Multimodal person authentication on a smartphone under realistic conditions

Andrew C. Morris[1a], Sabah Jassim[2b], Harin Sellahewa[2], Lorene Allano[3c], Johan Ehlers[2], Dalei Wu[1], Jacques Koreman[1], Sonia Garcia-Salicetti[3], Bao Ly-Van[3], Bernadette Dorizzi[3]

[1]Saarland University, FR 4.7 Phonetics, Bldg. C7.2, Box 151150, D-66041, Saarbrücken, Germany

[2]Department of Information Systems, University of Buckingham, Buckingham MK18 1EG, UK

[3]GET Instutut National des Télécommunications, 9 rue Charles Fourrier, 91011 Évry cedex, France

ABSTRACT

Verification of a person's identity by the combination of more than one biometric trait strongly increases the robustness of person authentication in real applications. This is particularly the case in applications involving signals of degraded quality, as for person authentication on mobile platforms. The context of mobility generates degradations of input signals due to the variety of environments encountered (ambient noise, lighting variations, etc.), while the sensors' lower quality further contributes to decrease in system performance. Our aim in this work is to combine traits from the three biometric modalities of speech, face and handwritten signature in a concrete application, performing non intrusive biometric verification on a personal mobile device (smartphone/PDA).

Most available biometric databases have been acquired in more or less controlled environments, which makes it difficult to predict performance in a real application. Our experiments are performed on a database acquired on a PDA as part of the SecurePhone project (IST-2002-506883 project "Secure Contracts Signed by Mobile Phone"). This database contains 60 virtual subjects balanced in gender and age. Virtual subjects are obtained by coupling audio-visual signals from real English speaking subjects with signatures from other subjects captured on the touch screen of the PDA. Video data for the PDA database was recorded in 2 recording sessions separated by at least one week. Each session comprises 4 acquisition conditions: 2 indoor and 2 outdoor recordings (with in each case, a good and a degraded quality recording). Handwritten signatures were captured in one session in realistic conditions. Different scenarios of matching between training and test conditions are tested to measure the resistance of various fusion systems to different types of variability and different amounts of enrolment data.

Keywords: multimodal person authentication, smartphone implementation, Gaussian mixture model, biometrics, verification

1 INTRODUCTION

This article describes a multi-modal user authentication system which has been implemented on a PDA as one of the two main objectives of the SecurePhone project. The other objective is to implement a prototype system on an existing PDA to enable the secure exchange of written and spoken documents between mobile devices. When the document has

[a] Email {morris, daleiwu, koreman}@coli.uni-saarland.de; Tel. +49 681 302 4696; Fax. +49 681 302 4684

[b] Email {saba.jassim, harin.sellahewa, johan-hendrik.ehlers}@buckingham.ac.uk

[c] Email {lorene.allano, sonia.salicetti, bao.ly_van, bernadette.dorizzi }@int-evry.fr

been agreed by two PDA users it can then be e-signed to form a legally binding contract. For increased security, as well as ease of use, biometric authorisation (rather than a PIN) is then requested to permit e-signing. Authorisation combines the three modalities of voice, face and signature. These familiar biometrics can all be acquired on a standard PDA and for this reason they also have high user acceptance. Although none of these biometrics can provide very accurate verification on their own, in combination they should be able to provide a high enough degree of security for a wide range of transactions. In this article we describe how this authentication system was designed to provide maximum security and ease of use within the constraints imposed by a presently available PDA.

In order to provide maximum protection for the client's biometric profile, with present PDA technology this means that all storage and processing of biometric data must be done on the SIM card within the PDA, and not on the PDA main processor. This places severe constraints both on the total size of the client's biometric profile and on the amount of computation which can be used in verification. Fortunately the memory of a standard SIM card is large enough to hold the required combined biometric profiles. Furthermore, almost all of the computational load in biometrics processing is due to feature extraction, and this can be done on the PDA main processor, which is much faster than the SIM.

In Section 2 we briefly describe the dedicated PDA database (PDAtabase) and test procedure which was created to facilitate development of this system. Section 3 presents each of the main features in the working system. Section 4 explains the final system design and presents performance results on the PDAtabase. Section 5 shows how performance is affected if any of the configuration parameters is changed. This is followed by a discussion of outstanding development issues and a conclusion.

2 PDATABASE

From the security point of view it would be an advantage for a speaker recognition system to be text independent. However, for the present application all biometric profiles must be stored on the PDA and the model size required for a text independent system, which must capture every kind of phonetic variation, would be too great. It was therefore decided from the beginning that the system should be text dependent. While a server based system would have the advantage of greater computational resources, it would also have the disadvantage of having to compress and transmit biometric data before it was processed. It would furthermore have to face legal issues concerning data privacy which the present P2P system avoids because no biometric data ever needs to leave the SIM card, where it is secure.

For the purpose of initial system development we made use of the BANCA audio-visual database [11] together with the BIOMET on-line signature database [7]. After some further experimentation with the CSLU speaker recognition database [3] (which is better suited for testing voice authentication with text dependent prompts), and selecting a suitable PDA (the Qtek2020) [14], a database tailored exactly to the requirements of this project was recorded directly on this PDA [10]. This database was specifically designed for the development of multi-modal fixed prompt based user authentication, using voice, face and handwritten signature.

Voice and face data. Data was collected for 30 male and 30 female subjects from three age groups (under 30, 30-45, over 45) and consists of 5-digit prompts, 10-digit prompts and short phrases (six examples of each), recorded in quiet and noisy environments, both inside and outside. 5-digit and phrase prompts were taken from [3]. Data was recorded in two sessions separated by at least one week. Like BANCA, subjects were divided into three groups: one (g0) for UBM (Universal Background Model) training, and two other groups, g1 and g2. For any given FA/FR (False Acceptance to False Rejection) cost ratio, thresholds can then be optimised on g1 and evaluated on g2, and vice versa. Voice and face data was sampled at the maximum possible rate of 44 KHz audio and around 19.6 640x480 pixel frames per second (frame rate was not exactly constant). While 44 KHz data could be useful for some types of voice analysis, for all of our tests this data was downsampled to 22 KHz because for the models we use this reduces processing time without compromising verification accuracy.

Signature data. Signature data was captured from the PDA touch screen at 100 (x,y) samples per second. Unlike the signature data in BIOMET, which was recorded on a specialised writing tablet, the Qtek does not capture either pen pressure or pen angles. These extra measurements would certainly improve verification precision if available. The subjects whose signatures were recorded were different from those whose face and voice were recorded, but as signatures are independent from face or voice, a virtual coupling between these two sets of people could be used with no problem.

Test protocol. As with BANCA, a set of fixed test protocols was established. Unlike BANCA, in which each standard protocol tests a different combination of training and test conditions, the PDAtabase test protocol provides a large number of design choices which have too many combinations to permit the definition of any set of standard tests. In order to make it practical to test any one of this large number of possible configurations, a script was created which enables the user to plug in user routines to perform each core function and then run any given configuration test automatically, completely hiding the processes of training and test set selection, match scores generation, scores fusion and performance evaluation from the user. The database, protocol documentation and test procedure are publicly available at cost price.

Four different choices of training/test data are summarised in Table 1 [10]. Dividing recording of voice and face data into well separated sessions can improve system performance. However, in this paper we consider only data divisions 1 and 2 because it would not be acceptable to ask a client to enrol in two separate sessions.

Data division D	Training		Testing	
	Session 1	Session 2	Session 1	Session 2
1	I1 I2	none	none	I1 I2 O1 O2
2	I1 I2 O1 O2	none	none	I1 I2 O1 O2
3	I1 I2	none	O1 O2	I1 I2
4	I1 I2	O1 O2	O1 O2	I1 I2

Table 1. Choice of data divisions for training and testing models for voice and face.
"I1, I2" = (inside, light, clean), (inside, dark noisy). "O1, O2" = (outside, light, noisy), (outside, dark noisy).
Of the 20 chimera signatures for each subject, 8 are used for training with D1 and D3 and 16 with D2 and D4.
Tests 1 & 2 look at effect of increasing amount of training data within one enrolment session.
Tests 3 & 4 look at effect of increasing amount of training data across two enrolment sessions.

Switches used to specify different test configurations in the automatic test procedure are summarised in Table 2 [10].

Tag	value range	Meaning
M	1,2,3	Biometric mode = voice, face, signature
T	1,2,3	Prompt type = 5-digit, 10-digit, phrase
D	1,2,3,4	Train/test data selection (see Table 1)
G	0,1	UBM for model initialisation is gender indep./dep.
S	0,1	UBM for score normalisation is gender indep./dep.
H	0,1	Thresholds are gender indep./dep.
U	0,1	UBM not used/used for model init. & score norm.

Table 2. Key to switches used for scores generation
Each of these flags, and the value assigned to it, is used to construct a test name identifier, e.g. M1.T2.D4.S0.H1.U1

3 AUTHENTICATION SYSTEM

3.1 Feature processing

Voice. The PDA is set to sample speech at 22 KHz. Voice features use 19 Mel-frequency cepstral coefficients (MFCC, without c0), with cepstral mean subtraction (CMS) to remove convolutive noise, and non-speech removal to reduce uninformative data. First order time difference features are then added [12]. All processing is online, so that feature processing can start before the utterance has been completed. As speech feature processing is the major computational bottleneck, MFCC generation was coded for both 16 bit and 32 bit fixed point arithmetic. Compared with floating point emulation, the 16 bit code speeded up calculation by a factor of 3.5 with only 3% relative error degradation. The 32 bit code gave a factor of 1.5 speed up with no loss in performance. This 32 bit code is now in use on the PDA.

Face. The face is represented by low-low Haar wavelet filters with lighting normalisation by histogram equalization [15]. As the user is expected to locate their face within the box displayed on the screen, and the location of facial features is computationally expensive, no image registration is applied. The wavelet features used are relatively insensitive to small variations in face position. Only a random selection of 10 face frames from each video is used. Other methods for face feature processing are discussed in [15].

Signature. Signature data is captured from the PDA touch screen at 100 (x,y) samples per second. This sequence of 2 dimensional data is then processed to give a sequence of 19 dimensional feature vectors, including velocity, acceleration, curvature, direction and several other features [5]. The glass touch screen is not an ideal surface for writing on. PDAtabase tests showed that signatures obtained in this way could give good verification accuracy, but not as good as signatures obtained from a dedicated writing tablet which also measures pen pressure and two pen angles [7].

3.2 Data modelling

All three of the biometrics modalities selected are more effectively represented by statistical data models than simple templates. While state of the art models differ between modalities, we have found that (diagonal covariance) Gaussian mixture models (GMM) [12][6] (Eq.2), used together with a GMM universal background model (UBM), give performance which is close to state of the art for all three modalities. While this is the model of choice for voice based authentication, the high performance which this model also gave for face and signature verification was unexpected.

As in [12], for each data frame, *x*, the client likelihood is given by (Eq.2). The log of the client likelihood for the whole utterance *X* of *T* frames is given by (Eq.3) and the match score is given by (Eq.4), the log of the ratio of the client likelihood to the impostor likelihood.

$$N(\vec{x},\vec{\mu},\Sigma) = \exp\left(-0.5(\vec{s}-\vec{\mu})'\Sigma^{-1}(\vec{s}-\vec{\mu})\right) \Big/ \sqrt{(2\pi)^d|\Sigma|} \tag{1}$$

$$p(\vec{x}|client) \cong \sum_i w_i N(\vec{x},\vec{\mu}_i,\Sigma_i), \quad f_{client}(\vec{x}) = \log p(\vec{x}|client) \tag{2}$$

$$s_{client} = \log p(\vec{X}|client) = \sum_t f_{client}(\vec{x}_t) \Big/ T \tag{3}$$

$$s = s_{client} - s_{impostor} \tag{4}$$

The reason for the success of the GMM with UBM model for all three modalities is probably because in each case the amount of enrolment data available for model training is very restricted. The GMM with MAP adaptive training (updating the Gaussian means only) from a UBM [13][9] is well suited to small amounts of training data. The UBM serves two purposes. It is used to initialise the client model before adaptive training with the enrolment data, and it is also used as a universal impostor model for score normalisation (the score used is proportional to the logarithm of the ratio of the posterior client probability to the posterior impostor probability). In any case, all three modalities on the PDA presently use a GMM to model biometric data features. Models were trained using the Torch machine learning API [4] (see Appendix B for GMM tuning parameters). A UBM, pre-trained on data from one prompt from g0, the set of UBM subjects, is installed both on the PC where enrolment takes place, and on the SIM card. Enrolment comprises 8 simulated client accesses, during which time the lighting and background noise conditions are varied to reflect the range of conditions expected during use. After biometric features have been extracted from this data, they are used to train a client model for each modality on the PC, which is then installed on the client's SIM card by the service provider [14].

Voice, signature and face models presently require (23.0, 2.9, 11.6) Kb respectively, totalling 37.5 Kb. This is rather large compared to the size of a set of iris or fingerprint templates (about 1.0 Kb), but this is the price we must pay for making use of highly variable but user friendly biometrics.

3.3 Score normalisation

Apart from subtracting the UBM log likelihood from the client log likelihood, no other form of score normalisation is currently applied. Several normalisations were been tested during the course of development (including Z-norm, Min-

Max and conversion from LLR to posterior client probability) [1], but none made a very great effect on verification precision.

3.4 Score fusion

Tests showed that fusion by the concatenation of voice and face features led to substantially lower performance than voice verification alone. In any case, as signatures cannot usefully be time aligned with video recordings, combination of all three modalities must use some form of late fusion. In late fusion the biometric data from each modality is processed to produce a measure of the fit of the data to the client profile which we refer to here as the "score" for each modality. These scores are then combined into a single score and the claimed identity is accepted if this combined score is above some preset threshold.

After testing several fusion methods suited to the limited computing capability of the PDA, the method selected for implementation was GMM based fusion [1] [8]. In this model two GMMs (Eq.5) installed (together with the client model and UBM) in the SIM estimate client and impostor likelihoods from the joint score vector, from which a final log likelihood ratio score is obtained. Optimum fusion GMMs used 4 diagonal covariance Gaussians. GMMs were trained by the usual maximum likelihood rule because in this case no advantage was found using MAP adaptive training.

$$\vec{s} \equiv (s_1, s_2, s_3)$$

$$p(\vec{s}|client) \equiv \sum_i w_i N(\vec{s}, \vec{\mu}_i, \Sigma_i), \quad f_{client}(\vec{s}) \equiv \log p(\vec{s}|client) \tag{5}$$

$$s_{fused} = \log \frac{p(\vec{s}|client)}{p(\vec{s}|impostor)} \cong f_{client}(\vec{s}) - f_{impostor}(\vec{s}) \tag{6}$$

3.5 Acceptance decision

A claimant is then accepted if the fused client-match score exceeds a given threshold. Here the match score is the log of the ratio of the client likelihood to the impostor likelihood, where the impostor likelihood is modelled by the UBM [12].

$$s_{fused} > \theta \tag{7}$$

The threshold, θ, is set to minimise the DCF (detection cost function) for a given effective cost ratio, R. If $FAR(\varphi)$ and $FRR(\varphi)$ are the FAR and FRR values which result for a given threshold φ, for some given development test set, g1, then the a-posteriori acceptance threshold is determined as follows.

$$\theta = \arg\min_{\varphi} DCF(\varphi) = \arg\max_{\varphi} (FAR(\varphi) + R \cdot FRR(\varphi)) / (1 + R) \tag{8}$$

This is then used as an a-priori threshold on the evaluation test set, g2 (and vice versa). The optimal value to use for R is not well defined as it depends on both the FA/FR cost ratio (which is application dependent), and the client prior probability (which is unknown and should change in the advent of suspicious behaviour). As in BANCA, we generate test results using 3 different values for R, (0.1, 1.0 and 10.0).

4 OPTIMAL CONFIGURATION

Prior to implementation on the PDA, the authentication system was first tuned on a desktop workstation to give optimum performance on the PDAtabase. Performance statistics reported are the conventional EER (Equal Error Rate) and the less conventional AvWER (the Average of the Weighted Error Rate for the three standard values of the effective cost ratio). The AvWER score should be low if all 3 WER scores are low, i.e. when the error rate is not too sensitive to small changes in the value used for the acceptance threshold. All results reported are averaged over separate tests for six different examples of a given prompt type. The system referred to below as "selected" is that which has been implemented on the PDA. It was selected as the configuration giving the best verification performance, but subject to certain practical constraints. For example, performance could have been improved significantly using T=2 (10-digit

prompts) instead of T=1 (5-digits). However, this would have doubled the preprocessing time and required more effort to speak. Selected face and voice configurations use T1.D2.G0.S0.H0.U1 (c.f. Table 2). Signature uses D1.G0.S0.H0.U1.

Modality	Config.	EER	AvWER
Voice	Selected	6.12	4.24
	U=0	30.44	15.58
	G=1	7.99	5.05
	S=1	6.45	9.68
	H=1	5.90	4.45
	D=1	16.21	9.62
	T=2	3.38	2.38
	NG=50	6.52	4.57
	NG=150	6.32	4.75
Face	Selected	28.57	15.62
	U=0	37.13	19.36
	G=1	46.63	20.25
	S=1	48.55	21.76
	H=1	29.03	16.41
	NG=3	30.55	16.86
	NG=5	29.80	16.38
Signature	Selected	6.19	6.12
	U=0	29.28	17.83
	G=1	18.06	11.77
	S=1	21.71	11.07
	H=1	7.48	6.77
	NG=50	10.83	7.14
	NG=150	6.36	5.76

Table 3. Baseline performance vs. performance if any *one* configuration variable is changed (NG = number of Gaussians)

From Table 3 we can observe that the use of a UBM (U=1) was crucial to obtaining satisfactory performance with all three modalities. This reflects the fact that the amount of enrolment data available for client model training is very limited for voice and signature modalities. Notice how performance falls also sharply for voice mode if D=1 instead of 2 (i.e. if the number of training utterances is reduced from 4 to 2). In the case of face data, which currently uses static face recognition only, it is rather the intra subject data variety which is limited. Any form of gender dependence appears to be harmful. With unlimited training data gender dependence would normally improve performance. The number of Gaussians used is not far from optimal for all three modalities. This is just as well because there is no more storage space on the SIM card. If more training data was available, the optimum number of Gaussians might be greater.

5 PDATABASE TEST RESULTS

Table 4 shows PDAtabase performance results for the selected system, for each modality alone, for score fusion from each pair of modalities, and for score fusion from all three modalities. DET curves and threshold sensitivity plots for the selected system are given in Appendix A. Performance results for each of the 6 prompts tested are given in Appendix C.

	EER	AvWER	FAR			FRR		
			R=0.0	R=1.0	R=10.0	R=0.0	R=1.0	R=10.0
Voice	6.12	4.24	19.10	4.81	0.86	2.08	8.33	19.10
Face	28.57	15.62	93.77	26.44	1.18	1.16	30.44	85.53
Signature	6.19	6.12	13.61	6.94	4.31	2.78	4.86	52.78
V+F	4.78	3.15	14.49	4.31	0.81	1.27	5.56	14.58
V+S	1.13	1.03	3.03	1.37	0.30	0.93	1.50	2.89
S+F	4.83	3.75	12.20	5.23	2.94	1.74	4.17	13.19
V+F+S	0.85	1.06	2.15	1.90	0.39	0.81	1.16	3.94

Table 4. 5-digit EER, average WER, FAR and FRR % scores (for 3 typical values of the cost ratio, R) obtained with the PDAtabase. Scores were obtained using a threshold optimised for data from one set of speakers while testing on another set.

While the selected system uses a 5-digit prompt, Table 5 below shows that performance could be significantly increased if it was possible to use a 10-digit prompt.

	EER	AvWER	FAR			FRR		
			R=0.0	R=1.0	R=10.0	R=0.0	R=1.0	R=10.0
Voice	3.38	2.38	9.42	2.62	0.62	1.27	4.63	10.42
Face	29.87	16.24	94.33	26.11	0.97	1.62	33.91	85.42
Signature	6.19	6.12	13.61	6.94	4.31	2.78	4.86	52.78
V+F	3.01	2.40	7.27	2.38	0.72	1.27	5.32	9.61
V+S	0.74	0.81	1.60	0.81	0.42	0.58	1.04	4.86
S+F	4.74	3.37	13.66	5.05	0.69	1.50	5.79	16.09
V+F+S	0.56	0.73	0.97	0.79	0.44	0.69	1.04	1.85

Table 5. 10-digit EER, average WER, FAR and FRR % scores (for 3 typical values of the cost ratio, R) obtained with the PDAtabase. Scores were obtained using a threshold optimised for data from one set of speakers while testing on another set.

6 DISCUSSION

The PDAtabase test results presented indicate that the level of verification accuracy attainable with an easy to use system is around 1% EER. The marked performance improvements resulting from scores fusion also underlines the fact that the three modalities chosen are highly complementary. We have seen that performance is very sensitive to the amount and variety of enrolment data available for client model training and for this reason we have decided that it would be worth extending the number of repetitions during enrolment from 4 to 8, while varying the background noise and lighting conditions.

Initial field test. With this enrolment we have made a (very) preliminary field test in which just one male enrolled and made 5 accesses, and 3 other males and one dummy (face=floor, voice=silence, signature=circle) each made 5 impostor accesses, with all processing taking place on the PDA. The result of this trial is displayed in Appendix D. This shows that, while the client and impostor fused scores are approximately as well separated in the field test as they were in the PDAtabase tests, the distribution of client and impostor scores is very different in the field tests (for voice and face modes, but not for signature) to the distribution in PDAtabase tests, with the result that the acceptance threshold optimised for the PDAtabase test is totally unsuitable for the field tests. This can be explained by a difference in lighting and acoustic noise conditions during database recording and during field tests. As the fusion GMMs only need to be trained once and for all, a solution to this mismatch problem could possibly be obtained by gathering a large set of client and impostor scores directly from the PDA, under many varied conditions, after which the fusion GMMs could be retrained and an optimal threshold could be set to best separate clients from impostors in this development set.

Integerisation. For security reasons mentioned earlier, both scores calculation for each modality and scores fusion should run on the SIM card where the client's biometric profile is to be securely stored and the accept/reject decision is to be made. The SIM card we have selected provides only 16-bit integer arithmetic operations, no fixed-point arithmetic and no functions such as log or exp. While a suitable verification procedure has been implemented on the SIM, until now these calculations have all been done on the PDA main processor. This delay is partly due to the fact that we are still running experiments to improve verification accuracy on the easily programmable PDA, and partly because of the unexpected length of time it is taking us to obtain information enabling us to program the required SIM operations. We do not yet know the effect which 16-bit integer calculations will have on computational speed or verification accuracy.

Security level attainable. If a photograph of the owner's face and signature plus an audio recording of their reading the fixed prompt was obtained, then successful impostorisation may be possible. While a liveness test we have proposed, based on correlation between mouth opening and speech energy [2] should counter this possibility, the computational load would be too great for the present PDA. An alternative solution would be to use random prompts with text independent verification, but the PDA would not be able to store the required large client voice model. However, for any type of transaction for which the traditional level of security given by a PIN or signature is acceptable, the security provided by the SecurePhone should be sufficient. For higher security applications it would not be so hard to add special purpose hardware to enable further modalities, such as iris or fingerprint, but this would risk alienating the casual user.

7 CONCLUSION

Our PDAtabase test results have shown that, using state-of-the-art verification techniques, the combination of the non-intrusive biometrics of voice, face and signature can achieve a level of authorisation accuracy on a mobile device which should be acceptable for the wide range of applications which is normally secured by a PIN or signature. This will permit multi-modal biometric authorization to give access to an e-signature facility which will enable legally binding contracts to be signed after they have been interactively modified and agreed in a mobile environment. With its short enrolment procedure, ease of use (no PIN or password to remember) and very high confidence that private biometric data cannot be accessed or misused by anyone, the system should expect very high user acceptance. While it remains to carry out more extensive field tests with verification implemented on the PDA SIM card, all the indications are that this system will very soon open up a new way of performing business transactions while on the move which is widely acceptable and advantageous to both the user and the service provider.

AKNOWLEDGEMENTS

This work was supported by the EC SecurePhone project IST-2002-506883.

APPENDIX A. DET CURVES AND THRESHOLD SENSITIVITY PLOTS

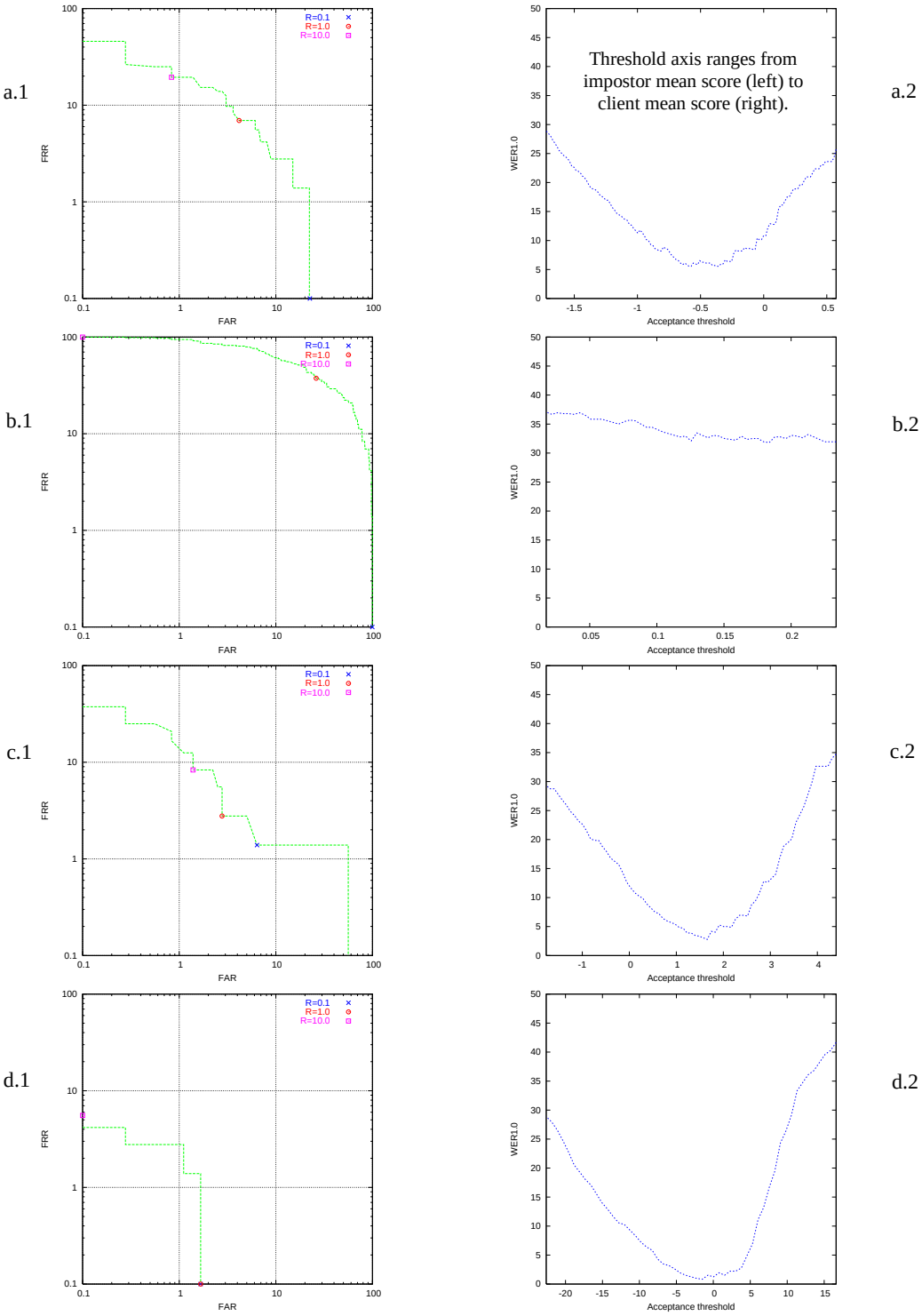

Fig.A.1. DET curves (left) and corresponding HTER (Half Total Error Rate) sensitivity plots (right), for the 5-digit prompt “28376” (test group g1). **a**) voice (top), **b**) face (upper middle), **c**) signature (lower middle) and **d**) fusion (bottom).

APPENDIX B. GMM TRAINING PARAMETERS

	Voice	Face	Signature	Fusion
Number of Gaussians	100	4	100	4
Number of features	38	120	19	3
Variance floor factor	0.6	12.0	0.063	0.016
Min. Gaussian weight	0.063	0.063	2.0	0.125
Weight on prior PDF	0.2	0.8	0.0	-
Max. N^o k-means iterations	1	1	1	5
Max. N^o EM iterations	5	5	5	10
Use UBM?	yes	yes	yes	no

Table B.1 GMM training parameters tuned for each modality and for fusion

Note that during client model training by MAP adaptation, initialized equal to the UBM, only the Gaussian means are updated. All of the Gaussian variances and mix weights are therefore shared between GMM and UBM, so these do not have to be stored twice. This means also that when the weight on the prior PDF is zero, adaptive training has the effect of using the UBM mix weights and variances for the client model, while the Gaussian means are trained by ML.

APPENDIX C. PDATABASE RESULTS SUMMARY FOR SCORES FUSION

Results for SecurePhone PDA verification test: fused_scores.all3a

	WER			FAR			FRR			Threshold		
G	EER	0.10	1.00	10.00	0.10	1.00	10.00	0.10	1.00	10.00	0.10	1.00	10.00
prompt numbers_01, gender = both													
1	1.39	1.49	1.94	2.40	2.50	2.50	2.50	1.39	1.39	1.39	0.04	0.04	0.04
2	0.28	0.10	2.08	0.38	1.11	0.00	0.00	0.00	4.17	4.17	-0.84	2.32	2.32
0	0.83	0.80	2.01	1.39	1.81	1.25	1.25	0.69	2.78	2.78			
prompt numbers_02, gender = both													
1	1.39	2.58	1.67	0.51	0.56	0.56	0.28	2.78	2.78	2.78	1.39	1.39	1.87
2	0.23	0.18	0.97	0.51	1.94	1.94	0.00	0.00	0.00	5.56	-1.07	-1.07	2.98
0	0.81	1.38	1.32	0.51	1.25	1.25	0.14	1.39	1.39	4.17			
prompt numbers_03, gender = both													
1	0.93	0.28	0.56	0.63	3.06	1.11	0.28	0.00	0.00	4.17	-3.09	-0.15	2.79
2	1.39	1.46	1.81	0.88	2.22	2.22	0.56	1.39	1.39	4.17	-0.29	-0.29	1.71
0	1.16	0.87	1.18	0.76	2.64	1.67	0.42	0.69	0.69	4.17			
prompt numbers_04, gender = both													
1	1.39	1.36	1.25	0.51	1.11	1.11	0.28	1.39	1.39	2.78	0.18	0.18	1.13
2	0.25	0.28	1.53	0.88	3.06	3.06	0.00	0.00	0.00	9.72	-1.49	-1.49	3.79
0	0.82	0.82	1.39	0.69	2.08	2.08	0.14	0.69	0.69	6.25			
prompt numbers_05, gender = both													
1	1.16	1.31	0.97	0.63	0.56	0.56	0.56	1.39	1.39	1.39	-0.01	-0.01	-0.01
2	0.28	0.56	3.06	0.76	6.11	6.11	0.28	0.00	0.00	5.56	-2.20	-2.20	2.02
0	0.72	0.93	2.01	0.69	3.33	3.33	0.42	0.69	0.69	3.47			
prompt numbers_06, gender = both													
1	0.83	1.31	0.97	0.13	0.56	0.56	0.00	1.39	1.39	1.39	0.58	0.58	2.29
2	0.69	0.28	1.53	0.38	3.06	3.06	0.00	0.00	0.00	4.17	-1.20	-1.20	2.12
0	0.76	0.80	1.25	0.25	1.81	1.81	0.00	0.69	0.69	2.78			
average over all prompts													
1	1.18	1.39	1.23	0.80	1.39	1.06	0.65	1.39	1.39	2.31			
2	0.52	0.48	1.83	0.63	2.92	2.73	0.14	0.23	0.93	5.56			
@0	0.85	0.93	1.53	0.72	2.15	1.90	0.39	0.81	1.16	3.94			

Table C.1 PDAtabase results summary for scores fusion (as produced by the automatic test script).

For each subtest in Table C.1, one row of statistics is given for g1 (using thresholds minimising WER for g2), and row for g2 (using thresholds minimising WER for g1), and one for the average between these. WER, FAR and FRR % are given for the 3 common values of the effective cost ratio, *R*. Results shown are for the six 5-digit prompts, for fusion of "selected" unimodal systems in Sec.4.

APPENDIX D. PRELIMINARY FIELD TEST RESULTS

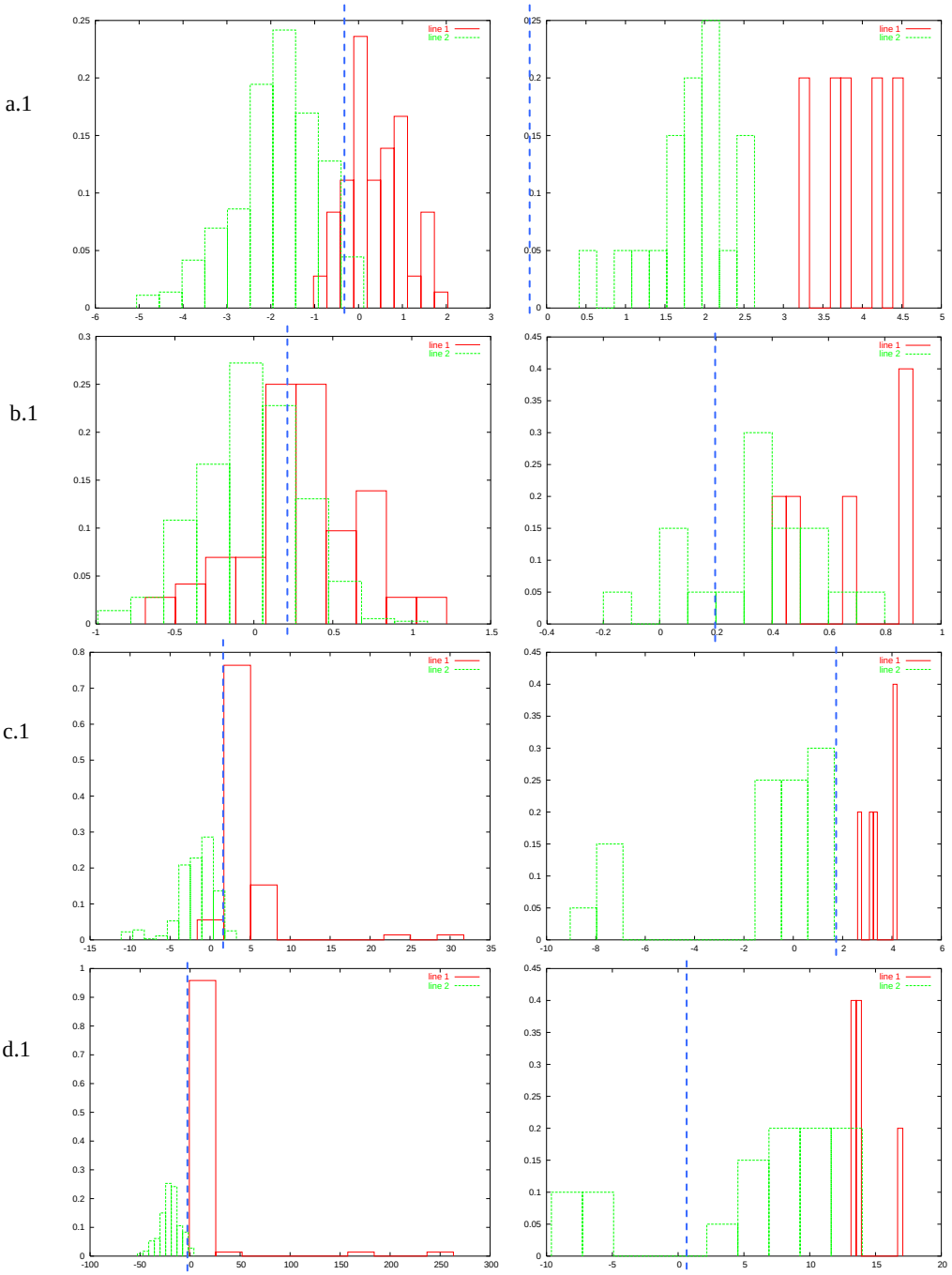

Fig.D.1. Client and impostor LLR scores on PDAtabase (left) and in (very limited) initial field trials (right). **a**) voice (top), **b**) face (upper middle), **c**) signature (lower middle) and **d**) fused (bottom). Vertical line shows optimal position of acceptance threshold determined for unimodal PDAtabase tests. The a-priori optimal threshold for PDAtabase (for R=1.0) was (-0.31, 0.19, 1.59, 0.51)

REFERENCES

[1] Allano, L., Morris, A.C., Sellahewa, H., Garcia-Salicetti, S., Koreman, J., Jassim, S., Ly-Van, B., Wu, D. & Dorizzi, B., "Non intrusive multi-biometrics on a mobile device: a comparison of fusion techniques", Proc. SPIE Conference on Biometric Techniques for Human Identification III, Orlando (in press).

[2] Bredin, H., Miguel, A., Witten, I.H. & Chollet, G., "Detecting replay attacks in audiovisual identity verification", Proc. ICASSP 2006 (in print).

[3] Cole, R., Noel, M. & Noel, V., "The CSLU Speaker Recognition Corpus", Proc. ICSLP, Sydney, pp.3167-3170, 1998.

[4] Collobert, R., Bengio, S. & Mariéthoz, J., "Torch: a modular machine learning software library", Technical Report IDIAP-RR 02-46, 2002.

[5] Dolfing, J.G.A., "Handwriting recognition and verification, a Hidden Markov approach", Ph.D. thesis, Philips Electronics N.V., 1998.

[6] Duda, O., Hart, P.E. & Stork, D.G., Pattern classification, Wiley, 2001.

[7] Garcia-Salicetti, S., Beumier, C., Chollet, G., Dorizzi, B., Leroux-Les Jardins, J., Lunter, J., Ni, Y. & Petrovska-Delacretaz, D., "BIOMET: a Multimodal Person Authentication Database Including Face, Voice, Fingerprint, Hand and Signature Modalities", Proc. 4th Conf. on AVBPA, pp. 845-853, Guildford, UK, July 2003.

[8] Koreman, J., Morris, A.C., Wu, D., Jassim, S., Sellahewa, H., Ehlers, J., Chollet, G., Aversano, G., Bredin, H. Garcia-Salicetti, S., Allano, L. Ly Van, B. & Dorizzi, B., " multi-modal biometric authentication on the SecurePhone PDA", Proc. MMUA, 2006 (in press).

[9] Mariéthoz, J. & Bengio, S., "A comparative study of adaptation methods for speaker verification", Proc. ICSLP 2002.

[10] Morris, A.C., Koreman, J., Sellahewa, H., Ehlers, J, Jassim, S., Allano, L. & Garcia-Salicetti, S., "The SecurePhone PDA database, experimental protocol and automatic test procedure for multimodal user authentication", Tech Report, Jan. 2006.

http://www.coli.uni-saarland.de/SecurePhone/documents/PDA_database_and_test_protocol.pdf

[11] Porée, F., Mariéthoz, J., Bengio, S. & Bimbot, F., "The BANCA Database and experimental protocol for speaker verification", IDIAP-RR 02-13, 2002.

[12] Reynolds, D.A., "Speaker identification and verification using Gaussian mixture speaker models", Speech Communication, Vol.17, pp.91-108, 1995

[13] Reynolds, D.A., Quatieri, T.F. & Dunn, R.B., "Speaker verification using adapted gaussian mixture models", Digital Signal Processing, Vol. 10, Issues 1-3, pp. 19-41, 2000.

[14] Ricci' R., Chollet, G., Crispino' M.V., Jassim, S., Koreman, J., Olivar-Dimas, M., García-Salicetti' S. & Soria-Rodríguez, P., "SecurePhone: a mobile phone with biometric authentication and e-signature support for dealing secure transactions on the fly", Proc. SPIE Conference on Mobile Multimedia/Image Processing for Military and Securty Applications, Orlando (in press).

[15] Sellahewa, H. & Jassim, S.,"Wavelet-based Face Verification for constrained platforms", Proc. SPIE Conference on Biometric Technology for Human Identification II, Florida, Vol. 5779, pp 173-183, 2005.