A Score-Level Fusion Benchmark Database For Biometric Authentication

Norman Poh and Samy Bengio

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

Abstract. Fusing the scores of several biometric systems is a very promising approach to improve the overall system's accuracy. Despite many works in the literature, it is surprising that there is no coordinated effort in making a benchmark database available. It should be noted that fusion in this context consists not only of multimodal fusion, but also intramodal fusion, i.e., fusing systems using the same biometric modality but different features, or same features but using different classifiers. Building baseline systems from scratch often prevents researchers from putting more efforts in understanding the fusion problem. This paper describes a database of scores taken from experiments carried out on the XM2VTS face and speaker verification database. It then proposes several fusion protocols and provides some state-of-the-art tools to evaluate the fusion performance.

1 Motivation

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. Examples of biometric modalities are fingerprint, face, voice, hand-geometry and retina scans [1]. 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. Biometric data is often noisy because of deformable nature of biometric traits, corruption by environmental noise, variability over time and occlusion by the user's accessories. The higher the noise, the less reliable the biometric system becomes.

One very promising approach to improve the overall system's accuracy is to fuse the scores of several biometric systems [2]. Despite many works in the literature, e.g. [3, 4], it is surprising that there is no coordinated effort in making a benchmark database available for such task. This work is one step towards better sharing of scores to *focus* on better understanding of the fusion mechanism.

In the literature, there are several approaches towards studying fusion. One practice is to use virtual identities whereby a biometric modality from one person is paired with the biometric modality of another person. From the experiment point of view, these biometric modalities belong to the same person. While this practice is somewhat accepted in the literature, it was questioned whether this was a right thing to do or not during the 2003 Workshop on Multimodal User Authentication [5]. The fundamental issue here is the independence assumption that two or more biometric traits of a single person are independent from each other[1]. Another practice is more reasonable: use off-the-

[1] To the best of our knowledge, there is no work in the literature that approves or disapproves such assumption.

shelf biometric systems [6] and quickly acquire scores. While this is definitely a better solution, committing to acquire the systems and to collect the data is admittedly a very time-consuming process. None of the mentioned approaches prevails over the others in understanding the problem of fusion. There are currently on-going but independent projects in the biometric community to acquire multimodal biometric databases, e.g., the BANCA [7], XM2VTS [8], BIOMET [9], MYCT [10] and University of Notre Dame Biometrics multimodal databases[2]. BANCA and XM2VTS contain face and speech modalities; BIOMET contains face, speech, fingerprint, hand and signature modalities; MYCT contains ten-print fingerprint and signature modalities and University of Notre Dame Biometrics Database contains face, ear profile and hand modalities acquired using visible, Infrared-Red and range sensors at different angles. Taking multimodal biometrics in a wider context, i.e., in the sense that it involves different sensors, the FRGC[3] database can also be considered as "multimodal". It contains face modality captured using camera (at different angles) and range sensors in different (controlled or uncontrolled) settings.

As a matter of fact, most reported works in the literature about fusion often concentrates on treatment of the baseline systems. While baseline systems are definitely important, the subject of fusion is unfortunately downplayed. Hence, we propose here not only to publish scores resulted from biometric authentication experiments, but also to provide a clear documentation of the baseline systems and well-defined *fusion protocols* so that experimental results can be compared. To the best of our knowledge, this is first ever published score data set. It is intended for comparison of different fusion classifiers *on a common setting*. We further provide a set of evaluation tools such as the DET [11] curve and the recent Expected Performance Curve (EPC) [12], visualisation of False Acceptance and False Rejection Rates versus threshold, distribution of client and impostor scores, and the HTER significance test [13], among others.

The scores are taken from the publicly available XM2VTS face and speech database[4]. It should be mentioned here that there exists another software tool that analyses biometric error rate called PRESS[14]. However, it does not include the DET curve. The tools proposed here, together with the database, provide a new plot called Expected Performance Curve (EPC) [12] and a significance test specially designed to test the Half Total Error Rate (HTER) [13].

Section 2 explains the XM2VTS database, the Lausanne Protocols and the proposed Fusion Protocols. Section 3 documents the 8 baseline systems that can be used for fusion. Section 4 presents the evaluation criteria, i.e., how experiments should be reported and compared. This is followed by conclusions in Section 5.

2 Database and Protocols

2.1 The XM2VTS database and the Lausanne Protocols

The XM2VTS database [15] contains synchronised video and speech data from 295 subjects, recorded during four sessions taken at one month intervals. On each session, two recordings were made, each consisting of a speech shot and a head shot. The speech shot consisted of frontal face and speech recordings of each subject during the recital of a sentence. The database is divided into three sets: a training set, an evaluation set and a test set. The training set (LP Train) was used to build client models, while the evaluation set (LP Eval) was used to compute the decision thresholds (as well as other hyper-parameters) used by classifiers. Finally, the test set (LP Test) was used to estimate the performance.

[2] Accessible from http://www.nd.edu/∼cvrl/UNDBiometricsDatabase.html

[3] Accessible from http://www.frvt.org/FRGC/

[4] The database of scores as well as the tools mentioned are freely available for download at http://www.idiap.ch/∼norman/fusion

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, denoted as **LP1** and **LP2** in this paper. In both configurations, the test set remains the same. Their difference is that there are three training shots per client for LP1 and four training shots per client for LP2. Table 1 is the summary of the data. The last column of Table 1 is explained in Section 2.2. Note that LP Eval's of LP1 and LP2 are used to calculate the optimal thresholds that will be used in LP Test. Results are reported only for the test sets, in order to be as unbiased as possible (using an *a priori* selected threshold). More details can be found in [8].

2.2 The Fusion Protocols

The fusion protocols are built upon the Lausanne Protocols. Before the discussion, it is important to distinguish two categories of approaches: *client-independent* and *client-dependent* fusion approaches. The former approach has only a global fusion function that is common to *all* identities in the database. The latter approach has a different fusion function for a different identity. It has been reported that client-dependent fusion is better than client-independent fusion, given that there are "enough" client-dependent score data. Examples of client-dependent fusion approach are client-dependent threshold [16], client-dependent score normalisation [17] or different weighing of expert opinions using linear [18] or non-linear combination [19]. The fusion protocols that are described here can be client-dependent or client-independent.

It should be noted that one can fuse any of the 8 baseline experiments in LP1 and 5 baseline experiments in LP2 (to be detailed in Section 3). We propose a full combination of all these systems. This protocol is called **FP-full**. Hence, there are altogether $2^8 - 8 - 1 = 248$ possible combinations for LP1 and $2^5 - 5 - 1 = 26$ for LP2. The reasons for minus one and minus the number of experts are that using zero expert and using a single expert are not valid options. However, some constraints are useful. For instance, in some situations, one is constrained to using a single biometric modality. In this case, we propose an intramodal fusion (**FP-intramodal**) . When no constraint is imposed, we propose a full combination (**FP-multimodal**). FP-intramodal contains $2^5 - 5 - 1 = 26$ face-expert fusion experiments for LP1, $2^3 - 3 - 1 = 4$ speech-expert fusion experiments for LP1, 1 face-expert fusion experiment for LP2 and $2^3 - 3 - 1 = 4$ speech expert-fusion experiments for LP2. Hence, FP-intramodal contains 35 fusion experiments. The second protocol contains $\sum_{m=1}^{5}\sum_{n=1}^{3}({}^5C_m\,{}^3C_n) = 217$ combinations, where ${}^nC_k$ is "n choose k". As can be seen, the first three fusion protocols contain an exponential number of combinations. For some specific study,

Table 1. The Lausanne Protocols of XM2VTS database. The last column shows the terms used in the fusion protocols presented in Section 2.2. LP Eval corresponds to the Fusion protocols' development set while LP Test corresponds to the Fusion Protocols' evaluation set.

Data sets	Lausanne Protocols		Fusion
	LP1	LP2	Protocols
LP Train client accesses	3	4	NIL
LP Eval client accesses	600 (3×200)	400 (2×200)	Fusion dev
LP Eval impostor accesses	40,000 ($25 \times 8 \times 200$)		Fusion dev
LP Test client accesses	400 (2×200)		Fusion eva
LP Test impostor accesses	112,000† ($70 \times 8 \times 200$)		Fusion eva

†: Due to one corrupted speech file of one of the 70 impostors in this set, this file was deleted, resulting in 200 less of impostor scores, or a total of 111,800 impostor scores.

it is also useful to introduce a smaller set of combinations, each time using only two baseline experts, according to the nature of the base-expert. This protocol is called **FP-2**. Three categories of fusion types have been identified under FP-2, namely multimodal fusion (using different biometric traits), intramodal fusion with *different* feature sets and intramodal fusion with the *same* feature set but *different* classifiers. There are altogether 32 such combinations (not listed here; see [20] for details).

Note that there are 8 biometric samples in the XM2VTS database on a per client basis. They are used in the following decomposition: 3 samples are used to train the baseline experts in LP1 (and 4 in LP2) on LP Train. There are remaining 3 samples in the in LP1 Eval (and only 2 in LP2 Eval). Finally, for both protocols, 2 client accesses for testing in the *test set*. Because fusion classifiers cannot be trained using scores from the *training set*, or they are simply not available in the current settings, we are effectively using the LP Eval to train the fusion classifiers and then LP Test to test the fusion classifiers' performance on the LP Test. To avoid confusion in terminology used, we call LP Eval as the *fusion development set* and LP Test as the *fusion evaluation set*.

3 Baseline System Description

There are altogether 8 baseline systems[5] All the 8 baseline systems were used in LP1. On the other hand, 5 out of 8 were used in LP2. This results in 13 baseline experiments (for LP1 and LP2). The following explanation describes these systems in terms of their features, classifiers, and the complete system which is made up of the pair (feature type, classifier).

3.1 Face and Speech Features

The face baseline experts are based on the following features:

1. **FH**: normalised **f**ace image concatenated with its RGB **H**istogram (thus the abbreviation **FH**) [21].
2. **DCTs**: DCTmod2 features [22] extracted from face images with a size of 40×32 (rows $\times$ columns) pixels. The Discrete Cosine Transform (DCT) coefficients are calculated from an 8×8 window with horizontal and vertical overlaps of 50%, i.e., 4 pixels in each direction. Neighbouring windows are used to calculate the "delta" features. The result is a set of 35 feature vectors, each having a dimensionality of 18. (**s** indicates the use of this small image compared to the bigger size image with the abbreviation **b**.)
3. **DCTb**: Similar to DCTs except that the input face image has 80×64 pixels. The result is a set of 221 feature vectors, each having a dimensionality of 18.

The speech baseline experts are based on the following features:

1. **LFCC**: The Linear Filter-bank Cepstral Coefficient (LFCC) [23] speech features were computed with 24 linearly-spaced filters on each frame of Fourier coefficients sampled with a window length of 20 milliseconds and each window moved at a rate of 10 milliseconds. 16 DCT coefficients are computed to decorrelate the 24 coefficients (log of power spectrum) obtained from the linear filter-bank. The first temporal derivatives are added to the feature set.
2. **PAC**: The Phase Auto-Correlation Mel Filter-bank Cepstral Coefficient (PAC-MFCC) features [24] are derived with a window length of 20 milliseconds and each window moves at a rate of 10 milliseconds. 20 DCT coefficients are computed to decorrelate the 30 coefficients obtained from the Mel-scale filter-bank. The first temporal derivatives are added to the feature set.

[5] Public contribution of score files is welcome. More will be released in the future as they become available.

3. **SSC**: Spectral Subband Centroid (SSC) features, originally proposed for speech recognition [25], were used for speaker authentication in [26]. It was found that mean-subtraction could improve these features significantly. The mean-subtracted SSCs are obtained from 16 coefficients. The γ parameter, which is a parameter that raises the power spectrum and controls how much influence the centroid, is set to 0.7 [27]. Also, the first temporal derivatives are added to the feature set.

3.2 Classifiers

Two different types of classifiers were used for these experiments: Multi-Layer Perceptrons (MLPs) and a Bayes Classifier using Gaussian Mixture Models (GMMs) [28]. While in theory both classifiers could be trained using any of the previously defined feature sets, in practice MLPs are better at matching feature vectors of fixed-size while GMMs are better at matching sequences (feature vectors of unequal size). Whatever the classifier is, the hyper-parameters (e.g. the number of hidden units for MLPs or the number of Gaussian components for GMMs) are tuned on the evaluation set LP1 Eval. The same set of hyper-parameters are used in both LP1 and LP2 configurations of the XM2VTS database.

For each client-specific MLP, the feature vectors associated to the client are treated as positive patterns while all other feature vectors *not* associated to the client are treated as negative patterns. All MLPs reported here were trained using the stochastic version of the error-back-propagation training algorithm [28].

For the GMMs, two competing models are often needed: a world and a client-dependent model. Initially, a world model is first trained from an external database (or a sufficiently large data set) using the standard Expectation-Maximisation algorithm [28]. The world model is then adapted for each client to the corresponding client data using the Maximum-A-Posteriori adaptation [29] algorithm.

3.3 Baseline Systems

The baseline experiments based on DCTmod2 feature extraction were reported in [30] while those based on normalised face images and RGB histograms (FH features) were reported in [21]. Details of the experiments, coded in the pair **(feature, classifier)**, for the face experts, are as follows:

1. **(FH, MLP)** Features are normalised **F**ace concatenated with **H**istogram features. The client-dependent classifier used is an MLP with 20 hidden units. The MLP is trained with geometrically transformed images [21].
2. **(DCTs, GMM)** The face features are the DCTmod2 features calculated from an input face image of 40×32 pixels, hence, resulting in a sequence of 35 feature vectors each having 18 dimensions. There are 64 Gaussian components in the GMM. The world model is trained using *all the clients* in the training set [30].
3. **(DCTb, GMM)** Similar to (DCTs,GMM), except that the features used are DCTmod2 features calculated from an input face image of 80×64 pixels. This produces in a sequence of 221 feature vectors each having 18 dimensions. The corresponding GMM has 512 Gaussian components [30].
4. **(DCTs, MLP)** Features are the same as those in (DCTs,GMM) except that an MLP is used in place of a GMM. The MLP has 32 hidden units [30]. Note that in this case a training example consists of a *big single* feature vector with a dimensionality of 35×18. This is done by simply concatenating 35 feature vectors each having 18 dimensions[6].

[6] This may explain why MLP, an inherently discriminative classifier, has worse performance compared to GMM, a generative classifier. With high dimensionality yet having only a few training

5. **(DCTb, MLP)** The features are the same as those in (DCTb,GMM) except that an MLP with 128 hidden units is used. Note that in this case the MLP in trained on a *single* feature vector with a dimensionality of 221×18 [30].

and for the speech experts:

1. **(LFCC, GMM)** This is the Linear Filter-bank Cepstral Coefficients (LFCC) obtained from the speech data of the XM2VTS database. The GMM has 200 Gaussian components, with the minimum relative variance of each Gaussian fixed to 0.5, and the MAP adaptation weight equals 0.1. This is the best known model currently available [31] under clean conditions.
2. **(PAC, GMM)** The same GMM configuration as in LFCC is used. Note that in general, 200-300 Gaussian components would give about 1% of difference of HTER [31]. This system is particularly robust to very noisy conditions (less than 6 dBs, as tested on the NIST2001 one-speaker detection task).
3. **(SSC, GMM)** The same GMM configuration as in LFCC is used [27]. This system is known to provide an optimal performance under moderately noisy conditions (18-12 dBs, as tested on NIST2001 one-speaker detection task).

3.4 Preliminary Correlation Analysis

A preliminary analysis was carried out on the FP-2 protocol. There are 32 fusion data sets here and each data set contains scores of two experts. Each data set contains two classes: client or impostor scores. For each class of each data set, we computed the correlation between scores of two experts in the linear space. The GMM and SVM scores are used as they are. Since correlation measures the linear relationship among variables, it fails to measure the MLP scores which are trained using a tanh or a sigmoid *nonlinear* activation function. An inverse of tanh or sigmoid function is applied to the scores prior to computing the correlation values. With the absence of this *corrective* procedure, the strong correlation is *systematically* under-estimated for the intramodal fusion datasets. The resultant distribution of these correlation values, categorised into intramodal and multimodal fusion datasets, are shown in Fig. 1. As can be observed, the multimodal fusion datasets have correlation around zero whereas the intramodal fusion datasets have relatively high correlation values. This is a strong indication that the gain from fusion using the intramodal data sets will be less than that from using the multimodal data sets.

4 Performance Evaluation

There are three important concepts about evaluation of a biometric system: (1) types of errors in biometric authentication, namely false acceptance, false rejection and their combined error called Weighted Error Rate (WER), (2) threshold criterion and (3) evaluation criterion. A *threshold criterion* refers to a strategy to choose a threshold to be applied on *an evaluation (test) set*. It is necessarily tuned on *a development (training) set*. An *evaluation criterion* is used to measure the final generalisation performance and is necessarily calculated on *an evaluation set*. A fully operational biometric system makes a decision using the following *decision function*:

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

examples, the MLP cannot be trained optimally. This may affect its generalisation on unseen examples. By treating the features as a sequence, GMM was able to generalise better and hence is more adapted to this feature set.

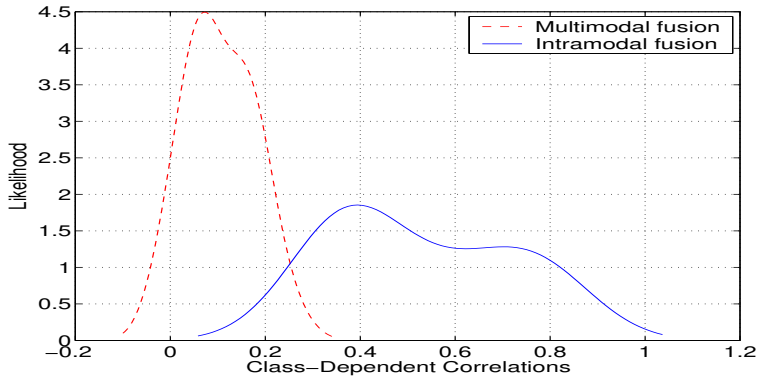

Fig. 1. Smoothed distribution (measured as unnormalised likelihood using Parzen window technique) of class-dependent correlations on the 32 fusion data sets of the FP-2 protocol according to the two categories of fusion: multimodal or intramodal. Since each data set has two classes (client and impostor), there are altogether $2 \times 32 = 64$ correlation values.

where $y(\mathbf{x})$ is the output of the underlying expert supporting the hypothesis that the biometric sample received $\mathbf{x}$ belongs to a client. The variables that follow will be derived from $y(\mathbf{x})$. For simplicity, we write y instead of $y(\mathbf{x})$. The same convention applies to variables that follow. Because of the accept-reject outcomes, the system may make two types of errors, i.e., false acceptance (FA) and false rejection (FR). 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.

To choose an "optimal threshold" Δ, it is necessary to define a threshold criterion. This has to be done on a development set. Two commonly used criteria are the Weighted Error Rate (WER) and Equal Error Rate (EER). WER is defined as:

$$\text{WER}(\alpha, \Delta) = \alpha\text{FAR}(\Delta) + (1 - \alpha)\ \text{FRR}(\Delta), \tag{4}$$

where $\alpha \in [0, 1]$ balances between FAR and FRR. A special case of WER is EER, which assumes that the costs of FA and FR are equal. It further assumes that the class prior distributions of client and impostor accesses are equal. As a result $\alpha = 0.5$. Let Δ^*_α be the optimal threshold that *minimises* WER on a *development set*. It can be calculated as follows:

$$\Delta^* = \arg\min_{\Delta} \text{WER}(\alpha, \Delta). \tag{5}$$

Note that the EER criterion can be calculated similarly by fixing $\alpha = 0.5$.

Having chosen an optimal threshold using the WER threshold criterion discussed previously, the final performance is measured using Half Total Error Rate (HTER). Note that the threshold is found with respect to a given α. It is defined as:

$$\text{HTER}(\Delta^*_\alpha) = \frac{\text{FAR}(\Delta^*_\alpha) + \text{FRR}(\Delta^*_\alpha)}{2}. \tag{6}$$

It is important to note that the FAR and FRR do not have the same *resolution*. Because there are more simulated impostor accesses than the client accesses, FRR changes more drastically when falsely rejecting a client access whereas FAR changes less drastically when falsely accepting an impostor access. Hence, when comparing the performance using HTER(Δ^*_α) from two systems (at the *same* Δ^*_α), the question of whether the HTER difference is significant or not has to take into account the imbalanced numbers of client and impostor accesses. This issue was studied in [13], and as a result, the HTER significance test was proposed. Finally, it is important to note that HTER in Eqn. (6) is identical to EER (WER with $\alpha = 0$) except that HTER is a *performance measure* (calculated on an *evaluation set* whereas EER is a *threshold criterion* optimised on a *development set*. Because of their usage in different context, EER should not be interpreted as a performance measure (in place of HTER) to compare the performance of different systems. Such practice, to our opinion, leads to an *unrealistic* comparison. The argument is that in an actual operating system, the threshold has to be fixed *a priori*. To distinguish these two concepts, when discussing HTER calculated on a development set using a threshold criterion also calculated on the same set, the HTER should be called *a posteriori* HTER. When discussing HTER calculated on an evaluation set with a threshold optimised on a development set, the HTER should be called *a priori* HTER.

The most commonly used performance visualising tool in the literature is the Decision Error Trade-off (DET) curve [11] and Receiver's Operating Characteristic (ROC) curve[7]. A DET curve is a ROC curve plotted in normal probability co-ordinate scales in its X- and Y-axes. It has been pointed out [12] that two DET curves resulting from two systems are not comparable because such comparison does not take into account how the thresholds are selected. It was argued [12] that such threshold should be chosen *a priori* as well, based on a given criterion. This is because when a biometric system is operational, the threshold parameter has to be fixed *a priori*. As a result, the Expected Performance Curve (EPC) [12] was proposed. This curve is constructed as follows: for various values of α between 0 and 1, select the optimal threshold Δ on a development (training) set, apply it on the evaluation (test) set and compute the HTER on the evaluation set. This HTER is then plotted with respect to α. The EPC curve can be interpreted similarly to the DET curve, i.e., the lower the curve, the better the generalisation performance. Although EPC is *recommended*, due to the popularity of ROC and DET curves, it is reasonable to report experimental results with these curves as well alongside with EPC. In this case, the pair of FAR and FRR values that constitute a point in ROC can be derived from the FAR and FRR terms in Eqn. (6), i.e., with the threshold Δ^*_α derived from the development (training) set.

5 Conclusions

In this study, we presented a score-level fusion database, several fusion protocols in different scenarios and some evaluation tools to encourage researchers to focus on the problem of biometric authentication score-level fusion. To the best of our knowledge, there has been no work in the literature that provides a benchmark database for score-level fusion. Hence, the efficiency of fusion classifiers can now be compared on equal platforms. We also further encourage contribution of scores following the *same* Lausanne Protocols to enrich this corpus. An extended version of this report, which includes a greater level of details on the evaluation tools, can be found in [20]. Finally, some baseline results on this data set using the fusion protocol with two experts (FP-2) can be found in [32].

[7] A good introduction can be found in "http://www.anaesthetist.com/mnm/stats/roc/".

6 Acknowledgment

This work was supported in part by the IST Program of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, funded in part by the Swiss Federal Office for Education and Science (OFES) and the Swiss NSF through the NCCR on IM2. The authors thank Julian Fierrez-Aguilar and the anonymous reviewers for giving suggestions and constructive comments, and Fabien Cardinaux and Sébastien Marcel for providing the data sets. This publication only reflects the authors' view.

References

1. A.K. Jain, R. Bolle, and S. Pankanti, *Biometrics: Person Identification in a Networked Society*, Kluwer Publications, 1999.
2. 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.
3. J. Kittler, K. Messer, and J. Czyz, "Fusion of Intramodal and Multimodal Experts in Personal Identity Authentication Systems," in *Proc. Cost 275 Workshop*, Rome, 2002, pp. 17–24.
4. J. Fierrez-Aguilar, J. Ortega-Garcia, D. Garcia-Romero, and J. Gonzalez-Rodriguez, "A Comparative Evaluation of Fusion Strategies for Multimodal Biometric Verification," in *Springer LNCS-2688, 4th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2003)*, Guildford, 2003, pp. 830–837.
5. J-L. Dugelay, J-C. Junqua, K. Rose, and M. Turk (Organizers), *Workshop on Multimodal User Authentication (MMUA 2003)*, no publisher, Santa Barbara, CA, 11–12 December, 2003.
6. M. Indovina, U. Uludag, R. Snelick, A. Mink, and A. Jain, "Multimodal Biometric Authentication Methods: A COTS Approach," in *Workshop on Multimodal User Authentication (MMUA 2003)*, Santa Barbara, 2003, pp. 99–106.
7. E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran, "The BANCA Database and Evaluation Protocol," in *Springer LNCS-2688, 4th Int. Conf. Audio- and Video-Based Biometric Person Authentication, AVBPA'03*. 2003, Springer-Verlag.
8. J. Lüttin, "Evaluation Protocol for the XM2FDB Database (Lausanne Protocol)," Communication 98-05, IDIAP, Martigny, Switzerland, 1998.
9. S. Carcia-Salicetti, C. Beumier, G. Chollet, B. Dorizzi, J. Leroux les Jardins, J. Lunter, Y. Ni, and D. Petrovska-Delacrtaz, "BIOMET: A Multimodal Person Authentication Database Including Face, Voice, Fingerprint, Hand and Signature Modalities," in *Springer LNCS-2688, 4th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA'03)*, Guildford, 2003, pp. 845–853.
10. J. Ortega-Garcia, J. Fierrez-Aguilar, D. Simon, J. Gonzalez, M. Faundez-Zanuy, V. Espinosa, A. Satue, I. Hernaez, J.-J. Igarza, C. Vivaracho, D. Escudero, and Q.-I. Moro, "Biometric on the Internet MCYT Baseline Corpus: a Bimodal Biometric Database," *IEE Proc. Visual Image Signal Processing*, vol. 150, no. 6, pp. 395–401, December 2003.
11. 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.
12. S. Bengio and J. Mariéthoz, "The Expected Performance Curve: a New Assessment Measure for Person Authentication," in *The Speaker and Language Recognition Workshop (Odyssey)*, Toledo, 2004, pp. 279–284.
13. S. Bengio and J. Mariéthoz, "A Statistical Significance Test for Person Authentication," in *The Speaker and Language Recognition Workshop (Odyssey)*, Toledo, 2004, pp. 237–244.

14. M. E. Schuckers and C. J. Knickerbocker, *Documentation for Program for Rate Estimation and Statistical Summaries PRESS*, Department of Mathematics, Computer Science and Statistics St Lawrence University, Canton, NY 13617 and Center for Identification Technology Research, West Virginia University.
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*, Barcelona, 2000, vol. 4, pp. 858–863.
16. 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.
17. 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.
18. 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.
19. 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.
20. 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.
21. S. Marcel and S. Bengio, "Improving Face Verification Using Skin Color Information," in *Proc. 16th Int. Conf. on Pattern Recognition*, Quebec, 2002, p. unknown.
22. C. Sanderson and K.K. Paliwal, "Fast Features for Face Authentication Under Illumination Direction Changes," *Pattern Recognition Letters*, vol. 24, no. 14, pp. 2409–2419, 2003.
23. L. Rabiner and B-H Juang, *Fundamentals of Speech Recognition*, Oxford University Press, 1993.
24. S. Ikbal, H. Misra, and H. Bourlard, "Phase Auto-Correlation (PAC) derived Robust Speech Features," in *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP-03)*, Hong Kong, 2003, pp. 133–136.
25. K. K. Paliwal, "Spectral Subband Centroids Features for Speech Recognition," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Seattle, 1998, vol. 2, pp. 617–620.
26. N. Poh, C. Sanderson, and S. Bengio, "An Investigation of Spectral Subband Centroids For Speaker Authentication," in *Springer LNCS-3072, Int'l Conf. on Biometric Authentication (ICBA)*, Hong Kong, 2004, pp. 631–639.
27. N. Poh, C. Sanderson, and S. Bengio, "An Investigation of Spectral Subband Centroids For Speaker Authentication," Research Report 03-62, IDIAP, Martigny, Switzerland, 2003.
28. C. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1999.
29. J.L. Gauvain and C.-H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Obervation of Markov Chains," *IEEE Tran. Speech Audio Processing*, vol. 2, pp. 290–298, 1994.
30. F. Cardinaux, C. Sanderson, and S. Marcel, "Comparison of MLP and GMM Classifiers for Face Verification on XM2VTS," in *Springer LNCS-2688, 4th Int'l Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA'03)*, Guildford, 2003, pp. 911–920.
31. N. Poh and S. Bengio, "Noise-Robust Multi-Stream Fusion for Text-Independent Speaker Authentication," in *The Speaker and Language Recognition Workshop (Odyssey)*, Toledo, 2004, pp. 199–206.
32. 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.