IDIAP RESEARCH REPORT

Improving Single Modal and Multimodal Biometric Authentication Using F-ratio Client-Dependent Normalisation

Norman Poh [a] Samy Bengio [a]

IDIAP–RR 04-52

September 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 04-52

Improving Single Modal and Multimodal Biometric Authentication Using F-ratio Client-Dependent Normalisation

Norman Poh Samy Bengio

September 2004

Abstract. This study investigates a new *client-dependent normalisation* to improve a single biometric authentication system, as well as its effects on fusion. There exists two families of client-dependent normalisation techniques, often applied to speaker authentication. They are client-dependent score and threshold normalisation techniques. Examples of the former family of techniques are Z-Norm, D-Norm and T-Norm. There is also a vast amount of literature on the latter family of techniques. Both families are surveyed in this study. Furthermore, we also provide a link between these two families of techniques and show that one is a dual representation of the other. These techniques are intended to adjust the variation across different client models. We propose "F-ratio" normalisation, or F-Norm, applied to face and speaker authentication systems in two contexts: single modal and fusion of multi-modal biometerics. This normalisation requires that only *as few as* two client-dependent accesses are available (the more the better). Different from previous normalisation techniques, F-Norm considers the client and impostor distributions *simultaneously*. We show that F-ratio is a natural choice because it is directly associated to Equal Error Rate. It has the effect of centering the client and impostor distributions such that a global threshold can be easily found. Another difference is that F-Norm actually "interpolates" between client-independent and client-dependent information by introducing two mixture parameters. These parameters *can be optimised* to maximise the class dispersion (the degree of separability between client and impostor distributions) while the aforementioned normalisation techniques cannot. The results of 13 single modal experiments and 32 fusion experiments carried out on the XM2VTS multimodal database show that in both contexts, F-Norm is advantageous over Z-Norm, client-dependent score normalisation with EER and no normalisation.

1 Introduction

Biometric authentication (BA) is a process of verifying an identity claim using a person's behavioral and physiological characteristics. BA is becoming an important alternative to traditional authentication methods such as keys ("something one has", i.e., by possession) or PIN numbers ("something one knows", i.e., by knowledge) because it is essentially "who one is", i.e., by biometric information. Therefore, it is not susceptible to misplacement or forgetfulness. However, today, biometric-based security systems (devices, algorithms, architectures) still have room for improvement, particularly in their accuracy, tolerance to various noisy environments and scalability as the number of individuals increases.

In this paper, we study the effect of client-dependent variations and show how client-dependent normalisation techniques can be used to improve the overall system accuracy. Examples of work in this direction are client-dependent threshold [20], client-dependent score normalisation [8], and different weighing of expert opinions using linear [11] and non-linear combinations [12].

There also exists a vast literature on score normalisation, such as Z-Norm, T-Norm [1] (for Test Normalisation), D-norm [2] (for Distance Normalisation) and more recently U-norm [10]. They are commonly applied to speaker verification problems where client-dependent Gaussian Mixture Models are used. The core idea about client-dependent normalisation is that there are possible variations among different client models and that such variation can be normalised so that a global threshold can be found. In short, all these normalisation techniques and weight change in one way or another indeed change the *final* decision function.

In this study, a thorough survey was conducted on client-dependent score normalisation and client-dependent threshold normalisation. They are surveyed in Section 2. These two families of normalisation techniques are different in that the former normalises the score such that a global threshold can be easily found; the latter directly manipulates the threshold. This implies that the latter approach has to be computed for given specific operating costs of false acceptance and false rejection whereas the former is automatically handled by the global threshold, which remains to be optimised.

In the terms used in [8, 1], Z-Norm is impostor-centric (i.e, normalisation is carried out with respect to the impostor distributions calculated "offline" by using additional data), T-Norm is also impostor-centric (but with respect to a given utterance calculated "online" by using additional cohort impostor models). D-Norm is neither client- nor impostor-centric; it is specific to the GMM architecture and is based on Kullback-Leibler distance between two GMM models.

In this paper, we propose to implement client-dependent normalisation using F-ratio. The advantage of F-ratio normalisation, or F-Norm, is that it considers client and impostor score distributions *simultaneously*. It is hence client-impostor centric. The F-Norm that we propose here is different from Z,T,D-Norms or client-dependent threshold techniques in that few of these techniques exploit *global* (or client-independent) client and impostor distributions. Furthermore, it requires only two client accesses to obtain the normalising parameters.

The client-impostor centric normalisation was somewhat studied by [8] but the normalisation used is actually subtracting the empirical (and theoretical) client-dependent threshold from the expert opinion. Hence, this technique is additive and has no multiplicative effect, i.e, it does not change the variance of the score. Another work that exploits global information was studied in [7], whereby client-dependent and client-independent information sources are fused using Support Vector Machines. The authors called this technique *user-adapted fusion*. This approach is different from F-Norm in that the issue of normalisation is considered as being part of the optimising parameter for fusion. In this work, F-Norm can be treated as a pre-processing step just before a decision threshold is chosen. Hence, it can be readily applied to a unimodal biometric system. Of course, the F-Norm normalised scores can also be used for fusion. In any case, the impact of normalisation on fusion is also examined to evaluate the possible gain of improvement.

We explicitly compared F-Norm with Z-Norm and client-dependent threshold normalisation and found that F-Norm is in overall superior. The experimental results based on the average of 13 unimodal and 32 intramodal and multimodal fusion experiments carried out on the XM2VTS multimodal

database support our hypothesis.

Section 3 explains how F-Norm is derived. Database and evaluation criteria used are explained in Sections 4 and 5. Experimental results are presented in Section 6. This is followed by conclusion in Section 7.

2 A Survey of Current Normalisation Techniques

This section reviews some of the most relevant normalisation techniques that we have found in the literature. Before presenting the survey, it is useful to present some notations and statistical background related to normalisation in Section 2.1. The two families of client-dependent techniques reviewed are score normalisation (presented in Section 2.2) and threshold normalisation (presented in Section 2.3). This list is by no means complete but it is definitely representative of some of the-state-of-the-art techniques. We also show the link between these two families of normalisation in Section 2.4.

2.1 Useful Notations and Definitions

In biometric authentication, there are only two classes: client or impostor. A fully operational biometric system makes a decision using the following *decision function*:

$$F(\mathbf{x}) = \begin{cases} accept & \text{if } y(\mathbf{x}) > \Delta \\ reject & \text{otherwise,} \end{cases} \tag{1}$$

where $y(\mathbf{x})$ is the output of the underlying expert supporting the hypothesis that the biometric sample $\mathbf{x}$ received belongs to a client. For convenience, instead of writing $y(\mathbf{x})$, we use y. All variables derived from y are thus implicitly related to $\mathbf{x}$. Because of the accept-reject outcomes, the system may make two types of errors, i.e., false rejection (FR) and false acceptance (FA), depending on the true class-label of the respective access (client or impostor, respectively). Normalised versions of FA and FR are often used and called false acceptance rate (FAR) and false rejection rate (FRR), respectively. They are defined as:

$$\text{FAR}(\Delta) = \frac{\text{FA}(\Delta)}{NI}\ , \tag{2}$$

$$\text{FRR}(\Delta) = \frac{\text{FR}(\Delta)}{NC}\ . \tag{3}$$

where FA and FR count the number of FA and FR accesses, respectively; and NI and NC are the total number of impostor and client accesses, respectively.

Suppose that the client score distribution is Gaussian, with mean μ^C and standard deviation σ^C, i.e., $y^C \sim \mathcal{N}(\mu^C, (\sigma^C)^2)$. The impostor score distribution is defined similarly, i.e., $y^I \sim \mathcal{N}(\mu^I, (\sigma^I)^2)$. Equal Error Rate (EER) is defined as the point where FAR=FRR. By assuming Gaussian distributions on the scores, it can be shown [19] that the theoretical EER can be calculated as:

$$\text{EER} = \frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{\text{F-ratio}}{\sqrt{2}}\right), \tag{4}$$

where

$$\text{F-ratio} = \frac{\mu^C - \mu^I}{\sigma^C + \sigma^I}, \tag{5}$$

and

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp\left[-t^2\right] dt. \tag{6}$$

The optimal threshold is:

$$\Delta = \frac{\mu^I \sigma^C + \mu^C \sigma^I}{\sigma^I + \sigma^C}. \tag{7}$$

It can therefore be seen that F-ratio occurs naturally. The term F-ratio is used here because this value is somewhat *similar to* the standard Fisher ratio. In a two-class problem, the Fisher ratio [5, pg. 107] is defined as

$$\frac{\mu^C - \mu^I}{(\sigma^C)^2 + (\sigma^I)^2}. \tag{8}$$

In the client-dependent context, we will assume

$$y^k(j) \ \sim \ \mathcal{N}(\mu^k(j), (\sigma^k(j))^2) \tag{9}$$

for $k = \{C, I\}$ and j is the identity of a given client. Eqns. (5, 7 and 8) can be deduced similarly. Suppose $\Psi_j(y)$ is a client-dependent normalisation function for client identity j, the final decision function in this context becomes:

$$F(\mathbf{x}) = \begin{cases} accept & \text{if } \Psi_j(y) > \Delta \\ reject & \text{otherwise,} \end{cases} \tag{10}$$

which is slightly different from Eqn. (1). Note that the threshold Δ has still to be found.

2.2 Survey on Client-Dependent Score Normalisation

In the literature [1], Z-Norm is defined as:

$$\Psi_j^Z(y) = \frac{y - \mu^I(j)}{\sigma^I(j)}, \tag{11}$$

T-Norm is defined similarly. They differ in the ways these parameters are derived. The parameters in T-Norm are derived from scores obtained from the *same* access data but from *different* classifier models of *other clients* (online). The parameters in Z-Norm are derived from additional data samples (not used to train the classifier models) of other *simulated impostors* (offline). We are interested in Z-Norm here, assuming that a *few* additional data samples are available from client for implementing the normalisation.

There exists many methods that propose client-dependent normalisation assuming additional data samples. In [8], these normalisation techniques are categorised into client-centric, impostor-centric and client-impostor centric, each normalising using the client distribution, impostor distribution and both distributions, respectively.

In [20], a similar version of Z-Norm but using only the client distribution was reported, i.e., the terms $\mu^I(j)$ and $\sigma^I(j)$ are replaced by $\mu^C(j)$ and $\sigma^C(j)$ in Eqn. (11). However, this technique requires much more client accesses. The authors' experiments were based on 5 accesses per client. To increase the robustness of the estimated client distribution parameters, score-pruning was proposed. This is an iterative procedure that prunes scores that are deemed unreliable in each cycle until stable mean and standard deviation are found. In [10], this client-centric score-pruning technique is called U-Norm. It is applied to PIN-based (or password-based) speaker verification. The authors claimed that U-Norm works in this specific context because PIN-based speech has few phonetic contents and U-Norm effectively normalises the client score distribution against the phonetic contents. Classical *a priori* Z-Norm (using uninformed casual impostors) and a more realistic *a posteriori* Z-Norm (using real impostors who know the PIN) were also compared with U-Norm. For both cases, U-Norm was shown to be superior. It should be mentioned that *a priori* Z-Norm performed worse than no normalisation at all. In this experiment, as we understood, 6 utterances (from 3 other sessions; each session has 2 access claims) were used in U-Norm. Again, *significantly* more client accesses are needed to implement client-centric normalisation.

Client-impostor centric normalisation was also studied in [8] and has two variants:

$$y^{TI1} = y - S_{EER}(j) \tag{12}$$

$$y^{TI2} = y - \Delta(j) \tag{13}$$

where $S_{EER}(j)$ is a threshold found empirically (directly estimated from the data) and $\Delta(j)$ is defined in Eqn. (7), both calculated from a given training set of client identity j. The difference between these two normalisation techniques is that the latter relies on the Gaussian assumption whereas the former does not.

2.3 Survey on Client-Dependent Threshold Normalisation

There exists also another category of approaches that directly estimates the *client-dependent threshold* and is surveyed in [20, Sec. 2]. The decision function can be written as:

$$F(\mathbf{x}) = \begin{cases} accept & \text{if } y > \Psi'_j(\Delta) \\ reject & \text{otherwise.} \end{cases} \tag{14}$$

Eqn. (14) implies that the client-dependent threshold has to be tuned to specific operating costs of false acceptance and false rejection. This is a major difference with techniques mentioned in Section 2.2. Nevertheless, considering the vast amount of works in this domain, they are surveyed here. Several forms of client-dependent threshold functions were proposed in the literature. For instance,

$$\Psi'_j(\Delta) = \alpha\left((\mu^I(j) - \sigma^I(j)\right) + \beta \tag{15}$$

$$\Psi'_j(\Delta) = \mu^I(j) + \alpha\left(\sigma^I(j)\right)^2 \tag{16}$$

$$\Psi'_j(\Delta) = \alpha\mu^I(j) + (1-\alpha)\mu^C(j) \tag{17}$$

$$\Psi'_j(\Delta) = \Delta + \alpha\left(\mu^C(j) - \mu^I(j)\right) \tag{18}$$

$$\Psi'_j(\Delta) = \alpha(\mu^I(j) + \beta\sigma^I(j) + (1-\alpha)\mu^C(j) \tag{19}$$

$$\Psi'_j(\Delta) = \mu^C(j) - \alpha\sigma^C(j) \tag{20}$$

Eqn. (15) was reported in [9]; Eqns. (16–18) in [13, 16]; Eqn. (19) in [6]; and Eqn. (20) in [20]. Note that Eqns. (15–16) are client-centric and Eqns. (17–20) are client-impostor centric. Note also that the global threshold Δ is in general not considered in client-dependent threshold except Eqns. (18). This approach is called a "threshold refinement" procedure by the author. When committing to client-dependent (local) threshold normalisation, one does not necessarily ignore the global threshold.

2.4 Link Between Score Normalisation and Threshold Normalisation

The client-dependent score normalisation in Section 2.2 and client-dependent threshold normalisation in Section 2.3 are strongly related. Taking the right-hand sides of Eqn. (10) and Eqn. (14), we have:

$$\Psi_j(y) > \Delta, \tag{21}$$

$$y > \Psi'_j(\Delta). \tag{22}$$

To show that they are dual, we will re-express Eqn. (22) into the form of Eqn. (21). To do so, it is necessary to assume that $\Psi'_j(\Delta)$ takes the following form, as a function of Δ:

$$\Psi'_j(\Delta) = a\Delta + b. \tag{23}$$

Note that Eqns. (15–20) can all be expressed by Eqn. (23) for variables a and b. In particular, for Eqns. (15–17, 19–20), b takes on the right-hand sides of these equations and $a = 0$. For Eqn. (18), $a = 1$ and $b = \alpha(\mu^C(j) - \mu^I(j))$. Replacing Eqn. (23) into Eqn. (22), we obtain:

$$y > a\Delta + b. \tag{24}$$

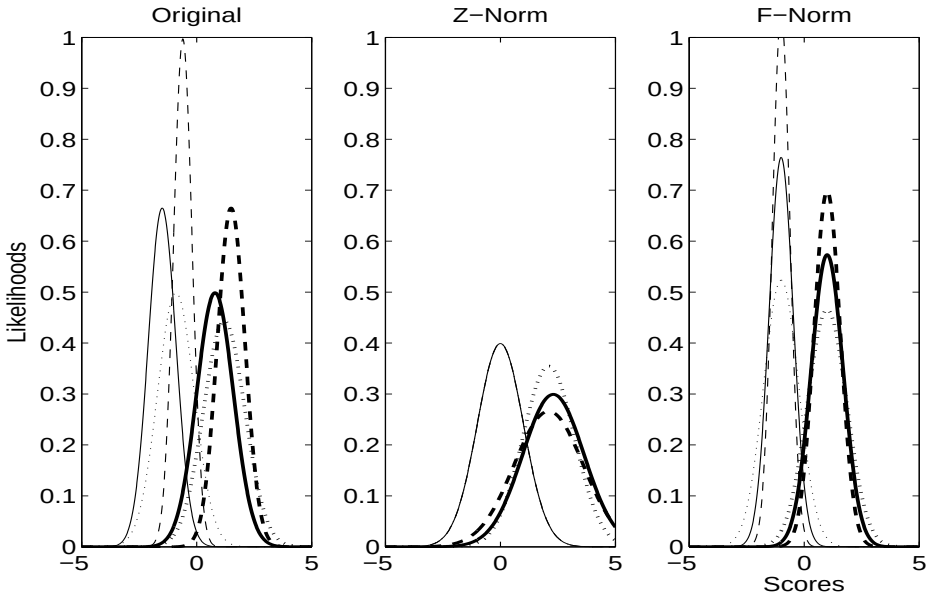

Figure 1: Comparison of the effects of F-Norm and Z-Norm. Left: the original distributions containing 3 client models (each represented by continuous, dotted and dashed lines; client score distributions are plotted with thin lines and impostor score distributions with bold lines). An global threshold may not be optimal. Middle: after applying Z-Norm, the impostor distributions become normal whereas the client distributions vary. Right: after applying F-Norm, all the client and impostor distributions are aligned so that a global threshold can be found easily.

Rearranging this equation will give:

$$\frac{y-b}{a} > \Delta \tag{25}$$

From Eqn. (21) and Eqn. (25), we see that:

$$\Psi(y) = \frac{y-b}{a}, \tag{26}$$

for Eqn. (18). For all other equations, where $a = 0$, we have:

$$\Psi(y) = y - b, \tag{27}$$

and $\Delta = 0$. As a result, manipulating the threshold or the score y has *exactly the same* effect. Note that the dual form of Eqn. (18), when written in Eqn. (26) is:

$$\Psi(y) = y - \alpha(\mu^C(j) - \mu^I(j)), \tag{28}$$

Hence, the threshold refinement procedure is just another score normalisation technique. The additional advantage of score normalisation over threshold normalisation is the additional flexibility provided by the global threshold which can still be adjusted to different operating costs of false acceptance and false rejection.

3 F-Ratio Normalisation

To give a quick idea about F-ratio normalisation, we will consider the effect of Z-Norm and the *desired* effect of F-Norm in Figure 1. In the left, there are 3 client score distributions and their respective

impostor score distributions, respectively modeled from the output of 3 client models. Z-Norm has the effect of normalising the varying impostor distributions into a single canonical impostor distribution so that decisions can be taken more easily. Unfortunately, it introduces variations into the client distributions. The objective of F-Norm is to fix both distributions, such that their means are "locked" into some pre-designated locations. For instance, it is intuitive to assign 1 to the client mean and -1 to the impostor mean. An immediate problem that may emerge is that the client mean cannot be estimated reliably because there are not enough client accesses. Here, we assume that at least as few as two samples are available. Under such limitation, we propose to use some prior information in a discriminative way.

To begin with, suppose that the "desired" mean c_k for $k = \{C, I\}$, i.e. client and impostor, respectively. $c_k|\forall_k$ are defined as:

$$c_k = \begin{cases} a & \text{if } k = C \\ -a & \text{if } k = I, \end{cases}$$

for a positive constant a. To ensure that the F-ratio value will not change, the corresponding σ^k for $k = \{C, I\}$ will have to be changed accordingly. Let $\sigma^{k'}$ be the modified standard deviations. We can then write the constraint as:

$$\text{F-ratio} = \frac{\mu^C - \mu^I}{\sigma^C + \sigma^I} = \frac{c_C - (-c_I)}{\sigma^{C'} + \sigma^{I'}} = \frac{2a}{\sigma^{C'} + \sigma^{I'}}. \tag{29}$$

The solution to this equation is:

$$\sigma^{k'} = \alpha' \sigma^k, \tag{30}$$

where,

$$\alpha' = \frac{2a}{\mu^C - \mu^I}, \tag{31}$$

for $k = \{C, I\}$. By taking the square of Eqn. (30) and applying the definition of variance of y, we obtain

$$\begin{aligned} (\sigma^{k'})^2 &= (\alpha')^2 E\left[(y^k - E[y^k])^2\right] \\ &= E\left[\left(\alpha'(y^k - E[y^k])\right)^2\right] \end{aligned} \tag{32}$$

Since α' is not dependent on the class label k, Eqn. (32) is also valid when applying to y, instead of y^k. Therefore, to map the client and impostor means to canonical values, one needs to modify the variance *without affecting* the F-ratio and the corresponding EER. This simply translates into multiplying score y with α', i.e.,

$$y^{k,\prime} = \alpha' y^k. \tag{33}$$

However, we still need to center the mean of the transformed scores, so that they are exactly $c_k|\forall_k$. The expected value of the distribution sampled from $y^{k,\prime}$ is:

$$\mu^{k,\prime} \equiv \alpha' \mu^k \tag{34}$$

Hence, the desired transformation, i.e., the F-ratio normalisation, can be achieved by shifting $y^{k,\prime}$ by $\mu^{k,\prime}$ and adding c_k. This can be done as follows:

$$y^F \equiv y' - \mu^{k,\prime} + c_k \tag{35}$$

Note that we have a choice between $k = C$ and $k = I$ to perform F-Norm. In biometric authentication task, one often does not have enough data to estimate the client mean reliably whereas one often has enough simulated impostor accesses to estimate the impostor mean more reliably. Therefore, $k = I$ is chosen.

By replacing Eqn. (33) in a class-independent manner (removing the superscript k) and Eqn. (34) into Eqn. (35), we obtain:

$$\begin{aligned} y^F &= \alpha' y - \alpha' \mu^I + c_I \\ &= \alpha'(y - \mu^I) + c_I. \end{aligned} \tag{36}$$

As a result, we obtain the F-Norm.

Until now, all variables related to y have not been tied to a particular client. Suppose that client j consists of a total of M_j scores that can be used for normalisation and that $M_j \geq 2$, i..e, there are at least 2 client scores available (apart from those used to train the baseline systems associated to client j). Let $\mu_C(j)$ be the client-dependent mean and μ_C be the client-independent mean of these scores. $\mu^I(j)$ and μ^I are defined similarly. Because each client has few scores, $\mu^C(j)$ cannot be estimated reliably, at least not as reliably as $\mu^I(j)$ (assuming that many more simulated impostor scores are available). Hence, we need some prior information. One such prior is the overall client and impostor means. We incorporate these client-independent information sources into Eqn. (31) as follows:

$$\alpha' = \frac{2a}{\beta(\mu^C(j) - \mu^I(j)) + (1-\beta)(\mu^C - \mu^I)}, \tag{37}$$

The β parameter weighs the mean difference between the client- independent mean difference and the client-dependent mean difference. It is tuned by cross-validation[1]. Similarly, Eqn. (36) can be incorporated with client-independent information as follow:

$$y^F = \alpha' \left(y - \left(\gamma \mu^I(j) + (1-\gamma)\mu^I \right) \right). \tag{38}$$

Note that slightly different from Eqn. (36), Eqn. (38) does not include c_I. This constant does not add any additional information. When applied in a client-independent manner, c_I actually ensures that the impostor mean is exactly $-a$ and the client mean is exactly a. The absence of c_I implies that the impostor distribution is centered around zero whereas the client distribution is centered around $2a$, as given by the constraint in Eqn. (29).

Preliminary experiments show that having c_I in a class-dependent context can adversely affect the resultant score. Hence, the final F-Norm function is defined by Eqns. (37 and 38). Preliminary experiments show that $\gamma = 1$ is often optimal, indicating that the shift introduced by client-dependent impostor mean is useful and very often reliable. This shift is exactly the same as in Z-Norm. Furthermore, these experiments also show that β can take a value of 1 and 0 and any values in between. This shows that incorporating β as an extra parameter, tuned in a discriminative way, can automatically adjust to the nature of the scores (which is somewhat experiment-dependent). $\beta = 1$ and $\gamma = 1$ implies that client-dependent information is beneficial whereas $\beta = 0$ and $\gamma = 0$ implies that no client-dependent normalisation is needed. The former case is actually equivalent to client-dependent threshold normalisation. This can be shown mathematically by finding F-ratio of F-normalised scores and showing that this value is equivalent to F-ratio of client-dependent threshold normalised scores (see Appendix A). In the latter case ($\beta = 0$), it can also be shown mathematically that the effect is equivalent to no normalisation at all (see Appendix B).

Hence, effectively, F-Norm is an interpolation between client-dependent threshold normalisation and no normalisation at all. It is different from Z-Norm, however, because Z-Norm does not make use of the client distributions.

4 XM2VTS Database and Systems

The XM2VTS database [15] contains synchronized video and speech data from 295 subjects, recorded during four sessions taken at one month intervals. The database is divided into three sets: a training

[1] In our implementation, we choose β to maximise the F-ratio, which is the same as minimising EER assuming that the client and impostor scores are each normally distributed, as shown in Eqn. (4).

set, an evaluation set and a test set. The training set was used to build client models, while the evaluation set was used to compute the decision thresholds as well as other hyper-parameters used by classifiers and normalisation. Finally, the test set was used to estimate the performance. The 295 subjects were divided into a set of 200 clients, 25 evaluation impostors and 70 test impostors. There exists two configurations or two different partitioning approaches of the training and evaluation sets. They are called Lausanne Protocol I and II (LP1 and LP2). The most important thing to note here is that there are only 3 samples in LP1 and 2 samples in LP2 for client-dependent adaptation and fusion training. We used altogether 7 face experts and 6 speech experts for LP1 and LP2, respectively. By combining 2 baseline experts at a time according multimodal or intramodal fusion problems, 32 fusion experiments are further identified. These experiments were reported in [17]. The 13 baseline experiments have $400 \times 13 = 5{,}200$ client accesses and $11800 \times 13 = 1{,}453{,}400$ impostor accesses. The 32 fusion experiments have $400 \times 32 = 12{,}800$ client accesses and $11{,}800 \times 32 = 3{,}577{,}600$ impostor accesses. The score files are made publicly available and are documented in [18][2].

5 Evaluation Using Pooled EPC Curve

Perhaps the most commonly used performance visualising tool in the literature is the Decision Error Trade-off (DET) curve [14]. It has been pointed out [3] that two DET curves resulted from two systems are not comparable because such comparison does not take into account how the thresholds are selected. It was argued [3] 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) [3] was proposed. We will adopt this evaluation method, which is also in coherence with the original Lausanne Protocols defined for the XM2VTS database. The criterion to choose an optimal threshold is called weighted error rate (WER), defined as follows:

$$\mathrm{WER}(\alpha, \Delta) = \alpha \mathrm{FAR}(\Delta^*) + (1 - \alpha)\ \mathrm{FRR}(\Delta^*), \tag{39}$$

where FAR and FRR are False Acceptance Rate and False Rejection Rate, respectively. Note that WER is optimised for a given $\alpha \in [0, 1]$. Let Δ^*_α be the threshold that *minimises* WER on a *development set*. The performance measure tested on an *evaluation set* at a given Δ^*_α is called Half Total Error Rate (HTER), which is defined as:

$$\mathrm{HTER}(\alpha) = \frac{\mathrm{FAR}(\Delta^*_\alpha) + \mathrm{FRR}(\Delta^*_\alpha)}{2}. \tag{40}$$

The EPC curve simply plots HTER versus α, since different values of α give rise to different values of HTER. The EPC curve can be interpreted in the same manner as the DET curve, i.e., the lower the curve is, the better the performance but for the EPC curve, the comparison is done at a given cost (controlled by α). One advantage of EPC curve is that it can plot a pooled curve from several experiments. For instance, to compare two methods over M experiments, only one pooled curve is necessary. This is done by calculating HTER at a given α point by taking into account all the false acceptance and false rejection accesses over all M experiments. The pooled FAR and FRR across $j = 1, \ldots, M$ experiments for a given $\alpha \in [0, 1]$ is defined as follow:

$$\mathrm{FAR}^{pooled}(\alpha) = \frac{\sum_{j=1}^{M} \mathrm{FA}(\Delta^*_\alpha(j))}{NI \times M}, \tag{41}$$

and

$$\mathrm{FRR}^{pooled}(\alpha) = \frac{\sum_{j=1}^{M} \mathrm{FR}(\Delta^*_\alpha(j))}{NC \times M}, \tag{42}$$

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

where $\Delta^*_\alpha(j)$ is the optimised threshold at a given α, NI is the number of impostor accesses and NC is the number of client accesses. FA and FR count the number of false acceptance and the number of false rejection at a given threshold $\Delta^*_\alpha(j)$. The pooled HTER is defined similarly as in Eqn. (40).

6 Experimental Results

Figure 2 shows the pooled EPC curve of 13 baseline experiments without applying normalisation, applying Z-Norm and applying F-Norm. Note that we could not compare F-Norm with T-Norm using the current database because we could not have access to the cohort models. As can be seen, F-Norm improves steadily over Z-Norm. The pooled EPC curve should be interpreted as the average performance over 13 baseline experiments. Of course, when analysed separately on a per experiment basis, the performance difference between F-Norm and Z-Norm is not always significant according to the HTER significance test [4] at 90% of confidence[3]. However, on average over the 13 experiments, the gain brought by F-Norm is *consistently positive* and *significant* for some large range of operating costs.

Figure 3 shows the results of 32 intramodal and multimodal fusion of face and speech experiments using the mean operator. There are six sets of 32 experiments here. The first three sets of experiments (labeled by "xxx-mean(norm)") fuse expert scores with a simple mean operator. The next three sets of experiments (labeled by "xxx-mean") fuse expert scores with *zero-mean unit variance normalisation* before applying the mean operator. Applying *zero-mean and unit variance normalisation* is often necessary when the expert opinions to fuse do not have the same variance. The normalisation parameters are obtained in a class-independent manner, similar to Z-Norm in Eqn. (11) but using global μ and σ.

The three sets of experiments use the original expert opinion, Z-Norm and F-Norm transformed opinions, respectively. From the figure, Z-Norm is sensitive to *zero-mean unit variance normalisation* normalisation. Without such normalisation fusion with Z-Norm performs worse than using the original opinions (with our without normalisation). The fusion performance of original opinions with normalisation is somewhat better than without normalisation (the two EPC curves cut each other at some points). Finally, fusion with Z-Norm transformed scores, with or without normalisation performs the best, in particular the one *without* normalisation.

F-Norm consistently outperforms Z-Norm and Z-Norm consistently outperforms no normalisation, over the 32 experiments. While this is true, on a per experiment basis, there are cases where Z-Norm outperforms F-Norm and no normalisation is better than applying any normalisation. Hence, one future direction is to examine when F-Norm should be applied and when it should not. Actually another set of experiment using client-impostor centric with theoretical EER as in Eqn. (12) was conducted as well. Its performance is so far away from the mentioned experiments that its EPC curve is not shown. More sophisticated non-linear fusion could have been used. Here, it is enough to show that the benefits obtained by F-Norm "carries through" the fusion phase.

7 Conclusions

In this paper, we proposed F-ratio normalisation, or F-Norm. This normalisation includes β and γ parameters that can balance the use of client dependent and client-independent information. It can be shown that when $\beta = 1, \gamma = 1$, F-Norm is equivalent to client-dependent threshold. When $\beta = 0, \gamma = 0$, F-Norm does not apply any normalisation. β balances the contribution between the client-dependent and client-independent mean difference (between the client and impostor distributions). γ balances the contribution between the client-dependent impostor mean and client-independent impostor mean. Because many simulated impostor accesses are available in the experimental setting, preliminary

[3]The individual experimental comparisons between Z-Norm and F-Norm are accessible in "http://www.idiap.ch/∼norman/myphp/expe/fratio".

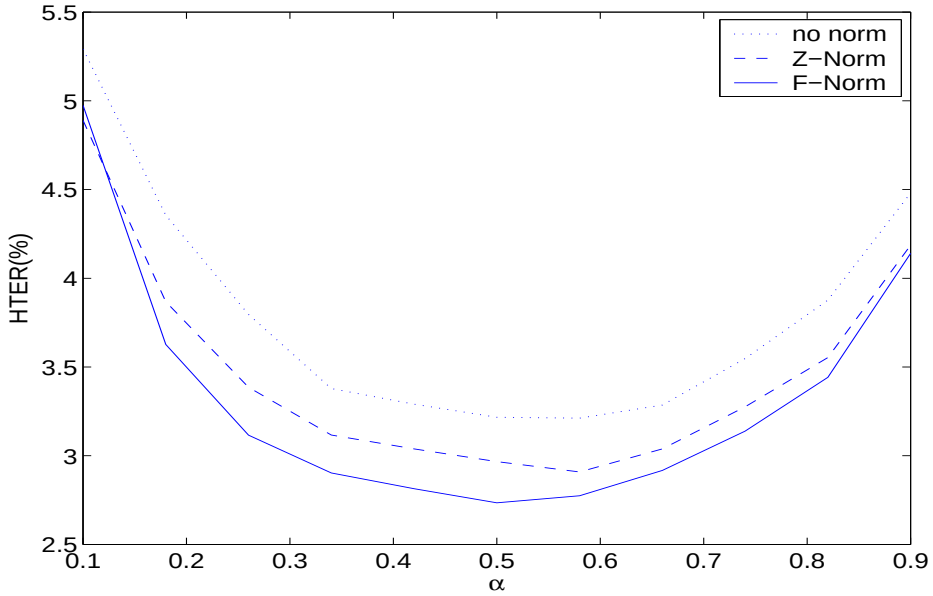

Figure 2: EPC curves of 13 baseline (face and speech) experts taken from the XM2VTS database with no normalisation, Z-Norm and F-Norm. $\gamma = 1$ and β was tuned automatically to maximise F-ratio. The improvement due to F-Norm is 95% significant compared to Z-Norm for α between 0.2 and 0.5. The client-dependent threshold normalisation using Eqn. (13) is in the range between 6.2% and 8.5% of HTER.

experiments show that $\gamma = 1$ is optimal. Hence, when applying F-ratio, one needs only to estimate the β parameter. This can be done by cross-validation or directly optimising on the training scores. With the β and γ parameter, F-Norm provides the right balance between the client-dependent and client-independent information in an experiment-dependent manner. We compared F-Norm with Z-Norm on 13 baseline biometric authentication systems on the XM2VTS database and found that on average, F-Norm is consistently superior over Z-Norm. Furthermore, for some large range of operating costs, the improvement is significant according to the HTER significance test [4]. We also further carried out 32 fusion experiments, each time taking 2 of the 13 baseline systems, based on the criterion that the experiment is either multimodal fusion or intramodal fusion. Based on these experiments, it was found that F-Norm is overall superior over no normalisation, Z-Norm and client-dependent EER normalisation. Future research will determine which normalisation techniques should be used under some specific conditions. Another direction of research is to integrate over no normalisation, Z-Norm and F-Norm such that final fused score will be optimal at a given operational false acceptance and false rejection cost.

Acknowledgement

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. The aurhots also thank Johnny Mariéthoz for fruitful discussions.

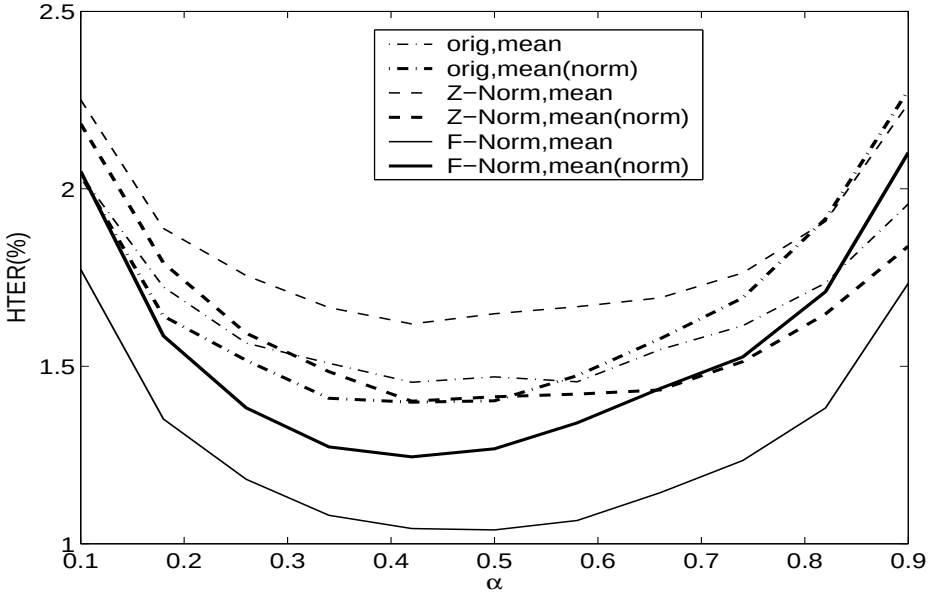

Figure 3: 32×6 (sets of) intramodal and multimodal fusion of face and speech experiments involving 2 experts carried out on the XM2VTS database using the mean operator with original expert opinions (dash-dotted line), Z-Norm (dashed line) normalised opinions, and F-Norm normalised opinions (continuous line). Another 3 sets of experiments were repeated but applying *zero-mean and unit-variance normalisation* (with parameters derived from the development set) just before fusion using the mean operator (all three plotted in bold lines). For all experiments, $\gamma = 1$ and β was tuned automatically to maximise F-ratio.

A Equivalence between F-Norm Scores with $\beta = 1, \gamma = 1$ and Threshold Dependent Normalised Scores

Suppose that a give client model j has a set of scores generated by

$$y^k \sim \mathcal{N}\left(\mu^k(j), (\sigma^k(j))^2\right),$$

for $k = \{C, I\}$, when the scores are known to belong to either the client himself $k = C$ or impostors $k = I$. When $\beta = 1$ and $\gamma = 1$, the score y^k is transformed into $y^{k,F}$ by:

$$y^{k,F} = \frac{2a}{\mu^C(j) - \mu^I(j)} \left(y^k - \mu^I(j)\right). \tag{43}$$

Suppose that normalised scores are generated by

$$y^{k,F} \sim \mathcal{N}\left(\mu^{k,F}(j), (\sigma^{k,F}(j))^2\right),$$

and we wish to find the parameters $\mu^{k,F}(j)$ and $\sigma^{k,F}(j)$, for $k = \{C, I\}$. Their solutions are:

$$\begin{aligned} \mu^{k,F}(j) &= E[y^{k,F}] \\ &= \frac{2a}{\mu^C(j) - \mu^I(j)} \left(\mu^k(j) - \mu^I(j)\right), \end{aligned} \tag{44}$$

using Eqn. (43)) and

$$\sigma^{k,F}(j) = \frac{2a}{\mu^C(j) - \mu^I(j)} \sigma^k(j), \tag{45}$$

using Eqn. (30). Note that $\mu^{I,F} = 0$ and $\mu^{C,F} = 2a$. This verifies that the client and impostor means are centered correctly. The F-ratio of F-Norm normalised scores is:

$$\text{F-ratio}^F(j) = \frac{\mu^{C,F} - \mu^{I,F}}{\sigma^{C,F} + \sigma^{I,F}}. \tag{46}$$

Replacing Eqn. (44) and Eqn. (45) into Eqn. (46), we obtain:

$$\text{F-ratio}^F(j) \quad = \quad \frac{\mu^C(j) - \mu^I(j)}{\sigma^C(j) + \sigma^I(j)} \tag{47}$$

whereby the factor $\frac{2a}{\mu^C(j) - \mu^I(j)}$ was canceled out.

The client-dependent threshold normalisation using Eqn. (12) or Eqn. (13) will transform y into y^T as follows:

$$y^T = y - \Delta(j), \tag{48}$$

where $\Delta(j)$ is a client-dependent threshold found by Eqns. (15–20), surveyed in Section 2.3, with the help of Eqn. (27).

Suppose that these normalised scores are generated by

$$y^{k,T} \sim \mathcal{N}\left(\mu^{k,T}(j), (\sigma^{k,T}(j))^2\right),$$

and we wish to find the parameters $\mu^{k,T}(j)$ and $\sigma^{k,T}(j)$, for all $k = \{C, I\}$. These solutions are:

$$\mu^{k,T}(j) = \mu^k(j) - \Delta(j) \tag{49}$$

and

$$\sigma^{k,T}(j) = \sigma^k(j). \tag{50}$$

The F-ratio of the threshold normalised scores is:

$$\text{F-ratio}^T(j) = \frac{\mu^{C,T}(j) - \mu^{I,T}(j)}{\sigma^{C,T}(j) + \sigma^{I,T(j)}}. \tag{51}$$

Replacing Eqn. (49) and Eqn. (50) into Eqn. (51), we obtain:

$$\text{F-ratio}^T(j) \quad = \quad \frac{\mu^C(j) - \mu^I(j)}{\sigma^C(j) + \sigma^I(j)} \tag{52}$$

in a user-dependent manner. Hence, from Eqns. (47 and 52), we see that:

$$\text{F-ratio}^F(j) = \text{F-ratio}^T(j) \tag{53}$$

As a result, we can conclude that F-Norm with $\beta = 1$ and $\gamma = 1$ has the same effect as applying a client-dependent threshold approach. It is important to note that the demonstration above is only true for a particular client. It does not imply that the effect of F-Norm *across all clients* is equivalent to the client-dependent threshold approach. □

B Equivalence between F-Norm Scores with $\beta = 0, \gamma = 0$ and Unnormalised Scores

When $\beta = 0$ and $\gamma = 0$, the score y^k is transformed into $y^{k,F}$ by:

$$y^{k,F}_{\beta=0} = \frac{2a}{\mu^C - \mu^I}\left(y^k - \mu^I\right). \tag{54}$$

Note that this is similar to Eqn. (44) except that the means are client-independent. Suppose that normalised scores are generated by

$$y_{\beta=0}^{k,F} \sim \mathcal{N}\left(\mu_{\beta=0}^{k,F}, (\sigma_{\beta=0}^{k,F})^2\right),$$

and we wish to find the parameters $\mu_{\beta=0}^{C,F}$ and $\sigma_{\beta=0}^{C,F}$. Their solutions are:

$$\mu_{\beta=0}^{k,F} = \frac{2a}{\mu^C - \mu^I}\left(\mu^k - \mu^I\right), \tag{55}$$

and

$$\sigma_{\beta=0}^{k,F} = \frac{2a}{\mu^C - \mu^I}\sigma^k \tag{56}$$

The F-ratio of F-Norm normalised scores is:

$$\text{F-ratio}_{\beta=0}^{F} = \frac{\mu_{\beta=0}^{C,F} - \mu_{\beta=0}^{I,F}}{\sigma_{\beta=0}^{C,F} + \sigma_{\beta=0}^{I,F}}. \tag{57}$$

Replacing Eqn. (55) and Eqn. (56) into Eqn. (57), we obtain:

$$\begin{aligned} \text{F-ratio}_{\beta=0}^{F} &= \frac{\mu^C - \mu^I}{\sigma^C + \sigma^I} \\ &= \text{F-ratio}. \end{aligned} \tag{58}$$

whereby the factor $\frac{2a}{\mu^C - \mu^I}$, μ^I and a were all canceled out. Note that we obtain the client-independent F-ratio. In short, F-Norm with $\beta = 0$ and $\gamma = 0$ has the same effect as no normalisation at all. □

References

[1] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score Normalization for Text-Independant Speaker Verification Systems. *Digital Signal Processing (DSP) Journal*, 10:42–54, 2000.

[2] M. Ben, R. Blouet, and F. Bimbot. A Monte-Carlo Method For Score Normalization in Automatic Speaker Verification Using Kullback-Leibler Distances. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 689–692, Orlando, 2002.

[3] S. Bengio and J. Mariéthoz. The Expected Performance Curve: a New Assessment Measure for Person Authentication. In *The Speaker and Language Recognition Workshop (Odyssey)*, pages 279–284, Toledo, 2004.

[4] S. Bengio and J. Mariéthoz. A Statistical Significance Test for Person Authentication. In *The Speaker and Language Recognition Workshop (Odyssey)*, pages 237–244, Toledo, 2004.

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

[6] K. Chen. Towards Better Making a Decision in Speaker Verification. *Pattern Recognition*, 36(2):329–346, 2003.

[7] J. Fierrez-Aguilar, D. Garcia-Romero, J. Ortega-Garcia, and J. Gonzalez-Rodriguez. Exploiting General Knowledge in User-Dependent Fusion Strategies For Multimodal Biometric Verification. In *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, pages 617–620, Montreal, 2004.

[8] J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez. Target Dependent Score Normalisation Techniques and Their Application to Signature Verification. In *Springer LNCS-3072, Int'l Conf. on Biometric Authentication (ICBA)*, pages 498–504, Hong Kong, 2004.

[9] S. Furui. Cepstral Analysis for Automatic Speaker Verification. *IEEE Trans. Speech and Audio Proc*, 29(2):254–272, 1981.

[10] D. Garcia-Romero, J. Gonzalez-Rodriguez, J. Fierrez-Aguilar, and J. Ortega-Garcia. U-Norm Likelihood Normalisation in PIN-Based Speaker Verification Systems. In *Springer LNCS-2688, 4th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2003)*, pages 208–213, Guildford, 2003.

[11] A. Jain and A. Ross. Learning User-Specific Parameters in Multibiometric System. In *Proc. Int'l Conf. of Image Processing (ICIP 2002)*, pages 57–70, New York, 2002.

[12] A. Kumar and D. Zhang. Integrating Palmprint with Face for User Authentication. In *Workshop on Multimodal User Authentication (MMUA 2003)*, pages 107–112, Santa Barbara, 2003.

[13] J. Lindberg, J.W. Koolwaaij, H.-P. Hutter, D. Genoud, M. Blomberg, J.-B. Pierrot, and F. Bimbot. Techniques for a priori Decision Threshold Estimation in Speaker Verification. In *Proc. of the Workshop R̈econnaissance du Locuteur et ses Applications Commerciales et Criminalistiques(RLA2C)*, pages 89–92, Avignon, 1998.

[14] A. Martin, G. Doddington, T. Kamm, M. Ordowsk, and M. Przybocki. The DET Curve in Assessment of Detection Task Performance. In *Proc. Eurospeech'97*, pages 1895–1898, Rhodes, 1997.

[15] 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*, volume 4, pages 858–863, Barcelona, 2000.

[16] J.B. Pierrot, J. Lindberg, J.W. Koolwaaij, H.P. Hutter, D. Genoud, M. Blomberg, and F.Bimbot. A Comparison of *a priori* Threshold Setting Procedures for Speaker Verification in the CAVE Project. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pages 125–128, Seattle, 1998.

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

[18] 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.

[19] N. Poh and S. Bengio. Why Do Multi-Stream, Multi-Band and Multi-Modal Approaches Work on Biometric User Authentication Tasks? In *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pages vol. V, 893–896, Montreal, 2004.

[20] J.R. Saeta and J. Hernando. On the Use of Score Pruning in Speaker Verification for Speaker Dependent Threshold Estimation. In *The Speaker and Language Recognition Workshop (Odyssey)*, pages 215–218, Toledo, 2004.