Open access • Proceedings Article • DOI:10.1109/ICMEW.2012.116

Bi-Modal Person Recognition on a Mobile Phone: Using Mobile Phone Data

Chris McCool, Sébastien Marcel, Abdenour Hadid, Matti Pietikäinen ...+10 more authors

Institutions: Idiap Research Institute, University of Oulu, Brno University of Technology, University of Surrey ...+2 more institutions

Published on: 09 Jul 2012 - International Conference on Multimedia and Expo

Topics: Mobile phone, Speaker recognition, Mobile device and Facial recognition system

Bi-Modal Person Recognition on a Mobile Phone: using mobile phone data

Chris Mccool, Sébastien Marcel, Abdenour Hadid, Matti Pietikainen, Pavel Matějka, Jaň Černocký, Norman Poh, Josef Kittler, Anthony Larcher, Christophe Levy, et al.

Bi-Modal Person Recognition on a Mobile Phone: using mobile phone data

Chris McCool*, Sébastien Marcel*, Abdenour Hadid†, Matti Pietikäinen†, Pavel Matějka‡, Jan Černocký‡, Norman Poh§, Josef Kittler§, Anthony Larcher¶, Christophe Lévy¶, Driss Matrouf¶, Jean-François Bonastre¶, Phil Tresadern‖, and Timothy Cootes‖

*Idiap Research Institute, Switzerland (Email: csmccool79@gmail.com, sebastien.marcel@idiap.ch)

†University of Oulu, Finland

‡Brno University of Technology, Czech Republic

§University of Surrey, England

¶University of Avignon, France

‖University of Manchester, England

***Abstract*—**

This paper presents a novel fully automatic bi-modal, face and speaker, recognition system which runs in real-time on a mobile phone. The implemented system runs in real-time on a Nokia N900 and demonstrates the feasibility of performing both automatic face and speaker recognition on a mobile phone. We evaluate this recognition system on a novel publicly-available mobile phone database and provide a well defined evaluation protocol. This database was captured almost exclusively using mobile phones and aims to improve research into deploying biometric techniques to mobile devices. We show, on this mobile phone database, that face and speaker recognition can be performed in a mobile environment and using score fusion can improve the performance by more than 25% in terms of error rates.

***Keywords*-bi-modal authentication; mobile biometrics; speaker recognition; face recognition**

I. Introduction

Mobile phones are now a part of many peoples daily lives. They help us to communicate with people by making phone calls and also by sending emails. Given the ubiquitous nature of these devices it is obvious that a real-world test of a pattern recognition algorithm is whether or not it can run on a mobile phone and in a mobile phone environment.

Only recently have researchers considered putting complex pattern recognition algorithms, such as speaker recognition and face recognition, on a mobile phone and this has led to a fractured approach to this problem. In [1], a bi-modal, face and speaker, authentication system was evaluated on Phase I of the MOBIO database [2], but this system never ran on a mobile device. By contrast, in [3] a face, speaker and teeth authentication system was developed to run on a HP iPAQ rw6100 (a portable digital assistant), however, the authors evaluated this system on an in-house database consisting of only 50 people. Other examples include performing only face recognition on a mobile device [4] and evaluating on a small in-house database of 20 people or performing only speaker authentication and evaluating on a small in-house mobile phone database of 50 people [5]. This fractured approach to applying biometric recognition to mobile devices has occurred primarily because of the lack of a large publicly available database and associated protocols.

This paper presents a novel mobile phone database to help evaluate mobile-based face, speaker, and bi-modal authentication algorithms as well as an initial, and novel, bi-modal recognition system. By doing this we hope to stimulate research in the field of multi-modal recognition in a mobile environment. We present our novel publicly-available database (the full MOBIO database) in Section II and the associated protocols are presented in Section III. We then outline our novel bi-modal recognition system, which runs in real-time on a Nokia N900, in Section IV. We then evaluate this bi-modal recognition system on the full MOBIO database in Section V to provide baseline results. We then conclude and outline directions of future work in Section VI.

II. Overview of Database

In this section we present the full MOBIO database. This database is unique because it is a bi-modal database that was captured almost exclusively on mobile phones. It consists of over 61 hours of audio-visual data with 12 distinct sessions usually separated by by several weeks. In total there are 192 unique audio-video samples for each of the 150 participants, this is almost twice the size of Phase I of the MOBIO database. This data was captured at 6 different sites over one and a half years with people speaking English[1].

Capturing the data on mobile phones makes this database unique because the acquisition device is given to the user, rather than being in a fixed position. This means that the microphone and video camera are no longer fixed and are now being used in an interactive and uncontrolled manner. This presents several challenges, including:

- high variability of pose and illumination conditions, even during recordings,

[1]The data was recorded at the following sites: University of Oulu (OULU), Idiap Research Institute (IDIAP), University of Avignon (LIA), University of Manchester (UMAN), University of Surrey (UNIS) and the Brno University of Technology (BUT).

- high variability in the quality of speech, and
- variability in the acquisition environments in terms of acoustics as well as illumination and background.

Below we describe how the database was acquired, including the recording protocols that were followed.

(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 1: Example images from two individuals. It can be seen that there is significant pose and illumination variation between all of the images. Also, it can be seen that the facial hair, (e)-(h) as well as hair style and make-up, (a)-(d), fluctuates significantly.

A. Database Acquisition

The mobile phone database was captured using two mobile devices: one being a Nokia N93i mobile phone and the other being a standard 2008 MacBook laptop. The laptop computer was used to capture only one session (the very first session) while all the other data was captured on the mobile phone, including the first session. The first session was a joint session with two separate recordings, one on the laptop computer and a second recording on the mobile phone captured on the same day at approximately the same time. Capturing the database on a mobile phone meant that the database acquisition was inherently uncontrolled because the mobile phone was given to the user and so the recording device was no longer in a fixed position. To ensure that we obtained useful data we enforced some minimal constraints upon the participants and validated the recorded data using automated tools.

Two constraints were placed upon the users when recording their data. First, we asked the user to ensure that most of the their face was in the image. To help with this, we provided live video feedback as shown in Figure 2 (a), with an ellipse (drawn in red), see Figure 2 (b), being the preferred position of the face. Second, we asked the users to be seated and that they generally talked to the phone. In addition to these minimal constraints, we also validated the data.

To ensure that the acquired data was meaningful and useful two automatic tools were used to validate the audio-video samples. The first tool was a simple speech/non-speech detector[2] that allowed us to measure the signal-to-noise ratio (SNR). If the SNR for an audio-video sample was higher than 6 dB then the sample was inspected by a human

[2]Available at www.irisa.fr/metiss/guig/spro/

(a) (b)

Figure 2: Example images showing how the data was captured when using a mobile phone. In (a) there is an example of a user holding the mobile phone for a recording and in (b) there is an example of the video feedback to the user with the ellipse (in red) in which the face is meant to be within; note that the image in (b) has been digitally altered to enhance the ability of the reader to see the red ellipse.

operator to ensure its intelligibility. The second tool was a simple face detector, similar to the one described in [6]. If this face detector gave a response in less than half of the frames of an audio-video sample then it was inspected by a human operator to ensure that a face was present in the majority of the audio-video frames. If the data failed to pass the inspection of the human operator it was recorded again.

B. Recording Protocols

The database was captured over one and a half years and so it was captured in two phases, Phase I and Phase II. In total more than 61 hours of audio-video data was captured for 150 participants. A summary of the database is given in Table I. To ensure that the data was recorded in a consistent manner at each site for each user a dialogue manager (DM) was implemented on the mobile phone. The DM prompted the participants with: *short response questions*, *free speech questions*, and to read a *pre-defined text*. The DM also presented instructions to the participants such as the pre-defined, and fictitious answers, for the *short response questions*.

As already noted, the data collection was conducted in two phases: Phase I and Phase II. Each session of Phase I consisted of 5 *short response questions*, 5 short *free speech questions*, 1 *pre-defined text*, and 10 *free speech questions*. Phase II was made shorter and consisted of 5 *short response questions*, 1 *pre-defined text*, and 5 *free speech questions*. In all cases the users were asked to speak in a natural manner and so people often correct themselves or add extra words, as is normal for realistic free speech. In addition, the responses to the *free speech questions* are always considered to be fictitious and so *do not* necessarily relate to the question as the sole purpose is to have the subject uttering free speech.

1) Short Response Questions: The short response questions consisted of 5 pre-defined questions, which were: (i) "What is your name?", (ii) "What is your address?", (iii) "What is your birth date?", (iv) "What is your license number?", and (v) "What is your credit card number?". Fictitious responses were provided for each participant and were constant throughout the recordings. The response for the license number was the same for each participant.

Site	Phase I			Phase II		
	NB subjects (female/male)	NB sessions	NB shots	NB subjects (female/male)	NB sessions	NB shots
BUT	**33** (15/18)	6	21	**32** (15/17)	6	11
IDIAP	**28** (7/21)	6	21	**26** (5/21)	6	11
LIA	**27** (9/18)	6	21	**26** (8/18)	6	11
UMAN	**27** (12/15)	6	21	**25** (11/14)	6	11
UNIS	**26** (6/20)	6	21	**24** (5/19)	6	11
UOULU	**20** (8/12)	6	21	**17** (7/10)	6	11

Table I: We present a summary of the number (NB) of: subjects per site, sessions per subject and shots (questions) per session.

2) Free Speech Questions: The free speech questions were designed to ensure that the participant talked for approximately 10 seconds; for the short *free speech questions* we only required that they speak for approximately 5 seconds. The questions were randomly selected from a list of approximately 40 questions and the answer did not have to relate to the question.

3) Pre-Defined Text: The users were asked to read out a pre-defined text. This text was designed to take longer than 10 seconds to utter and the participants were allowed to correct themselves while reading these sentences. The text was the same for all of the recordings.

III. Experimental Protocols

In this section we present the authentication protocol for the full MOBIO database. The database is split into three non-overlapping partitions for: training, development and evaluation. The data is split so that two of the six sites, used for data collection, are used in totality for one partition. This ensures that no identity in one partition will be present in the other partitions.

The **training partition** can be used in any way deemed appropriate and all of the audio-video samples (192 separate audio-video samples for each participant) are available for use. The normal use of the training set would be to derive background models or sub-spaces. For instance the training partition could be used to derive a linear discriminant analysis (LDA) sub-space. This partition consists of the data collected from LIA and UNIS and has a total of 50 (13 female and 37 male) subjects.

The development and evaluation partitions are used to evaluate the performance of the systems. The **development partition** consists of data acquired from UMAN and UOULU and has a total of 42 (18 female and 24 male) subjects. It can be used to derive hyper-parameters, such as the optimal number of feature dimensions or weights to perform fusion of the scores from the face and speaker authentication systems, such as for the system described in Section IV-C. The development partition is also used to derive thresholds that are then applied to the evaluation partition to obtain the final system performance, see Section III-A for details. The **evaluation partition** should be used sparingly as it provides the final, unbiased, performance of the system. It consists of the data collected from BUT and IDIAP and has a total of 58 (20 female and 38 male) subjects.

The development and evaluation partitions have a similar protocol which consist of: (i) making a model for each client (client enrolment), and (ii) testing client models to produce true access scores and false access scores. We make a model for each client by using the first 5 questions from the their first session, these are the *short response questions* and we refer to this as the enrolment data. Testing of the client models is performed in a gender-dependent manner using the *free speech questions* from the remaining 11 sessions. We use the *free speech questions* because it is different to the enrolment data and thus represents a harder and more realistic problem. This leads to 105 true access scores, or samples, for each client; a true access score corresponds to a sample (video or audio) of the i^{th} client that is compared against the model of the i^{th} client. The false access scores are obtained by using the 105 samples from all of the other clients of the same gender; a false access score corresponds to a sample (video or audio) of the i^{th} client that is compared against a model(s) that is not the i^{th} client. Thus, if there are 30 models for a particular gender then false scores would be produced by using the 105 test samples from the other 29 users. For clarity and ease of use, the database, protocol files and annotations have been made publicly-available at http://www.idiap.ch/dataset/mobio[3].

A. Performance Evaluation

In this section we outline how to measure the performance of an authentication system. This should not be confused with a recognition system, the recognition system consists of multiple steps one of which is authentication. For instance face recognition consists of face detection followed by face authentication. Authentication is the process of confirming if the sample t came from the claimed identity i, by comparing it against their model θ_i. The comparison against the model produces a score s which is compared against a threshold τ and is accepted as confirming their identity if $s > \tau$.

To measure the performance of an authentication system we propose three methods. The first is to present a single number to represent the authentication system's performance using the half total error rate (HTER). The second is to present a summary of the authentication system's performance using a detection error tradeoff (DET) plot. The third is to present the unbiased performance of the authentication system using an expected performance curve (EPC).

[3]The development partition consists of $1,890$ true and $32,130$ false access scores for female trials and $2,520$ true and $57,960$ false access male trials. The evaluation partition consists of $2,100$ true and $39,900$ false access scores for female trials and $3,990$ true and $147,630$ false access scores for male trials.

The HTER is used to represent the performance of an authentication system with a single number. The HTER is obtained by deriving a threshold, τ^*_{eer}, on an independent data set at the equal error rate (EER) point. This threshold is then applied to the evaluation partition ($\mathcal{D}_{evl}$). The HTER is the weighted error rate (WER) [7] of the false alarm rate (FAR) and the false rejection rate (FRR) when $\beta = 0.5$,

$$WER(\tau^*_{eer}, \mathcal{D}_{evl}, \beta) = \beta \cdot FAR(\tau^*_{eer}, \mathcal{D}_{evl}) + (1-\beta) \cdot FRR(\tau^*_{eer}, \mathcal{D}_{evl}), \quad (1)$$

The DET [8] and EPC [7] plots provide a quick summary of the overall system performance. The DET plot consists of the miss probability (FRR) versus the probability of false acceptance (FAR). However, the DET plot can be misleading because it presents the authentication system performance without an associated threshold. Therefore, we also propose the use of the EPC which presents a summary of the WER for different operating points [7]. When producing an EPC an optimal threshold τ^*_o has to be derived for multiple operating points, o, corresponding to a tradeoff between the FAR and FRR. Thus, the optimal threshold τ^*_o is computed using different values of $\beta_o \in [0; 1]$ corresponding to different operating points,

$$\tau^*_o = \underset{\tau}{\operatorname{argmin}} \; WER(\tau, \mathcal{D}_{dev}, \beta_o), \quad (2)$$

where $\mathcal{D}_{dev}$ denotes the development set. Finally, the performance for different values of β_o is then computed on $\mathcal{D}_{evl}$ using the threshold tuned for each operating point.

IV. Bi-Modal Recognition System

The bi-modal recognition system consists of three parts: face recognition, speaker recognition, and fusion. Each of these parts is described in detail below.

A. Face Recognition System

Face recognition consists of two components: **face localisation** and **face authentication**. Both components rely on using local binary patterns (LBPs) [9], or a variant called the modified census transform (MCT) [10], to efficiently encode the face image.

The **face localisation** component consists of finding the largest face in each video frame. To do this we apply a face detector that is trained as a cascade of classifiers using MCTs. We took the implementation of [6] and re-implemented it to use integer arithmetic; an integer arithmetic approach was taken to reduce the computation overhead of running on a mobile phone. This detector outputs a single detection, or localisation result, by considering the largest detected face as being the face of interest, this simplification is valid because it is expected that the user is interacting with the device and consequently their face should be the largest detected face. After the face localisation stage all detected frames are normalised to be 64×80 pixel gray scale images, with a distance between the eyes of 33 pixels. The detected frames are then passed onto the face authentication component.

The **face authentication** component divides the detected faces into non-overlapping regions and each region is then represented using a histogram of LBP features. This is a parts-based scheme which is similar to [11] but avoids the need to compute the 40 different Gabor wavelets. For each region, r, we calculate an average histogram, θ_r, by taking the average over all of the detected frames from the enrolment or test videos. Thus, given the enrolment histograms for client i, θ_i, and the test histograms for the video t, θ_t, the similarity for the r^{th} region is,

$$s_r\left(\theta_{i,r}, \theta_{t,r}\right) = \frac{\sum_{j=1}^{J} min\left(\theta_{i,r}(j), \theta_{t,r}(j)\right)}{\sum_{j=1}^{J} \theta_{i,r}(j)}, \quad (3)$$

where J is the total number of histogram bins corresponding to the distinct number of LBP codes ($J = 256$ in this work). The similarity is then summed for all of the regions to produce a single similarity measure, or score, $s = \sum_{r=1}^{R} s_r$. In our work we had $R = 80$ non-overlapping regions.

B. Speaker Recognition System

Speaker recognition consists of two components: voice activity detection (**VAD**) and **speaker authentication**.

The role of **VAD** is to select only speech and to discard silence and noise. The VAD component used in this work uses a Hungarian downscaled phoneme recogniser which is the cascade of 3 neural networks. The input is the long temporal context of mel-filter banks and the output is a posterior probability of detected phonemes [12]. The posterior probability of all silence classes are summed for each frame to form a silence class and the same is applied for phonemes to form a speech class. Viterbi decoding is applied to smooth the output. After VAD all valid frames are passed to the speaker authentication component.

The **speaker authentication** component uses an i-vector extractor [13] to obtain features which are then modelled using probabilistic linear discriminant analysis (PLDA) [14]. The i-vector approach obtains a low dimensional fixed length feature vector to represent a variable length utterance. It does this by taking a set of observations from an utterance which are then used to obtain a point MAP estimate [13], this point MAP estimate then yields the low dimensional i-vector.

The fixed length i-vectors extracted per utterance can be used as input to a standard pattern recognition algorithm. We use a PLDA model [14] that provides a probabilistic framework applicable to fixed-length input vectors. The i-vectors $\mathbf{w}_{\mathcal{X}}$ are assumed to be distributed according to the form

$$\mathbf{w}_{\mathcal{X}} = \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{U}\mathbf{x} + \epsilon \quad (4)$$

where $\mathbf{V}$ is a sub-space of identity variability and $\mathbf{U}$ is a sub-space of channel (or session) variability. Using the PLDA model, one can directly evaluate the log-likelihood ratio, to obtain a score, for the hypothesis test corresponding

to "whether the two i-vectors were or were not generated by the same speaker." Note that the difference between the *enrolment segment* (on which a model was created) and the *test segment* (which is scored against the model) vanishes – scoring with i-vectors is symmetrical.

For the system to work on the mobile device we chose the following parameters. We obtained observations from the audio signal using 19 Mel-frequency cepstrum coefficients [15] and C0 features. This base feature vector was augmented with their delta and double delta coefficients resulting in 60 dimensional feature vectors which were then processed by short time Gaussianisation [16] over a 50 frame window. A universal background model (UBM) [17] with 128 Gaussians was used (but usually it is 2048). The i-vector was trained such that it resulted in a $M = 400$ dimensional vector. We used one simplification (constant GMM component alignment) for running an i-vector extractor on the mobile phone because of memory consumption issues [18]. The parameters for PLDA were 90 dimensions for the speaker sub-space $\mathbf{V}$ and full-rank (400) for the channel subspace $\mathbf{U}$.

C. Score Fusion

Score fusion consists of normalising the individual face and speaker authentication scores and then fusing them. The similarity scores, for face authentication, and the log-likelihood scores, for speaker authentication, were normalised to produce probabilities such that the scores lie in the range $[0, 1]$. We achieve this by using logistic regression, hence taking y_i as the input and producing probability, P_i, as the output:

$$P_i = (1 + \exp(-ay_i + b))^{-1} \quad (5)$$

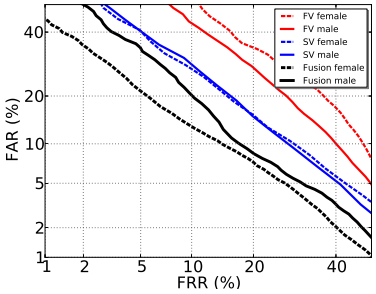
where a is a scaling factor and b is a bias.

We chose to use logistic regression for two reasons. First, it is a well establishe method in statistics, belonging to a family of methods called generalized linear models. Its optimisation procedure, known as "gradient ascent" [19], is well understood; it converges to a global minimum, meaning that there is a unique solution. Second, its output can be interpreted as a probability and so presenting this information to the user is more meaningful than just the raw score.

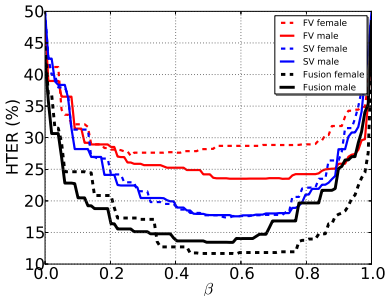
The final combined score is obtained by taking the product of the two scores, P_{video} for face authentication and P_{audio} for speaker authentication. Although the sum rule could have also been used. Both rules result in similar generalisation performance.

V. Experimental Results

In this section we present the results for our fully automatic bi-modal system on the full MOBIO database. This system runs in real-time on a Nokia N900 and demonstrates that automatic bi-modal authentication can be deployed on a mobile phone; the system processes approximately 15 frames per second. To evaluate the performance of our face, speaker, and bi-modal systems we processed the MOBIO database on the mobile phone (a Nokia N900). A demonstration of our system can be found at http://www.mobioproject.org/demonstrations.



(a) DET plot



(b) EPC

Figure 3: We present in (a) a DET plot and (b) an EPC plot of the results obtained on the mobile phone database. Results for the face (FV), speaker (SV) and bi-modal (Fusion) authentication systems are presented in red, blue and black respectively. The gender-dependent trials for males and females are presented as the solid and dashed lines respectively.

We present the result for all three authentication components of our system, face, speaker, and bi-modal authentication, in Figure 3. The DET plot in Figure 3 (a) shows that the performance of the face and speaker authentication components are quite modest, while the EPC in Figure 3 (b) shows that their performance is consistent across the development and evaluation partitions. The modest performance of the face and speaker authentication systems can be attributed to the fact that simplified systems have been used in order to run in the mobile environment. Despite this, the fusion (bi-modal) system appears to perform quite well.

It can be seen in Figure 3 that the bi-modal system outperforms either modality on its own. The single best performing modality is audio (speaker authentication) which achieves an HTER of 17.7% for female trials and 18.2% for male trials, see Table II. Despite the fact that the audio modality performs as much as 37% better than the video modality it can be seen that performing bi-modal authentication, score fusion, leads to substantial improvements in performance. For instance, the performance is improved by 25% for female trials, to 13.3%, and by 35% for male trials, to 11.9%.

From these results we have shown that it is possible to implement a fully automatic bi-modal recognition system

	Video (FV)		Audio (SV)		Fusion	
	male	fem.	male	fem.	male	fem.
Dev Set (EER %)	21.6%	20.9%	18.0%	15.1%	**10.9**%	**10.5**%
Test Set (HTER %)	24.1%	28.2%	18.2%	17.7%	**11.9**%	**13.3**%

Table II: The results obtained on the mobile phone database for the face (video), speaker (audio) and bi-modal (fusion) authentication system, for male and female (fem.). All of these systems run in real-time on a Nokia N900 mobile phone.

which can run in real-time on a mobile phone. To the best of our knowledge this is the first such system that has been benchmarked using a publicly-available database captured on mobile phones.

VI. Conclusions and Future Work

One of the main outcomes of this work has been to acquire and make publicly-available a mobile phone database. This database has associated protocols, that are well defined and so will allow for easy comparison between different systems.

To encourage the development of other mobile phone based recognition systems we have described and evaluated our own novel bi-modal recognition systems. This system runs in real-time on a Nokia N900 and performs face, speaker, and bi-modal recognition. We have shown that this can be effectively deployed on a mobile phone and can be used for authentication in real-world conditions for a mobile phone. Finally, we have also shown that performing the fusion of face and speaker recognition can provide a significant performance improvement with a relative gain in performance of more than 25%.

We hope that future work will examine other techniques to perform accurate face, speaker and bi-modal authentication on a mobile phone using the public and free database described in this paper.

Acknowledgements

This work has been performed by the MOBIO and Tabula Rasa projects of the 7th Framework Research Programme of the European Union (EU) grant agreement numbers 214324 and 257289. For more information about the MOBIO consortium please visit http://www.mobioproject.org.

References

[1] Linlin Shen, Nengheng Zheng, Songhao Zheng, and Wei Li, "Secure mobile services by face and speech based personal authentication," in *IEEE International Conference Intelligent Computing and Intelligent Systems*, 2010, pp. 97–100.

[2] S. Marcel and et al., "On the results of the first mobile biometry (mobio) face and speaker verification evaluation," in *Proceedings of the ICPR 2010 contests*, 2010.

[3] Dong-Ju Kim, Kwang-Woo Chung, and Kwang-Seok Hong, "Person authentication using face, teeth and voice modalities for mobile device security," in *IEEE Transactions on Consumer Electronics*, 2010, pp. 2678–2685.

[4] Tao Qian and R. Veldhuis, "Biometric authentication system on mobile personal devices," in *IEEE Transactions on Intrumentation and Measurement*, 2010, pp. 763–773.

[5] K. S. Rao, A. K. Vuppala, S. Chakrabarti, and L. Dutta, "Robust speaker recognition on mobile devices," in *International Conference on Signal Processing and Communications*, 2010.

[6] Y. Rodriguez, *Face detection and verification using local binary patterns*, Ph.D. thesis, Idiap Research Institute and École Polytechnique Fédérale de Lausanne, 2006.

[7] S. Bengio, J. Mariéthoz, and M. Keller, "The Expected Performance Curve," in *International Conference On Machine Learning*, 2005.

[8] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Eurospeech*, 1997, vol. 4, pp. 1895–1898.

[9] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[10] B. Froba and A. Ernst, "Face detection with the modified census transform," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2004, pp. 91–96.

[11] Wenchao Zhang, Shiguang Shan, Wen Gao, Xilin Chen, and Hongming Zhang, "Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition," in *IEEE International Conference on Computer Vision*, 2005, pp. 786–791.

[12] P. Schwarz, P. Matějka, and J. Černocký, "Towards lower error rates in phoneme recognition," in *Lecture Notes in Computer Science*, 2004, pp. 465–472.

[13] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. PP, no. 99, pp. 1 –1, 2010.

[14] S. J. D. Prince and S. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE International Conference on Computer Vision*, Rio de Janeiro, Brazil, 2007.

[15] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[16] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *A Speaker Odyssey - The Speaker Recognition Workshop*, 2001.

[17] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[18] O. Glembek, L. Burget, P. Matějka, M. Karafiat, and P. Kenny, "Simplification and optimization of i-vector extraction," in *International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 4516–4519.

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