On the Improvements of Uni-modal and Bi-modal Fusions of Speaker and Face Recognition for Mobile Biometrics

Elie Khoury, Manuel Günther, Laurent El Shafey and Sébastien Marcel

Idiap Research Institute, Martigny, Switzerland

{elie.khoury,manuel.guenther,laurent.el-shafey,sebastien.marcel}@idiap.ch

Abstract

The MOBIO database provides a challenging test-bed for speaker and face recognition systems because it includes voice and face samples as they would appear in forensic scenarios. In this paper, we investigate uni-modal and bi-modal multi-algorithm fusion using logistic regression. The source speaker and face recognition systems were taken from the 2013 speaker and face recognition evaluations that were held in the context of the last International Conference on Biometrics (ICB-2013). Using the unbiased MOBIO protocols, the employed evaluation measures are the equal error rate (EER), the half-total error rate (HTER) and the detection error trade-off (DET). The results show that by uni-modal algorithm fusion, the HTER's of the speaker recognition system are reduced by around 35 %, and of the face recognition system by between 15 % and 20 %. Bi-modal fusion drastically boosts recognition by a relative gain of 65 % - 70 % of performance compared to the best uni-modal system.

1. Introduction

During the last years, more and more surveillance devices are installed in public places or in private properties to possibly capture crime scenes. Some of them are able to record both image and voice data. Similarly, instant messaging systems such as Skype, Google talk, Yahoo messenger and Facebook messenger support both video and audio plugins. In forensic investigations, usually a human operator would compare these recordings to samples from suspects. The studies in [1, 2] have shown that automatic speaker and face recognition algorithms can outperform humans in comparing speech utterances or images from unfamiliar identities. Therefore, using automatic algorithms to verify suspects' identities based on their voice and their face are favorable in these cases.

Automatic speaker recognition is investigated since the 1970s [3] and regularly evaluated by the National Institute of Standards and Technology (NIST)[1] since 1996. Similarly, automatic face recognition started in the late 1980s [4], and many evaluations were conducted. Since 2000, face recognition vendor tests (FRVT) [5], which are also executed by NIST,[2] evaluate capabilities of automatic face recognition applications under controlled conditions.

The appropriate framework for the admissibility of the results of automatic forensic face and speaker recognition systems in front of court is to use evidence interpretation, and to standardize the procedures and the protocols for testing [6]. This allows a scientific and logical methodology to clearly determine the capability of the systems and to be conscious of the error rates. Recent scientific efforts have converged towards exploiting the Bayesian approach for the analysis of evidence, such that opinions about the prosecution and defense hypotheses are expressed in the form of posterior probabilities [7]. In this sense, using log-likelihood ratio (LLR) as a degree of support of one hypothesis over the other has become a crucial demand that successful forensic speaker and face recognition systems should afford.

After the success of the first edition in 2010 [8], the Biometric Group at the Idiap Research Institute organized the second edition of speaker [9] and face [10] recognition evaluations in mobile environment. These evaluations are conducted on the MOBIO database, which provides the unique opportunity to analyze two mature biometrics, i. e., speaker and face recognition side by side in a challenging environment. The conditions in MOBIO are closer to forensic scenarios than during NIST evaluations and, hence, it is more suitable to show algorithm capacities for forensic investigations. Two unbiased evaluation protocols exist for MOBIO, which allow a direct comparison of results of different algorithms with figures published in literature.

In total, 12 institutions participated to the speaker recognition evaluation [9], while the face recognition evaluation [10] analyzed 9 systems. All participants of both evaluations had to strictly follow the unbiased evaluation protocols. To assure a fair comparison, during the evaluations the file names of the test data were anonymized so that the participants could not use the name of the probe file to infer identity. In [9, 10], the speaker and face recognition systems were assessed, and [9] already showed that fusing different speaker recognition systems

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

[2] http://www.nist.gov/itl/iad/ig/frvt-docs.cfm

outperforms each single system. In this paper, we investigate whether the integration of state-of-the-art speaker and face recognition systems can further improve performance.

The remainder of the paper is as follows. Section 2 introduces the MOBIO database and the evaluation protocols. In section 3, the techniques that were used in systems submitted to the speaker and face evaluations are described briefly, and the employed multi-algorithm and bi-modal fusion technique is detailed. The experimental evaluation is provided in section 4, while section 5 concludes the paper.

2. MOBIO Database

The MOBIO database is a bi-modal, face and speaker, video database recorded from 152 people. MOBIO is challenging since the data is acquired on mobile devices with real noise. The extracted images contain faces with uncontrolled illumination, facial expression, near-frontal pose and occlusion, while the extracted speech segments are relatively short, sometimes less than 2 seconds. Therefore, this database is suited to evaluate algorithms under uncontrolled conditions as they would appear in surveillance scenarios. More technical details about the MOBIO database can be found in [11] and on its official web page.[3]

Based on the gender of the clients, two different evaluation protocols *male* and *female* exist. These protocols are identical for speaker and face recognition. Each five recordings per client are used to enroll client models, the remaining recordings serve as probes. Similarity scores are computed between all models and all probes. In order to have unbiased protocols, the clients of the database are split up into 3 different sets: training, development and evaluation, which are statistically detailed in table 1. The use of the sets is restricted to:

Training set The data of this set is used to learn the background parameters of the algorithm (projection matrices, background models, etc.). It can also be used as a cohort for score normalization.

Development set The data of this set is used to optimize meta-parameters of the algorithm. Scores produced with this set can be exploited to train calibration parameters for system fusion.

Evaluation set The data of this set is used for computing the final evaluation performance. No training or tuning is allowed to be performed on this set.

3. Multi-algorithm and Bi-modal Fusion

The MOBIO database permits to integrate information from voice and face samples to recognize clients. In [9] and [10], several state-of-the-art speaker and face recognition systems provided score files for MOBIO. In this section, we give a short summary of the systems that participated in the evaluations.[4] For more detailed information, please refer to [9, 10].

3.1. Speaker Recognition

A text-independent speaker recognition system generally contains 3 main modules: feature extraction, modeling and scoring. The feature extraction modules employed in the speaker recognition evaluation [9] included feature computation (MFCC, LFCC, PLP, F0, etc.), voice activity detection (energy-based, phoneme-based, etc.), speech enhancement (spectral subtraction, Wiener filtering, etc.), and feature post-processing (feature warping, cepstral mean and variance normalization, etc.).

The modeling and scoring modules were often related. Different techniques were used in the evaluation. They can be divided into 4 mains groups:

- Gaussian mixture modeling (GMM) [12] first estimates a universal background model (UBM), which is then adapted to each client using maximum *a posteriori* (MAP). Scores are computed by estimating the log-likelihood ratio of the probe with regards to the client model and the UBM.

- In Gaussian super-vector (GSV) modeling [13], the mean vectors of the Gaussians are concatenated. Nuisance attribute projection (NAP) [14] is generally used for session compensation. The scores are computed using support vector machines (SVM).

- Inter-session variability (ISV) modeling [15] aims to estimate and eliminate the effects of the session variability. The scoring employed a linear approximation of the log-likelihood ratio.

- Total variability modeling (i-vector) [16] extracts a low-dimensional vector from each of the speech segments. Different i-vector post-processing types like whitening [17], length normalization [18], linear discriminant analysis (LDA) and within-class covariance normalization (WCCN) [19] were used. At scoring level, between i-vectors the cosine distance or probabilistic linear discriminant analysis (PLDA) [20, 21] was computed.

The best system in the evaluation, i. e., Alpineon (cf. table 2(a)) is based on the fusion of 9 different i-vector sub-systems, each with a different set of features. More details about this and the other speaker recognition systems can be found in [9].

[3] http://www.idiap.ch/dataset/mobio

[4] Note that after the evaluations some of the participants submitted corrected score files so that the entries in table 2 partially do not correspond to the results published in [9, 10].

Table 1: PARTITIONING OF THE MOBIO DATABASE. *This table details the number of clients and recordings of the training set, as well as the number of clients and enrollment recordings, and the number of probes for the development and the evaluation set, for the* male *and* female *protocols of the MOBIO database.*

	Training		Development				Evaluation			
			Enrollment		Probe		Enrollment		Probe	
	Clients	*Files*	*Clients*	*Files*	*Files*	*Scores*	*Clients*	*Files*	*Files*	*Scores*
male	37	7104	24	120	2520	60480	38	190	3990	151620
female	13	2496	18	90	1890	34020	20	100	2100	42000
Total	50	9600	42	210	4410	94500	58	290	6090	193620

3.2. Face Recognition

In the algorithms submitted to the face recognition evaluation [10], the first step of all participants was to align the faces using the provided hand-labeled eye positions. Afterward, different image normalization techniques [22, 23, 24] were used to reduce illumination effects. Various kinds of features like edge information (POEM) [25], Gabor features [26, 27], local binary patterns (LBP) [28], local phase quantization (LPQ) [29] and color information were extracted. The best single-feature based system, i. e., UC-HU (see table 2(b)) learned how to extract features using a convolutional neural network [30].

On top of these features, different kinds of face recognition systems were executed. Several algorithms computed histograms of various kinds and used histogram comparisons to compute scores. Other systems used principal component analysis (PCA) or linear discriminant analysis (LDA) to reduce feature dimensionality. Furthermore, partial least squares (PLS) classifiers or support vector machines (SVM) were trained to enroll models and compare them to probe features.

Additionally, some participants fused face recognition systems of different kinds. The best performing system in [10], i. e., UNILJ-ALP (cf. table 2(b)) was the multi-representation PCA, which fused in total 30 different face recognition sub-systems. For more detailed descriptions of this and all the other submitted face recognition systems, please refer to [10].

3.3. Fusion

To fuse different recognition systems, we take the well-known *linear logistic regression* approach, which has successfully been employed to combine heterogeneous speaker [31, 32] and face [33] authentication systems, as well as for bi-modal authentication [34].

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 $h_q(\boldsymbol{\mathcal{O}}_t, \boldsymbol{g}_i)$ between the current probe sample $\boldsymbol{\mathcal{O}}_t$ and a given client model $\boldsymbol{g}_i$. These scores are fused using a linear combination:

$$h_{\boldsymbol{\beta}}(\boldsymbol{\mathcal{O}}_t, \boldsymbol{g}_i) = \beta_0 + \sum_{q=1}^{Q} \beta_q h_q(\boldsymbol{\mathcal{O}}_t, \boldsymbol{g}_i) \qquad (1)$$

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 $\mathcal{X}_{\text{cli}}$ be the set of true client access trials, i. e., the set of those pairs $\boldsymbol{x} = \{\boldsymbol{\mathcal{O}}_t, \boldsymbol{g}_i\}$ where the identities of test sample $\boldsymbol{\mathcal{O}}_t$ and client $\boldsymbol{g}_i$ match. Let furthermore $\mathcal{X}_{\text{imp}}$ be the set of impostor trials, i. e., those pairs where the identities of $\boldsymbol{\mathcal{O}}_t$ and $\boldsymbol{g}_i$ differ. Then, the objective function to maximize is [35]:

$$\begin{aligned} L(\boldsymbol{\beta}) = &- \sum_{\boldsymbol{x} \in \mathcal{X}_{\text{imp}}} \log\left(1 + \exp\left(h_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{\beta})\right)\right) \\ &- \sum_{\boldsymbol{x} \in \mathcal{X}_{\text{cli}}} \log\left(1 + \exp\left(-h_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{\beta})\right)\right) \end{aligned} \qquad (2)$$

The maximum likelihood estimation procedure converges to a global maximum. In our work, this optimization is done using the conjugate-gradient algorithm [35].

One important fact of the fusion procedure is that the fused scores are already in form of log-likelihood ratios. Hence, after fusing scores from different systems, not only the verification accuracy is increased, but the scores can directly be used to present evidence in front of court.

4. Experiments

For several different speaker and face verification algorithms, scores for the MOBIO database are provided in the speaker and face evaluations [9, 10]. To evaluate multi-algorithm and bi-modal system fusion, we executed several experiments using these scores. All experiments are run using the open source software library Bob [36], and we provide both scripts[5] and data[6] for the scientific community affording reproducible research. The first set of experiments assessed the uni-modal multi-algorithm fusion, while the second set of tests fused algorithms of both data types.

4.1. Evaluation metrics

The metrics that we use to evaluate verification performance are based on the false acceptance rate (FAR) and the false rejection rate (FRR), which are calculated for

[5] http://pypi.python.org/pypi/xbob.paper.BTFS2013
[6] http://www.idiap.ch/dataset/mobio

Table 2: RESULTS FROM ICB. *These tables repeat[4] the results from [9] and [10],* ***ordered by EER on the development set of the*** **male** ***protocol*** *from the MOBIO database.*

(a) Speaker recognition

Id	System	*male*		*female*	
		EER	HTER	EER	HTER
S-1	Alpineon	**5.04**	**7.08**	**7.98**	**10.68**
S-2	L2F-EHU	7.89	8.14	11.01	13.59
S-3	Phonexia	9.60	10.78	8.36	14.18
S-4	GIAPSI	9.68	8.86	11.59	12.81
S-5	IDIAP	9.96	10.03	12.01	14.27
S-6	Mines-Telecom	10.20	9.11	11.43	11.63
S-7	L2F	10.60	11.05	13.48	14.73
S-8	EHU	11.31	10.06	17.94	19.51
S-9	CPqD	11.82	10.21	14.35	15.99
S-10	CTDA	12.74	19.40	19.47	22.64
S-11	RUN	13.73	12.13	13.39	14.09
S-12	ATVS	14.88	15.43	16.84	17.86

(b) Face recognition

Id	System	*male*		*female*	
		EER	HTER	EER	HTER
F-1	UNILJ-ALP	**1.71**	7.45	**2.75**	**10.46**
F-2	GRADIANT	3.14	9.52	5.38	12.27
F-3	UC-HU	3.49	**6.21**	4.71	10.83
F-4	CPqD	5.48	7.67	6.30	11.21
F-5	TUT	5.48	10.02	7.35	12.05
F-6	UTS	6.11	11.96	7.46	13.57
F-7	Idiap	6.63	10.29	6.24	12.51
F-8	CDTA	7.65	11.93	10.74	15.90
F-9	baseline	14.80	17.11	14.71	20.94

Table 3: UNI-MODAL FUSION. *These tables show the results of uni-modally fusing the N best speaker or face recognition systems from table 2.*

(a) Speaker recognition

Pool of classifiers	*male*		*female*	
	EER	HTER	EER	HTER
S-1	5.04	7.08	7.98	10.68
+ S-2	3.81	5.92	6.14	8.85
+ S-3	3.41	5.43	4.18	7.97
+ S-4	3.05	4.76	3.59	6.91
+ S-5	2.86	4.70	3.65	6.87
+ S-6	2.86	4.75	3.65	**6.73**
+ S-7	2.90	4.75	**3.54**	6.87
+ S-8	2.90	4.75	3.54	6.89
+ S-9	2.90	4.76	3.59	6.96
+ S-10	2.98	4.97	3.61	7.04
+ S-11	2.81	4.69	3.60	6.87
all	**2.78**	**4.63**	3.60	6.87

(b) Face recognition

Pool of classifiers	*male*		*female*	
	EER	HTER	EER	HTER
F-1	1.71	7.45	2.75	10.46
+ F-2	1.59	7.30	2.59	10.03
+ F-3	1.50	6.62	2.44	9.99
+ F-4	1.47	6.69	2.49	9.97
+ F-5	1.51	6.89	2.37	10.10
+ F-6	1.47	6.90	2.22	9.31
+ F-7	1.51	6.83	2.12	9.29
+ F-8	1.51	6.72	2.02	9.01
all	**1.39**	**6.27**	**1.97**	**8.47**

the development and evaluation sets independently. The definition of these rates depends on a *threshold* θ:

$$\begin{aligned} \mathrm{FAR}(\theta) &= \frac{|\{s_{\mathrm{imp}} \mid s_{\mathrm{imp}} \geq \theta\}|}{|\{s_{\mathrm{imp}}\}|} \\ \mathrm{FRR}(\theta) &= \frac{|\{s_{\mathrm{cli}} \mid s_{\mathrm{cli}} < \theta\}|}{|\{s_{\mathrm{cli}}\}|} \end{aligned} \tag{3}$$

Here, s_{cli} are client (true target) and s_{imp} impostor (non-target) scores, both of which might come from a single speaker or face recognition system, or might have been created by fusing scores of many systems using Eq. (1).

The first evaluation metric is based on the *equal error rate* (EER) on the development set and the *half total error rate* (HTER) on the evaluation set. Particularly, the optimal threshold θ^* is based on the EER of the development set, and the HTER is computed using this threshold:

$$\begin{aligned} \theta^* &= \arg\min_{\theta} |\mathrm{FAR}_{\mathrm{dev}}(\theta) - \mathrm{FRR}_{\mathrm{dev}}(\theta)| \\ \mathrm{EER} &= \frac{\mathrm{FAR}_{\mathrm{dev}}(\theta^*) + \mathrm{FRR}_{\mathrm{dev}}(\theta^*)}{2} \\ \mathrm{HTER} &= \frac{\mathrm{FAR}_{\mathrm{eval}}(\theta^*) + \mathrm{FRR}_{\mathrm{eval}}(\theta^*)}{2} \end{aligned} \tag{4}$$

The second type of evaluation is based on the detection error trade-off (DET) [37]. In this curve, the FRR is plotted against the FAR in normal deviate scale. In opposition to receiver operating characteristics (ROC), DET curves allow easy observation of system contrasts, especially for low FAR values. The DET curve of a system is linear when client and impostor score are Gaussian distributed, and with an angle of 45° the variances of the Gaussians are equal.

4.2. Uni-modal fusion

The final verification results of the evaluations are given in table 2, where the different systems are sorted according to the EER of the male protocol. Apparently, compared to the best speaker recognition system S-1, the best face recognition system F-1 has lower error rates on the development set, but similar evaluation set errors.

For both the speaker and face recognition systems, we assessed how the fusion of the N best systems improves performance, varying N from 1 to 12 for the speaker and to 9 for the face recognition systems. The results of these experiments can be found in table 3. Clearly, fusing scores from multiple algorithms improves performance. For speaker recognition, incorporating scores from the best two

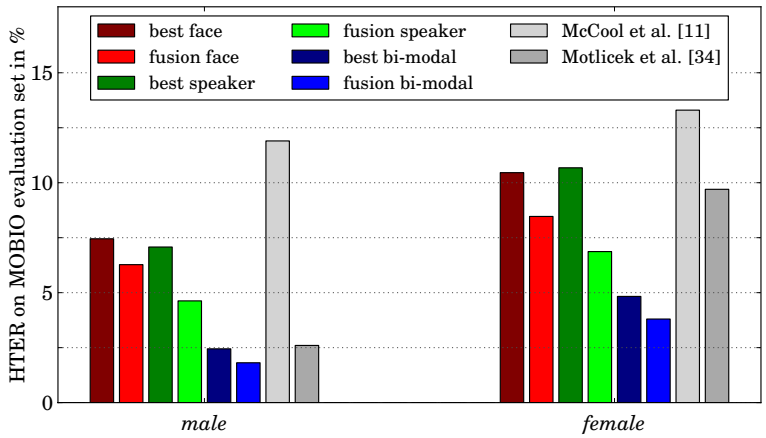


Figure 1: SYSTEM COMPARISON. *This figure shows the HTER on the evaluation set of MOBIO for the best speaker and face recognition system and their fusion, as well as the fusions of all speaker, all face and all recognition systems. Additionally, results of the bi-modal systems from [11] and [34] are included.*

systems already gains more than one percent of absolute error rate, for the female protocol even around two percent. Incorporating all speaker recognition systems results in a overall relative reduction of around 50% in the development set, and still 30% in the evaluation set.

For face recognition, a similar trend is obtainable, though in general there is a higher difference between development and evaluation set. The best face recognition system has already quite low error rates on the development set. Still, the fusion of all face recognition algorithms results in a relative reduction of around 20% - 30% for both protocols and both sets.

Note that the difference in error rates between the development and evaluation set is an artifact of the small data set and confirms previous findings [34].

4.3. Bi-modal fusion

As we showed in [34], fusion of speaker and face recognition systems can tremendously improve performance on the MOBIO database. A similar behavior is observed combining the speaker and face recognition systems provided by the participants of the evaluations. Two different configurations were evaluated:

4.3.1. Fusion of all speaker and face systems

In figure 1, the HTER results on the evaluation set are displayed for the best speaker and face recognition systems, as well as their bi-modal fusion, and the uni-modal and bi-modal fusion of all speaker and face systems. Clearly, uni-modal fusion gives improvements over the best uni-modal system. Bi-modally fusing the best speaker and the best face recognition system outperforms the uni-modal systems significantly, and the fusion of all systems gives by far the best results, which is 0.16 % EER and 1.78 % HTER for male and 0.16 % EER and 3.80 % HTER for female clients.

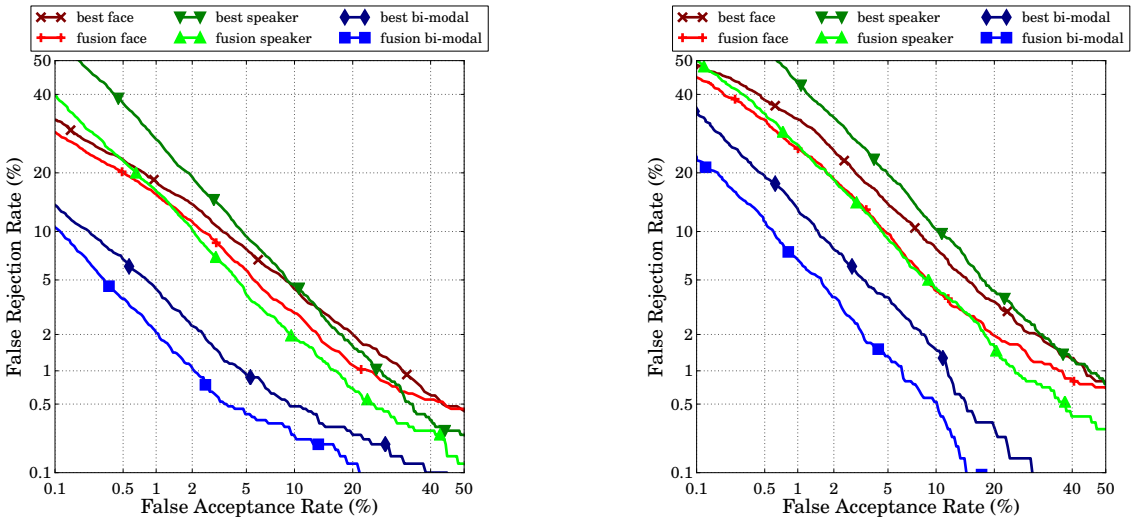
In figure 1, the results from [11, 34] are added. While [34] exploits bi-modal single-algorithm fusion, [11] fuses bi-modal multi-algorithm recognition systems using the sum rule. Clearly, both systems are outperformed by our bi-modal multi-algorithm fusion strategy of systems with various type of features that is based on linear logistic regression.

The DET curves, which are shown in figure 2, reveal similar trends. For different working points (different thresholds) the order of the systems is stable, ranging from the single uni-modal systems over the uni-modal fusions to the bi-modal fusions.

4.3.2. Optimal bi-modal fusion

In the second experiment we assessed, which of the submitted algorithms are best suited for fusion. Starting with the best performing system on the development set, i. e., the face recognition system F-1, we tested which other algorithm gets the highest improvement in EER on the development set, i. e., is most complementary to the F-1 system. After finding the speaker recognition systems S-1 (male) and S-5 (female) to be most suitable, we added the next best algorithm and so forth. Figure 3 shows the EER and the according HTER values of fusing the optimal set of systems. From left to right, the indicated system was added to the set of fused algorithms.

Apparently, the biggest gain is in fusing one face and one speaker recognition system, and further adding more systems improves performance only moderately. After fusing approximately 10 systems, which differ between the male and female protocols, performance on the development set settles. Adding more systems does not improve the EER on the development set any more, though



(a) *male* (b) *female*

Figure 2: DET CURVES. *This figure shows DET curves for uni-modal and bi-modal speaker and face recognition systems and their fusions on the evaluation set of the MOBIO database for the* male *and* female *protocols.*

the evaluation set performance varies slightly.

Note that speaker and face recognition systems contribute similarly. For the *male* protocol, the top 10 systems contain 6 face and 4 speaker recognition systems, while for *female* the top 10 comprise speaker and face recognition systems in an equal number.

5. Conclusions

In the present paper, we tested how the decision level fusion of several state-of-the-art speaker and face recognition algorithms can help to improve recognition performance. We used the 12 speaker and 9 face recognition systems that were submitted to the speaker and face recognition evaluations in mobile environments [9, 10]. We showed that the uni-modal fusion of speaker or face recognition systems is able to improve performance moderately, i. e., by approximately 35 % or 15 % relative gain, respectively. Already fusing the best speaker and the best face recognition algorithm improved recognition performance by more than 55 % compared to the best uni-modal system. By integrating more speaker and face recognition systems into the fusion process, this gain can be increased up to 70 %. The final 1.78 % HTER for the *male* protocol and 3.80 % HTER for the *female* protocol outperform previously published bi-modal fusion algorithms that used the same database with the same evaluation protocols.

These findings lead to the conclusion that multi-modal multi-algorithm fusion should be the future trends in biometric recognition. It can also help forensic applications, especially since the resulting fused scores are already in terms of log-likelihood ratios.

6. Acknowledgment

The research leading to the results presented in this paper has received funding from the European Community's Seventh Framework Program (FP7) under grant agreements 238803 (BBfor2) and from the Swiss National Science Foundation under the LOBI project.

7. References

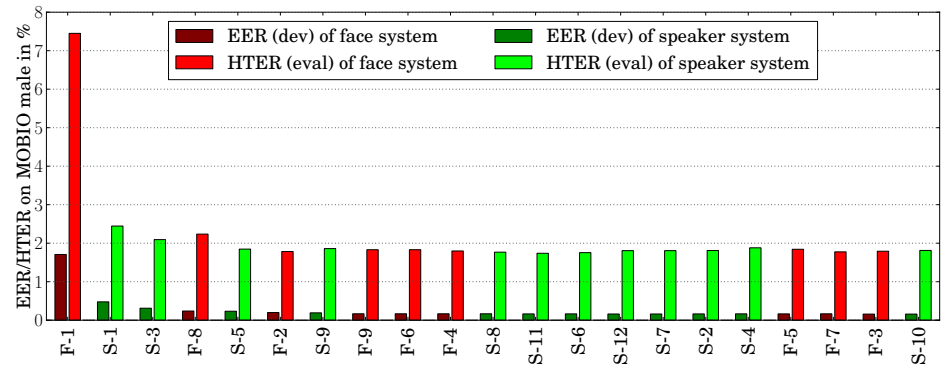
[1] V. Hautamäki, T. Kinnunen, M. Nosratighods, K.-A. Lee, B. Ma, and H. Li, "Approaching human listener accuracy with modern speaker verification," in *INTERSPEECH*, 2010, pp. 1473–1476.

[2] A.J. O'Toole, P.J. Phillips, F. Jiang, J. Ayyad, N. Penard, and H. Abdi, "Face recognition algorithms surpass humans matching faces over changes in illumination," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 1642–1646, 2007.

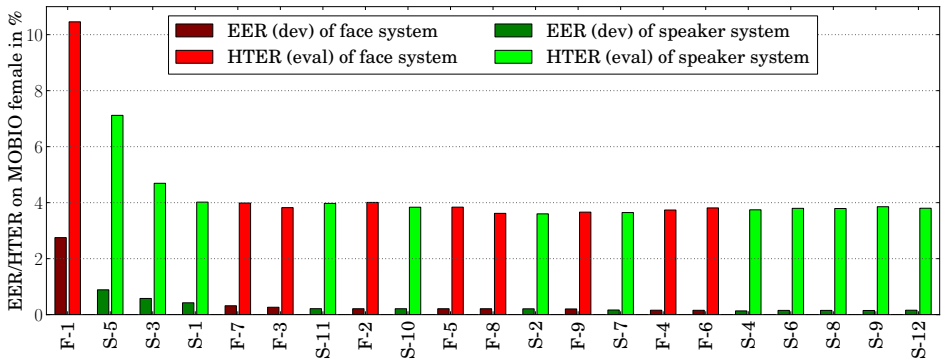
[3] J.P. Campbell Jr., "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[4] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *Journal of the Optical Society of America A*, vol. 4, no. 3, 1987.

[5] D.M. Blackburn, M. Bone, and P.J. Phillips, *Face Recognition Vendor Test 2000: Evaluation Report*, Storming Media, 2001.



(a) Pool of classifiers for *male*



(b) Pool of classifiers for *female*

Figure 3: OPTIMAL BI-MODAL FUSION. *This figure shows the improvements in EER and HTER of optimally fusing speaker and face recognition systems. From left to right, the indicated system is added to the set of fused systems, starting with the best system F-1.*

[6] O. Ribaux, S.J. Walsh, and P. Margot, "The contribution of forensic science to crime analysis and investigation: Forensic intelligence," *Forensic science international*, vol. 156, no. 2, pp. 171–181, 2006.

[7] D. Ramos-Castro, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "Likelihood ratio calibration in a transparent and testable forensic speaker recognition framework," in *Speaker and Language Recognition Workshop*. 2006, pp. 1–8, IEEE.

[8] S. Marcel et al., "On the results of the first mobile biometry (MOBIO) face and speaker verification evaluation," in *International Conference on Pattern Recognition*. 2010, pp. 210–225, Springer-Verlag.

[9] E. Khoury et al., "The 2013 speaker recognition evaluation in mobile environment," in *International Conference on Biometrics*, 2013.

[10] M. Günther et al., "The 2013 face recognition evaluation in mobile environment," in *International Conference on Biometrics*, 2013.

[11] C. McCool et al., "Bi-modal person recognition on a mobile phone: Using mobile phone data," in *IEEE International Conference on Multimedia and Expo, Workshop on Hot Topics in Mobile Multimedia*, 2012.

[12] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[13] W.M. Campbell, D.E. Sturim, D.A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006.

[14] A. Solomonoff, C. Quillen, and W.M. Campbell, "Channel compensation for SVM speaker recognition," in *Proceedings of Odyssey, Speaker and Language Recognition Workshop*, 2004, pp. 57–62.

[15] R. Vogt and S. Sridharan, "Explicit modelling of session variability for speaker verification," *Computer Speech & Language*, vol. 22, no. 1, pp. 17–38, 2008.

[16] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker

verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 788–798, 2011.

[17] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka, and N. Brümmer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2011, pp. 4832–4835.

[18] D. Garcia-Romero and C.Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *INTERSPEECH*, 2011, pp. 249–252.

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

[20] S.J.D. Prince and J.H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[21] L. El Shafey, C. McCool, R. Wallace, and S. Marcel, "A scalable formulation of probabilistic linear discriminant analysis: Applied to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1788–1794, 2013.

[22] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1635–1650, 2010.

[23] G. Heusch, Y. Rodriguez, and S. Marcel, "Local binary patterns as an image preprocessing for face authentication," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2006, pp. 9–14.

[24] N. Vu and A. Caplier, "Illumination-robust face recognition using retina modeling," in *IEEE International Conference on Image Processing*, 2009, pp. 3289–3292.

[25] N. Vu and A. Caplier, "Face recognition with patterns of oriented edge magnitudes," in *European Conference on Computer Vision*. 2010, pp. 313–326, Springer.

[26] L. Wiskott, J.-M. Fellous, N. Krüger, and C. v.d. Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Recognition and Artificial Intelligence*, vol. 19, pp. 775–779, 1997.

[27] M. Günther, D. Haufe, and R.P. Würtz, "Face recognition with disparity corrected Gabor phase differences," in *Artificial Neural Networks and Machine Learning*, 2012, pp. 411–418.

[28] T. Ahonen, A. Hadid, and M. Pietikainen, "Face recognition with local binary patterns," in *European Conference on Computer Vision*, 2004, vol. 3021, pp. 469–481.

[29] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkil, "Recognition of blurred faces using local phase quantization," in *International Conference on Pattern Recognition*, 2008, pp. 1–4.

[30] D.D. Cox and N. Pinto, "Beyond simple features: A large-scale feature search approach to unconstrained face recognition," in *IEEE International Conference on Automatic Face Gesture Recognition*, 2011, pp. 8–15.

[31] S. Pigeon, P. Druyts, and P. Verlinde, "Applying logistic regression to the fusion of the NIST'99 1-speaker submissions," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 237–248, 2000.

[32] N. Brümmer et al., "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Transactions on Speech, Audio and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[33] C. McCool and S. Marcel, "Parts-based face verification using local frequency bands," in *IEEE/IAPR Third International Conference on Advances in Biometrics*. 2009, pp. 259–268, Springer-Verlag.

[34] P. Motlicek, L. El Shafey, R. Wallace, C. McCool, and S. Marcel, "Bi-modal authentication in mobile environments using session variability modelling," in *International Conference on Pattern Recognition*, 2012, pp. 1100–1103.

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

[36] A. Anjos, L. El Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel, "Bob: a free signal processing and machine learning toolbox for researchers," in *ACM International Conference on Multimedia*, 2012, pp. 1449–1452.

[37] A.F. Martin, G.R. Doddington, T. Kamm, M. Ordowski, and M.A. Przybocki, "The DET curve in assessment of detection task performance," in *EUROSPEECH*, 1997.