MODELLING SPOKEN SIGNATURES WITH GAUSSIAN MIXTURE MODEL ADAPTATION

Jean Hennebert, Andreas Humm and Rolf Ingold

Université de Fribourg, Boulevard de Pérolles 90, 1700 Fribourg, Switzerland

ABSTRACT

We report on our developments towards building a novel user authentication system using combined acquisition of online handwritten signature and speech modalities. In our approach, signatures are recorded by asking the user to say what she/he is writing, leading to the so-called spoken signatures. We have built a verification system composed of two Gaussian Mixture Models (GMMs) sub-systems that model independently the pen and voice signal. We report on results obtained with two algorithms used for training the GMMs, respectively Expectation Maximization and Maximum A Posteriori Adaptation. Different algorithms are also compared for fusing the scores of each modality. The evaluations are conducted on spoken signatures taken from the MyIDea multimodal database, accordingly to the protocols provided with the database. Results are in favor of using MAP adaptation with a simple weighted sum fusion. Results show also clearly the impact of time variability and of skilled versus unskilled forgeries attacks.

Index Terms— Handwriting recognition, speaker recognition, pattern classification

1. INTRODUCTION

Multimodal biometrics has raised a growing interest in the industrial and scientific communities. The potential increase of accuracy combined with better robustness against forgeries makes indeed multimodal biometrics a promising field. In our work, we are interested in building multimodal authentication systems using speech and signatures as modalities. Speech and signatures are indeed two major modalities used by humans in their daily transactions and interactions. Many automated biometric systems based on signature or speech alone have been studied and developed in the past [1] [2]. However, there are still few deployments in commercial applications. Three reasons can be proposed to explain this: (1) negative impact of time-variability [3], (2) degraded performances in the case of trained forgeries [4][5], (3) decreased performances in mismatched conditions, such as mismatched sensors or environments [5]. Several attempts have already been reported to improve signature verification systems using speech as an extra modality. In [6], a tablet PC system based on online signature and speech is proposed to ensure the security of electronic medical records. In [3], an online signature verification system and a speaker verification system are also combined to reach better authentication performances. The main difference between these works and our approach lies in the acquisition procedure that is, in our case, simultaneous.

Our proposal is here to record bimodal signatures by asking the user to simultaneously say and write the signature. Such bimodal signatures are referred here as CHASM signatures for **c**ombined **h**andwriting **a**nd **s**peech **m**odalities signatures[1], or more simply referred as, **spoken signatures**. The motivation of performing a synchronized acquisition is multiple. Firstly, it avoids doubling the acquisition time. Secondly, the synchronized acquisition will probably give better robustness against intentional imposture as imitating simultaneously the voice and the writing of somebody else has a larger cognitive load. Finally, the synchronization patterns (i.e. where do users synchronize) or the intrinsic deformation of the inputs (mainly the slowdown of the speech) may be dependent on the user, therefore bringing useful biometrics information.

Our previous works on spoken signatures have been dedicated to data acquisition [7], survey and definition of realistic scenario of use [8] and early experiments on a baseline system [9]. We report in this paper on the continuation of the development of this system, with the introduction of more advanced modelling techniques and fusion strategies. More precisely, we report on results obtained with two algorithms used for training our GMM based system, respectively Expectation Maximization (EM) and Maximum A Posteriori Adaptation (MAP). Different algorithms are also compared for fusing the scores of each modality. Interesting conclusions are also drawn regarding the impact of time variability and on the degradation due to skilled versus unskilled forgeries attacks.

The remainder of this paper is organized as follows. In section 2, we give an overview of MyIDea, the database used for this work and of the evaluation protocols. In section 3 we present our modelling system based on a fusion of GMMs. Section 4 presents the experimental results. Finally, conclusions, discussions and future work are presented.

2. SPOKEN SIGNATURE DATABASE

2.1. MyIDea Database

Spoken signature data have been acquired in the framework of the MyIDea biometric data collection [7] [10]. MyIDea database is a multimodal database that contains other modalities such as fingerprint, talking face, etc. MyIDea contains about 70 users that have been recorded over three sessions spaced in time. The data set used to perform the experiments reported in this article has been given the reference MYIDEA-CHASM-SET1 by the distributors of MyIDea. This set should be considered as a development set. A second set of data is planned to be recorded in a near future and will be used as evaluation set. In MyIDea, spoken signatures have been acquired with a WACOM Intuos2 graphical tablet and a standard computer headset microphone (Creative HS-300). For the tablet stream, x,y-coordinates, pressure, azimuth and elevation angles of the pen are sampled at 100 Hz. The speech waveform is recorded at 16 kHz and coded linearly on 16 bits. Fig. 1 shows an example of spoken signature. The grey areas on the figure correspond to inter-stroke moments, when the user lift the pen out of the range of the tablet.

[1] In a similar way, we have also defined CHASM handwriting where the user reads what he is writing. CHASM handwriting could be used for user authentication or for enhanced content recognition, but this is out of the scope of this paper where we focus on spoken signatures.

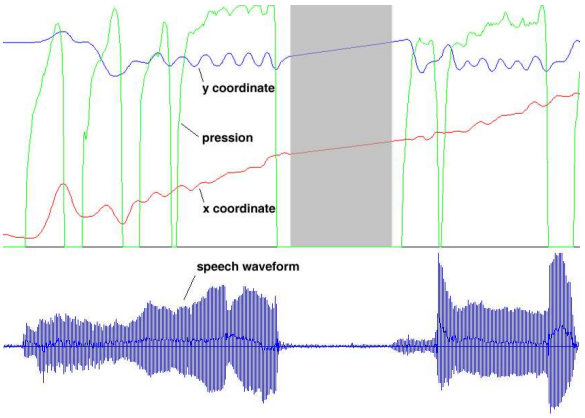

Fig. 1. Synchronized visualization of handwriting (upper part including x, y and p, not including angles for sake of clarity) and speech signals (bottom part).

In [9], we provide more comments on spoken signature data and on the way users synchronize their acoustic events with signature strokes. In [11], we report on a usability survey conducted on the subjects of MyIDea. The main conclusions of the survey are the following. First, all recorded users were able to perform the signature acquisition. Speaking and signing at the same time did not prevent any acquisition to happen. Second, the survey shows that such acquisitions are acceptable from a usability point of view.

2.2. Evaluation Protocols

In MyIDea, six *genuine* spoken signatures are acquired for each subject per session. This leads to a total of 18 true acquisitions after the three sessions. After acquiring the genuine signatures, the subject is also asked to imitate six times the signature of another subject. Spoken signature imitations are performed by letting the subject having an access to the static image and to the textual content of the signature to be forged. This procedure leads to a total of 18 *skilled forgeries* after the three sessions, i.e. six impostor signatures on three different subjects. Spoken signature assessment protocols have been defined on MyIDea [11]. The protocols have been crafted to be as realistic as possible and to put in evidence difficulties tied to time variability. Two protocols have been defined. The first one is called **without time variability** where user models are built using three spoken signatures of the first session. For testing, the three remaining signatures of the first session are used. The same procedure is repeated for sessions two and three, leading to a total of 70 users $\times$ 3 accesses $\times$ 3 sessions = 630 *genuine* tests. For impostor attempts, *random forgeries* are considered using one signature for each of the remaining subjects in the database, giving a total of 70 users $\times$ 69 accesses $\times$ 3 sessions = 14490 random forgeries. Impostor tests are also performed using *skilled forgeries* for which the 18 available skilled forgeries are used against each user, giving a total of 70 users $\times$ 18 accesses $\times$ 3 sessions = 3780 skilled forgeries. The second protocol is called **with time variability** where the six signatures from the first session are used to build client models. Genuine tests are performed on the six signatures of session two and three, giving a total of 70 users $\times$ 12 accesses = 840 genuine tests. Random and skilled impostor attempts are performed in the similar manner as for the protocol *without time variability* with the distinction that models are here trained on the first session only, giving a total of 70 users $\times$ 69 accesses = 4830 random forgeries and 70 users $\times$ 18 accesses = 1260 skilled forgeries. The amounts of tests mentioned above are approximative as some users did not complete all sessions.

3. SYSTEM DESCRIPTION

We have chosen to use standard GMMs to model independently both streams of data, followed by a simple fusion at the score level (see Fig. 2).

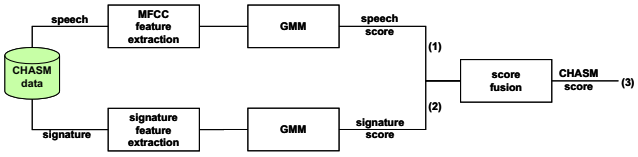

Fig. 2. Baseline CHASM signature verification system.

3.1. Feature extraction

For each point of the signature, we extract 25 dynamic features based on the x and y coordinates, the pressure and angles of the pen in a similar way as what is described in [12] and [9]. The features are mean and standard deviation normalized on a per user basis. For the speech signal, we use 12 Mel Frequency Cepstral Coefficients (MFCC) and the energy extracted every 10 ms on a window of 25.6 ms. An energy-based speech detection module based on a bi-Gaussian model is applied to remove the silence from the data. MFCC coefficients are mean and standard deviation normalized using normalization values computed on the speech part of the data. We can already mention that we performed experiments including delta and delta-delta coefficients without further improvements of the results. Delta features were then left apart in our configuration.

3.2. GMMs System

GMMs are used to model the likelihoods of the features extracted from the signature and from the speech signal. One could argue that GMMs are actually not the most appropriate models in this case as they are intrinsically not capturing the time-dependant specificities of speech and signatures. HMMs would be potentially more adequate in this case. However, GMMs have been reported to compare reasonably well to HMMs in terms of signature verification [13] and are often considered as baseline systems in speaker verification. Furthermore, GMMs are well-known flexible modelling tools able to approximate any probability density function. With GMMs, the probability density function $p(x_n|M_{client})$ or *likelihood* of a D-dimensional feature vector x_n given the model of the client M_{client}, is estimated as a weighted sum of multivariate Gaussian densities :

$$p(x_n|M_{client}) = \sum_{i=1}^{I} w_i \mathcal{N}(x_n, \mu_i, \Sigma_i) \quad (1)$$

in which I is the number of mixtures, w_i is the weight for mixture i and the Gaussian densities $\mathcal{N}$ are parameterized by a mean $D \times 1$ vector μ_i, and a $D \times D$ covariance matrix, Σ_i. In our case, we make the hypothesis that the features are uncorrelated and we use diagonal covariance matrices. By making the hypothesis of observation independence, the global *likelihood* score for the sequence of feature vectors, $X = \{x_1, x_2, ..., x_N\}$ is computed with

$S_c = p(X|M_{client}) = \prod_{n=1}^{N} p(x_n|M_{client})$. The likelihood score S_w of the hypothesis that X is **not** from the given client is here estimated using a world GMM model M_{world} or *universal background model* trained by pooling the data of many other users. The decision whether to reject or to accept the claimed user is performed comparing the ratio of client and world score against a global threshold value T. The ratio is here computed in the log-domain with $R_c = \log(S_c) - \log(S_w)$. The training of the client and world models is performed with the Expectation-Maximization (EM) algorithm that iteratively refines the component weights, means and variances to monotonically increase the likelihood of the training feature vectors. In our setting, we apply a simple binary splitting procedure to increase the number of Gaussian components through the training procedure. The world model is trained by pooling the available genuine accesses in the database. The skilled forgeries attempts are excluded for training the world model as it would lead to optimistic results. Ideally, a fully independent set of users would be preferable, but this is not possible considering the small number of users ($\approx$ 70) available. In this paper, we compare client models trained from scratch using the EM algorithm with client models obtained from an adaptation of the world model using a Maximum A Posteriori criterion [14]. As suggested in many paper using MAP, we perform only the adaptation of the mean vector μ_i, leaving untouched the covariance matrix Σ_i and the mixture coefficient w_i.

3.3. Score Fusion

We obtain the spoken signature (ss) score by applying a weighted sum of the signature (si) and speech (sp) log-likelihood ratios with $R_{c,ss} = W_{sp}R_{c,sp} + W_{si}R_{c,si}$. This is a reasonable procedure if we assume that the local observations of both sub-systems are independent. This is however clearly not the case as the users are intentionally trying to synchronize their speech with the signature signal. Time-dependent score fusion procedures or feature fusion followed by joint modelling would be more appropriate than the approach taken here. More advanced score recombination could also be applied such as, for example, using classifier-based score fusion. We report here our results with or without using a *z-norm* score normalization preceding the summation. As the mean and standard deviation of the z-norm are estimated a posteriori on the same data set, z-norm results are of course unrealistic but give however an optimistic estimation of what could be the performances.

4. EXPERIMENTAL RESULTS

We report our results in terms of Equal Error Rates (EER) which are obtained for a value of T where the impostor False Acceptation and client False Rejection error rates are equal. Table 1 shows the evolution of the EER as a function of the number of mixtures in the client and world models trained using the EM algorithm and using protocol *with time variability and random forgeries*. We tested with 8, 16, 32 and 64 Gaussian mixtures in the client and world model. Increasing the number of Gaussian further to 16 is actually showing a performances degradation, probably due to the limited amount of training data. The optimal model size for the EM algorithm seems to lie around 16 mixtures. Similar conclusions were obtained for the other protocols. As suggested in [3], we have also tried to vary the number of mixtures around 16 as a function of the number of feature vectors available for training the model of a specific user. Doing this, we obtained user-dependent numbers of mixtures. However, our experiments did not show any significant improvements against using a fixed-number of mixtures.

Table 1. EM algorithm, EER as a function of the number of Gaussian mixtures in the client and world models, protocol with time variability, random forgeries, equal weights for the fusion.

# of mixtures c/w	8/8	16/16	32/32	64/64
signature	7.4	6.1	6.2	6.8
speech	14.4	14.2	14.5	16.3
sum fusion (0.5/0.5)	**5.4**	**4.1**	**4.6**	**6.0**

Table 2 show results obtained on the same protocol but this time using the MAP adaptation training. As the MAP algorithm leaves untouched the Gaussian mixtures for which no or few training points are associated, we could here increase the number of mixtures to larger values. When comparing with the EM algorithm for equal configurations, the improvement is very much significant. Our best results are obtained for the 128/128 configuration with an overall performance of 1.7%. While MAP adaptation is actually known to improve results for modelling speech with GMMs, the results reported here are, to the best of our knowledge, the first results reported using GMM MAP adaptation to model signatures. One can conclude from these results that MAP adaptation applies better than EM when few training data is available to build the model.

Table 2. MAP algorithm, EER as a function of the number of Gaussian mixtures in the client and world models, protocol with time variability, random forgeries, equal weights for the fusion.

# of mixtures c/w	32/32	64/64	128/128
signature	3.1	3.2	2.7
speech	11.8	13.3	12.4
sum fusion (0.5/0.5)	**1.8**	**2.0**	**1.7**

Table 3 summarizes the results with our best MAP 128/128 system in terms of ERR for the different protocols. The following conclusions can be drawn. The speech modelisation performs equally well than the signature for single session experiments (without time variability). However, when multi-session accesses are considered, signature performs better than speech. Signature and speech modalities suffer from time-variability but in different degrees. It is probable that users show a larger intra-variability for the speech than for the signature modality. Another explanation could be in the acquisition conditions that are more difficult to control in the case of the speech signal: different position of the microphone, environmental noise, etc. Another conclusion from table 3 is that skilled forgeries decreases systematically and significantly the performance in comparison to random forgeries. For the protocol *with time variability*, a drop of about 200% relative performance is observed for the signature modality and about 50% for the speech modality. We have to note here that the skilled forger do not try to imitate the voice of the user but actually say the genuine verbal content. The sum fusion, although very straightforward, brings systematically a clear improvement of the results. Interestingly, the z-norm fusion is better than the sum fusion for the protocol without time variability and is worse in the case of the protocol with time variability. A visual analysis of the score distribution of both modalities, before z-norm and after z-norm, lead us to a potential intuitive interpretation of this behavior. The application of the z-norm is, by nature, aligning the score distributions of both modalities. While this is good to fuse

Table 3. MAP adaptation, protocol with and without time variability, skilled and unskilled forgeries, 128 Gaussian mixtures for the client and 128 for the world, MAP adaptation, equal weights for the fusion.

time variability	without		with	
forgeries	random	skilled	random	skilled
signature	0.4	3.9	2.7	7.3
speech	0.8	2.7	12.4	17.1
sum fusion (.5/.5)	**0.2**	**0.9**	**1.7**	**5.0**
z-norm fusion (.5/.5)	**0.1**	**0.7**	**2.3**	**8.6**

scores that lies in different ranges, the z-norm is also giving equal importance to each modalities. This is of course not favorable in the case of systems showing very different individual performances.

Figure 3 shows the evolution of the EER for different combinations of the weights used for the sum fusion in the case of the MAP 64 GMM system. As what could be expected, there are optimal weight values that minimize the EER. For protocol without time variability the optimal values are 0.3 and 0.7 for W_{si} and W_{sp} respectively. For protocol with time variability, the optimal values are 0.5 and 0.5 for W_{si} and W_{sp}. Curves for the z-norm, although not reported here, were similar as the one for the sum fusion but with different optimal weight values. While improving further the performances of our system, this optimization of the weights is optimistic as it is done a posteriori on the scores. These values should be validated on an independent evaluation set. We could also notice that, when optimal weight values are used, there is no clear advantage of using z-norm instead of the sum based fusion.

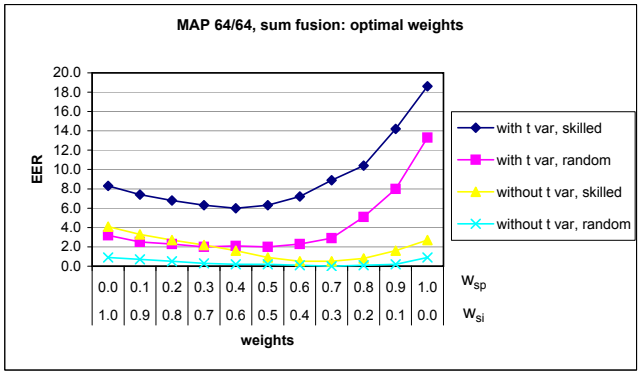

Fig. 3. Evolution of the EER as a function of fusion weights.

5. CONCLUSIONS AND FUTURE WORK

A verification system using GMMs for modelling spoken signatures has been presented and evaluated. Results obtained with this system show that the use of both modalities outperforms these modalities used alone. Results also show that there is a clear impact of time variability and skilled forgeries on the performances. The best results were obtained with a MAP adaptation procedure used to train the system and a weighted sum fusion. In our future work, we plan to investigate the use of more robust modelling techniques against time variability and forgeries. In this direction, we have identified potential modelling techniques such as HMMs, time-dependent score fusion, fusion at the feature level followed by joint modelling, etc. Also, as soon as an extended set of spoken signature data will be available, experiments will be conducted according to a development/evaluation set framework.

Acknowledgments. We warmly thank Asmaa El Hannani for her precious feedbacks when we were experimenting with GMMs based systems. This work was partly supported by the Swiss NSF program "Interactive Multimodal Information Management (IM2)", as part of NCCR and by the EU BioSecure NoE project.

6. REFERENCES

[1] F. Leclerc and R. Plamondon, "Automatic signature verification: the state of the art–1989-1993," *Int'l J. Pattern Recognition and Artif. Intelligence*, vol. 8, no. 3, pp. 643–660, 1994.

[2] D. Reynolds, "An overview of automatic speaker recognition technology," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2002, vol. 4, pp. 4072–4075.

[3] B. Ly-Van et al, "Signature with text-dependent and text-independent speech for robust identity verification," in *Proc. Workshop on Multim. User Auth. (MMUA)*, 2003, pp. 13–18.

[4] A. Humm A. Wahl, J. Hennebert and R. Ingold, "Generation and evaluation of brute-force signature forgeries," in *Int'l Workshop on Multimedia Content Representation, Classification and Security*, Istanbul, September 2006, pp. pp. 2–9.

[5] Claus Vielhauer, *Biometric User Authentication for IT Security*, Springer, 2006.

[6] S. Krawczyk and A. K. Jain, "Securing electronic medical records using biometric authentication," in *Audio- and Video-based Biometric Person Authentication (AVBPA)*, Rye Brook, NY, 2005, pp. 1110–1119.

[7] B. Dumas et al, "Myidea - multimodal biometrics database, description of acquisition protocols," in *In proc. of Third COST 275 Workshop (COST 275)*, 2005, pp. 59–62, Hatfield (UK).

[8] A. Humm, J. Hennebert, and R. Ingold, "Scenario and survey of combined handwriting and speech modalities for user authentication," in *6th Int'l Conf. on Recent Advances in Soft Computing* , Canterburry, UK, 2006, pp. 496–501.

[9] A. Humm, J. Hennebert, and R. Ingold, "Gaussian mixture models for chasm signature verification," in *3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, Washington, 2006.

[10] J. Hennebert et al, "Myidea multimodal database," http://diuf.unifr.ch/go/myidea, 2005.

[11] A. Humm, J. Hennebert, and R. Ingold, "Combined handwriting and speech modalities for user authentication," Tech. Rep. 06-05, University of Fribourg, Informatics, 2006.

[12] B. Ly Van, S. Garcia-Salicetti, and B. Dorizzi, "Fusion of hmm's likelihood and viterbi path for on-line signature verification," in *Biometrics Authentication Workshop*, May 15th 2004, Prague.

[13] J. Richiardi and A. Drygajlo, "Gaussian mixture models for on-line signature verification," in *Proc. ACM SIGMM workshop on Biometrics methods and app.*, 2003, pp. 115–122.

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