An Algorithm to Estimate the Biometric Performance Change Over Time

Norman Poh, Josef Kittler, Chi-Ho Chan, and Medha Pandit

Abstract

We present an algorithm that models the rate of change of biometric performance over time on a subject-dependent basis. It is called "homomorphic users grouping algorithm" or HUGA. Although the model is based on very simplistic assumptions that are inherent in linear regression, it has been applied successfully to estimate the performance of talking face and speech identity verification modalities, as well as their fusion, over a period of more than 600 days. Our experiments carried out on the MOBIO database show that subjects exhibit very different performance trends. While the performance of some users degrades over time, which is consistent with the literature, we also found that for a similar proportion of users, their performance actually improves with use. The latter finding has never been reported in the literature. Hence, our findings suggest that the problem of biometric performance degradation may be not as serious as previously thought, and so far, the community has ignored the possibility of improved biometric performance over time. The findings also suggest that adaptive biometric systems, that is, systems that attempt to update biometric templates, should be subject-dependent.

Index Terms

Biometric authentication, score normalization

I. Introduction

With the increasing need for identity authentication in our daily life – from making purchase online and unlocking personal devices to accessing secure premises – the biometric technology certainly has an important role. Although physical lock, PIN and gesture (which is common for mobile devices) offer convenient solutions, only biometrics-based authentication can provide the ultimate means of validating the identity credential. The use of biometric authentication entails many questions, especially when it is used over a long period of time and in different locations: Will the system performance degrade over time? What kind of metrics can best describe the performance over time? How will its performance be affected by different acquisition environments, such as the office environment, public locations, and outdoor?

In this study, we attempt to gain understanding whether or not a change in performance of a biometric system is *subject-dependent*. We will not attempt to explain the cause of performance change; we know that it is not only age-related, but is also dependent on biometric sample quality and habituation. As will become clear, measuring the change in biometric performance is challenging, let alone explaining the causes of this change. However, we intend to challenge the common assertion that the performance of a biometric system systematically degrades overtime.

In fact we will show that the performance change is subject dependent. For some individuals it degrades, but for others it improves and for some it remains stable. Our approach to answering the above question is, first of all, to model the performance change in a subject-dependent manner. Thus, for each enrollee, we attempt to model his/her performance change. Then, the performance change is classified as upward, stable, or downward trend. Once the subjects have been classified, it is now possible to aggregate their matching scores at the beginning and at the end of a study period and compare the performance metrics derived from these two points in time.

We validated our approach on the MOBIO database. It contains both talking face and speech biometrics of 150 enrollees collected using a Nokia mobile phone, covering over 600 days of recording. The 600-day time span is sub-divided into three time periods each having a length of 200 days. They constitute the "initial ($<$ 200 days), "middle" (between 200 and 400 days), and "end" ($>$ 400 days) time periods. Our experimental results show that for some subjects, the system does better in the initial period than in the end period. This implies that the system gets worse over time. This is consistent with the literature [1], [2], including our own work [3]. However, we also observed that there exists a set of subjects for whom the system does better in the end period than the initial period. This means that the system improves with use. To our knowledge, this has never been observed or reported in the literature. A logical explanation of this is habituation. As subjects get used to the system, they can provide biometric samples of higher quality or in a consistent way.

Whatever the underlying causes of performance change, that is, ageing, biometric sample quality, or familiarity with a particular biometric device or mode of acquisition, our study entails several implications. First, a subject-dependent strategy has to be adopted when maintaining a biometric system. Second, the study of biometrics ageing demands more research as any degradation in performance may be compensated by the subject's familiarity. Furthermore, it appears to be important to

Norman Poh is with Department of Computing; and Josef Kittler, Chi-Ho Chan and Medha Pandit are with CVSSP, University of Surrey, Guildford, GU2 7XH, Surrey, UK. E-mails: {n.poh,j.kittler, chiho.chan, m.pandit}@surrey.ac.uk

separately study the effect of ageing and the effect of degradation due to mismatched acquisition conditions, and/or devices. Finally, more longitudinal biometric databases are required to understand the effect of ageing.

The rest of the paper is organised as follows: Section II motivates the need for subject-specific performance characterisation. Section IV then extends the notion of subject-specific performance to grouping the subjects by their performance evolution over time. A novel algorithm known as "homomorphic user grouping algorithm" will be introduced. The remaining sections present the MOBIO database (Section V) and the experimental results (Section VI). This is followed by conclusions in Section VII.

II. Background Literature on User-specific Performance Characterisation

Doddington *et al* [4] defined four categories of animals based on the mean of the subject's genuine or impostor scores; they are:

- **sheep**: subjects who can easily be recognised – they have high mean genuine scores;
- **goats**: subjects who are particularly difficult to recognise – they have low mean genuine scores;
- **lambs**: subjects who are easy to imitate – they have high mean impostor scores; and
- **wolves**: subjects who are particularly successful at imitating others – they consistently produce high impostor scores for all enrollees.

Goats contribute significantly to the False Reject Rate (FRR) of a system while wolves and lambs increase its False Acceptance Rate (FAR). For a lamb, the subject's template has the tendency of producing high nonmatch scores with any biometric sample belonging to other subjects. This higher-than-normal nonmatch score is captured by the higher-than-normal nonmatch *mean* score statistic. As a result, the presence of lambs will induce many high false acceptance instances, leading to high FAR. The wolves are statistics centred on the zero-effort nonmatch (impostor) attack. A biometric sample is classified as a wolf if the sample consistently produces high nonmatch scores when matched against *any* template. The definition of wolves captures this notion by computing the mean nonmatch score (when matched with all the available templates) and then identifying those samples that have the highest mean nonmatch score. Indeed, it is conceivable to identify wolves using this simple statistics in order to cause a biometric system to fail with high false acceptance instances. This attack is known as the wolf's attack, which also leads to high FAR.

Yager and Dunstone [5] further distinguish four other semantic categories of users by considering both the genuine and impostor matching 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 matching scores.

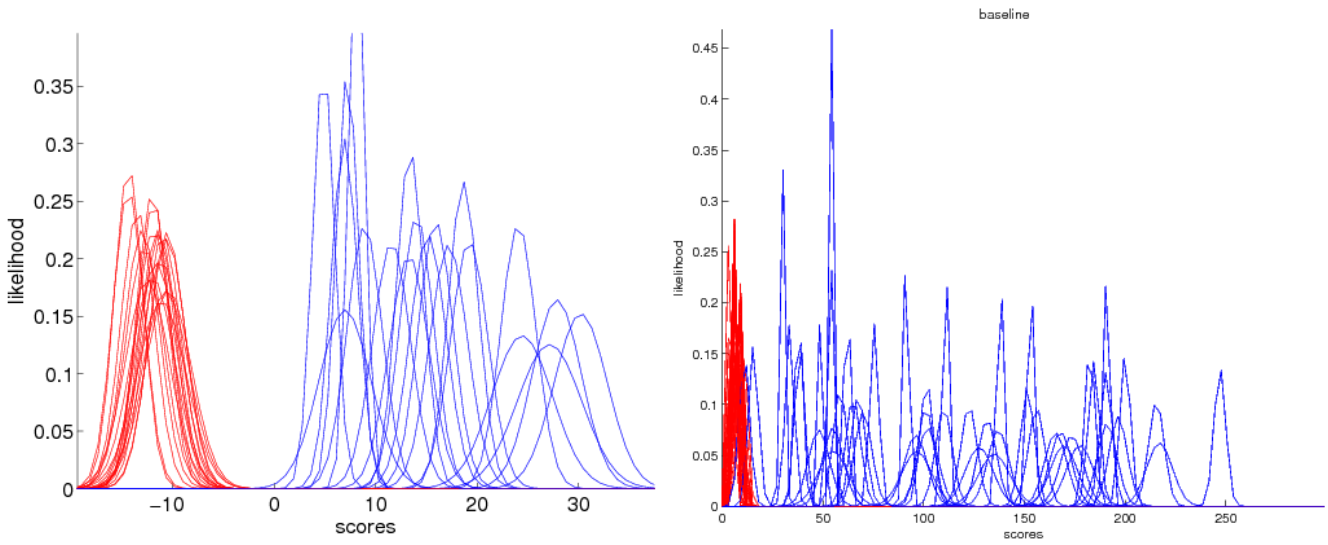
Poh and Kittler [6] further consider the second order moments from which several client-specific class-separability criteria are derived. Among the six criteria studied, three are found to be useful to rank the user models according to their performance, hence providing a means to separate well-behaved models from the badly behaved ones (in terms of performance). These criteria are the Fisher ratio [7], the F-ratio [8] and the d-prime statistics [9].

Referring to Doddington's menagerie, sheep are characterized by high genuine (similarity) matching scores whereas goats are characterized by low genuine matching scores. Lambs have similar matching problems as goats, by having high impostor matching scores. Finally, wolves are persons who can consistently give high impostor similarity scores when matched against all the references (i.e., enrolled templates/models in the gallery). While sheep dominate the population of client models, goats (resp. lambs) constitute only a small fraction of the population. However the latter category constitutes a disproportionately large portion of false rejection (resp. acceptance) errors.

Although the original Doddington's study was applied to speaker verification, the same phenomenon was independently observed in [6] using the face, fingerprint and iris biometric modalities; [10] using the fingerprint modality; [11] using the face modality; and many others, e.g. [5]. Using finger-vein and fingerprint as case studies, Une et al [12] 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 recognized that fine-tuning the system parameters (including feature extraction and classifier or distance matching parameters) and the decision threshold for each individual reference (model, classifier) can greatly boost the recognition (identification and verification) performance further. For instance, *lowering* the similarity decision threshold (in relation to a globally pre-set value) for the goats is likely to compensate for their disproportionately high false rejection errors. Similarly, *increasing* the decision threshold for the lambs will also compensate for their disproportionately high false acceptance errors. This strategy is called client (model/template) specific decision. Examples of such strategies abound: [13], [14], [15], [16], [17]. Rather than adjusting the thresholds, one can instead transform the matching score distribution. This alternative strategy is called *client-specific score normalization*. Examples are Z-norm [18], F-norm [8], EER-norm [19] and model-specific log-likelihood ratio (LLR)-based normalization [20]. Both categories of approaches have been discussed and summarized in [21].

Clearly, a fundamental understanding of Doddington's menagerie is important for designing and optimizing 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. Yet, we know that it is certainly dependent on the choice of



(a) Face biometrics (XM2VTS)

(b) Fingeprrint biometrics (Biosecure DS2)

Fig. 1. User-specific class-conditional score distributions of (a) a speech verification system taken from the XM2VTS database and (b) a fingerprint verification system taken from the Biosecure DS2 data set. Shown here are the distributions of 20 enrollees. The right clusters (in blue) are for the genuine class whereas the left ones (in red) are for the impostor class. For (a), only the scores associated with 20 randomly selected enrolled subjects out of the total of 200 are used in order to avoid cluttering. Therefore, there are 20 pairs of score distributions in (a).

biometric device, the result of the user's interaction with the device and the acquisition environment, hence related to the quality of biometric samples. Although Grother and Tabassi [22] 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. As has been pointed out by Ross *et al* [23] and Wittman *et al* [11], there is an important causal relationship between the quality of reference and the Doddington's menagerie.

Hiclin and Ulery [10] argued that the presence of these animals may be due to the poor quality of template. However, Wittman *et al* [11] showed that even in a very controlled scenario, subjects still show some tendency towards a particular animal species. Teli *et al* [24] further suggested that goats and lamb always exist for a given data set but a subject is likely to change animal species membership when the matching algorithm changes.

The above literature suggests that there is definitely an interplay between subject, biometric sample quality, and matching algorithms that affect biometric performance. The subject-dependency of biometric performance also suggests that tailoring a biometric system by adjusting its decision threshold to each subject, or otherwise normalising the score distribution using subject-specific score normalisation can improve the system performance [21].

In order to motivate the need for user-specific performance characterisation, we fitted a Gaussian distribution to the match scores produced by a template (or reference model) *of each subject*, for genuine or impostor comparisons, respectively. For this exercise, we shall use speech and fingerprint biometrics in order to illustrate that the biometric menagerie is a phenomenon that permeates all biometrics. For the speech biometrics, we shall use a subset of a face matcher evaluated on the XM2VTS database [25] whereas for the fingerprint biometrics, we will use the Biosecure DS2 dataset [26]. The choice of Gaussian distribution is dictated by the small sample size of the data, especially the genuine matching scores (to be further explained in Section III). The result is shown in Figure 1(a). This figure clearly shows that the client-specific genuine and impostor score distributions are very different from one enrolled subject to another. It further points to the need of characterising the system performance on a subject-by-subject basis. Therefore, it is important to optimize the decision threshold to each enrollee instead of having a global decision threshold.

III. Scope and Assumptions

A. Scope

The biometric menagerie phenomenon may not entirely be caused by the user himself or herself. Instead, the phenomenon is certainly associated with the template or (statistical) reference model that represents the user. This is where the quality of the template may have impact on the recognisability of the user which explains why the system performance is subject-dependent.

Suppose that a user has four biometric samples, T_1, T_2, T_3, and T_4. Let $M(T_i, T_j)$ represent the matching score between template T_i and query T_j. Since the matching scores $M(T_1, T_2)$ and $M(T_1, T_3)$ are generated from template T_1, they are likely to be dependent on each other. This dependency is exploited by a user-specific score normalisation or fusion in order to enhance the accept/reject decision.

This paper does not consider the case where two different templates are used to represent a user. Therefore, we do not offer an explanation as to whether or not there is a dependency (positive correlation) between $M(T_1, T_2)$ and $M(T_3, T_4)$. Answering this question would address whether or not biometric menagerie is indeed user-dependent. Although this is an important research question, we do not intend to study this scientific problem in-depth.

On the contrary, we are interested to find out the generalisation ability of a user-specific strategy when the same template is used over a period of time. For instance, we want to find out if the dependency between $M(T_1, T_2)$ and $M(T_1, T_3)$ still holds when T_2 and T_3 have been collected with a gap of several weeks or months apart, and the template T_1 has been kept the same throughout this period. This use-case scenario is of practical importance because this is how nearly all biometric systems operate. This is the main research topic being pursued here.

Another line of research concerns solutions mitigating the gross subject dependent score variability via client-specific score normalisation such as F-norm, Z-norm, etc. Although we do not consider these schemes here, as will become clear toward the end of this paper, our findings will have implications on how these schemes are implemented.

B. Assumption

All literature on Doddington's zoo or biometric menagerie needs to characterise client-specific score distributions. The most common assumption is that each client-specific score distribution is normally distributed. Our study is no exception here. It is, therefore, imperative to examine this assumption closely.

Let y be a matching score and ω be the class of matching, which indicates whether the matching score is a non-match or a match, $\omega \in \{0, 1\}$. Furthermore, let the class-conditional distribution for the j-th enrolled subject be denoted by $p(y|\omega, j)$. When this distribution is assumed to be Gaussian, it is described by $p(y|\omega, j) = \mathcal{N}(y|\mu_j^\omega, (\sigma_j^\omega)^2)$ where μ_j^ω is the mean or the first order moment; and $(\sigma_j^\omega)^2$, the second order moment or variance. The Gaussian assumption implies that $p(y|\omega, j)$ is completely specified by the first two orders of moment. One way to relax this assumption is by approximating the true underlying, unknown distribution $p(y|\omega, j)$ with a sufficient number of moments.

Of course, if the form of distribution is known, then, the problem reduces to finding the correct parameters of the distribution for a given set of observed scores. In reality, the form of score distribution is conditional upon the acquisition conditions of the query samples. This implies that a single parametric form of (class-conditional) score distribution may be insufficient to characterize the true underlying distribution. For instance, the output of Daugman's iris recognition system is known to follow a Binomial distribution. However, a comparison between a pair of iris depends on the number of bits actually used in the comparison [27]. In this case, the matching score can be normalised via simple rescaling involving this parameter – the number of bits used for comparison. For more complex systems such as those that are based on statistical models, e.g., Gaussian Mixture Model for speaker verification [28], or neural network for face recognition [29], it is extremely difficult to define a parametric formula to compute a matching score that is invariant to the acquisition conditions. As a result, the (class-conditional) score distributions cannot be adequately characterised by a simple parametric form.

In our client-specific study here, not only the parametric form of (class-conditional) score distributions is generally unknown, the number of available samples to estimate the distributions is also very limited. For non-match (impostor) scores, one typically has a couple of hundreds of scores whereas for the match (genuine) scores, there could be as few as one score but typically not more than five. Recall that the size of the match score set depends only on the number of genuine samples. If there are g samples for a subject, then there will be at most $g - 1$ scores because one of the samples is reserved as the template and the remaining samples are treated as query samples.

Faced with the limited number of samples, in spite of the fact that the actual class-conditional score distributions are unlikely to follow a single parametric form, it is certainly justifiable to approximate $p(y|\omega, j)$ using the first two moments.

This approximation will hold only when the score distribution is centred around the mean. There are a number of techniques that can improve the central tendency of scores. One way is to use the Box-Cox transform [30]. This technique was used to improve the central tendency of both the class-conditional scores jointly in our previous work [31]. Another method is to use the generalized logit transform. If the output of a biometric matcher is bounded in $[a, b]$, the following order-preserving transformation is recommended [32]:

$$y' = \log\left(\frac{y-a}{b-y}\right) \tag{1}$$

Note that when $a = 0$ and $b = 1$, this reduces to the logit transform that is commonly used in statistics.

In summary, the assumption that the class-conditional score distributions $p(y|\omega, j)$ is Gaussian, as adapted in the majority of papers in the literature on biometric menagerie and score normalisation (e.g., Z-norm and T-norm), is motivated by what could practically work, rather than by theory. Indeed, the popularity of Z-norm and T-norm [18] implies that the Gaussian assumption works well in practice.

In the next section, we augment the subject-specific characterisation of matching scores with the time notion.

IV. Our Framework: A Homomorphic Users Grouping Algorithm

In the previous section, a subject's biometric performance is characterised by four parameters, namely, the mean and standard deviation of his/her impostor and genuine scores, i.e., $\{\mu_j^\omega, \sigma_j^\omega\}$ for $\omega \in \{0,1\}$ and each and every user $j \in \mathcal{J}$. In order to consider the time domain, we fit a regression line to a time-series of matching scores, for genuine ($\omega = 1$) or impostor ($\omega = 0$) matching. Let the score time-series for subject j be composed of samples $\{y_{j,t}^\omega\}$ collected at time $t = 1, \ldots, T$. This time-series is approximated by:

$$y_{j,t}^\omega \approx f(\mathbf{b}_j, t) + \eta_t$$

where η_t is noise and $f(\mathbf{b}_j, t)$ is a polynomial regression function of degree D for each of the two possible classes of matching scores, ω; it has the following form:

$$f(\mathbf{b}_j, t) = [b_D^{(j)}, b_{D-1}^{(j)}, \ldots, b_0^{(j)}]\,[t^D, t^{D-1}, \ldots, t^0]^\intercal$$

where $\mathbf{b}_j$ is a vector of coefficients with $D+1$ elements indexed by $\{b_d^{(j)}\}$ and the symbol $\intercal$ denotes a vector transpose operation.

The use of regression implies a number of important assumptions, that is, (1) linearity in the parameter space $[t^D, t^{D-1}, \ldots, t^0]$, (2) independence of the observations, (3) constant variance across the period observed $\{t\}$, and (4) normality of the distribution of fitting errors. These assumptions are worth commenting here. First, the use of polynomial degree D relaxes the assumption that the scores $\{y_{j,t}^\omega\}$ change linearly over time. However, the use of higher D is simply infeasible when one has very few observations $\{y_{j,t}^\omega\}$. This imposes on us a realistic compromise of using relatively small degree of freedom.

The second assumption implies that knowing one particular value of $y_{j,t}^\omega$ does not allow us fully to predict the value of another. This assumption is not entirely satisfied considering that if two matching scores come from biometric samples obtained from the same session (collected seconds apart), they are likely to be correlated. However, if two biometric samples are obtained from two sessions that are weeks apart, then, their matching scores are likely to be less correlated. Hence, there is some weak temporal correlation among matching scores.

Given the few observations in $\{y_{j,t}^\omega\}$, the time-dependent variance is hard to estimate. However, even if it could be estimated, the variance is unlikely to be constant over time. Consider a biometric authentication application for mobile devices. Since the acquisition environments are likely to be different from one session to another, for instance, in two different locations, the noise as observed in the matching scores is likely to be different, too. Hence, in general, the assumption of constant variance is likely to be violated.

Finally, as the above example shows, the normality assumption of the distribution of error (with zero mean and constant variance) for the observed $\{y_{j,t}^\omega\}$ is unlikely to be true. However, this assumption will be required in order to derive confidence intervals around the fitted curve.

All the above assumptions appear to suggest that regression is an unlikely candidate for our choice. However, we have to adopt a rather pragmatic view of data modelling that is best expressed by George Box who famously quoted that "All models are wrong; some models are useful." Our view is that any regression models are likely to violate one or more of the assumptions above. We need to understand the limitations of the model and at the same time, recognise that the paucity of data does not allow us to use complicated models. Our approach is to use a linear regression model to approximate the score trend which we then use to approximate the evolution of biometric error over time for each subject. The final use of the model is merely to identify if a subject's performance increases, decreases, or remains stable over time. The fitted models are not used to quantify performance, for instance. Therefore, for our purpose of estimating biometric performance trend, an approximate model is deemed to be acceptable.

The estimated regression line $f(\mathbf{b}_j, t)$ gives us the expected value $\mu_{j,t}^\omega$ as well as its spread $\sigma_{j,t}^\omega$ (thanks to assumption (4) above) at time t for each of the matching classes $\omega \in \{0,1\}$. From these four time-varying parameters, an instantaneous Equal Error Rate (EER) at time (EER_t) can be calculated as [33]:

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

where

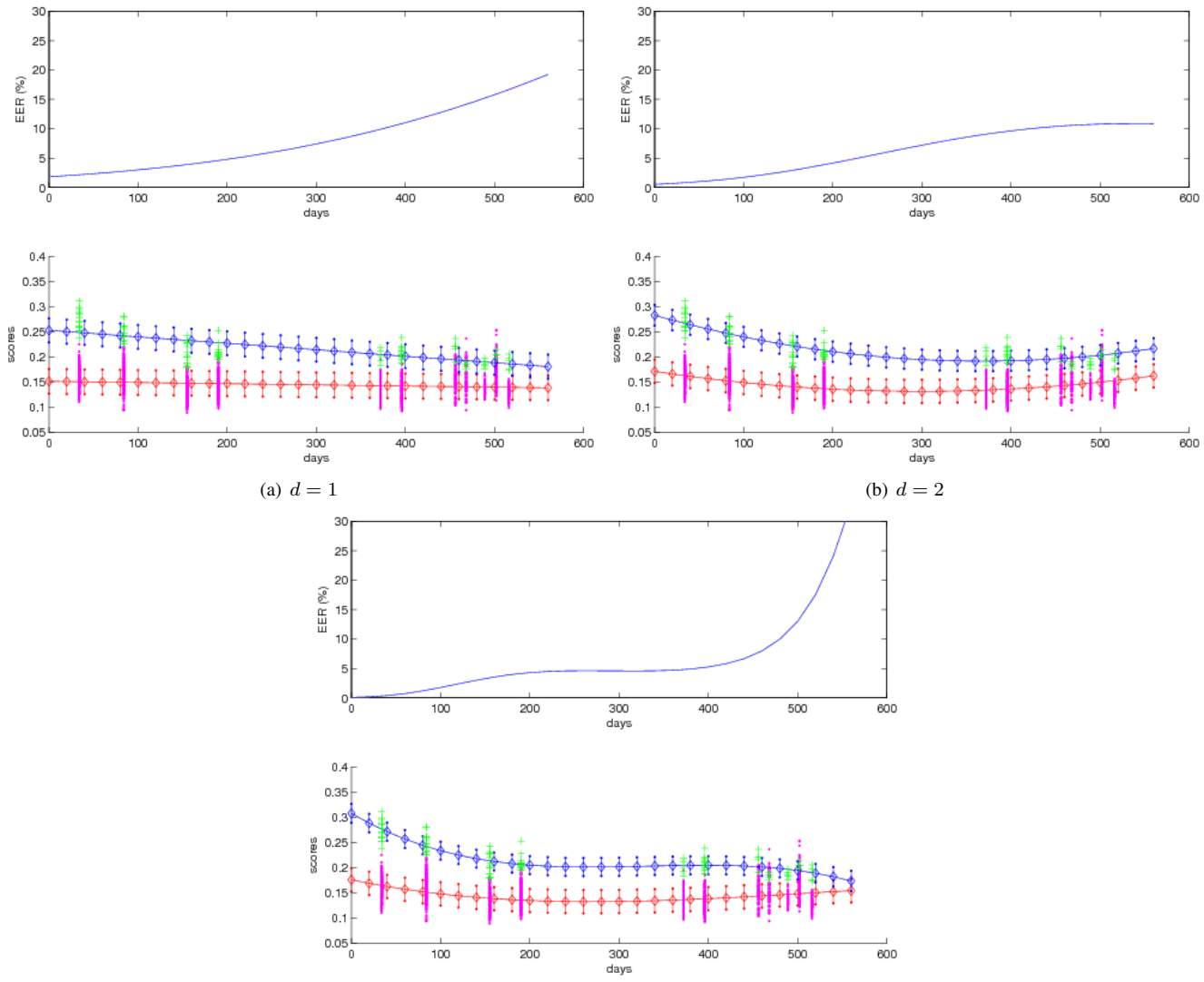
$$\text{F-ratio}_{j,t} = \frac{\mu_{j,t}^1 - \mu_{j,t}^0}{\sigma_{j,t}^1 + \sigma_{j,t}^0}, \tag{3}$$

and the error function is defined as:

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp\left[-x^2\right] dx. \tag{4}$$

Throughout this paper, we will use polynomial regression. A polynomial function of degrees 1, 2 and 3 have been fitted to a subject's score time-series and the results are shown in Figure 2.

As can be observed, while the different choices of the degree of freedom can affect the magnitude of EER, it has little impact on the overall trend which is an upward trend in this case. Since our ultimate aim is the latter, that is, to detect upward,



(a) $d = 1$ (b) $d = 2$

(c) $d = 3$

Fig. 2. Fitting with different degrees of freedom. While the choice of degree of freedom is difficult, it is obvious that Equal Error Rate (EER) increases over time for this subject.

downward, or stable trends, we will use polynomial degree 1. This reduces any unnecessary risk of over-fitting whilst at the same time addresses our need.

The derived subject-specific F-ratio time-series is then further subject to a linear regression fitting so that the gradient so obtained can be used to characterise the performance trend.

$$\text{F-ratio}_{j,t} \approx f(\mathbf{a}_j, t),$$

where

$$f(\mathbf{a}_j, t) = a_j^1 t + a_j^0.$$

and a_j^i for $i \in \{0, 1\}$ are elements of $\mathbf{a}$, and j is the index of each subject j. Since F-ratio is inversely and non-linearly proportional to EER, an increasing F-ratio time-series actually implies better performance (see Figure 3). To detect the trend of subject-specific F-ratio over time, we need only to infer the value of a_j^1, and we do so for all subjects $j \in \mathcal{J}$.

The next step consists of partitioning the subjects. A simple way is to partition them at a fixed percentile in the empirical cumulative distribution function of $\{\beta \in a_j^1 | \vee_j\}$ presented in the right of the last row in Figure 8. For this purpose, we assign the subjects who have a_j^1 between 0% and 20% percentiles to the first partition, namely Group A, those who have a_j^1 between 20% and 40% percentiles to the second partition, namely Group B, and so on, until the fifth partition of subjects, namely Group E who have a_j^1 between 80% and 100% percentiles. Therefore, we expect Group A should have *increasing* EER over time because they all have negative a_j^1 values. Conversely, we expect Group E to have *decreasing* EER over time because they all have positive a_j^1 values.

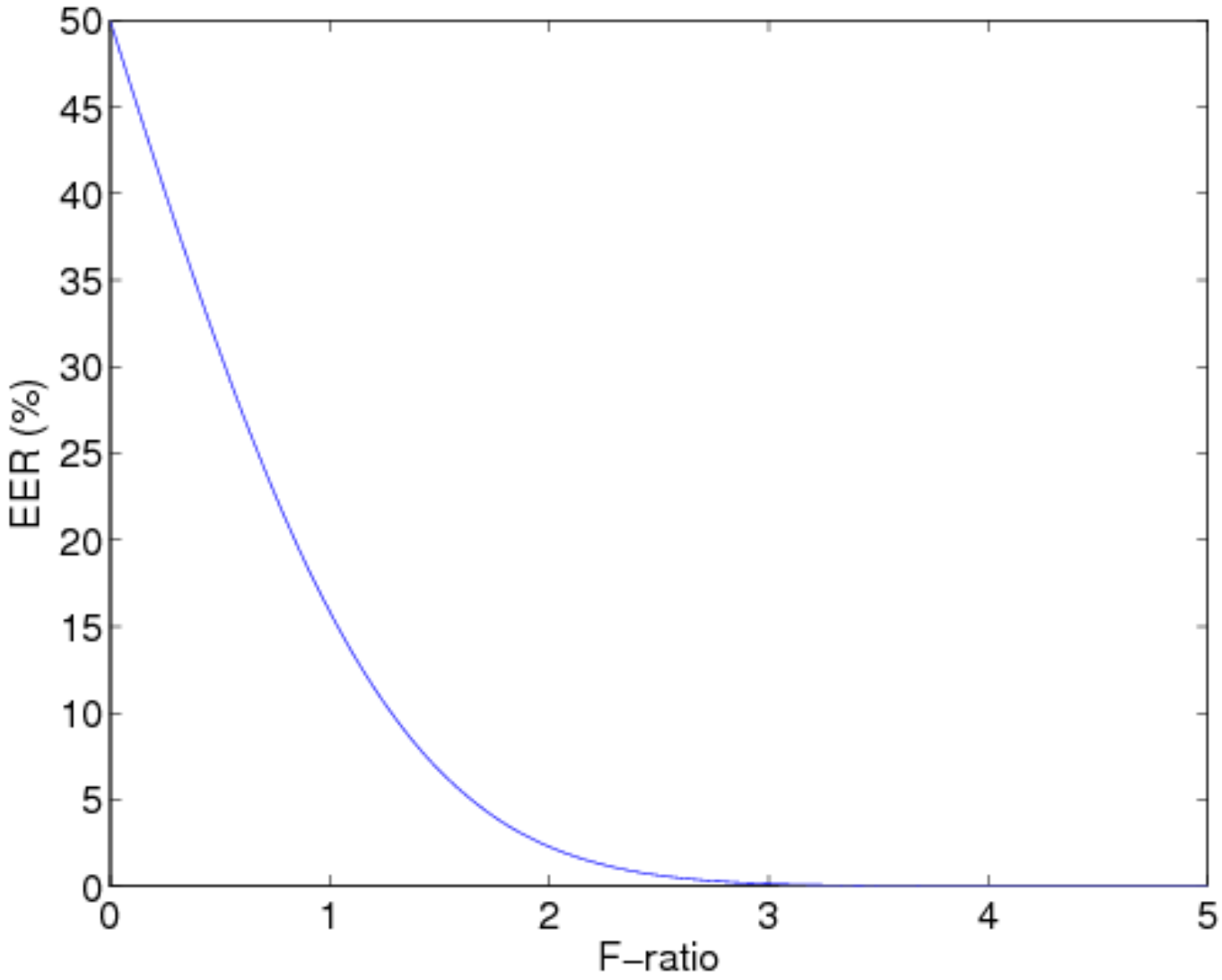


Fig. 3. The relationship between F-ratio and EER.

In order to validate the partitioned subjects, we need to plot the EER of subjects in the first and the fifth partition at different time periods. The MOBIO database that we used cover a period of roughly 600 days. It is convenient to divide the time into three periods, each lasting 200 days. Let us define these three time periods as "initial, "middle", and "end" periods as discussed before.

For each group and each time period, one can calculate the EER values. We would expect that for Group A, the "end" period EER is higher than the "initial" period, hence, providing an indication of performance degradation. This would be consistent with the findings in the literature.

However, for Group E, the findings that the EER calculated for the "end" period is *lower* than the "initial" period is rather "surprising". The results of our findings will further be discussed in Section VI.

The algorithm above is called "homomorphic user grouping algorithm" (HUGA) because all the subjects are grouped in a common *parametric score space*; and the final output of the algorithm is a number of subject partitions. The scores associated with these subject partitions are aggregated so that coarsely quantised time-dependent EERs can be estimated. The proposed system architecture is summarised in Figure 4. The numbers in each box show the sequence of operations. The algorithms can be stated as:

- For each subject $j \in \mathcal{J}$:
  - Fit the regression function f to $\{y^{\omega}, t\}$ for each class of matching, ω.
  - Calculate the F-ratio over time for subject j
  - Calculate the gradient of F-ratio, a_j^1
- Cluster $\{a_j^1 | \forall_j\}$ into 5 partitions by their percentiles in a regular intervals of 20% percentiles.

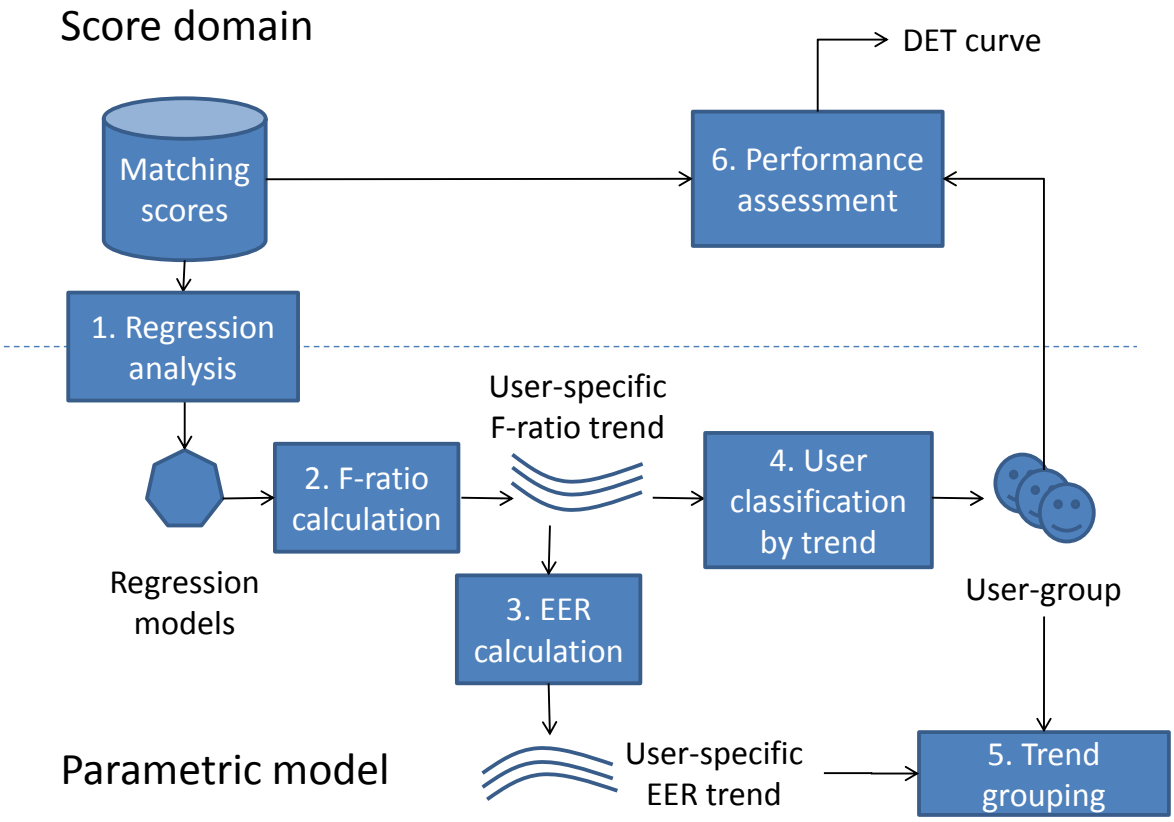


Fig. 4. An overview of the architecture

- Estimate the error for subjects in partitions 1 and 5 in the "initial" and "end" periods.

V. Experiment Setup

A. Database and Experimental Protocol

The MOBIO database [34], a bimodal face and speech database, is used for evaluating our approach. It consists of roughly 600 days of speaking face and speech data covered by 12 distinct sessions, recorded using a mobile phone (Nokia N900) at different locations simulating a typical office environment. In total, there are 192 unique video samples for each of the 150 subjects. This data was captured at six different sites with people speaking English.

Following the Phase II protocol presented in Figure 5, the database is divided into background, development and evaluation non-overlapping sets. While the training set, containing 50 subjects, is used to derive background models for the speech classifier, the development set, containing 42 subjects and the evaluation set, containing 58 subjects, are used uniquely to measure the system performance over time.

For the multi-modality experiment, the fusion tuning set is used for optimising the fusion parameters. In that case, g1 and g2 are used alternatively as development and evaluation sets; when g1 as development section is used as fusion tuning set, g2 is used as evaluation set, and vice versa. Therefore in this 2-fold validation scheme all subjects are used to measure the system performance. The enrolment data is the first session of the MOBIO database, while the probe (query) data is constituted by the remaining sessions of the MOBIO database.

B. Face classifiers

A face image is represented by two descriptors: Multiscale local binary pattern, MLBP (denoted as F1 in our experiment) and Multiscale local phase quantisation, MLPQ (F2). While the former descriptor attempts to capture variation in local image

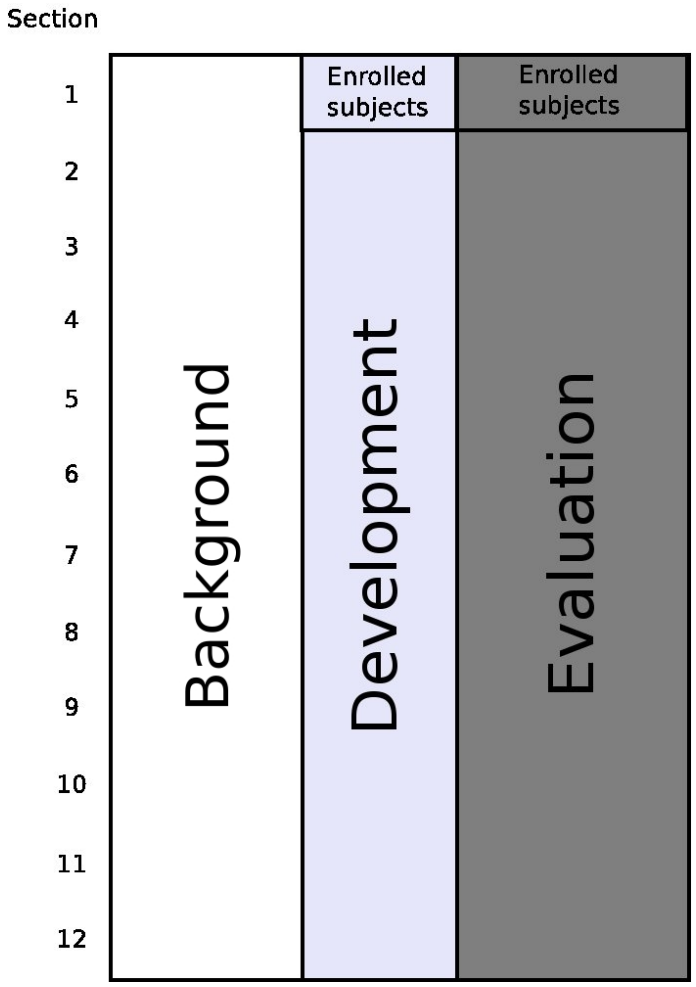


Fig. 5. Diagram showing the partitioning of the Mobio database according to the phase II protocol.

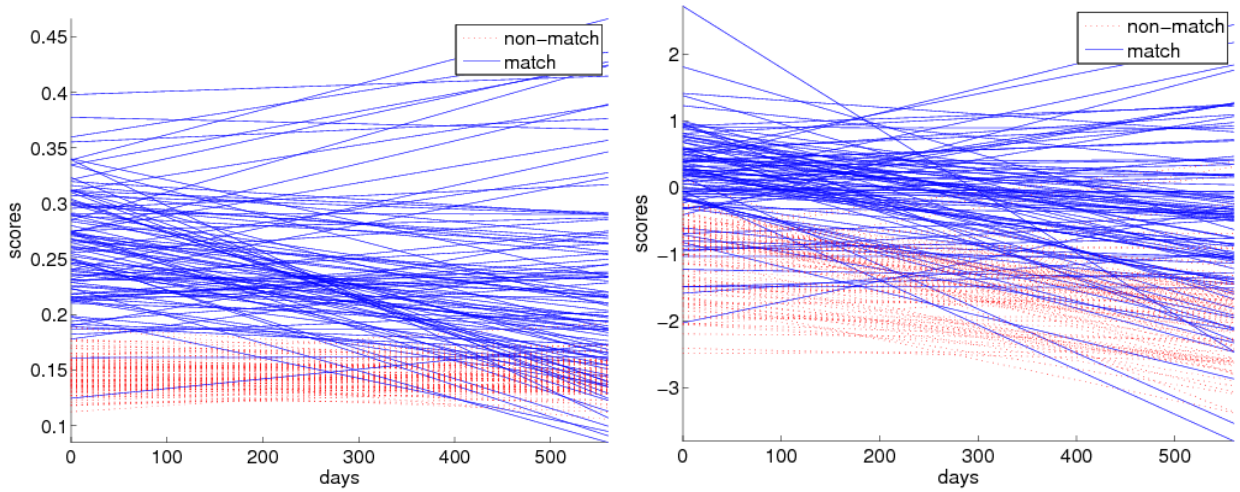
texture, the latter exploits the phase information and is, therefore, expected to be robust to image blurring. For both features, an image is encoded as a histogram convering a statistical summary of these features derived at different image resolutions. The back-end classifier for both descriptors attempt to compare a pair of histograms. This is done first by linearly projecting them into a more compact space via Linear Discriminant Analysis (LDA). The distance between a pair of histograms is defined in terms of a normalised correlation metric in the LDA space.

C. Speech classifier

A long-standing state-of-the-art classifier in speaker verification is based on Gaussian Mixture Models (GMM). A pair of GMM models is used: one to model the distribution of the speech features for the target user/subject, and another to model the distribution of these features for an arbitrary number of non-target subjects. While the former model is subject-dependent, the latter model, also called the "world model" or "universal background model", is common to all enrollees. In order to use them to verify whether a speaker is the target subject or not, a standard approach is to compute the likelihood ratio of the observed speech features given the two distributions. Modern speaker verification classifiers are still based on the two GMM models but involve more complex classification rule.

Features that have been used for speech recognition also find their use for speaker verification. This is unexpected (but works extremely well) because in speaker verification, it is desirable to retain as much speaker variation as possible whereas in speech recognition, it is better to suppress the variation. For our experiment, linear-frequency cepstral coefficient (LFCC) that have evenly spaced frequency bands are used [35],

An important assumption when using a GMM is that the LFCC features (of roughly 13 dimensions) from a speech sequence are independent of each other. Because this assumption is violated for features in close temporal proximity, it is common to use a first order time-difference features (called "delta" features) as well as second order ones ("delta-delta"). All these features are used in our system.



(a) Face (b) Speech

Fig. 6. The fitted linear curves across all the subjects for the face classifier MLPQ and the speaker verification classifier based on GMM.

D. Fusion classifier

The fusion classifier used throughout this paper is logistic regression. It can be viewed as a weighted sum rule fusion whose weights are tuned in the framework of logistic regression, using the "gradient ascent" algorithm [36]. The performance of this classifier has been shown to give good results [37].

VI. Empirical Analysis

A. Results I: Model fitting

This section presents the intermediate results that show the fitted score models. Regression is extensively used at two different levels: modelling the score density over time; and modelling the F-ratio time series.

Figure 6 shows the fitted linear regression lines conditioned on each subject and each matching type (genuine or impostor), using a face authentication system and the speech authentication system as examples. Each continuous line (plotted in blue) represents the mean score trend of the match comparison whereas each dashed line (red) represents the trend of the non-match comparison. A linear regression line is used in order to coarsely model the trend as upwards, downwards, or stable. As can be observed, the fitted mean score trends for the match (genuine) comparison exhibit much higher variation than their non-match comparison counterparts. However, it is difficult to observe the performance of the system over time for each subject.

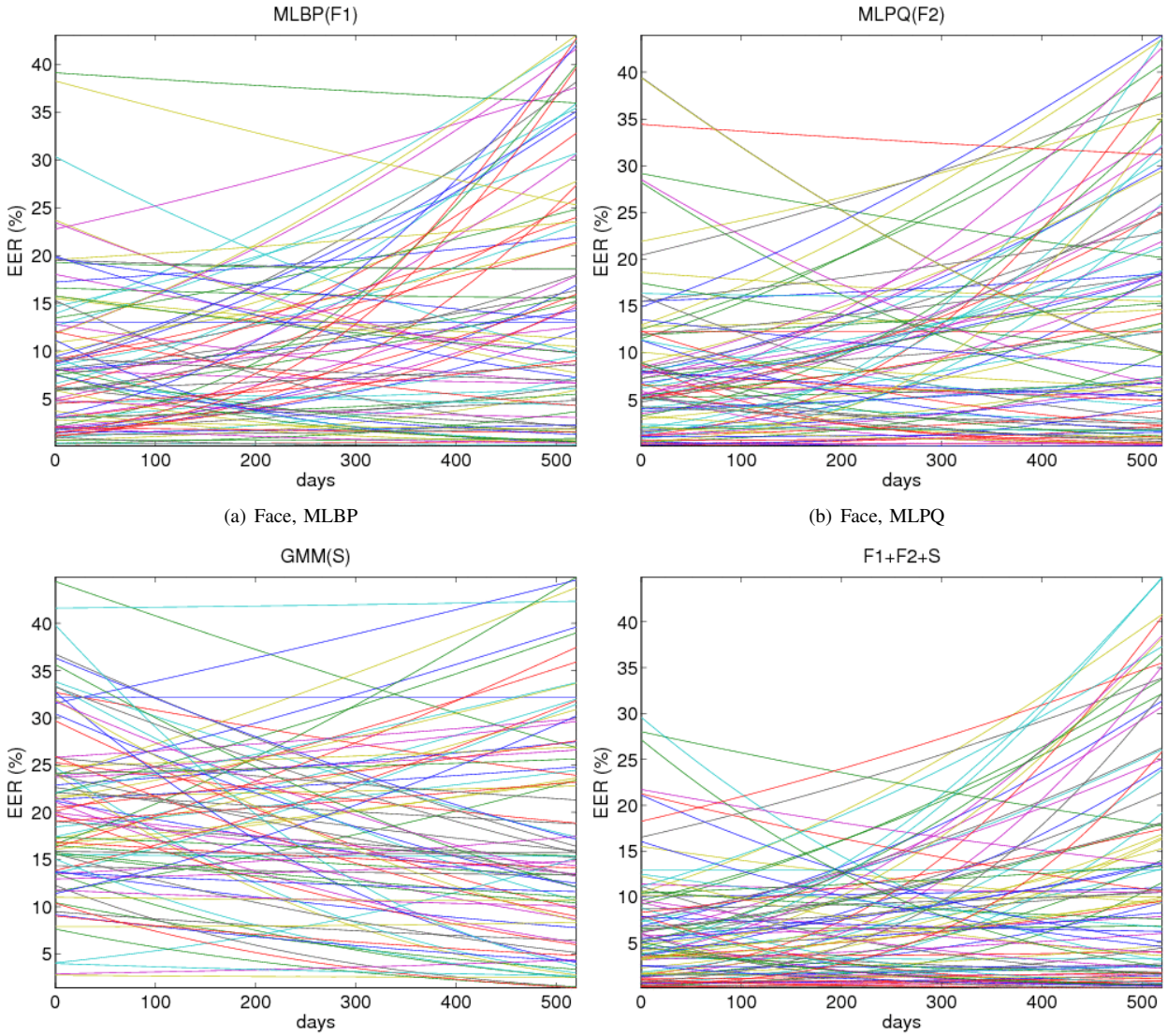
To remedy this we propose the use of EER curves presented in Figure 7 which are obtained by applying (2) to the parameters $\{\mu^{\omega}_{j,t}, \sigma^{\omega}_{j,t}\}$ for all t associated with each subject j. Each curve in this figure is a EER time-series for each subject. It is obvious from this figure that the fusion system, as shown in Figure 7(d), has better performance on day 1 as none of the subjects have EER exceeding 30%. However, the multi-modal fusion system still degrades in performance just like its constituent systems. In Figure 7, the EER curves for different subjects have three possible trends which are stable, downwards and upwards.

B. Results II: Partitioned Groups

In order to summarize the performance trends presented in Figure 7, HUGA, based on the gradient of F-ratio, is proposed to group the subjects into 5 partitions in ascending order. Figure 8 shows how the EER trends are clustered sensibly for the MLBP classifier. The other classifiers behave similarly. The first three partitions of subjects, (Groups A, B and C), have negative gradients (decreasing performance or increasing EER) shown in the top row of Figure 8, the fourth partition of subjects, Group D, exhibits a stable performance over time, whereas the last partition of subjects, Group E, has positive gradient (increasing performance or decreasing EER).

The time-dependent EERs of the three unimodal systems, as well as their combined systems, are shown in Figure 9. As can be observed, our subject-dependent analysis can identify subjects whose performance improves over time as well as the subjects whose performance degrades over time.

Note that, for all systems, including the fusion systems (with different constituent sub-systems), the subjects in Group A exhibit increasing EER over the three time periods. Conversely, the subjects in Group E have decreasing EER.



(a) Face, MLBP

(b) Face, MLPQ

(c) Speech, LPPC-GMM

(d) Fusion

Fig. 7. The evolution of EER curves over time for different systems. (a) and (b) are two face biometric systems, (c) is a speaker verification system, and (d) is the fusion of all three systems.

VII. Conclusions

This paper presents a new methodology for estimating the biometric performance over time. A novel model called "homomorphic user grouping algorithm" (HUGA) was introduced in order to clarify the subject-specific performance over time. Although the modelled performance trend is approximate, it allows us to detect changes in the score trend and to improve our understanding of the impact of time on biometric performance. Our analysis suggests that biometric performance does change over time. In the literature, it is widely claimed that biometric performance degrades over time. Our experiments support this observation only partially. A rather surprising finding is that the biometric performance can also improve with use. A logical explanation of this is that as the subjects get used to the system, they learn to use the device in an optimal way. However, it is not possible to quantify the degree of familiarity with the system, nor separating the influence of ageing from the influence of the acquisition environment. The MOBIO database that we used was designed to be as realistic as possible. As a consequence, there was no effort to correct for these factors. However, despite such large variations, the proposed methodology, HUGA, can still sensibly identify the subject-specific performance trends. In order to encourage the research community to reproduce the results on longitudinal biometric databases, and ideally using other biometric modalities, the tool is made available for the research community.

Acknowledgments

This work was supported by the Biometrics Evaluation and Testing (BEAT), an EU FP7 project with grant no. 284989.

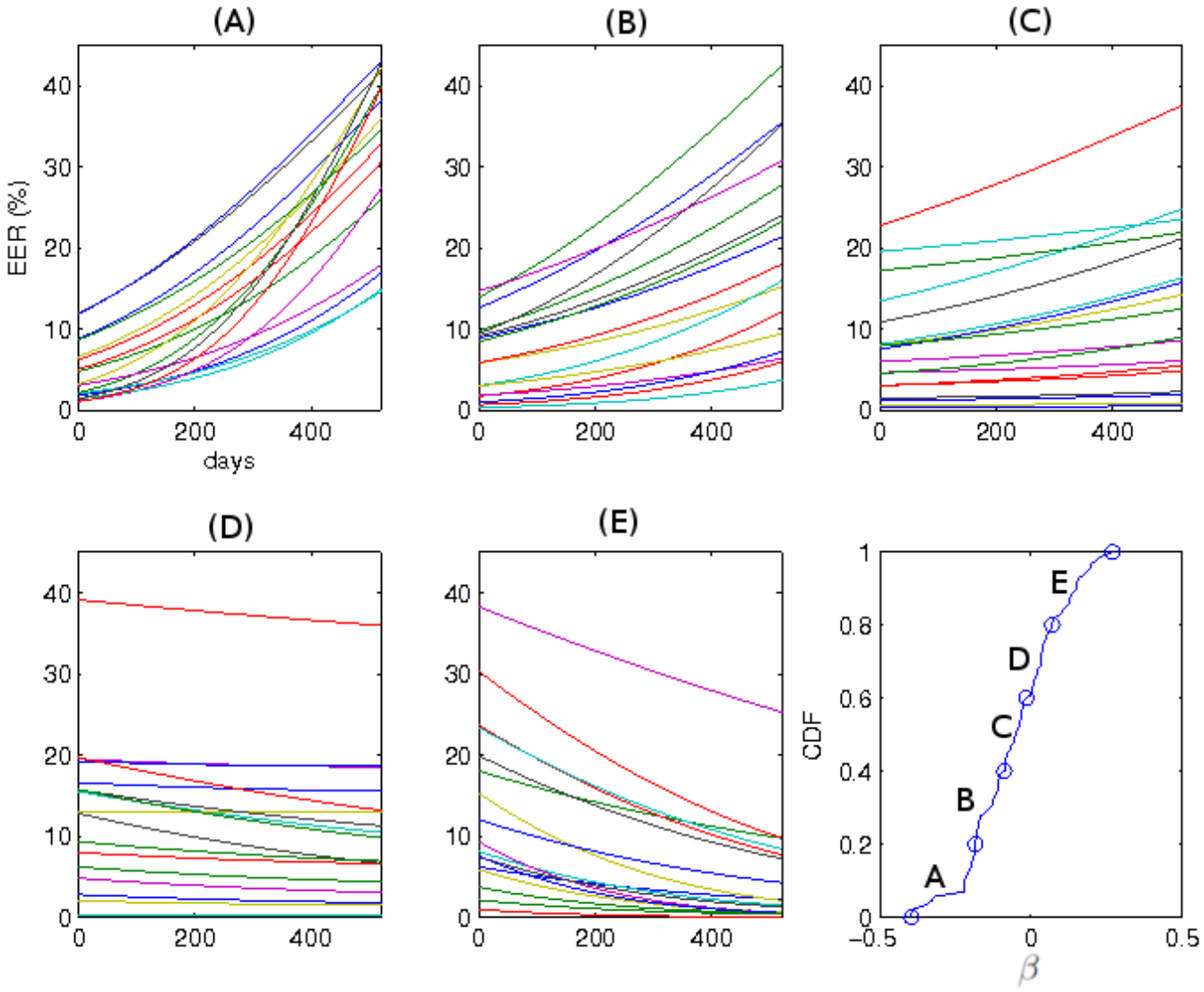


Fig. 8. Categorisation of the EER trends for the MLBP classifier.

REFERENCES

[1] P. J. Flynn, K. W. Bowyer, and P. J. Phillips, "Assessment of Time Dependency in Face Recognition: An Initial Study," in *LNCS 2688, 4th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2003)*, Guildford, 2003, pp. 44–51.

[2] H. Wang and P. J. Flynn, "Experimental Evaluation of Eye Location Accuracies and Time-Lapse Effects on Face Recognition Systems," in *LNCS 3546, 5th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2005)*, New York, 2005, pp. 627–636.

[3] N. Poh and J. Kittler, "A method for estimating authentication performance over time, with applications to face biometrics," in *12th IAPR Iberoamerican Congress on Pattern Recognition (CIARP)*, Via del Mar-Valparaiso, Chile, 2007, pp. 360–369.

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

[5] N. Yager and T. Dunstone, "Worms, chameleons, phantoms and doves: New additions to the biometric menagerie," *Automatic Identification Advanced Technologies, 2007 IEEE Workshop on*, pp. 1–6, June 2007.

[6] N. Poh and J. Kittler, "A methodology for separating sheep from goats for controlled enrollment and multimodal fusion," in *Proc. of the 6th Biometrics Symposium*, Tampa, 2008, pp. 17–22.

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

[8] 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.

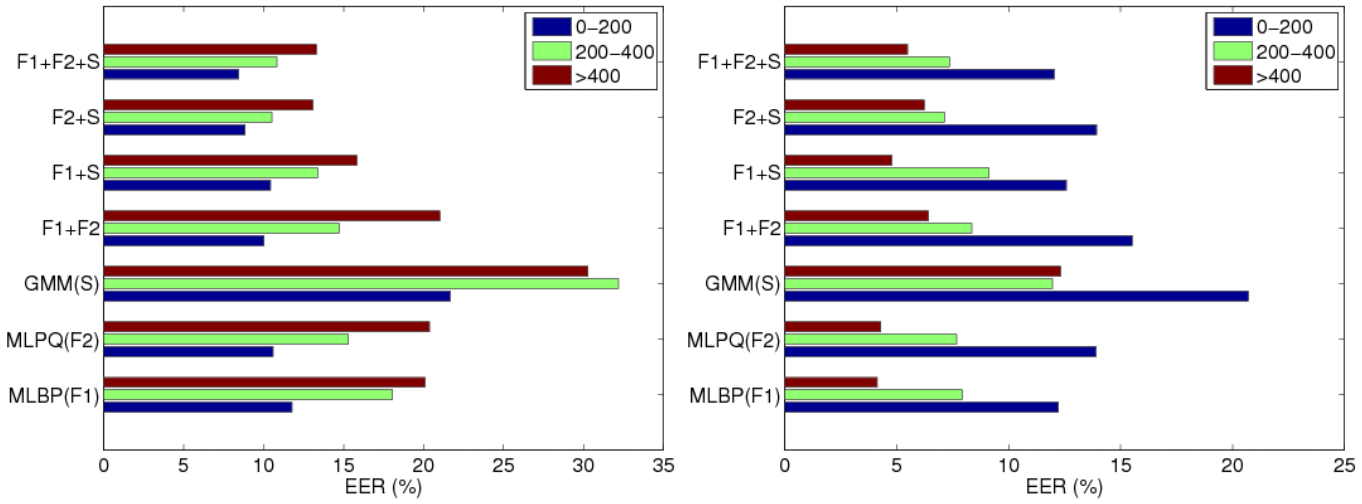
[9] J. Daugman, "Biometric decision landscapes," University of Cambridge Computer Laboratory, Tech. Rep. TR482, 2000.

[10] A. Hicklin and B. Ulery, "The myth of goats: How many people have fingerprints that are hard to match?" National Institute of Standards and Technology, Tech. Rep. NISTIR 7271, 2005.

[11] M. Wittman, P. Davis, and P. Flynn., "Empirical studies of the existence of the biometric menagerie in the frgc 2.0 color image corpus," *Conf. on Computer Vision and Pattern Recognition Workshop*, pp. 33–33, June 2006.

[12] M. Une, A. Otsuka, and H. Imai, "Wolf attack probability: A theoretical security measure in biometric authentication systems," *IEICE-Transactions on Info and Systems*, vol. E91-D, no. 5, pp. 1380–1389, 2008.

[13] 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.



(a) Group A (b) Group E

Fig. 9. EER of different partitions for different systems in the three time periods.

[14] J.-B. Pierrot, "Elaboration et Validation d'Approaches en Vérification du Locuteur," Ph.D. dissertation, ENST, Paris, September 1998.

[15] K. Chen, "Towards Better Making a Decision in Speaker Verification," *Pattern Recognition*, vol. 36, no. 2, pp. 329–346, 2003.

[16] J. 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] D. Genoud, "Reconnaissance et Transformation de Locuteur," Ph.D. dissertation, Ecole Polythechnique Fédérale de Lausanne (EPFL), Switzerland, 1998.

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

[19] J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Target Dependent Score Normalisation Techniques and Their Application to Signature Verification," in *LNCS 3072, Int'l Conf. on Biometric Authentication (ICBA)*, Hong Kong, 2004, pp. 498–504.

[20] N. Poh and J. Kittler, "On the use of log-likelihood ratio based model-specific score normalisation in biometric authentication," in *LNCS 4542, IEEE/IAPR Proc. Int'l Conf. Biometrics (ICB'07)*, Seoul, 2007, pp. 614–624.

[21] ——, "Incorporating variation of model-specific score distribution in speaker verification systems," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 3, pp. 594–606, 2008.

[22] P. Grother and E. Tabassi, "Performance of Biometric Quality Measures," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 531–543, 2007.

[23] A. Rattani, G.-L. Marcialis, and F. Roli, "An experimental analysis of the relationship between biometric template update and the doddington's zoo: A case study in face verification," in *ICIAP '09: Proceedings of the 15th International Conference on Image Analysis and Processing*. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 434–442.

[24] M. N. Teli, J. R. Beveridge, P. J. Phillips, G. H. Givens, D. S. Bolme, and B. A. Draper, "Biometric zoos: Theory and experimental evidence," in *Biometrics: Theory, Applications, and Systems, 2009. BTAS '09. IEEE 3rd International Conference on*, 2011, pp. 1 –8.

[25] 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.

[26] N. Poh, T. Bourlai, and J. Kittler, "A multimodal biometric test bed for quality-dependent, cost-sensitive and client-specific score-level fusion algorithms," *Pattern Recognition*, vol. 43, no. 3, pp. 1094–1105, 2010.

[27] J. Daugman, "New methods in iris recognition," *IEEE Trans. on Systems, Man, and Cybernetics, Part B*, vol. 37, no. 5, pp. 1167–1175, Oct. 2007.

[28] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[29] F. Cardinaux, C. Sanderson, and S. Marcel, "Comparison of MLP and GMM Classifiers for Face Verification on XM2VTS," in *LNCS 2688, 4th Int'l Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2003)*, Guildford, 2003, pp. 911–920.

[30] G. E. Box and D. R. Cox, "An Analysis of Transformations," *Automatic Identification Advanced Technologies, 2007 IEEE Workshop on*, vol. B, no. 26, pp. 211–246, 1964.

[31] N. Poh and J. Kittler, "On using error bounds to optimize cost-sensitive multimodal biometric authentication," in *Proc. 19th Int'l Conf. Pattern Recognition (ICPR)*, 2008, pp. 1–4.

[32] S. C. Dass, Y. Zhu, and A. K. Jain, "Validating a Biometric Authentication System: Sample Size requirements," *IEEE Trans. on Pattern Analysis and Machine*, vol. 28, no. 12, pp. 1902–1319, 2006.

[33] 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.

[34] C. McCool, S. Marcel, A. Hadid, M. Pietikainen, P. Matejka, J. Cernocky, N. Poh, J. Kittler, A. Larcher, C. Levy, D. Matrouf, J.-F. Bonastre, P. Tresadern, and T. Cootes, "Bi-modal person recognition on a mobile phone: using mobile phone data," in *IEEE ICME Workshop on Hot Topics in Mobile Multimedia*, July 2012.

[35] Q. Le and S. Bengio, "Client dependent gmm-svm models for speaker verification," in *ICANN*, 2003, pp. 443–451.

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

[37] "Benchmarking quality-dependent and cost-sensitive score-level multimodal biometric fusion algorithms," *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 4, no. 4, pp. 849–866, 10 2009.