# On the Use of Log-Likelihood Ratio Based Model-Specific Score Normalisation in Biometric Authentication

Norman Poh and Josef Kittler

CVSSP, University of Surrey, Guildford, GU2 7XH, Surrey, UK
normanpoh@ieee.org, j.kittler@surrey.ac.uk

**Abstract.** It has been shown that the authentication performance of a biometric system is dependent on the models/templates specific to a user. As a result, some users may be more easily recognised or impersonated than others. We propose a *model-specific* (or user-specific) likelihood based score normalisation procedure that can reduce this dependency. While in its original form, such an approach is not feasible due to the paucity of data, especially of the genuine users, we stabilise the estimates of local model parameters with help of the user-independent (hence global) parameters. The proposed approach is shown to perform better than the existing known score normalisation procedures, e.g., the Z-, F- and EER-norms, in the majority of experiments carried out on the XM2VTS database . While these existing procedures are *linear* functions, the proposed likelihood based approach is *quadratic* but its complexity is further limited by a set of constraints balancing the contributions of the local and the global parameters, which are crucial to guarantee good generalisation performance.

## 1 Introduction

An automatic biometric authentication system works by first building a model or template for each user. During the operational phase, the system compares a scanned biometric sample with the registered model to decide whether an identity claim is authentic or fake. Typically, the underlying class-conditional probability distributions of scores have a strong user dependent component, modulated by within model variations. This component determines how easy or difficult it is to recognise an individual and how successfully he or she can be impersonated. The practical implication of this is that some user models (and consequently the users they represent) are systematically better (or worse) in authentication performance than others. The essence of these different situations has been popularized by the so called Doddington's zoo, with each of them characterized by a different animal name such as lamb, sheep, wolf or goat [1]. A sheep is a person who can be easily recognized; a goat is a person who is particularly difficult to be recognized; a lamb is a person who is easy to imitate; and a wolf is a person who is particularly successful at imitating others.

In the literature, there are two ways to exploit the Doddington'zoo effect to improve the system performance by using model-specific threshold and by model-specific score normalisation. The term *client-specific* is more commonly used than *model-specific*.

S.-W. Lee and S.Z. Li (Eds.): ICB 2007, LNCS 4642, pp. 614–624, 2007.

However, we prefer the latter because the source of variability is the model and *not* the user (or client). For instance, if one constructs two biometric models to represent the same person, these two models may exhibit different performance.

In model-specific thresholding, one employs a different decision threshold for each user, e.g. [2,3,4,5]. The model-specific threshold can be a function of a *global* decision threshold [6,7,8]. In model-specific score normalisation, one uses a one-to-one mapping function such that after this process, only a global threshold is needed. Examples of existing methods are Z-, D- (for Distance), T- (for Test), EER- (for Equal Error Rate) and more recently, F-Norms (for F-ratio). According to [9,10], Z-Norm [10] is impostor-centric, i.e, normalisation is carried out with respect to the impostor distributions calculated "offline" by using additional data. T-Norm [10] is also impostor-centric and its normalisation is a function of a given utterance calculated "online" by using additional cohort impostor models. D-Norm [11] is neither client- nor impostor-centric; it is specific to the Gaussian mixture model (GMM) architecture and is based on Kullback-Leibler distance between two GMM models. EER-norm [9] is client-impostor centric. In [5], a client-centric version of Z-Norm was proposed. However, this technique requires as many as five client accesses. As a consequence of promoting user-friendliness, one does not have many client-specific biometric samples. F-norm [12] is client-impostor centric; it is designed to cope with learning using as few as one sample per client (apart from those used to build the model).

In this paper, we propose a model-specific log-likelihood ratio (MS-LLR) based model-specific score normalisation procedure. While the existing Z-, D- and T-norms are linear functions, the proposed MS-LLR procedure is quadratic. Note that directly estimating the model-specific class-conditional score distributions is difficult because the number of samples available for each user is often very small. As a result, the estimated parameters of the distributions are very unreliable and this leads to unsatisfactory generalisation. We overcome this problem by adapting the model-specific (hence local) parameters from the model-independent (hence global) parameters. An important assumption in MS-LLR is that the class conditional score distributions are Gaussian. When this assumption is likely to be violated, we first transform the scores to exhibit distribution that is closer to Gaussian. The rationale is as follows: if the global (user-independent) class conditional score distributions are obviously violating the Gaussian, e.g., highly skewed, one cannot expect that the MS-LLR will be Gaussian.

When we applied the MS-LLR procedure to the individual systems in the XM2VTS score-level fusion benchmark database [13], almost all the systems showed *systematic improvement* over the baseline and more than half of them were better than the existing normalisation procedures in terms of *a posteriori* equal error rate (EER). The overall result is that better generalisation performance is obtained in terms of DET curve and of expected performance curve (EPC) where *a priori* threshold is used. This means that improvement is likely over various operating thresholds.

## 2 Methodology

Let $y$ be the output score of a biometric system and $p(y|j,k)$ be its model-specific class-conditional score distribution, where $j \in \{1,\ldots,J\}$ is a model identity and there

are $J$ models. $k$ is the class label which can be client (genuine user) or impostor, i.e., $k \in \{\mathtt{C}, \mathtt{I}\}$. A score normalisation procedure based on the log-likelihood ratio framework can be realised as follow:

$$y^{norm} = \Psi_j(y) = \log \frac{p(y|j, \mathtt{C})}{p(y|j, \mathtt{I})} \tag{1}$$

We will assume that $p(y|j,k)$ is a Gaussian, i.e., $p(y|j,k) = \mathcal{N}\left(\mu_j^k, (\sigma_j^k)^2\right)$, where $\mu_j^k$ and $\sigma_j^k$ are the class conditional mean and standard deviation of user $j$ for $k = \{\mathtt{C}, \mathtt{I}\}$. We refer to $\mu_j^k$ and $\sigma_j^k$ as *user-specific statistics*. In this case, $\Psi_j(y)$ can be written as:

$$\Psi_j(y) = \frac{1}{2(\sigma_j^{\mathtt{C}})^2}(y - \mu_j^{\mathtt{C}})^2 - \frac{1}{2(\sigma_j^{\mathtt{I}})^2}(y - \mu_j^{\mathtt{I}})^2 + \log \frac{\sigma_j^{\mathtt{C}}}{\sigma_j^{\mathtt{I}}}, \tag{2}$$

Being an LLR, such a user-specific normalization procedure is optimal (i.e., results in the lowest Bayes error) when

1. the parameters $\mu_j^k, \sigma_j^k$ for $k \in \{\mathtt{C}, \mathtt{I}\}$ and for all $j$ are estimated correctly.
2. the class-conditional scores can be described by the first and second order statistics.

The first condition is unlikely to be fulfilled in practice because there is always lack of user-specific training data. For instance, one has only two or three genuine scores to estimate $p(y|j, \mathtt{C})$ but may have more simulated impostor scores, e.g., in the order of hundreds, to estimate $p(y|j, \mathtt{I})$. As a result, in its original form, (2) is not a practical solution. The second condition can be fulfilled by converting any score such that the resulting score distribution confirms better to a Gaussian distribution.

In Section 2.1, we present the Z-norm and its variants (D- and T-norms). Other existing score normalisation procedures will also be discussed. In Section 2.2, we will show how to estimate robustly the parameters in (2) in order to fulfill the first condition. We then deal with the second condition in Section 2.3.

### 2.1 Some Existing Score Normalisation Procedures

Three types of score normalisation will be briefly discussed here. They are Z-, EER- and F-norms.

Z-norm [2] takes the form.:

$$y_j^Z = \frac{y - \mu_j^{\mathtt{I}}}{\sigma_j^{\mathtt{I}}}. \tag{3}$$

Z-norm is *impostor* centric because it relies only on the impostor distribution. In fact, it can be verified that after applying Z-norm, the resulting expected value of the impostor scores will be zero across all the models $j$. The net effect is that applying a global threshold to Z-normalised scores will give better performance than doing so with the baseline unprocessed scores.

An alternative procedure that is *client-impostor* centric is called the EER-norm [9]. It has the following two variants:

$$y^{TI1} = y - \Delta_j^{theo} \tag{4}$$

$$y^{TI2} = y - \Delta_j^{emp} \tag{5}$$

where $\Delta_j^{theo} = \frac{\mu_j^I \sigma_j^C + \mu_j^C \sigma_j^I}{\sigma_j^I + \sigma_j^C}$ is a threshold found as a result of assuming that the class-conditional distributions, $p(y|j,k)$ for both $k$, are Gaussian and $\Delta_j^{emp}$ is found empirically. In reality, the empirical version (5) cannot be used when only one or two user-specific genuine scores are available[1]. Another study conducted in [14] used a rather heuristic approach to estimate the user-specific threshold. This normalization is defined as:

$$y^{mid} = y - \underbrace{\frac{\mu_j^I + \mu_j^C}{2}} \tag{6}$$

The rest of the approaches in [14] can be seen as an approximation to this one. The under-braced term is consistent with the term $\Delta_j^{theo}$ in (4) when one assumes that $\sigma_j^C = \sigma_j^I = 1$.

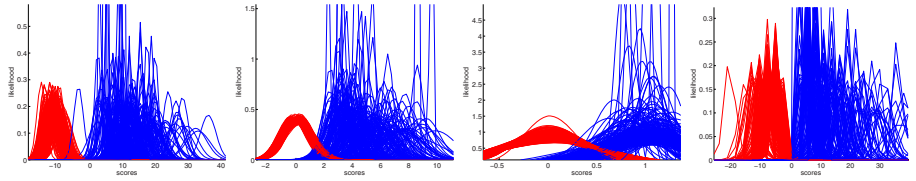
A significantly different normalisation procedure than the above two is called F-norm [12]. It is designed to project scores into another score space where the expected client and impostor scores will be the same, i.e., one for client and zero for impostor, across all $J$ models. Therefore, F-norm is also client-impostor centric. This transformation is:

$$y_j^F = \frac{y - \mu_j^I}{\gamma \mu_j^C + (1-\gamma)\mu^C - \mu_j^I}. \tag{7}$$

where $\gamma$ has to be tuned. Two sensible default values are $0$ when $\mu_j^C$ cannot be estimated because no data exists and at least $0.5$ when there is only a single user-specific sample. $\gamma$ thus accounts for the degree of reliability of $\mu_j^C$ and should be close to $1$ when abundant genuine samples are available. In all our experiments, $\gamma = 0.5$ is used when using F-norm.

In order to illustrate why the above procedures may work, we carried out an experiment on the XM2VTS database (to be discussed in Section 3). This involved training the parameters of the above score normalisation procedures on a development (training) set and applied it to an evaluation (test) set. We then plotted the model-specific class conditional distribution of the *normalised* scores, $p(y^{norm}|j,k)$, for all $j$'s and the two $k$'s. The distributions are shown in Figure 1. Since there are 200 users in the experiment, each sub-figure shows 200 Gaussian fits on the impostor distributions (the left cluster) and another 200 on the client distributions (right cluster). The normalisation procedures were trained on the development set and were applied on the evaluation set. The figures shown here are the *normalised score distributions* on the evaluation set. Prior to any normalisation, in (a), the model-specific class conditional score distributions are very different from one model to another. In (b), the impostor score distributions are aligned to centre close to zero. In (c), the impostor distributions centre around zero whereas the client distributions centre around one. Shown in (d) is the proposed MS-LLR score normalisation (to be discussed). Its resulting optimal decision boundary is located close to zero. This is a behaviour similar to EER (which was not shown here due to bad generalisation). Since the distributions in (b), (c) and (d) are better aligned than (a), improvement is expected.

[1] In our experiments, due to too few user-specific genuine scores, (4) results in poorer performance than the baseline systems without normalisation. Following this observation, the performance of EER-norm and its variants will not be reported in the paper.



(a) baseline (b) Z-norm (c) F-norm (d) MS-LLR

**Fig. 1.** Model-specific distributions $p(y^{norm}|j,k)$ for (a) the baseline system, (b) Z-norm, (c) F-norm and (d) our proposed MS-LLR using one of the 13 XM2VTS systems

### 2.2 User-Specific Parameter Adaptation

In order to make (2) practical enough as a score normalisation procedure, we propose to use the following *adapted* parameters:

$$\mu^k_{adapt,j} = \gamma^k_1 \mu^k_j + (1-\gamma^k_1)\mu^k \tag{8}$$

$$(\sigma^k_{adapt,j})^2 = \gamma^k_2 (\sigma^k_j)^2 + (1-\gamma^k_2)(\sigma^k)^2 \tag{9}$$

where $\gamma^k_1$ weighs the first moment and $\gamma^k_2$ weighs the second moment of the model-specific class-conditional scores. $\gamma^k_t$ thus provides an *explicit* control of contribution of the user-specific information against the user-independent information. Note that while (8) is found by the maximum *a posteriori* adaptation [15], (9) is not; (9) is motivated by parameter regularisation as in [16] where, in the context of classification, one can adjust between the solution of a linear discriminative analysis and that of a quadratic discriminative analysis.

We used a specific set of $\gamma^k_t$ values as follows:

$$\gamma^{\mathtt{I}}_1 = 1, \gamma^{\mathtt{I}}_2 = 1, \gamma^{\mathtt{C}}_1 = 0.5, \gamma^{\mathtt{C}}_2 = 0 \tag{10}$$

The rationale for using the first two constraints in (10) is that the model-specific statistics $\mu^{\mathtt{I}}_j$ and $\sigma^{\mathtt{I}}_j$ can be estimated reliably since a sufficiently large number of simulated impostor scores can be made available by using a development population of users. The rationale of the third (10) and fourth constraints is exactly the opposite of the first two, i.e., due to the lack of user-specific genuine scores, the statistics $\mu^{\mathtt{C}}_j$ and $\sigma^{\mathtt{C}}_j$ cannot be estimated reliably. Furthermore, between these two parameters, the second order moment ($\sigma^{\mathtt{C}}_j$) is more affected than its first order counterpart ($\mu^{\mathtt{C}}_j$). As a result, if one were to fine tune $\gamma^k_t$, the most likely one should be $\gamma^{\mathtt{C}}_j$. Our preliminary experiments on the XM2VTS database (to be discussed in Section 3) show that the value of $\gamma^{\mathtt{C}}_j$ obtained by the cross-validation procedure is not necessarily optimal. Furthermore, in the case of having only one observed genuine training score, cross-validation is impossible. For this reason, we used the default $\gamma^{\mathtt{C}}_j = 0.5$ in all our experiments. This hyper-parameter plays the same role as that of $\gamma$ in the F-norm in (7). Although the F-norm and the proposed MS-LLR are somewhat similar, MS-LLR is a direct implementation of (1) whereas the

F-norm, as well as other normalisation procedures surveyed in Section 2.1, are, at best, approximations to (2).

In brief, the proposed MS-LLR is based on (2) whose model-specific statistics are obtained via adaptation, i.e., (8) and (9). To further constrain the model, we suggest to use (10). When only one genuine samples is available, we recommend $\gamma_j^{\mathtt{C}} = 0.5$. However, when more user-specific genuine samples are available, $\gamma_j^{\mathtt{C}} > 0.5$ generalises probably better.

### 2.3 Improving the Estimate of Parametric Distribution By Score Transformation

All the existing procedures mentioned in Section 2.1, as well as our proposed one based on LLR, i.e., (2), strongly rely on the Gaussian assumption on $p(y|j,k)$. There are two solutions to this limitation. Firstly, if the physical characteristic of scores is known, the associated theoretical distribution can be used so that one replaces the Gaussian assumption with the theoretical one in order to estimate $p(y|j,k)$. Unfortunately, very often, the true distribution is not known and/or there is always not enough data to estimate $p(y|j,k)$, especially for the case $k = \mathtt{C}$.

Secondly, one can improve the parametric estimation of $p(y|j,k)$ by using an order preserving transformation that is applied globally (independent of any user). When the output score is bounded in $[a,b]$, the following transformation can be used [17]:

$$y' = \log\left(\frac{y-a}{b-y}\right) \tag{11}$$

For example, if $y$ is the probability of being a client given an observed biometric sample $x$, i.e., $y = P(\mathtt{C}|x)$, then $a = 0$ and $b = 1$. The above transformation becomes:

$$\begin{aligned} y' &= \log\left(\frac{y}{1-y}\right) = \log\left(\frac{P(\mathtt{C}|x)}{P(\mathtt{I}|x)}\right) \\ &= \log\left(\frac{p(x|\mathtt{C})}{p(x|\mathtt{I})}\right) + \log\left(\frac{P(\mathtt{C})}{P(\mathtt{I})}\right) \\ &= \underbrace{\log\left(\frac{p(x|\mathtt{C})}{p(x|I)}\right)} + const \end{aligned} \tag{12}$$

The function $\log\left(\frac{y}{1-y}\right)$ is actually an inverse of a sigmoid (or logistic) function. The underbraced term is called a log-likelihood ratio (LLR). Therefore, $y'$ can be seen as a shifted version of LLR. When the output score is not bounded, in our experience, we do not need to apply any transformation because assuming $p(y|j,k)$ to be Gaussian is often adequate. We believe that the Gaussian distribution exhibits such a good behaviour because it effectively approximates the *true* distribution using its first two moments. It should be emphasized here that the order preserving transformation discussed here does not guarantee that the resulting score distribution to be Gaussian. In fact, this is not the goal because $p(y|k)$ is in fact a mixture $p(y|j,k)$ for all $j$'s by definition. Conversely, if $p(y|k)$ is higly skewed, one cannot expect that $p(y|j,k)$ to be Gaussian.

## 3 Database, Evaluation and Results

The publicly available[2] XM2VTS benchmark database for score-level fusion [13] is used. The systems used in the experiments are shown in the first column of Table 1. For each data set, there are two sets of scores, i.e., the *development* and the *evaluation* sets. The development set is used *uniquely* to train the parameters of a given score normalisation procedure, including the threshold (bias) parameter, whereas the evaluation set is used uniquely to evaluate the generalisation performance. The fusion protocols were designed to be compatible with the originally defined Lausanne Protocols [18] (LPs). In order to train a user-specific procedure, three user-specific genuine scores are available per client for LP1 whereas only two are available for LP2.

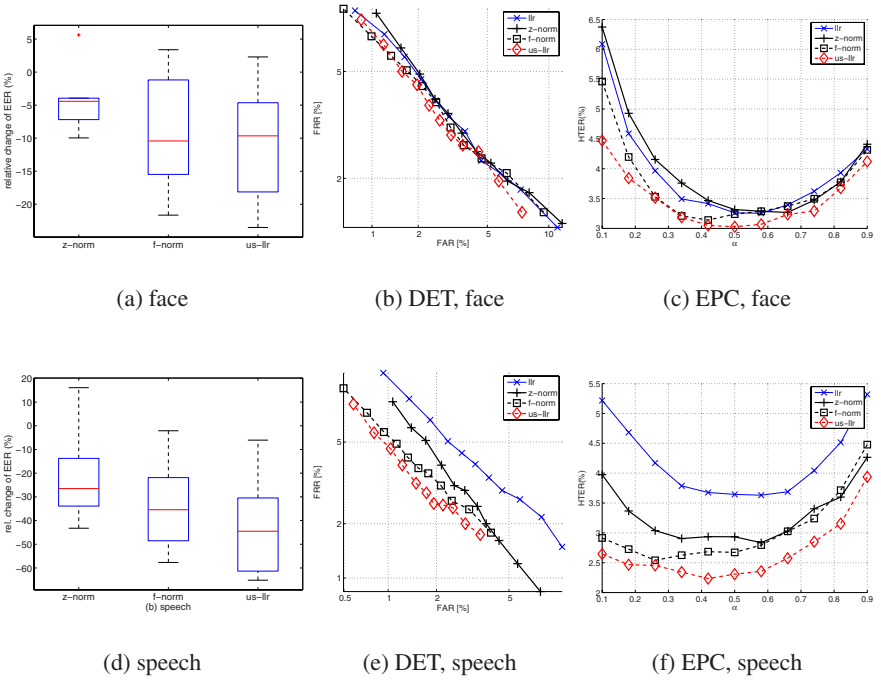
**Table 1.** Absolute performance for the *a posteriori* selected threshold calculated on the evaluation (test) score set of the 11 XM2VTS systems as well as two whose outputs are post-processed according to the techniques described in Section 2.3

| no. | system (modality, feature, classifier) | *a posteriori* EER (%) | | | |
|---|---|---|---|---|---|
| | | baseline | Z-norm | F-norm | MS-LLR |
| 1 | (F,DCTs,GMM) | 4.22 | 4.04 | ∗ 3.57 | 3.79 |
| 2 | (F,DCTb,GMM) | 1.82 | 1.92 | ∗ 1.43 | 1.65 |
| 3 | (S,LFCC,GMM) | 1.15 | 1.34 | 0.68 | ∗ 0.44 |
| 4 | (S,PAC,GMM) | 6.62 | 4.96 | 4.63 | ∗ 4.37 |
| 5 | (S,SSC,GMM) | 4.53 | 2.57 | 2.33 | ∗ 2.03 |
| 6 | (F,DCTs,MLP) | 3.53 | 3.28 | 3.14 | ∗ 2.89 |
| 7 | (F,DCTs,iMLP) | 3.53 | 3.18 | 3.19 | ∗ 2.70 |
| 8 | (F,DCTb,MLP) | 6.61 | 6.32 | 6.53 | ∗ 6.31 |
| 9 | (F,DCTb,iMLP) | 6.61 | ∗ 6.35 | 6.84 | 6.77 |
| 10 | (F,DCTb,GMM) † | ∗ 0.55 | 0.97 | 0.79 | 0.78 |
| 11 | (S,LFCC,GMM) | 1.37 | 0.99 | 0.58 | ∗ 0.48 |
| 12 | (S,PAC,GMM) | 5.39 | ∗ 4.65 | 5.28 | 5.07 |
| 13 | (S,SSC,GMM) | 3.33 | ∗ 2.20 | 2.60 | 2.32 |

Note: Rows 1–9 are from LP1 where 3 genuine samples per client are used for training; whereas rows 10–13 are from LP2 where only two are available for training. ∗ denotes the smallest EER in a row. †: We verified that for this system, the scores between the development and evaluation sets are somewhat different, thus resulting in poor estimation of the parameters of the score normalisation procedures.

The most commonly used performance visualising tool in the literature is the Decision Error Trade-off (DET) curve [19]. It has been pointed out [20] 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 [20] 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) [20] was proposed and the following criterion is used:

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




(a) face (b) DET, face (c) EPC, face

(d) speech (e) DET, speech (f) EPC, speech

**Fig. 2.** Performance of the baseline, Z-norm, F-norm and MS-LLR score normalisation procedures on the 11+2 XM2VTS systems in terms of the distribution of relative change of *a posteriori* EERs for (a) the face and (d) the speech systems shown here in boxplots; in *pooled* DET curves (b and e); and in *pooled* EPC curves (c and f). A box in a boxplot contains the first and the third quantile of relative change of *a posteriori* EERs. The dashed lines ending with horizontal lines show the 95% confidence of the data. Outliers are plotted with "+". The statistics in (a–c) are obtained from the 7 face systems shown in Table 1 whereas those in (d–f) are obtained from the remaining 6 speech systems.

$$\mathrm{WER}_\alpha(\Delta) = \alpha \mathrm{FAR}(\Delta) + (1-\alpha)\mathrm{FRR}(\Delta), \tag{13}$$

where $\alpha \in [0,1]$ balances FAR and FRR.

An EPC is constructed as follows: for various values of $\alpha$ in (13) between 0 and 1, select the optimal threshold $\Delta$ on the development (training) set, apply it on the evaluation (test) set and compute the half total error rate (HTER) on the evaluation set. HTER is the average of false acceptance rate (FAR) and false rejection rate (FRR). This HTER (in the Y-axis) is then plotted with respect to $\alpha$ (in the X-axis). The EPC curve can be interpreted similarly to the DET curve, i.e., the lower the curve, the better the generalisation performance. In this study, the *pooled* version of EPC is used to visualise the performance. This is a convenient way to compare methods on several data sets by viewing only a single curve per method. This is done by calculating the *global* FAR and FRR over a set of experiments for *each* of the $\alpha$ values. The pooled EPC curve and its implementation can be found in [13].

We applied the Z-norm, F-norm and the proposed MS-LLR score normalisation procedures on the 11+2 XM2VTS systems, i.e., the 11 original systems and two of which are based on the transformed output using (11). The *a posteriori* EER's are shown in Table 1. The improvement of each system, i.e., $\frac{\text{EER}_{norm}}{\text{EER}_{orig}} - 1$, is shown as boxplots in Figures 2(a and c), *pooled* DET curves in Figures 2(b and d), and *pooled* EPC curves in Figures 2(c and f), for the face and the speech systems, respectively. As can be observed, in all experiments, normalised scores give almost always better improvement but there is one exception, notably with system (F,DCTb,GMM). The degradation is possibly due to the mismatch between $p(y|j,k)$ in the development set and the same distribution in the evaluation set.

## 4 Conclusions

The XM2VTS database is collected under relatively controlled conditions. Although baseline performance is already very good, we show that by applying model-specific score normalisation on the output of the resulting systems, one can further improve the system performance. In particular, among the few score normalisation procedures tested, our proposed model-specific log-likelihood ratio-based (MS-LLR) approach performs best . For the speech systems, the reduction of *a posteriori* EER is 40% on average and can be as high as 60%. For the face systems, this improvement is only up to 10% on average. From the *pooled* DET and EPC curves, the average results show that MS-LLR performs best; this is followed by F-norm and Z-norm. The EER-norm performs worse than the baseline systems due to overfitting on the development set. This is because only two or three genuine samples are available. Nevertheless, for the F-norm and the proposed MS-LLR, thanks to parameter adaptation, the additional genuine scores are fully exploited. This is contrary to the Z-norm which does not make use of such information.

We conjecture that the proposed MS-LLR works best because it combines the following strategies: the general LLR framework shown in (1), the Gaussian assumption on the model-specific class conditional score distribution and the constraints in (10).

We also observe that when there is a mismatch between the development and the evaluation sets, e.g., due to different noise factors to which a biometric system is vulnerable, the model-specific class conditional distributions will change. As a result, without taking this change into account, any model-specific score normalisation may fail. This calls for predicting this change in order to take the effect of Doddington's zoo fully into account. An interesting observation is that the speech systems improve much better than the face systems. Finding out why is beyond the scope of this paper and it will be the subject of future investigation.

Another potential research direction is to combine the system outputs *after* applying model-specific score normalisation. Fusion at this level can be intramodal, i.e., involving a single biometric modality, or multimodal, i.e., involving more than one biometric modalities. Since we have already observed somewhat systematic improvement of performance after the score normalisation process, further improvement is to be expected when these outputs are used in the context of fusion. This subject is currently being investigated.

## Acknowledgment

This work was supported partially by the prospective researcher fellowship PBEL2-114330 of the Swiss National Science Foundation, by the BioSecure project (www.biosecure.info) and by the Engineering and Physical Sciences Research Council (EPSRC) Research Grant GR/S46543. This publication only reflects the authors' view.

## References

1. Doddington, G., Liggett, W., Martin, A., Przybocki, M., Reynolds, D.: Sheep, Goats, Lambs and Woves: A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation. In: Int'l Conf. Spoken Language Processing (ICSLP), Sydney (1998)
2. Furui, S.: Cepstral Analysis for Automatic Speaker Verification. IEEE Trans. Acoustic, Speech and Audio Processing / IEEE Trans. on Signal Processing 29(2), 254–272 (1981)
3. Pierrot, J.-B.: Elaboration et Validation d'Approaches en Vérification du Locuteur, Ph.D. thesis, ENST, Paris (September 1998)
4. Chen, K.: Towards Better Making a Decision in Speaker Verification. Pattern Recognition 36(2), 329–346 (2003)
5. Saeta, J.R., Hernando, J.: On the Use of Score Pruning in Speaker Verification for Speaker Dependent Threshold Estimation. In: The Speaker and Language Recognition Workshop (Odyssey), Toledo, pp. 215–218 (2004)
6. Jonsson, K., Kittler, J., Li, Y.P., Matas, J.: Support vector machines for face authentication. Image and Vision Computing 20, 269–275 (2002)
7. Lindberg, J., Koolwaaij, J.W., Hutter, H.-P., Genoud, D., Blomberg, M., Pierrot, J.-B., Bimbot, F.: Techniques for a priori Decision Threshold Estimation in Speaker Verification. In: Proc. of the Workshop Ëconnaissance du Locuteur et ses Applications Commerciales et Criminalistiques¨(RLA2C), Avignon, pp. 89–92 (1998)
8. Genoud, D.: Reconnaissance et Transformation de Locuteur, Ph.D. thesis, Ecole Polythechnique Fédérale de Lausanne (EPFL), Switzerland (1998)
9. Fierrez-Aguilar, J., Ortega-Garcia, J., Gonzalez-Rodriguez, J.: Target Dependent Score Normalisation Techniques and Their Application to Signature Verification. In: Zhang, D., Jain, A.K. (eds.) ICBA 2004. LNCS, vol. 3072, pp. 498–504. Springer, Heidelberg (2004)
10. Auckenthaler, R., Carey, M., Lloyd-Thomas, H.: Score Normalization for Text-Independant Speaker Verification Systems. Digital Signal Processing (DSP) Journal 10, 42–54 (2000)
11. Ben, M., Blouet, R., Bimbot, F.: 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, vol. 1, pp. 689–692 (2002)
12. Poh, N., Bengio, S.: F-ratio Client-Dependent Normalisation on Biometric Authentication Tasks. In: IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, pp. 721–724 (2005)
13. Poh, N., Bengio, S.: Database, Protocol and Tools for Evaluating Score-Level Fusion Algorithms in Biometric Authentication. Pattern Recognition 39(2), 223–233 (2005)
14. Toh, K.-A., Jiang, X., Yau, W.-Y.: Exploiting Global and Local Decision for Multimodal Biometrics Verification. IEEE Trans. on Signal Processing 52(10), 3059–3072 (2004)
15. Gauvain, J.L., Lee, C.-H.: Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Obervation of Markov Chains. IEEE Tran. Speech Audio Processing 2, 290–298 (1994)
16. Friedman, J.: Regularized discriminant analysis. J. American Statiscal Association 84, 165–175 (1989)

17. Dass, S.C., Zhu, Y., Jain, A.K.: Validating a biometric authentication system: Sample size requirements. IEEE Trans. Pattern Analysis and Machine Intelligence 28(12), 1302–1319 (2006)
18. Matas, J., Hamouz, M., Jonsson, K., Kittler, J., Li, Y., Kotropoulos, C., Tefas, A., Pitas, I., Tan, T., Yan, H., Smeraldi, F., Begun, J., Capdevielle, N., Gerstner, W., Ben-Yacoub, S., Abdeljaoued, Y., Mayoraz, E.: Comparison of Face Verification Results on the XM2VTS Database. In: Proc. 15th Int'l Conf. Pattern Recognition, Barcelona, vol. 4, pp. 858–863 (2000)
19. Martin, A., Doddington, G., Kamm, T., Ordowsk, M., Przybocki, M.: The DET Curve in Assessment of Detection Task Performance. In: Proc. Eurospeech'97, Rhodes, pp. 1895–1898 (1997)
20. Bengio, S., Mariéthoz, J.: The Expected Performance Curve: a New Assessment Measure for Person Authentication. In: The Speaker and Language Recognition Workshop (Odyssey), Toledo, pp. 279–284 (2004)