Hidden Markov Models for Spoken Signature Verification

Andreas Humm, Jean Hennebert and Rolf Ingold

Abstract— **In this paper we report on the developments of an efficient user authentication system using combined acquisition of online signature and speech modalities. In our project, these two modalities are simultaneously recorded by asking the user to utter what she/he is writing. The main benefit of this multimodal approach is a better accuracy at no extra costs in terms of access time or inconvenience. More specifically, we report in this paper on significant improvements of our initial system that was based on Gaussian Mixture Models (GMMs) applied independently to the pen and voice signal. We show that the GMMs can be advantageously replaced by Hidden Markov Models (HMMs) provided that the number of state used for the topology is optimized and provided that the model parameters are trained with a Maximum a Posteriori (MAP) adaptation procedure instead of the classically used Expectation Maximization (EM). The evaluations are conducted on spoken signatures taken from the MyIDea multimodal database. Consistently with our previous evaluation of the GMM system, we observe for the HMM system that the use of both speech and handwriting modalities outperforms significantly these modalities used alone. We also report on the evaluations of different score fusion strategies.**

I. 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. On the one hand, handwritten signatures are nowadays legally and socially accepted means for user authentication and contractual terms acceptance. On the other hand, producing a speech signal is a very natural non-intrusive gesture.

Many automated biometric systems based on signature or speech alone have been studied and developed in the past (many surveys are available, see for example [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. It is also worth mentioning the work presented in [7], where a similar approach is used, not for biometric aspects but to enhance the recognition of spoken content for noisy mobile environment. In this approach, the user simultaneously writes the first characters of a spoken utterance. The recognition of the first characters is injected in the HMM decoding of the speech part and allows to enhance the speech detection while eliminating less probable hypothesis.

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 to as, **spoken signatures**. The motivations of performing a synchronized acquisition are 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 [8], survey and definition of realistic scenario of use [9] and experiments on a baseline system based on Gaussian Mixture Models (GMM) applied independently to the pen and voice signal [10][11]. We report in this paper on improvements of this baseline system with an attempt to replace the GMMs by Hidden Markov Models (HMMs). On the one hand, the HMMs should present the advantage of allowing for more detailed modelling of the data, incorporating sequential information of the strokes for signature and of the phonemes for speech. On the other hand, HMMs have more parameters to tune such as the choice of the topology and the number of states. An extensive evaluation carried on the MyIDea database compares the GMMs and HMMs approaches. Different algorithms are also compared for training the models, namely the Expectation Maximization (EM) and the Maximum a Posteriori (MAP)

A. Humm, J. Hennebert and R. Ingold are with:
Université de Fribourg, Boulevard de Pérolles 90, 1700 Fribourg, Switzerland, {andreas.humm, jean.hennebert, rolf.ingold}@unifr.ch

[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.

First IEEE International Conference on Biometrics: Theory, Applications, and Systems, BTAS 2007, pp. 1 -6

adaptation procedure. Finally, two score fusion strategies are compared and conclusions are drawn regarding the impact of skilled versus random 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 score-level fusion of HMMs. Section 4 presents the experimental results. Finally, conclusions, discussions and future work are presented.

II. SPOKEN SIGNATURE DATABASE

A. MyIDea Database

Spoken signature data have been acquired in the framework of the MyIDea biometric data collection [8][12]. MyIDea 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 area on the figure correspond to inter-stroke moments, when the user lift the pen out of the range of the tablet. We have to note that these kind of events are not very frequent for signatures and are more frequent for handwriting.

In [10], we provide more comments on spoken signature data and on the way users synchronize their acoustic events with signature strokes. In [13], 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.

B. 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 [13]. The protocols have been crafted to be as realistic as possible and to put in evidence difficulties tied to time variability and skilled impostors.

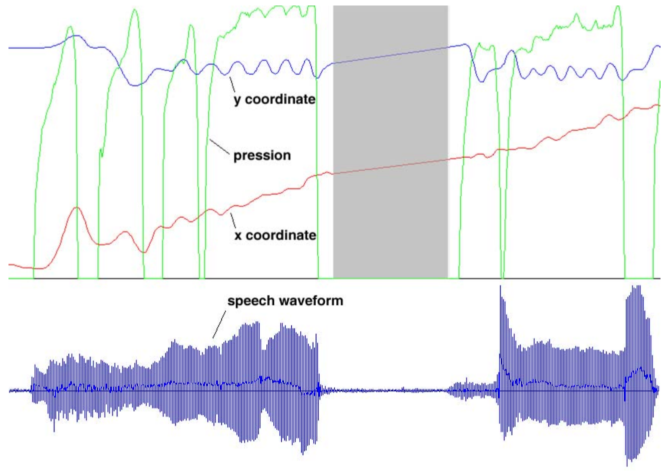

Fig. 1. Synchronized visualization of handwriting (upper part including x, y and p, not including angles for sake of clarity) and speech signals (bottom part).

Two protocols have been defined. The first one is called **without time variability** where user models are built and tested using signatures from the same session. The second protocol is called **with time variability** where user models are built using data from one session and tested using data from the other two sessions that are spaced in time. In this paper, we focus on the protocol *with time variability* which is corresponding to a more realistic situation. For extended results on the comparison between the two protocols, we refer to [10][11].

Regarding the protocol *with time variability*, 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. 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 amounts of tests mentioned above are approximative as some users did not complete all sessions.

III. SYSTEM DESCRIPTION

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

A. 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 [14] and [10]. 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

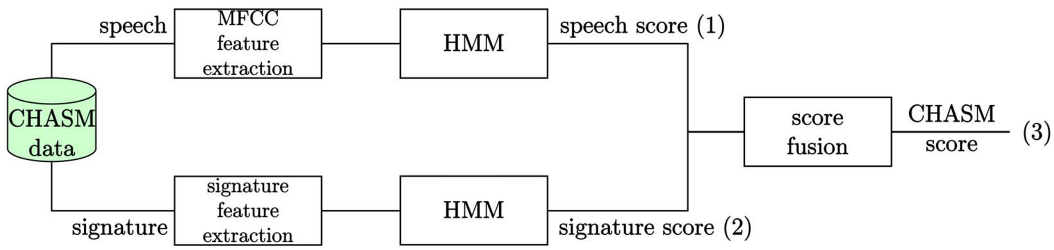

Fig. 2. CHASM signature verification system.

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. This was probably due to the small amount of training data available in our case. Delta features were then left apart in our configuration.

B. HMMs System

HMMs have been extensive used to model the likelihoods of the features extracted from signatures [14][15] or handwriting [16] and from the speakers [17]. Our motivations to study HMMs are multiple. First, they are the natural extension of the GMMs that were used in our initial baseline system. Second, they should allow more detailed modelling of the data, incorporating sequential information of the strokes for handwriting and of the phonemes for speech. While HMMs are richer than GMMs in terms of modelling capabilities, they have more parameters to tune such as the choice of the topology and the number of states. Related works have actually compared HMMs to GMMs in the framework of signature modelling with no clear advantages of the HMMs against GMMs [18].

The client score S_{client} is here the likelihood of the observation sequence X given the HMM parameters associated to a client. By applying the usual simplifying assumption of HMM based modelling (see for example [17]), the likelihood of X given the model M_{client} can be written

$$\begin{aligned} S_{client} &= P(X|M_{client}) \\ &= \sum_{allpaths} \prod_{n=1}^{N} \underbrace{P(x_n|q_n, M_{client})}_{em.\ probs} \\ & \quad \underbrace{P(q_n|q_{n-1}, M_{client})}_{trans.\ probs} \end{aligned} \tag{1}$$

which expresses the likelihood as the sum, over all possible state paths of length N in the model, of the product of emission probabilities and transition probabilities measured along the paths. The value $P(x_n|q_n, M_{client})$ is the so-called *emission probability* and represents the probability to observe a feature vector x_n when visiting state q_n. The value $P(q_n|q_{n-1}, M_{client})$ is the *transition probability* and represents the probability to go from state q_{n-1} to state q_n in the HMM. Alternatively to equation 1, the Viterbi criterion can also be used, stating that instead of considering all potential paths through the HMM, only the best path is taken into account, i.e. the path that maximizes the product of emission and transition probabilities.

We have chosen to use continuous HMMs where the emission probability is modelled using a probability density function computed with weighted Gaussian mixtures. Additionally, we have made the usual assumption that the components of the feature vector are uncorrelated. This presents the advantage to let the covariance matrix be diagonal and to be more computationally efficient. As illustrated on Fig. 3, we have opted to use a strictly left-right topology for the HMM where transitions are only allowed from each state to itself and to its immediate right-hand neighbors. Such a topology is widely used for modelling speech as states will naturally correspond to the sequence of phonemes. For the signature, the state sequence will model the sequence of strokes. As explained in more details in section 4, we investigated different strategies to find out the optimal number of states for each client.

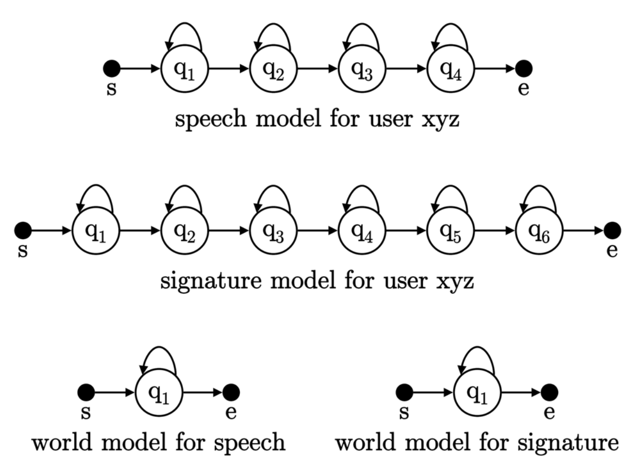

Fig. 3. HMM topology.

The likelihood score S_{world} 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

$$R_{client} = \log(S_{client}) - \log(S_{world}) \quad (2)$$

The training of each client's model is done in several iterations. In each iteration two steps are performed. In the first step, a Viterbi forced alignment is performed [17] to find the most likely sequence of states given the parameters of the HMM. In the second step, the Gaussian mixture densities of each state are re-estimated 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 accumulated in a given state. In our setting, we apply a simple binary splitting procedure to increase the number of Gaussian mixtures through the training procedure for the client's model and world models. Transition probabilities are also updated simply counting the accumulated number of passages on a given transition.

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 (≈ 70) available.

Alternatively to the EM based training described above, we also investigated the use of a Maximum A Posteriori criterion [19] to adapt the client model from the world model. As suggested in many papers, we perform only the MAP adaptation of the mean vector, leaving untouched the covariance matrix and the mixture coefficient.

C. 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 $R_{client,ss} = W_{sp}R_{client,sp} + W_{si}R_{client,si}$. 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 could be more appropriate than the approach taken here, but we leave this for future work. 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.

IV. 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 Acceptance and client False Rejection error rates are equal.

We have chosen to build the HMMs with variable number of states for each user. Intuitively, as users have different sizes of signatures and also different number of phonemes in their name, the number of states should then be different for each user. Also, the optimal number of states for the speech part and for the signature part will probably be different as the respective signal production processes are different. As we don't know a priori what is actually pronounced and written in spoken signature, we have chosen to compute a number of states proportionally to the number of signature and speech feature vectors. More precisely, for the signature part, the number of states K_{si} in the HMM is computed proportionally to the average number N_a of feature vectors in the six available genuine signatures:

$$K_{si} = \frac{N_a}{\alpha} \quad (3)$$

where α is a dividing factor that needs to be tuned. There is a similar computation for the speech modelling part, taking into account the number of speech frames instead of the number of signature points.

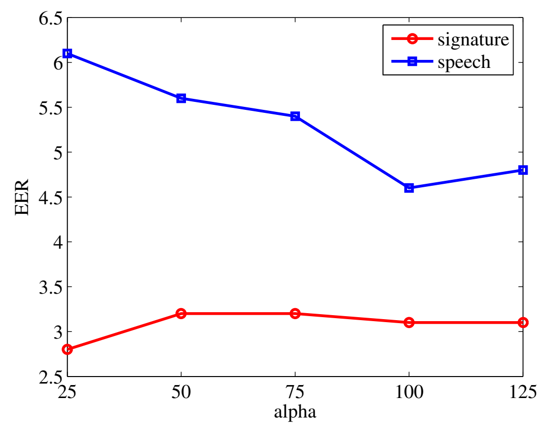

Fig. 4. Equal Error Rates (random forgeries, MAP) as a function of the dividing factor α, 16 Gaussian mixtures in each state of the HMM and 16 Gaussian mixtures in the world model.

The first set of experiments aimed at finding the optimal value of the α parameter. We used the MAP adaptation algorithm with 16 Gaussian mixtures in each state and 16 Gaussian mixtures in the world model. As it can be observed from Fig. 4, there is an optimal value of α and this optimal value is different for the speech and signature parts. As we have more or less the same number of feature vectors for the signature part and for the speech part (one vector every 10 ms for both streams, the silence parts being removed from the speech signal), we can then conclude that these optimal values of α will lead to HMMs with more states for the signature part than for the speech part. This result is actually in accordance with the observation that there are generally more strokes in a signature than there are phonemes in the spoken name. We compared this approach (variable number of states) to the approach of using a fixed number of states for all users. We tried with 1, 3, 5, 7 and 9 states per HMM,

TABLE I

RESULTS WITH GMM AND HMM MODELLING FOR EM AND MAP ADAPTATION.

modelling	GMM (%EER)				HMM (%EER)			
training	EM		MAP		EM		MAP	
forgeries	random	skilled	random	skilled	random	skilled	random	skilled
signature	5.5	9.0	2.6	7.4	3.8	7.3	2.8	5.6
speech	7.6	12.7	5.2	13.5	8.0	11.7	4.6	12.7
sum fusion	**3.7**	**5.8**	**1.8**	**5.6**	**2.8**	**5.0**	**1.5**	**4.2**
z-norm fusion	**2.7**	**6.0**	**1.5**	**5.6**	**2.1**	**4.7**	**1.1**	**5.0**

16 Gaussian per states. The results were in favor of using a variable number of states as described above.

In a second set of experiments, our objective was to perform an extensive comparison of GMMs with HMMs for different training algorithms (EM vs. MAP) and for different strengths of forgeries (random vs skilled). In these experiments, the configuration of the GMMs was 16 Gaussians for the client and 16 Gaussians for the world model. The HMM system was using 16 Gaussians in each states and with a variable number of states for each client such as described above. The values of α were respectively equal to 25 and 100 for the signature and speech parts using MAP. For the EM algorithm, the best α values were found slightly different, both equal to 100 for the signature and speech part. Table I summarizes the results in terms of ERR. The following conclusions can be drawn.

1) **Comparison GMM - HMM**. When considering the fusion of both modalities, the HMM modelling is leading consistently to better accuracy than the GMM modelling. When the signature and speech modalities are considered separately, the HMM modelling is, in most of the cases, leading to better results than the GMM modelling. Only two of the configurations show a slight advantage for GMMs but probably, the difference are not significant in these cases.
2) **Comparison EM - MAP**. As it was already reported in many previous work (including ours), GMMs benefit significantly from a MAP adaptation instead of a full EM training. Interestingly, we see the same tendency for the HMMs. The MAP adaptation is also better in terms of cpu usage as typically fewer iterations on the training set are required to reach convergence.
3) **Comparison random - skilled forgeries**. We can observe that skilled forgeries decrease systematically and significantly the performance in comparison to random forgeries and this for both modalities. This result is clearly understandable for the signature part where the forger is training to imitate the genuine signature. For the speech part, the impact is also understandable even though the forger does not try to imitate the voice of the user. Indeed, the forger is actually saying the genuine verbal content, i.e. producing a speech signal phonetically close to the genuine enrollment data.
4) **Comparison sum fusion - z-norm fusion**. As what could be expected, the z-norm fusion is better than the sum fusion for most configurations[2]. However, we can notice that the simple sum fusion is giving fairly good results. This is probably due to the fact that we are fusing scores computed with very similar systems.
5) **Comparison signature - speech**. For all configurations, the signature modality performs better than the speech one. Signatures are probably more discriminative and more stable through time than speech for the protocols used in these experiments.

As a general comment on the approach, the score fusion of both modalities, even for the very straightforward sum based procedure, brings systematically a clear improvement of the results in comparison to the modalities used alone. These results are in favor of the multimodal spoken signature methodology in terms of accuracy.

V. CONCLUSIONS AND FUTURE WORK

A verification system using HMMs for modelling spoken signatures has been presented and evaluated on a realistic multi-session protocol. Results obtained with this system show that the use of both modalities outperforms these modalities used alone. Furthermore, as both modalities are acquired simultaneously, the gain in accuracy is not at the cost of any extra time for the user to enroll or to perform an access. The current best performance of our spoken signature verification system on random forgeries is 1.1% EER, obtained using the HMMs trained with a MAP adaptation procedure.

When considering the fusion of both modalities, the HMM modelling is leading consistently to better accuracy than the baseline GMM modelling. The HMM system also shows better performance when the training is performed using a MAP adaptation instead of the classically used EM training. As already observed in our previous works, results show a clear impact of skilled forgeries on the performances.

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

[2] We can note that in the configuration HMM-MAP-skilled, the z-norm fusion performs significantly worse than the sum fusion. 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 result. The application of the z-norm is, by nature, aligning the score distributions of both modalities through mean normalization. While this is beneficial when fusing scores that lies in different ranges, the z-norm is also giving equal importance to each modalities through the standard deviation normalization. This is of course not favorable in the case of systems showing very different individual performances.

modelling techniques 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 signature data will be available, experiments will be conducted according to a development/evaluation set framework.

VI. ACKNOWLEDGMENTS

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] 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.

[2] 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.

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

[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 (MRCS'06)*, pages 2–9, Istanbul, September 2006.

[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)*, pages 1110–1119, Rye Brook, NY, 2005.

[7] Y. Watanabe, K. Iwata, R. Nakagawa, K. Shinoda, and S. Furui. Semi-synchronous speech and pen input. In *32nd International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, 2007.

[8] 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).

[9] 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 (RASC 2006)*, pages 496–501, Canterburry, Kent, United Kingdom, 2006.

[10] 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.

[11] J. Hennebert, A. Humm, and R. Ingold. Modelling spoken signatures with gaussian mixture model adaptation. In *32nd International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, 2007.

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

[13] 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.

[14] 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.

[15] J. G. A. Dolfing. *Handwriting Recognition and Verification, a Hidden Markov Approach*. PhD thesis, Philips Electronics N. V., 1998.

[16] M. Liwicki and H. Bunke. Hmm-based on-line recognition of handwritten whiteboard notes. In *Proceedings of the 10th Int. Workshop on Frontiers in Handwriting Recognition*, pages 595–599, 2006.

[17] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals Of Speech Recognition*. Prentice Hall, 1993.

[18] J. Richiardi and A. Drygajlo. Gaussian mixture models for on-line signature verification. In *Proc. 2003 ACM SIGMM workshop on Biometrics methods and applications*, pages 115–122, 2003.

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