A Biometric Menagerie Index for Characterising Template/Model-Specific Variation

Norman Poh and Josef Kittler

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

Abstract. An important phenomenon influencing the performance of a biometric experiment, attributed to Doddington et al (1998), is that the match scores (whether under genuine or impostor matching) are strongly dependent on the model or template from which the match scores have been derived. Although there exist studies to classify the characteristic of the template/model, as well as the query data, into animal names such as sheep, goats, wolves and lambs – so-called Doddington's menagerie, or higher semantic categories considering simultaneously both genuine and impostor match scores, due to Yager and Dunstone (2008), there is currently absence of means to characterise the extent of Doddington's menagerie. This paper aims to design such an index, called the biometric menagerie index (BMI). It is defined as the ratio of the between-client variance and the expectation of the total variance. BMI has three desirable properties. First, it is invariant to shifting and scaling of the match scores. Second, its value lies between zero and one, with zero implying the absence of Doddington's menagerie effect, and one signifying its strong presence. Third, it is experimentally verified that BMI generalizes to *different choices* of impostor population. Our findings based on the XM2VTS benchmark score database suggest the followings: First, the BMI of genuine match scores is generally higher than that of the impostor match scores. Second, two different matching algorithms observing the same biometric data may have significantly different BMI values, hence suggesting that the biometric menagerie is algorithm-dependent.

1 Motivation

An automatic biometric system works by first building a model or template specific to a user. During the operational phase, the system compares a scanned biometric sample with the registered model to decide whether an identity claim is genuine or not (from a different person). As a result, each user model/template exhibits different behaviours in terms of output scores when being presented genuine and impostor biometric samples. The consequence is that some user models are better than others in representing the user's identity. Doddington et al's initial study [1] attempted to characterise the user models by how easy or difficult they can be recognised, and how easy query samples can imitate others, hence, introducing user categories by names such as sheep, goats, lambs and wolves. A sheep is a person who can be easily recognised; a goat is a person who is particularly difficult to be recognised; a lamb is a person who is easy to imitate; and a wolf is a person who is particularly successful at imitating others. A more recent

M. Tistarelli and M.S. Nixon (Eds.): ICB 2009, LNCS 5558, pp. 816–827, 2009.

study by Yager and Dunstone [2] further distinguishes four other semantic categories of users by considering both the genuine and impostor match scores, for *each* claimed identity, simultaneously. However, their approach considers only the *client-specific* first order moments (i.e., for each claimed identity) of the match scores. Poh and Kittler [3] further considered the second order moments from which several client-specific class-sepabarability criteria can be derived. Thanks to these criteria, the authors deomnstrated empirically that client models/templates (in the gallery) can be ranked according to their performance, hence separating well-behaved models from badly behaved models (in terms of performance).

Referring to Doddington's menagerie, sheep are characterised by high genuine similarity match scores whereas goats are characterised by low genuine similarity match scores. Lambs are defined as a symmetry of goats, i.e., having high impostor similarity match scores. Finally, wolves are persons who can consistently give high impostor similarity scores when matched against the client models (enrolled templates in the gallery). While sheep dominate the population of client models, goats constitute only a small fraction of the population. However the latter category constitutes disproportionately a large portion of false rejection error. Although the original Doddington's study was applied to speaker verification, the same phenomenon was independently observed in [3] using the face, fingerprint and iris biometric modalities; [4] using the fingerprint modality; [5] using the face modality; and many others, e.g., [2]. Using finger-vein and fingerprint as case studies, Une et al [6] proposed a measure known as the wolf attack probability, which quantifies the maximum probability of success of impersonating a victim by feeding wolves in a biometric system. These studies provide a mounting evidence that the biometric menagerie is a general phenomenon inherent in all biometric experiments.

As a result of these menagerie studies, it is beginning to be recognised that fine-tuning the system parameters (including feature extraction and classifier or distance matching parameters) and the decision threshold for each individual client model (classifier) can greatly boost the recognition (identification and verification) performance further. For instance, *lowering* the similarity decision threshold (than a globally pre-set value) for the goats is likely to compensate for their disproportionately high false rejection errors, and thus *increasing* the similarity decision threshold for the lambs in order to compensate for their disproportionately high false acceptance error. This strategy is called client (model/template) specific decision. Examples of such strategies abound: [7,8,9,10,11]. Rather than adjusting the thresholds, one can instead transform the match score distribution. This alternative strategy is called *client-specific score normalisation*. Examples are Z-norm [12], F-norm [13], EER-norm [14] and model-specific log-likelihood ratio (LLR)-based normalisation [15]. Both categories of approaches have been discussed and summarised in [16].

Clearly, a fundamental understanding of Doddington's menagerie has an important impact on designing and optimising a biometric system as a whole. There is, however, a certain lack of understanding of this phenomenon. For instance, we ignore the reason for the existence of wolves, as well as of lambs and goats. However, we know that it is certainly dependent on the choice of biometric device, the result of the user's interaction with the device and the acquisition environment, hence related to the quality of

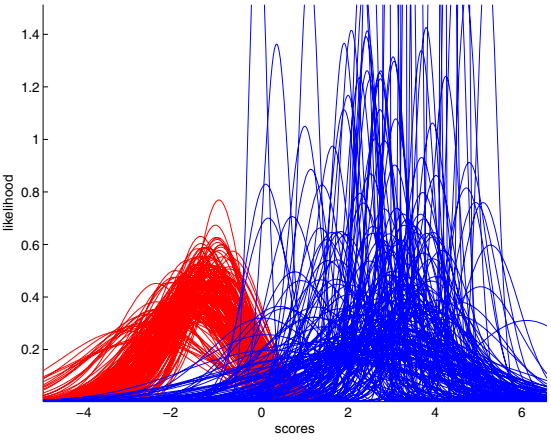


Fig. 1. An example of (client) model-specific likelihood plots (conditioned on client/genuine (blue) and impostor (red) classes, for all the 200 users using the first data set. The zoo index for the genuine match scores is 0.96 whereas for the impostor match scores, it is 0.23.

biometric samples. Although Grother and Tabassi [17] have attempted to define *quality* as a scalar summary of a sample's appropriateness to be used for matching, quality itself is clearly multi-facetted. There is currently a missing link between template-related quality measures and the Doddington's menagerie.

The goal of this paper is to advance the understanding of Doddington's menagerie by proposing a statistical measure to quantify the extent of this phenomenon. To date, there simply exists no such measure. The only means thus far has been applying a client-specific score normalisation/decision threshold procedure and observe if the procedure indeed yields an improved performance, or not, e.g., [12,13,14,15]. Attempting to diagnose such procedure is difficult since this can be due to inaccurate estimation of the parameters, especially when dealing with the genuine match score distribution (which has very few samples compared to its impostor counterpart).

The proposed measure is called the *biometric menagerie index* (BMI) or simply the *zoo index*, as a means to directly measure the dependency of a match score on a user model/template. In order to motivate our approach, we illustrate in Figure 1 the model/template-specific class-conditional score distribution, i.e., one distribution for each template and for each class, subject to being involved in genuine (same-person) and impostor (different-person) matching. An important observation here is that the genuine and impostor match scores have very different characteristics. In particular, the genuine match scores exhibit higher model-dependency than the impostor match scores for this example. This figure also shows that the compounded genuine (impostor) similarity score distribution is a mixture of *client-specific* genuine (impostor) similarity score distributions, hence, can have multiple modes.

This paper is organised as follows: Section 2 explains how BMI was derived. Section 3 shows the BMI values for face and speech biometrics and this is followed by conclusions.

2 Deriving a Zoo Index

In this section, we shall analyse the match scores conditioned on each class and each claimed identity, which is necessarily drawn from a distribution specified by $p(y|k,j)$. This distribution can be estimated from the score set $\mathcal{Y}_j^k$ introduced in the previous section. The goal of this section is to derive a zoo index. Section 2.1 will first introduce the notation. Section 2.2 defines BMI mathematically. We then present the properties of BMI in Section 2.3. Finally, the estimation of the relevant statistics is described in Section 2.4.

2.1 Notation

We shall denote $j' \in \mathcal{J}'$ to be the *true* identity whereas $j \in \mathcal{J}$ to be the *claimed* identity which must come from a client set (gallery) $\mathcal{J} = [1, \ldots, J\}$ and there are J clients (for whom templates/models are built). The set $\mathcal{J}'$ is a set of identities/users not in $\mathcal{J}$, i.e., the two populations do not overlap: $\mathcal{J} \cap \mathcal{J}' = \emptyset$. Thus, throughout this study, we deal with *open-set* recognition, and an extension to closed-set recognition is straightforward.

In order to simplify the notation, we also assume that there is only a single model associated with each client. In a usual identification setting, one compares all samples belonging to j' against the model of j in order to obtain a score set $\mathcal{Y}(j,j')$. When $j = j'$, the match score set is genuine (client), whereas when $j \neq j'$, it is impostor. We further introduce two *client-specific* score sets, both dependent on the claimed identity: $\mathcal{Y}_j^{\mathtt{C}} \equiv \mathcal{Y}'(j,j)$ for the client class, and $\mathcal{Y}_j^{\mathtt{I}}$ for the impostor class. In this study, for the impostor class, the scores are a *union* or aggregation of all other users (from $\mathcal{J}'$) claiming to be j, i.e., $\mathcal{Y}_j^{\mathtt{I}} \equiv \bigcup_{j' \in \mathcal{J}, j' \neq j} \mathcal{Y}'(j,j')$.

2.2 Deriving the Zoo Index Using the Within- and Between-Class-Variances

Using the above notations, we shall use the score variable y to reprensent an element in the set $\mathcal{Y}_j^k$ for a given class $k = [\mathtt{C}, \mathtt{I}\}$ (client or impostor) and a given claimed identity j. The unknown disribution from which $\mathcal{Y}_j^k$ was generated is denoted by $p(y|k,j)$. Thus, the unconditional distribution of y is

$$p(y) = \sum_{k,j} p(y|k,j) P(j|k) P(k)$$

where $P(j|k)$ is the prior class-conditional probability claiming identity j and $P(k)$ is the prior class probability.

The class-conditional expected value (global mean) of y is defined by:

$$\begin{aligned} \mu^k &= \sum_{j=1}^{J} \left(\int_y p(y|k,j) y \, dy \right) P(j|k) \\ &\equiv \mathbb{E}_j \left[\mathbb{E}_y \left[y|k,j \right] |k \right] \end{aligned} \tag{1}$$

where $P(j|k) = P(j,k)/P(k)$ and we have used the following notation

$$\mathbb{E}_y \left[y|k, j = j_* \right] \equiv \int_y y p(y|k, j = j_*) \, dy$$

to denote the expectation of y conditioned on a particular claimed identity $j = j_*$ as well as the class k, and

$$\mathbb{E}_j[\bullet|k] \equiv \sum_{j=1}^{J} \bullet P(j|k)$$

to denote the expectation over all clients, conditioned on k.

In contrast to (1), the (class-conditional) client-specific mean is defined as:

$$\mu_j^k \equiv \mathbb{E}_y[y|k,j]$$

which is related to the (class-conditional) global mean by:

$$\mu^k = \mathbb{E}_j\left[\mu_j^k|k\right]$$

as is evident from (1).

The (class-conditional) global score variance, for a given claimed identity j_* can be calculated as follows:

$$\begin{aligned}\mathbb{E}_y\left[\left(y-\mu^k\right)^2|k,j_*\right] &= \mathbb{E}_y\left[\left(y-\mu_{j_*}^k+\mu_{j_*}^k-\mu^k\right)^2|k,j_*\right] \\ &= \mathbb{E}_y\left[\left(y-\mu_{j_*}^k\right)^2|k,j_*\right]+\mathbb{E}_y\left[\left(\mu_{j_*}^k-\mu^k\right)^2|k,j_*\right] \\ &\quad +2\underbrace{\left(\mu_{j_*}^k-\mu^k\right)\mathbb{E}_y\left[\left(y-\mu_{j_*}^k\right)|k,j_*\right]}\end{aligned}$$

where we introduced the term $\mu_{j_*}^k$. Note that since under the expectation $\mathbb{E}_y[\bullet|k,j_*]$, the second term is invariant, and the third (underbraced) term vanishes, we have:

$$\underbrace{\mathbb{E}_y\left[\left(y-\mu^k\right)^2|k,j_*\right]}_{V_{k,j_*}^{tot}} = \underbrace{\mathbb{E}_y\left[\left(y-\mu_j^k\right)^2|k,j_*\right]}_{V_{k,j_*}^{W}} + \underbrace{\left(\mu_j^k-\mu^k\right)^2}_{V_{k,j_*}^{B}} \tag{2}$$

where V_{k,j_*}^{tot} is the *total variance*, V_{k,j_*}^{W} is the *within-client* variance and $V_B^k(j_*)$ is the *squared bias* (or squared distance) between the global mean and the client specific mean, all conditioned on the class label k (which can be genuine or impostor matching) and a particular claimed identity j_*. If one takes the expecation $\mathbb{E}_j[\bullet|k]$ on both sides of (2), then the two terms on the right-hand sides are known as variance and squared bias, respectively. The resultant two components constitutes the bias-variance decomposition of Krogh and Vedelsby [18].

Using the fact that j is a variable, and j_* being its realisation, we shall similarly define $V_{k,j}^{tot}$ to be the variable of V_{k,j_*}^{tot} (the latter being a realization). We shall also define $V_{k,j}^{W}$ and $V_{k_j}^{B}$ in a similar manner. We then propose to define the *client-specific* biometric menagerie index (BMI) (for a given client with index j) to be:

$$\mathrm{BMI}_{k,j} \equiv \frac{V_{k,j}^{B}}{V_{k,j}^{tot}}.$$

Likewise, the BMI index for the whole population of clients is:

$$\bar{\mathrm{BMI}}_k \equiv \mathbb{E}_j\left[\mathrm{BMI}_{k,j}|k\right]$$

with variance $\mathbb{E}_j\left[\left(\mathrm{BMI}_{k,j}-\bar{\mathrm{BMI}}_k|k\right)^2\right]$.

2.3 Properties of BMI

Because of (2), $V^B_{k,j} \leq V^{tot}_{k,j}$ since $V^W_{k,j}$ is always *positive*. As a result, the client-specific BMI index is always bounded:

$$0 \leq \text{BMI}_{k,j} \leq 1,$$

and so is the BMI index for the whole population

$$0 \leq \bar{\text{BMI}}_k \leq 1.$$

When $\text{BMI}_{k,j} = 0$, this implies that $V^B_{k,j} = 0$. This can only happen when the client-specific mean μ^k_j coincides with the global mean μ^k. If this is true over the whole population, then $\bar{\text{BMI}}_k = 0$. This strongly implies that the Doddington's menagerie contains only one species.

On the other hand, when $\bar{\text{BMI}}_k = 1$, this implies that $\mathbb{E}_j[V^W_{k,j}|k] = 0$. In other words, the total variance is fully determined by the expected squared bias $\mathbb{E}_j[V^B_{k,j}|k]$ (or the between class variance), and the expected within-client variance has little importance. This signifies the presence of widely different species in the Doddington's menagerie. In practice, we have never observed that the BMI takes on the extreme values of zero or one. Therefore, we emphasize that the BMI value should *always* be expressed along with its confidence interval (characterised by its variance).

Apart from the bounded property, BMI is also invariant to shifting and scaling. It is trivial to show the invariance of BMI to shifting. Suppose y is shifted by a constant. Then, its expected value will be shifted by the same amount. This constant simply gets canceled out when calculating the variance of the shifted score, because this operation involves only finding the difference between the shifted score and shifted expected value. Hence, all the terms in (2) are not affected by the constant shift. It follows that BMI is invariant to shifting.

BMI is also invariant to scaling. Suppose that y is scaled by a constant a, i.e., ay. Then, all the terms in (2) will be scaled by a^2. The resultant BMI will be:

$$\text{BMI}_{k,j} = \frac{a^2 V^B_{k,j}}{a^2 V^{tot}_{k,j}} = \frac{V^B_{k,j}}{V^{tot}_{k,j}}$$

hence invariant to the constant scaling.

We simulated four different possible scenarios giving different BMI values, all as a function of the following *client-specific* means and variances (arranged in a vector):

$$\boldsymbol{\mu} = [-7.00, -3.50, 0, 3.50, 7.00]' \text{ and } \boldsymbol{\sigma} = [0.20, 1.00, 3.00, 0.70, 4.00]'.$$

The results are shown in Figures 2 (a)–(d). The actual mean and variance used for each sub-figure are shown in the caption. In (a), all client-specific means are aligned to the global mean (zero in our case), so that the between-client variance is set to zero, hence, giving BMI=0. The other extreme, giving BMI=1 is shown in (d). This was done by decreasing the variance to very small values so that the between-client variance dominates the total variance. Figures (b) and (c) are two intermediate cases sampled from the range of possibilities.

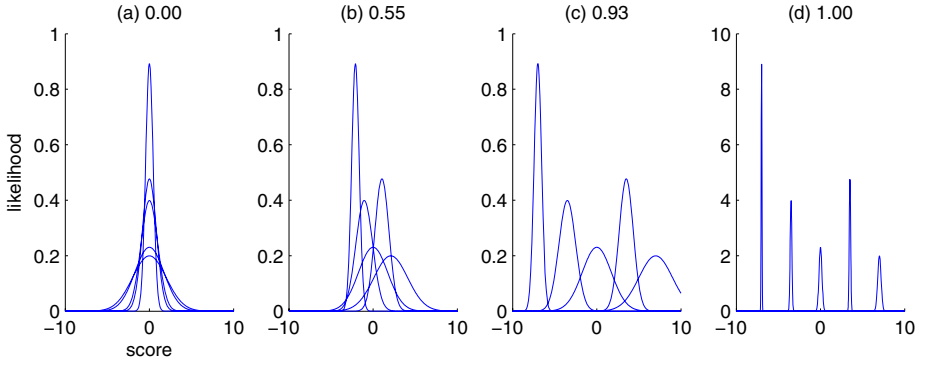


Fig. 2. Examples of client-specific distributions for various BMI values (in the subtitle). The following means and variances are used: (a) [$\boldsymbol{\mu} \times 0$, $\boldsymbol{\sigma}$], (b) [$\boldsymbol{\mu} \times 0.3$, $\boldsymbol{\sigma}$], (c) [$\boldsymbol{\mu}$, $\boldsymbol{\sigma}$], and (d) [$\boldsymbol{\mu}$, $\boldsymbol{\sigma} \times 0.01$].

2.4 Empirical Estimates

In the followings, we will work with the sample $y_{ij}^k \in \mathcal{Y}_j^k$ for any claimed identity j, conditioned on the class k, and $i = 1, \ldots, N^k$, and there are N^k samples. The estimated global and client-specific means are:

$$\hat{\mu}^k = \frac{1}{JN^k} \sum_{j=1}^{J} \sum_{i=1}^{N^k} y_{ij}^k \quad \text{and} \quad \hat{\mu}_j^k = \frac{1}{N^k} \sum_{i=1}^{N^k} y_{ij}^k$$

The estimated (squared) bias term is:

$$\hat{V}_{k,j}^B = (\hat{\mu}_j^k - \hat{\mu}^k)^2$$

The estimated total variance is:

$$\hat{V}_{k,}^{tot} = \frac{1}{N^k} \sum_{i=1}^{N^k} (y_{ij}^k - \hat{\mu}^k)^2.$$

Assuming all hypotheses to be equiprobable, the expected mean and variance of BMI index can be estimated as follow:

$$\widehat{\bar{\text{BMI}}}^k = \frac{1}{J} \sum_{j=1}^{J} \left(\frac{\hat{V}_B^{k,j}}{\hat{V}_{k,j}^{tot}} \right)$$

$$\widehat{var}(\text{BMI}^k) = \frac{1}{J} \sum_{j=1}^{J} \left(\frac{\hat{V}_B^k(j)}{\hat{V}_{tot}^k(j)} - \widehat{\bar{\text{BMI}}}^k \right)^2$$

The variance of BMI is a useful indicator of the confidence of the estimated BMI. The smaller the variance, the higher the confidence of the estimate.

3 Experiments

3.1 Data-Set

We use the publicly available[1] XM2VTS multimodal score-level fusion benchmark database [19] for this purpose. 13 sets of experiments were available, among which 6 (resp. 7) are on the speech (resp. face) biometrics. The distribution of match scores, in general, exhibits high central tendency except for two comparison subsystems which are based on Multi-layer Perceptrons (MLPs) having tangent hyperbolic $\tanh$) output. For these systems, their outputs are mapped into a linear scale using the inverse of $\tanh$, hence, denoted by "MLPi". This one-to-one transformation is non-linear and is essential to guarantee the central tendency of the match scores. Since the BMI index is invariant to shifting and scaling, no further prepossessing is needed, and the experiments can be compared directly using BMI.

There are 200 client models that are assessed on two data sets, termed dev (for development) and eva (for evaluation) sets. There are 3 genuine samples per client for the dev set under Lausanne Protocol I (LP1) and 4 for Lausanne protocol II (LP2). These protocols define what data the expert systems should use for enrolling the clients. Although these protocols entail changes in the number of genuine match scores for the dev set, they remain the same for the eva set, which is two genuine match scores per client model, hence a total of 200×2 genuine match scores for eva but 200×3 for dev LP1 and 200×4 for dev LP2.

It is worth noting that the impostor population of dev and eva sets are different, thus, $\mathcal{J}'_{dev}$ (the former) consists of 25 impostors and $\mathcal{J}'_{eva}$ (the latter) consists of 70 impostors. Each impostor contributes 8 samples. This results in 25×8 impostor scores for the impostor dev and 70×8 for the impostor eva set. Using different impostor population is beneficial for testing the stability of BMI under changes in impostor population.

3.2 Results

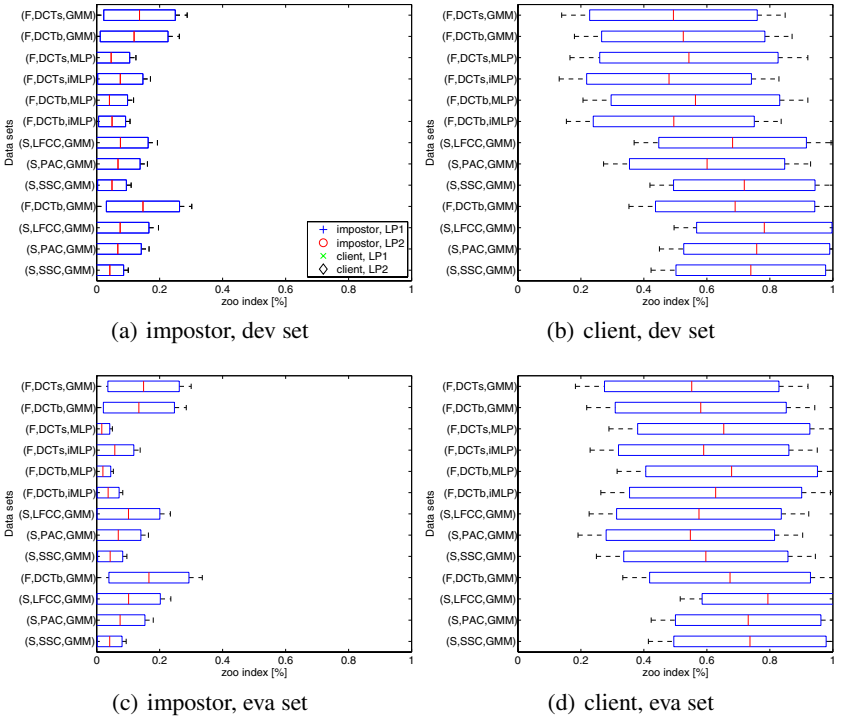
Figure 3 shows the BMI values and its variance for the 13 score data sets conditioned on the class label (genuine or impostor), and on the data set used, i.e., dev or eva. We can observe the following:

(i) The variances of impostor BMI match scores are generally much higher than their genuine counter parts. This is expected since there are many more impostor match scores than the genuine match scores. For instance, $200 \times 75 \times 8$ impostor match scores versus 200×2 genuine match scores for the eva set. As a result, the impostor BMI can be estimated with *higher* confidence.

(ii) The *expected* genuine match scores have, in general, higher BMI values than their corresponding impostor match scores.

(iii) The BMI values between the development and the evaluation sets are somewhat consistent in pattern. In order to verify this futher, we plotted BMIs of the dev set versus their corresponding values on the eva set in Figure 4.

[1] http://personal.ee.surrey.ac.uk/Personal/Norman.Poh/web/fusion



(a) impostor, dev set (b) client, dev set

(c) impostor, eva set (d) client, eva set

Fig. 3. The BMI of 13 systems calculated on the four data sets formed by the following dichotomies: client vs impostor, dev vs. eva. Each figure shows 13 BMI values – with the first 9 (from top to bottom) from LP1 systems whereas the remaining 4 from LP2 – conditioned on dev and eva sets. Each BMI was estimated from 200 client-specific BMI values (since there are 200 client models) in each system.

Concerning observation (iii), it is worth recalling that the impostor populations in the dev and the eva sets consist of *different* subjects (persons), and that the client models (in the gallery) remain the same. By comparing BMI with different impostor data sets, this experiment confirms that BMI is indeed *insensitive* to the choice of impostor population.

The client BMI for LP1 is *inconsistent* between dev and the eva sets, with correlation as low as -0.11. In comparison, the client BMI for LP2 is more consistent (giving correlation as high as 0.95; see Figure 4 caption). This is because the genuine match scores between the dev and eva sets for LP1 are *biased*. In LP1, the dev score sets were derived from the same session (in the same visit) of data set as the one used to build the base-classifier model, but this is not the case for LP2 (where a different session of data was used). As a result, for LP1, the genuine match scores in dev is *positively biased* (resulting in better recognition than expected). For LP2, there is no such bias, since the dev and eva sets contains similar *cross session* variability, with dev scores derived from the third session and eva from the fourth session. Recall also that in both LP1 and LP2, the data used to generate the eva score set is the same (the forth session/visit). Two

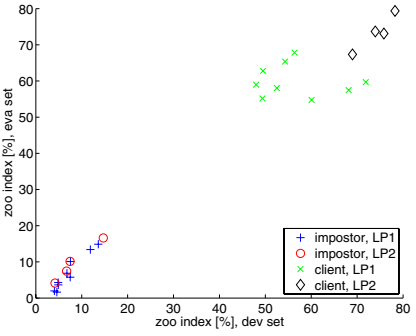


Fig. 4. Comparison of BMI calculated on the development set (X-axis) and the same statistic calculated on the evaluation set (Y-axis) for (a) impostor BMI (on LP1 data set), (b) impostor BMI (LP1), (c) client BMI (LP1), and (d) client BMI (LP1). (a) and (b) consists of 9 systems and LP2 consists of 4 systems. The correlation of BMI between the dev and eva sets for these four data sets are 0.96, 0.98, -0.11 and 0.95, respectively.

consecutive sessions were separated by about one month's interval, but this is certainly not consistent across all the users. This observation highlights the importance of using *unbiased* (cross-session) data set to estimate the BMI values.

4 Conclusions

In this paper, we proposed a biometric menagerie index (BMI), based on the expected within-client variance and expected squared bias (or between-client variance). The current study is somewhat preliminary; the following extensions are currently being investigated:

- **Wider range of biometrics:** While the current study is limited to face and speech, more biometric modalities will be considered
- **Multivariate extension of BMI:** The current study is limited to calculating BMI on a single modality. A future study will include an analysis of user-specific class-conditional BMI *across different modalities/systems*. This will allow to resolve whether two BMI values, each obtained from their respective user model, are correlated or not. If they are correlated, then the presence of lamb may be attributed to data of insufficient quality. On the other hand, if the BMI values are not correlated, then the presence of lamb may not be caused by the data, but by the choice of classifier or feature representation, which may systematically fail to discriminate a certain group of users.
- **Pre-processing techniques:** The current study relies on the central tendency of class-conditional match scores, essentially relying on the good estimates of the first and second order moments. It is therefore vital to ensure that the data adhere to this assumption. A promising technique to achieve this is by means of Gaussian Copula [20,21], which has been used successfully to estimate the effective biometric sample size. An extension of BMI to multimodal biometrics by means of multivariate Gaussian Copula is straightforward.

Acknowledgement

This work was supported partially by the advanced researcher fellowship PA0022_121477 of the Swiss National Science Foundation and by the EU-funded Mobio project (www.mobioproject.org) grant IST-214324.

References

1. Doddington, G., Liggett, W., Martin, A., Przybocki, M., Reynolds, D.: Sheep, Goats, Lambs and Wolves: A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation. In: Int'l. Conf. Spoken Language Processing (ICSLP), Sydney (1998)
2. Yager, N., Dunstone, T.: Worms, chameleons, phantoms and doves: New additions to the biometric menagerie. In: IEEE Workshop on Automatic Identification Advanced Technologies, June 2007, pp. 1–6 (2007)
3. Poh, N., Kittler, J.: A Methodology for Separating Sheep from Goats for Controlled Enrollment and Multimodal Fusion. In: Proc. of the 6th Biometrics Symposium, Tampa, pp. 17–22 (2008)
4. Hicklin, A., Ulery, B.: The myth of goats: How many people have fingerprints that are hard to match? Tech. Rep. NISTIR 7271, National Institute of Standards and Technology (2005)
5. Wittman, M., Davis, P., Flynn, P.J.: Empirical studies of the existence of the biometric menagerie in the frgc 2.0 color image corpus. In: Conf. on Computer Vision and Pattern Recognition Workshop, June 2006, p. 33(2006)
6. Une, M., Otsuka, A., Imai, H.: Wolf attack probability: A theoretical security measure in biometric authentication systems. IEICE-Transactions on Info and Systems E91-D(5), 1380–1389 (2008)
7. 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)
8. Pierrot, J.-B.: Elaboration et Validation d'Approaches en Vérification du Locuteur, Ph.D. thesis, ENST, Paris (September 1998)
9. Chen, K.: Towards Better Making a Decision in Speaker Verification. Pattern Recognition 36(2), 329–346 (2003)
10. 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)
11. Genoud, D.: Reconnaissance et Transformation de Locuteur, Ph.D. thesis, Ecole Polythechnique Fédérale de Lausanne (EPFL), Switzerland (1998)
12. Auckenthaler, R., Carey, M., Lloyd-Thomas, H.: Score Normalization for Text-Independant Speaker Verification Systems. Digital Signal Processing (DSP) Journal 10, 42–54 (2000)
13. 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)
14. 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)
15. Poh, N., Kittler, J.: On the Use of Log-likelihood Ratio Based Model-specific Score Normalisation in Biometric Authentication. In: Lee, S.-W., Li, S.Z. (eds.) ICB 2007. LNCS, vol. 4642, pp. 614–624. Springer, Heidelberg (2007)
16. Poh, N., Kittler, J.: Incorporating Variation of Model-specific Score Distribution in Speaker Verification Systems. IEEE Transactions on Audio, Speech and Language Processing 16(3), 594–606 (2008)

17. Grother, P., Tabassi, E.: Performance of Biometric Quality Measures. IEEE Trans. Pattern Analysis and Machine Intelligence 29(4), 531–543 (2007)
18. Krogh, A., Vedelsby, J.: Neural Network Ensembles, Cross-Validation and Active-Learning. In: Advances in Neural Information Processing Systems, vol. 7 (1995)
19. Poh, N., Bengio, S.: Database, Protocol and Tools for Evaluating Score-Level Fusion Algorithms in Biometric Authentication. Pattern Recognition 39(2), 223–233 (2005)
20. Song, P.X.-K.: Multivariate dispersion models generated from gaussian copula. Scandinavian Journal of Statistics 27(2), 305–320 (2000)
21. Dass, S.C., Zhu, Y., Jain, A.K.: Validating a Biometric Authentication System: Sample Size requirements. IEEE Trans. on Pattern Analysis and Machine 28(12), 1319–1902 (2006)