Bi-Modal Authentication in Mobile Environments Using Session Variability Modelling

Petr Motlicek, Laurent El Shafey, Roy Wallace, Christopher McCool, Sébastien Marcel
Idiap Research Institute, Martigny, Switzerland *
{*petr.motlicek,laurent.el-shafey,roy.wallace,christopher.mccool,sebastien.marcel*}*@idiap.ch*

Abstract

We present a state-of-the-art bi-modal authentication system for mobile environments, using session variability modelling. We examine inter-session variability modelling (ISV) and joint factor analysis (JFA) for both face and speaker authentication and evaluate our system on the largest bi-modal mobile authentication database available, the MOBIO database, with over 61 hours of audio-visual data captured by 150 people in uncontrolled environments on a mobile phone. Our system achieves 2.6% and 9.7% half total error rate for male and female trials respectively – relative improvements of 78% and 27% compared to previous results.

1. Introduction

Mobile phones are an increasingly ubiquitous part of our daily lives. Biometric authentication on these devices is particularly challenging because the environments in which they are used are, by definition, changeable and uncontrolled. As they typically incorporate both a microphone and a camera, mobile devices present a unique opportunity to apply a bi-modal approach, using face and speech data, to biometric authentication. This is a relatively new challenge that has received limited attention, with most previous studies using small in-house databases [4, 7, 8]. An international competition was organized in 2010 for bi-modal authentication on the largest publicly-available database of audio-visual samples collected on mobile phones: the MOBIO database [5]. However, only a small subset, namely Phase I, was available during that evaluation.

More recently, [6] presented the first benchmark results of bi-modal authentication on the complete MOBIO (Phase I and II) protocol, using a fully automatic real-time system running within the hardware constraints of a Nokia N900 mobile phone.

This paper differs from and extends on the work of [6] in several ways. Firstly, rather than imposing the hardware constraints of a mobile phone, this work focuses purely on improving authentication accuracy on bi-modal data captured in challenging mobile environments. Our approach is to exploit recent advances in inter-session variability modelling (ISV) and joint factor analysis (JFA) using Gaussian mixture models (GMMs) for both face and speaker authentication, to reach state-of-the-art performance on the MOBIO database. Secondly, while [6] made use of automatic face detection from videos, here we use the facial image data specified by the MOBIO Still-Image protocol, for improved reproducibility and easier comparison of face recognition algorithms. The MOBIO Still-Image protocol has been utilised before for the task of face authentication in [11]. In this paper we extend this previous work to the task of bi-modal authentication.

The following are the major contributions of this paper. We are the first to propose a fused face and speaker authentication system that uses session variability modelling techniques in both modalities. Secondly, we show that fusing the results improves authentication accuracy over using either of the modalities alone. Thirdly, we compare session variability modelling techniques for speaker authentication in mobile environments, using only the training data set of the MOBIO database, and show that ISV performs favourably. Finally, the proposed bi-modal system achieves the most accurate results so far on the complete MOBIO authentication protocol, providing 78% and 27% relative improvements for male and female trials over previous results [6].

Section 2 describes the session variability modelling framework. Section 3 describes our approaches to feature extraction and score fusion, followed by experimental protocols and results.

*The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7) under grant agreements 238803 (BBfor2) and 257289 (TABULA RASA).

2. GMM-based face and speech modelling

Bi-modal authentication requires the processing of both image data (faces) and audio data (speech). Naturally, two separate feature extraction processes are used for the two modalities, as will be described in Section 3.1. In both cases, however, when feature extraction is performed for a biometric sample $\boldsymbol{\mathcal{X}}$, which we use interchangeably to represent a facial image or spoken utterance, the output is a set of K feature vectors, $\boldsymbol{X} = \{\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^K\}$. For a video containing face and speech samples, each modality is modelled separately, and an overall authentication score is derived at the last stage using score fusion (see Section 3.2).

In this work, we use very similar modelling techniques for both face and speech modalities. Specifically, we use a generative probabilistic framework that models the observed feature vectors using GMMs. This framework remains the basis of state-of-the-art speaker recognition [10, 3]. For face recognition, it was found to offer the best trade-off in terms of complexity, robustness and discrimination [1], and was further developed in [11] to incorporate session varability modelling. In the remainder of this section we first describe, very briefly, the baseline GMM-based approach, followed by the session variability modelling approaches that build on this baseline and are evaluated in this work.

2.1 Baseline GMM-based approach

In the baseline approach, the distribution of feature vectors for each client is modelled by a GMM estimated using background model adaptation [1], which utilises a universal background model (UBM) as a prior for deriving client models with maximum *a posteriori* (MAP) adaptation. Typically, covariance matrices are assumed to be diagonal and only the means of the GMM components are adapted, as this has been consistently shown to be most effective for both speaker and face recognition. We use GMMs with 512 components as in [11]. Given a test sample, $\boldsymbol{\mathcal{X}}_t$, a client is authenticated by comparing the extracted features $\boldsymbol{X}_t$ to the model of the claimed client identity, $\boldsymbol{s}_i$, and calculating an average log likelihood ratio score with respect to the UBM producing a score $h(\boldsymbol{X}_t, \boldsymbol{s}_i)$. An efficient approximation to the log-likelihood ratio known as linear scoring is used to improve authentication speed without loss of accuracy [2]. Finally, ZT-norm score normalisation is applied as in [11] to produce $\overline{h}(\boldsymbol{X}_t, \boldsymbol{s}_i)$. The client is authenticated if and only if $\overline{h}(\boldsymbol{X}_t, \boldsymbol{s}_i)$ is greater than a tuned decision threshold.

2.2 Session variability modelling

In the baseline approach using mean-only MAP adaptation, a client's model, $\boldsymbol{s}_i$, is effectively the result of adding an offset, $\boldsymbol{d}_i$, to the UBM, $\boldsymbol{m}$, in a high-dimensional GMM mean supervector space[1].

$$\boldsymbol{s}_i = \boldsymbol{m} + \boldsymbol{d}_i \tag{1}$$

This offset is difficult to estimate reliably with limited enrolment data because it is sensitive to the conditions in which the data was captured. To address this problem, *session variability modelling* techniques were developed that constrain mean offsets to lie within linear, low-dimensional subspaces. In this way, small amounts of enrolment data can be utilised more reliably by appropriately constraining the directions of adaptation within supervector space.

Specifically, the observations of the j'th image of client i, $\boldsymbol{X}_{i,j}$, are assumed to be drawn from a distribution specified not by (1) but instead by $\boldsymbol{\mu}_{i,j}$, where

$$\boldsymbol{\mu}_{i,j} = \boldsymbol{m} + \boldsymbol{U}\boldsymbol{x}_{i,j} + \boldsymbol{d}_i \tag{2}$$

and $\boldsymbol{U}$ is a low-dimensional session variability subspace trained with an expectation-maximisation (EM) algorithm on a large training set. Finally, $\boldsymbol{x}_{i,j}$ are latent factors with standard normal priors [10].

The first work to apply session variability modelling to face authentication was [11], using two techniques known as ISV [10] and JFA [3]. As described in more detail in [11], ISV and JFA differ in their definition of the client-dependent offset $\boldsymbol{d}_i$. Briefly, given latent factors $\boldsymbol{z}_i$ and $\boldsymbol{y}_i$ with standard normal priors, for ISV, $\boldsymbol{d}_i = \boldsymbol{D}\boldsymbol{z}_i$, where the diagonal matrix $\boldsymbol{D}$ is defined as a function of the UBM covariance. For JFA, $\boldsymbol{d}_i = \boldsymbol{V}\boldsymbol{y}_i + \hat{\boldsymbol{D}}\boldsymbol{z}_i$, where $\boldsymbol{V}$ is a low-dimensional client variability subspace and both $\boldsymbol{V}$ and $\hat{\boldsymbol{D}}$ are learnt from training data using EM. In this work, for the first time, we evaluate the use of ISV and JFA for bi-modal (face and speech) authentication.

3. Experiments

In this section we describe our face and speech features, our approach to score fusion, experimental protocols and, finally, results on the MOBIO database.

3.1 Face and speech feature extraction

For face authentication we use the GMM-based system of [11]. Feature extraction is based on the approach

[1] A GMM mean supervector is an NM-dimensional vector formed by concatenating the M-dimensional means from each of the N components of a GMM.

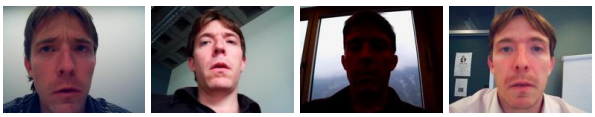


Figure 1: MOBIO database example images.

of [9] that divides the face in a set of overlapping blocks, assumed to be independent observations of local face features. Each block is mean and variance normalised before extracting 2D discrete cosine transform coefficients. The lowest frequency coefficients are retained excluding the zeroth coefficient. As in [11] we use a block size of 12×12 pixels and retain 44 coefficients.

For speaker authentication, speech segments are first isolated using energy-based voice detection. Then mel frequency cepstral coefficient (MFCC) features are extracted for 25ms frames with 10ms overlap and a 24-band filter bank. The resulting 60-dimensional feature vectors contain 19 MFCCs plus energy, delta and double delta coefficients.

3.2 Score fusion

For bi-modal authentication, scores are first derived using the uni-modal sub-systems, giving $\overline{h}_{\text{face}}(\boldsymbol{X}_t^{\text{face}}, \boldsymbol{s}_i^{\text{face}})$ and $\overline{h}_{\text{speech}}(\boldsymbol{X}_t^{\text{speech}}, \boldsymbol{s}_i^{\text{speech}})$, before being fused to produce a single score for the trial, $\overline{h}_{\text{fused}}(t, i)$. Fusion is performed by a weighted sum of the two scores. We examine two different approaches to obtain the weights. The first is to use equal weights and the second is to use linear logistic regression (LLR) to learn the optimal weights on the development set[2].

3.3 Database and protocols

We evaluate authentication accuracy on the largest publicly available bi-modal mobile phone database, the MOBIO database [6]. This database was recorded almost exclusively on mobile phones and contains audio-video data of 150 people captured over one and a half years. Figure 1 demonstrates typical session variability observed in this database. We use the facial image data specified by the MOBIO Still-Image protocol[3] as in [11]. The MOBIO protocol defines separate training, development and testing sets (see Section III of [6]). For UBM training and score normalisation we use the same subsets of the training data as [11], except that gender-dependent T-norm cohorts are used for the speaker authentication sub-system. The development set is used to tune the dimensionality of the ISV and JFA subspaces, and find the decision threshold that minimises the equal error rate (EER) on the development set. This threshold is then applied to the test set to calculate the half total error rate (HTER)[4]. We also provide test set detection error tradeoff (DET) plots and expected performance curves (EPCs) as in [6].

[2]We use the implementation of LLR from `http://niko.brummer.googlepages.com/focalmulticlass`

[3]The protocol and manual annotations are available from `http://www.idiap.ch/resource/biometric`

Modality	System	Male		Female	
		Dev	Test	Dev	Test
Face	McCool et al. [6]	21.6	24.1	20.9	28.2
	GMM	9.2	10.5	10.7	20.4
	ISV	**3.6**	7.5	**6.7**	**12.2**
	JFA	4.0	**7.3**	7.7	13.2
Speaker	McCool et al.[6]	18.0	18.2	15.1	17.7
	GMM	12.6	15.8	20.0	22.6
	ISV	**8.2**	**8.9**	**11.9**	**15.3**
	JFA	15.5	14.7	23.1	19.4
Fusion	McCool et al. [6]	10.9	11.9	10.5	13.3
	ISV (sum)	2.1	3.3	3.8	11.0
	ISV (LLR)	**1.2**	**2.6**	**2.3**	**9.7**

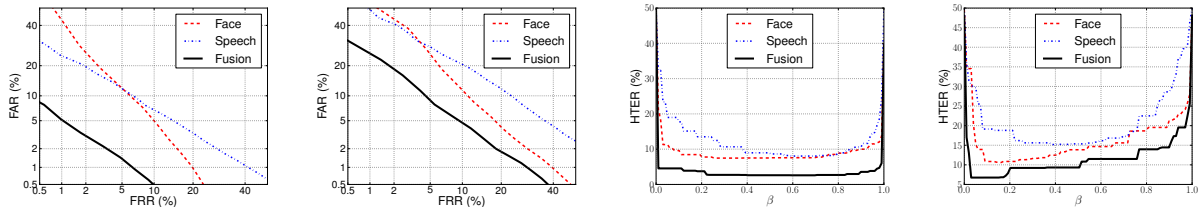
Table 1: Authentication error rates (Dev set EER, Test set HTER in %) on MOBIO using either a uni-modal face, uni-modal speaker, or bi-modal (fused) system.

3.4 Results

In this section, we present the results of uni-modal face and speaker authentication systems as well as the bi-modal fused systems. The resulting error rates are reported in Table 1, with corresponding DET and EPC plots in Figure 2. The reference systems for face, speaker and bi-modal authentication [6] are evaluated on the same data. Those systems were optimized to run in real-time on a mobile phone. The face authentication component was based on histograms of local features obtained from faces detected automatically in videos. The speaker authentication component was based on i-vector features modelled using probabilistic linear discriminant analysis. The fusion was performed using a product rule on normalized scores.

In our experiments, both face and speaker uni-modal authentication systems were built using GMM, ISV and JFA algorithms. It can be seen that ISV performed the best in both face and speaker uni-modal systems on the development set and test set, with only one exception (the face system on the male test set). In fact, using ISV instead of the baseline GMM technique consistently reduced the HTER by around 30% to 40% across both modalities. In almost all cases, ISV significantly outperformed JFA, possibly due to the limited amount of training data in the MOBIO database.

[4]The average of false acceptance and false rejection rates



(a) Male Test DET (b) Female Test DET (c) Male Test EPC (d) Female Test EPC

Figure 2: DET and EPC plots comparing ISV systems: uni-modal face, speech, and bi-modal fused (LLR).

The dimensionalities of the $\boldsymbol{U}$ and $\boldsymbol{V}$ subspaces used in ISV and JFA algorithms were optimized on the development set. It is worth noting that, in the case of ISV, results were not very sensitive to the dimensionality of $\boldsymbol{U}$. The optimal value for the face system was 320 while for the speaker system it was 40 and 320 for males and females respectively.

The best fused bi-modal system was obtained by combining the face and speaker ISV systems. Using LLR score fusion, the HTER was further reduced by 65% for males and 20% for females. We also experimented with fusing the uni-modal JFA systems but found no additional improvements. Our ISV-based bi-modal system thus achieves 2.6% HTER for males and 9.7% for females, which is a relative improvement of 78% and 27% respectively compared to [6].

Finally, we analysed score calibration by calculating the increase in test HTER with respect to the corresponding test set EER. For males and females, the test set HTER was 2.6% and 9.7% respectively, as shown in Table 1. In contrast, we found that the corresponding test set EER values were 2.6% and 6.8%. Thus, the difference between test set HTER and EER was negligible for male trials but for females there was a substantial difference (2.9% absolute). This suggests that the female test set HTER could be considerably reduced by addressing an apparent score calibration problem.

4. Conclusions

This paper presented a state-of-the-art bi-modal authentication system robust to challenging mobile environments. For the first time, session variability modelling techniques were evaluated for both face and speaker authentication on the MOBIO database. Results proved the effectiveness of the techniques, with ISV performing particularly well for both face and speaker authentication. By using LLR score fusion, the error rate was further reduced substantially, showing the complementarity of the two modalities. The resulting bi-modal authentication system provides the most accurate result by far (relative improvements of at least 30% for the fused system) on the MOBIO database when compared with previously published results.

In future work we plan to investigate gender-dependent training, the use of additional training databases, and the aforementioned score calibration issues. While this paper focused on the accuracy of algorithms, future work includes development to run within the practical constraints of mobile platform hardware.

References

[1] F. Cardinaux et al. User authentication via adapted statistical models of face images. *IEEE Trans. Signal Process.*, 54:361–373, 2006.

[2] O. Glembek et al. Comparison of scoring methods used in speaker recognition with joint factor analysis. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pages 4057–4060, 2009.

[3] P. Kenny et al. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Trans. Audio, Speech, Lang. Process.*, 15(4):1435–1447, May 2007.

[4] D.-J. Kim et al. Person authentication using face, teeth and voice modalities for mobile device security. In *IEEE Trans. Consum. Electron.*, pages 2678–2685, 2010.

[5] S. Marcel et al. On the results of the first mobile biometry (MOBIO) face and speaker verification evaluation. In *Proc. Int. Conf. Pattern Recognition contests*, 2010.

[6] C. McCool et al. Bi-modal person recognition on a mobile phone: using mobile phone data. In *Proc. IEEE ICME Workshop on Hot Topics in Mobile Multimedia*, Melbourne, July 2012. Available at http://publications.idiap.ch.

[7] T. Qian and R. Veldhuis. Biometric authentication system on mobile personal devices. In *IEEE Trans. Instrum. Meas.*, pages 763–773, 2010.

[8] K. S. Rao et al. Robust speaker recognition on mobile devices. In *Proc. Int. Conf. Signal Processing and Communications*, 2010.

[9] C. Sanderson and K. Paliwal. Fast features for face authentication under illumination direction changes. *Pattern Recognition Letters*, 24:2409–2419, 2003.

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

[11] R. Wallace et al. Inter-session variability modelling and joint factor analysis for face authentication. In *Proc. Int. Joint Conference on Biometrics*, 2011.