Spoken Handwriting Verification using Statistical Models

Andreas Humm, Rolf Ingold and Jean Hennebert
Université de Fribourg, Boulevard de Pérolles 90, 1700 Fribourg, Switzerland
{andreas.humm, rolf.ingold, jean.hennebert}@unifr.ch

Abstract

We are proposing a novel and efficient user authentication system using combined acquisition of online handwriting and speech signals. In our approach, signals are recorded by asking the user to say what she or he is simultaneously writing. This methodology has the clear advantage of acquiring two sources of biometric information at no extra cost in terms of time or inconvenience. We have built a straightforward verification system to model these signals using statistical models. It is composed of two Gaussian Mixture Models (GMMs) sub-systems that takes as input features extracted from the pen and voice signals. The system is evaluated on MyIdea, a realistic multimodal biometric database. Results show that the use of both speech and handwriting modalities outperforms significantly these modalities used alone. We also report on the evaluations of different training algorithms and fusion strategies.

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 handwriting as modalities. Speech and handwriting are indeed two major modalities used by humans in their daily transactions and interactions. Also, these modalities can be acquired simultaneously with no inconvenience, just asking the user to say what she/he is writing. Finally, speech and handwriting taken alone do not compare well in terms of performance against more classical biometric systems such as iris or fingerprint. Merging both biometrics will potentially lead to a competitive system.

Many automated biometric systems based on speech alone have been studied and developed in the past. See, for example, [12] for a review of such systems. Biometric systems based on online handwriting were not so numerous, however, we can mention [8] or [11] as examples of state-of-the-art systems. Close to handwriting, signature based systems have also been proposed [7].

There were also several systems that were proposed in order to use both speech and handwriting signals. In [6], a tablet PC system is using a combination of online signature and speech to ensure the security of electronic medical records. In [9], a similar system using signature and speech is also proposed to reach better authentication performances. The main difference between these works and our approach lies in the acquisition procedure that is, in their case differed and in our case, simultaneous.

Our proposal is indeed to record speech and handwriting signals where the user reads what she or he is writing. Such acquisitions are referred here and in our related works as CHASM handwriting for **c**ombined **h**andwriting **a**nd **s**peech **m**odalities handwritings[1], or more simply referred as, **spoken handwriting**. We note here that such signals could also be used to recognize the content of what is said or written. However, we focus here on the task of user authentication.

Our motivation to perform a synchronized acquisition is multiple. Firstly, it avoids doubling the acquisition time. Secondly, the synchronized acquisition will probably give better robustness against intentional imposture. Indeed, imitating simultaneously the voice and the writing of somebody has a much higher cognitive load than for each modality taken separately. Finally, the synchronization patterns (i.e. where do users synchronize) or the intrinsic deformation of the inputs (mainly the slowdown of the speech signal) may be dependent on the user, therefore bringing an extra piece of useful biometrics information.

Our previous works on spoken handwriting have been dedicated to data acquisition [1], survey and definition of realistic scenario [5] and experiments with spoken signatures [4]. In this paper, we report more specifically on the development of our spoken handwriting system and on its

[1] In a similar way, we have also defined CHASM signatures where we record a bimodal signature by asking the user to simultaneously say and write the signature, but this is out of the scope of this paper where we focus on spoken handwriting.

Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02, ICDAR'07. 2007. pp. 999-1003

evaluation using a realistic database. Conclusions on the use of different modelling algorithms and fusion strategies are also drawn.

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 and future work are presented.

2. Spoken Handwriting Database

2.1. MyIDea Database

Spoken handwriting data have been acquired in the framework of the MyIDea biometric data collection [1] [2]. MyIDea is a multimodal database that contains many other modalities such as fingerprint, talking face, etc. The "set 1" of MyIDea is already available for research institution. It includes about 70 users that have been recorded over three sessions spaced in time. 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 our future work[2].

Spoken handwriting 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. The data samples are also provided with timestamps to allow a precise synchronization of both streams. The timestamps are especially important for the handwriting streams as the graphical tablet does not send data samples when the pen is out of range. Fig. 1 shows an example of spoken handwriting, synchronized thanks to the timestamps (upper part including x, y and p, not including angles for sake of clarity) and speech signals (bottom part). The grey area on the figure corresponds to inter-stroke moments, when the user lift the pen out of the range of the tablet.

In [3], we report on a usability survey conducted on the subjects that took part to MyIDea recordings. The main conclusions of the survey are the following. First, all users were able to perform the handwriting acquisition. Speaking and writing at the same time did not prevent any acquisition to happen. Second, the answers to the survey show that simultaneous acquisitions are acceptable from a usability point of view.

[2]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.

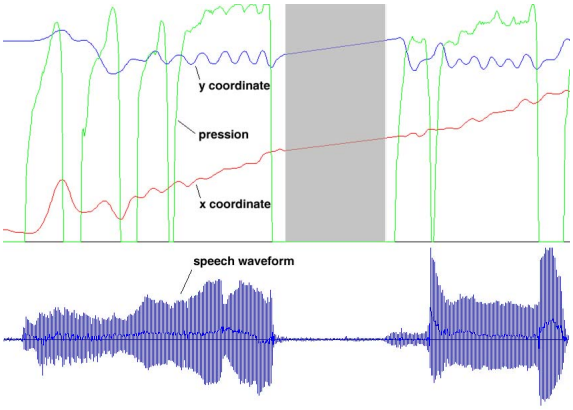

Figure 1. Synchronized visualization.

2.2. Recording and Evaluation Protocols

In MyIDea, for each of the three sessions, the subject is asked to read and write a random text fragment. The subject is allowed to train for a few lines on a separated sheet in order to accustom with the procedure of talking and writing at the same time. After acquiring the genuine handwriting, the subject is also asked to imitate the handwriting of another subject and to synchronously utter the content of the text. In order to do this, the imitator has access to the *static* handwriting data of the subject to imitate. The access to the voice recording is not given for imitation as this would lead to a too difficult cognitive load, practically infeasible in the limited time frame of the acquisition. This procedure leads to a total of three impostor attempts on different subjects after the three sessions.

A spoken handwriting assessment protocol has already been defined on MyIDea [3] and will be followed for the realization of the tests in this paper. In short, this protocol is following a **text-prompted scenario** where we assume that the system prompts the subject to write and say a random piece of text each time an access is performed. This kind of scenario allows to make the system more secure against spoofing attacks where the forger plays back a pre-recorded version of the genuine data. This scenario has also the advantage to be very convenient for the subject who does not need to remember any password phrase.

For each subject in the database, the text from the first session is used to train the system. The available text for training is, on average, composed of 5 lines for a total of 50 to 100 words. Each genuine test uses the data available from session two and session three. Therefore, 2 genuine tests can be performed per user, giving a total of 70 users * 2 accesses = 140 genuine tests. We consider two kinds of forgeries. Random forgeries are performed using one recording from the remaining subjects, giving 70 users * 69 accesses = 4830 random forgeries. Skilled forgeries are

performed using the 3 available imitations for a total of 70 users * 3 accesses = 210 skilled forgeries.

3. System Description

As illustrated on Fig. 2, our system models independently the speech and handwriting signals to obtain a score that is finally fused.

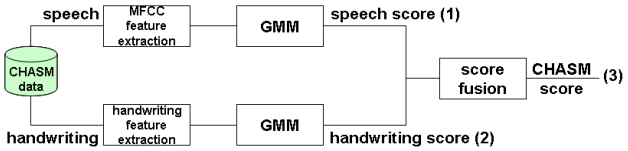

Figure 2. CHASM handwriting system.

3.1. Feature extraction

For each point of the handwriting, we extract 25 dynamic features based on the x and y coordinates, the pressure and angles of the pen in a similar way as in [10] and [4]. This feature extraction was actually proposed to model signatures, however it can be used without modification in our case as the signals are similar (coming from a graphical tablet) and as nothing specific to signature was included in the computation of the features. The features are mean and standard deviation normalized on a per user basis.

For the speech signal, we compute 12 Mel Frequency Cepstral Coefficients (MFCC) and the energy every 10 ms on a window of 25.6 ms. We realized that the speech signal contains a lot of silence which is due to the fact that writing is usually more slow than speaking. It is known, in the speech domain, that silence parts impair the estimation of reliable models. We therefore implemented a procedure to remove all the silence parts of the speech. This silence removal component is using an energy-based speech detection module based on a bi-Gaussian model. MFCC coefficients are mean and standard deviation normalized using normalization values computed on the speech part of the data.

3.2. GMMs System

GMMs are used to model the likelihoods of the features extracted from the handwriting 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 handwritings. However, we are here in the context of a text-independent scenario where the vocabulary is a priori open. A GMM is well appropriated to handle this constraint. 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

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

in which I is the number of Gaussians, w_i is the weight for Gaussian 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 use diagonal covariance matrices as approximation of the full covariance matrices. This approximation is classically done when using GMMs for two reasons. First it allows to reduce the amount of parameters to estimate, taking into account the small quantity of data available to train the biometric models. Second it is a way to reduce drastically the cpu time needed for the inversion of the covariance matrix. 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

$$S_c = p(X|M_{client}) = \prod_{n=1}^{N} p(x_n|M_{client}) \qquad (2)$$

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 usually 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. Another way to train the client model is to adapt the world model using a Maximum A Posteriori criterion (MAP) [13].

In our experiments we tried using both training algorithms. For the EM, 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[3]. For the MAP, as suggested in many papers, 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]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 (≈ 70) available.

3.3. Score Fusion

We obtain the spoken handwriting (sh) score by applying a weighted summation of the handwriting (hw) and speech (sp) log-likelihood ratios with $R_{c,sh} = W_{sp}R_{c,sp} + W_{hw}R_{c,hw}$. 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 handwriting 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 an optimistic estimation of what could be the fusion performances with such a normalisation.

4. Experimental Results

We report our results in terms of Equal Error Rates (EER) which are obtained for a value of the threshold T where the impostor False Acceptation and client False Rejection error rates are equal.

First, we performed a set of tests using 16, 32, 64, 128 and 256 Gaussian mixtures in the client and world model trained with the EM algorithm. From these tests, we observed that the optimal model size seems to lie around 256 mixtures. Increasing the number of Gaussian further to 256 is actually showing a performances degradation, probably due to the limited amount of training data. Coincidentally, the optimal sizes for the handwriting and speech GMMs were both equal to 256. Therefore, for the rest of the paper, we will report results obtained using models of size 256. Table 1 shows a comparison of results using the EM versus the MAP algorithm and random forgeries for testing. The fusion is, in this case, the simple summation fusion, without any z-norm nor weighting. Several interesting conclusions can be drawn from these results.

1. The fusion of speech and handwriting can really improve the performance of the biometric system. This gain of performance can be obtained at no extra cost for the user as both stream of data is recorded simultaneously.

2. The sum fusion that is applied here is extremely simple and requires actually no further estimation of parameters. This result can be explained considering that the models used for speech and handwriting are very much similar in architecture and order.

3. The MAP adaptation algorithm is leading to better results than the EM algorithm. 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 handwriting. The reasoning is probably similar, i.e. it is better to adapt from a well-trained world model than to build from scratch a GMM using few data.

Table 1. Comparison EM/MAP algorithms on random forgeries.

algorithm	EM	MAP
handwriting	6.8 %	4.0 %
speech	7.5 %	1.8 %
sum fusion (0.5/0.5)	**2.3 %**	**0.7 %**

Table 2 summarizes the results with our best MAP 256/256 system but this time comparing random versus skilled forgeries. The following conclusions can be drawn. For the handwriting, skilled forgeries decreases the performances in a significant manner. This result could be expected as the forger is imitating the handwriting of the genuine user. For the speech signal, skilled forgeries also decreases the performance. As the forger do not try to imitate the voice of the genuine user, this result can be surprising. However, it can be explained as the forger is actually saying the exact same verbal content as the one used by the user at training time. When building a speaker model, the characteristics of the speaker are of course captured, but also, to some extend, the content of the speech signal itself.

Results using the z-norm fusion are also reported in table 2, showing an advantage against the sum fusion. The application of the z-norm is, by nature, aligning the score distributions of both modalities which, as expected, leads to better fusion results without needing to tune the weights.

Table 2. Random versus skilled forgeries.

forgeries	random	skilled
handwriting	4.0 %	13.7 %
speech	1.8 %	6.9 %
sum fusion (.5/.5)	**0.7 %**	**6.9 %**
z-norm fusion (.5/.5)	**0.3 %**	**4.0 %**

Figure 3 shows the evolution of the EER for different combinations of the weights used for the sum fusion in the case of our best MAP 256/256 GMM system. As what could be expected, there are optimal weight values that minimize the EER. The optimal values are 0.2 and 0.8 for W_{hw}

and W_{sp} respectively, giving an EER of 0.5 and 2.9 for random and for skilled forgeries. 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.

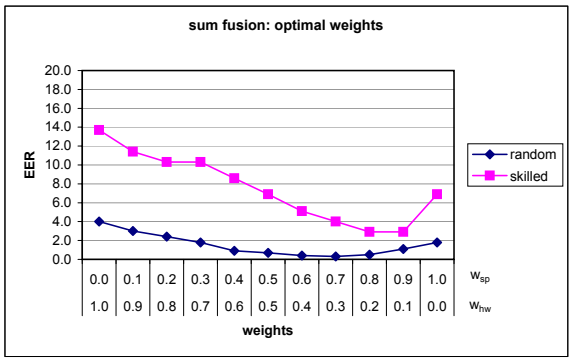

Figure 3. Evolution of the EER according to different values of the fusion weights.

As general conclusion of these experiments, we can reasonably say that the speech modelisation performs on average better than the handwriting. Intuitively, one could argue that this is understandable as the handwriting is a gesture that is more or less fully learned (behavioral biometric) while speech contains information that are dependent to learned and physiological features (behavioral and physiological biometric). However, we should pay attention to the fact that performance of speech and handwriting biometric systems are dependent to the quantity of data available for training. This means that if we would have more handwriting material, the conclusion could potentially be reversed. Also, using 50 to 100 words to perform the access is probably not applicable for a commercial scenario. We plan, in our future work, to analyze the impact on the performances using less words for testing.

5. Conclusions and Future Work

A verification system using GMMs for modelling spoken handwritings has been presented and evaluated. Results obtained with this system show that the use of both modalities outperforms these modalities used alone. The results also show that there is a clear impact of 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 forgeries. In this direction, we have identified potential directions such as time-dependent score fusion, fusion at the feature level followed by joint modelling, etc. Also, as soon as an extended set of spoken handwriting 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.

References

[1] B. Dumas et al. Myidea - multimodal biometrics database, description of acquisition protocols. In *In proc. of Third COST 275 Workshop (COST 275)*, pages 59–62, October 27 - 28 2005. Hatfield (UK).

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

[3] A. Humm, J. Hennebert, and R. Ingold. Combined handwriting and speech modalities for user authentication. Technical Report 06-05, University of Fribourg, Department of Informatics, 2006.

[4] A. Humm, J. Hennebert, and R. Ingold. Gaussian mixture models for chasm signature verification. In *3rd Joint Workshop on MLMI*, Washington, 2006.

[5] 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 RASC 2006*, pages 496–501, Canterburry, Kent, United Kingdom, 2006.

[6] S. Krawczyk and A. K. Jain. Securing electronic medical records using biometric authentication. In *AVBPA*, pages 1110–1119, Rye Brook, NY, 2005.

[7] F. Leclerc and R. Plamondon. Automatic signature verification: the state of the art–1989-1993. *Int'l J. Pattern Recognition and Artificial Intelligence*, 8(3):643–660, 1994.

[8] M. Liwicki, A. Schlapbach, H. Bunke, S. Bengio, J. Mariéthoz, and J. Richiardi. Writer identification for smart meeting room systems. In *Proceedings of the 7th International Workshop on Document Analysis Systems*, page 186195, 2006.

[9] B. Ly-Van et al. Signature with text-dependent and text-independent speech for robust identity verification. In *Proc. Workshop on MMUA*, pages 13–18, 2003.

[10] 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.

[11] Y. Nakamura and M. Kidode. Online writer verification using kanji handwriting. In *Int'l Workshop on MRCS*, pages 207–214, Istanbul, September 2006.

[12] D. Reynolds. An overview of automatic speaker recognition technology. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 4, pages 4072–4075, 2002.

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