TABULA RASA

Trusted Biometrics under Spoofing Attacks

http://www.tabularasa-euproject.org/

Funded under the 7th FP (Seventh Framework Programme)
Theme ICT-2009.1.4
[Trustworthy Information and Communication Technologies]

D3.2: Evaluation of baseline non-ICAO biometric systems

Due date: 30/09/2011 **Submission date:** 30/09/2011
Project start date: 01/11/2010 **Duration:** 42 months
WP Manager: Abdenour Hadid **Revision:** 0

Author(s): Federico Alegre, Xuran Zhao, Nick Evans (EURECOM); John Bustard, Mark Nixon (USOU); Abdenour Hadid (UOULU); William Ketchantang, Sylvaine Picard, Stéphane Revelin (MORPHO); Alejandro Riera, Aureli Soria-Frisch (STARLAB); Gian Luca Marcialis (UNICA)

Project funded by the European Commission in the 7th Framework Programme (2008-2010)		
Dissemination Level		
PU	Public	Yes
RE	Restricted to a group specified by the consortium (includes Commission Services)	No
CO	Confidential, only for members of the consortium (includes Commission Services)	No

D3.2: Evaluation of baseline non-ICAO biometric systems

Abstract:

To gain insight into the performance of current biometric systems when not confronted to spoofing attacks, we report in this deliverable the results of some baseline systems. This will provide valuable baseline performance for later investigating the performance and the vulnerabilities of biometric systems when confronted to spoofing attacks. This document presents the evaluation and performance profiles for all non-ICAO mono-modal and multi-modal biometrics addressed within the TABULA RASA project. The baseline results presented here will form a cornerstone of all subsequent work related to spoofing and countermeasures, the performance of which will be compared to that presented here. All biometrics are assessed with a common, minimal protocol involving independent development and evaluation datasets and a common operating point, namely the equal error rate. Detection error trade-off curves are also reported to illustrate the dynamic performance of all biometric systems for alternative operating points and different applications.

Contents

1 Introduction

The TABULA RASA project aims to assess the threat of direct, sensor-level spoofing to biometric systems and then to propose novel countermeasures. These activities require the comparison of system performance under spoofing conditions, first without and then with countermeasures, to baseline scores for standard, state-of-the-art biometric systems. The purpose of this deliverable (and also of the companion deliverable D3.1) is thus to establish the baseline scores to which all later experimental results with spoofing and countermeasures will potentially be compared. This particular document presents baseline scores for all non-ICAO biometric modalities considered within the TABULA RASA project. Baseline scores for ICAO biometrics are presented within the companion document, D3.1. Both documents relate to the biometric database and system specifications previously reported in D2.2.

All results are presented in terms of detection error trade-off (DET) profiles which illustrate the dynamic behaviour of a biometric system as the decision threshold is changed, i.e. how the false acceptance rate varies according to the false rejection rate. While there are boundless variations on assessment approaches in the literature they all pertain to specific operating conditions, i.e. prior probabilities, false rejection and/or false acceptance costs. Most of these are furthermore dependent on specific applications and biometric modalities. The TABULA RASA project addresses numerous different use-cases and numerous mono-modal and multi-modal biometrics. Thus, in the absence of a single dominant application and multiple biometric combinations, all systems are optimised so as to minimise the equal error rate (EER) which is the sole, standard metric for all biometric modalities in the TABULA RASA project. Results reported here are therefore not necessarily directly comparable to other published results on the same datasets. This, however, does not detract from the impact of the work presented here since the focus in later deliverables, and of the wider project, is on the degradation in performance caused by spoofing relative to the baseline and thus the precise operating point is not of critical importance. Naturally, further work would be required to assess impacts on alternative operating points and specific applications but this is beyond the scope of the work in TABULA RASA. The illustration of DET plots nonetheless gives an idea of performance for other operating points even if they are not that used for optimisation. DET plots are presented for evaluation datasets only.

In all cases experiments relate to a standard, minimal specification which involves multiple sessions and independent development and evaluation datasets. Auxiliary datasets used for the learning of world or universal models and normalisation procedures etc. are independent from those used for development and evaluation. Independent datasets do not contain any overlap in terms of clients/subjects. Full details of database and system specifications are presented in D2.2 though a brief summary of both is provided here so that the document can be read independently.

Before starting with a biometric-by-biometric treatment of baseline results we present here a brief summary of future work in order to show how the material presented here fits into the wider picture and direction of the TABULA RASA project. Figure 1 illustrates

three example DET profiles for any hypothetical mono-modal or multi-modal biometric and aims to distinguish the results in this deliverable from those which can be expected in future work. The lowest profile (solid black) is that of the baseline: a standard, state-of-the-art biometric system with no spoofing and no countermeasures. For this hypothetical biometric modality the EER is in the order of 10%. When spoofing attacks are applied performance is expected to degrade, perhaps considerably. Many such profiles will be derived for different spoofing attacks (whereas there is generally only one baseline). One such profile is illustrated and is the highest (dashed red) in Figure 1. It corresponds to an EER in the order of 40%. Finally, the third, middle profile (dashed blue) illustrates performance once spoofing countermeasures are applied. Once again, multiple profiles will be derived for different countermeasure strategies. That illustrated corresponds to an EER in the order of 20% thus showing an improvement over the higher spoofing profile but still a gap to the original baseline. The goal of the latter work is thus to reduce this gap as much as possible and thus the establishment of the lower baseline is a cornerstone of the project.

The only profile in Figure 1 which relates to the work presented in this document is the lowest, namely that of the baseline. Research related to other profiles are the subject of future work and deliverables. For example, D3.4 is the subject of performance under spoofing attacks (D3.3 for ICAO biometrics) due in M21 whereas D4.2 and D4.4 are the subject of first initial, and then advanced countermeasures (D4.1 and D4.3 for ICAO biometrics) due in M25 and M33 respectively.

In the following sections we present the baseline scores and DET profiles for each of the non-ICAO biometric modalities (voice, gait, vein and fingerprint, electro-physiology) and then the numerous multi-modal combinations (2D-face and voice, 2D-face and fingerprint, 2D-face and 3D-face, ECG and EEG).

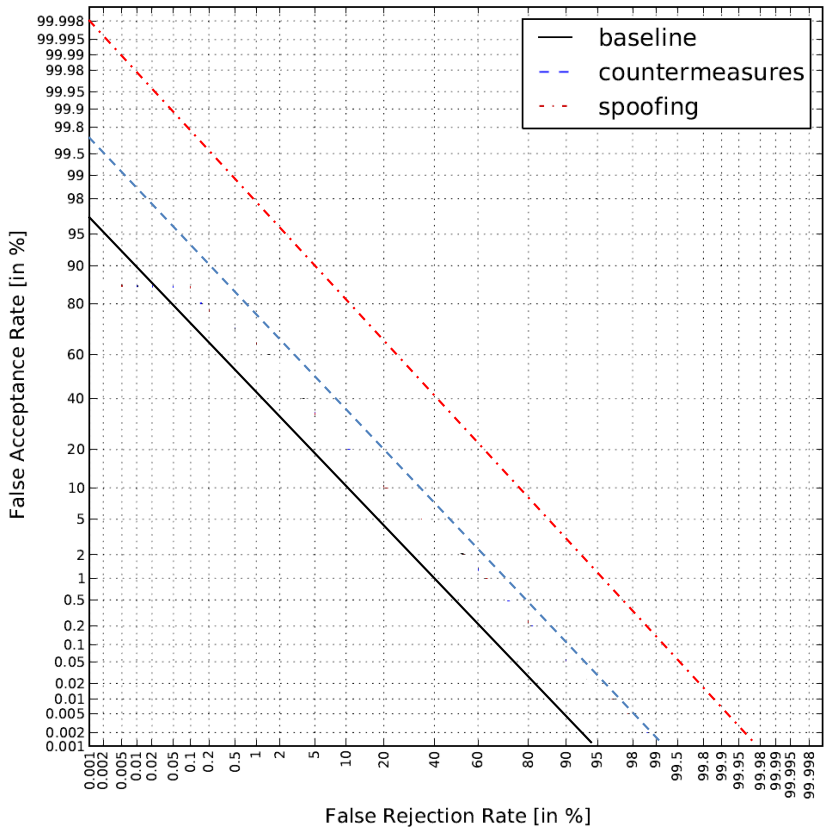

Figure 1: An overview of the different stages in the TABULA RASA project. The objective of the work reported in this deliverable is to establish the lower baseline performance of state-of-the-art biometrics systems for all non-ICAO mono-modal and multi-modal modalities.

2 Voice Biometrics

The voice biometric will be assessed both standalone and in combination with the 2D-face modality. Mono-modal and multi-modal assessments are performed on different databases. The NIST speaker recognition evaluation datasets are the de facto standard, are used in all state-of-the-art research and are thus used here for mono-modal work. The datasets are not multi-modal, however, and thus the MOBIO database is used for all multi-modal work. In the following we concentrate on baseline results related to the NIST speaker recognition evaluation (SRE) datasets. They are are telephony-based and relate arguably to the most appealing use of voice recognition, namely remote recognition over the telephone. The scenario is one of the most challenging in terms of spoofing and countermeasures since it is entirely unsupervised and is thus particularly prone to spoofing attacks.

Also reported here are the differences between the speaker recognition systems optimised for the NIST SRE datasets and the multi-modal MOBIO database. Speaker recognition results for the MOBIO database are also presented. The database itself, however, is not described here; brief details are presented in Section 6 and in deliverables D2.2 and D3.1.

2.1 The ALIZE Speaker Recognition System

All speaker recognition work to be conducted in the TABULA RASA project will be undertaken using the state-of-the-art 'ALIZE' speaker recognition system [3] using the implementation described in [11]. As described in D2.2 the SPro toolkit system is used for feature extraction whereas ALIZE itself is used to perform various forms of feature normalization such as cepstral mean and variance normalisation. The standard approach to statistical speaker modelling is based on Gaussian mixture models (GMMs) [28] and generally employs some form of world model [8] or universal background model (UBM) trained using the expectation maximisation (EM) algorithm [9] and large amounts of data from a pool of background speakers. Target speaker models are generally adapted from the UBM during enrollment through maximum a posteriori (MAP) adaptation [13]. Scores correspond to the log-likelihood ratio of the target model and the test segment, normalised with respect to the background model. Various score-level normalisation procedures such as test-normalisation (TNorm) [2] can be applied. Final decision logic is based on a threshold which is empirically determined using a large, representative development set. The ALIZE framework also provides for more recent, advanced approaches including support vector machines (SVMs) [36], e.g. generalised linear discriminant sequence kernel (GLDS) [6] and the GMM super-vector linear kernel (GSL) [7], nuisance attribute projection (NAP) [5] and factor analysis (FA) [19].

2.2 The NIST SRE Datasets

The NIST datasets contain several hundreds of hours of speech data collected from telephone conversations. Common to each dataset is a single compulsory, 'core' evaluation

condition which typically involves in the order of a few minutes of speech per training and testing segment. Whilst at a later date we may turn our attention to different conditions involving more training data, or fewer test data, the NIST-defined core condition is the default condition used for all experimental work in TABULA RASA. Training and testing protocols are defined by NIST and allow for different systems and technologies from different research groups to be readily and meaningfully compared according to standard experimental and evaluation protocols and metrics. We note, however, that this is not strictly the case for TABULA RASA work since our system will be optimised to minimise the equal error rate (EER) and not on the standard operating conditions (costs) defined by NIST. A typical speaker recognition system requires an independent development set in addition to independent auxiliary data which is needed for background model training and the learning of normalisation strategies. This data typically comes from other NIST datasets, such as the 2004 dataset which will be the case for all TABULA RASA work. Basic evaluation rules relate to the independence of trial decisions, permitted normalisation procedures, human interaction and the use of additional data, such as speech transcripts, etc.

2.3 Performance Evaluation

Within TABULA RASA we aim to assess the effect of spoofing on a range of systems based on recent developments in the field of automatic speaker recognition. All of them lead to state-of-the-art performance as judged by the series of NIST SREs and are all based upon the ALIZE toolkit [3]. The inclusion of multiple baseline systems in the case of speaker recognition is motivated by the likely impact of different channel compensation algorithms which may be of assistance to a would-be spoofer. As is common practice separate systems are independently optimised for both male and female data subsets. All systems are optimised according to the standard EER metric with dynamic performance assessed according to the standard detection error trade-off (DET) plots. Note that, in the case of the NIST SRE datasets, this is in contrast to convention which dictates optimisation according to the minimum decision cost function (minDCF). Results presented in the context of TABULA RASA are thus not directly comparable to those in the open literature.

2.3.1 Setup

The two different setups are described here, one for the NIST SRE datasets and another for the MOBIO database.

NIST SRE
In all experiments the NIST'04 dataset is used for background data, e.g. that used for learning the universal background model (UBM) and that used in the application of test normalisation (TNorm), nuisance attribute projection (NAP), and factor analysis (FA). The NIST'05 dataset is used for development whereas the NIST'06 dataset is used for

evaluation[1]. All experiments relate to the core condition (1conv4w-1conv4w) which involves approximately 5 minutes of data for model training and testing.

Features are composed of 16 linear frequency cepstral coefficients (LFCCs), their first derivatives and delta energy, thereby producing a feature vector with 33 coefficients which are computed from Hamming windowed frames of 20ms and with a frame rate of 10ms. Voice activity detection is then applied using energy coefficients which are first normalized to fit a zero-mean and unity-variance distribution. They are used to train a three-component GMM which aims to classify acoustic frames into speech/non-speech according to acoustic energy. Speaker modelling is applied only to speech frames; non-speech frames are discarded.

A total of five different systems were assessed and optimised in a similar fashion to the work reported in [11]. All systems were tested with and without the application of TNorm using an impostor cohort from the NIST'04 database. All systems have their roots in the standard GMM:

- **GMM-UBM:** The classical system is the GMM-UBM approach with TNorm likelihood score normalization. The UBM model is trained with an EM algorithm. Speaker models are adapted from the UBM via the maximum a posteriori (MAP) adaptation of the GMM mean vectors. Diagonal covariance matrices are not adapted. A top ten component selection is used for likelihood computation. The GMM-UBM system is used as a basis for all other systems.

- **GMM supervector linear kernel (GSL):** The GSL system uses an SVM classifier which is applied to GMM supervectors. The approach is based on that described in [7]. The GSL system is known to outperform other related approaches such as the generalized linear discriminant sequence kernel (GLDS). Supervectors come directly from the GMM-UBM system.

- **GMM supervector linear kernel + nuisance attribute projection (GSL-NAP):** The GSL-NAP system is identical to the GSL system but is enhanced with nuisance attribute projection to attenuate intersession (interchannel) variability [5]. Performance is dependent on the rank of the NAP matrix. All experiments reported here correspond to NAP matrices of rank 40 and are learned on the full male and female subsets of the NIST'04 database.

- **Factor Analysis (FA):** A FA-based system is also proposed. It is implemented by following the novel latent symmetrical approach [23] to Kenny's original work presented in [19]. This new strategy allows the results (GMM models without session effects) to be used directly in a SVM classifier, among other advantages. All work reported here relates to an intersession matrix corresponding to the 40 most significant eigenchannels.

[1]Some of this work was conducted through external collaboration with Swansea University

- **GSL with FA Supervectors (GSL-FA):** The GSL-FA system aims to exploit the complementarity in the FA and GSL-NAP systems. The system, reported in [11] is a discriminative SVM approach applied to mean supervectors evaluated in the factor analysis framework.

MOBIO

Both different features and modelling approaches were optimised for the MOBIO database; here we describe only the differences between the two setups. For the MOBIO database the best performance was obtained with the standard GMM-UBM system. We note that similar setups were reported in related work [21]. Feature extraction is applied to the original 48kHz-sampled signal using a mel-scaled filterbank. While frame rates, sizes and windowing are the same as for the NIST SRE setup, feature vectors are composed of 53 components including 26 MFCC coefficients which are augmented with 26 delta coefficients and the delta energy. A 256-component UBM is trained using the EM algorithm and speaker models are adapted from the UBM via MAP adaptation. Test-normalisation (TNorm) is applied as before using impostor segments from the MOBIO database.

2.3.2 Results

Results are presented here for the two different datasets.

NIST SRE

We ran a series of experiments designed to compare the performance of the GMM, GSL, GSL-NAP, FA and GSL-FA systems for both male and female subsets. All systems were optimised independently on the development set and were then applied without modification to the evaluation set. The EERs for each system are presented in Table 2.3.2 and show the evolution in performance with different approaches to compensate for intersession variation. Results are reported for both development and evaluation subsets. This is because the best performing systems are judged from their performance on the development set and not on the evaluation set. The differences in performance are thus of interest and both are essential to make the link between EERs and DET plots presented below.

For the male subset the best performing FA system (judged from the development set) gives an EER of 5.9% on the evaluation set whereas for the female subset the best performing SGL-FA system gives an EER of 5.4%. These compare well to the GMM-UBM system where the respective EERs are 9.3% and 10.7%. Upon comparison of these results to those reported in the most recent NIST SRE campaigns we note that the tested systems represent the state-of-the-art in current automatic speaker recognition technology and are therefore suitable candidates for assessing the threat from spoofing and for testing countermeasure developments.

DET plots for the evaluation set are illustrated in Figure 2. They relate to the FA system for the male subset and the SGL-FA system for the female subset (even though they are not necessarily the systems which give the best performance on the evaluation

System	Development		Evaluation	
	male	female	male	female
GMM-UBM	8.2	11.0	9.3	10.7
SGL	7.8	8.9	7.2	7.8
SGL-NAP	5.9	8.7	5.7	7.0
FA	4.7	7.8	5.9	6.9
SGL-FA	5.1	7.1	4.7	5.4

Table 1: Equal error rate (EER) scores (%) for each of the five speaker verification systems and for both male and female subsets. Results are illustrated for the development (NIST'05) and evaluation/test (NIST'06) datasets.

set). The two profiles are relatively close except in the low false rejection region where the false acceptance rate is higher for the female subset than for the male subset. It will be of interest to explore further the effect of different systems on the dynamics of the DET profile under spoofing.

MOBIO

The performance of the experiments on the MOBIO test data set are presented by DET curves in Figure 3. We report EERs of 15.2% for male speakers and 18.4% for female speakers. The difference between the two subsets is thus more pronounced for the MOBIO database than for the NIST SRE datasets. We note that similar levels of performance were obtained on the same MOBIO database in related work [21].

Results are noticeably worse than that for the NIST SRE datasets. This is only to be expected, however, given that the NIST SRE datasets contain greater amounts of data per trial and higher quality telephone recordings. The MOBIO database, in contrast, is far more challenging and contains fewer data per trial and lower-quality data recorded from a more distant microphone. Given that the MOBIO database is multi-modal it is of interest to see if the fusion of speaker recognition scores with 2D-face recognition scores can improve the EER in this case. This work is reported in Section 6.

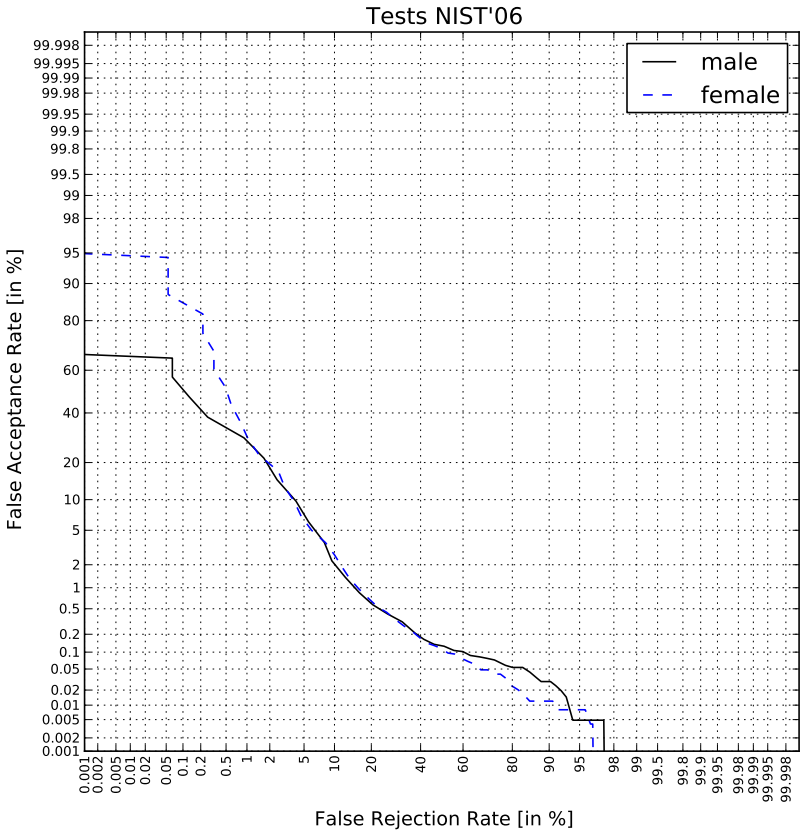

Figure 2: Detection error trade-off (DET) profiles for male and female subsets of the evaluation/test NIST'06 dataset.

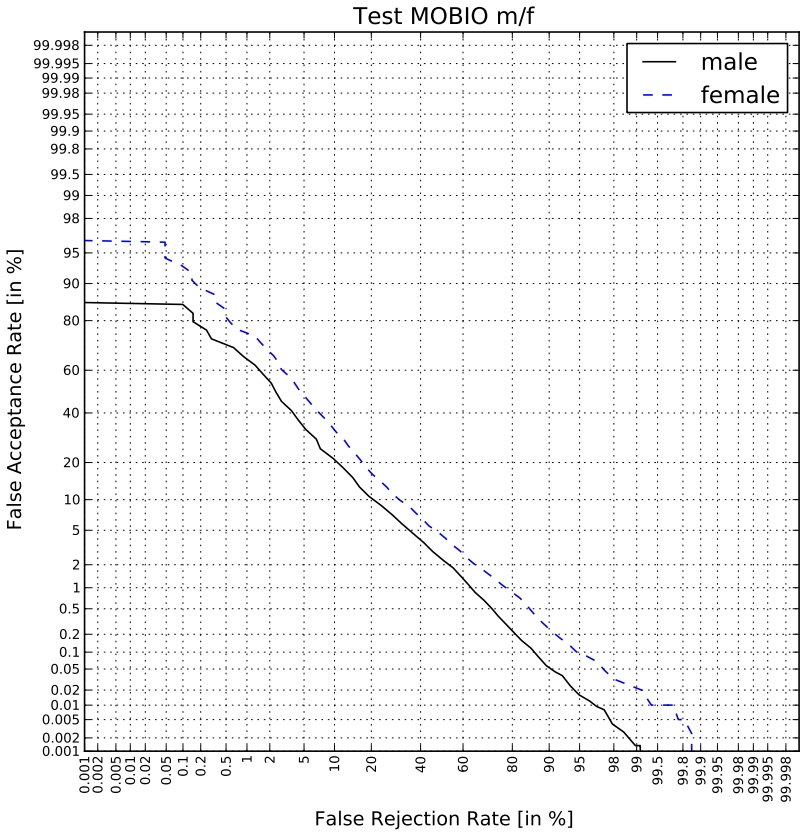

Figure 3: Detection error trade-off (DET) profiles for male and female subsets of the evaluation/test MOBIO dataset.

3 Gait Biometrics

Biometric gait recognition refers to recognizing people from the way they walk. Although at a much earlier stage of development than ICAO biometrics, such as face, iris and fingerprint, gait recognition has recently become a topic of great interest in biometric research. The current state-of-art is that people can be recognized by gait by silhouette or by model based approaches [24]. This has required development of techniques to analyze video data for the purpose of recognizing the walking subject. There have been more approaches which use the human silhouette, and of these, approaches which use the averaged silhouette have proved most popular [25]. The earliest approaches achieved recognition rates exceeding 90% and this is matched by the most recent approaches on databases extending to 300 subjects. Much of the earlier work was conducted on data acquired using controlled conditions but later recognition was demonstrated on outdoor derived data, though with slightly lower performance. There have been many databases for evaluating the progress in gait recognition research such as HiD (NIST, US) [31], Soton (Southampton UK) [33], and CASIA (CAS, China) [40] databases. The earliest databases comprised of only tens of subjects, sometimes wearing specified clothing. More recent databases include outdoor as well as indoor data (thus, with uncontrolled illumination) and with variation in camera viewpoint.

Gait is a relatively new biometric, as such there has been relatively little investigation into spoofing attacks. Some preliminary investigations point out that gait is potentially difficult to hide or spoof as it is behavioural and encompasses the whole body. In order to allow investigating, in later deliverables, the vulnerability of gait biometrics against spoofing attacks, we start by providing in this section the performance of baseline gait biometric systems when not confronted to spoofing attacks. We report the results of two baseline systems developed at the universities of Southampton and Oulu and evaluate them on The USOU gait database, one of the largest gait databases containing multiple gait views.

3.1 Baseline Systems

As with face recognition, gait recognition can be performed in 2D or 3D. For comprehensiveness, we investigate both approaches as they may have different vulnerabilities to spoofing. To provide baseline performance, two systems developed at the universities of Southampton and Oulu are considered. The USOU recognition system is a 3D gait approach and the UOULU system uses 2D data. 3D Gait recognition systems have the advantage of using multiple synchronised and calibrated cameras, making video based replay attacks impractical. 3D approaches also address the difficulty of recognising subjects from different viewpoints therefore requiring that any spoofing strategy is effective when viewed from any direction. 2D approaches, using only one camera, are usually more practical as they are simpler to implement and to deploy in real-world applications. When efficiently combining the shape and motion cues, the performance of 2D approaches may easily reach that of 3D counterparts.

3.1.1 USOU Gait Recognition System

3D volumetric data is used to synthesise silhouettes from a fixed viewpoint relative to the subject. The 3D volume is synthesised from eight sychnronised camera views using shape from silhouette applied to the results of standard background subtraction approaches. The resulting silhouettes are then passed to a standard gait analysis technique; in this case the average 3D silhouette [37]. The advantage of using three-dimensional data is that silhouettes from any arbitrary viewpoint can be synthesised, even if the viewpoint is not directly seen by a camera.

As the subject walks through the tunnel the swinging of their legs alters the frontal depth of the stride. Gait sequences are detected by finding minima in this depth. A complete gait cycle consists of a left leading and right leading step. This is obtained from three consecutive minima.

Silhouettes are taken from a side-on viewpoint normal to the plane of walking. This view is not seen by any camera and so must be synthesised. The use of a side-on viewpoint facilitates comparison with previous non-3D results. To generate the average silhouette images the centre of mass is found for each frame. The average silhouette is then found by summing the centre of mass aligned silhouettes.

The derived average silhouette is scale normalised so that it is 50 pixels high, whilst preserving the aspect ratio. The average silhouette is treated as the feature vector and used for verification via the Euclidean distance metric between samples.

3.1.2 UOULU Gait Recognition System

The dynamic texture based gait recognition system [17, 18] of the University of Oulu uses dynamic texture descriptors, Local Binary Patterns from Three Orthogonal Planes (LBP-TOP), to describe human gait in a spatio-temporal way.

Firstly, a video sequence of a person's walking can be thought as spatio-temporal volume. The volume is partitioned into sub-volumes. Using the sub-volume representation, motion and shape are encoded on three different levels: pixel-level (single bins in the histogram), region-level (sub-volume histogram) and global-level (concatenated sub-volume histograms).

Secondly, LBP-TOP description is formed by calculating the LBP features from XY, XT and YT planes of volumes and concatenating the histograms to catch the transition information in spatio-temporal domain. The LBP-TOP features from each sub-volume are extracted and concatenated to encode motion and shape characteristics.

Thirdly, the length of the LBP-TOP histogram representation can be quite large depending on the number of sampling points and number of sub-volumes that are used. A better and more compact representation can be obtained by using feature selection methods. Gentle AdaBoost was used to perform feature selection and to build a strong classifier. Instead of building a classifier that gives the identity of the person from one sample, a two-class classifier was trained, which classifies weather two samples come from the same person or not.

3.2 The USOU Gait Database

The Southampton gait database [32] is one of the largest gait databases and crucially contains multiple views and detailed camera calibration information. This enables 3D reconstruction from the data, and as such provides valuable information that can be used for examining potential spoofing and countermeasure techniques. Because of this it will be used as the basis for this project.

The USOU Gait Database consists of 2705 separate recordings taken from 227 Subjects. Each recording consists of 8 synchronised video sequences of approximately 140 frames. Each subject was recorded walking through the tunnel at least 9 times.

The enrollment and test sequences were obtained on the same day, when the subject was wearing the same clothing. Significantly lower performance could be expected if subjects were to change their clothing between enrollment and validation [22]. Similarly no subjects were carrying any objects in the recorded data which could also degrade the system's recognition capability.

3.3 Performance Evaluation

50% of the dataset was selected at random to be excluded so that results from this section could be used for training purposes for later fusion experiments. Nine recordings of each of the remaining 113 subjects were selected, one for enrollment and eight as a validation test.

This leads to one enrollment video for each user and 8×113 test client (positive sample) videos for each user. When producing impostor scores all the other clients are used, yielding in 8×112×113 impostor attacks.

3.3.1 Setup

USOU Gait Recognition System:
The first stage of processing the recorded data was to convert the captured images into colour from their original raw Bayer format, using nearest-neighbour interpolation. Background estimation and segmentation was then performed to find the subject's silhouette; modelling each background pixel with a single Gaussian distribution per colour channel. The distribution for each pixel was found using previously captured video footage, where no subject was present. The background segmentation was performed by calculating the distance between a pixel and its corresponding background distribution, where a pixel would be marked as background if its distance was less than a global threshold; linked to the standard-deviation found by the background estimation. Shadow labelling and removal was performed to reduce the number of pixels incorrectly marked as foreground. Binary morphological post-processing was then performed to reduce noise levels and smooth the silhouette's shape. Finally, all regions except that with the greatest area were removed and any holes in the remaining region were filled. Radial distortion caused by the camera optics was removed by the use of a non-linear transformation.The reconstructed volumetric

data was smoothed using binary erosion and dilation morphological operators to reduce the level of noise and reconstruction artefacts. Gait cycles were detected by finding the instances where the length of the bounding box encompassing a subject is minimal in the direction of travel. This was found by fitting local polynomials to the length variation and locating potential maxima values. The time for a gait cycle was determined by finding the first maxima in the cross correlation of the length variation. This corresponds to the first half gait cycle. The first set of length peaks that were separated by the half gait cycle were identified. The first and third peaks were selected to identify a complete gait cycle. Sagittal silhouettes for each frame within this cycle were calculated. The center of mass of each of the voxels was used to align each frame on top of one another and the sagittal silhouettes were then averaged to produce a signature for each recording. Each signature was then compared to determine the average pixel difference between them. DET curves were then calculated to identify how false accept and false reject validation rates change with respect to one another, as a result of different validation thresholds.

UOULU Gait Recognition System:
The silhouette extraction process is the same as used with USOU Gait recognition system. Gait cycles were estimated using the width of the bottom part (feet) of the silhouette. The sagittal silhouette images were stacked to a space time volume and LBP-TOP features (radius 3 and number of sampling points 8) were calculated using a grid defined by the centroid of the silhouette in each frame. Histogram of the LBP-TOP features was used to represent each gait sample. A boosted classifier was trained on the training set to get a matching function between two histograms. Based on the matching scores of all samples in the test data, DET curve was calculated to identify the validation performance on different thresholds.

3.3.2 Results

USOU Gait Recognition System:
By this baseline approach, the EER is around 6%. This is an encouraging result, reflecting a high CCR in recognition based on this data (91%). We anticipate that covariate structure will reduce this capability and that the effect of clothes could be to reduce validation capability and thus would be a potential avenue for a spoofing attack. Provided subjects are recorded and validated wearing the same clothing and not carrying objects, the 3D gait recognition approach is comparable to other widely used biometrics such as 2D-face, provided such face images are recorded in a similarly unconstrained lighting setup as the gait tunnel.

From a manual examination of each of the recordings that were incorrectly classified, there are two main causes of failure: shape from silhouette distortion, and variation in arm swing. The distortion is caused by inaccurate camera calibration, which produces different body shapes at different points in the tunnel. Arm swing magnitude appears less constrained than leg dynamics. Weighting to the silhouette could be used to address this issue.

UOULU Gait Recognition System:
With this approach, EER of about 4.5% is achieved. This result is indeed encouraging. This performance can be explained by the use of spatio-temporal analysis that combines both motion and shape cues. However, the result is obtained with data that is recorded while the subjects are wearing the same clothing in all samples. We believe the performance may decrease when more covariate conditions and spoofing attacks are included, as has been observed in many studies on other databases such as the USF gait database [31]. This will be verified in next deliverables when the system will be confronted to spoofing attacks.

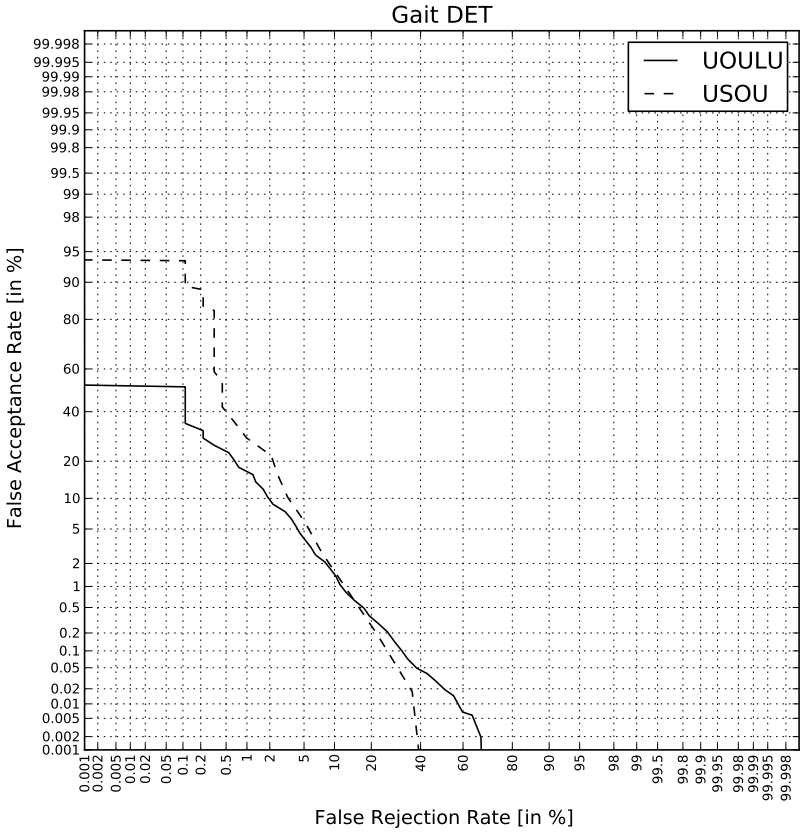

Figure 4: Detection error trade-off (DET) profiles for the UOULU and USOU gait recognition systems on the USOU gait database.

4 Vein and Fingerprint Biometrics

The fingerprint modality is well known and provides good accuracy. Unfortunately though, it can be spoofed relatively easily using fake finger. On the other hand, the emerging vein modality is very difficult to spoof but nonetheless also provides good accuracy. The main reason for this is that this modality cannot be observed without owner complicity. The fusion of fingerprint and vein modalities therefore has the potential to considerably improve recognition accuracy, and robustness to spoofing attacks. Even though the fingerprint vein modality is already reasonably mature, it is still relatively new compared to some other modalities discussed in this document and large-scale recognition evaluations are still to be organized. Several projects based on the finger vein biometric already exist and the modality itself is currently at the exchange format normalization stage (ISO-IEC 19794-9 PDAM1). Even so, it is difficult to find reliable references regarding simultaneous finger vein and fingerprint recognition. In the following we aim to establish a performance base line in the case of real fingerprint vein data that was acquired specifically for the TABULA RASA project. We describe the performance of the Morpho FingerVP (fingerprint vein) system in authentication mode.

4.1 The FingerVP System

The FingerVP system will be used for all TABULA RASA work. To develop this brand new product, Morpho implemented a multi-modal biometric recognition module. This module combines vein imaging technology (VeinID) and fingerprint identification technology. The FingerVP system performs simultaneous finger vein and fingerprint recognition. The two modalities are acquired by the same scanner at the same time. A user's vein and fingerprints can be compared to a database containing data from up to 10 000 persons. In the work reported here it is used only in authentication mode.

The complementary nature of the two modalities allows the system user to choose between different security policies. The combined identification method involving the simultaneous recognition of blood vessel patterns under the skin and fingerprints aims to offer levels of security and accuracy unrivaled worldwide [16]. Designed to be easily integrated in any type of identification system, the module meets requirements for a wide range of applications, including access control, identity checks and secure payments.

4.2 The TabulaRasaVP Database

The TabulaRasaVP database was acquired using the MORPHO FingerVP scanner. The sensor allows simultaneous acquisition of fingerprint and second phalanx finger vein patterns within two sessions. The collection of samples during two sessions aims to take into account some variability in the acquisition process: finger pose, ambient temperature, previous activity of the persons from whom samples are acquired.

The database is composed of references (templates for enrollment) and searches (templates for queries). Both contain vein and fingerprint data. References are acquired only

during session 1. Search data are captured during both session 1 and session 2. The elapsed time between session 1 and session 2 is approximately one month. Session 2 contains data from a subset of clients in session 1. Time between data collection for sessions 1 and 2 is approximately one month. During the first session however, two samples are acquired in order to confirm effective enrollment and verification within a short time span. During the acquisition sessions 6 finger images are acquired for each person. They include images of the forefinger, the middle finger and the ring finger of both hands. The little fingers were not acquired because they are very unlikely to be used in a practical scenario situation. Both genders are represented among the data.

The database is composed of a set of two images for each finger acquired: one fingerprint image and one finger vein image. In total there are 204 reference samples (102 fingers) and 288 search samples (102 fingers for session 1 and 42 fingers for session 2) which were collected from 17 persons. There are no public databases available for finger vein and fingerprints biometrics. Morpho owns some internal databases but for confidentiality reasons it is not possible to disclose any information about them. As a consequence, the TabulaRasaVP database has been collected especially for the TABULA RASA project. As described in deliverable D2.2, the initial objective was to collect data from 30 persons. But this data collection is strongly based on the voluntary participation of Morpho employees. Thus, it is highly dependent to the participation rate during data collection and unfortunately we couldn't reach the 30 subjects objective. This database is used only for evaluation. The development and training of the system were performed with Morpho's own additional, independent databases.

4.3 Performance Evaluation

The evaluation has been performed using the database and system described above.

4.3.1 Setup

Biometric samples are analyzed by a coding algorithm and a template is created for each modality from the features extracted. Similarly to fingerprints, for which minutiae are used to create templates (usually positions and orientations), vein templates are constructed using a number of singular points in the vein network. Extracted vector patterns are used to identify the coordinates of branch points and to generate a unique vein map signature based on the shape and location of the vein structure. Then, a matching is performed for the two biometrics by comparing the templates in the search database to each template in the references database. The template matching algorithm provides a consolidated score that can be compared to a given threshold in order to obtain a decision.

Note that the FingerVP system only outputs the fusion result (score) and no access to individual scores of each modality can be provided. The FingerVP system is a commercial product. Thus, for confidentiality concerns, it is not possible to disclose any detailed information about the algorithms and their precise function, in particular regarding image processing/feature extraction and templates computation.

4.3.2 Results

Figure 5 presents verification results for the Morpho FingerVP system evaluated on 98 genuine and 9702 impostors tests corresponding to session 1, and 40 genuine and 4018 impostors tests corresponding to session 2, where data were acquired one month after the enrollment session.

As we see in Figure 5, the FingerVP system delivers good performances for both session 1 and session 2, in fact the system gives perfect accuracy in the case of session 2. Concerning session 1, the first false rejections appear at around 0.01% FAR. However, it must be highlighted that FAR values lower than 0.01% are not statistically significant because of the small number of impostor tests (equal to 9702). The EER is 0.01% for session 1 whereas for session 2 the EER is 0%. Naturally these results do not mean that the system is error-free but rather that the number of tests is not sufficient to provide a reliable estimate of the EER.

We accept that the size of the test database is quite small. In consequence, the reported performance does not strictly reflect the levels of performance that can be expected in operational scenarios (i.e. deployment in the field). Even if these results therefore cannot be generalized, they represent a useful baseline to further evaluate the degradation when the system is subjected to spoofing attacks.

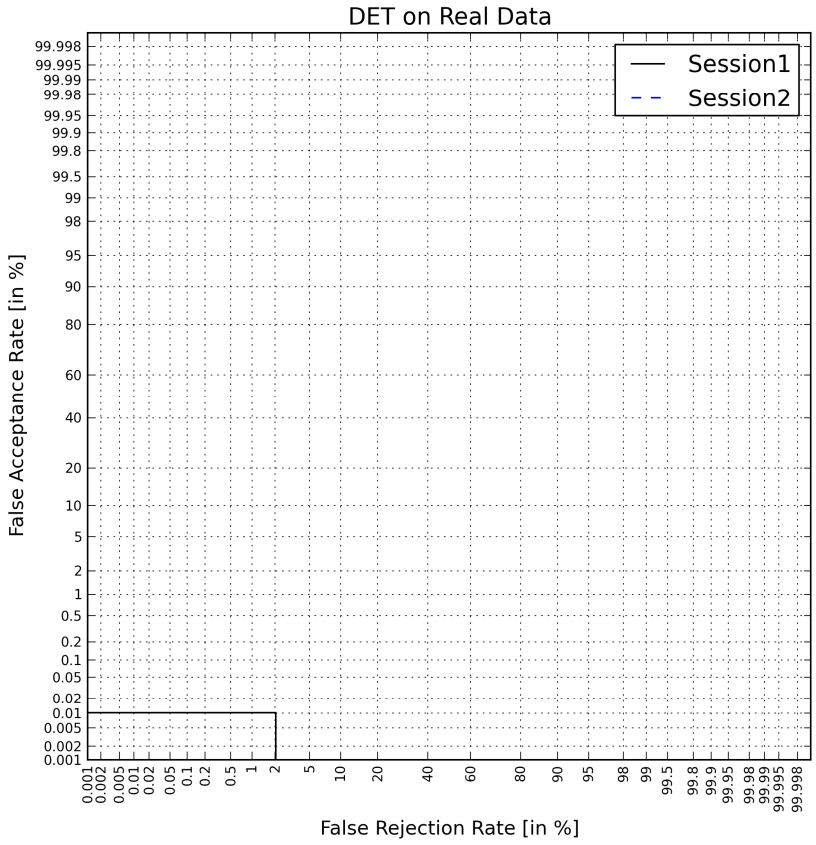

Figure 5: Detection error trade-off (DET) profile for the vein and fingerprint biometric. For session 2 the EER is 0% and thus the profile is not visible.

5 Electro-physiology Biometrics

Electroencephalogram (EEG) and electrocardiogram (ECG) biometric modalities [15, 27, 29] are quite new compared to more traditional biometric systems, such as fingerprint or voice recognition. Electro-physiological biometric systems have advantages compared to other systems. If the user is wearing an EEG/ECG recording system, such as the ENOBIO wireless and wearable amplifier developed by Starlab Barcelona SL, electro-physiological biometrics can be performed in a continuous manner. Moreover, more information besides the 'identity' can be extracted from EEG and ECG signals such as mental alertness and emotions for instance. The main challenge regarding this biometric modality is that the sensor should be as small as possible and comfortable to wear, so the user acceptance to use such a system increases.

It is important to know that, although EEG and ECG biometrics are newer biometric modalities compared to others, Starlab have been researching on this for the last 6 years and has implemented a ready to use system called Starfast. This system has been tested in real life scenarios in several pilots. The work of Starlab has focused on both Software (machine learning and methodologies to extract the best features per subject) and Hardware (mainly the ENOBIO sensor has been designed to be wireless, wearable and unobtrusive, see [30]).

Another important point is that EEG and ECG should be considered as 2 different biometric modalities. They are both categorized as electro-physiological biometrics, and both signals can be recorded at the same time with the same ENOBIO device. But in any case both signals are independent and the Starfast system can work only with EEG, only with ECG or with both. That is why the results we present here are mono-modal, whereas multi-modal results with fused EEG and ECG modalities are presented in Section 9.

5.1 The StarFast System

The Starfast system combines both EEG and ECG biometric modalities, recorded by the ENOBIO device at the same time, allowing to fuse the results of each individual modality and providing a more reliable biometric score. Of course, the Starfast system can work using only one of the two modalities. Regarding the EEG biometric system, several features are extracted (26) and, during the enrollment, the best 5 features for each individual are selected. Finally a fusion among those best features is performed. Regarding the ECG modality, the methodology is similar, but in this case only the best feature out of the 4 extracted ones is the selected one during the enrollment stage. This original methodology is explained in more detail in Section 5.3.1 below and in [29].

5.2 The Starlab Databases

Starlab has gathered 2 different databases: Eyes-closed DB and Task-Performing DB. In both cases EEG and ECG data were collected using the ENOBIO sensor, which consists of 4 active electrodes: 2 in the forehead (Fp1 and Fp2 locations) for EEG, 1 in the left wrist for ECG and finally the last one in the right earlobe as reference.

The fist dataset (Eyes-closed), was recorded in a controlled environment where the subjects were asked to sit, relax and close their eyes. The enrollment consisted of four 3-minutes takes. Then the authentication tests are about 1 minute long, again asking the subjects to relax and close their eyes. We also recorded three 3-minute takes to a set of 32 subjects. This database, that we call external database, is used as reference subjects in the classification stage. The other enrolled subjects are the ones used to test the system. We have 8 enrolled subjects with several authentication takes for each one of them. We have a total of 50 legal transactions, in which a subject claim to be himself and a total of 366 impostor transactions, in which a subject claims to be another subject from the enrolled set of subjects.

For the second dataset (Task-performing), the subjects were recorded keeping their eyes open, sitting on a chair, and they were free to perform a number of office activities (such as answering the phone, keystroke, drinking water, using the mouse...) during the authentication phase. For the enrollment, in this case, the subjects were asked to sit keeping their eyes open, but without performing any tasks. 29 subjects were enrolled in both the EEG and the ECG module and then they where authenticated while performing the different tasks defined above. From this set of subjects we successfully could evaluate 80 legal transaction and 2188 impostor transactions.

As a note, both the legal and the impostor transactions are needed in order to compute the Equal Error Rate (EER) and to extract the DET curves in Section 5.3.2.

5.3 Performance Evaluation

The evaluation has been done using both databases briefly described above: the Eyes-closed DB and the Task-performing DB. In the later, we have to take into account that we will have more movement and ocular artifacts than in the Eyes-closed DB, and thus the biometric performance will be affected. Within the scope of TABULA RASA it will be of interest later on to determine whether these differences impact on spoofing and countermeasures performance.

5.3.1 Setup

The Starfast matching algorithm uses an original methodology we call 'personal classifier'. For instance in the case of EEG, we record 4 enrollment takes of 3 minutes each. We cut each one of the takes in 4 second epochs, and out of each epoch we extract 5 different features. For each channel we compute the Auto-regression Coefficients and the Fourier Transform and for each pair of channel we compute the Mutual Information, Correlation and Coherence. So we end up with 7 different features. We use a Linear Discriminant Analysis with 4 different Discriminant Functions (see [10]). By combining classifiers and features we have 7*4=28 possible combination of features/classifiers.

Applying a leave-one-out approach to our 4 enrollment takes we compute the 5 best combination of feature/classifier (out of 28) for each subject. This information is stored in the template of each subject, and thus, in the authentication stage, the corresponding

Database	Modality	EER (%)
Eyes-closed DB	EEG	20
Eyes-closed DB	ECG	4
Task-performing DB	EEG	27
Task-performing DB	ECG	14

Table 2: Equal error rate (EER) scores (%) for EEG and ECG modalities and for both eyes-closed and task-performing databases.

template is loaded and the classification is performed using the 5 best combination of feature/classifier of the claimed subject.

Regarding the ECG modality, the approach is quite similar. But in this case, we only use one feature related with the shape of the ECG waveform but again we use 4 different discriminant functions. In this case we store in the template the best discriminant function for each subject.

5.3.2 Results

As we have 2 different databases recorded in different conditions, we decided to evaluate both separately. We should take into account that the recording conditions are different in both cases, that is in the Eyes-closed set the subjects are relaxed, sitting down and keeping their eyes closed, while in Task-performing set the subjects are free to move while sitting down in a office scenario. In the second case there are more artifacts that in the former case, and thus the performance is expected to drop.

A summary of the results follows in terms of the Equal Error Rate (EER). The EER is reached when the False Rejection Rate (FRR) equals the False Acceptance Rate (FAR).

As expected, we see that the performance decreases in the Task-performing data set. This is because in the Task-performing dataset, the subjects keep their eyes open and thus eye movement and blink artifacts are present in the EEG signal. As the subjects in this dataset are also moving (performing normal office activities) there are also movement artifacts in the ECG channel. The presence of this artifacts cause this decrease of performance. We can also see that the ECG modality is more robust that the EEG one. Finally, it is interesting to note that in the case of the Eyes-closed database and the ECG modality, we can keep a FRR equal to 4% while decreasing the FAR to 2,5% as we can see in Figure 6.

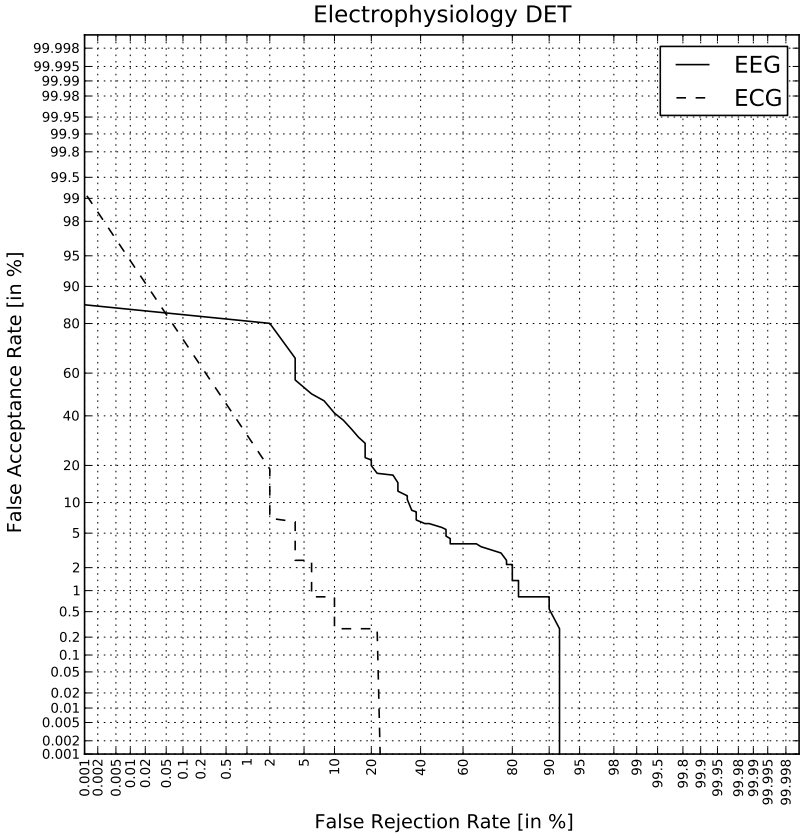

Figure 6: Detection error trade-off (DET) profile for ECG and EEG modalities and the eyes-closed database.

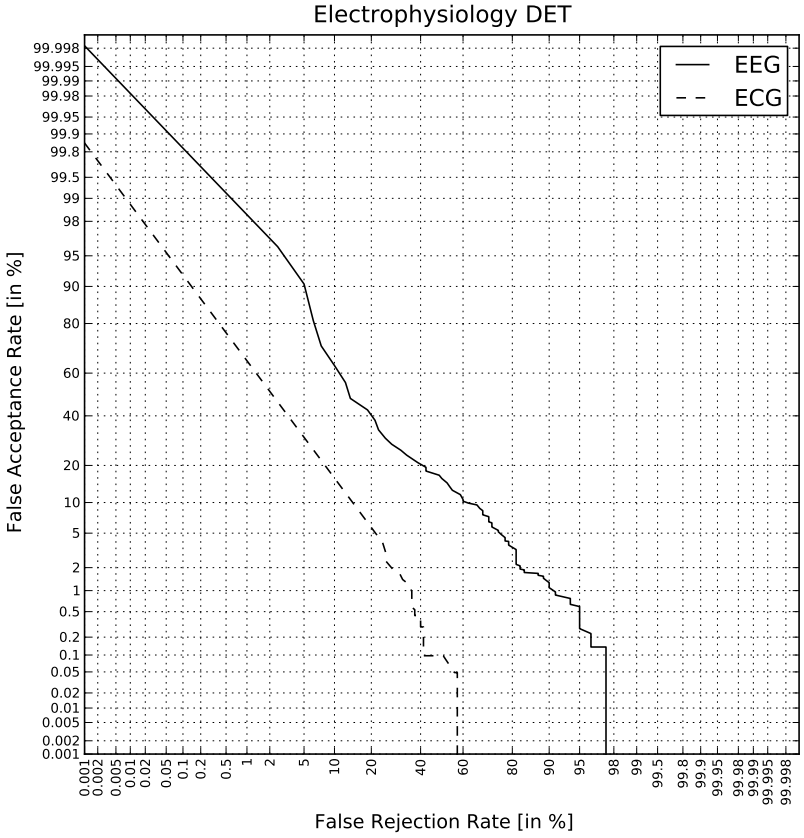

Figure 7: Detection error trade-off (DET) profile for ECG and EEG modalities and the task-performing database.

6 Multi-Modal Biometrics: 2D-Face and Voice

In this section we show the performance of the 2D-face and voice multi-modal biometric. The aim is not strictly to demonstrate the gain in performance over the individual modalities (though it is certainly expected) but to establish baseline multi-modal performance for subsequent work in spoofing and countermeasures. Multi-modal biometrics are considered within the context of TABULA RASA since they provide an inherent level of protection against spoofing; the presence of more than one modality might require all biometric traits to be spoofed simultaneously in order to fool the system.

The performance of multi-modal fusion depends fundamentally on that of individual systems but also requires that the set of biometric traits are decorrelated in order to exploit their complementarity. While correlated features often arise in pattern recognition systems based on multiple classifiers, many publications have shown that, statistically, multiple biometric traits are mostly decorrelated and thus there is significant potential for multi-modal biometric fusion.

In the following we describe the two basic fusion systems used for 2D-face and voice multi-modal biometrics. Both approaches, as is the case with all multi-modal work in TABULA RASA, involve so-called 'score-level fusion', i.e. scores provided by the individual systems are combined through an additional function, and a single, unique score is derived.

6.1 The Systems

In this Section, we first briefly summarise the individual systems and fusion approaches. Fusion produces a multi-modal score which is used to compute the performance of the multi-modal biometric. Only the 2D-face recognition system is summarised here; the voice recognition system is presented in Section 2.1.

6.1.1 2D-face recognition system

The face authentication system is the baseline system described in Section 2.2 of D2.2. This system consists of dividing the face into overlapping blocks (or parts) and a GMM is then trained by considering each block (part) to be a separate observation of the same underlying distribution (a face). Two models are trained, one is a world model (Ω_{model}) that describes the distribution of all faces and the second is a client model (Ω^i_{client}) which describes the distribution for a particular client's face. To verify if an observation belongs to a client, or not, it is scored against both the client (Ω^i_{client}) and world (Ω_{model}) model. The two models, Ω^i_{client} and Ω_{world}, produce a log-likelihood score which is then combined using the log-likelihood ratio (LLR):

$$h(x) = \ln(p(x \mid \Omega^i_{client})) - \ln(p(x \mid \Omega_{world})), \tag{1}$$

to produce a single score.

This score is used to assign the observation to the world class of faces (not the client) or the client class of faces (it is the client) and consequently a threshold τ has to be applied

to the score $h(x)$ to declare (verify) that x matches to the i^{th} client model Ω^i_{client}, i.e if $h(x) \geq \tau$. More details for this system can be found in Section 2.2 of D2.2.

6.1.2 The MITSfusion system

The MITSfusion system, which has been used in the evaluation of both 2Dface-voice and EEG-ECG fusion, implements a so-called iterative fusion operator tree. Being the system applied herein to bi-modal data, the tree presents just a node. An early version of the system is presented in [34].

The unique node of the tree is first characterized by the used fusion operator. In its current implementation the methodology allows the selection of one fusion operator among the following: power mean, weighted sum, order weighted averaging (OWA), Yager S-norm, uni-norm based on Yager norms and arithmetic mean, Choquet fuzzy integral, and Sugeno fuzzy integral. The selected fusion operator is denoted as $\mathbf{FO}_0$. Second, the node is characterized by the set of fusion operator parameters, which is denoted herein as $\mathbf{p}_0$. The parameter set is selected through an extensive search in the parameter space of the corresponding fusion operator.

The final feature of the MITSfusion system indicates if the parameter set is either personalized for each of the subjects of the database or if a general parameter set is used. In the first case we assign a parameter set to each of the subjects, so that $\mathbf{p}_0$ becomes a superset defined as:

$$\mathbf{p}_0 = \{\mathbf{p}_{01}, \ldots, \mathbf{p}_{0L}\}, \tag{2}$$

where L is the number of subjects. In the second case there is a unique parameter set $\mathbf{p}_0$ for all subjects. The search of all the optimal operators, parameters, and the usage or not of personalized fusion operators is driven by the maximization of the Area Under the Curve (AUC).

As mentioned in the former paragraphs we have evaluated the following fusion operators $\mathbf{FO}_0$:

Power or Generalized Mean. The mean is one of the most well-know fusion operators. Beside the most used arithmetic mean, there are other mean operators like the geometric mean or the harmonic mean. A parametric generalization of all these expressions has been proposed [4], which is known as the power or generalized mean,

$$z = \Big(\frac{1}{n}\sum_{i=1}^{n} x_i^m\Big)^{1/m}, \tag{3}$$

whose value depends on the real-valued parameter m, e.g. for $m = 1$ results in the arithmetic mean and for $m = 2$ is denoted as the quadratic mean.

Yager S-norm. T- and S-norms are aggregation operators related with the concept of statistical metrical spaces. T- and S-norms were adopted in fuzzy systems for operating

with fuzzy membership functions [20]. The Yager S-norm has been selected herein after a preliminary study taking the diversity of operators to be analyzed into consideration. This S-norm presents the following expression:

$$z = \min\{1, (x_1^p + x_2^p)^{1/p}\}, \tag{4}$$

where $p \in [0, \infty]$.

Weighted Sum. The weighted sum is an operator used in different application domains, e.g. descriptive statistics, neural networks. It is a further generalization of the arithmetic mean. In this case the generalization is done by weighting the input values, i.e.

$$z = \sum_{i=1}^{n} w_i x_i. \tag{5}$$

Usually the sum of the weights is normalized to sum to 1, which ensures that we are working in the unit hypercube.

Uninorm Based On Yager Norms. Uni-norms were introduced in [39]. Uni-norms generalize T- and S-norms by introducing an arbitrary neutral element denoted as e defined in $[0, 1]$ such that $U(x, e) = x$. One can see uni-norms and absorbing-norms as two different ways of combining T- and S-norms in the unit hypercube. Thus in the uni-norms the subspace $[0, e] \times [0, e]$ is occupied by a T-norm, whereas $[e, 1] \times [e, 1]$ by a S-norm. In the remaining two sub-spaces there is a compensatory operator, although this is not a condition of the operator (i.e. the only condition is that the resulting operator must be commutative and associative). Moreover these two quadrants have to be filled by compromise operators like means or min/max itself. We have selected a uni-norm based on the Yager T- and S-norms, and on the arithmetic mean in the U-quadrant. This can be expressed as:

$$z = \begin{cases} \max\{0, 1 - ((1 - x_1)^p + (1 - x_2)^p)^{1/p}\} & : \quad x_1, x_2 \in [0, e] \times [0, e,] \\ \min\{1, (x_1^p + x_2^p)^{1/p}\} & : \quad x_1, x_2 \in [e, 1] \times [e, 1] \\ \frac{x_1 + x_2}{2} & : \quad otherwise \end{cases}, \tag{6}$$

where (e, p) are the parameters of this uni-norm.

Ordered Weighted Averaging. A generalization of the average, where the weighting is established after sorting the input data, was proposed in [38] and denoted as Ordered Weighted Averaging (OWA). The OWA presents the following expression:

$$z = \sum_{i=1}^{n} w_{(i)} x_{(i)}, \tag{7}$$

where $w_{(i)}$ are the weights of the operator. The bracketed subindices state for a sorting operation that is applied on x_i before aggregating their values, e.g. (1) state for the larger

x_i, (n) for the lowest one.

Sugeno Fuzzy Integral. The concept of fuzzy integral was first proposed in [35] as a means of fusing data simulating subjective multi-criteria evaluation undertaken by humans. The operator present some similarities to the OWA described in the former section, since the applied weighting depends on the particular canonical subspace of the input variables in the unit hypercube, i.e. on the result of a sorting operation. In contrast to the OWA, the weighting set in the fuzzy integral is not unique, but it changes in each canonical region.

The fuzzy integral uses as fusion operators a combination of T- and S-norms [20]. There are different types of fuzzy integrals, which are defined by the used T- and S-norm. The Sugeno fuzzy integral [35] combines the minimum and maximum operators as expressed for the case of n operands operands in:

$$z = \bigvee_{i=1}^{n} \mu(A_{(i)}) \wedge x_{(i)}, \tag{8}$$

where $\vee$ states for the maximum operator, $\wedge$, for the minimum, and $\mu()$, for the coefficients of the so-called fuzzy measures, i.e. the weighting coefficients in the fuzzy integral. There are 2^{n-1} coefficients, one for each subset that can be established on the n information sources to be fused. As formerly mentioned, the bracketed indices represent the result of a sorting operation. Hence only n coefficients of the fuzzy measure are selected for the aggregation. These coefficients correspond to the subsets: $A_{(1)} = \{x_{(1)}\}, A_{(2)} = \{x_{(1)}, x_{(2)}\}, \ldots, A_{(n)} = \{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$. Therefore the actual weight set for each aggregation depends on the canonical region defined by the input variables.

Choquet Fuzzy Integral. In case of the Choquet fuzzy integral [14] the maximum and the minimum are respectively substituted by the sum and the product:

$$z = \sum_{i=1}^{n} x_{(i)} \cdot [\mu(A_{(i)}) - \mu(A_{(i-1)})], \tag{9}$$

where $\mu(A_{(0)}) = \mu(\emptyset) = 0$. As it can be proofed, the Choquet fuzzy integral generalises the weighted sum and the OWA operators.

6.1.3 UNICA fusion

UNICA's fusion system again operates at the score level by applying different fixed and trained fusion rules. The prime rules are referred to as 'fixed' since they are non-parametric in contrast to the trained rules.

The fixed fusion rules are given by:

Simple Average Rule (**SAR**):

$$s_{sar} = s_{voice} + s_{face} \tag{10}$$

Simple Product Rule (**SPR**):

$$s_{spr} = s_{voice} \cdot s_{face} \tag{11}$$

Bayes Rule (**BayesR**):

$$s_{bayes} = \frac{s_{voice} \cdot s_{face}}{s_{voice} \cdot s_{face} + (1 - s_{voice}) \cdot (1 - s_{face})} \tag{12}$$

Minimum Rule (**MinR**):

$$s_{min} = min\{s_{voice}, s_{face}\} \tag{13}$$

Maximum Rule (**MaxR**):

$$s_{max} = max\{s_{voice}, s_{face}\} \tag{14}$$

where s_{voice} and s_{face} are scores for the 2D-face and voice modalities, respectively.

The trained fusion rules are given by:

Likelihood Ratio Rule (**LRR**):

$$s_{lrr} = \frac{p(s_{voice}|genuineusers) \cdot p(s_{face}|genuineusers)}{p(s_{voice}|impostors) \cdot p(s_{face}|impostors)} \tag{15}$$

where e.g. $p(s_{voice}|impostors)$ is the score of probability distribution for impostors, estimated on the training dataset, calculated on s_{voice}.

Weighted Average Rule (**WAR**):

$$s_{war} = w_{voice} \cdot s_{voice} + w_{face} \cdot s_{face} \tag{16}$$

Weighted Product Rule (**WPR**):

$$s_{wpr} = s_{voice}^{w_{voice}} \cdot s_{face}^{w_{face}} \tag{17}$$

Logistic Based Fusion (**LBF**):

$$s_{lbf} = \frac{1}{e^{-(w_{voice} \cdot s_{voice} + w_{face} \cdot s_{face})}} \tag{18}$$

where $w_{voice}, w_{face} \in [0, 1]$ and $w_{voice} + w_{face} = 1$ are optimised through training.

6.2 The MOBIO Database

The challenging MOBIO database is used to evaluate the systems. This publicly available bi-modal (audio and video) database was captured at six different sites across five different countries. It was captured over a one-and-a-half year period and consists of 150 participants with a female to male ratio of approximately 1:2 (99 males and 51 females). The database was recorded using two mobile devices: a mobile phone and a laptop computer (the laptop was only used to capture part of the first session).

Standard MOBIO protocols are used to evaluate both multi-modal systems. This protocol divides the database into three distinct sets: one for training, one for development and one for testing. The splitting was done so that each set is composed of the totality of the recording from two sites. This means that there is no information regarding the individuals or the conditions for a site between sets. We note that only scores from the development set were used for optimisation. The third dataset, normally used for the learning of background models etc. was not used in any way for multi-modal work. More details about this database can be found in Section 2.1.2 of D2.2.

6.3 Performance Evaluation

For both systems under evaluation here, the MOBIO development dataset was used to optimise all algorithms whereas all DET curves reported below correspond to the test dataset. EERs are, however, given for both development and test sets where appropriate.

6.3.1 Setup

MITS Fusion:
The MITSfusion system was used with just one node, where features are determined in the training phase. Hence we have to assess:

- Employed fusion operator $\mathbf{FO}_0$;
- Usage of personalized operators vs. a general one;
- Optimal parameter set $\mathbf{p}_0$ (or superset in case of personalized ones).

We applied the following procedure:

1. We apply all available fusion operators as $\mathbf{FO}_0$.
2. For each fusion operator we look for the parameter set $\mathbf{p}_0$ and superset $\{\mathbf{p}_{01}, \ldots, \mathbf{p}_{0L}\}$ that maximize the AUC.
3. We compute the average AUC over subjects when:
   (a) Applying a unique parameter set $\mathbf{p}_0$, which we denote as $\overline{AUC}_g$.

Dataset	Female	Male
2D-Face (Dev)	10.0%	10.8%
2D-Face (Test)	19.8%	12.0%
Voice (Dev)	21.5%	16.6%
Voice (Test)	18.4%	15.9%

Table 3: Baseline equal error rates (EERs) (%) for the development and evaluation/test subsets of the MOBIO database for both 2D-face and voice modalities.

(b) Applying a parameter superset $\{\mathbf{p}_{01}, \ldots, \mathbf{p}_{0L}\}$, which we denote as $\overline{AUC}_p$.

4. We select the fusion operator with maximal $\overline{AUC}_p$.

5. If the difference between $\overline{AUC}_p$ and $\overline{AUC}_g$ is less than 3%, we personalize the operator, i.e. $\mathbf{FO}_0$ is parameterized with $\mathbf{p}_0$. If not, we parameterized it with one parameter for each subject, i.e. $\{\mathbf{p}_{01}, \ldots, \mathbf{p}_{0L}\}$.

The development dataset was used to optimise the MITSfusion system (see procedure described above). Once the fusion operator, optimal parameters and personalization setting are determined, they are applied to the test dataset. The setup is the same for both male and female datasets.

UNICA Fusion:
All scores are first normalized according to the Max-Min rule. Fixed fusion rules are then applied directly to the test dataset whereas trained fusion rules are first trained on the MOBIO development dataset. Likelihood Ratio Rule training consists of estimating the probability of genuine and impostor classes where scores are normalized on a logarithmic scale. Other trained rules are optimised by setting the weights to minimize the error prediction function with 1000 epochs. This function evaluates distances from the impostor and genuine scores to 0 and 1. Finally, for all experiments, scores are further normalized.

6.3.2 Results

While it is not strictly the objective of this work to compare mono-modal and multi-modal biometric performance, the EERs for the 2D-face and voice modalities are presented in Table 3 for both development and test subsets of the MOBIO database and for both the female and male subsets. We present below multi-modal performance according to the two different fusion systems.

MITS Fusion:
As a result of the application of the assessment procedure on the development dataset we selected the weighted sum as fusion operator. The operator is not personalized and the

Dataset	Gender	EER (%)
Development	female	5.5
Development	male	6.0
Test	female	10.2
Test	male	6.0

Table 4: Equal error rate (EER) scores (%) for fused 2D-face and voice modalities of the MOBIO test dataset using MITS Fusion.

optimised parameter set corresponds to 0.9 weighting factor for the 2D-face modality and 0.1 for the voice modality.

A summary of the results in terms of the EER is given in Table 4. Figure 8 illustrates the corresponding DET curves for the test dataset and both gender subsets. Upon comparison with baseline performance in Table 3 we note greatly improved performance for the multi-modal biometric.

UNICA Fusion:
DET curves for the UNICA fusion systems and female datasets are illustrated in Figures 9 and 10. Corresponding EERs are illustrated in Tables 5 and 6. All DET plots and EERs relate to the MOBIO test set. Figure 9 and Table 5 report results for fixed rules. We note an improvement with respect to the best mono-modal biometric. This is particularly true for the BayesR approach. The improvement is notable especially given the difference in performance among the mono-modal biometrics.

Figure 10 and Table 6 report results for trained rules. Even in this case, we see that EERs are lower than the best baseline biometric, but the performance is lower than that of fixed fusion rules. This is maybe due to the performance unbalance among baseline biometrics, but also due to the fact that training patterns (from the MOBIO development set) are not representative of testing patterns, as can be noticed by observing the difference in performance for development and test subsets reported in Table 3. This can be seen from the performance of the face biometric in particular. Another interesting result is the one related to the LLR rule. This rule is claimed to be the 'optimal' one, according to hypothesis verification test theory. However, it can be noticed that this is not true due, in our opinion, to the presence of poorly representative patterns in the training set.

DET curves for the male datasets are illustrated in Figures 11 and 12. Corresponding EERs are included in Tables 7 and 8. All DET plots and EERs again refer to the MOBIO test set. Figure 11 and Table 7 report results of fixed rules. All rules perform better than baseline systems. In particular, BayesR gives the best results. Even in this case, a performance imbalance is noticeable among the baseline biometrics. However, it is not so high as in the case of the female subset.

Figure 10 and Table 6 report results of trained rules. EERs are lower than that of the best baseline matcher and the performance is not so different with respect to fixed fusion

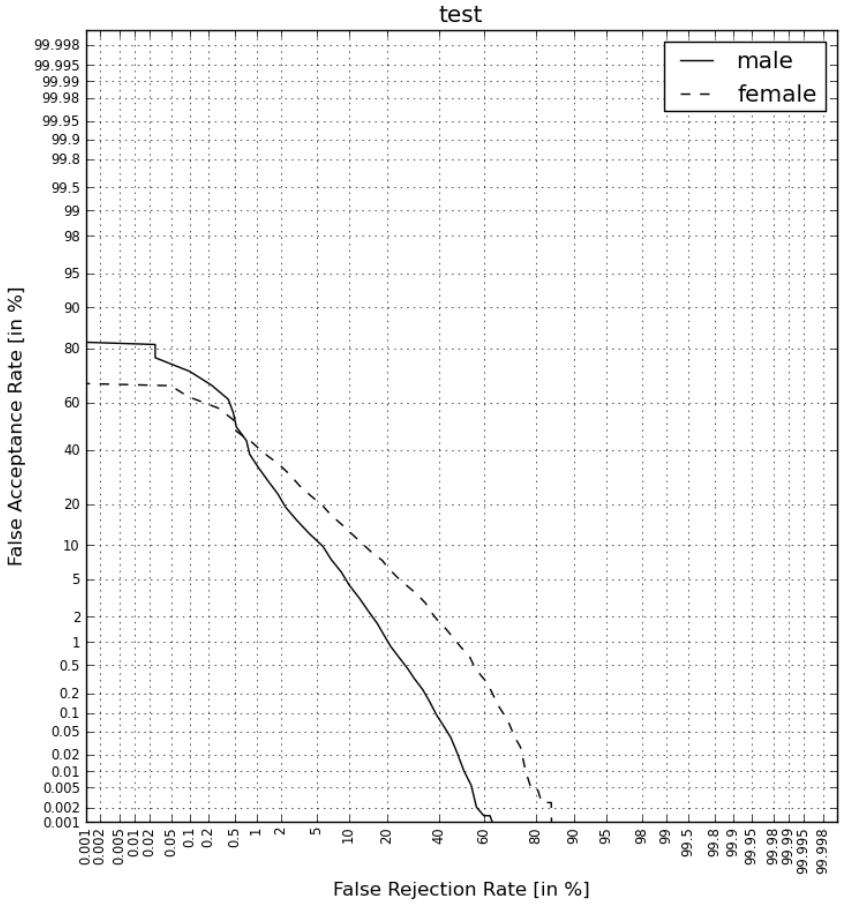

Figure 8: Detection error trade-off (DET) profile for fused 2D-face and voice modalities of the MOBIO test dataset using MITS Fusion.

rules. We notice in this case that LLR performs better than in the case of the female dataset; this may be due to the similar levels of performance for the two mono-modal biometrics. This can be also noticed by observing the performance of WAR, which is the best trained rule among those investigated.

These results show that almost all fusion rules lead to improved classification performance. Only maximum and minimum rules show no improvements in EER but, even in these cases DET curves show significant improvements in FAR_{zero} and FRR_{zero}. We can hypothesise that trained rules could perform better by using a different estimation error function but, in general, they strongly depend on training samples, which must be representative of test samples. Results with an LLR rule, which is commonly considered a reference decision rule even for baseline systems (see for example the 2D-face system used here), exhibit slightly worse performance than SAR and BayesR.

Dataset (Female)	EER
SAR	11.8%
SPR	11.8%
BayesR	11.8%
MinR	17.1%
MaxR	14.3%

Table 5: Equal error rate (EER) scores (%) for fused 2D-face and voice modalities using fixed UNICA fusion rules and the female subset of the MOBIO test dataset.

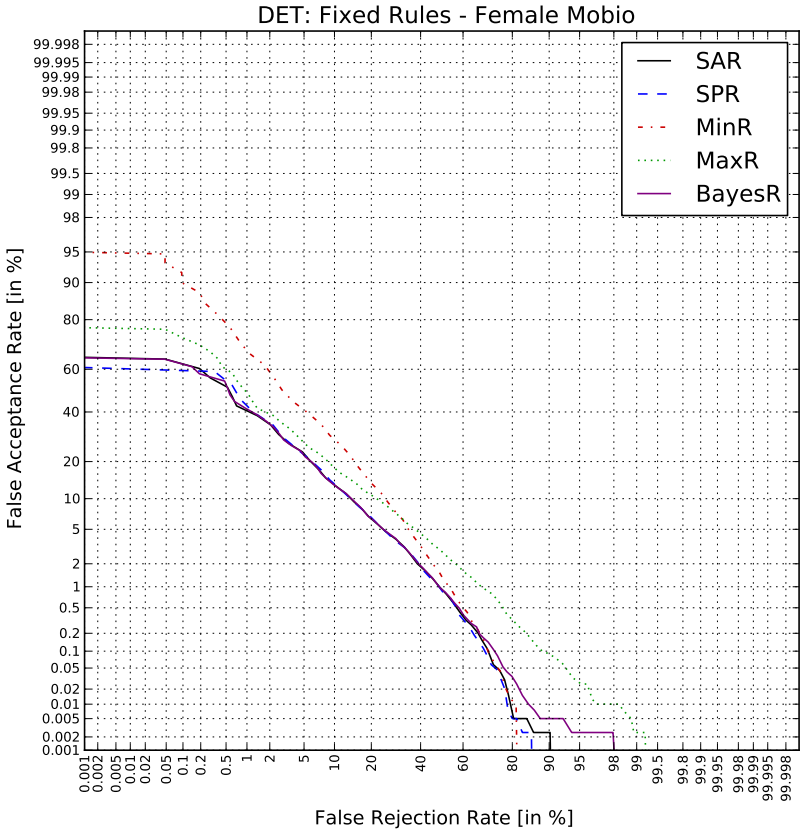

Figure 9: Detection error trade-off (DET) profile for fused 2D-face and voice modalities using fixed UNICA fusion rules and the female subset of the MOBIO test dataset.

Dataset (Female)	EER
WAR	12.0%
WPR	12.8%
LRR	14.3%
LBF	12.4%

Table 6: Equal error rate (EER) scores (%) for fused 2D-face and voice modalities using trained UNICA fusion rules and the female subset of the MOBIO test dataset.

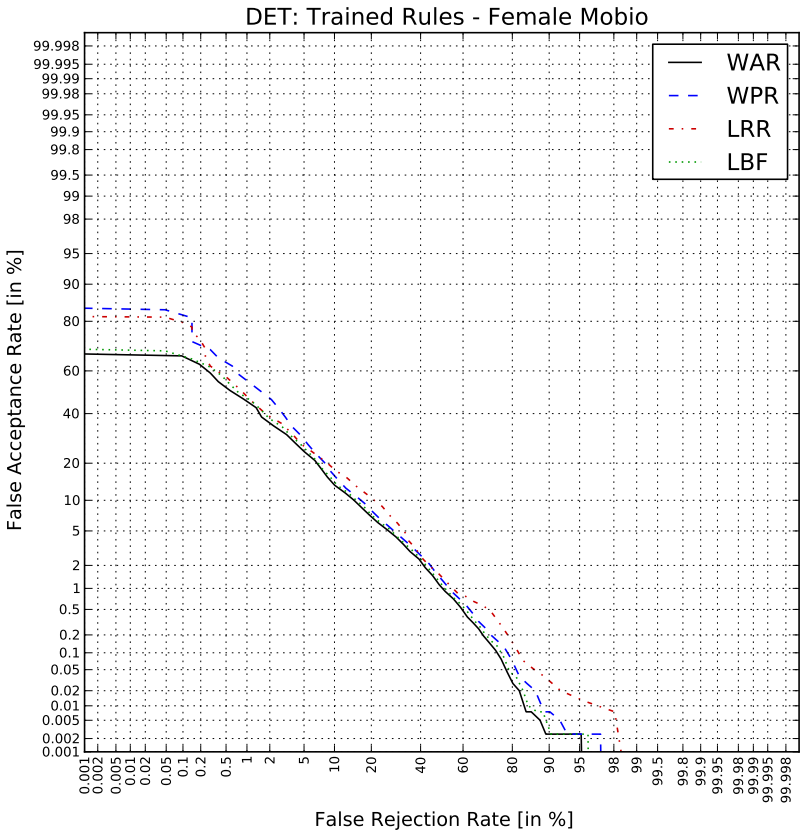

Figure 10: Detection error trade-off (DET) profile for fused 2D-face and voice modalities using trained UNICA fusion rules and the female subset of the MOBIO test dataset.

Hence, fixed fusion rules show the best performance. Simple average and 'Bayes' rules are the most effective in EER reduction. Among trained rules, LLR maintains good performance, similar to that of the strongest fixed rules. We noticed that, for both male and female data subsets, more that 5% EER improvement over the best baseline systems

are obtained. This confirms what has been concluded in several publications, i.e. that score-level fusion is able to recover errors caused by the lack in performance of individual biometric verification systems.

Dataset (Male)	EER
Voice	15.9%
Face	12.0%
SAR	7.3%
SPR	7.9%
BayesR	7.3%
MinR	11.9%
MaxR	12.8%

Table 7: Equal error rate (EER) scores (%) for fused 2D-face and voice modalities using fixed UNICA fusion rules and the female subset of the MOBIO test dataset.

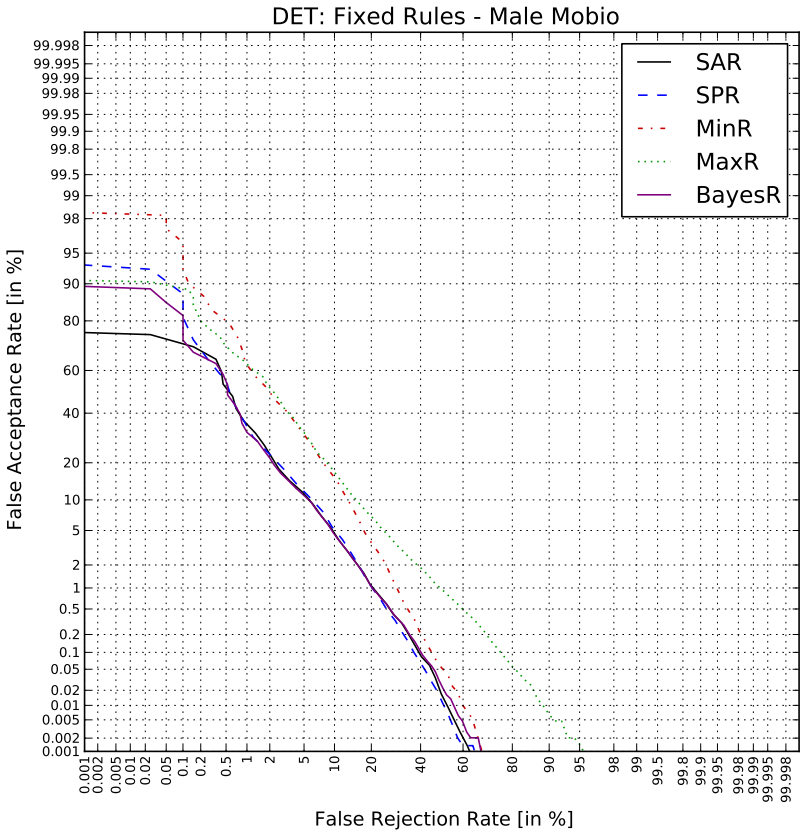

Figure 11: Detection error trade-off (DET) profile for fused 2D-face and voice modalities using fixed UNICA fusion rules and the male subset of the MOBIO test dataset.

Dataset (Male)	EER
WAR	7.9%
WPR	8.4%
LRR	8.0%
LBF	11.0%

Table 8: Equal error rate (EER) scores (%) for fused 2D-face and voice modalities using trained UNICA fusion rules and the male subset of the MOBIO test dataset.

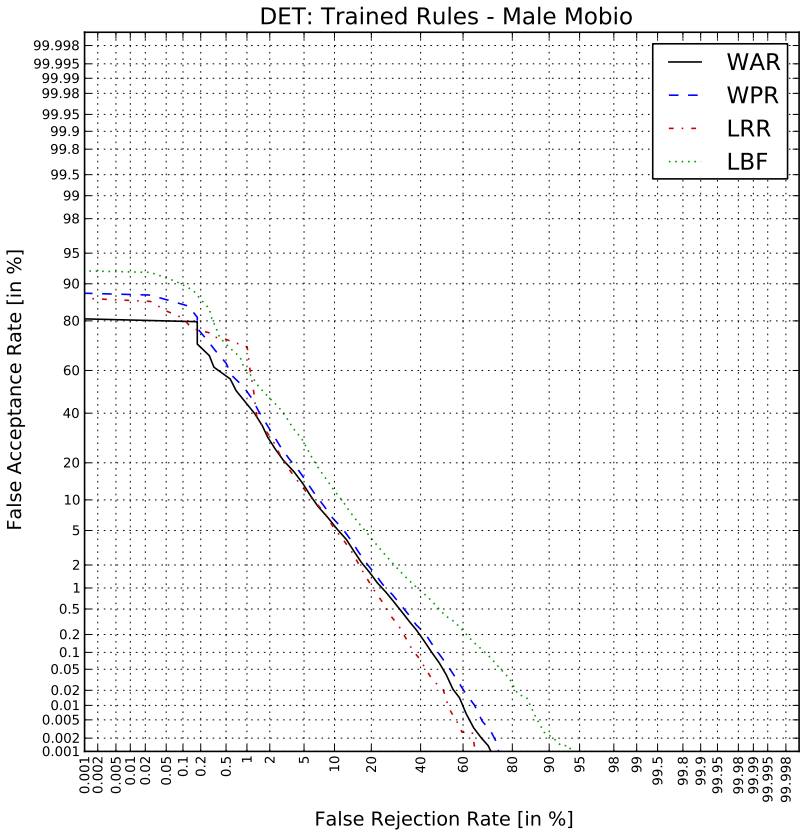

Figure 12: Detection error trade-off (DET) profile for fused 2D-face and voice modalities using trained UNICA fusion rules and the male subset of the MOBIO test dataset.

7 Multi-Modal Biometrics: 2D-Face and Fingerprint

In this section we show the performance of multi-modal fusion systems when combining fingerprint and face modalities. Fusion is once again performed at the score-level by applying different fixed and trained fusion rules.

7.1 The Systems

The 2D-face recognition system is the same as that summarised in Section 6.1.1. A baseline evaluation of the fingerprint recognition system is presented in the companion deliverables D3.1; a summary of the recognition system is presented below. The fusion system is exactly the same as the UNICA fusion system described in Section 6.1.3 of this document except where s_{voice} and w_{voice} are replaced by scores and weights corresponding to the fingerprint modality.

7.1.1 Fingerprint recognition system

The minutiae-based NIST Fingerprint Image Software 2 (NFIS2) [12] is a minutiae-based fingerprint processing and recognition system formed from independent software, which constitutes a de facto standard reference system used in many fingerprint-related research contributions.

From the different software modules that are comprised within NFIS2, the most relevant for evaluation purposes are: MINDTCT for minutiae extraction, and BOZORTH3 for fingerprint matching.

MINDTCT
The MINDTCT system takes a fingerprint image and locates all minutiae in the image, assigning to each minutia point its location, orientation, type, and quality. Together with the minutiae set, a quality map of the input image is also generated using characteristics such as low contrast, incoherent ridge flow, and high curvature. Additionally MINDTCT computes ridge counts between a minutia point and each of its nearest neighbors.

BOZORTH3
The BOZORTH3 matching algorithm computes a match score between the minutiae from two fingerprints to help determine if they are from the same finger. It uses only the location and orientation of the minutiae points to match the fingerprints, and it is rotation and translation invariant.

7.2 The BioSecure Database

The evaluation was performed on 8 datasets of the BioSecure database. It is a non-chimerical dataset, where fingerprints and faces are captured from the same subjects, but, depending on each biometric (face/fingerprint), according to different environmental

EER (Equal Error Rate)		
Dataset	dev	test
ca0	0.7%	1.4%
ca0_PCA	21.5%	18.6%
caf	1.4%	0.7%
caf_PCA	18.6%	19.3%
wc	2.1%	1.4%
wc_PCA	27.9%	24.2%
op	0.8%	0.0%
th	0.7%	0.7%

Table 9: Baseline equal error rates (EERs) (%) for each of the 8 different subsets of the BioSecure database for both development and evaluation/test datasets.

conditions/type of sensor. Accordingly, there are 6 2D-face datasets and 2 fingerprint datasets giving a total of 12 face-fingerprint fusion conditions. The face datasets are denoted $ca0$, caf and wc, and are used with or without principal component analysis (PCA), whereas the fingerprint datasets are denoted op and th. Each dataset is divided into development and test subsets which are used to train, validate and test the trained rules, though fixed rules need only the test dataset. Each development and test dataset consists of 9800 samples: 9660 impostors and 140 genuine users. Table 9 shows the baseline mono-modal EER for each dataset.

7.3 Performance Evaluation

We investigated the same score-level fusion rules outlined in Section 6, where the outputs of individual recognition systems are combined through a 'fusion rule'.

7.3.1 Setup

The experimental setup is exactly as described in Section 6.3.1 for UNICA fusion except that s_{voice} and w_{voice} are replaced by scores and weights corresponding to the fingerprint modality.

7.3.2 Results

Significant performance improvements are achieved for the SAR, SPR and BayesR fixed fusion rules and for the WAR, WPR and LBF trained fusion rules. Other rules did not lead to better performance or results were not consistent in all experiments. Tables of corresponding EERs for each fusion rule are presented below with DET curves for the best and worst fusion results, in order to give an indication of the range in performance. In Tables 10 and 11 the first column indicates which datasets were used for fusion and the

subsequent columns show the EER for each fusion rule. Figures 13 and 14 show DET curves for fixed and trained fusion rules for the *ca*0 face dataset and the *op* fingerprint dataset. In this case genuine and impostors classes are so well separated that they do not figure in the plot. On the other hand, we show the worst DET curves in Figures 15 and 16. Here plots relate to the fusion of *wc_PCA* (2D-face) and *th* (fingerprint) BioSecure datasets.

The main problem with the BioSecure datasets is the high level of imbalance between fingerprint and 2D-face PCA-based recognition algorithms; additionally, fingerprint recognition performance is very good. Consequently, the relative reduction in EER is very low: from 0.7% to 0.0% when fusing fingerprint and face modalities. On the other hand, results can be interpreted differently: even with very good fingerprint recognition performance, fusion still results in an EER of 0% (see for example results related to fusion in Tables 10-11).

EER of fixed fusion rules					
face & finger	SAR	SPR	MinR	MaxR	BayesR
ca0 & op	0.0%	0.0%	0.0%	0.0%	0.00%
ca0_PCA & op	0.7%	0.7%	0.7%	8.6%	0.7%
ca0 & th	0.0%	0.0%	0.7%	0.0%	0.0%
ca0_PCA & th	1.4%	0.7%	1.4%	9.2%	0.7%
caf & op	0.0%	0.0%	0.0%	0.0%	0.0%
caf_PCA & op	0.6%	0.1%	0.0%	6.5%	0.0%
caf & th	0.0%	0.0%	0.7%	0.0%	0.0%
caf_PCA & th	0.7%	0.7%	0.7%	6.6%	0.7%
wc & op	0.0%	0.0%	0.0%	0.0%	0.0%
wc_PCA & op	0.6%	0.0%	0.0%	13.7%	0.0%
wc & th	0.0%	0.0%	0.7%	0.0%	0.0%
wc_PCA & th	2.1%	1.4%	0.7%	14.3%	1.4%

Table 10: Baseline equal error rates (EERs) (%) for fixed fusion of 2D-face and fingerprint modalities.

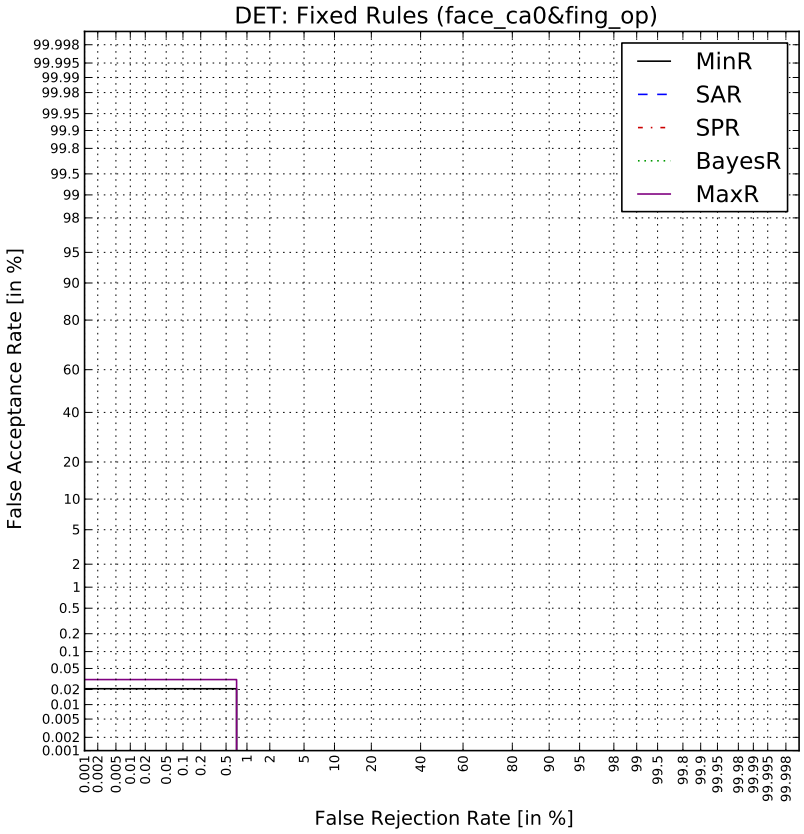

Figure 13: Detection error trade-off (DET) profiles for best fixed fusion rules and *ca*0 2D-face and *op* fingerprint subsets. Only two profiles are visible due to the very low degree of overlap between target and impostor score distributions in other cases which leads to 0% EER.

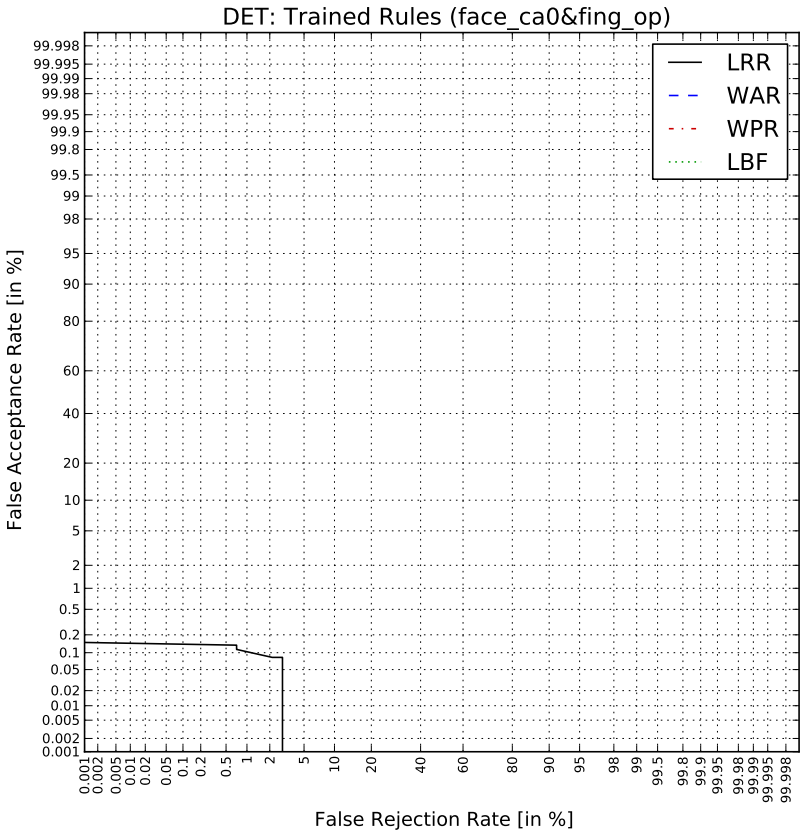

Figure 14: Detection error trade-off (DET) profiles for best trained fusion rules and *ca*0 2D-face and *op* fingerprint subsets. Only one profile is visible due to the very low degree of overlap between target and impostor score distributions in other cases which leads to 0% EER.

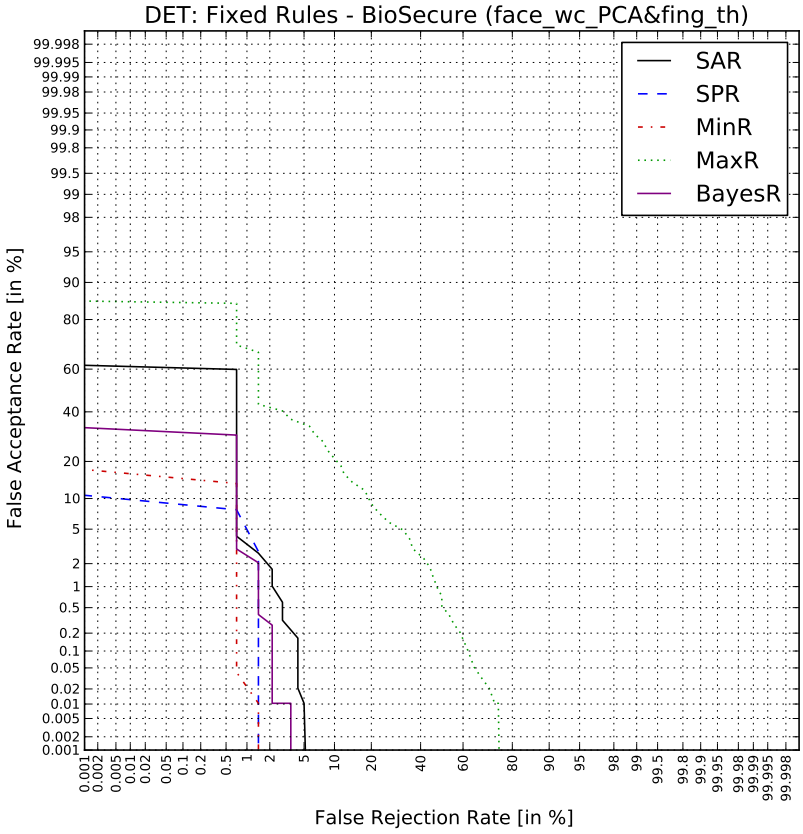

Figure 15: Detection error trade-off (DET) profiles for worst fixed fusion rules and *wc_PCA* face and *th* fingerprint subsets.

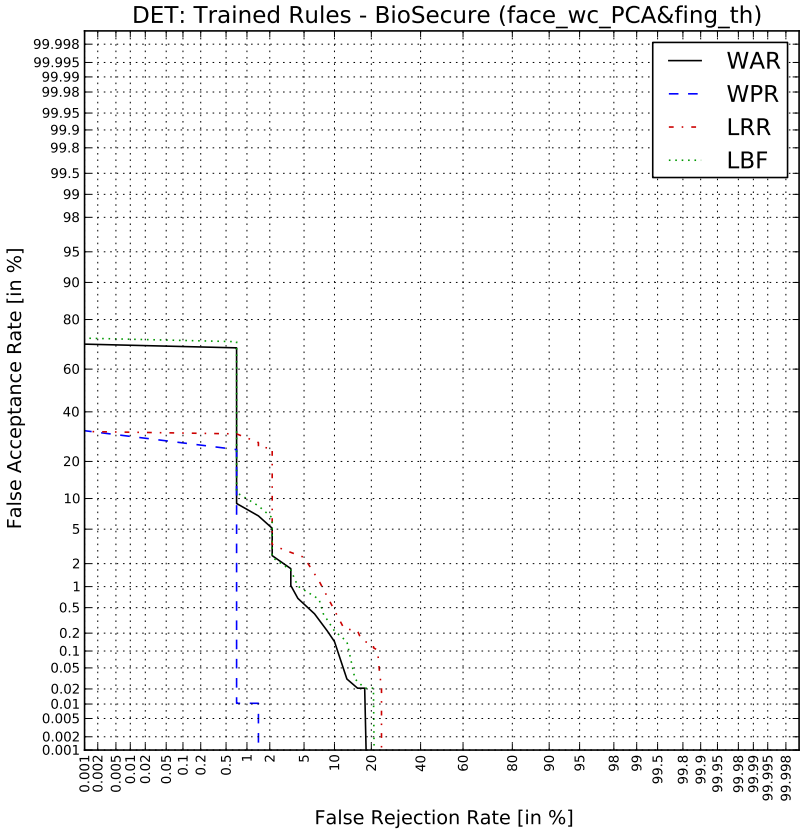

Figure 16: Detection error trade-off (DET) profiles for worst trained fusion rules and *wc_PCA* face and *th* fingerprint subsets.

EER of trained fusion rules				
face & finger	WAR	WPR	LRR	LBF
ca0 & op	0.0%	0.0%	0.1%	0.0%
ca0_PCA & op	0.7%	0.7%	2.8%	0.7%
ca0 & th	0.0%	0.0%	0.7%	0.0%
ca0_PCA & th	1.4%	0.1%	2.8%	1.4%
caf & op	0.0%	0.0%	0.1%	0.0%
caf_PCA & op	0.7%	0.0%	1.4%	0.7%
caf & th	0.0%	0.0%	0.7%	0.0%
caf_PCA & th	0.7%	0.7%	2.7%	0.7%
wc & op	0.0%	0.0%	0.0%	0.0%
wc_PCA & op	0.7%	0.0%	1.9%	0.7%
wc & th	0.0%	0.7%	0.7%	0.0%
wc_PCA & th	2.3%	0.7%	2.9%	2.3%

Table 11: Baseline equal error rates (EERs) (%) for trained fusion of 2D-face and fingerprint modalities.

8 Multi-Modal Biometrics: 2D-Face and 3D-Face

In this section we explore the score-level fusion of 2D and 3D-face recognition systems. A multi-modal system comprising different approaches to face recognition is especially attractive because the degree of cooperation required from the user is lower than that required by a multi-modal system using different modalities, for example iris and fingerprint recognizers. However, systems using the same modality (the face) may yield correlated scores, thus limiting to some degree the potential benefits of multi-modal fusion.

8.1 The Systems

We again applied fixed and trained fusion rules as described in Section 6.1.3 except where s_{voice} and w_{voice} are replaced by scores and weights corresponding to the 3D-face modality. The 2D-face recognition system is again the same as that reported in Section 6.1.1 whereas the 3D-face recognition system is summarised below. The fusion system is the UNICA fusion system described in Section 6.1.3.

8.1.1 3D-Face System

The 3D-face recognition system for the baseline evaluation was developed in the Multimedia Image Group, EURECOM. This system uses a sparser representation of dense 3D facial scans and hence makes the comparison between faces easier for recognition. First a generic face is warped using the Thin Plate Spline (TPS) method for each 3D scan to remove the 'common' face shape information. For each face, 15 fiducial points are considered. Next, the generic face is aligned and scaled on to each face based on the set of points. Then it is coarsely warped to make the two surfaces as close as possible. Assuming that the two surfaces are in sufficient alignment and the correspondences are found as the closest vertices, 136 more point-pairs are obtained. Finally, the generic face is warped based on a total of 151 point-pairs. Thus, each 3D-face model can be represented with the 3D vector of size 151×3 which is obtained from the warping parameters in x, y and z directions for each control point.

In order to measure the similarity between facial surfaces, the angle between the two warping vectors and the difference between their magnitudes and angles are calculated. This results in two distance vectors of size 151×1 for any pair of faces. Central tendencies of these vectors are fused and a decision is made based on the nearest neighbour approach.

8.2 The FRGC Database

The specification of the FRGC database can be found in [26]. A subject session consists of four controlled still images, two uncontrolled still images, and one three-dimensional image. The controlled images were acquired in a studio setting and are full frontal facial images taken under two lighting conditions. Facial expressions are neutral or smiling. On the other hand, the uncontrolled images were taken in varying illumination conditions. Each set of

Dataset	Total patterns	Genuine Users	Impostors	EER
Training 2D	6,420,817	18,770 0.29%	6,402,047 99.71%	8.82%
Training 3D	6,420,817	18,770 0.29%	6,402,047 99.71%	15.97%
Testing 2D	9,631,225	28,142 0.29%	9,603,083 99.71%	8.78%
Testing 3D	9,631,225	28,142 0.29%	9,603,083 99.71%	15.82%

Table 12: FRGC datasets: samples, classes and baseline equal error rates (EERs).

uncontrolled images contains smiling or neutral expressions. A Vivid 900/910 sensor is used to capture 3D images. It is a structured light sensor that captures a 640×480 range sampled and registered colour images. All 3D images were taken under controlled illumination conditions appropriate to the sensor. Subjects stood or were seated approximately 1.5 meters from the sensor.

The database consists of training and validation sets. The training set also consists of two parts which are a large still training set and a 3D training set. There are 222 subjects in the large still training set and 466 subjects in the validation set. The database includes $12,776$ images/videos in the large still training set, with 6,388 controlled still images and 6,388 uncontrolled still images. It includes 943×8 images/videos in the 3D training set that contains 3D scans, and controlled and uncontrolled still images. The 3D training set is for training 3D and 3D-to-2D algorithms. The validation set contains images from 466 subjects collected in 4007 subject sessions. In the validation set, there are 4007×8 images/videos. Finally, the database contains static and colourful subjects and consists of single faces. Images are in JPEG format and the resolution is 1704×2272 or 1200×1600. The FRGC database can be obtained by contacting the FRGC Liaison at frgc@nist.gov. More information on how to access this database can be obtained from the FRGC website [1].

A summary of the database structure is illustrated in Table 12 which shows the number of samples, the samples per class and the baseline EER for each dataset and for both 2D and 3D modes.

8.3 Performance Evaluation

We investigated the same score-level fusion rules outlined in Section 6, where the outputs of individual recognition systems are combined through a 'fusion rule'.

8.3.1 Setup

The experimental setup is exactly as described in Section 6.3.1 for UNICA fusion except that s_{voice} and w_{voice} are replaced by scores and weights corresponding to the 3D-face modality.

8.3.2 Results

Results are illustrated in Table 13 and in the DET plots in Figures 17 and 18. Improvements are not as significant as in other modalities, such as the fusion of face and voice recognition systems reported in Section 6. In this case the high correlation between the two sets of scores impacts on the potential of fusion. Among explored rules, only LRR obtains a small performance improvement. For trained rules the DET curves for WAR, WPR and LBF often overlap.

Note that these observations do not necessarily detract from the potential of fused 2D and 3D-face recognition systems as a countermeasure against spoofing, since different attacks may not be as successful in fooling one system as another. In other words, a multi-modal biometric systems employing 2D and 3D-face images, may be viewed as an 'intrinsically robust' biometric system against spoofing attacks. These aspects will be addressed later in the project and in future deliverables.

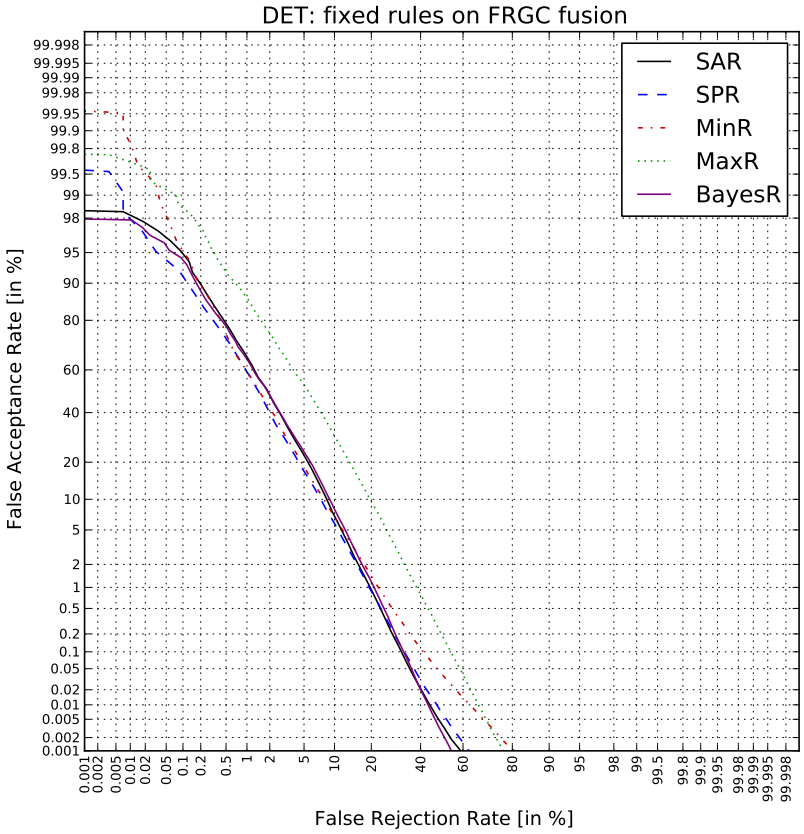

Figure 17: Detection error trade-off (DET) profiles for 2D and 3D-face multi-modal biometric performance using fixed fusion rules.

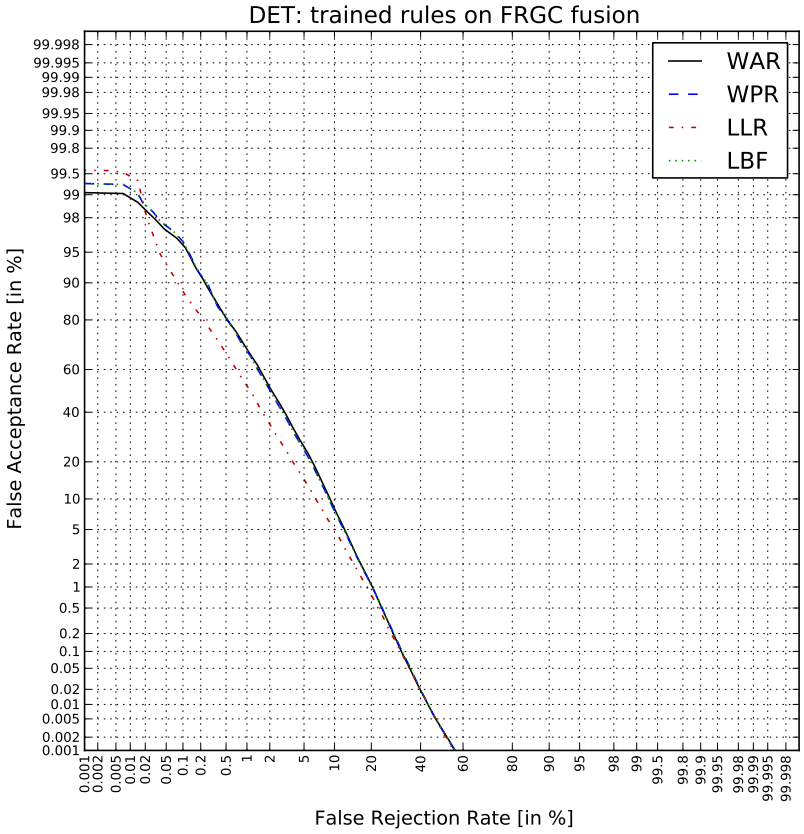

Figure 18: Detection error trade-off (DET) profiles for 2D and 3D-face multi-modal biometric performance using trained fusion rules.

Dataset	EER
SAR	9.0%
SPR	8.4%
MinR	8.8%
MaxR	15.8%
BayesR	9.4%
WAR	9.5%
WPR	9.3%
LRR	7.9%
LBF	9.3%

Table 13: EER for different fusion rules applied to the 2D and 3D-face multi-modal biometric.

9 Multi-Modal Biometrics: ECG and EEG

We evaluate in this section the performance of the fusion of electro-physiological modalities, namely EEG and ECG. A similar evaluation was conducted in [34] but on a limited database, which took into account data from only 4 subjects. We have significantly increased the number of subjects as described in Section 5 and recalled herein.

Electro-physiological authentication is an emergent biometric modality, taking into account at this moment EEG and ECG signals. Their main feature is that these signals can only be generated by a living subject and are not based on external physical traits as most of current biometric technologies. The purpose herein is to generate a baseline for the posterior evaluation of the spoofing attack effect on the fusion of both modalities. The main challenge of the EEG-ECG fusion is to tackle very different performance levels of the mono-modal authentication, where ECG presents significantly outperforms EEG (see Table 5.3.2).

9.1 The MITSfusion system

A description of the MITSfusion system can be found in Section 6.1.2.

9.2 The Databases

Starlab has gathered 2 different databases: Eyes-closed DB and Task-Performing DB. In both cases EEG and ECG data were collected using the ENOBIO sensor, which consists of 4 active electrodes: 2 in the forehead (Fp1 and Fp2 locations) for EEG, 1 in the left wrist for ECG and finally the last one in the right earlobe as a reference. The first dataset (Eyes-closed), was recorded in a controlled environment where the subjects were asked to sit, relax and close their eyes. In the second dataset (Task-performing), signals were gathered while subjects kept their eyes open, sitting on a chair, and were free to perform a number of office activities (such as answering the phone, keystroke, drinking water, using the mouse...). The reader is referred to Section 5 for more details. The two databases were used for the evaluation of ECG and EEG as a multi-modal biometric.

9.3 Performance Evaluation

The evaluation was conducted using the two databases in a similar fashion to the approach described in Section 6.3.1. We used the MITSfusion system with just one node, where its features are determined in the training phase. Hence we have to assess:

- Employed fusion operator $\mathbf{FO}_0$.
- Usage of personalized operators vs. a general one.
- Optimal parameter set $\mathbf{p}_0$ (or superset in case of personalized ones).

9.3.1 Setup

We undertook the following procedure:

1. We apply all available fusion operators as $\mathbf{FO}_0$.
2. For each fusion operator we look for the parameter set $\mathbf{p}_0$ and superset $\{\mathbf{p}_{01}, \ldots, \mathbf{p}_{0L}\}$ that maximize the AUC.
3. We compute the average AUC over subjects when:
   (a) Applying a unique parameter set $\mathbf{p}_0$, which we denote as $\overline{AUC}_g$.
   (b) Applying a parameter superset $\{\mathbf{p}_{01}, \ldots, \mathbf{p}_{0L}\}$, which we denote as $\overline{AUC}_p$.
4. We select the fusion operator with maximal $\overline{AUC}_p$.
5. If the difference between $\overline{AUC}_p$ and $\overline{AUC}_g$ is less than 3%, we personalize the operator, i.e. $\mathbf{FO}_0$ is parameterized with $\mathbf{p}_0$. If not, we parameterized it with one parameter for each subject, i.e. $\{\mathbf{p}_{01}, \ldots, \mathbf{p}_{0L}\}$.

In this case we used the Eyes-Closed DB as a development dataset and the Task-Performing DB as a test/evaluation dataset. The development dataset was used for optimising the features used in the MITSfusion system (see procedure in Section 6.3.1). This selection is motivated by its applicability in a real use case scenario, where the subjects enroll in eyes-closed condition, but are authenticated later in a more uncontrolled protocol.

It is worth pointing out that the evaluation setup used for the multi-modal evaluation is by far more pessimistic than the one used in the mono-modal EEG and ECG evaluation. This is because we used a setup closer to use-case conditions for the multi-modal evaluation. Hence two data sets acquired in two very different conditions, namely eyes closed and task performing, are used respectively for development and test/evaluation in the multi-modal setup, which becomes therefore more risky. In case a direct comparison between mono-modal results and multi-modal fusion one is needed, the reader can hypothesize with a high degree of confidence that multi-modal performance would improve with respect to the level given as outlined below if using exactly the same setup. However it is not the objective of this project to compare mono-modal/multi-modal performance, but the computation of a baseline performance for comparison with the results under spoofing attacks and with countermeasures.

9.3.2 Results

As a result of the application of the assessment procedure on the development data set, i.e. Eyes-Closed DB, we selected the uni-norm fusion operator. The operator is not personalized and the optimal parameter set is $(e, p) = (0.9, 3)$, where e defines the application

Data Set	modality	EER (%)
Eyes-Closed	EEG	20
Eyes-Closed	ECG	4
Eyes-Closed	*fusion dev*	*4*
Task-Performing	EEG	27
Task-Performing	ECG	14
Task-Performing	*fusion test*	*0.6*

Table 14: Equal Error Rate (EER) of the multi-modal fusion of EEG and ECG modalities in the development (*Eyes-Closed - fusion dev*) and test (*Task-Performing - fusion test*) data sets. Performance of the fused modalities is given for the sake of comparison.

scope of each norm, and p is the exponent in the Yager norms expression as defined in equation (6):

$$z = \begin{cases} \max\{0, 1 - ((1-x_1)^p + (1-x_2)^p)^{1/p}\} & : \quad x_1, x_2 \in [0, e] \times [0, e,] \\ \min\{1, (x_1^p + x_2^p)^{1/p}\} & : \quad x_1, x_2 \in [e, 1] \times [e, 1] \\ \frac{x_1+x_2}{2} & : \quad otherwise \end{cases}$$

A summary of the results is given in Table 14 which shows the EERs for the two datasets. Figure 19 shows the corresponding DET curve for the test data set. Results are seen to improve compared to performance for the two individual modalities. In case of the development data set, and due to the much better performance of the ECG modality with respect to the EEG (see Table 14), the fusion adapts to emphasise the ECG modality. In the test data set, where EEG and ECG present a more similar performance in the mono-modal authentication (see Table 14), the fusion result clearly outperforms both modalities in spite of using a riskier setup as explained in Section 9.3.1.

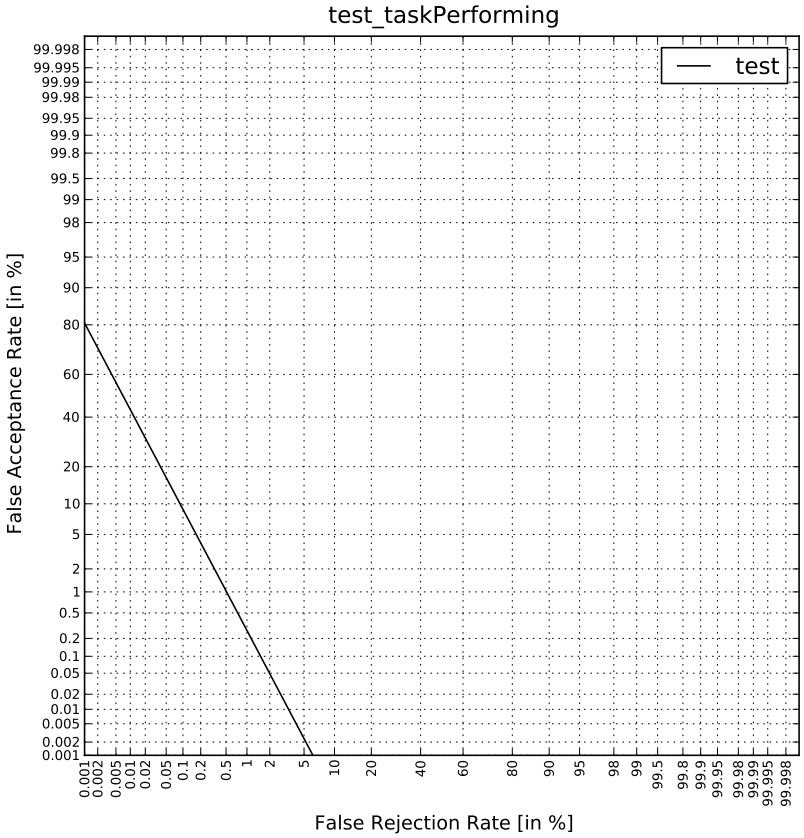

Figure 19: DET curve on the test data set, i.e. Task-performing DB.

10 Summary

Presented in this document are equal error rates (EERs) and detection error trade-off (DET) profiles for all non-ICAO mono-modal and multi-modal biometrics assessed in the TABULA RASA project. Together they form the baselines for all future work in spoofing and countermeasures. We re-iterate that traditional biometrics research is not the goal in this project and thus we do not necessarily seek recognition performance which is superior to the baseline. We seek to show how (i) the baseline performance deteriorates in the face of spoofing and, more importantly, (ii) how countermeasures may be effectively harnessed to reduce the gap between performance under spoofing and the baseline performance presented here.

The different baseline systems show a variation in performance between 0%[2] and 27% EER. While it is not the objective to compare the performance of each modality a summary of the EERs for each biometric is listed in Table 15.

Modality	Database	System	Subset	EER (%)
Voice	NIST'06	ALIZE	male	5.9
			female	5.4
	MOBIO	ALIZE	male	15.2
			female	18.4
Gait	USOU	USOU	-	6.0
		UOULU	-	4.5
Vein and Fingerprint	TabulaRasaVP	FingerVP	-	0.0
Electro-physiology	Eyes-closed	StarFast	EEG	20.0
			ECG	4.0
	Task-performing	StarFast	EEG	27.0
			ECG	14.0
2D-Face and Voice	MOBIO	MITSfusion	male	6.0
			female	10.2
		UNICA fusion	male	7.3
			female	11.8
2D-Face and Fingerprint	BioSecure	UNICA fusion	ca0/op	0.0
2D-Face and 3D-Face	FRGC	UNICA fusion	-	0.9
ECG and EEG	Task-performing	MITSfusion	-	0.6

Table 15: A summary of EERs for all non-ICAO mono-modal and multi-modal biometrics addressed in TABULA RASA.

[2]Note the discussion of statistical significance in Section 4.3.2

References

[1] NIST. face recognition grand challenge(FRGC). page http://face.nist.gov/frgc/.

[2] R. Auckenthaler and J. S. Mason. Gaussian selection applied to text-independent speaker verification. In *Proc. Speaker Odyssey 2001*, pages 83–88, 2001.

[3] J.-F. Bonastre, F. Wils, and S. Meignier. ALIZE, a free toolkit for speaker recognition. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, volume 1, pages 737 – 740, 18-23, 2005.

[4] P. Bullen. *Handbook of Means and their Inequalities*. Kluwer, 2003.

[5] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff. Svm based speaker verification using a gmm supervector kernel and nap variability compensation. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, page I, may 2006.

[6] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-carrasquillo. Support vector machines for speaker and language recognition. *Computer Speech and Language*, 20:210–229, 2006.

[7] W. M. Campbell, D. E. Sturim, and D. A. Reynolds. Support vector machines using gmm supervectors for speaker verification. *IEEE Signal Processing Letters*, 13:308–311, 2006.

[8] M. J. Carey, E. S. Parris, and J. S. Bridle. A speaker verification system using alphanets. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 397–400, April 1991.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[10] R. Duda, P. Hart, , and D. Stork. *Pattern Classification*. John Wiley and Sons, New York, NY, USA, 2001.

[11] B. G. B. Fauve, D. Matrouf, N. Scheffer, J.-F. Bonastre, and J. S. D. Mason. State-of-the-art performance in text-independent speaker verification through open-source software. *IEEE Transactions on Audio Speech and Language processing*, 15(7):1960–1968, 2007.

[12] M. Garris, C. Watson, R. McCabe, and C. Wilson. User's guide to NIST fingerprint image software 2 (nfis2). Technical report, NIST, 2004.

[13] J.-L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 2:291–298, 1994.

[14] M. Grabisch, H. T.Nguyen, and A. A. Walker. *Fundamentals of Uncertainty Calculi with Applications to Fuzzy Inference*. Kluwer Academic Pub., Dordrecht, 1995.

[15] S. Israel, J. Irvine, A. Cheng, M. Wiederhold, and B. Wiederhold. Ecg to identify individuals. *Pattern Recognition*, 38:133–142, 2005.

[16] H. Jarosz. Biomtrie et multi-modalit : un saut technologique vient-il d'tre franchi. ExpoProtection2010.

[17] V. Kellokumpu, G. Zhao, S. Z. Li, and M. Pietikäinen. Dynamic texture based gait recognition. In *IAPR/IEEE International Conference on Biometrics 2009*, pages 1000–1009, 2009.

[18] V. Kellokumpu, G. Zhao, and M. Pietikäinen. Dynamic textures for human movement recognition. In *ACM International Conference on Image and Video Retrieval*, CIVR '10, pages 470–476, 2010.

[19] P. Kenny. Joint factor analysis of speaker and session variability: theory and algorithms. Technical Report 06/08-13, CRIM, 2006.

[20] E. P. Klement, R. Mesiar, and E. Pap. *Triangular Norms (Trends in Logic, Vol. 8)*. Springer, 1 edition, July 2000.

[21] S. Marcel, C. McCool, P. Matejka, T. Ahonen, and J. Cernocky. Mobile biometry face and speaker verification evaluation. Technical Report Idiap-RR-09-2010, IDIAP, 2010.

[22] D. Matovski, M. Nixon, S. Mahmoodi, and J. Carter. The effect of time on gait recognition performance. *IEEE Transactions on Information Forensics and Security*, 2011.

[23] D. Matrouf, N. Scheffer, B. Fauve, and J.-F. Bonastre. A straightforward and efficient implementation of the factor analysis model for speaker verification. In *Proc. Interspeech*, 2007.

[24] M. Nixon and J. Carter. Automatic Recognition by Gait. *Proceedings of the IEEE*, 94(11):2013–2024, 2006.

[25] M. S. Nixon, T. N. Tan, and R. Chellappa. *Human Identification based on Gait*. International Series on Biometrics. Springer, 2005.

[26] P. J. Phillips, P. J. Flynn, and T. Scruggs. Overview of the face recognition grand challenge. *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR05)*, 1:947–954, 2005.

[27] M. Poulos, M. Rangoussi, N. Alexandris, and A. Evangelou. Person identification from the eeg using nonlinear signal classification. *Methods of Information in Medicine*, 41:64–75, 2002.

[28] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1-3):19 – 41, 2000.

[29] A. Riera, A. Soria-Frisch, M. Caparrini, C. Grau, and G. Ruffini. Unobtrusive biometric system based on electroencephalogram analysis. *EURASIP Journal on Advances in Signal Processing*, 2008, 2008.

[30] G. Ruffini, S. Dunne, and E. F. et al. Enobio dry electrophysiology electrode; first human trial plus wireless electrode system. *Proceedings of the 29th IEEE EMBS Annual International Conference*, 2007.

[31] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:162–177, 2005.

[32] R. D. Seely, S. Samangooei, L. Middleton, J. Carter, and M. Nixon. The university of southampton multi-biometric tunnel and introducing a novel 3d gait dataset. In *Biometrics: Theory, Applications and Systems*. IEEE, September 2008.

[33] J. Shutler, M. Grant, M. Nixon, and J. Carter. On a large sequence-based human gait database. In *Conf. Recent Advances in Soft Computing*, pages 66–72, 2002.

[34] A. Soria-Frisch, A. Riera, and S. Dunne. Fusion operators for multi-modal biometric authentication based on physiological signals. In *Proc. 2010 IEEE International Conference on Fuzzy Systems (FUZZ'IEEE)*, pages 1 –7, july 2010.

[35] M. Sugeno. *The Theory of Fuzzy Integrals and Its Applications*. PhD thesis, Tokyo Institute of Technology, Japan, 1974.

[36] J. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[37] G. Veres, L. Gordon, J. Carter, and M. Nixon. What image information is important in silhouette-based gait recognition. In *IEEE Computer Vision and Pattern Recognition conference*, 2004.

[38] R. R. Yager. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Trans. Syst. Man Cybern.*, 18(1):183–190, 1988.

[39] R. R. Yager and A. Rybalov. Uninorm aggregation operators. *Fuzzy Sets Syst.*, 80(1):111–120, 1996.

[40] S. Yu, T. Tan, K. Huang, K. Jia, and X. Wu. A study on gait-based gender classification. *IEEE Transactions on Image Processing*, 18:1905–1910, August 2009.