] (in face verification) provide strong evidences to support this claim. Moreover, the gain α is often very small (near 1) compared to N [9].

3 Variance Reduction and EER Reduction

Until now, it is not clear how variance reduction can lead to better classification, in terms of false rejection rate (FRR) and false acceptance rate (FAR) in a biometric authentication system. Figure 1 illustrates the effect of averaging scores in a two-class problem, such as in BA where an identity claim could belong either to a client or an impostor. Let us assume that the genuine user scores in a situation where 3 samples are available but are used separately, follow a normal distribution of mean 1.0 and variance ($\text{VAR}_{AV}(\mathbf{x})$ of genuine users) 0.9, denoted as $\mathcal{N}(1, 0.9)$, and that the impostor scores (in the mentioned situation) follow a normal distribution of $\mathcal{N}(-1, 0.6)$ (both graphs are plotted with "+"). If for each access, the 3 scores are used, according to Eqn. (13), the variance of the resulting distribution will be reduced by a factor of 3 or less. Both resulting distributions are plotted with "o". Note the area where both the distributions overlap before and after. The latter area is shaded in Figure 1. This area corresponds to the zone where minimum amount of mistakes will be committed given that the threshold is optimal[1]. Decreasing this area implies an improvement in the performance of the system.

Let the scores' probability density function (*pdf*) be $P(y|\mathbf{x} \in \mathbf{x}_C)$ for the client set C and $P(y|\mathbf{x} \in \mathbf{x}_I)$ similarly for the impostor set I. Let us first assume that these *pdf*s are Gaussians. FRR and FAR can then be defined as:

$$\begin{aligned} \text{FRR}(\theta) &= \int_{-\infty}^{\theta} P(y|\mathbf{x} \in \mathbf{x}_C)dy \\ &= \int_{-\infty}^{\theta} \frac{1}{\sigma_C\sqrt{2\pi}} \exp\left[\frac{-(y-\mu_C)^2}{2\sigma_C^2}\right] dy \\ &= \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{\theta - \mu_C}{\sigma_C\sqrt{2}}\right), \text{ and} \end{aligned} \tag{14}$$

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

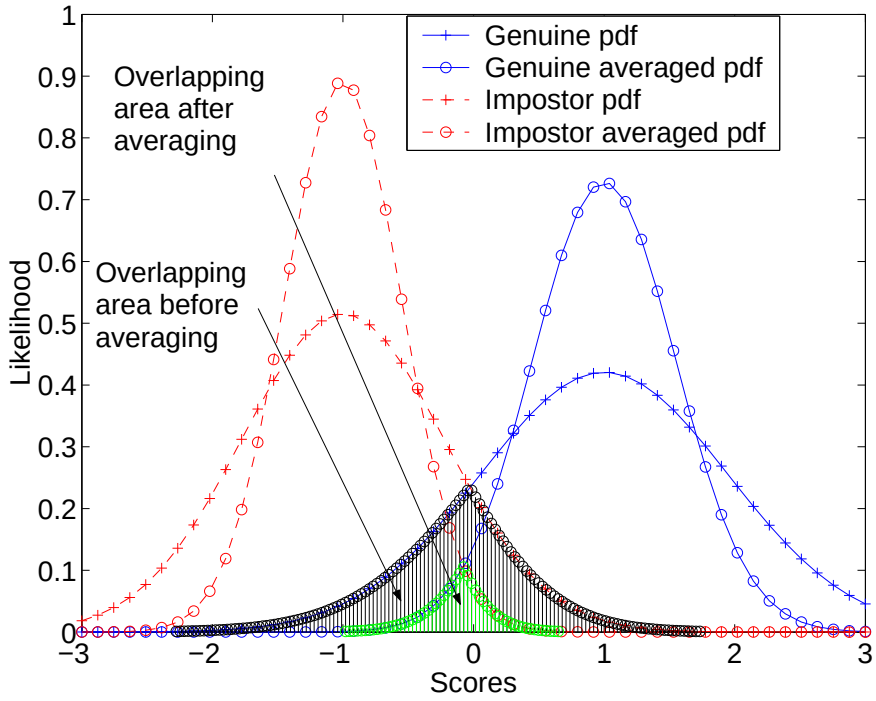

Figure 1: Averaging score distributions in a two-class problem

$$\begin{aligned}
\mathrm{FAR}(\theta) &= \int_{\theta}^{\infty} P(y|\mathbf{x} \in \mathbf{x}_I)dy \\
&= 1 - \int_{-\infty}^{\theta} P(y|\mathbf{x} \in \mathbf{x}_I)dy \\
&= 1 - \left[\frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{\theta - \mu_I}{\sigma_I\sqrt{2}}\right)\right] \\
&= \frac{1}{2} - \frac{1}{2}\mathrm{erf}\left(\frac{\theta - \mu_I}{\sigma_I\sqrt{2}}\right), \qquad (15)
\end{aligned}$$

where

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

which is the so-called error function. μ_C and σ_C are the expected value and the standard deviation of scores belonging to the client set C and similarly μ_I and σ_I for the impostor set I. Note that the use of an error function for such analysis has been reported in [10], but with differences in the definition of the error function. In another similar work (but limited to the context of combining multiple samples) [3], the Equal Error Rate (EER) curve was not calculated explicitly and validated via experiments as done here. Furthermore, the issue on how the dependency among samples affects the resultant variance was not studied theoretically as done in Section 2.

The minimal error happens when $\mathrm{FAR}(\theta) = \mathrm{FRR}(\theta) = \mathrm{EER}$, i.e., the Equal Error Rate. Making these two terms equal (Eqns (14) and (15)) and using the property that $\mathrm{erf}(-z) = -\mathrm{erf}(z)$, we can deduce that:

$$\theta = \frac{\mu_I\sigma_C + \mu_C\sigma_I}{\sigma_I + \sigma_C}. \qquad (16)$$

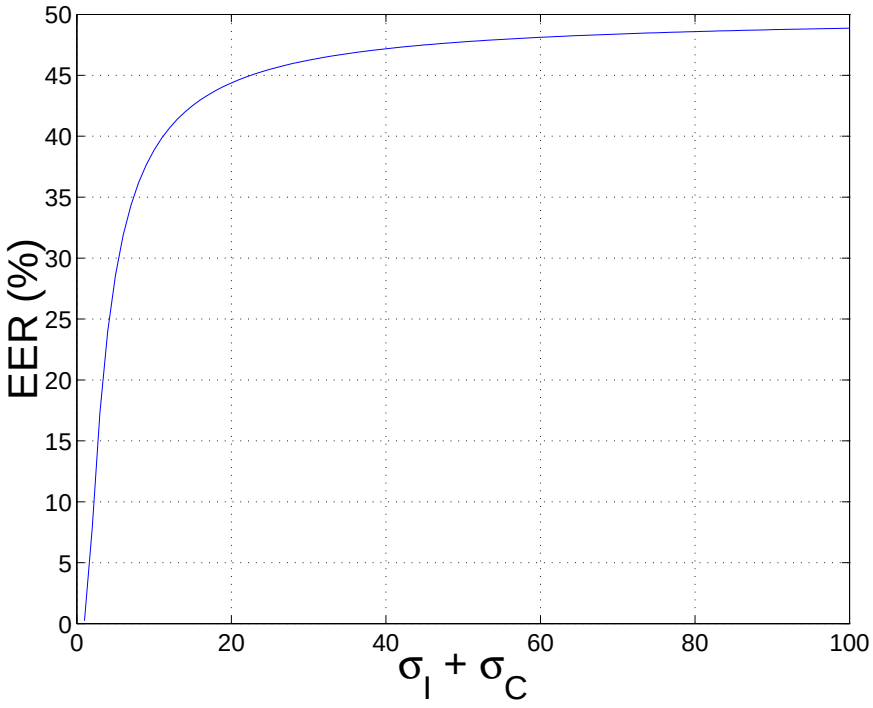

Figure 2: Equal error rate versus the sum of standard deviations of client and impostor scores

By introducing Eqn. (16) into Eqn. (15) (or equivalently into Eqn. (14)), we obtain:

$$\text{EER} = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{\mu_I - \mu_C}{(\sigma_C + \sigma_I)\sqrt{2}}\right). \tag{17}$$

To check the validity of Eqn. (17), we actually compared this theoretical EER with the empirical EER, calculated by using the optimal threshold:

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

and approximated by the commonly used Half Total Error Rate:

$$\text{HTER} = (\text{FAR}(\theta^*) + \text{FRR}(\theta^*))/2.$$

The difference between the theoretical EER and HTER is actually very small, as shown in Figure 3. This difference is due to the fact that the client and impostor distributions are not truly Gaussian. On the other hand, it also reveals that the Gaussian assumption is acceptable in practice.

Assuming that $\mu_C = 1$ and $\mu_I = -1$, we plot the graph EER by varying the term $\sigma_I + \sigma_C$ in Figure 2. EER is therefore a monotonically increasing function as $\sigma_I + \sigma_C$ increases.

Let σ'_I and σ'_C be the new σ_I and σ_C due to variance reduction for the impostor and client set, respectively. Using the annotations in Section 2, $\sigma' = \sqrt{\text{VAR}_{COM}}$ and $\sigma = \sqrt{\text{VAR}_{AV}}$. These definitions apply for both the client and impostor distributions. From Eqn. (11), we can deduce that:

$$\sigma'_I \leq \sigma_I \text{ and } \sigma'_C \leq \sigma_C.$$

Since EER is a monotonically increasing function as shown in Figure 2, these inequalities imply that:

$$\text{EER}(\sigma'_I, \sigma'_C) \leq \text{EER}(\sigma_I, \sigma_C),$$

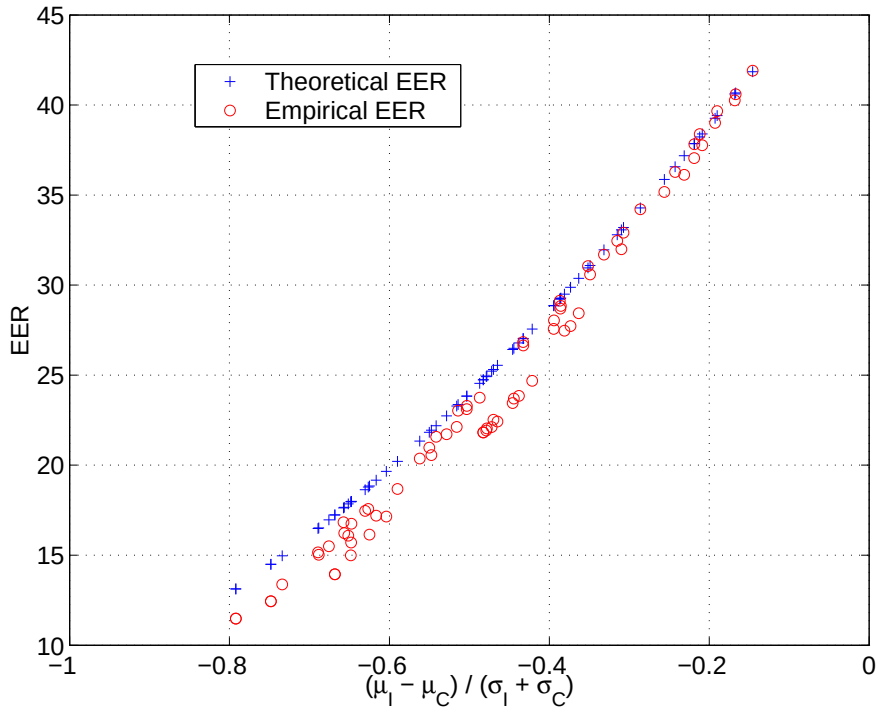

Figure 3: The theoretical and empirical EER as a function of ratio $(\mu_I - \mu_C)/(\sigma_I + \sigma_C)$, carried out on 72 independent experiments on the NIST2001 database with HTER ranging from 10% to 45%

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

In fact, without assuming the Gaussian distribution, as long as the EER function has a monotonically increasing behaviour with respect to $\sigma_I + \sigma_C$, the above conclusions remain valid. To require that EER be a monotonically increasing function, the necessary condition is that the right tail of the impostor *pdf* is a decreasing function and the left tail of the client *pdf* is an increasing function. A Gaussian function exhibits such behaviour on its left and right tails. Unfortunately, in the case of non-Gaussian *pdf*s, the analytical analysis such as the one done here is more difficult.

To evaluate the improvement due to variance reduction, we can define a gain factor β, similar to α defined in Eqn. (12), as follows:

$$\beta_{mean} = \frac{\text{mean}_i(\text{EER}_i)}{\text{EER}_{COM}} \tag{18}$$

where EER_{COM} is the EER of the combined system (with reduced variance) and EER_i is the EER of the i-th system. In our previous work [9] in the context of biometric authentication, *all experiments* verified that $\beta_{mean} \geq 1$, which is theoretically achievable. β_{mean} can only measure the relative improvement with respect to the average EER of the underlying expert. In practice, one wishes to know whether the resultant combined expert is better than the best underlying expert. This can be measured using:

$$\beta_{min} = \frac{\min_i(\text{EER}_i)}{\text{EER}_{COM}}, \tag{19}$$

which is defined very similarly to β_{mean}, except that the minimum EER of the underlying experts is used. $\beta_{min} \geq 1$ implies that the resultant expert is better than the best underlying expert. In fact, for both β_{mean} and β_{min}, $(\beta^{-1} - 1) \times 100\%$ measures the relative reduction of the combined expert with respect to the EER of the mean or the mininum EER's of the underlying experts.

4 Non-linear Combination Strategies

The analysis in the previous section is indeed based on the combination using the mean operator, which is a special case of a weighted sum with equal weights. One can also use non-linear methods such as Multi-Layer Perceptrons (MLPs) and Support Vector Machines (SVMs).

It is obvious that by using higher capacity (flexibility of a classifier to represent the underlying function), variance can be further reduced on the training set; see for instance [8, Chap 9] which demonstrated that a weighted sum reduces more variance than the mean operator. However, it is less obvious how this variance is reduced on *unseen* data. Hence, using an empirical procedure such as cross-validation to find the suitable capacity is of pivotal importance [7]. In this work, non-linear combination mechanisms such as MLPs and SVMs are superior over the average operator *most* of the time. Furthermore, the higher the independence of the underlying experts, the greater the β values. In this study, based on the XM2VTS database, combining face and speech experts can yield β_{mean} as high as 5.56 (and β_{min} as high as 3.10), whereas combining experts due to different features of the same modalities yields β_{mean} as high as 1.84 (and β_{min} as high as 1.12). Finally, diversity due to classifiers (therefore same features) yields β_{mean} as high as 2.05 (and β_{min} as high as 1.22). All these experiments show that $\beta_{mean} \geq 1$ and non-linear combination mechanisms, such as MLP and SVM, are *often* (there are exceptions) better than the mean operator, i.e., β_{min} of MLP and SVM $\geq \beta_{min}$ of the mean operator.

5 Conclusions

This study contributes to fusion field in several aspects. Firstly, it clarifies the intuition that independence of streams, subbands or modalities (as observed in each individual expert hypothesis/score) is crucial in determining the success of posterior combination. This is explained by variance reduction due to the combination. Secondly, variance reduction can be derived in many ways, other than streams, bands (both are considered features) and modalities: samples, virtual samples and classifiers [7, 9]. Thirdly, analytical analysis shows that the more hypotheses that are available the more robust the system will be. This is confirmed by experiments as reported in [1]. Finally, the successful use of non-linear techniques in combining scores really depends on the correct estimate of the underlying hyperparameters using techniques such as cross-validation, as supported by evidences in [7]. Although the study here concerns only classification of two-class problems, extending the analysis to N-class problems is straightforward, e.g., by using one-against-all encoding scheme. This theoretical study is certainly limited in scope as it does not provide a means to predict the best combination out of N streams/bands/modalities.

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

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

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

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

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

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

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

[2] S. Dupont, H. Bourlard, and C. Ris, "Robust Speech Recognition Based on Multi-Stream Features," in *Proc. ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, April 1997, pp. 95–98, IDIAP-RR 97-01.

[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] S. Bengio, C. Marcel, S. Marcel, and J. Mariéthoz, "Confidence Measures for Multimodal Identity Verification," *Information Fusion*, vol. 3, no. 4, pp. 267–276, 2002.

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

[6] D. Ellis, "Improved Recognition by Combining Different Features and Different Systems," in *AVIOS Speech Developers Conference and Expo*, San Jose, California, USA, May 2000.

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

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

[9] N. Poh and S. Bengio, "Variance Reduction Techniques in Biometric Authentication," IDIAP, Martigny, Switzerland, Researh Report 03-17, 2003.

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