Revisiting Doddington's Zoo: A Systematic Method to Assess User-dependent Variabilities

[1,2]Norman Poh, [1,2]Samy Bengio and [3]Arun Ross,
[1]IDIAP Research Institute, CP 592, 1920 Martigny, Switzerland,
[2]Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, and
[3] Lane Department, West Virginia University, Morgantown, WV 26506, USA.
norman@idiap.ch, bengio@idiap.ch, ross@csee.wvu.edu

Abstract

A systematic analysis of user-dependent performance variability in the context of automatic speaker verification was first studied by Doddington et al(1998). Different categories of users were identified and labeled as sheep, goats, lambs and wolves. While this categorization is significant, it does not provide a criterion to rank the users in a database based on their variability in performance. In this work we design and evaluate a user-dependent performance criterion that requires only a limited number of client (i.e., genuine) training scores. We then extend such a study to formulate a user-specific score normalization scheme (a variant of the classical F-norm) and show that user-dependent variabilities can be reduced by employing such a scheme. The results of 13 experiments confirm the efficacy of the proposed scheme.

1 Introduction

User-specific biometric schemes was, perhaps, first exploited by Furui [1] who introduced user-dependent score normalization schemes to enhance the matching performance of automatic speaker verification systems. Later, Doddington *et al* [2] developed a statistical framework to identify different categories of individuals based on the matching performance of individual users. While there are several factors that impact matching performance (e.g., environmental mismatch between training and test sessions), their work focused on determining user-induced variability. In particular, they identified four categories of users: (a) sheep – users who can be easily recognized, (b) goats – users who are particularly difficult to be recognized, (c) lambs – users who are easy to be imitated, and (d) wolves – users who are particularly successful at imitating others. Thus, goats contribute significantly to the False Reject Rate (FRR) of a system while wolves and lambs increase its False Accept Rate (FAR). The intent of this paper is to develop a criterion to rank users based on their "recognizability" *after* mitigating the effect due to the user-induced variability. Developing such a criterion is challenging because the criterion has to (i) be based on *very few* user-specific genuine samples, (ii) generalize well on unseen data (stable), and (iii) be unbiased. To the best of our knowledge, this is the first attempt in the literature to determine the "recognizability" index of a user in a quantitative fashion. This work is different from Doddington*et al*'s [2] at least in two aspects: (i) their analysis focused on designing statistical procedures to identify wolves, goats, lambs and sheep based on match score data, whereas our focus is on designing a user-specific performance criterion; and (ii) the criterion developed here *reduces* the user-induced variability prior to sorting the users based on their recognizability.

Section 2 describes the database that was used to conduct experiments reported in this paper; Section 3 outlines the user-specific LLR framework; Section 4 proposes and investigates several criteria for ranking users and evaluates their usefulness in terms of stability and unbiasedness; and Section 5 concludes the paper.

2 Database and Data Preparation

We used the XM2VTS multimodal fusion benchmark database[1] documented in [3] to conduct the experiments reported in this paper. The database has match scores corresponding to seven face systems and six voice systems. The database was divided into training (development) and test sets according to the LP1 and LP2 protocols discussed in [3]. The label assigned to each system (Table 1) has the format Pn:m where n denotes the protocol number (1 or 2) and m denotes the order in which the respective system is invoked for an individual. For MLP-based classifiers,

[1]Available at http://www.idiap.ch/∼norman/fusion.

Table 1. The characteristics of 11 (+2 modified) systems taken from the XM2VTS benchmark fusion database.

Labels	Modalities	Features	Classifiers
P1:1	face	DCTs	GMM
P1:2	face	DCTb	GMM
P1:3	speech	LFCC	GMM
P1:4	speech	PAC	GMM
P1:5	speech	SSC	GMM
P1:6	face	DCTs	MLP
P1:7	face	DCTs	MLPi
P1:8	face	DCTb	MLP
P1:9	face	DCTb	MLPi
P2:1	face	DCTb	GMM
P2:2	speech	LFCC	GMM
P2:3	speech	PAC	GMM
P2:4	speech	SSC	GMM

DCTx is Discrete Cosine Transform coefficients and x is the size of the image block, i.e., either small (s) or big (b). LFCC is Linear Frequency Cepstral Coefficient. PAC is Phase-AutoCorrelation. SSC is Spectral Subband Centroids. Details of the systems can be found in [3]. MLPi denotes the output of MLP converted to LLR using inverse hyperbolic tangent function. P1:6 and P1:7 (resp. P1:8 and P1:9) are the *same* systems except that the scores of the latter are inversed.

their associated class-conditional scores have a skewed distribution due to the use of the logistic activation function in the output layer. Since the Log-Likelihood Ratio (LLR) is used in this paper, these scores are converted to a LLR-compatible domain by merely inverting the logistic function. This ensures that all the sub-systems (i.e, modalities and algorithms) can be studied in a common framework.

3 Towards a Robust User-specific Score Normalization Procedure

Let $y \in \mathcal{Y}$ be a realization of a match score after processing and matching a biometric sample claiming identity $j \in \{1, \dots, J\}$. This is accomplished by comparing the procured biometric sample against the template feature set corresponding to identity j in the database. The user-specific transformation into the Log-Likelihood Ratio (LLR) domain, in its most general form, can be written as:

$$y_j^{LLR} = \log \frac{p(y|C,j)}{p(y|I,j)} \tag{1}$$

where $p(y|k,j)$ is the likelihood of y being in class $k = \{C, I\}$, i.e., either *client* (C) or *impostor* (I), given the identity claim j. Note that class C is also referred to as the *genuine* class. The associated decision function is:

$$\text{decision}(y_j^{LLR}) = \begin{cases} \text{Client} & \text{if } y_j^{LLR} > \Delta_j \\ \text{Impostor} & \text{otherwise,} \end{cases} \tag{2}$$

where Δ_j is a user-specific threshold. In practice, one has extremely few scores to estimate $p(y|k,j)$, especially for the client class. Typically, the size of $\{y|C,j\}$, for any given j, is in the order of tens[2] whereas the size of $\{y|I,j\}$ is in the order of hundreds when an additional (and often external) database of users is used. For the same reason, the decision function in Eqn. (2) is user-independent, i.e., the score y^{LLR} or y (without the index j) is used along with a common threshold Δ no matter what the claimed identity is.

Due to the limited availability of user-specific data, it is sometimes assumed that $p(y|k,j)$ (i.e., the user-specific distribution) is Gaussian, i.e., $p(y|k,j) = \mathcal{N}\big(y|\mu_j^k, (\sigma_j^k)^2\big)$, where the mean is $\mu_j^k \equiv E_{y \in \mathcal{Y}|k,j}[y]$ and its corresponding variance is $(\sigma_j^k)^2 \equiv E_{y \in \mathcal{Y}|k,j}[(y - \mu_j^k)^2]$. Note that such a solution is not practical for two reasons: i) the conditional scores may not be normally distributed and ii) one always lacks user-specific training data and hence the corresponding parameters μ_j^k, σ_j^k for $k \in \{C, I\}$ and for all j cannot be estimated reliably. If one further imposes the constraint that the user-specific client information is *non-informative*, Eqn. (1) can be written as (see [4] for the derivation):

$$y_j^{LLR} = \frac{(y - \mu_j^I)^2}{2(\sigma_j^I)^2},$$

which is proportional to the square of the Z-norm [1] having the form:

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

Our goal here is to estimate Eqn. (1) *after* relaxing the Gaussian assumption. To the best of our knowledge, there are two methods to do so. In the first method, one can impose the constraint $\sigma_j^I =$ *a constant* because it is non-informative. In this way, we obtain:

$$y_j^{Z'} = y - \mu_j^I. \tag{4}$$

We call this expression the *Z-shift*. Note that the constant can be discarded as the threshold in the decision function, i.e., Eqn. (2) can be adjusted accordingly.

[2] In the database that we work on, only *two or three* scores are available for training the user-specific score normalization procedure.

The second method, known as F-norm [5], has two objectives: i) it *avoids* the need to estimate the second order conditional estimates $\sigma_j^k, \forall_{k,j}$, and ii) it takes the user-specific client information into account. A useful result from [6] is to make use of the F-ratio, which is defined as:

$$\text{F-ratio}_j = \frac{\mu_j^C - \mu_j^I}{\sigma_j^C + \sigma_j^I}. \quad (5)$$

Note that Eqn. (5) is user-specific whereas the same equation in [6] is user-independent. It is related to the Equal Error Rate (EER) as:

$$\text{EER}_j = \frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{\text{F-ratio}_j}{\sqrt{2}}\right), \quad (6)$$

where

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp\left[-t^2\right] dt. \quad (7)$$

In order to make F-ratio a useful user-specific score normalization procedure, we impose the following constraint:

$$\frac{\mu_j^C - \mu_j^I}{\sigma_j^C + \sigma_j^I} = \frac{1-0}{\sigma'^C_j + \sigma'^I_j}, \quad (8)$$

where the numerator of the RHS term is the *desired* difference in mean after the transformation and the denominator is the sum of standard deviations as a result of the transformation. Solving this constraint yields:

$$\sigma'^k_j \quad = \quad \alpha \sigma_j^k, \quad (9)$$

where $\alpha = (\mu_j^C - \mu_j^I)^{-1}$. Using the definition of variance and taking the square of Eqn. (9), we obtain:

$$(\sigma'^k_j)^2 \quad = \quad E\left[\left(\alpha(y - \mu_j^k)\right)^2\right]. \quad (10)$$

Note that the factor α is not dependent on y. This implies that one needs to multiply the score by the factor $(\mu_j^C - \mu_j^I)^{-1}$ so as to fulfill the constraint in Eqn. (8). When this transformation is carried out on the primitive form of Eqn. (4), we obtain the desired transformation as:

$$y_j^F = \frac{y_j^{Z'}}{\mu_j^C - \mu_j^I} = \frac{y - \mu_j^I}{\mu_j^C - \mu_j^I}. \quad (11)$$

We verify that the following constraints are fulfilled (by design):

$$\mu_j^{F,C} \equiv E[y^F|C,j] = \frac{E[y|C,j] - \mu_j^I}{\mu_j^C - \mu_j^I} = 1, \quad (12)$$

and

$$\mu_j^{F,I} \equiv E[y^F|I,j] = \frac{E[y|I,j] - \mu_j^I}{\mu_j^C - \mu_j^I} = 0. \quad (13)$$

In practice, however, μ_j^C cannot be estimated reliably. To account for such unreliability, a possible solution is to weigh user-specific μ_j^C with user-independent μ^C via a free parameter $\gamma \in [0, 1]$. Such a solution is classical and can be found in [7]. The final form is:

$$y_j^F = \frac{y - \mu_j^I}{\gamma \mu_j^C + (1-\gamma)\mu^C - \mu_j^I}. \quad (14)$$

A similar form of normalization was proposed in [5] and has the following form:

$$y_j^{F'} = \frac{y - \mu_j^I}{\gamma \underbrace{(\mu_j^C - \mu_j^I)} + (1-\gamma) \underbrace{(\mu^C - \mu^I)}}. \quad (15)$$

Note that the latter weighs the *difference of means* between user-specific and user-independent parameters (the underbraced terms in Eqn. (15)) whereas the former (the one proposed here) weighs between the genuine user-specific and user-independent *means*. Since the latter was previously called F-norm [5], the former is considered a *variant* of F-norm. The F-norm's variant is consistently used in this paper. Note that by setting $\gamma = 1$, both F-norm and its variant converge to the same solution. Their difference is thus rather subtle[3].

As before, the combined result using the independence assumption is $\sum_i y_{i,j}^m$ for $m \in \{Z, Z', F\}$, using Eqns. (3, 4 and 14), respectively. Among these three methods, F-norm's variant can be regarded as an *improved* procedure over Z-norm, because the former does not take the second order moment into account, thus requiring significantly fewer data points for reliable estimation. Furthermore, it is *client-impostor* centric, i.e., it relies on both genuine and impostor parameters, while Z-norm does not take the parameter of the genuine distribution into account. The proposed F-norm's variant is also an improved version of Z-shift as appeared in Eqn. (4) since the former incorporates Z-shift. The superiority of F-norm (not its variant) over Z-norm was shown in [5] empirically using the same database as the one used in this paper. We will use F-norm's variant for the rest of this paper.

4 In Search of a Stable User-Specific Criterion

In the previous section, it was mentioned that in order to make a user-specific LLR transformation practical, assuming a parametric distribution such as Gaussian is important

[3]Our intent here is not to claim that one is better than the other but to propose a unifying framework. Both procedures were observed to perform equally well on the XM2VTS fusion benchmark database. The variant of the F-norm will be used in Section 4 when developing the user-specific performance criterion.

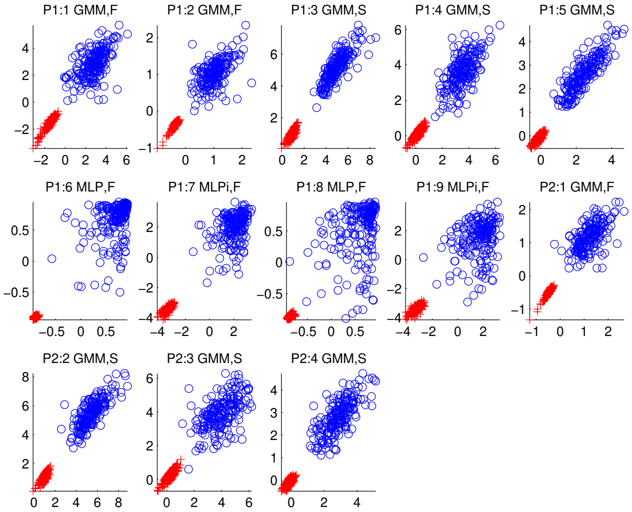


Figure 1. User-specific conditional score mean of development set versus that of evaluation set, i.e., $\mu_j^k|dev$ versus $\mu_j^k|eva$, for $k = \{C, I\}$, of 13 systems carried out on the XM2VTS. There are 200 data points for each statistic because there are 200 users. Blue circles are genuine means whereas red plus signs are impostor mean.

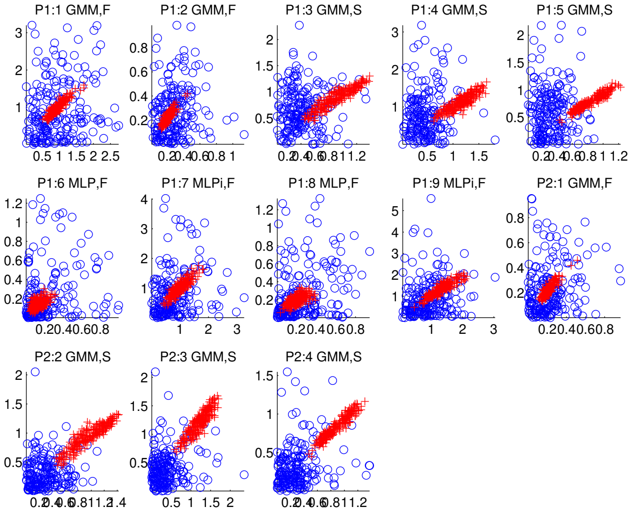


Figure 2. User-specific conditional score standard deviation of development set versus that of evaluation set, i.e., $\sigma_j^k|dev$ versus $\sigma_j^k|eva$, for $k = \{C, I\}$, of 13 systems carried out on the XM2VTS. There are 200 data points for each statistic because there are 200 users. Blue circles are genuine standard deviations whereas red plus signs are impostor standard deviations.

so that the underlying parameters can be estimated reliably based on few data samples (scores). In this section, we will analyze the scores of the 13 systems mentioned in Section 2. First, the scores are divided into two subsets: a development (training) set and an evaluation set. In our case, this task has been pre-defined by the XM2VTS fusion protocols documented in [3]. For each set of scores and for each user, we computed the class-conditional (genuine and impostor) first and second-order moments. The results are presented by plotting $\mu_j^k|dev$ versus $\mu_j^k|eva$ shown in Figure 1 and by plotting $\sigma_j^k|dev$ versus $\sigma_j^k|eva$ shown in Figure 2 for all k. Our goal here is to find out if the conditional μ_j^k or σ_j^k is reliable enough to generalize to unseen data. Note that in the XM2VTS database, the impostors in the development and in the evaluation sets are from two different sets of populations. Hence, we are actually measuring if the systems *behave in a predictive manner* for different sets of impostors. One way to measure the degree of generalization or "agreement" is by computing correlation ρ_t^k between the parameter $t \in \{\mu, \sigma\}$ estimated on a development set and the one estimated on an evaluation set, for each class $k = \{C, I\}$. We summarize ρ_t^k of the 13 systems in Figure 3 as a boxplot. Each box indicates the bound of upper and lower quantiles. The two horizontal lines at the top and the bottom of a box covers the 95% confidence bound. Any data points (correlation in this case) beyond this bound is denoted with a plus sign and is considered an outlier. Each bar contains 13 data samples. The higher the correlation, the more stable the parameter is. As can be observed and as expected, the user-specific impostor parameters are likely to be more stable than that of genuine, independent of the underlying systems. Note that there are only 2 or 3 samples (depending on whether it is the LP1 or LP2 protocol) to estimate the user-specific genuine Gaussian parameters. Despite this fact, μ_j^C is still informative. On the other hand, σ_j^C is not at all informative, judging from its relatively low correlation (about 0.2). Note that the outliers (with extremely low correlation values; indicated by plus signs) are due to the MLP systems prior to applying the inversion function of hyperbolic tangent. This is somewhat expected because the user-specific class-conditional scores are not normally distributed but are known to have a skewed distribution due to the nature of the non-linear activation function. As a result, their associated user-specific parameters generalize poorly compared to the rest of the systems. This shows that the inversion process is *effective* in mitigating this undesired effect.

A good user-specific criterion should be able to generalize over unseen novel data. Furthermore, one should be able to estimate it based on as few samples as possible. Finally,

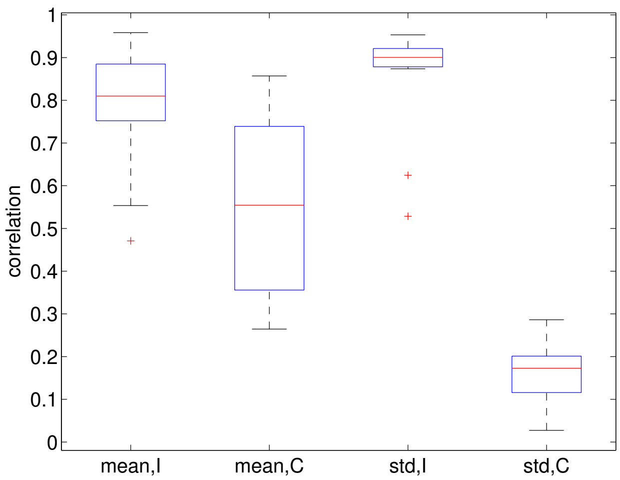


Figure 3. Boxplot of the conditional correlation $\rho^k, \forall_k$ of the four parameters, μ_j^I, μ_j^C, σ_j^I and σ_j^C of the 13 face and speech systems XM2VTS. Each correlation value is measured on 200 users. The two outliers in σ_j^I are due to (MLP,F) of P1:6 and P1:8, respectively. Similarly the outlier in μ_j^I is due to (MLP,F) of P1:6.

it has to be based on the four (or less) parameters analyzed earlier, i.e., $\mu_j^k, \sigma_j^k | k = \{C, I\}$ for each user j. An intuitive way to do so is to quantify the degree of dispersion. One such quantity is F-ratio, as defined in Eqn. (5). Other such measures are d-prime statistics used in [8] and two-class Fisher-ratio [9, Sec. 3.6]. While the two latter measures are as good as the former, we prefer F-ratio because it is functionally related to EER, as shown in Eqn. (6).

We plot the user-specific F-ratio of the 13 systems given the development set versus its evaluation set counterpart in Figure 4. As can be seen, using the original form, this quantity is very noisy and does not generalize well. Again, note that the goodness of prediction can be measured by a correlation index. Hence, user-specific F-ratio (similarly d-prime and two-class Fisher ratio) is not a good criterion because it is not stable. An intuitive way to "stabilize" such a measure is by weighting the user-specific parameters (μ and σ) in the criterion with use-independent ones as follow:

$$\mu^k_{adjusted,j} = \mu^k_j \gamma^k_1 + \mu^k (1 - \gamma^k_1) \tag{16}$$

$$\sigma^k_{adjusted,j} = \sigma^k_j \gamma^k_2 + \sigma^k (1 - \gamma^k_2) \tag{17}$$

where $\gamma^k_i \in [0, 1]$ for $i = 1, 2$ (first and second moments), are parameters to be estimated. These four parameters can be plugged into the F-ratio equation in Eqn. (5). By tuning γ^k_i, one compensates between the user-specific and user-independent information sources.

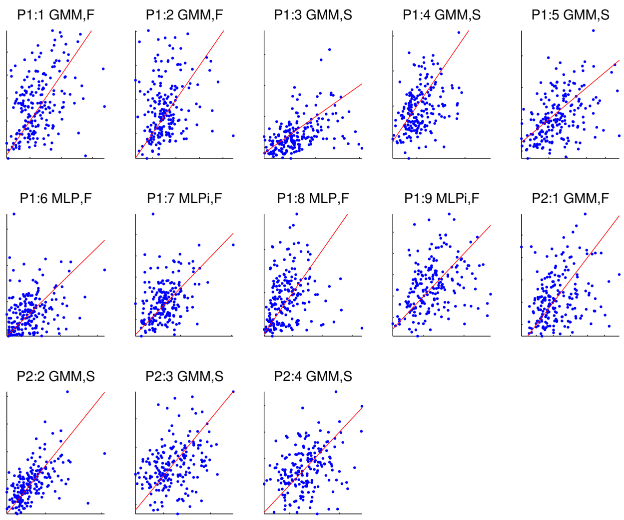


Figure 4. User-specific F-ratio as appeared in Eqn. (5) of development set versus that of evaluation set of the 13 face and speech based XM2VTS systems .

Note that from previous experiments (see Figures 2 and 3), σ_j^C is likely to contain no information.

A more conservative construction of F-ratio is to set $\gamma^C_2 = 0$, $\gamma^I_1 = 0$, $\mu^I_2 = 1$ and leaving only $\gamma^C_1 \equiv \gamma$ to be tuned. Note that this can be the same γ parameter used in Eqn. (14). This results in a criterion of the form:

$$\frac{\gamma \mu^C_j + (1-\gamma)\mu^C - \mu^I_j}{\sigma^C + \sigma^I_j} \tag{18}$$

Unfortunately, preliminary experiments show that this criterion is not satisfactory as it is heavily biased, although setting γ close to 1 does help to increase the correlation and so does its generalization ability (a figure similar to Figure 4 is not shown here but its statistics are shown in Figure 5).

We then examined the possibility of evaluating the criterion in the LLR normalized domain using F-norm's variant. Evaluating the criterion in the F-norm domain consists of finding the equivalence of the F-ratio in the F-norm domain. From the constraint in Eqn. (8) and using the weighing scheme in Eqn. (17), the F-norm of score normalized by F-norm's variant is:

$$\frac{1 - 0}{\sigma^{F,C}_{adjusted,j} + \sigma^{F,I}_{adjusted,j}}, \tag{19}$$

where $1 - 0$ is the difference between the two conditional means *after applying* F-norm. By setting $\gamma^C_2 = 0$ and $\gamma^I_2 = 1$, we obtain:

$$\frac{1}{\sigma^{F,C} + \sigma^{F,I}_j} \tag{20}$$

If we assume that $\sigma^{F,C}$ is not informative (since $\sigma_j^{F,C}$ and σ_j^C are not), we can drop the term to obtain only:

$$\frac{1}{\sigma_j^{F,I}}. \tag{21}$$

Due to Eqn. (9), this term can *equivalently* be computed using:

$$\frac{1}{(\gamma\mu_j^C + (1-\gamma\mu^C) - \mu_j^I)\sigma_j^I}. \tag{22}$$

Note that estimating $\sigma^{F,C}$ in Eqn. (20) involves actually transforming the scores into the F-norm domain. Using Eqn. (22) effectively *avoids* the need to do so and, hence, is computationally more effective.

It is obvious that γ is crucial to the success of this procedure. Preliminary experiments in [5] shows that $\gamma = 0.5$, i.e., assigning a non-informative prior to both user-specific and user-independent information, when the genuine scores are scarce, is close to optimal. Fine-tuning γ by cross-validation in the case when two user-specific genuine scores are available did not help. $\gamma \geq 0.5$ on the other hand is beneficial when abundant genuine scores are available.

We objectively compare the criteria discussed thus far using correlation and bias. Bias is defined as the arithmetic difference between a given criterion estimated on a development set and its counterpart estimated on an evaluation (test) set, or:

$$\text{bias} \equiv E_j[\text{F-ratio}_j|dev - \text{F-ratio}_j|eva]$$

Figure 5 summarizes the comparison using two boxplots: one for correlation and the other for bias. As can be observed, the constrained F-norm ratio has the best generalization ability while having an acceptable level of bias.

Before concluding this section, we evaluate the goodness of the constrained F-norm ratio by excluding the worst contributing users. The results are shown in Figure 6. Note that each DET curve is a composite DET due to all the 13 systems[4].

5 Summary and Future Work

In this work we have demonstrated that it is possible to derive a criterion to rank users according to the "strength" of their performance. Such a criterion has to (i) be based on very few user-specific genuine samples, (ii) generalize well on unseen data (stable), and (iii) be unbiased. Guided by some preliminary experiments, we first surmised that such a criterion is best evaluated in the user-specific LLR domain. In particular, three different user-specific LLR procedures were discussed, viz., Z-norm, Z-shift and F-norm. The constrained F-norm ratio was observed to exhibit the desired properties. Such a criterion is only meaningful when scores are transformed into the F-norm domain. We demonstrated the usefulness of this criterion by filtering away badly behaved users in terms of their contribution to the overall system error.

Presently, we are working on utilizing this information in a fusion framework. A suitably developed "OR"-switcher will be used to invoke only a subset of the modalities in a multimodal biometric system. The constrained F-norm ratio will be used as the criterion function to determine which of the modalities will participate in the final decision process.

[4]A composite DET is derived based on summing all FAR (resp. FRR) of each DET aligned by a criterion known as a Weighted Error Rate parameterized by $\alpha \in [0, 1]$. The pooled/composite Expected Performance Curve (EPC) is derived similarly. The details of composite DET and pooled EPC can be found in [3].

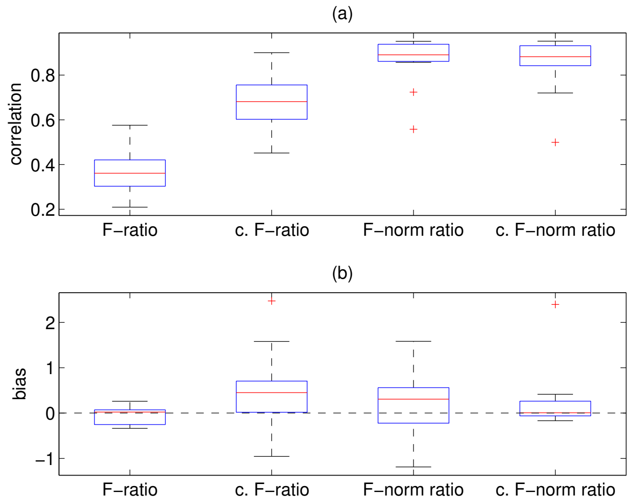


Figure 5. Comparison of the four criteria, viz., F-ratio as in Eqn. (5), constrained F-ratio as in Eqn. (18), F-norm ratio as in Eqn. (20) and constrained F-norm ratio as in Eqn. (22), using (a) correlation (which measures the generalization ability) and (b) bias (between a given criterion on the development and evaluation sets) of the 13 XM2VTS face and speech systems. Each bar summarizes 13 (correlation or bias) statistics. Higher correlation and bias around zero are desirable properties.

Acknowledgement

This work was supported in part by the IST Program of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, funded in part by the Swiss Federal Office for Education and Science (OFES) and the

Swiss NSF through the NCCR on IM2. This publication only reflects the authors' view.

References

[1] S. Furui, "Cepstral Analysis for Automatic Speaker Verification," *IEEE Trans. Acoustic, Speech and Audio Processing / IEEE Trans. on Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[2] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "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.

[3] N. Poh and S. Bengio, "Database, Protocol and Tools for Evaluating Score-Level Fusion Algorithms in Biometric Authentication," *Pattern Recognition*, vol. 39, no. 2, pp. 223–233, February 2005.

[4] Johnny Mariéthoz and Samy Bengio, "A Bayesian Framework for Score Normalization Techniques Applied to Text Independent Speaker Verification," *IEEE Signal Processing Letters*, vol. 12, no. 7, pp. 532–535, 2005.

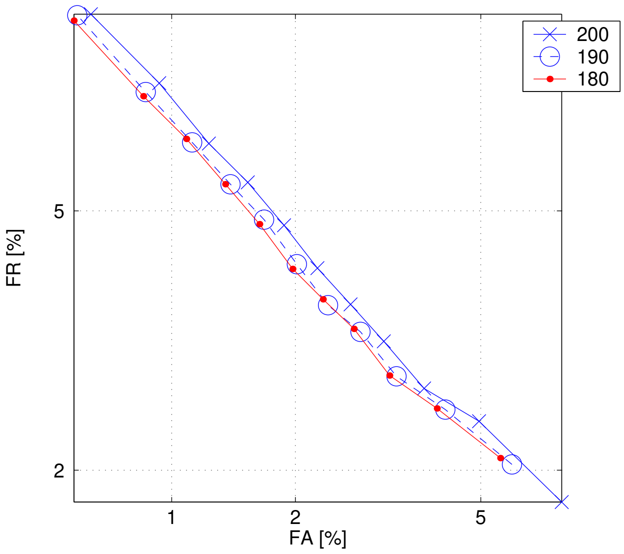
[5] N. Poh and S. Bengio, "F-ratio Client-Dependent Normalisation on Biometric Authentication Tasks," in *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, 2005, pp. 721–724.

[6] 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.

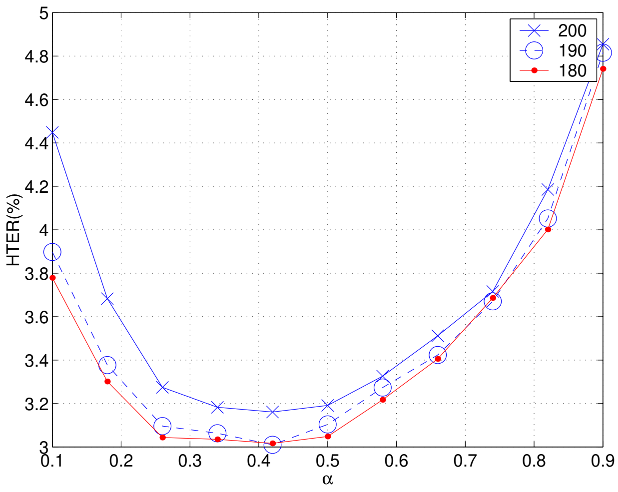
[7] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer-Verlag, 2001.

[8] J. Daugman, "High confidence visual recognition of persons by a test of statistical independence," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 11, pp. 1148–1161, 1993.

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



(a) DET



(b) EPC

Figure 6. A composite (a) DET and (b) EPC of all the 13 systems on the *evaluation* set, *after* transforming the scores into the F-domain. The original curve contains all the 200 users. The two other curves show a performance improvement due to excluding the worst 10 and 20 users identified by the constrained F-norm ratio criterion on the *development* set.