Modelling Combined Handwriting and Speech Modalities

Andreas Humm, Jean Hennebert, and Rolf Ingold

Université de Fribourg, Boulevard de Pérolles 90, 1700 Fribourg, Switzerland
{andreas.humm, jean.hennebert, rolf.ingold}@unifr.ch

Abstract. We are reporting on consolidated results obtained with a new user authentication system based on 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 are proposing here two scenarios of use: spoken signature where the user signs and speaks at the same time and spoken handwriting where the user writes and says what is written. These two scenarios are implemented and fully evaluated using a verification system based on Gaussian Mixture Models (GMMs). The evaluation is performed on MyIdea, a realistic multimodal biometric database. Results show that the use of both speech and handwriting modalities outperforms significantly these modalities used alone, for both scenarios. Comparisons between the spoken signature and spoken handwriting scenarios are also drawn.

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 signing or 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.

1.1 Motivations

Many automated biometric systems based on speech alone have been studied and developed in the past, as reviewed previously [1]. Numerous biometric systems based on signature have also been studied and developed in the past [2][3]. Likewise biometric systems based on online handwriting were not so numerous, however, we can refer to [4] or [5] as examples of state-of-the-art systems.

S.-W. Lee and S.Z. Li (Eds.): ICB 2007, LNCS 4642, pp. 1025–1034, 2007.

Our proposal here is to record speech and handwriting signals where the user reads aloud what she or he is writing. Such acquisitions are referred here and in our related works as CHASM for **c**ombined **h**andwriting **a**nd **s**peech **m**odalities[1]. In this work, we have been defining two scenarios. In the first one, called **spoken signatures**, a bimodal signature with voice is acquired. In this case, the user is simply asked to say the content of the signature, corresponding in most of the case to his or her name. This scenario is similar, in essence, to text-dependent password based systems where the signature and speech content remains the same from access to access. Thanks to the low quantity of data requested to build the biometric templates, this scenario would fit in commercial applications running, for example, in banks. In the second scenario, called **spoken handwriting**, the user is asked to write and read synchronously the content of several lines of a given random piece of text. This scenario is less applicable in the case of commercial applications because of the larger quantity of data requested to build models. However, it could be used for forensic applications. Comparisons that we will draw between these scenarios will, of course, have to be weighted due to the difference of quantity of data.

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.

1.2 Related Work

Several related works have already shown that using speech and signature modalities together permits significant improvements in authentication performances in comparison to systems based on speech or signature alone. In [6], a tablet PC system based on online signature and voice modalities is proposed to ensure the security of electronic medical records. In [7], an online signature verification system and a speaker verification system are also combined. Both sub-systems use Hidden Markov Models (HMMs) to produce independent scores that are then fused together. In [8], tests are reported for a system where the signature verification part is built using HMMs and the speaker verification part uses either dynamic time warping or GMMs. The fusion of both systems is performed at the score level and results are again better than for the individual systems. In [9], the SecurePhone project is presented where multimodal biometrics is used to secure access and authenticate transactions on a mobile device. The biometric modalities include face, signature and speech signals.

The main difference between these works and our CHASM approach lies in the acquisition procedure. In our case, the speech and signature data streams

[1] 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.

are recorded simultaneously, asking the user to actually say the content of the signature or text. Our procedure has the advantage of shortening the enrollment and access time for authentication and will potentially allow for more robust fusion strategies upstream in the processing chain. This paper is actually reporting on consolidated evaluation results of our CHASM approach. It presents novel conclusions regarding comparison of performance of spoken signature and spoken handwriting. Individual analysis and performance evaluation of spoken signatures and spoken handwriting have been presented in our related works [10][11][12].

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 CHASM Database

2.1 MyIDea Database

CHASM data have been acquired in the framework of the MyIDea biometric data collection [13][14]. 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 institutions. It includes about 70 users that have been recorded over three sessions spaced in time. This set is here 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].

CHASM data 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.

In [15], we provide more comments on spoken signature and spoken handwriting data and on the way users synchronize their acoustic events with signature strokes. In [16], 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 or handwriting acquisition. Speaking and signing or writing at the same time did not prevent any acquisition from happening. Second, the survey shows that such 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.

2.2 Recording and Evaluation Protocols

Spoken signatures. 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 in a gender dependent way by letting the subject having an access to the static image and to the textual content of the signature to be forged. The access to the voice recording is not given for imitation as this would lead to a too difficult task considering the high cognitive load and would be practically infeasible in the limited time frame of the acquisition. This procedure leads to a total of 18 *skilled forgeries* after the three sessions, i.e. six impostor signatures on three different subjects. Two assessment protocols have been defined on MyIDea with the objective of being as realistic as possible (see [16] for details). The first one is called **without time variability** where signatures for training and testing are taken from the same session. The second protocol is called **with time variability** where the signatures for training are taken from the first session while for testing they are taken from a different session. To compare with the skilled forgeries described above, we also test with *random forgeries* taking the accesses from the remaining users. These protocols are strictly followed here.

Spoken handwriting. For each of the three sessions, the subject is asked to read and write a random text fragment of about 50 to 100 words. The subject is allowed to train for a few lines on a separated sheet in order to be accustomed to 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 (same gender) and to synchronously utter the content of the text (skilled forgeries). 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 also not given for imitation. This procedure leads to a total of three impostor attempts on different subjects after the three sessions. An assessment protocol for spoken handwriting is also available with MyIDea [16] and is followed for the realization of the tests in this paper. In short, this protocol trains the models on data from session one and test it on data from sessions two and three. As for spoken signatures, we also test against skilled forgeries and random forgeries. It actually corresponds to a **text-prompted scenario** where 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 the system to be more secure against spoofing attacks where the forger plays back a pre-recorded version of the genuine data. This scenario also has the advantage of being very convenient for the subject who does not need to remember any password phrase.

3 System Description

As illustrated on Fig. 1, our system models independently the speech and handwriting signals to obtain a score that is finally fused.

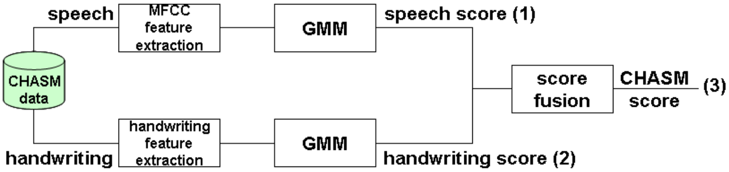

Fig. 1. CHASM handwriting system

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 [17] and [10]. This feature extraction was actually proposed to model signatures. However it can be used without modification in the case of handwriting 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 models. We therefore implemented a procedure to remove all the silence parts of the speech signal. This silence removal component is using a classical 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-dependent specificities of speech and handwriting. However, a GMM is well appropriated to handle the text-independent constraint of the spoken handwriting scenario. We also wanted to have similar types of models for both scenarios to draw fair comparisons. 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 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

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

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) [18].

In our experiments we used the EM algorithm to build the word model by applying 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]. In the results reported here, we used MAP adaptation to build the client models. 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.

3.3 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. The z-norm is here applied globally on both speech and signature scores for all test accesses, in a user-independent way. 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.

[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.

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.

4.1 Spoken Signature

Table 1 summarizes the results with our best MAP system (128 Gaussians for the client and world models) in terms of ERR for the different protocols. The following conclusions can be drawn. The speech modelisation performs equally well as the signature in the case of 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. This could be here even more amplified as users are probably not used to slow down the speech to the pace of handwriting. 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 1 is that skilled forgeries decrease 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 forgers do not try to imitate the voice of the user but actually say the genuine verbal content which is very probably the source of the loss of performance. Also from Table 1, we can conclude that the sum fusion, although very straightforward, brings systematically a clear improvement in the results, in comparison to the modalities taken alone. 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*. An interpretation of this is proposed in [11].

4.2 Spoken Handwriting

Table 2 summarizes the results with our best MAP system (256 Gaussians for the client and world models), comparing random versus skilled forgeries. The

Table 1. Summary of spoken signature results in terms of terms of Equal Error Rates. Protocol with and without time variability, skilled and unskilled forgeries.

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 %**

Table 2. Spoken handwriting results in terms of terms of Equal Error Rates, with time variability. Comparison of 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 %**

following conclusions can be drawn. For the handwriting, skilled forgeries decrease the performances in a significant manner. This result is actually understandable as the forger is intentionally 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 extent, 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.

As a conclusion of these experiments with spoken handwriting, 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 on learned and physiological features (behavioral and physiological biometric).

4.3 Comparison of Spoken Signatures and Spoken Handwriting

We are able here to do a comparison of results obtained with spoken signatures and spoken handwriting data as our experiments are performed using the same database, with the same users and the same acquisition conditions. Results of spoken handwriting in Table 2 can be compared with results of spoken signatures in Table 1, for the protocol *with time variability*. The signature modality of spoken signatures provides better results than the handwriting modality of spoken handwriting. This can be explained in the following way. Handwriting is a taught gesture that is crafted to be understood by every person. In school, every child learns in more or less the same way to write the different characters. In contrast, a signature is built to be an individual characteristic of a person that should not be imitable and that is used for authentication purposes. A comparison of the speech modality of Table 1 and 2 shows that spoken handwriting provides better results than spoken signatures. An explanation for this lies in the quantity of speech data available. While the average length of the speech is about two seconds for signature, spoken handwriting provides about two minutes of speech. The speech model is therefore more precise for spoken handwriting than for spoken signature. Now, if we compare the z-norm fusion of Table 1 and 2, we can observe that spoken handwriting performs better than spoken signatures. However, we should pay attention that this conclusion is also dependent on the quantity of data. If we would have less handwriting data, the conclusion may also be reversed.

5 Conclusions and Future Work

We presented consolidated results obtained with a new user authentication system based on combined acquisition of online handwriting and speech signals. It has been shown that the modelling of the signals can be performed advantageously using GMMs trained with a MAP adaptation procedure. A simple fusion of GMM scores lead to significant improvements in comparison to systems where the modalities would be used alone. From a usability point of view, this gain of performance is obtained at no extra cost in terms of acquisition time, as both modalities are recorded simultaneously. The proposed bi-modal speech and handwriting approach seems then to be a viable alternative to systems using single modalities. In our future work, we plan to investigate the use of more robust modelling techniques against time variability and forgeries. We have identified potential directions such as HMMs, time-dependent score fusion, 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. We will also investigate if the biometrics performances are impaired due to the signal deformations induced by the simultaneous recordings.

Acknowledgments. We warmly thank Asmaa El Hannani for her precious help 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 IST-2002-507634 NoE project.

References

1. Reynolds, D.: An overview of automatic speaker recognition technology. In: Proc. IEEE ICASSP, vol. 4, pp. 4072–4075 (2002)
2. Plamondon, R., Lorette, G.: Automatic signature verification and writer identification - the state of the art. Pattern Recognition 22(2), 107–131 (1989)
3. Leclerc, F., Plamondon, R.: Automatic signature verification: the state of the art–1989-1993. Int'l J. Pattern Rec. and Art. Intelligence 8(3), 643–660 (1994)
4. Liwicki, M., Schlapbach, A., Bunke, H., Bengio, S., Mariéthoz, J., Richiardi, J.: Writer identification for smart meeting room systems. In: Proceedings of the 7th International Workshop on Document Analysis Systems, pp. 186–195 (2006)
5. Nakamura, Y., Kidode, M.: Online writer verification using kanji handwriting. In: Gunsel, B., Jain, A.K., Tekalp, A.M., Sankur, B. (eds.) MRCS 2006. LNCS, vol. 4105, pp. 207–214. Springer, Heidelberg (2006)
6. Krawczyk, S., Jain, A.K.: Securing electronic medical records using biometric authentication. In: Kanade, T., Jain, A., Ratha, N.K. (eds.) AVBPA 2005. LNCS, vol. 3546, pp. 1110–1119. Springer, Heidelberg (2005)

7. Fuentes, M., et al.: Identity verification by fusion of biometric data: On-line signature and speech. In: Proc. COST 275 Workshop on The Advent of Biometrics on the Internet, Rome, Italy, November 2002, pp. 83–86 (2002)
8. Ly-Van, B., et al.: Signature with text-dependent and text-independent speech for robust identity verification. In: Proc. Workshop MMUA, pp. 13–18 (2003)
9. Koreman, J., et al.: Multi-modal biometric authentication on the securephone pda. In: Proc. Workshop MMUA, Toulouse (2006)
10. Humm, A., Hennebert, J., Ingold, R.: Gaussian mixture models for chasm signature verification. In: 3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, Washington (2006)
11. Hennebert, J., Humm, A., Ingold, R.: Modelling spoken signatures with gaussian mixture model adaptation. In: 32nd ICASSP, Honolulu (2007)
12. Humm, A., Ingold, R., Hennebert, J.: Spoken handwriting verification using statistical models. In: Accepted for publication ICDAR (2007)
13. Dumas, B., et al.: Myidea - multimodal biometrics database, description of acquisition protocols. In: proc. of Third COST 275 Workshop (COST 275), Hatfield (UK) (October 27 - 28 2005), pp. 59–62 (2005)
14. Hennebert, J., et al.: Myidea database (2005), `http://diuf.unifr.ch/go/myidea`
15. Humm, A., Hennebert, J., Ingold, R.: Scenario and survey of combined handwriting and speech modalities for user authentication. In: 6th Int'l. Conf. on Recent Advances in Soft Computing (RASC 2006), Canterburry, Kent, UK, pp. 496–501 (2006)
16. Humm, A., Hennebert, J., Ingold, R.: Combined handwriting and speech modalities for user authentication. Technical Report 06-05, University of Fribourg, Department of Informatics (2006)
17. Van Ly, B., Garcia-Salicetti, S., Dorizzi, B.: Fusion of hmm's likelihood and viterbi path for on-line signature verification. In: Biometrics Authentication Workshop, Prague (May 15th 2004)
18. Reynolds, D., Quatieri, T., Dunn, R.: Speaker verification using adapted gaussian mixture models. Digital Signal Processing 10, 19–41 (2000)