Marcel, S., McCool, C., Matjka, P., Ahonen, T., ernocký, J., Chakraborty, S., ... (2010). On the Results of the First Mobile Biometry (MOBIO) Face and Speaker Verification Evaluation. En Recognizing Patterns in Signals, Speech, Images and Videos: ICPR 2010 Contests, Istanbul, Turkey, August 23-26, 2010, Contest Reports. Springer Verlag (Germany). 210-225. doi:10.1007/978-3-642-17711-8_22.

# On the Results of the First Mobile Biometry (MOBIO) Face and Speaker Verification Evaluation

Sébastien Marcel[1], Chris McCool[1], Pavel Matějka[3], Timo Ahonen[2], Jan Černocký[3], Shayok Chakraborty[4], Vineeth Balasubramanian[4], Sethuraman Panchanathan[4], Chi Ho Chan[5], Josef Kittler[5], Norman Poh[5], Benoît Fauve[6], Ondřej Glembek[3], Oldřich Plchot[3], Zdeněk Jančík[3], Anthony Larcher[7], Christophe Lévy[7], Driss Matrouf[7], Jean-François Bonastre[7], Ping-Han Lee[8], Jui-Yu Hung[8], Si-Wei Wu[8], Yi-Ping Hung[8], Lukáš Machlica[9], John Mason[10], Sandra Mau[11], Conrad Sanderson[11], David Monzo[12], Alberto Albiol[12], Antonio Albiol[12], Hieu Nguyen[13], Bai Li[13], Yan Wang[13], Matti Niskanen[14], Markus Turtinen[14], Juan Arturo Nolazco-Flores[15], Leibny Paola Garcia-Perera[15], Roberto Aceves-Lopez[15], Mauricio Villegas[16], Roberto Paredes[16]

[1]Idiap Research Institute, CH, [2]University of Oulu, FI, [3]Brno University of Technology, CZ, [4]Center for Cognitive Ubiquitous Computing, Arizona State University, USA, [5]Centre for Vision, Speech and Signal Processing, University of Surrey, UK, [6]Validsoft Ltd., UK, [7]University of Avignon, LIA, FR, [8]National Taiwan University, TW, [9]University of West Bohemia, CZ, [10]Swansea University, UK, [11]NICTA, AU, [12]iTEAM, Universidad Politecnica de Valencia, ES, [13]University of Nottingham, UK, [14]Visidon Ltd, FI, [15]Tecnologico de Monterrey, MX, [16]Instituto Tecnológico de Informática, Universidad Politécnica de Valencia, ES

**Abstract.** This paper evaluates the performance of face and speaker verification techniques in the context of a mobile environment. The mobile environment was chosen as it provides a realistic and challenging test-bed for biometric person verification techniques to operate. For instance the audio environment is quite noisy and there is limited control over the illumination conditions and the pose of the subject for the video. To conduct this evaluation, a part of a database captured during the "Mobile Biometry" (MOBIO) European Project was used. In total there were nine participants to the evaluation who submitted a face verification system and five participants who submitted speaker verification systems.

**Keywords:** mobile, biometric, face recognition, speaker recognition, evaluation

## 1 Introduction

Face and speaker recognition are both mature fields of research. Face recognition has been explored since the mid 1960's [5]. Speaker recognition by humans has been done since the invention by the first recording devices, but automatic speaker recognition is a topic extensively investigated only since 1970 [6]. However, these two fields have often been considered in isolation to one another as very few joint databases exist.

For speaker recognition there is a regular evaluation organised by the National Institute of Standards and Technology (NIST) [1] called the NIST Speaker Recognition

[1] *h*ttp://www.nist.gov

Evaluation. NIST has been coordinating SRE since 1996 and since then over 50 research sites have participated in the evaluations. The goal of this evaluation series is to contribute to the direction of research efforts and the calibration of technical capabilities of text independent speaker recognition. The overarching objective of the evaluations has always been to drive the technology forward, to measure the state-of-the-art, and to find the most promising algorithmic approaches.

Although there is no regular face recognition competition, there have been several competitions and evaluations for face recognition. These include those led by academic institutions, such as the 2004 ICPR Face Verification Competition [25], in addition to other major evaluations such as the Face Recognition Grand Challenge [27] organised by NIST.

The MOBIO Face and Speaker Verification Evaluation provides the unique opportunity to analyse two mature biometrics side by side in a mobile environment. The mobile environment offers challenging recording conditions including adverse illumination, noisy background and noisy audio data. This evaluation is the first planned of a series of evaluations and so only examines uni-modal face and speaker verification techniques.

In the next section, we briefly present the state-of-the-art in face and speaker verification. Then, we introduce in section 3 the MOBIO database and its evaluation protocol. In sections 4 and 5, we shortly describe the individual face and speaker verification systems involved in this evaluation. The reader can be referred to [24] for a more detailed description of these systems. Finally in section 6, we present the results obtained and discuss them.

## 2 Face and Speaker Verification

### 2.1 Face Verification

The face is a very natural biometric as it is one that humans use everyday in passports, drivers licences and other identity cards. It is also relatively easy to capture the 2D face image as no special sensors, apart from a camera that already exist on many mobile devices, are needed.

Despite the ease with which humans perform face recognition the task of automatic face recognition (for a computer) remains very challenging. Some of the key challenges include coping with changes in the facial appearance due to facial expression, pose, lighting and aging of the subjects.

There have been surveys of both face recognition [40] [34] and video based analysis [35]. From all of these it can be seen that there are many different ways to address the problem of face recognition in general, and more particularly of face verification in theis paper. Some of the solutions can include (but are not limited to) steps such as image preprocessing, face detection, facial feature point detection, face preprocessing for illumination and 2D or 3D geometric normalisation, quality assessment feature extraction, score computation based on client-specific and world models, score normalisation and finally decision making. However, the actual steps taken vary drastically from one system to another.

### 2.2 Speaker Verification

The most prevalent technique for speaker verification is the Gaussian Mixture Model (GMM) paradigm that uses a Universal Background Model (UBM). In this paradigm a UBM is trained on a set of independent speakers. Then a client is enroled by adapting from this UBM using the speaker specific data. When testing two likelihoods are produced, one for the UBM and one for the client specific model, and these two scores are combined using the log-likelihood ratio and compared to a threshold to produce a "client/imposter" decision [29].

Many other techniques for speaker verification have been proposed. These techniques range from Support Vector Machines [9], Joint Factor Analysis [20] and other group based on Large Vocabulary Continuous Speech Recognition systems [33] through to prosodic and other high level based features for speaker verification [32]. One common thread with the speaker verification techniques proposed nowadays is the ability to cope with inter-session variability which can come from the: communication channel, acoustic environment, state of the speaker (mood/health/stress), and language.

## 3 MOBIO Database and Evaluation Protocol

### 3.1 The MOBIO Database

The MOBIO database was captured to address several issues in the field of face and speaker recognition. These issues include: (1) having consistent data over a period of time to study the problem of model adaptation, (2) having video captured in realistic settings with people answering questions or talking with variable illumination and poses, (3) having audio captured on a mobile platform with varying degrees of noise.

The MOBIO database consists of two phases, only one of which was used for this competition. The first phase (Phase I) of the MOBIO database was captured at six separate sites in five different countries. These sites are at the: University of Manchester (UMAN), University of Surrey (UNIS), Idiap Research Institute (IDIAP), Brno University of Technology (BUT), University of Avignon (LIA) and University of Oulu (UOULU). It includes both native and non-native English speakers (speaking only English).

The database was acquired primarily on a mobile phone. The Phase I of the database contains 160 participants who completed six sessions. In each session the participants were asked to answer a set of questions which were classified as: i) set responses, ii) read speech from a paper, and iii) free speech. Each session consisted of 21 questions: 5 set response questions, 1 read speech question and 15 free speech questions. In total there were five **Set responses** to questions and **fake responses** were supplied to each user. **Read speech** was obtained from each user by supplying the user with three sentences to read. **Free speech** was obtained from each user by prompting the user with a random question. For five of these questions the user was asked to speak for five seconds (short free speech) and for ten questions the user was asked to speak for ten seconds (long free speech), this gives a total of fifteen such questions.

### 3.2 The MOBIO Evaluation Protocol

The database is split into three distinct sets: one for training, one for development and one for testing. The data is split so that two sites are used in totality for one set, this means that the three sets are completely separate with no information regarding individuals or the conditions being shared between any of the three sets.

The training data set could be used in any way deemed appropriate and all of the data was available for use. Normally the training set would be used to derive background models, for instance training a world background model or an LDA sub-space. The development data set had to be used to derive a threshold that is then applied to the test data. However, for this competition it was also allowed to derive fusion parameters if the participants chose to do so. To facilitate the use of the development set, the same protocol for enroling and testing clients was used in the development and test splits. The test split was used to derive the final set of scores. No parameters could be derived from this set, with only the enrolment data for each client available for use; no knowledge about the other clients was to be used. To help ensure that this was the case the data was encoded so that the filename gave no clue as to the identity of the user.

The protocol for enroling and testing were the same for the development split and the test split. The first session is used to enrol the user but only the five set response questions can be used for enrolment. Testing is then conducted on each individual file for sessions two to six (there are five sessions used for development/testing) and only the free speech questions are used for testing. This leads to five enrolment videos for each user and 75 test client (positive sample) videos for each user (15 from each session). When producing imposter scores all the other clients are used, for instance if in total there were 50 clients then the other 49 clients would perform an imposter attack.

### 3.3 Performance Evaluation

Person verification (either based on the face, the speech or any other modality) is subject to two type of errors, either the true client is rejected (false rejection) or an imposter is accepted (false acceptance). In order to measure the performance of verification systems, we use the Half Total Error Rate (HTER), which combines the False Rejection Rate (FRR) and the False Acceptance Rate (FAR) and is defined as:

$$HTER(\tau, \mathcal{D}) = \frac{FAR(\tau, \mathcal{D}) + FRR(\tau, \mathcal{D})}{2} \quad [\%] \tag{1}$$

where $\mathcal{D}$ denotes the used dataset. Since both the FAR and the FRR depends on the threshold $\tau$, they are strongly related to each other: increasing the FAR will reduce the FRR and vice-versa. For this reason, verification results are often presented using either Receiver Operating Characteristic (ROC) or Detection-Error Tradeoff (DET) curves, which basically plots the FAR versus the FRR for different values of the threshold. Another widely used measure to summarise the performance of a system is the Equal Error Rate (EER), defined as the point along the ROC or DET curve where the FAR equals the FRR.

However, it was noted in [4] that ROC and DET curves may be misleading when comparing systems. Hence, the so-called Expected Performance Curve (EPC) was proposed, and consists in an unbiased estimate of the reachable performance of a system

at various operating points. Indeed, in real-world scenario, the threshold $\tau$ has to be set a priori: this is typically done using a development set (also called validation set). Nevertheless, the optimal threshold can be different depending on the relative importance given to the FAR and the FRR. Hence, in the EPC framework, $\beta \in [0; 1]$ is defined as the tradeoff between FAR and FRR. The optimal threshold $\tau^*$ is then computed using different values of $\beta$, corresponding to different operating points:

$$\tau^* = \underset{\tau}{\operatorname{argmin}} \quad \beta \cdot \text{FAR}(\tau, \mathcal{D}_d) + (1 - \beta) \cdot \text{FRR}(\tau, \mathcal{D}_d) \tag{2}$$

where $\mathcal{D}_d$ denotes the development set.

Performance for different values of $\beta$ is then computed on the test set $\mathcal{D}_t$ using the previously found threshold. Note that setting $\beta$ to 0.5 yields to the Half Total Error Rate (HTER) as defined in Equation (1).

## 4 Face Verification Systems

### 4.1 Idiap research institute (IDIAP)

The Idiap Research Institute submitted two face (video) recognition systems. The two used exactly the same verification method using a mixture of Gaussians to model a parts-based topology, as described in [10], and so differed only in the way in which the faces were found in the video sequence (the face detection method). The systems submitted by the Idiap Research Institute served as baseline systems for the face (video) portion of the competition.

**System 1** is referred to as a frontal face detector as it uses only a frontal face detector. **System 2** is referred to as a multi-view face detector as it uses a set of face detectors for different poses. Both frontal and multi-view face detection systems are taken from [30].

### 4.2 Instituto Tecnológico de Informática (ITI)

Two face recognition systems were submitted by the Instituto Tecnológico de Informática. Both systems, first detect faces every 0.1 seconds up to a maximum of 2.4 seconds of video. For enrolment or verification, only a few of the detected faces are selected based on a quality measure. The face verification approach was based on [37]. Each face is cropped to 64×64 pixels and 9×9 pixel patches are extracted at overlapping positions every 2 pixels, 784 features in total. The verification score is obtained using a Nearest-Neighbor classifier and a voting scheme. For further details refer to [37].

**System 1** used the *haarcascade_frontalface_alt2* detection model that is included with the OpenCV library, and as quality measure used the confidence of a face-not-face classifier learnt using [36]. For verification, 10 face images are used. **System 2** used the face detector from the commercial OmniPerception's SDK and as quality the average of the confidences of the detector and the face-not-face classifier. For verification, 5 face images are used.

### 4.3 NICTA

NICTA submitted two video face recognition systems. Both systems used OpenCV for face detection in conjunction with a modified version of the Multi-Region Hisotgram (MRH) face comparison method [31]. To extend MRH from still-to-still to video-to-video comparison, a single MRH signature was generated for each video sequence by averaging the histograms for each region over the available frames. Two signatures are then compared through an $L_1$-norm based distance. If a person has several video sequences for enrolment, multiple signatures are associated with their gallery profile, and the minimum distance of those to the probe video signature is taken as the final result. For normalisation, each raw measurement is divided by the average similarity of each probe-gallery pair to a set of cohort signatures from the training set [31].

**System 1** used only closely cropped faces (of size $64 \times 64$ pixels) which excluded image areas susceptible to disguises, such as hair and chin. **System 2** used information from those surrounding regions as well, resulting $96 \times 96$ pixel sized faces. The results show that the use of the surrounding regions considerably improved the recognition performance for the female set.

### 4.4 Tecnologico de Monterrey, Mexico and Arizona State University, USA (TEC-ASU)

The CUbiC-FVS (CUbiC-Face Verification System) was based on distance computations using a nearest neighbor classifier (similar to Das [14]). Each video stream was sliced into images and a face detection algorithm based on the mean-shift algorithm (similar to [13]) was used to localize a face in a given frame. The block based discrete cosine transform (DCT) was used to derive facial features (similar to Ekenel *et al*. [16]), since this feature is known to be robust to illumination changes.

For each user $U_i$, all the respective feature vectors were assembled into a training matrix $M_i$. A distance measure, $D_{true}$, was computed as the minimum distance of $T$ (the test data) from the feature vectors of matrix $M_k$ of the claimed identity $k$. Similarly, $D_{imp}$ was computed as the minimum distance of $T$ from the feature vectors of all matrices other than $M_k$. The ratio of $D_{true}$ to $D_{imp}$ was used to decide whether the claim has to be accepted or not. The scores were scaled so that clients have a positive score and imposters have a negative score.

### 4.5 University of Surrey (UNIS)

In total, UNIS submitted 4 systems which can be divided into two categories: a *fusion* systems (FS) as well as a *single descriptor* systems (SDS). FS is composed of two subsystems which differ mainly in the feature representation, one based on Multiscale Local Binary Pattern Histogram (MLBPH) [12] and the other based on Multiscale Local Phase Quantisation Histogram (MLPQH) [11]. SDS above refers to MLBPH. In each category, we have *basic* and *updated* versions. Hence, the 4 systems are: **System 1** (Basic+SDS), **System 2** (Updated+SDS), **System 3** (Basic+FS), **System 4** (Updated+FS). The *basic* and *updated* systems differ in terms of image selection strategies and data

sets for the LDA matrix training. Regarding the image selection strategy, a basic system chooses a single face image, while an updated system selects 15 images from the video sequence. For training the LDA matrix, the training set of the MOBIO database is used in the basic system, while the updated system uses an external database. In each version, we measure the difference between the results of those 4 systems (without score normalisation) and the results of these systems with test-normalisation, using the training set of the MOBIO database.

### 4.6 Visidon Ltd (VISIDON)

Visidon face identification and verification system is originally designed for embedded usage, in order to quickly recognize persons in still images using a mobile phone, for example [1]. Thanks to a real-time frame performance, additional information provided by video can be easily utilized to improve the accuracy.

Both object detector (used for face and facial feature detection) and person recognition modules are based on our patented technology.

### 4.7 University of Nottingham (UON)

We implemented two methods: video-based (**System 1**) and image-based (**System 2**). System 1 makes use of all frames in a video and bases on the idea of Locally Linear Embedding [18]. System 2 uses only a couple of frames in a video and bases on 4 facial descriptors: Raw Image Intensity, Local Binary Patterns [2], Gabor Filters, Local Gabor Binary Patterns [39, 19]; 2 subspace learning methods: Whitened PCA, One-shot LDA [38]; and Radial Basis Function SVM for verification.

In our experiments, system 2 performs much better than system 1. However, system 2 didn't perform as well as it could be because we made a mistake in the training process which makes the final SVM over-fitted. Another observation is that face detection is very important to get high accuracy.

### 4.8 National Taiwan University (NTU)

In each frame, we detected and aligned faces according to their eye and mouth positions. We also corrected the in-plane and out-of-plane rotations of the faces. We further rejected false face detections using a face-non face SVM classifier.

We proposed two systems: **System 1** applied the Facial Trait Code (FTC) [21]. FTC is a component based approach. It defines the $N$ most discriminative local facial features on human faces. For each local feature, some prominent patterns are defined and symbolized for facial coding. The original version of FTC encodes a facial image into a codeword composed of $N$ integers. Each integer represents a pattern for a local feature. In this competition, we used 100 local facial features, each had exactly 100 patterns, and it made up a feature vector of 100 integer numbers for each face. **System 2** applied the Probabilistic Facial Trait Code (PFTC), which is an extension of FTC. PFTC encodes a facial image into a codeword composed of $N$ probability distributions. These distributions gives more information on similarity and dissimilarity between a local

facial image patch and prominent patch patterns, and the PFTC is argued to outperform the original FTC. The associating study is currently under review. In this competition, we used 100 local facial features, each had exactly 100 patterns, and it made up a feature vector of 10000 real numbers for each face.

We collected at most 10 faces (in 10 frames) from an enrollment video. Each collected face was encoded into a gallery codeword. We collected at most 5 faces from a testing video. Each collected face was encoded into a probe codeword. Then, this probe codeword was matched against known gallery codewords.

### 4.9 iTEAM, Universidad Politecnica Valencia (UPV)

The UPV submitted two face recognition systems. Both systems use the same method for feature extraction and dimensionality reduction which are based on HOG-EBGM algorithm [3] and Kernel Fisher Analysis (KFA) [23] respectively. KFA was trained using face images from the FERET database [28] and ten face images of each person of the MOBIO training set. Similarity measurements are computed using the cosine distance. Our systems differed only in the way in which the faces were extracted from the video sequence. **System 1** extracts faces from each frame independently using the OpenCV AdaBoost implementation [22] . **System 2** uses a commercial closed solution [26] for face detection and also introduces a Kalman filter to track the eyes and reduce the eye detection noise.

## 5 Speaker Verification Systems

### 5.1 Brno University of Technology (BUT)

Brno University of Technology submitted two audio speaker verification systems and one fusion of these two systems. The first system is Joint Factor Analysis and the second one iXtractor system. Both systems used for training the MOBIO data but also other data mainly from NIST SRE evaluations. Both system use 2048 Gaussians in UBM.

**System 1** – Joint factor analysis (JFA) system closely follows the description of "Large Factor Analysis model" in Patrick Kenny's paper [20]. **System 2** – I-vector system was published in [15] and is closely related to the JFA framework. While JFA effectively splits model parameter space into wanted and unwanted variability subspaces, i-vector system aims at describing the subspace with the highest overall variability.

### 5.2 University of Avignon (LIA)

The LIA submitted two speakers recognition systems. Both are based on the UBM/GMM (Universal Background Model / Gaussian Mixture Model) paradigm without factor analysis. During this evaluation, development and training (even UBM training) were processed by using only MOBIO corpus.

The two systems, **LIA system 1** and **LIA system 2** differ by the acoustic parametrisation and the number of Gaussian components into the UBM. For the LIA system 1, the acoustic vectors are composed of 70 coefficients and the UBM has 512 components while LIA system 2 has only 50 coefficients, a bandwidth limited to the 300-3400Hz range and a UBM with 256 Gaussian components.

### 5.3 Tecnologico de Monterrey, Mexico and Arizona State University, USA (TEC-ASU)

Our speaker verification system, named TECHila, is based on a Gaussian Mixture Model (GMM) framework. The speech signal was downsampled to $8$ KHz and a short-time 256-pt Fourier analysis is performed on a 25ms Hamming window (10ms frame rate). Every frame log-energy was tagged as high, medium and low (low and $80\%$ of the medium log-energy frames were discarded). The magnitude spectrum was transformed to a vector of Mel-Frequency Cepstral Coefficients (MFCCs). Further, a feature warping algorithm is applied on the obtained features. Afterwards, a gender-dependent 512-mixture GMM UBM was initialised using k-means algorithm and then trained by estimating the GMM parameters via the EM (expectation maximization) algorithm. Target-dependent models were then obtained with MAP (maximum a posteriori) speaker adaptation. Finally, the score computation followed a hypothesis test framework.

Two approaches were used: a) *System 1* composed of 16 static Cepstral, 1 log Energy, and 16 delta Ceptral coefficient and single file adaptation (7 seconds of speech). b) *System 2* composed of 16 static Cepstral, 1 log Energy, 16 delta Ceptral coefficient, 16 double delta coefficient and all file adaptation (using the set of all target files).

### 5.4 University of West Bohemia (UWB)

Systems proposed by UWB made use of Gaussian Mixture Models (GMMs) and Support Vector Machines (SVMs), 4 systems were submitted. In the feature extraction process the speech signal was downsampled to 16kHz and voice activity detector was applied to discard non-speech frames. Subsystems exploited MFCCs extracted each 10 ms utilizing a 25 ms hamming window, delta's were added, simple mean and variance normalization was applied. GMMs were adapted from Universal Background Model (UBM) according to MAP adaptation with relevance factor 14. UBM consisted of 510 mixtures. UBM and impostors for SVM modeling were chosen from the world-set supplied by MOBIO in a gender specific manner. Score normalization was not utilized.

The specific systems were **System 1**: GMM-UBM [29], **System 2**: SVM-GLDS [7], **System 3**: SVM-GSV [8], and **System 4** was their combination. Regarding low amount of impostor data, the best performing system turned out to be **System 1** followed by **System 4**. However, for females **System 4** slightly outperformed **System 3**.

### 5.5 Swansea University and Validsoft (SUV)

The speaker verification systems submitted by Swansea University and Validsoft are based on standard GMM-MAP systems [29], whose originality lies in the use of wide band ($0$ to $24$ kHz) mel frequency cepstral coefficients (MFCCs) features, an idea already explored by Swansea University during the Biosecure evaluation campaign [17].

**System 1** is a GMM-MAP system with a large number filter bands ($50$) and cepstral coefficients ($29$). **System 2** is a GMM-MAP system based on a standard number of filter bands ($24$) and cepstral coefficients ($16$). **System 3** is a score level fusion of System 1 and System 2 after T-normalisation.

## 6 Discussion

In this section, the results of the MOBIO uni-modal face and speaker verification evaluation are summarised and discussed.

**Table 1.** Table presenting the results (HTER) of the best performing face verification systems for each participants on the Test set.

| | Male | Female | Average |
|---|---|---|---|
| IDIAP* | 25.45% | 24.39% | 24.92% |
| ITI* | 16.92% | 17.85% | 17.38% |
| NICTA* | 25.43% | 20.83% | 23.13% |
| TEC* | 31.36% | 29.08% | 30.22% |
| UNIS* | 9.75% | 12.07% | 10.91% |
| VISIDON* | 10.30% | 14.95% | 12.62% |
| UON* | 29.80% | 23.89% | 26.85% |
| NTU* | 20.50% | 27.26% | 23.88% |
| UPV* | 21.86% | 23.84% | 22.85% |

**Table 2.** Table presenting the results (HTER) of the best performing speaker verification systems for each participants on the Test set.

| | Male | Female | Average |
|---|---|---|---|
| BUT* | 10.47% | 10.85% | 10.66% |
| LIA* | 14.49% | 15.70% | 15.10% |
| SUV* | 13.57% | 15.27% | 14.42% |
| TEC* | 15.45% | 17.41% | 16.43% |
| UWB* | 11.18% | 10.00% | 10.59% |

### 6.1 Face verification

A summary of the results of the face verification systems can be found in Table 1. The results of the same systems are also presented in the DET plots in Figure 1 (male trials) and in Figure 2 (female trials).

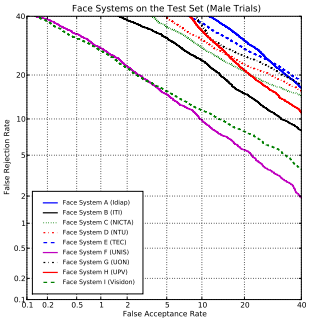


**Fig. 1.** DET plot of face verification systems on the test set (male trials).

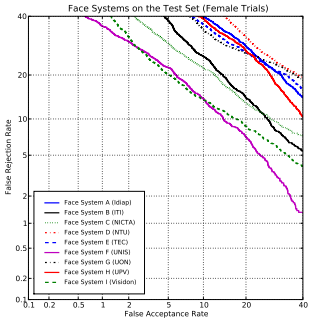


**Fig. 2.** DET plot of face verification systems on the test set (female trials).

From the plots, it can be observed mainly three groups of systems (more distinctly for female trials). The first group is composed by the two best performing systems. The best performance, with an HTER of 10.9%, is obtained by the UNIS System 4 (norm) which is fusing multiple cues and is post-processing the scores (score normalisation). This system without score normalisation, UNIS System 4, obtained an HTER of 12.9%. The second best performance is obtained by the VISIDON System 1 with an HTER of 12.6% and is using local filters but no score normalisation. Interestingly, it should be noticed that these systems use a proprietary software for the task of face detection. The second group is composed of two systems, ITI System 2 and NICTA System 2 (norm). ITI System 2 is also using a proprietary software for face detection (the same than UNIS System 4) while NICTA System 2 (norm) is using OpenCV for that task.

Interestingly, NICTA System 2 (with normalisation) performs better on the female test set than on the male test. This is the opposite trend to what occurs for most of the

other systems (such as the UNIS, VISIDON and ITI systems) where better results are obtained on the male test set than on the female test set. The third group is composed mainly by all the remaining systems and obtained an HTER of more than $20\%$. The majority of these systems uses an OpenCV like face detection scheme and all seem to have similar performance.

From these results we can draw two conclusions: (1) the choice of the face detection system can have an important impact on the face verification performance, and (2) the role of score normalisation on the performance is difficult to establish clearly.

The impact of the face detection algorithm can be seen clearly when examining the two systems from ITI. The difference between these two systems from ITI comes only from the use of a different face detection technique: ITI System 1 uses the frontal OpenCV face detector and ITI System 2 uses the OmniPerception SDK. The difference in face detector alone leads to an absolute improvement of the average HTER of more than $4\%$. This leads us to conclude that one of the biggest challenges for video based face recognition is the problem of accurate face detection.

A second interesting conclusion is that score normalisation can be difficult to apply to face recognition. This can be seen by examining the performance of the systems from UNIS and NICTA. The NICTA results show that score normalisation provides a minor but noticeable improvement in performance. However, the UNIS systems provide conflicting results as score normalisation on Systems 1 and 2 degrades performance whereas score normalisation on Systems 3 and 4 improves performance. The only conclusion that can be brought from this is that more work is necessary to be able to successfully apply score normalisation to face verification.

### 6.2 Speaker verification

A summary of the results for the speaker verification systems is presented in terms of HTER in Table 2 and also in DET plots in Figure 3 (male trials) and in Figure 4 (female trials). Generally, the audio systems exhibit smaller dispersion of HTER scores than their video counterparts, which can be attributed to smaller differences between individual audio systems than between those for videos.

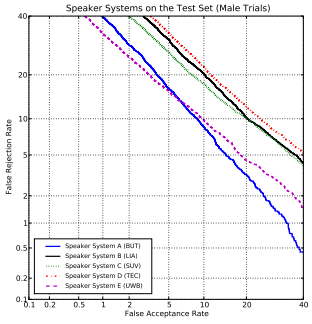


**Fig. 3.** DET plot of speaker verification systems on the test set (male trials).

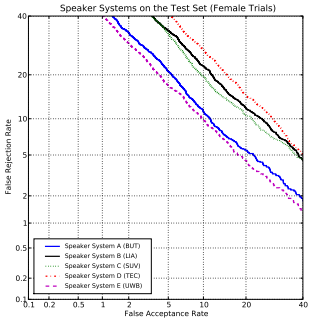


**Fig. 4.** DET plot of speaker verification systems on the test set (female trials).

From the results it can be seen that voice activity detection (VAD) is crucial for all audio systems (just as face detection is crucial for face verification). The participants use largely different approaches from classical energy based (LIA, TEC-ASU) through to sub-band quality measures (UWB) and the use of phone recognizers (BUT). By contrast , the variability in feature extraction is much smaller with most participants using standard MFCC coefficients with some variants.

For the speaker verification part, two approaches were adopted: GMM-UBM and SVM-based. The former ones were generally weaker in performances, with the exception of UWB System1 - a pure GMM-UBM based system that was the best performing single system. This performance is probably due to UWB VAD, their system is also fully trained on MOBIO 16kHz data.

The later approach (SVM) performed well both on standard GMM means (UWB) as well as on JFA-derived speaker factors (BUT System1). This supports the conclusion that SVMs provide superior performance on shorter segments of speech.

The importance of score normalisation was also confirmed, mainly for the systems not based on SVMs. However, it was hard to derive representative gender dependent ZT-norm cohorts, mainly because there were too few speakers in the world-set of the MOBIO database.

Another lesson learned was the importance of the target (MOBIO) data for training when compared to the hundreds hours of non-target (NIST) telephone data. It can be seen that the SVM-based techniques largely benefit from having this data in their imposter sets. On the other hand, JFA does not improve with this data as the utterances are too short and too few.

## 7 Conclusion

This paper presented the results of several uni-modal face and speaker verification techniques on the MOBIO database (Phase I). This database provides realistic and challenging conditions as it was captured on a mobile device and in uncontrolled environments.

The evaluation was organised in two stages. During the first stage, the training and development sets of the database was distributed among the participants (from December 1 2009 to January 27 2010). The deadline for the submission of the first results by the participants on the development set was February 1 2010. During the second stage, the test set was distributed only to the participants that met the first deadline. The deadline for the submission of the results on the test set was March 8 2010.

Out of the thirty teams that signed the End User License Agreement (EULA) of the database and downloaded it, finally, fourteen teams have participated to this evaluation. Eight teams participated to the face verification part of the evaluation, four teams participated to the speaker verification part of the evaluation and one team participated both to the face and the speaker part. Only one team dropped from the competition during the second stage. Each participant provided at least the results of one system but were allowed to submit the results of several systems.

This evaluation produced three interesting findings. First, it can be observed that face verification and speaker verification obtained the same level of performance. This

is particularly interesting because it is generally observed that speaker verification performs much better than face verification in general. Second, it has been highlighted that segmentation (face detection and voice activity detection) was critical both for face and speaker verification. Finally, it has been shown that the two modalities are complementary as a clear gain in performance can be obtained simply by fusing the individual face and speaker verification scores.

Overall, it was shown that the MOBIO database provides a challenging test-bed both for face verification, for speaker verification but also for bi-modal verification. This evaluation would have established baseline performance for the MOBIO database.

The MOBIO consortium is planning to distribute the database (Phase I) in August 2010 together with the results and the annotations (face detection output) generated by the participants during this evaluation. It is foreseen as well to distribute the Phase II of the MOBIO database before the end of 2010.

## Acknowledgements

This work has been performed by the MOBIO project 7th Framework Research Programme of the European Union (EU), grant agreement number: 214324. The authors would like to thank the EU for the financial support and the partners within the consortium for a fruitful collaboration. For more information about the MOBIO consortium please visit `http://www.mobioproject.org`.

The authors would also like to thank Phil Tresadern (University of Manchester), Bastien Crettol (Idiap Research Institute), Norman Poh (University of Surrey), Christophe Levy (University of Avignon), Driss Matrouf (University of Avignon), Timo Ahonen (University of Oulu), Honza Cernocky (Brno University of Technology) and Kamil Chalupnicek (Brno University of Technology) for their work in capturing this database and development of the protocol.

NICTA is funded by the Australian Government as represented by the *Department of Broadband, Communications and the Digital Economy* as well as the Australian Research Council through the *ICT Centre of Excellence* program.

## References

1. Visidon ltd., (http://www.visidon.fi)
2. Ahonen, T., Hadid, A., Pietikainen, M.: Face Recognition with Local Binary Patterns. Lecture Notes in Computer Science pp. 469–481 (2004)
3. Albiol, A., Monzo, D., Martin, A., Sastre, J., Albiol, A.: Face recognition using hog-ebgm. Pattern Recognition Letters 29(10), 1537–1543 (2008)
4. Bengio, S., Mariéthoz, J., Keller, M.: The Expected Performance Curve. In: Intl Conf. On Machine Learning (ICML) (2005)
5. Bledsoe, W.W.: The model method in facial recognition. Tech. rep., Panoramic Research Inc. (1966)
6. Campbell, J.P.: Speaker recognition: A tutorial. Proceedings of the IEEE 85(9) (Sep 1997)
7. Campbell, W.: Generalized linear discriminant sequence kernels for speaker recognition. IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP'02 1, I–161–I–164 (2002)

8. Campbell, W., Sturim, D., Reynolds, D.: Support vector machines using gmm supervectors for speaker verification. Signal Processing Letters, IEEE 13(5), 308–311 (2006)
9. Campbell, W., Sturim, D., Reynolds, D., Solomonoff, A.: Svm based speaker verification using a gmm supervector kernel and nap variability compensation. Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings 1, I–I (2006)
10. Cardinaux, F., Sanderson, C., Marcel, S.: Comparison of mlp and gmm classifiers for face verification on xm2vts. In: International Conference on Audio- and Video-based Biometric Person Authentication. pp. 1058–1059 (2003)
11. Chan, C., Kittler, J., Poh, N., Ahonen, T., Pietikäinen, M.: (multiscale) local phase quantization histogram discriminant analysis with score normalisation for robust face recognition. In: VOEC. pp. 633–640 (2009)
12. Chan, C.H., Kittler, J., Messer, K.: Multi-scale local binary pattern histograms for face recognition. In: Lee, S.W., Li, S.Z. (eds.) ICB. Lecture Notes in Computer Science, vol. 4642, pp. 809–818. Springer (2007)
13. Comaniciu, D., Ramesh, V., Meer, P.: Real-time tracking of non-rigid objects using mean shift. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR2000). pp. 142–149 (2000)
14. Das, A.: Audio visual person authentication by multiple nearest neighbor classifiers. In: SpringerLink (2007)
15. Dehak, N., Dehak, R., Kenny, P., Brmmer, N., Ouellet, P., Dumouchel, P.: Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In: Proc. International Conferences on Spoken Language Processing (ICSLP). pp. 1559–1562 (Sep 2009)
16. Ekenel, H., Fischer, M., Jin, Q., Stiefelhagen, R.: Multi-modal person identification in a smart environment. In: IEEE CVPR (2007)
17. Fauve, B., Bredin, H., Karam, W., Verdet, F., Mayoue, A., Chollet, G., Hennebert, J., Lewis, R., Mason, J., Mokbel, C., Petrovska, D.: Some results from the biosecure talking face evaluation campaign. In: Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP) (2008)
18. Hadid, A., Pietikäinen, M.: Manifold learning for video-to-video face recognition. In: COST 2101/2102 Conference. pp. 9–16 (2009)
19. Hieu, N., Bai, L., Shen, L.: Local gabor binary pattern whitened pca: A novel approach for face recognition from single image per person. In: The 3rd IAPR/IEEE International Conference on Biometrics, 2009. Proceedings. (2009)
20. Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Dumouchel, P.: A study of inter-speaker variability in speaker verification. In: IEEE Transactions on Audio, Speech and Language Processing (Jul 2008)
21. Lee, P.H., Hsu, G.S., Hung, Y.P.: Face verification and identification using facial trait code. IEEE Conference on Computer Vision and Pattern Recognition pp. 1613–1620 (2009)
22. Lienhart, R., Kuranov, A., Pisarevsky, V.: Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In: DAGM'03, 25th Pattern Recognition Symposium. pp. 297–304. Madgeburg, Germany (2003)
23. Liu, C.: Capitalize on dimensionality increasing techniques for improving face recognition grand challenge performance. IEEE Transactions on Pattern Analysis and Machine Intelligence 28, 725–737 (2006)
24. Marcel, S., McCool, C., Matějka, P., Ahonen, T., Černocký, J., Chakraborty, S., Balasubramanian, V., Panchanathan, S., Chan, C.H., Kittler, J., Poh, N., Fauve, B., Glembek, O., Plchot, O., Jančík, Z., Larcher, A., Lévy, C., Matrouf, D., Bonastre, J.F., Lee, P.H., Hung, J.Y., Wu, S.W., Hung, Y.P., Machlica, L., Mason, J., Mau, S., Sanderson, C., Monzo, D.,

Albiol, A., Albiol, A., Nguyen, H., Li, B., Wang, Y., Niskanen, M., Turtinen, M., Nolazco-Flores, J.A., Garcia-Perera, L.P., Aceves-Lopez, R., Villegas, M., Paredes, R.: Mobile biometry (MOBIO) face and speaker verification evaluation. Idiap-RR Idiap-RR-09-2010, Idiap Research Institute (May 2010)
25. Messer, K., Kittler, J., Sadeghi, M., Hamouz, M., Kostin, A., Cardinaux, F., Marcel, S., Bengio, S., Sanderson, C., Poh, N., Rodriguez, Y., Czyz, J., Vandendorpe, L., McCool, C., Lowther, S., Sridharan, S., Chandran, V., Palacios, R.P., Vidal, E., Bai, L., Shen, L., Wang, Y., Yueh-Hsuan, C., Hsien-Chang, L., Yi-Ping, H., Heinrichs, A., Muller, M., Tewes, A., von der Malsburg, C., Wurtz, R., Wang, Z., Xue, F., Ma, Y., Yang, Q., Fang, C., Ding, X., Lucey, S., Goss, R., Schneiderman, H.: Face authentication test on the banca database. In: Proceedings of the 17th International Conference on Pattern Recognition. vol. 4, pp. 523–532 (2004)
26. Neurotechnologija: Verilook SDK, neurotechnologija Biometrical and Artificial Intelligence Technologies (http://www.neurotechnologija.com)
27. Phillips, J., Flynn, P., Scruggs, T., Bowyer, K., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. In: IEEE Conference of Computer Vision and Pattern Recognition. vol. 1, pp. 947–954 (2005)
28. Phillips, J.P., Moon, H., Rizv, S., Rauss, P.J.: The FERET evaluation methodology for face-recognition algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(10), 1090–1104 (2000)
29. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted gaussian mixture models. Digital Signal Processing 10(1-3), 19 – 41 (2000)
30. Rodriguez, Y.: Face Detection and Verification using Local Binary Patterns. Ph.D. thesis, EPFL (2006)
31. Sanderson, C., Lovell, B.C.: Multi-region probabilistic histograms for robust and scalable identity inference. In: International Conference on Biometrics, Lecture Notes in Computer Science (LNCS). vol. 5558, pp. 199–208 (2009)
32. Shriberg, E., Ferrer, L., Kajarekar, S.: Svm modeling of snerf-grams for speaker recognition. In: International Conference on Spoken Language Processing (ICSLP). Jeju Island, Korea (Oct 2004)
33. Stolcke, A., Ferrer, L., Kajarekar, S., Shriberg, E., Venkataraman, A.: MLLR transforms as features in speaker recognition. In: International Conference on Spoken Language Processing (ICSLP). pp. 2425–2428. Lisbon, Portugal (Sep 2005)
34. Tan, X., Chen, S., Zhou, Z.H., Zhang, F.: Face recognition from a single image per person: A survey. Pattern Recognition 39(9), 1725–1745 (2006)
35. Tistarelli, M., Bicego, M., Grosso, E.: Dynamic face recognition: From human to machine vision. Image and Vision Computing 27(3), 222 – 232 (2009)
36. Villegas, M., Paredes, R.: Simultaneous learning of a discriminative projection and prototypes for nearest-neighbor classification. Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on pp. 1–8 (2008)
37. Villegas, M., Paredes, R., Juan, A., Vidal, E.: Face verification on color images using local features. Computer Vision and Pattern Recognition Workshops, 2008. CVPR Workshops 2008. IEEE Computer Society Conference on pp. 1–6 (Jun 2008)
38. Wolf, L., Hassner, T., Taigman, Y.: Descriptor based methods in the wild. In: Real-Life Images workshop at the European Conference on Computer Vision (ECCV) (October 2008)
39. Zhang, W., Shan, S., Gao, W., Chen, X., Zhang, H.: Local Gabor Binary Pattern Histogram Sequence (LGBPHS): A Novel Non-Statistical Model for Face Representation and Recognition. In: Proc. ICCV. pp. 786–791 (2005)
40. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. ACM Computing Surveys 35(4), 399–458 (Dec 2003)