F-RATIO CLIENT-DEPENDENT NORMALISATION FOR BIOMETRIC AUTHENTICATION TASKS

Norman Poh and Samy Bengio

IDIAP, CP 592, 1920 Martigny, Switzerland

ABSTRACT

This study investigates a new *client-dependent normalisation* to improve biometric authentication systems. There exists many client-dependent score normalisation techniques applied to speaker authentication, such as Z-Norm, D-Norm and T-Norm. Such normalisation is intended to adjust the variation across different client models. We propose "F-ratio" normalisation, or F-Norm, applied to face and speaker authentication systems. This normalisation requires only that *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 a mixture parameter. This parameter *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 unimodal experiments carried out on the XM2VTS multimodal database show that such normalisation is advantageous over Z-Norm, client-dependent threshold normalisation or 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 [1], model-dependent score normalisation [2] or different weighing of expert opinions using linear [3] or non-linear combination [4]. There also exists a vast literature on score normalisation, such as Z-Norm, T-Norm [5] (for Test Normalisation), D-norm [6] (for Distance Normalisation). 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. All these normalisation techniques and weight change in one way or another indeed change the *final* decision function. 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*. In the terms used in [5, 2], 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 [1], a similar version of Z-Norm but using only the client distribution was reported. However, this technique requires more client accesses. The authors' experiments were based on 5 accesses per client. To increase the robustness of the estimated parameters, F-ratio normalisation (F-Norm) that we propose is client-impostor centric. It requires only two client accesses to obtain the normalising parameters. This aspect was somewhat studied by [2] 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.

There exists also another category of approaches that directly estimates the *client-dependent threshold* and is surveyed in [1, Sec. 2]. These approaches are client-impostor centric as well but their output is either accept or reject. This implies that the client-dependent threshold has to be tuned to specific operating costs of false acceptance and false rejection. They are hence not considered here.

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. In our opinion, there are two similar works in the literature that exploit the global distributions. In [7], a global threshold is refined with a client-dependent threshold. In [8], 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. 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 biometric experiments (7 face verification and 6 speaker verification experiments) carried out on the XM2VTS multimodal database support our hypothesis.

2. F-RATIO NORMALISATION

In biometric authentication, there are only two classes: client or impostor. Suppose that the client score distribution is Gaussian, with mean μ^C and standard deviation σ^C, i.e., $\mathcal{N}(\mu^C, (\sigma^C)^2)$. The impostor score distribution is defined similarly, i.e., $\mathcal{N}(\mu^I, (\sigma^I)^2)$. By

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.

assuming Gaussian distributions on the scores, it has been shown [9] that the theoretical Equal Error Rate (EER) can be calculated as:

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

where

$$\text{F-ratio} = \frac{\mu^C - \mu^I}{\sigma^C + \sigma^I}, \tag{2}$$

and

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp\left[-t^2\right] dt. \tag{3}$$

The optimal threshold is:

$$\Delta = \frac{\mu^I \sigma^C + \mu^C \sigma^I}{\sigma^I + \sigma^C}. \tag{4}$$

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 [10, pg. 107] is defined as

$$\frac{\mu^C - \mu^I}{(\sigma^C)^2 + (\sigma^I)^2}. \tag{5}$$

In the literature [5], Z-Norm is defined as:

$$y^Z = \frac{y - \mu^I(j)}{\sigma^I(j)}, \tag{6}$$

where j indicates a client-dependent estimates of μ^I and $\sigma^I(j)$. 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.

In [2], *client-dependent threshold* normalisation, or *target-impostor* normalisation as called by the authors, was studied and has two variants:

$$y^{TI1} = y - S_{EER}(j) \tag{7}$$

$$y^{TI2} = y - \Delta(j) \tag{8}$$

where $S_{EER}(j)$ is a threshold found empirically (directly estimated from the data) and $\Delta(j)$ is defined in Eqn. (4), 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 Gaussianity assumption whereas the former does not.

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 thier 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:

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

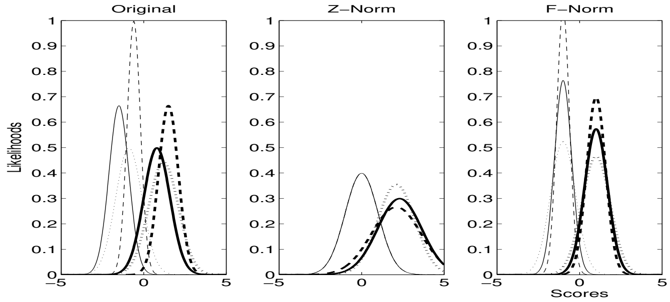

Fig. 1. Comparison of the effects of F-Norm and Z-Norm. Left: The original distributions containing 3 client models (each represented by continuous, dotted and dashed lines; client score distributions are plotted with bold lines and impostor score distributions with thin lines). A global threshold may 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.

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:

$$\text{F-ratio} = \frac{\mu^C - \mu^I}{\sigma^C + \sigma^I} = \frac{c_C - (-c_I)}{\sigma^{C'} + \sigma^{I'}} = \frac{2a}{\sigma^{C'} + \sigma^{I'}}. \tag{9}$$

The solution to this equation is:

$$\sigma^{k'} = \alpha' \sigma^k, \tag{10}$$

where,

$$\alpha' = \frac{2a}{\mu^C - \mu^I}, \tag{11}$$

for $k = \{C, I\}$. By taking the square of Eqn. (10) and applying the definition of variance of y, we obtain

$$\begin{aligned} (\sigma^{k'})^2 &= (\alpha')^2 E\left[(y^k - E[y^k])^2\right] \\ &= E\left[\left(\alpha'(y^k - E[y^k])\right)^2\right] \end{aligned} \tag{12}$$

Since α' is not dependent on the class label k, Eqn. (12) 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.,

$$y^{k,\prime} = \alpha' y^k. \tag{13}$$

However, we still need to centre 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:

$$\mu^{k,\prime} \equiv \alpha' \mu^k \tag{14}$$

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:

$$y^F \equiv y' - \mu^{k,\prime} + c_k \tag{15}$$

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. (13) in a class-independent manner (removing the superscript k) and Eqn. (14) into Eqn. (15), we obtain:

$$\begin{aligned} y^F &= \alpha' y - \alpha' \mu^I + c_I \\ &= \alpha'(y - \mu^I) + c_I. \end{aligned} \quad (16)$$

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. (11) as follows:

$$\alpha' = \frac{2a}{\beta(\mu^C(j) - \mu^I(j)) + (1-\beta)(\mu^C - \mu^I)}, \quad (17)$$

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. (16) can be incorporated with client-independent information as follow:

$$y^F = \alpha' \left(y - \left(\gamma \mu^I(j) + (1-\gamma)\mu^I \right) \right). \quad (18)$$

Note that slightly different from Eqn. (16), Eqn. (18) does not inclulde 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. (9).

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. (17 and 18). 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 [11]). In the latter case ($\beta = 0$), it can also be shown mathematically that the effect is equivalent to no normalisation at all (see [11]).

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.

3. XM2VTS DATABASE AND SYSTEMS

The XM2VTS database [12] 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 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. 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 [13]. The 13 baseline experiments have 400 × 13 = 5,200 client accesses and 11800 × 13 = 1,453,400 impostor accesses. The score files are made publicly available and are documented in [14][2].

4. EVALUATION USING POOLED EPC CURVE

Perhaps the most commonly used performance visualising tool in the literature is the Decision Error Trade-off (DET) curve [15]. It has been pointed out [16] 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 [16] 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) [16] 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 threhsold is called weighted error rate (WER), defined as follows:

$$\mathrm{WER}(\alpha, \Delta) = \alpha \mathrm{FAR}(\Delta^*) + (1-\alpha)\, \mathrm{FRR}(\Delta^*), \quad (19)$$

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:

$$\mathrm{HTER}(\alpha) = \frac{\mathrm{FAR}(\Delta^*_\alpha) + \mathrm{FRR}(\Delta^*_\alpha)}{2}. \quad (20)$$

The EPC curve simply plots HTER versus α, since different values of α give rise to different values of HTERs. 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:

$$\mathrm{FAR}^{pooled}(\alpha) = \frac{\sum_{j=1}^{M} \mathrm{FA}(\Delta^*_\alpha(j))}{NI \times M}, \quad (21)$$

and

$$\mathrm{FRR}^{pooled}(\alpha) = \frac{\sum_{j=1}^{M} \mathrm{FR}(\Delta^*_\alpha(j))}{NC \times M}, \quad (22)$$

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

[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. (1).

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

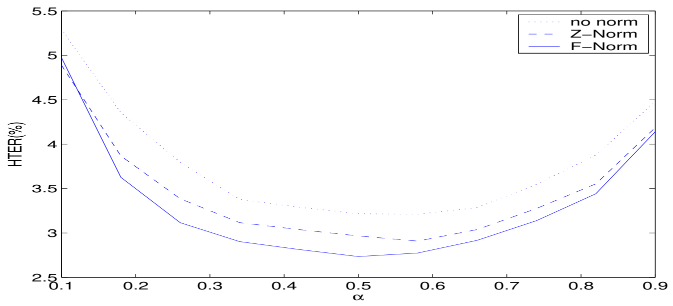

Fig. 2. EPC curves of 13 baseline (face and speech) experts taken from the XM2VTS database with no normalisation, Z-Norm and F-Norm. $\gamma = 1$ and β was tuned automatically to maximise F-ratio. The improvement due to F-Norm is 95% significant compared to Z-Norm for α between 0.2 and 0.5. The client-dependent threshold normalisation using Eqn. (8) is in the range between 6.2% and 8.5% of HTER.

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. (20).

5. 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 or they simply do not exist as T-Norm is specific to speaker verification (with Gaussian Mixture Models as classifier) while the data dealt with contains face verification systems (with Multi-Layer Perceptrons as classifier). 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 significant test [17] 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.

6. CONCLUSIONS

In this paper, we proposed F-ratio normalisation, or F-Norm. This normalisation includes a parameter β that can balance the use of client dependent and client-independent information. It can be shown that when $\beta = 1$, F-Norm is equivalent to client-dependent threshold. When $\beta = 0$, F-Norm does not apply any normalisation. Because β can be tuned, e.g., by cross-validation or directly optimising on the training data, F-Norm provides the right balance in an experiment-dependent manner. We compared F-Norm with Z-Norm on 13 baseline biometric authentication systems on the XM2VTS face and speaker verification 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 significant test [17]. Future research will determine which normalisation techniques should be used under some specific conditions. Finally, considering the popularity and rather dominant approach of T-Norm, which is specific to speaker verification tasks, a comparative study between F-Norm and T-Norm will be made. Since T-Norm takes on the form of Z-Norm (except that they are applied in different context; the former applies the *same* speech utterance on *several* cohort speaker models whereas the latter applies *several* speech utterances on the *same* speaker model), one further research direction is to implement T-Norm in the form of F-Norm (instead of the current Z-Norm).

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

7. REFERENCES

[1] 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)*, Toledo, 2004, pp. 215–218.

[2] 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)*, Hong Kong, 2004, pp. 498–504.

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

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

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

[6] 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)*, Orlando, 2002, vol. 1, pp. 689–692.

[7] 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)*, Avignon, 1998, pp. 89–92.

[8] 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)*, Montreal, 2004, vol. 5, pp. 617–620.

[9] 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)*, Montreal, 2004, pp. vol. V, 893–896.

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

[11] N. Poh and S. Bengio, "An Investigation of F-Norm in Biometric Authentication Tasks," Research Report 04-46, IDIAP, Martigny, Switzerland, 2004.

[12] J. Matas, M. Hamouz, K. Jonsson, J. Kittler, Y. Li, C. Kotropoulos, A. Tefas, I. Pitas, T. Tan, H. Yan, F. Smeraldi, J. Begun, N. Capdevielle, W. Gerstner, S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Comparison of Face Verification Results on the XM2VTS Database," in *Proc. 15th Int'l Conf. Pattern Recognition*, Barcelona, 2000, vol. 4, pp. 858–863.

[13] 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.

[14] 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.

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

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

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