IDIAP RESEARCH REPORT

BI-MODAL BIOMETRIC AUTHENTICATION ON MOBILE PHONES IN CHALLENGING CONDITIONS

Elie Khoury Laurent El Shafey Chris McCool
Manuel Günther Sébastien Marcel

Idiap-RR-30-2013

OCTOBER 2013

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Bi-Modal Biometric Authentication on Mobile Phones in Challenging Conditions

Elie Khoury[a], Laurent El Shafey[a], Christopher McCool[b], Manuel Günther[a], Sébastien Marcel[a]

[a]*Idiap Research Institute, Martigny, Switzerland*
{*elie.khoury,laurent.el-shafey,manuel.guenther,marcel*}*@idiap.ch*
[b]*NICTA, Queensland Research Laboratory, Australia*
chris.mccool@nicta.com.au

Abstract

This paper examines the issue of face, speaker and bi-modal authentication in mobile environments when there is significant condition mismatch. We introduce this mismatch by enrolling client models on high quality biometric samples obtained on a laptop computer and authenticating them on lower quality biometric samples acquired with a mobile phone. To perform these experiments we develop three novel authentication protocols for the large publicly available MOBIO database. We evaluate state-of-the-art face, speaker and bi-modal authentication techniques and show that inter-session variability modelling using Gaussian mixture models provides a consistently robust system for face, speaker and bi-modal authentication. It is also shown that multi-algorithm fusion provides a consistent performance improvement for face, speaker and bi-modal authentication. Using this bi-modal multi-algorithm system we derive a state-of-the-art authentication system that obtains a half total error rate of 6.3% and 1.9% for Female and Male trials, respectively.

Keywords: face authentication, speaker authentication, bi-modal authentication, Gaussian mixture model, session variability, inter-session variability, total variability, i-vector, fusion

1. Introduction

Mobile phones have become an integral part of many people's daily life. They are used not just for telephonic communication, but also to send and receive

emails, take photos or even have video conversations. This has led to the mobile phone being an inherently multimedia device, which often has a front-facing camera in addition to the standard microphone. Hence, it forms an exciting new device that allows researchers to explore the applicability of bi-modal (face and speaker) authentication in challenging mobile phone environments.

This exciting challenge of bi-modal authentication in the mobile phone environment has begun to receive more attention. An international competition was organised in 2010 [1], where researchers evaluated state-of-the-art algorithms for face and speaker authentication using Phase I of the MOBIO database [2]. In this evaluation, enrolment was exclusively performed with mobile phone data. It was shown that a combination of these systems produced an impressive bi-modal authentication system. Since then other researchers have examined methods to perform face [3, 4], speaker [5, 6] and bi-modal [7, 8] authentication in the challenging mobile phone environment.

A theme common to some of the prior work on biometric authentication in a mobile environment is the idea of *session variability* modelling [5, 4], which achieves state-of-the-art results for bi-modal authentication [8]. Session variability modelling aims to estimate and suppress any variability such as audio or image noise that may cause confusion between different observations of the same biometric identity. In [5] session variability modelling was used to cope with audio channel variability, while [4] introduced this concept to face authentication, where its application was supposed to reduce the impact of pose and illumination variation. Finally, in [8] state-of-the-art face and speaker authentication systems that used *inter-session variability* (ISV) modelling were combined to derive a state-of-the-art bi-modal authentication system. However, this prior work applied ISV modelling in the case of matched acquisition conditions, i. e., where biometric test samples are acquired using the same device as employed for client model enrolment. Furthermore, they did not use the most recent advances such as *total variability* (TV) modelling, which has been applied to speaker [9] and face [10] authentication.

In this paper we explore three issues of applying bi-modal authentication to the challenging mobile phone environment. First, we examine the issue of mismatched conditions between enrolment and testing. In particular, we examine the effect of enrolling users on high quality biometric samples acquired with a laptop computer and then authenticating them using lower quality biometric samples acquired with a mobile phone. As a significant contribution, we develop three new protocols for the MOBIO database [2] with respect to prior work [1, 2, 4, 8] that was exclusively using mobile phone data both for enrolment and testing.

Second, we extend the work of [8] by examining the effectiveness of TV modelling for bi-modal authentication. Third, we show the effectiveness of multi-algorithm fusion to further improve the results for face, speaker and bi-modal authentication in the mobile phone environment. The final outcome of this work is the development of a state-of-the-art bi-modal (face and speaker) authentication system that improves upon the previous state-of-the-art [8] with a relative performance gain of 35% for Female and 27% for Male trials on the MOBIO database.

The remainder of this paper is structured as follows: In Section 2 we outline the employed face and speaker authentication systems, while Section 3 combines these into bi-modal and multi-algorithm authentication systems. Section 4 presents the new protocols for the MOBIO database that are used in our experiments, which we discuss in Section 5. Finally, Section 6 concludes the paper.

2. Face and speaker authentication systems

We examine the effectiveness of state-of-the-art *Gaussian mixture model* (GMM) based approaches for face, speaker and bi-modal authentication. GMMs have formed the basis of state-of-the-art speaker authentication systems for over a decade [11, 9] and it was recently shown that incorporating session variability modelling into a GMM system produces state-of-the-art results for face authentication [12]. Also, the combination of GMM-based systems that use session variability modelling produces a state-of-the-art bi-modal (face and speaker) authentication system [8].

When using GMMs and session variability for speaker and face authentication, the same underlying approach is taken. The main difference is how the feature vectors are extracted from the image (face) and audio (speech) samples. Below we describe the feature extraction process for both face and speaker authentication followed by a description of the GMM and the associated session variability modelling approaches that we examine.

2.1. Feature extraction

Two separate feature extraction processes are used for image (face) and audio (speech) data. For both modalities, a biometric sample $\boldsymbol{\mathcal{O}}$ (image or audio) is decomposed into a set $\boldsymbol{O}$ of K feature vectors ($\boldsymbol{O} = \{\boldsymbol{o}^1, \boldsymbol{o}^2, \cdots, \boldsymbol{o}^K\}$), where each feature vector is of dimensionality M. This decomposition is performed in the spatial domain for the image data, and in the time domain for the audio data.

2.1.1. *Face-based features*

For the image data, we rely on parts-based features that were proposed for the task of face authentication in [13]. These features have since been successfully employed by several researchers [14, 15]. The key idea is to decompose the face image into a set of overlapping blocks before extracting a feature vector from each of them. The feature vectors extracted from these blocks are then considered as observations of the same signal (the same face), and can be modelled in a generative way.

The feature extraction process is similar to the approach described in [16]. First, each image is rotated, scaled and cropped to 64×80 pixels such that the eyes are 16 pixels from the top and separated by 33 pixels. Second, to reduce the impact of illumination, each cropped image is preprocessed with the multi-stage algorithm of Tan & Triggs [17], using their default parametrisation. Third, 12×12 blocks of pixel values are extracted from the preprocessed image using an exhaustive overlap, leading to $K = 3657$ blocks per image. Fourth, pixel values of each block are normalised to zero mean and unit variance, prior to extracting the $M + 1$ lowest frequency *2D discrete cosine transform* (2D-DCT) coefficients [13] and removing the zero frequency coefficient as it is redundant. Fifth, the resulting 2D-DCT feature vectors are normalised to zero mean and unit variance in each dimension with respect to the other feature vectors of the image. As in previous work [16, 8], M was set equal to 44.

2.1.2. *Speaker-based features*

For the audio data, observations are extracted at equally-spaced time instants using a sliding window approach. First, audio segments are denoised using the Qualcomm-ICSI-OGI front end [18]. Second, *voice activity detection* (VAD) is performed jointly using the normalised log energy and the 4 Hz modulation energy [19]. The aim of the 4 Hz modulation energy is to discriminate speech from other audio sources such as noise and music. An adaptive threshold is applied on both the 4 Hz modulation energy and the normalised log energy. In our experiments, this approach provided a relative improvement of up to 16% compared to the common energy-based VAD. Third, 19 *mel frequency cepstrum coefficient* (MFCC) and log energy features together with their first- and second-order derivatives are obtained by computing 24 filter bank coefficients over 20 ms Hamming windowed frames every 10 ms. This results in acoustic feature vectors of dimensionality $M = 60$. Finally, feature normalisation based on *cepstral mean and variance normalisation* (CMVN) is applied on the remaining speech. The number of feature vectors K extracted from each audio sample depends on the duration of

the sample and the number of segments that the VAD classifies to be speech.

2.2. GMM-based modelling

We use the same generative probabilistic framework that models the observed feature vectors using *Gaussian mixture models* (GMMs) for both image (face) and audio (speech) modalities. GMMs have been successfully applied first to speaker authentication [20, 11] and then to face authentication [13, 21, 14, 15, 16]. One of the main challenges with GMMs is to reliably estimate a client model with limited enrolment data. This enrolment process is sensitive to the conditions, in which the data was captured. To address this issue, several session variability modelling techniques built on the GMM baseline have been proposed that constrain client models to be in a restricted subspace. In this work, we consider two approaches to session variability modelling, *inter-session variability* (ISV) modelling [22] and *total variability* (TV) modelling [9]. Both methods were initially proposed for speaker authentication [22, 9] before being applied to face authentication [12, 10]. In the remainder of this section, we first describe the GMM baseline system, followed by the more advanced ISV and TV techniques.

2.2.1. Gaussian mixture modelling

The distribution of the observed feature vectors (face or speech) is modelled using a GMM. A GMM is the weighted sum of C multi-variate Gaussian components $\mathcal{N}$:

$$p(\boldsymbol{o}|\boldsymbol{\Theta}_{\text{gmm}}) = \sum_{c=1}^{C} \omega_c \mathcal{N}\left(\boldsymbol{o}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\right) , \tag{1}$$

where $\boldsymbol{\Theta}_{\text{gmm}} = \{\omega_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\}_{c=\{1,\ldots,C\}}$ are the parameters of this distribution: the weights, the means and the covariance matrices, respectively.

To use GMMs for authentication we need to learn a GMM $\boldsymbol{\mathcal{S}}_i$ for each subject i from a set of enrolment samples. One of the main challenges is that the number of enrolment images or audio recordings per client is usually limited, possibly to a single sample. In practice, it has been shown that for both speaker [11] and face authentication [14, 15] an efficient enrolment method is to use a subject-independent prior GMM $\boldsymbol{\mathcal{M}}$, called the *universal background model* (UBM), and to adapt this prior to the enrolment samples of the subject i to generate the client model $\boldsymbol{\mathcal{S}}_i$. The UBM $\boldsymbol{\mathcal{M}}$ is learnt beforehand by maximising the likelihood of observations extracted from a large independent training set of several identities using the iterative *expectation-maximisation* (EM) algorithm [23]. Afterwards, adaptation is achieved by using *maximum a posteriori* (MAP) estimation [20],

where only the means of the UBM are updated, as this has been shown to be efficient for both modalities [20, 14, 15]. As in previous work [11, 14, 15, 16] GMMs are assumed to have diagonal covariance matrices.

A convenient and compact representation of mean-only MAP adaptation and other session variability modelling techniques is the GMM super-vector notation. This notation consists of grouping the parameters of the various Gaussian components of a GMM (weights, means or covariance matrices) into single large vectors or matrices. For instance, the mean super-vector $\boldsymbol{m}$ of the UBM $\boldsymbol{\mathcal{M}}$ is obtained by concatenating the means $\boldsymbol{\mu}_c$ of all its components: $\boldsymbol{m} = \left[\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_c^T, \cdots, \boldsymbol{\mu}_C^T\right]^T$. In [22] it was shown that mean-only MAP adaptation can then be written as:

$$\boldsymbol{s}_i = \boldsymbol{m} + \boldsymbol{d}_i \,, \tag{2}$$

where $\boldsymbol{s}_i$ is the mean super-vector of the GMM $\boldsymbol{\mathcal{S}}_i$ and $\boldsymbol{d}_i$ is a client-specific offset for subject i. This offset $\boldsymbol{d}_i$ is given by:

$$\boldsymbol{d}_i = \boldsymbol{D}\boldsymbol{z}_i \,, \text{ with } \boldsymbol{D} = \sqrt{\frac{\boldsymbol{\Sigma}}{\tau}} \,, \tag{3}$$

where $\boldsymbol{\Sigma}$ is the variance super-vector of the UBM (recalling that covariance matrices are assumed to be diagonal), τ is a relevance factor that provides a weight for the prior UBM when performing MAP adaptation [20] and $\boldsymbol{z}_i$ is a latent variable, which is assumed to be normally distributed $\mathcal{N}\left(0, \boldsymbol{I}\right)$. In the following, since only the means of the UBM are adapted, we will use $\boldsymbol{s}_i$ and $\boldsymbol{m}$ to describe the client model $\boldsymbol{\mathcal{S}}_i$ and the UBM $\boldsymbol{\mathcal{M}}$, respectively, by abusing the notation.

Scoring. Once a client model is enrolled, a test sample $\boldsymbol{\mathcal{O}}_t$ (also called a probe sample) is authenticated against the model $\boldsymbol{s}_i$ by calculating a *log-likelihood ratio* (LLR) score with respect to the UBM $\boldsymbol{m}$ [11]:

$$h_{\text{gmm}}\left(\boldsymbol{\mathcal{O}}_t, \boldsymbol{s}_i\right) = \sum_{k=1}^{K_t} \left[\log\left(p\left(\boldsymbol{o}_t^k \mid \boldsymbol{s}_i\right)\right) - \log\left(p\left(\boldsymbol{o}_t^k \mid \boldsymbol{m}\right)\right)\right] . \tag{4}$$

With higher $h_{\text{gmm}}\left(\boldsymbol{\mathcal{O}}_t, \boldsymbol{s}_i\right)$ values, the probability increases that the observations $\boldsymbol{O}_t$ extracted from the sample $\boldsymbol{\mathcal{O}}_t$ were produced by the client model $\boldsymbol{s}_i$.

Recently, a linear approximation of Eq. (4), known as *linear scoring* [24], has been adopted in the speaker authentication literature and has also been applied to face authentication [16]. It relies on the mean centralised first order sufficient statistics of the UBM given the observations $\boldsymbol{O}_t$, as given in Equation (9) of [12].

This approximation was shown [24] to be orders of magnitude more efficient with no significant degradation in performance.

As a final step, we also perform *zt-score normalisation* [25] due to the consistent performance improvements that this gives for both face [16] and speaker authentication [26].

2.2.2. *Inter-session variability modelling*

One problem of the GMM mean-only MAP adaptation approach is that the client model $\boldsymbol{s}_i$ can be difficult to estimate reliably with limited enrolment data as it is sensitive to the conditions in which the data was captured. Part of the reason for this is that there is no explicit model to capture and suppress detrimental variations such as audio or image noise. *Session variability* modelling aims to estimate and suppress the effects of within-client variations in order to create more discriminant client models. For the face modality, within-client variations include variations of pose, illumination or expression of samples of a given subject, whereas for the speaker modality variations are, amongst others, caused by the sensor (microphone) or the environment (background noise or acoustic conditions).

Inter-session variability (ISV) modelling [22] and *joint factor analysis* (JFA) [27] are two session variability modelling techniques, used in the context of a GMM-based system, that have been successfully applied to both speaker [22, 28, 26] and face authentication [4, 12]. Both techniques aim to estimate session variation like audio or image noise in order to compensate for it. This compensation for the estimated session variation is the key difference between ISV and the classic mean-only MAP adaptation. Note that for these experiments we have not examined JFA as it was shown empirically that ISV outperforms JFA for both speaker [8] and face [8, 12] authentication.

As in [22] it is assumed that session variability results in an additive offset to the mean super-vector $\boldsymbol{s}_i$ of the client model. This offset can be added directly to the normal mean-only MAP adaptation representation. Given the j-th biometric sample $\boldsymbol{\mathcal{O}}_{i,j}$ of subject i the mean super-vector $\boldsymbol{\mu}_{i,j}$ of the GMM that best represents this biometric sample is:

$$\boldsymbol{\mu}_{i,j} = \boldsymbol{m} + \boldsymbol{U}\boldsymbol{x}_{i,j} + \boldsymbol{D}\boldsymbol{z}_i \,, \tag{5}$$

where $\boldsymbol{U}$ is a subspace that constrains the possible session effects, $\boldsymbol{x}_{i,j}$ is its associated latent session variable ($\boldsymbol{x}_{i,j} \sim \mathcal{N}(0, \boldsymbol{I})$), while $\boldsymbol{D}$ and $\boldsymbol{z}_i$ represent the client-specific offset in the same manner as for mean-only MAP adaptation, which is given in Eqs. (2) and (3).

The model enrolment of a client is performed in the following manner. Given a session subspace $\boldsymbol{U}$, which is learnt by maximising the likelihood of the training data, the latent variables $\boldsymbol{x}_{i,j}$ and $\boldsymbol{z}_i$ are jointly estimated using MAP. Afterwards, the session varying part is suppressed by retaining only the client-specific information:

$$\boldsymbol{s}_i^{(\text{isv})} = \boldsymbol{m} + \boldsymbol{D}\boldsymbol{z}_i^{(\text{isv})}. \tag{6}$$

For details on how to jointly estimate the latent variables and how to train the subspace $\boldsymbol{U}$, readers are referred to [12].

Scoring. ISV relies on a LLR score similar to Eq. (4). The main differences are that the session offsets of the enrolment samples have been compensated while generating the client model, and that session offsets of the probe sample $\boldsymbol{\mathcal{O}}_t$ are estimated prior to scoring. This means that the latent session variables $\boldsymbol{x}_{i,t}$ and $\boldsymbol{x}_{\text{ubm},j}$ of the observed feature vectors $\boldsymbol{O}_t = \{\boldsymbol{o}_t^1, \boldsymbol{o}_t^2, \cdots, \boldsymbol{o}_t^{K_t}\}$ extracted from a biometric sample $\boldsymbol{\mathcal{O}}_t$ are first estimated, before computing a LLR score:

$$\begin{aligned} h_{\text{isv}}\left(\boldsymbol{\mathcal{O}}_t, \boldsymbol{s}_i\right) = \sum_{k=1}^{K^t} \Big[& \log\left(p\left(\boldsymbol{o}_t^k \mid \boldsymbol{s}_i + \boldsymbol{U}\boldsymbol{x}_{i,t}\right)\right) \\ & - \log\left(p\left(\boldsymbol{o}_t^k \mid \boldsymbol{m} + \boldsymbol{U}\boldsymbol{x}_{\text{ubm},t}\right)\right) \Big]. \end{aligned} \tag{7}$$

In practice, simplifications have been proposed to speed up the process in [24], which consists of first approximating the session offset $\boldsymbol{U}\boldsymbol{x}_{i,t}$ of the client model, with the session offset $\boldsymbol{U}\boldsymbol{x}_{\text{ubm},t}$, and then using the linear scoring approximation as for the GMM-baseline.

Finally, as with the GMM baseline, we perform zt-score normalisation.

2.2.3. Total variability modelling

In [29] it was shown that JFA can fail to separate between-client and within-client variations into two different subspaces. This is potentially caused by the high dimensionality of the GMM mean super-vector space.

To address this issue, an alternative technique called *total variability* (TV) modelling was developed for speaker authentication [30, 31] and later applied to face authentication [10]. This framework is built on the GMM approach and relies on the definition of a single subspace that contains both identity and session variabilities. In particular, it aims to extract low-dimensional factors $\boldsymbol{w}_{i,j}$, so-called *i-vectors*, from biometric samples $\boldsymbol{\mathcal{O}}_{i,j}$. More formally, the TV approach

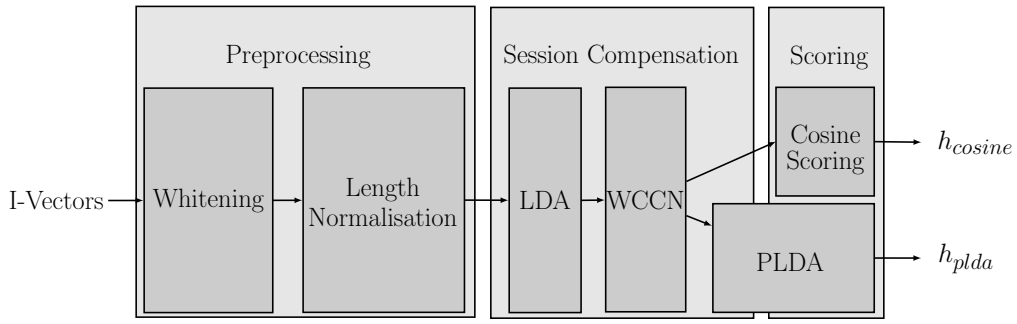


Figure 1: I-VECTOR PROCESSING TOOL CHAIN. *This figure shows the steps of the i-vector processing tool chain. Each preprocessing and session compensation step is optional.*

can be described in the GMM mean super-vector space by:

$$\boldsymbol{\mu}_{i,j} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w}_{i,j}\,, \tag{8}$$

where $\boldsymbol{T}$ is the low-dimensional total variability subspace and $\boldsymbol{w}_{i,j}$ the low-dimensional i-vector, which is assumed to follow a normal distribution $\mathcal{N}(0, \boldsymbol{I})$.

The TV subspace $\boldsymbol{T}$ is learnt by maximising the likelihood over a large training set. This algorithm is similar to the one used to estimate the identity (between-class) subspace in JFA [32], with one major difference: while JFA jointly consider the samples coming from a given subject, TV treats them as if they have been produced by different identities, which is an advantage when large unlabelled training datasets are used. In addition, the extraction of i-vectors requires the estimation of a covariance matrix $\boldsymbol{\Sigma}_T$ to model the residual variability that is not captured by the subspace $\boldsymbol{T}$.

In contrast to ISV, TV does not explicitly perform session compensation. TV is just a front-end that extracts a low dimensional i-vector $\boldsymbol{w}_{i,j}$ from each sample $\boldsymbol{\mathcal{O}}_{i,j}$ based on the total variability of the training set. As such, it is likely to capture both client-specific and session-specific information. Hence, TV requires to use separate session compensation and scoring techniques after the extraction of i-vectors. Additionally, a set of preprocessing algorithms have been proposed to map i-vectors into a more adequate space [33, 34, 10]. Possible variants of preprocessing, session compensation and scoring methods have been employed and combined in different manners [30, 31, 10]. Some of them are summarised in Fig. 1 and described in the remainder of this section.

I-vector preprocessing. First, i-vector *whitening* was proposed in [33, 10] and shown to boost classification performance. Whitening consists of normalising the

i-vector space such that the covariance matrix of the i-vectors, of a training set, is turned into the identity matrix. This is performed by applying:

$$\boldsymbol{w}_{i,j}^{(\text{whitened})} = \boldsymbol{W}^T \left(\boldsymbol{w}_{i,j} - \bar{\boldsymbol{w}} \right) , \tag{9}$$

where $\bar{\boldsymbol{w}}$ is the mean of a training set of i-vectors, $\boldsymbol{w}_{i,j}^{(\text{whitened})}$ the whitened i-vector, and $\boldsymbol{W}$ the whitening transform. This transform $\boldsymbol{W}$ is computed as the Cholesky decomposition of $\bar{\boldsymbol{\Sigma}}^{-1} = \boldsymbol{W}\boldsymbol{W}^T$, where $\bar{\boldsymbol{\Sigma}}$ is the covariance matrix of a training set of i-vectors.

Another efficient preprocessing technique is i-vector length normalisation [34, 10], which aims at reducing the impact of a mismatch between training and test i-vectors. It consists of mapping the i-vectors into a unit hypersphere:

$$\boldsymbol{w}_{i,j}^{(\text{l-norm})} = \frac{\boldsymbol{w}_{i,j}}{\|\boldsymbol{w}_{i,j}\|}, \tag{10}$$

which is very effective when using session compensation or scoring methods that assume Gaussian-like distributions.

Session compensation. A set of session compensation techniques have been proposed for both speaker [31] and face [10] authentication. *Linear discriminant analysis* (LDA) [35] is a popular algorithm that aims at learning a linear projection maximising between-class variations while minimising within-class variations. The projection matrix $\boldsymbol{A}$ is learnt from a training set of i-vectors extracted from samples coming from several identities, by first computing the between-class and within-class scatter matrices:

$$\boldsymbol{S}_W = \sum_i \sum_{j=1}^{J_i} \left(\boldsymbol{w}_{i,j} - \bar{\boldsymbol{w}}_i \right) \left(\boldsymbol{w}_{i,j} - \bar{\boldsymbol{w}}_i \right)^T \tag{11}$$

$$\boldsymbol{S}_B = \sum_i J_i \left(\bar{\boldsymbol{w}}_i - \bar{\boldsymbol{w}} \right) \left(\bar{\boldsymbol{w}}_i - \bar{\boldsymbol{w}} \right)^T \tag{12}$$

where J_i is the number of i-vectors from client i, $\bar{\boldsymbol{w}}_i$ is the mean of this client-specific i-vectors, and $\bar{\boldsymbol{w}}$ the means of all i-vectors in the training set. Next, LDA maximises the ratio of the determinants of these two scatter matrices. The solution is found by solving the generalised eigenvalue decomposition $\boldsymbol{S}_B\boldsymbol{v} = \lambda \boldsymbol{S}_W\boldsymbol{v}$. We then retain the n_{lda} eigenvectors with the greatest eigenvalues to build the projection matrix $\boldsymbol{A}$. An i-vector $\boldsymbol{w}_{i,j}$ is projected into the LDA space by:

$$\boldsymbol{w}_{i,j}^{(\text{lda})} = \boldsymbol{A}^T \boldsymbol{w}_{i,j} \, . \tag{13}$$

Within-class covariance normalisation (WCCN) is a technique initially introduced for SVM-based speaker authentication [36]. It has since be successfully applied to i-vectors for both speaker [30] and face authentication [10]. It aims to normalise the within-class covariance matrix of a training set of i-vectors. Given the within-class scatter matrix from Eq. (11) and the number of identities N in the training set, the WCCN linear transform $\boldsymbol{B}$ can be computed using the Cholesky decomposition of:

$$\left(\frac{1}{N}\boldsymbol{S}_W\right)^{-1} = \boldsymbol{B}\boldsymbol{B}^T \,. \tag{14}$$

An i-vector $\boldsymbol{w}_{i,j}$ is projected into the corresponding WCCN space by:

$$\boldsymbol{w}_{i,j}^{(\text{wccn})} = \boldsymbol{B}^T \boldsymbol{w}_{i,j} \,. \tag{15}$$

Scoring. Once session compensation has been performed, any scoring technique might be employed for authentication purposes. *Cosine similarity scoring* [9, 10] is a simple and efficient method used to estimate how close a (normalised) i-vector $\boldsymbol{w}_t$ extracted from a probe sample $\boldsymbol{\mathcal{O}}_t$ is to the i-vector $\boldsymbol{w}_i$ representing a client i:

$$h_{\text{cosine}}\left(\boldsymbol{w}_t, \boldsymbol{w}_i\right) = \frac{\boldsymbol{w}_t \cdot \boldsymbol{w}_i}{\|\boldsymbol{w}_t\| \|\boldsymbol{w}_i\|} \,. \tag{16}$$

Another technique commonly applied in the i-vector space is *probabilistic linear discriminant analysis* (PLDA) [37, 38, 39]. PLDA is a probabilistic framework that incorporates both between-class and within-class information and, therefore, performs session compensation. In addition, considering the authentication problem, this probabilistic approach allows the generation of LLR scores.

More formally, PLDA assumes that the j-th i-vector of client i is generated by:

$$\boldsymbol{w}_{i,j} = \boldsymbol{F}\boldsymbol{h}_i + \boldsymbol{G}\boldsymbol{k}_{i,j} + \boldsymbol{\epsilon}_{i,j} \,, \tag{17}$$

where $\boldsymbol{F}$ and $\boldsymbol{G}$ are the subspaces describing the between-class and within-class variations, respectively, $\boldsymbol{h}_i$ and $\boldsymbol{k}_{i,j}$ are the associated latent variables, which are assumed to be normally distributed $\mathcal{N}\left(0, \boldsymbol{I}\right)$, and $\boldsymbol{\epsilon}_{i,j}$ represents the residual noise, which is supposed to follow a Gaussian distribution $\mathcal{N}\left(0, \boldsymbol{\Sigma}_\epsilon\right)$.

The parameters $\boldsymbol{\Theta}_{\text{plda}} = \left\{\boldsymbol{F}, \boldsymbol{G}, \boldsymbol{\Sigma}_\epsilon\right\}$ of this model are learnt using an EM algorithm over a training set of i-vectors. Once the model has been trained, given an i-vector $\boldsymbol{w}_t$ extracted from a probe sample $\boldsymbol{\mathcal{O}}_t$ and an i-vector $\boldsymbol{w}_i$ representing a client i, authentication can be achieved by computing the LLR score:

$$h_{\text{plda}}\left(\boldsymbol{w}_t, \boldsymbol{w}_i\right) = \frac{p(\boldsymbol{w}_t, \boldsymbol{w}_i \mid \boldsymbol{\Theta})}{p(\boldsymbol{w}_t \mid \boldsymbol{\Theta}) p(\boldsymbol{w}_i \mid \boldsymbol{\Theta})} \,. \tag{18}$$

Here, $p(\boldsymbol{w}_t, \boldsymbol{w}_i \mid \boldsymbol{\Theta})$ is the log-likelihood that the i-vectors $\boldsymbol{w}_t$ and $\boldsymbol{w}_i$ share the same latent identity variable $\boldsymbol{h}_i$ and, hence, are coming from the same client, whereas $p(\boldsymbol{w}_t \mid \boldsymbol{\Theta})p(\boldsymbol{w}_i \mid \boldsymbol{\Theta})$ is the log-likelihood that the i-vectors $\boldsymbol{w}_t$ and $\boldsymbol{w}_i$ have different latent identity variables $\boldsymbol{h}_t$ and $\boldsymbol{h}_i$ and, therefore, are from different clients. For details on how to estimate these likelihoods and how to train the parameters $\boldsymbol{\Theta}_{\text{plda}}$, readers are referred to [37, 38, 39].

Finally, recent work [40] on speaker recognition at NIST SRE 2012[1] has shown that the duration mismatch between enrolment and test speech segments can tremendously affect the accuracy of the system. To cope with this variability, [40] proposed to truncate the speech signal into shorter segments. Then the i-vectors of the truncated versions together with the i-vectors of the original signals are used to train the PLDA. We also evaluate this technique on the MOBIO database (see Section 5.6).

3. Bi-modal and multi-algorithm authentication systems

Several fusion strategies are known in the literature [41]. They can be classified into three main categories:

Low-level fusion which is also known as *data fusion*, combines multiple sources of raw data to produce new raw data. The major problem of this fusion method comes with non-balanced dimensionalities of data from the multiple sources.

Intermediate-level fusion or *feature level fusion* combines various features that might come from several raw data sources or even from the same raw data. The drawback of feature-level fusion is that synchronisation between modalities [42] is required, which is not provided in the MOBIO database.

High-level fusion which is also called *decision level fusion*, *late fusion* or *score fusion*, combines decisions from several systems. This fusion strategy is very flexible and can be used for multi-modal (face and speaker) or multi-algorithm (for instance combining GMM and ISV) fusion. High-level fusion methods include majority voting methods, fuzzy logic based methods [43], and statistical methods.

In this work we choose the high-level fusion approach due to its ease of use for both multi-modal [8] and multi-algorithm [44, 45, 46] fusion.

[1] http://www.nist.gov/itl/iad/mig/sre12.cfm

3.1. Linear logistic regression

We take the well-known statistical *linear logistic regression* approach, which has been successfully employed for combining heterogeneous speaker and face authentication classifiers [44, 45, 46] and for bi-modal (face and speaker) authentication [8].

Linear logistic regression combines a set of Q classifiers using the sum rule. Let the probe $\boldsymbol{\mathcal{O}}_t$ be processed by Q classifiers, each of which produces an output score $h_q\left(\boldsymbol{\mathcal{O}}_t, \boldsymbol{s}_i\right)$. These scores are fused using a linear combination:

$$h_{\text{fusion}}\left(\boldsymbol{\mathcal{O}}_t, \boldsymbol{s}_i, \boldsymbol{\beta}\right) = \beta_0 + \sum_{q=1}^{Q} \beta_q h_q\left(\boldsymbol{\mathcal{O}}_t, \boldsymbol{s}_i\right), \tag{19}$$

where $\boldsymbol{\beta} = [\beta_0, \beta_1, ..., \beta_Q]$ are the fusion weights (also known as *regression coefficients*).

The coefficients $\boldsymbol{\beta}$ are computed by estimating the maximum likelihood of the logistic regression model on the scores of the development set. Let $\mathcal{X}_{\text{true}}$ be the set of true client access trials, i. e., the set of pairs $\boldsymbol{x} = \{\boldsymbol{\mathcal{O}}_t, \boldsymbol{s}_i\}$, where the identity of the test sample $\boldsymbol{\mathcal{O}}_t$ and of the client $\boldsymbol{s}_i$ is the same. Let furthermore $\mathcal{X}_{\text{imp}}$ be the set of impostor trials, i. e., the set of pairs $\boldsymbol{x} = \{\boldsymbol{\mathcal{O}}_t, \boldsymbol{s}_i\}$, where the identities of the test sample $\boldsymbol{\mathcal{O}}_t$ and of the client $\boldsymbol{s}_i$ are different. Let $\mathcal{X} = \mathcal{X}_{\text{true}} \cup \mathcal{X}_{\text{imp}}$. The objective function to maximise is:

$$L(\boldsymbol{\beta}) = -\sum_{\boldsymbol{x} \in \mathcal{X}} \log\left(1 + \exp\left(-y_{\boldsymbol{x}} h_{\text{fusion}}\left(\boldsymbol{x}, \boldsymbol{\beta}\right)\right)\right), \tag{20}$$

where:

$$y_{\boldsymbol{x}} = \begin{cases} +1, & \text{if } \boldsymbol{x} \in \mathcal{X}_{\text{true}} \\ -1, & \text{if } \boldsymbol{x} \in \mathcal{X}_{\text{imp}} \end{cases} \tag{21}$$

The maximum likelihood estimation procedure converges to a global minimum. In our work, this optimisation is done using the *conjugate-gradient* algorithm [47].

This approach performs best when the scores of the classifiers are statistically independent of each other. For this reason we measure the independence and, therewith, the complementary nature of our classifiers. We use the scatter plots (see Fig. 8) and the *relative common error* (RCE):

$$\text{RCE} = \text{CE} \times \max\left\{\frac{1}{\text{TE}_1}, \frac{1}{\text{TE}_2}, \ldots \frac{1}{\text{TE}_Q}\right\}, \tag{22}$$

where CE is the number of *common errors* between the Q classifiers and TE_q is the *total number of errors* of the q^{th} subsystem. The lower RCE is, the more independent the classifiers are.

In this work we evaluate the effectiveness of both bi-modal and multi-algorithm fusion. This leads to a number of different system combinations, which we outline in Fig. 2. The top row of Fig. 2 displays the three different bi-modal fusion systems, while the bottom row shows the two different multi-algorithm fusion systems and the bi-modal multi-algorithm fusion approach that we examine.

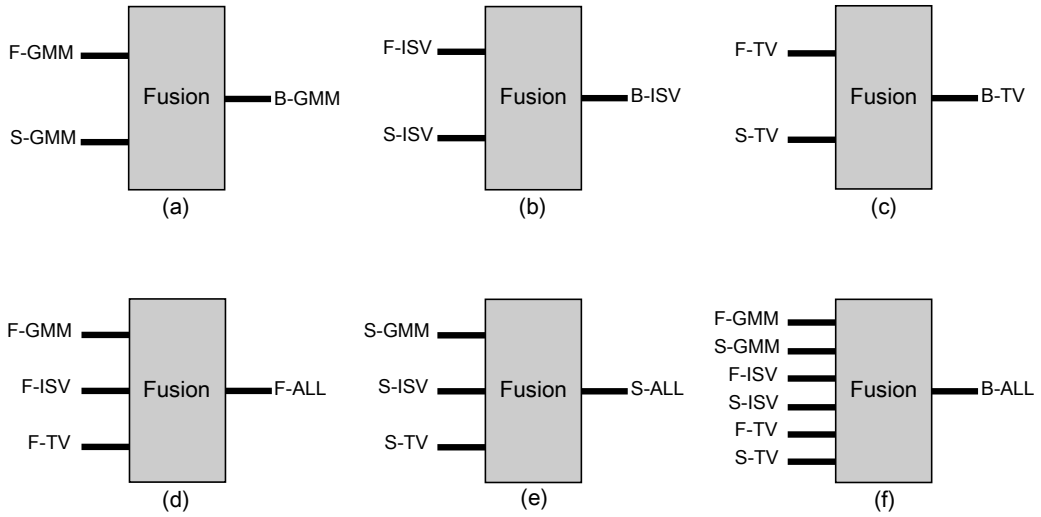


Figure 2: FUSION STRATEGIES. *This figure displays different fusion strategies used in this paper: (a) - (c) bi-modal fusion strategies, (d) - (e) multi-algorithm fusion strategies, (f) bi-modal multi-algorithm fusion.*

4. Database and protocols

The MOBIO database [2] is a unique bi-modal, face and speaker, database as it was captured almost exclusively using mobile phones. It consists of over 61 hours of audio-visual data of 150 people captured within twelve distinct sessions that are usually separated by several weeks. The users answered a set of questions, which varied in type, including:

1. *short response questions* (**p**) such as “what is your address”,

2. *free speech questions*, where the user speaks about any subject for approximately 10 seconds (**f**) or about 5 seconds (**r**), and

3. *pre-defined text* (**l**) that the user read out.

All of this data was captured on a mobile phone, except for the first session, where data was obtained using both a mobile phone and a laptop computer. One of the unique attributes of this database is that the acquisition device was held by the user, rather than being in a fixed position. As such, the microphone and camera are not fixed and used in an interactive and uncontrolled manner. This presents several challenges such as high variability of pose and illumination conditions, high variations in the quality of speech, and variability in terms of acoustics as well as illumination and background. Exemplary images of one identity are given in Fig. 3.

This challenging mobile phone database has been used to evaluate several face and speaker authentication systems [1] as well as bi-modal authentication systems [7, 8]. The database provides a well defined protocol, which was initially described for the full database in [4]. This protocol separates the clients of the database into three non-overlapping partitions for training, development (DEV) and evaluation (EVAL). The performance is measured in a gender-dependent manner (Female and Male, respectively). An overview of this initial protocol, which we refer to as **mobile-0**, is provided in Table 1. A limitation of this previously defined protocol is that only the lower quality biometric data acquired from the mobile phone was used, while the higher quality laptop data were ignored.

4.1. Evaluation protocols

In this work we extend the MOBIO protocol [2] and define three novel protocols that explore mismatched conditions by making use of the laptop data[2]. The mismatched conditions that we wish to investigate are the specific cases of enrolling a user with high quality biometric samples (for instance acquired from a laptop computer) and then compared, or tested, using lower quality biometric samples obtained using a mobile phone.

mobile-1 is identical to a **mobile-0**, except that it includes the laptop data in the training set. This ensures that the same training data is being used for

[2]The MOBIO database (videos, still images, eye locations and the four evaluation protocols) are available for free at `http://www.idiap.ch/dataset/mobio`

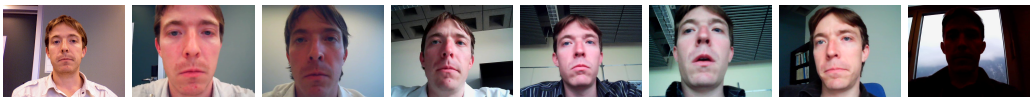


Figure 3: MOBIO DATABASE EXAMPLES. *This figure shows one image of the MOBIO database captured with a laptop on the left, and seven other images of the same identity captured with a mobile phone with significantly varying acquisition conditions.*

mobile and laptop evaluation (the next protocol that we present). It provides an additional $1,050$ training samples compared to **mobile-0**. Enrolment and testing is conducted using only mobile phone data. Please see Table 2 for more details.

laptop-1 contains the same training data as **mobile-1**, but enrolment is performed exclusively using laptop data, while testing is conducted exclusively with mobile phone data. See Table 3 for details on the kind of data used in this protocol.

laptop-mobile-1 also consists of the same training data as **mobile-1**. Here, enrolment is performed using both mobile and laptop data, while testing is still conducted exclusively on mobile phone data, see Table 4 for details.

4.2. Evaluation criteria

To measure the accuracy of the presented authentication systems, we use two different evaluation criteria that were previously defined in [2]. These measures are the *half total error rate* (HTER) and *detection error trade-off* (DET) plots.

The HTER is used to represent the performance of an authentication system on the unbiased evaluation partition as a single number. To compute the HTER, a threshold θ is defined on the development partition at the intersection point of the *false acceptance rate* (FAR) and the *false rejection rate* (FRR). The corresponding FAR (or FRR) value of the development partition at this threshold θ is known as the *equal error rate* (EER). The threshold is applied to the evaluation partition ($\mathcal{D}_{\text{EVAL}}$) to obtain the HTER:

$$\text{HTER} = \frac{\text{FAR}(\theta, \mathcal{D}_{\text{EVAL}}) + \text{FRR}(\theta, \mathcal{D}_{\text{EVAL}})}{2}, \tag{23}$$

which is the average of the FAR and the FRR at θ. Finally, we provide a complete overview of the system performances using a DET plot [48], which outlines the miss probability (FRR) versus the probability of false acceptance (FAR) on the evaluation set.

Table 1: THE **MOBILE-0** PROTOCOL. *This table gives an overview of the data used in the (original)* ***mobile-0*** *protocol of the MOBIO database.*

Set	Phase I			Phase II	Nb. videos/client	Nb. videos
	laptop data	mobile data	mobile data	mobile data		
	session-01	session-01	sessions 02-06	sessions 07-12		
	videos/client	videos/client	(videos/client)/sess.	(videos/client)/sess.		
Train	-	**5p**+10**f**+5**r**+1**l**	**5p**+10**f**+5**r**+1**l**	**5p**+5**f**+1**l**	192	9600
Enrol	-	**5p**	-	-	5	500
Test	-	-	10**f** + 5**r**	**5f**	105	10500

Table 2: THE **MOBILE-1** PROTOCOL. *This table gives an overview of the data used in the (novel)* ***mobile-1*** *protocol for the MOBIO database.*

Set	Phase I			Phase II	Nb. videos/client	Nb. videos
	laptop data	mobile data	mobile data	mobile data		
	session-01	session-01	sessions 02-06	sessions 07-12		
	videos/client	videos/client	(videos/client)/sess.	(videos/client)/sess.		
Train	**5p**+10**f**+5**r**+1**l**	**5p**+10**f**+5**r**+1**l**	**5p**+10**f**+5**r**+ 1**l**	**5p**+5**f**+1**l**	213	10650
Enrol	-	**5p**	-	-	5	500
Test	-	-	10**f** + 5**r**	**5f**	105	10500

Table 3: THE **LAPTOP-1** PROTOCOL. *This table gives an overview of the data used in the (novel)* ***laptop-1*** *protocol for the MOBIO database.*

Set	Phase I			Phase II	Nb. videos/client	Nb. videos
	laptop data	mobile data	mobile data	mobile data		
	session-01	session-01	sessions 02-06	sessions 07-12		
	videos/client	videos/client	(videos/client)/sess.	(videos/client)/sess.		
Train	**5p**+10**f**+5**r**+1**l**	**5p**+10**f**+5**r**+1**l**	**5p**+10**f**+5**r**+ 1**l**	**5p**+5**f**+1**l**	213	10650
Enrol	**5p**	-	-	-	5	500
Test	-	-	10**f** + 5**r**	**5f**	105	10500

Table 4: THE **LAPTOP-MOBILE-1** PROTOCOL. *This table gives an overview of the data used in the (novel)* ***laptop-mobile-1*** *protocol of the MOBIO database.*

Set	Phase I			Phase II	Nb. videos/client	Nb. videos
	laptop data	mobile data	mobile data	mobile data		
	session-01	session-01	sessions 02-06	sessions 07-12		
	videos/client	videos/client	(videos/client)/sess.	(videos/client)/sess.		
Train	**5p**+10**f**+5**r**+1**l**	**5p**+10**f**+5**r**+1**l**	**5p**+10**f**+5**r**+ 1**l**	**5p**+5**f**+1**l**	213	10650
Enrol	**5p**	**5p**	-	-	10	1000
Test	-	-	10**f** + 5**r**	**5f**	105	10500

5. Experimental Results

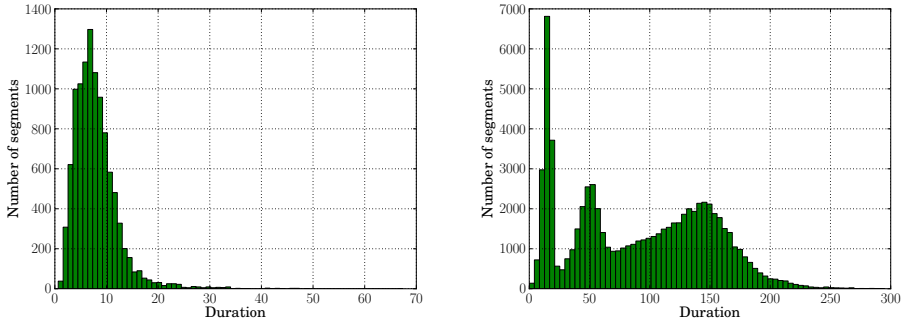
In this section, we evaluate the accuracy of the uni-modal and bi-modal authentication systems described in Sections 2 and 3, across the four protocols defined in Section 4. Global observations are first highlighted in Section 5.1. A detailed comparison between GMM, ISV and TV systems is presented in Section 5.2. Bi-modal and multi-algorithm experiments are examined in Sections 5.3 and 5.4, respectively, and the results are compared with other state-of-the-art face and speaker authentication fusion systems. The results obtained on the four protocols are presented and summarised in Section 5.5. Section 5.6 presents the results of gender-dependent systems and the use of extended training set. We present both the EER on the DEV set and the HTER on the EVAL set. Results for the best systems for each modality are highlighted in bold. We also distinguish the best uni-modal single algorithm systems by highlighting them in bold italics.

To make the comparison simple and the results reproducible we used the same parameters throughout the experiments and conducted all experiments with the open-source Bob toolbox[3] [49] to implement all of our systems. GMMs are composed of 512 Gaussian components and the UBMs are trained in the following manner: 25 iterations of k-means clustering are performed to initialise the UBM and then 100 iterations of maximum likelihood estimation are executed. For ISV, the rank n_U of the subspace $\boldsymbol{U}$ is set to 50 for speaker authentication system (**S-ISV**) and 160 for face authentication system (**F-ISV**), 10 iterations are performed to train the subspace $\boldsymbol{U}$. For TV, the rank n_T of the subspace $\boldsymbol{T}$ is set to 400, and 25 iterations are done for the subspace training. When using LDA with i-vectors the projection matrix $\boldsymbol{A}$ is limited to $n_{\text{lda}} = 200$ dimensions. For PLDA, the ranks n_F and n_G of the subspaces $\boldsymbol{F}$ and $\boldsymbol{G}$ are both set to 50. In addition, the cohort set for zt-normalisation is selected from the training data: two thirds are used for t-norm and the remaining third is used for z-norm. For t-models, we used one model per session (instead of one model per client), as in [4]. This copes with the limited number of clients in the cohort.

5.1. Global observations

Looking at the results provided in Tables 5, 6, 7, and 8, two general trends are emerging throughout the results. First, error rates on female clients are higher than

[3]Bob is a free signal processing and machine learning toolbox originally developed at Idiap Research Institute. The total variability system was incorporated especially for this paper. You can download Bob from: http://www.idiap.ch/software/bob



(a) Speech duration in MOBIO (b) Speech duration in NIST SRE 2012

Figure 4: MOBIO vs. NIST SRE 2012. *This figure compares the distributions of the speech duration (in seconds) over the probe files after applying VAD between the MOBIO database and the NIST SRE 2012 data.*

on male clients. This might be due to the fact that the training set contains more men than women. Second, comparing the results of face authentication (**Face**) and speaker authentication (**Speaker**) systems, it is obvious that error rates of **Face** systems are lower than error rates of **Speaker** systems. This is possibly caused by the fact that speech segments are relatively short. Indeed, the average duration of the MOBIO probes after VAD is 7.9 s (see their distribution in Fig. 4(a)), whereas the average duration of the probes after VAD in NIST SRE 2012 is 91.5 s (see Fig. 4(b)).

5.2. Comparison of the modelling techniques

To be comparable with previous work [12, 8], our analysis focuses on the results of the **mobile-0** experiments, which are summarised in Table 5. However, similar conclusions might be drawn from the experiments on the other protocols that are given in Tables 6, 7 and 8.

It can be seen that **F-ISV** and **S-ISV** (rows 2 and 7) outperform **F-GMM** and **S-GMM** (rows 1 and 6). This is also shown in [8]. Apparently, except for a few cases TV (rows 3, 4, 8, 9) provides superior performance to the standard GMM approach.

However, the comparison between ISV and TV is not so simple, although usually ISV provides similar or better performance. For **Speaker**, **S-TV** and **S-ISV** are comparable, even though **S-ISV** is slightly better. For **Face**, **F-ISV** is significantly better than **F-TV**. The DET curves plotted in Fig. 5 and Fig. 6 support

Table 5: PERFORMANCE SUMMARY ON **MOBILE-0**. *This table reports the EER (%) on* DEV *and* HTER *(%) on* EVAL *obtained with the* ***mobile-0*** *protocol using gender independent (GI) training.*

	Modelling technique	Scoring technique	Female		Male		
			DEV	EVAL	DEV	EVAL	
Face	F-GMM	linear scoring + zt-norm	8.20	17.92	8.49	10.81	1
	F-ISV	linear scoring + zt-norm	***5.41***	***10.63***	***3.45***	***6.54***	2
	F-TV	PLDA	16.60	19.34	10.12	11.94	3
		cosine scoring	10.95	16.16	5.55	8.99	4
	F-ALL	**logistic regression**	**5.03**	**11.62**	**2.77**	**6.06**	5
Speaker	S-GMM	linear scoring + zt-norm	17.94	17.68	13.41	12.12	6
	S-ISV	linear scoring + zt-norm	***12.22***	***16.23***	***10.40***	***10.36***	7
	S-TV	PLDA	12.59	17.36	11.31	11.11	8
		cosine scoring	15.93	23.58	12.66	12.87	9
	S-ALL	**logistic regression**	**9.21**	**14.65**	**7.31**	**7.89**	10
Bi-modal	B-GMM	linear scoring + zt-norm	4.50	12.32	3.69	4.80	11
	B-ISV	linear scoring + zt-norm	***2.01***	***7.16***	***1.59***	***2.42***	12
	B-TV	PLDA (speaker) + cosine (face)	4.29	9.93	2.29	3.77	13
	B-ALL	**logistic regression**	**1.43**	**6.30**	**0.92**	**1.89**	14

Table 6: PERFORMANCE SUMMARY ON **MOBILE-1**. *This table reports the EER (%) on* DEV *and* HTER *(%) on* EVAL *obtained with the* ***mobile-1*** *protocol using gender independent (GI) training.*

	Modelling technique	Scoring technique	Female		Male		
			DEV	EVAL	DEV	EVAL	
Face	F-GMM	linear scoring + zt-norm	8.24	18.13	8.41	10.79	1
	F-ISV	linear scoring + zt-norm	***5.61***	***11.25***	***3.41***	***6.46***	2
	F-TV	PLDA	16.63	19.23	9.71	11.57	3
		cosine scoring	12.32	16.25	6.03	10.09	4
	F-ALL	**logistic regression**	**5.67**	**11.85**	**2.90**	**6.27**	5
Speaker	S-GMM	linear scoring + zt-norm	17.73	17.73	13.21	12.05	6
	S-ISV	linear scoring + zt-norm	***11.43***	***16.02***	***10.16***	***10.35***	7
	S-TV	PLDA	13.86	18.18	12.86	10.86	8
		cosine scoring	24.02	29.53	11.98	12.78	9
	S-ALL	**logistic regression**	**9.94**	**14.43**	**7.38**	**7.68**	10
Bi-modal	B-GMM	linear scoring + zt-norm	4.23	12.35	3.53	4.68	11
	B-ISV	linear scoring + zt-norm	***2.17***	***7.61***	***1.55***	***2.43***	12
	B-TV	PLDA (speaker) + cosine (face)	5.03	9.43	3.97	4.15	13
	B-ALL	**logistic regression**	**1.64**	**6.32**	**0.75**	**2.06**	14

Table 7: PERFORMANCE SUMMARY ON **LAPTOP-1**. *This table reports the EER (%) on* DEV *and* HTER *(%) on* EVAL *obtained with the* ***laptop-1*** *protocol using gender independent (GI) training.*

	Modelling technique	Scoring technique	Female		Male		
			DEV	EVAL	DEV	EVAL	
Face	F-GMM	linear scoring + zt-norm	18.78	20.79	13.02	16.59	1
	F-ISV	linear scoring + zt-norm	***12.65***	***12.91***	***6.83***	***9.55***	2
	F-TV	PLDA	19.52	21.84	11.74	15.77	3
		cosine scoring	18.62	16.42	9.01	12.24	4
	F-ALL	**logistic regression**	**11.86**	**11.73**	**5.47**	**8.87**	5
Speaker	S-GMM	linear scoring + zt-norm	19.05	19.80	16.39	16.06	6
	S-ISV	linear scoring + zt-norm	13.00	***18.43***	***12.78***	13.54	7
	S-TV	PLDA	***12.70***	20.14	13.61	***11.89***	8
		cosine scoring	18.73	25.27	14.72	14.71	9
	S-ALL	**logistic regression**	**9.36**	**16.24**	**8.85**	**9.04**	10
Bi-modal	B-GMM	linear scoring + zt-norm	8.89	12.16	6.07	7.83	11
	B-ISV	linear scoring + zt-norm	***3.60***	***7.47***	***3.03***	***4.52***	12
	B-TV	PLDA (speaker) + cosine (face)	7.41	10.46	4.88	5.61	13
	B-ALL	**logistic regression**	**2.91**	**6.83**	**1.82**	**3.37**	14

Table 8: PERFORMANCE SUMMARY ON **LAPTOP-MOBILE-1**. *This table reports the EER (%) on* DEV *and* HTER *(%) on* EVAL *obtained with the* ***laptop-mobile-1*** *protocol using gender independent (GI) training.*

	Modelling technique	Scoring technique	Female		Male		
			DEV	EVAL	DEV	EVAL	
Face	F-GMM	linear scoring + zt-norm	8.15	17.46	7.65	10.43	1
	F-ISV	linear scoring + zt-norm	***5.24***	***10.44***	***3.21***	***5.99***	2
	F-TV	PLDA	14.97	17.55	8.06	11.45	3
		cosine scoring	10.89	14.26	5.75	8.67	4
	F-ALL	**logistic regression**	**5.24**	**10.32**	**2.34**	**5.54**	5
Speaker	S-GMM	linear scoring + zt-norm	15.04	15.98	11.42	10.55	6
	S-ISV	linear scoring + zt-norm	***9.31***	***15.18***	***8.17***	***9.40***	7
	S-TV	PLDA	10.95	15.54	10.40	***9.40***	8
		cosine scoring	19.89	26.31	11.03	12.27	9
	S-ALL	**logistic regression**	**7.30**	**12.48**	**5.68**	**6.39**	10
Bi-modal	B-GMM	linear scoring + zt-norm	3.08	9.69	2.10	4.24	11
	B-ISV	linear scoring + zt-norm	***1.22***	***6.37***	***1.11***	***2.26***	12
	B-TV	PLDA (speaker) + cosine (face)	4.07	7.25	2.14	3.41	13
	B-ALL	**logistic regression**	**1.11**	**6.32**	**0.64**	**1.77**	14

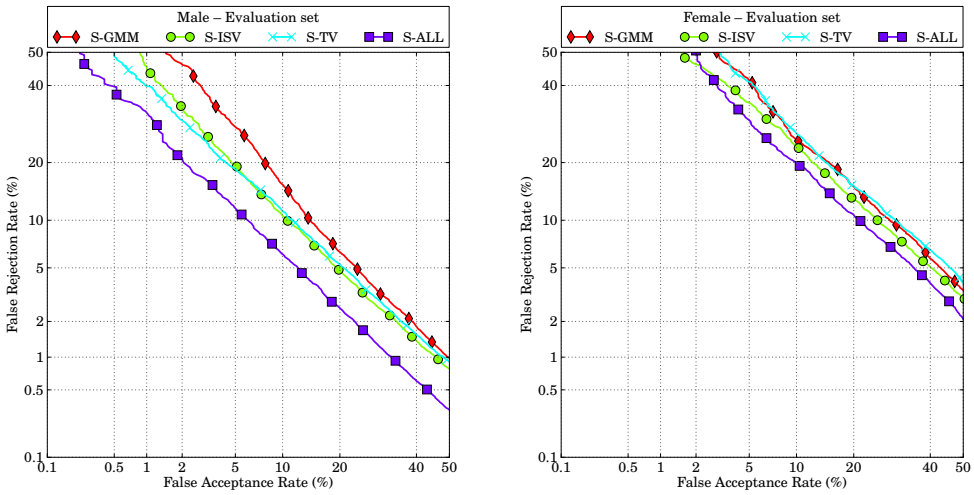


Figure 5: DET CURVES FOR **SPEAKER**. *This figure shows performances of GMM, ISV, TV and the multi-algorithm fusion on **mobile-0** for both Male and Female.*

this observation. In these curves we can see some intersection between the ISV and TV curves and between the GMM and TV curves. These curves also show that the systems are well calibrated (close to straight lines with angles close to $45\,°$) especially the ISV and TV systems.

The observation that ISV is at least as good as TV for speaker authentication is strange because TV is considered to be a state-of-the-art speaker authentication approach [9]. We believe that TV is limited by a lack of data needed to train the several steps such as: learning the TV matrix, whitening, LDA, WCCN and PLDA. For this reason we explore the use of additional training data in Section 5.6.

5.3. Bi-modal authentication

The bi-modal ISV system (**B-ISV**) outperforms both the **B-GMM** and **B-TV** systems. In Table 5 (rows 11, 12, and 13) and in the DET curves in Fig. 7 it can be seen that **B-ISV** clearly outperforms **B-TV**. The error rates drop significantly for all bi-modal systems. For example, on the Female **mobile-0** protocol the HTER of the ISV system drops from 10.6% (**F-ISV**) and 16.2% (**S-ISV**) to 7.2% (**B-ISV**), a relative performance gain of 33% compared to the best uni-modal system. The results on the Male **mobile-0** protocol are even more impressive with the HTER of the ISV system dropping from 6.5% (**F-ISV**) and 10.4% (**S-ISV**) to

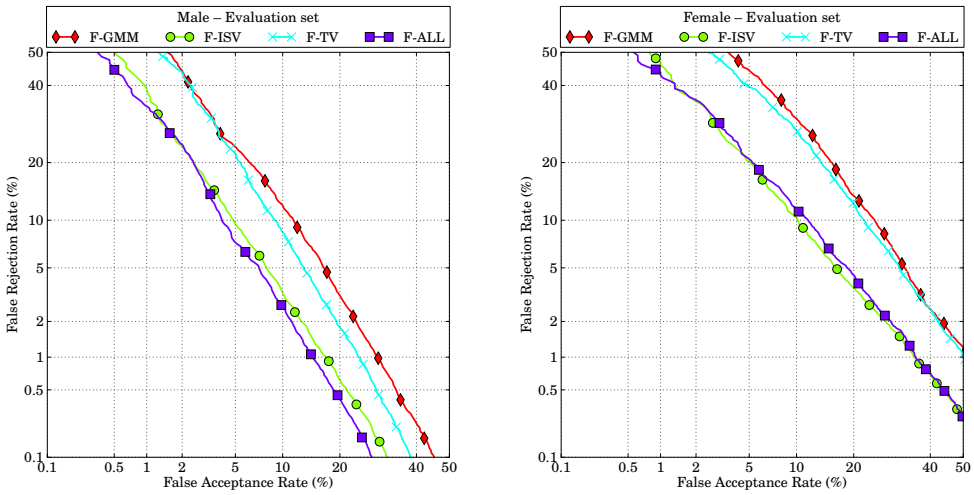


Figure 6: DET CURVES FOR **FACE**. *This figure displays performances of GMM, ISV, TV and the multi-algorithm fusion on **mobile-0** for both Male and Female.*

2.4% (**B-ISV**), a relative performance gain of 63% compared to the best uni-modal system. This improvement can be explained by the fact that image and audio modalities are complementary: when **Face** fails to take the right decision because of image variability (illumination, head pose, etc.), **Speaker** is available to rescue, and *vice versa*.

5.4. Multi-algorithm fusion

The fusion of multiple algorithms (GMM, ISV and TV) consistently outperforms single systems, as shown in Table 5 (rows 5, 10 and 14). For example, the HTER of the **Speaker** system on the Male **mobile-0** protocol drops from 10.4% (for ISV) to 7.9%, which corresponds to a relative improvement of 24%. The impact of the multi-algorithm fusion is higher for **Speaker** than **Face** because **Speaker** obtains a relative improvement of on average 19% compared to 3% for **Face** (the average is taken across all four protocols). We attribute this larger gain in performance for **Speaker** to the fact that TV has comparable results with ISV for **Speaker** but not for **Face**; TV performs much worse than ISV for **Face**. Finally, we note that the best bi-modal multi-algorithm fusion (**B-ALL**) outperforms the best uni-modal **Face** (**F-ALL**) and **Speaker** (**S-ALL**) systems with a relative improvement of up to 69% and 76%, respectively (for Male trials).

To explore the reason for the performance gains from multi-algorithm fusion

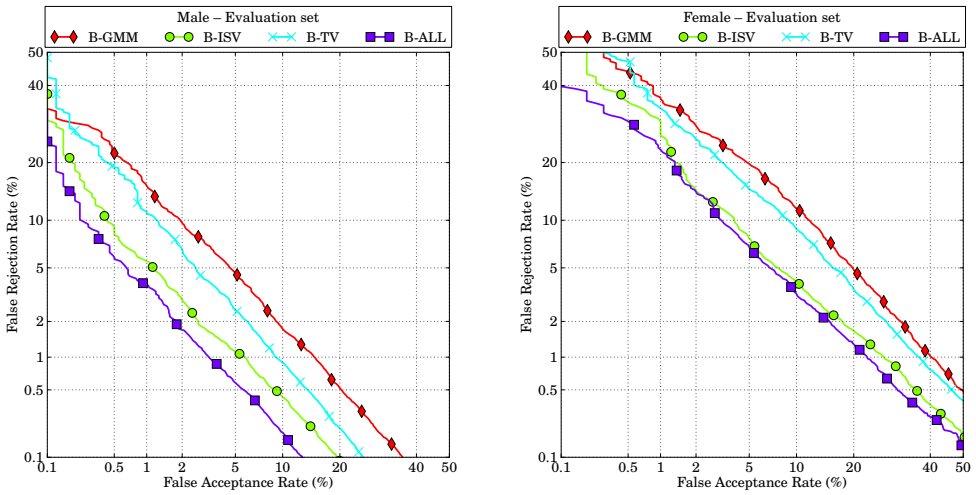
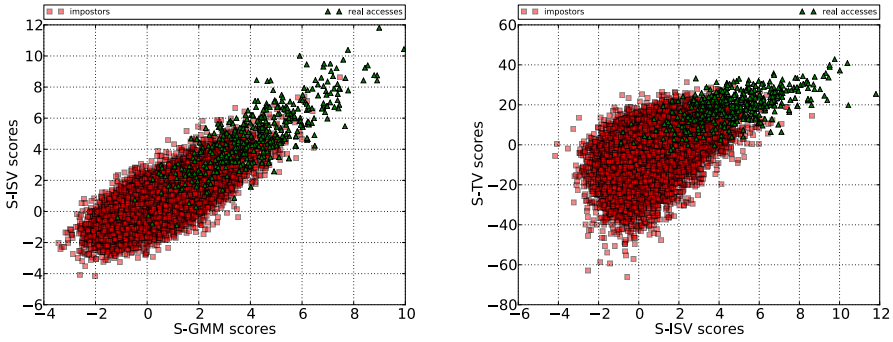


Figure 7: DET CURVES FOR **BI-MODAL**. *This figure shows performances of GMM, ISV, TV and the multi-algorithm fusion on **mobile-0** for both Male and Female.*

we examine scatter plots of the scores for the ISV, GMM and TV systems. Fig. 8 shows scatter plots that relate the ISV scores to GMM scores and TV scores to ISV scores for the **Speaker** system on the Male **mobile-0** protocol. The scatter plots indicate that fusing TV and ISV scores is the best strategy since the overlap between impostor and real client access classes is lower than for ISV and GMM. The small overlap can be explained by the fact that the scoring methods used for TV are significantly different from the ones used for ISV and GMM. This is supported by the observation that ISV and GMM scores are more correlated (linear distribution of the points) than TV and ISV scores (more wide-spread distribution).

In Table 9 we present the *relative common error* (RCE) of all of the multi-algorithm fusion systems. Apparently, ISV and TV have the lowest percentage of common errors: $\mathrm{RCE} = 31.6\%$, followed by TV and GMM with: $\mathrm{RCE} = 35.4\%$, while ISV and GMM have the highest common error: $\mathrm{RCE} = 54.7\%$. This result confirms that TV is the most helpful system for multi-algorithm fusion. Another finding of this table is that the fusion of the three systems is better ($\mathrm{RCE} = 21.7\%$) than the fusion of any two systems. It can also be seen that having a low percentage of relative common errors leads to an improved HTER, see row 3 of Table 9. Similar finding are obtained for **Face** and the bi-modal fusion systems as given in Table 9, though the improvement is not as significant as for **Speaker** since TV for



(a) ISV vs. GMM (b) TV vs. ISV

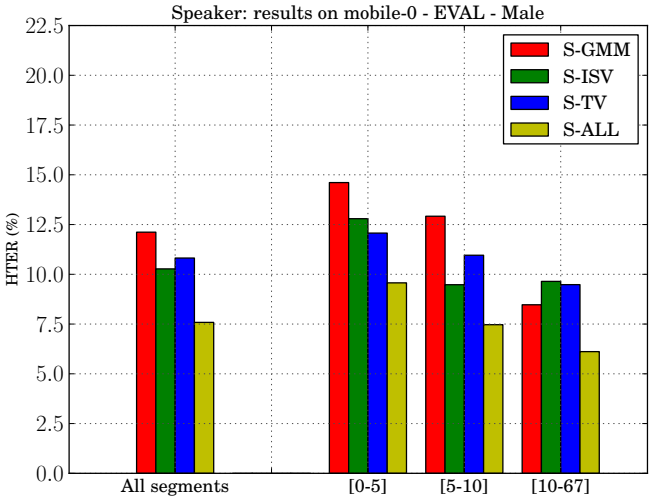
Figure 8: SCORE SCATTER PLOTS. *This figure displays scatter plots of scores obtained with each two* ***Speaker*** *systems.*

Table 9: MULTI-ALGORITHM FUSION. *This table displays common errors (*CE*), relative common errors (*RCE*) and half total error rates (*HTER*) on the Male* ***mobile-0*** *protocol.*

	Measure	GMM & ISV	GMM & TV	ISV & TV	GMM & ISV & TV	
Speaker	CE	9994	4827	4676	3253	1
	RCE (%)	65.05	44.25	***42.18***	**32.39**	2
	HTER (%)	10.17	9.36	***7.90***	**7.89**	3
Face	CE	5309	4371	2675	2385	4
	RCE (%)	79.95	45.56	***40.29***	**35.92**	5
	HTER (%)	6.47	7.72	***6.20***	**6.06**	6
Bi-modal	CE	2899	1509	1001	861	7
	RCE (%)	73.13	40.55	***26.90***	**23.14**	8
	HTER (%)	2.36	2.76	***1.99***	**1.89**	9

Face is not as good as for **Speaker**.

To better understand why multi-algorithm fusion significantly improves the results for **Speaker**, we group the audio probe files into three clusters depending on their duration as seen in Table 9(b). Fig. 9(a) displays the HTER of each of the groups. Although **S-ISV** is the best average system, Fig. 9(a) shows that **S-TV** is better for short duration (< 5 s) segments. Interestingly, **S-GMM** is performing better on relatively long duration (> 10 s) segments (this might be due to threshold tuning on DEV). Hence, **S-TV** performs better on the 22.6% audio samples that are less than 5s, while **S-GMM** leads on the 26.6% of audio samples longer 10s. We believe that these two observations are the major reasons for multi-algorithm fusion providing a significant boost in performance for **Speaker**.



(a) HTER for different durations

Speech duration (s)	Percentage of segments
$[0-5]$	22.6%
$[5-10]$	50.8%
$[10-67]$	26.6%

(b) Duration intervals

Figure 9: DURATION IMPACT ON MODELLING ALGORITHMS. *This figure shows performances of the **Speaker** algorithms on different speech durations and the distributions of speech duration of the Male **mobile-0** protocol.*

Comparison with existing work. By performing bi-modal multi-algorithm fusion we are able to develop a state-of-the-art bi-modal, face and speaker, authentication system. In Fig. 10 we compare our system against the results obtained in [2, 8] on the same **mobile-0** protocol. This figure shows that we obtain a relative improvement of 35% and 27% on Female and Male, respectively, compared to the results of [8].

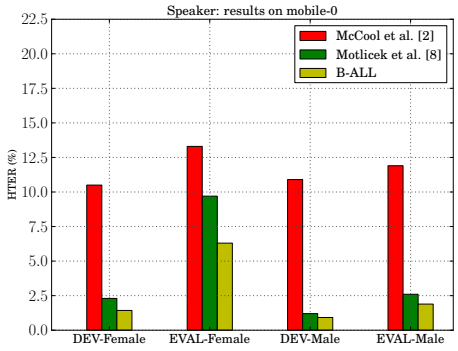
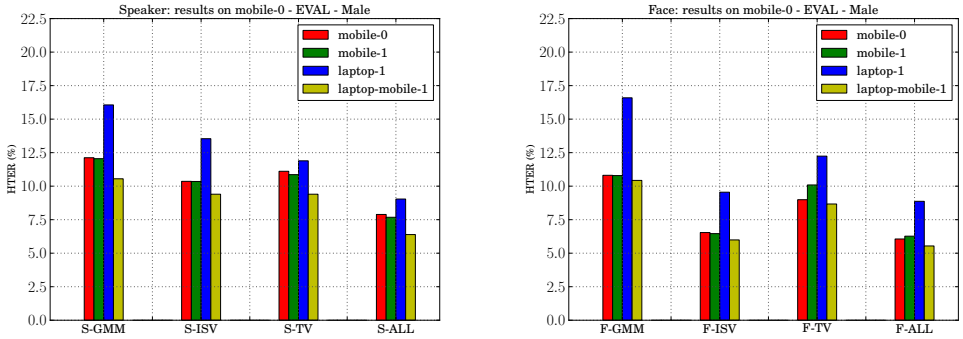


Figure 10: COMPARISON WITH EXISTING WORK. *This figure displays* HTER *of* ***B-ALL****, [2] and [8] on the* ***mobile-0*** *protocol.*

5.5. Comparison of the protocols

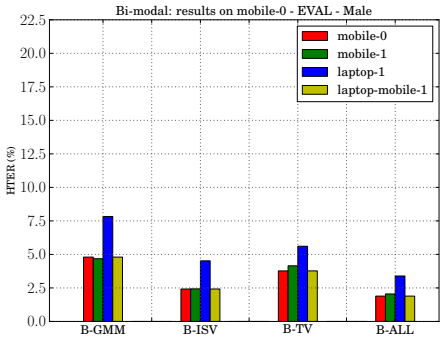
Fig. 11 displays the impact on enrolment condition mismatch on speaker, face and bi-modal authentication. It shows that GMM and ISV are significantly affected by changing the enrolment conditions (between **mobile-1** and **laptop-1**), whereas TV seems to be more robust to these changes. Indeed, for Male clients the **F-GMM**, **S-GMM** and **B-GMM** systems have a relative performance degradation of 54%, 33% and 67%, respectively, while **F-ISV**, **S-ISV** and **B-ISV** loose 47%, 30% and 86%, respectively. By contrast the **F-TV**, **S-TV** and **B-TV** systems have a relative performance decrease of only 9%, 21% and 35%, respectively. Another interesting result is that the degradation of **Face** systems is higher than on **Speaker** systems. This shows that **Face** is more affected by condition mismatch of high versus low image quality (see Fig. 3) and is an issue that deserves further investigation.

On the other hand, adding enrolment data as done in the **laptop-mobile-1** protocol improves authentication in all systems of **Speaker** and **Face**, even though the additional laptop data is quite different to the mobile phone data. In Figs. 11(a)



(a) **Speaker** authentication

(b) **Face** authentication



(c) **Bi-modal** authentication

Figure 11: BEHAVIOUR ON THE DIFFERENT PROTOCOLS. *This figure shows* HTER *of the modelling techniques across the different protocols.*

and 11(b) the **laptop-mobile-1** protocol outperforms the other protocols throughout. Interestingly, the bi-modal systems are reaching a performance plateau, since the results on the **laptop-mobile-1** are comparable to the results on **mobile-0** and **mobile-1** protocols (see Fig. 11(c)).

5.6. Additional training data

In this section we examine two issues that relate to the limited amount of training data available with MOBIO. These issues are: i) the use of gender-independent models versus gender-dependent models and ii) the performance difference between ISV and TV. For speaker authentication, gender-dependent models are usually derived as they provide improved performance [31, 26]. However, for MO-

Table 10: GENDER-INDEPENDENT VS. GENDER-DEPENDENT TRAINING. *This table reports EER (%) on* DEV *and* HTER *(%) on* EVAL *on protocol* ***mobile-0*** *using gender-independent (GI) or gender dependent (GD) training.*

	Modelling technique	Scoring technique	Female				
			GI		GD		
			DEV	EVAL	DEV	EVAL	
Face	F-ISV	linear scoring + zt-norm	***5.41***	***10.63***	6.52	12.63	1
	F-TV	cosine scoring	10.95	16.16	10.53	16.22	3
	F-ALL	**logistic regression**	**5.03**	**11.62**	**4.75**	**9.90**	3
Speaker	S-ISV	linear scoring + zt-norm	***12.22***	16.23	12.55	***13.50***	4
	S-TV	PLDA	12.59	17.36	20.83	22.04	5
	S-ALL	**logistic regression**	**9.21**	**14.65**	**11.53**	**13.70**	6
Bi-modal	B-ISV	linear scoring + zt-norm	***2.01***	***7.16***	3.32	8.48	7
	B-TV	PLDA (speaker) + cosine (face)	4.29	9.93	6.61	11.39	8
	B-ALL	**logistic regression**	**1.43**	**6.30**	**2.28**	**7.92**	9

BIO we found that with limited data, gender-independent models provided similar or better performance, see Table 10. In addition, current state-of-the-art speaker authentication systems use TV [9], but in our experiments TV does not provide better results than ISV. We attribute this lack of performance to the limited amount of data available with MOBIO. To explore both of these issues we use an external database to train gender-dependent models including: UBM, subspaces (subspace $\boldsymbol{U}$ for ISV and subspace $\boldsymbol{T}$ for TV), whitening, LDA and WCCN.

We conducted an experiment with additional audio data to train a gender-dependent Female model. The external data were collected from the Voxforge speech dataset.[4] The new audio training set contains 78 female clients, 65 of which belong to Voxforge.

The impact of extending the training data is examined in Table 11 for TV modelling. It can be seen that using more data reduces the EER and HTER from 20.3% and 21.1% to 14.8% and 18.7% respectively. It also shows that the main improvement is obtained on PLDA (row 1) with a relative performance gain of 18% (The HTER drops from 41.5% to 33.8%). Table 11 also shows that the duration variability [40] described in Section 2.2.3 is also helpful in the case of additional training data. In contrast to the performance gains of **S-TV**, **S-ISV** is not improved by adding external training data as can be seen in Fig. 12.

[4] http://www.voxforge.org This dataset was selected because of its similarity with MOBIO (short duration segments, different sessions of the same client, etc.). However, since the dataset is mainly dedicated to speech recognition (ASR) and is updated on the fly, it tolerates errors especially in the client identities, which could limit its usability for speaker authentication.

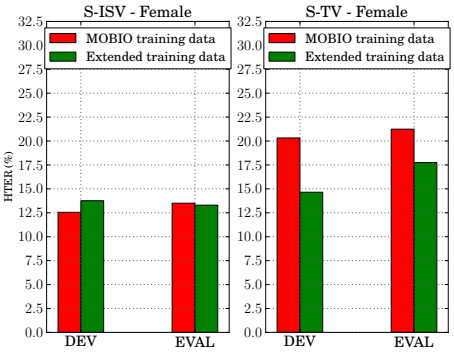


Figure 12: EXTENDED TRAINING DATA. *This figure shows the impact of extended training data on the Female* ***mobile-0*** *protocol using gender-dependent (GD) training.*

Table 11: EXTENDED TRAINING SETS IN I-VECTORS PREPROCESSING. *This table provides EER (%) on* DEV *and* HTER *(%) on* EVAL *for different combinations of i-vector preprocessing steps on the Female* ***mobile-0*** *protocol using gender-dependent (GD) training.*

Scoring method	i-vector processsing	MOBIO training set		Extended training set		
		DEV	EVAL	DEV	EVAL	
PLDA	-	40.48	41.47	33.61	33.84	1
	Whitening	22.07	21.23	16.93	20.09	2
	Whitening +Lnorm	25.36	23.50	18.16	20.64	3
	Whitening + Lnorm + LDA	22.17	23.01	16.24	19.20	4
	Whitening + Lnorm + LDA + WCCN	***20.32***	***21.24***	14.77	18.66	5
	Whitening + Lnorm + LDA + WCCN + DV	20.83	22.04	***14.65***	***17.75***	6

6. Conclusions

In this paper, we studied the problem of face, speaker and bi-modal authentication in the challenging mobile environment. The study was carried out on the MOBIO database, for which we proposed three new protocols. One of these new protocols, **laptop-1**, presents a significant challenge for both speaker and face authentication as there is a significant mismatch between enrolment and test conditions. Empirically, we found that both face and speaker authentication are adversely affected by this condition mismatch with the relative performance of the best uni-modal and uni-algorithm systems **F-ISV** and **S-ISV** degrading by 47% and 37%, respectively, for Male trials. The impact of this condition mismatch was extended to the bi-modal system, whose relative performance degraded by as much as 80% for Male trials.

We also examined several aspects of bi-modal and multi-algorithm fusion in the challenging mobile environment. We developed a state-of-the-art bi-modal multi-algorithm fusion system (**B-ALL**) that outperformed the state-of-the-art system of [8] obtaining a relative performance improvement of 35% and 27% on Female and Male trials, respectively. We found that multi-algorithm fusion provides a consistent performance improvement, particularly for the audio modality with average performance improvements of 3% for face authentication and 19% for speaker authentication across Male and Female trials for all of the protocols. In addition to this we showed empirically that ISV consistently outperforms not only GMM, but also TV with a limited amount of training data for both face and speaker authentication. We further explored this performance difference and found that TV provided improved performance for short utterances (less than 5 s) and ISV provided better performance for medium length utterances (between 5 s and 10 s).

7. Acknowledgement

The research leading to these results has received funding from the Swiss National Science Foundation under the LOBI project, from the European Community's Seventh Framework Programme (FP7) under grant agreements 238803 (BBfor2) and 284989 (BEAT), and from NICTA. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

[1] S. Marcel, et al., On the results of the first mobile biometry (MOBIO) face and speaker verification evaluation, in: Intl. Conf. on Pattern Recognition (ICPR), 2010.

[2] C. McCool, et al., Bi-modal person recognition on a mobile phone: using mobile phone data, in: IEEE Intl. Conf. on Multimedia and Expo (ICME), Workshop on Hot Topics in Mobile Multimedia, 2012.

[3] S. Mau, S. Chen, C. Sanderson, B. Lovell, Video face matching using subset selection and clustering of probabilistic multi-region histograms, in: Intl. Conf. of Image and Vision Computing (ICIVC), 2010.

[4] R. Wallace, M. McLaren, C. McCool, S. Marcel, Inter-session variability modelling and joint factor analysis for face authentication, in: Intl. Joint Conf. on Biometrics (IJCB), 2011.

[5] L. Perera, R. Lopez, J. Flores, Speaker verification in different database scenarios, Computación y Sistemas 15 (2011) 17–26.

[6] A. Roy, M. Magimai-Doss, S. Marcel, A fast parts-based approach to speaker verication using boosted slice classiers, IEEE Trans. on Information Forensics and Security 7 (2012) 241–254.

[7] L. Shen, N. Zheng, S. Zheng, W. Li, Secure mobile services by face and speech based personal authentication, in: IEEE Intl. Conf. on Intelligent Computing and Intelligent Systems (ICIS), 2010, pp. 97–100.

[8] P. Motlicek, L. E. Shafey, R. Wallace, C. McCool, S. Marcel, Bi-modal authentication in mobile environments using session variability modelling, in: 21st Intl. Conf. on Pattern Recognition (ICPR), 2012.

[9] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification, IEEE Trans. on Audio, Speech, and Language Processing 19 (2011) 788–798.

[10] R. Wallace, M. McLaren, Total variability modelling for face verification, IET Biometrics (2012) 188–199.

[11] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, Speaker verification using adapted Gaussian mixture models, Digital Signal Processing 10 (2000) 19–41.

[12] C. McCool, R. Wallace, M. McLaren, L. E. Shafey, S. Marcel, Session variability modelling for face authentication, IET Biometrics (2013) (To appear).

[13] C. Sanderson, K. K. Paliwal, Fast features for face authentication under illumination direction changes, Pattern Recognition Letters 24 (14) (2003) 2409–2419.

[14] S. Lucey, T. Chen, A GMM parts based face representation for improved verification through relevance adaptation, in: IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR), Vol. 2, 2004, pp. 855–861.

[15] F. Cardinaux, C. Sanderson, S. Bengio, User authentication via adapted statistical models of face images, IEEE Trans. on Signal Processing 54 (2006) 361–373.

[16] R. Wallace, M. McLaren, C. McCool, S. Marcel, Cross-pollination of normalisation techniques from speaker to face authentication using Gaussian mixture models, IEEE Trans. on Information Forensics and Security 7 (2) (2012) 553–562.

[17] X. Tan, B. Triggs, Enhanced local texture feature sets for face recognition under difficult lighting conditions, IEEE Trans. on Image Processing 19 (6) (2010) 1635–1650.

[18] A. Adami, et al., Qualcomm-ICSI-OGI features for ASR, in: Intl. Conf. on Spoken Language Processing (ICSLP), 2002, pp. 4–7.

[19] E. Scheirer, M. Slaney, Construction and evaluation of a robust multifeature speech/music discriminator, in: IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 2, 1997, pp. 1331–1334.

[20] D. Reynolds, R. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models, IEEE Trans. on Speech and Audio Processing 3 (1) (1995) 72–83.

[21] F. Cardinaux, C. Sanderson, S. Marcel, Comparison of MLP and GMM classifiers for face verification on XM2VTS, in: 4th Intl. Conf. on Audio- and Video-based Biometric Person Authentication (AVBPA), 2003, pp. 911–920.

[22] R. Vogt, S. Sridharan, Explicit modelling of session variability for speaker verification, Computer Speech & Language 22 (1) (2008) 17–38.

[23] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, series B 39 (1) (1977) 1–38.

[24] O. Glembek, L. Burget, N. Dehak, N. Brümmer, P. Kenny, Comparison of scoring methods used in speaker recognition with joint factor analysis, in: IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2009, pp. 4057–4060.

[25] R. Auckenthaler, M. Carey, H. Lloyd-Thomas, Score normalization for text-independent speaker verification systems, Digital Signal Processing 10 (1) (2000) 42–54.

[26] E. Khoury, L. El Shafey, S. Marcel, The Idiap speaker recognition evaluation system at NIST SRE 2012, in: NIST Speaker Recognition Workshop, 2012.

[27] P. Kenny, G. Boulianne, P. Ouellet, P. Dumouchel, Joint factor analysis versus eigenchannels in speaker recognition, IEEE Trans. on Audio, Speech, and Language Processing 15 (4) (2007) 1435–1447.

[28] M. McLaren, R. Vogt, B. Baker, S. Sridharan, A comparison of session variability compensation approaches for speaker verification, IEEE Trans. on Information Forensics and Security 5 (4) (2010) 802 –809.

[29] N. Dehak, Discriminative and generative approaches for long- and short-term speaker characteristics modeling: application to speaker verification, Ph.D. thesis (2009).

[30] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, P. Dumouchel, Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification, in: Interspeech, 2009, pp. 1559–1562.

[31] N. Brümmer, et al., ABC system description for NIST SRE 2010, in: NIST Speaker Recognition Workshop, 2010, pp. 1–20.

[32] P. Kenny, G. Boulianne, P. Dumouchel, Eigenvoice modeling with sparse training data, IEEE Trans. on Speech and Audio Processing 13 (3) (2005) 345–354.

[33] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka, N. Brümmer, Discriminatively trained probabilistic linear discriminant analysis for speaker verification, in: IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2011, pp. 4832–4835.

[34] D. Garcia-Romero, C. Espy-Wilson, Analysis of i-vector length normalization in speaker recognition systems, in: Interspeech, 2011, pp. 249–252.

[35] R. A. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics 19 (1936) 179–188.

[36] A. Hatch, S. Kajarekar, A. Stolcke, Within-class covariance normalization for SVM-based speaker recognition, in: 9th Intl. Conf. on Spoken Language Processing (ICSLP), 2006, pp. 1471–1474.

[37] S. J. D. Prince, J. H. Elder, Probabilistic linear discriminant analysis for inferences about identity, in: IEEE Intl. Conf. on Computer Vision (ICCV), 2007, pp. 1–8.

[38] P. Li, Y. Fu, U. Mohammed, J. H. Elder, S. J. D. Prince, Probabilistic models for inference about identity, IEEE Trans. in Pattern Analysis and Machine Intelligence 34 (2012) 144 – 157.

[39] L. E. Shafey, C. McCool, R. Wallace, S. Marcel, A scalable formulation of probabilistic linear discriminant analysis, IEEE Trans. in Pattern Analysis and Machine Intelligence (2013) (To appear).

[40] T. Hasan, R. Saeidi, J. H. L. Hansen, D. A. van Leeuwen, Duration mismatch compensation for i-vector based speaker recognition systems, in: IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2013.

[41] B. V. Dasarathy, Decision fusion, IEEE Computer Society Press, 1994.

[42] D. Shah, K. Han, S. Narayanan, A low-complexity dynamic face-voice feature fusion approach to multimodal person recognition, in: 11th IEEE Intl. Symposium on Multimedia, 2009, pp. 24–31.

[43] C. W. Lau, B. Ma, H. M. Meng, Y. S. Moon, Y. Yam, Fuzzy logic decision fusion in a multi-modal biometric system, in: Intl. Conf. on Spoken Language Processing (ICSLP), 2004.

[44] S. Pigeon, P. Druyts, P. Verlinde, Applying logistic regression to the fusion of the NIST'99 1-speaker submissions, Digital Signal Processing 10 (1–3) (2000) 237–248.

[45] N. Brümmer, et al., Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006, IEEE Trans. on Speech, Audio and Language Processing 15 (7) (2007) 2072–2084.

[46] C. McCool, S. Marcel, Parts-based face verification using local frequency bands, in: IEEE/IAPR International Conference on Biometrics, 2009.

[47] T. P. Minka, Algorithms for maximum-likelihood logistic regression, Tech. Rep. 758, CMU Statistics Department (2001).

[48] A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, The DET curve in assessment of detection task performance, in: 5th Eur. Conf. on Speech Communication and Technology (EUROSPEECH), Vol. 4, 1997, pp. 1895–1898.

[49] A. Anjos, L. E. Shafey, R. Wallace, M. Günther, C. McCool, S. Marcel, Bob: a free signal processing and machine learning toolbox for researchers, in: 20th ACM Intl. Conf. on Multimedia (ACMMM), 2012.