WHY DO MULTI-STREAM, MULTI-BAND AND MULTI-MODAL APPROACHES WORK ON BIOMETRIC USER AUTHENTICATION TASKS?

Norman Poh and Samy Bengio

IDIAP, CP 592, 1920 Martigny, Switzerland

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. While there exists literature discussing this aspect, a study on the relationship between correlation, variance reduction and Equal Error Rate (often used in biometric authentication) has not been treated theoretically as done here, using the mean operator. 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 feature sets. There are very few literature reported the use of multi-stream for speaker authentication although both applications use conceptually similar techniques. One such example can be found in [2]. Multi-modal is yet another technique that is applied in Biometric Authentication (BA), 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, [1, 3] for speech recognition and [4–6] 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 [7]. In a separate study in BA [8], these two approaches are called Variance Reduction (VR) via extractors and VR via classifiers. The term variance reduction is originated from [9, 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. Pankanti et al [10] shaded some lights on this subject using AND and OR operator. Unfortunately, their proof requires the assumption that the scores due to the underlying experts are independent (not correlated), which is often not true when the underlying experts receive the *same* biometric data. Sanchez et al [4] showed both theoretically and empirically that fusing multiple instances of biometric traits can indeed reduce the system error by as much as 40%. The theoretically analysis, unfortunately, again did not deal with the case when the expert opinions are correlated.

The authors 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)" and Le Quan, Jean-Marc Adobez, Conrad Sanderson and the anonymous reviewers for constructing comments.

Hence, the central issues examined here are: (i) how correlation in the classifier scores affects the combination mechanism, and (ii) how this correlation affects the classification accuracy in terms of Equal Error Rate (EER; although there exists other variant of criterion such as decision cost function, EER is a well-accepted criterion evaluation and is very often used in the literature) . It should be underlined that there exists many applications of fusion in BA but the theoretical aspect of fusion, particularly dealing with correlation and in terms of Equal Error Rate, has not been treated elsewhere in the literature. In this study, the mean operator is used as a case study for studying these issues because it can be interpreted theoretically. In practice, non-linear trainable functions such as Multi-Layer Peceptrons and Support Vector Machines can also be used but their analysis requires more efforts than done here. 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, $y^j(\mathbf{x})$ be the j-th measured relationship between the biometric trait $\mathbf{x}$ and the person of a *single access*, and there are N such measurements per access, i.e., $j = 1, \ldots, N$. For example, j could denote the j-th subband of a spectrogram representing the speech of a person, the j-th stream or type of feature (e.g. Mel-scale Frequency Cepstrum Coefficients), the j-th biometric modality (e.g., speech, face or fingerprint), the j-th sample, the j-th classifier (but for the *same* access). In this context, $y^j(\mathbf{x})$ is referred to as an instance of the j-th *response* of the biometric measurement $\mathbf{x}$ given by an expert system (often called a score in the literature). Typically, this output (e.g. score) is compared with a predefined threshold to make the accept/reject decision. Let $h(\mathbf{x})$ to be a deterministic function or an ideal function that consistently gives $+1$ when $\mathbf{x}$ corresponds to the client and -1 when it corresponds to the impostor. Then we can write the mapping function of each response as the summation between the desired function and an error $w^j(\mathbf{x})$:

$$y^j(\mathbf{x}) = h(\mathbf{x}) + w^j(\mathbf{x}). \quad (1)$$

Note that the error term $w^j(\mathbf{x})$ follows an unknown distribution $W^j(\mathbf{x})$ with zero mean. Since $w^j(\mathbf{x})$ is dependent on $\mathbf{x}$, it is obvious that $y^j(\mathbf{x})$, which follows the distribution $Y^j(\mathbf{x})$, is also dependent on $\mathbf{x}$. Dropping $\mathbf{x}$ for clarity (since it is *present* in every term discussed), one can write the expectation of Y^j, $E[Y^j]$, as:

$$E[Y^j] = E[h] + E[W^j] = h, \quad (2)$$

Assuming that Y^j and Y^k can be correlated, the covariance between them can be written as follows:

$$\begin{aligned} \mathrm{Cov}(Y^j, Y^k) &= E\left[(Y^j - E[Y^j])(Y^k - E[Y^k])\right] \\ &= E\left[(Y^j - h)(Y^k - h)\right] = E[W^j W^k]. \quad (3) \end{aligned}$$

where Eqns. (1) and (2) are used. We would like to compare the variance of two cases: (i) N responses are available per access and they are used separately, and (ii) all N responses are used together. For the first case, this variance is called the *average of variance* over all N, and is denoted as σ^2_{AV}. It can be calculated as follows:

$$\sigma^2_{AV} = \frac{1}{N}\sum_{j=1}^{N}\mathrm{Cov}(Y^j,Y^j) = \frac{1}{N}\sum_{j=1}^{N}E[W^jW^j], \quad (4)$$

where Eqn. (3) is used. To calculate the variance of the second case, one needs to determine how the responses are combined. One easy way to combine them is to use the mean operator (more complicated, linear (e.g. [9, Sec. 9.6]) and non-linear functions can also be used but the use of mean operator is particularly useful for this discussion). The resultant averaged response, denoted as $\bar{Y}$, is defined as follow:

$$\bar{Y} \equiv \frac{1}{N}\sum_{j=1}^{N}Y^j, \quad (5)$$

Note that according to this definition, $E[\bar{Y}] = \frac{1}{N}\sum_j^N E[Y^j] = h$. The variance of $\bar{Y}$ (over many accesses), denoted as σ^2_{COM}, is called the *variance of average*, and can be calculated as follows:

$$\begin{aligned}
\sigma^2_{COM} &= \mathrm{Cov}(\bar{Y},\bar{Y}) = E\left[(\bar{Y}-E[\bar{Y}])(\bar{Y}-E[\bar{Y}])\right] \\
&= E\left[(\bar{Y}-h)(\bar{Y}-h)\right] \\
&= E\left[\left(\frac{1}{N}\sum_{m=1}^{N}Y^m - h\right)\left(\frac{1}{N}\sum_{n=1}^{N}Y^n - h\right)\right] \\
&= E\left[\left(\frac{1}{N}\sum_{m=1}^{N}(Y^m - h)\right)\left(\frac{1}{N}\sum_{n=1}^{N}(Y^n - h)\right)\right] \\
&= E\left[\left(\frac{1}{N}\sum_{m=1}^{N}W^m\right)\left(\frac{1}{N}\sum_{n=1}^{N}W^n\right)\right] \\
&= E\left[\frac{1}{N^2}\left(\sum_{i=m}^{N}\sum_{n=1}^{N}W^mW^n\right)\right], \quad (6)
\end{aligned}$$

where Eqns. (2) and (5) are used. The index m and n are introduced to take into account the possible covariance of error among different W^m and W^n. Before expanding Eqn. (6) further, let us define the correlation among different W^m and W^n as follows:

$$\rho = \frac{E[W^mW^n]}{\sigma_m\sigma_n}, \quad (7)$$

where σ_m and σ_n are the standard deviations of W^m and W^n. Note that correlation has the property that $-1 \le \rho \le +1$. Going back to Eqn. (6), we have:

$$\begin{aligned}
\sigma^2_{COM} &= E\left[\frac{1}{N^2}\left(\sum_{j=1}^{N}W^jW^j + 2\sum_{m=1,m<n}^{N}W^mW^n\right)\right] \\
&= \frac{1}{N^2}\sum_{j=1}^{N}E[W^jW^j] + \frac{2}{N^2}\sum_{m=1,m<n}^{N}E[W^mW^n] \\
&= \frac{1}{N^2}\sum_{j=1}^{N}\sigma_j^2 + \frac{2}{N^2}\sum_{m=1,m<n}^{N}\rho\sigma_m\sigma_n, \quad (8)
\end{aligned}$$

since $E[W^jW^j] = \sigma_j^2$ by definition. Now, we need to consider two cases: when W_m and W_n are independent from each other (i.e., $\rho = 0$) and when they are not (i.e., $\rho \neq 0$).

2.1. Independence Assumption: $\rho = 0$

In this case, $E[W^mW^n] = 0$, hence $\rho = 0$. As a consequence, the right term in Eqn. (8) will be zero. In the same notation, Eqn. (4) can be rewritten as:

$$\sigma^2_{AV} = \frac{1}{N}\sum_{j=1}^{N}\sigma_j^2, \quad (9)$$

Comparing Eqns. (8) and (9), it can be easily seen that:

$$\sigma^2_{COM} = \frac{1}{N}\sigma^2_{AV}, \quad (10)$$

which is true when W^m and W^n are not correlated. This is the lowest theoretical bound that σ^2_{COM} can achieve. Basically, this shows that by averaging N scores, the *variance of average* (σ^2_{COM}) can be reduced by a factor of N with respect to the *average of variance* (σ^2_{AV}), when two instances of Y^m and Y^n are not correlated.

2.2. Dependence Assumption: $\rho \neq 0$

The upper bound can be derived from the second assumption that W_m and W_n are correlated, i.e. $\rho \neq 0$. This worst-case bound is in fact equal to σ^2_{AV}, i.e., there is no gain. To be more explicit, we wish to test the hypothesis that $\sigma^2_{COM} \le \sigma^2_{AV}$. By using Eqns. (8) and (9), this can be shown as follows:

$$\begin{aligned}
\sigma^2_{COM} &\le \sigma^2_{AV} \\
\frac{1}{N^2}\sum_{j=1}^{N}\sigma_j^2 + \frac{2}{N^2}\sum_{m=1,m<n}^{N}\rho\sigma_m\sigma_n &\le \frac{1}{N}\sum_{j=1}^{N}\sigma_j^2 \quad (11)
\end{aligned}$$

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

$$0 \le (N-1)\sum_{j=1}^{N}\sigma_j^2 - 2\sum_{m=1,m<n}^{N}\rho\sigma_m\sigma_n.$$

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:

$$\begin{aligned}
0 &\le \sum_{m=1,m<n}^{N}(\sigma_m^2+\sigma_n^2) - 2\sum_{m=1,m<n}^{N}\rho\sigma_m\sigma_n \\
0 &\le \sum_{m=1,m<n}^{N}\left(\sigma_m^2 - 2\rho\sigma_m\sigma_n + \sigma_n^2\right) \\
0 &\le \sum_{m=1,m<n}^{N}\left((\sigma_m^2 - 2\rho\sigma_m\sigma_n + \rho^2\sigma_n^2 + (1-\rho^2)\sigma_n^2)\right) \\
0 &\le \sum_{m=1,m<n}^{N}\left((\sigma_m - \rho\sigma_n)^2 + (1-\rho^2)\sigma_n^2\right). \quad (12)
\end{aligned}$$

In other words, hypothesis in Eqn. (11) is always true, regardless of the value ρ. As a consequence, we have just shown that $\sigma^2_{COM} \le \sigma^2_{AV}$. Taking this conclusion and that of Eqn. (10), one can conclude that:

$$\frac{1}{N}\sigma^2_{AV} \le \sigma^2_{COM} \le \sigma^2_{AV}. \quad (13)$$

Referring back to Eqn. (8), if $\rho < 0$, i.e., W_i is negatively correlated, then the right hand term in this equation would be negative and consequently $\sigma^2_{COM} \le \frac{1}{N}\sigma^2_{AV}$! Obviously, negative correlation would help improve the results. However, and unfortunately, in reality, negative correlation will not happen if the underlying experts are trained separately, i.e., for a given 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). One possible exception will be that the experts are specifically trained to be decorrelated or even negatively correlated *in a collaborative way*. By fusing scores obtained from experts that are trained independently (which is often so in multimodal fusion), one can almost be certain that $0 \le \rho \le 1$.

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:

$$\alpha = \frac{\sigma^2_{AV}}{\sigma^2_{COM}}. \quad (14)$$

By dividing Eqn. (13) by σ^2_{COM} and rearranging it, we can deduce that

$$1 \le \alpha \le N. \quad (15)$$

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 [4] (in face verification) provide strong evidences to support this claim. Moreover, the gain (measured using β which is non-linearly but monotonically proportional to α, as defined in Section 3) is often very small (near 1) compared to N [8].

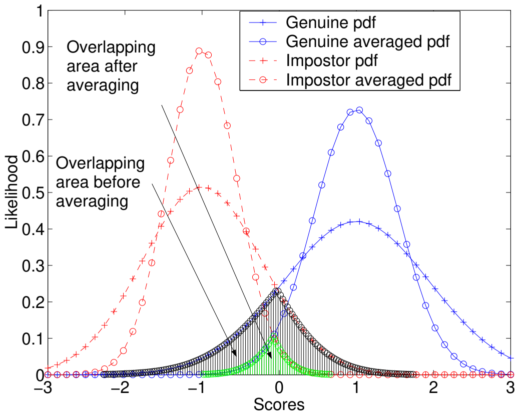

Fig. 1. Averaging score distributions in a two-class problem

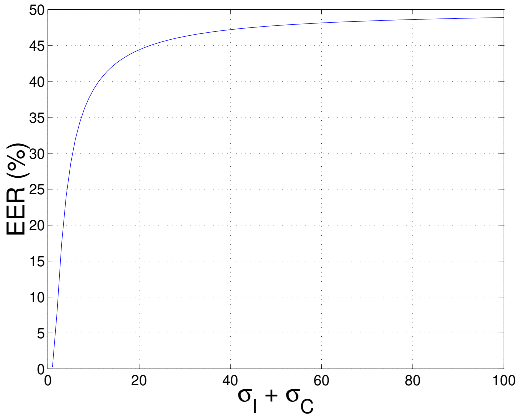

Fig. 2. Equal error rate versus the sum of standard deviations of client and impostor scores

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 ($\sigma^2_{AV}(\mathbf{x})$ 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. (15), 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:

$$\begin{aligned} \mathrm{FRR}(\theta) &= \int_{-\infty}^{\theta} P(y|\mathbf{x} \in \mathbf{x}_C)dy \\ &= \int_{-\infty}^{\theta} \frac{1}{\sigma_C\sqrt{2\pi}} \exp\left[\frac{-(y-\mu_C)^2}{2\sigma_C^2}\right] dy \\ &= \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{\theta-\mu_C}{\sigma_C\sqrt{2}}\right), \text{ and} \end{aligned} \quad (16)$$

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

$$\begin{aligned} \mathrm{FAR}(\theta) &= \int_{\theta}^{\infty} P(y|\mathbf{x} \in \mathbf{x}_I)dy \\ &= 1 - \int_{-\infty}^{\theta} P(y|\mathbf{x} \in \mathbf{x}_I)dy \\ &= 1 - \left[\frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{\theta-\mu_I}{\sigma_I\sqrt{2}}\right)\right] \\ &= \frac{1}{2} - \frac{1}{2}\mathrm{erf}\left(\frac{\theta-\mu_I}{\sigma_I\sqrt{2}}\right), \end{aligned} \quad (17)$$

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 [11], but with differences in the definition of the error function. In another similar work (but limited to the context of combining multiple samples) [4], 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 FAR(θ) = FRR(θ) = EER, i.e., the Equal Error Rate. Making these two terms equal (Eqns (16) and (17)) and using the property that $\mathrm{erf}(-z) = -\mathrm{erf}(z)$, we can deduce that:

$$\theta = \frac{\mu_I\sigma_C + \mu_C\sigma_I}{\sigma_I + \sigma_C}. \quad (18)$$

By introducing Eqn. (18) into Eqn. (17) (or equivalently into Eqn. (16)), we obtain:

$$\mathrm{EER} = \frac{1}{2} - \frac{1}{2}\mathrm{erf}\left(\frac{\mu_C - \mu_I}{(\sigma_C + \sigma_I)\sqrt{2}}\right). \quad (19)$$

To check the validity of Eqn. (19), we actually compared this theoretical EER with the empirical EER, calculated by using the optimal threshold:

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

and approximated by the commonly used Half Total Error Rate:

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

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.

Using the notation in Section 2, let σ^I_{COM} and σ^C_{COM} be the standard deviations of the fused scores (using the mean operator) of both the impostor and client distributions, respectively. These definitions also apply the average of the standard deviations σ^I_{AV} and σ^C_{AV}. From Eqn. (13), we can deduce that:

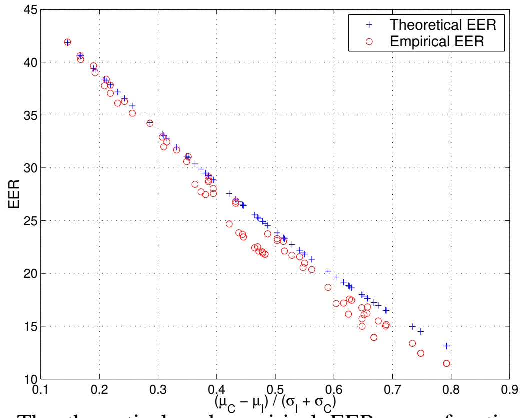

Fig. 3. The theoretical and empirical EER as a function of ratio $(\mu_I - \mu_C)/(\sigma_I + \sigma_C)$, carried out on 72 independent experiments on the NIST2001 database with HTER ranging from 10% to 45%

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

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

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

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

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. (14), as follows:

$$\beta_{mean} = \frac{\text{mean}_i(\text{EER}_i)}{\text{EER}_{COM}} \tag{20}$$

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 [8] 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:

$$\beta_{min} = \frac{\text{min}_i(\text{EER}_i)}{\text{EER}_{COM}}, \tag{21}$$

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 [9, 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 [8]. 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. In the case when they are dependent, fusion will also lead to improved results but the gain will be smaller. 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 [8]. 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 [8]. 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, $(N-1)\sum_{i=1}^N \sigma_i^2 = \sum_{i=1,i<j}^N(\sigma_i^2 + \sigma_j^2)$. $\square$

7. 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] N. Poh and S. Bengio, "Noise-robust multi-stream fusion for text-independent speaker authentication," Martigny, Switzerland, IDIAP-RR 04-01, 2004.

[3] 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.

[4] 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.

[5] 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.

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

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

[8] 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.

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

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

[11] 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.