Spoken signature for user authentication

Andreas Humm
Université de Fribourg
Boulevard de Pérolles 90
1700 Fribourg, Switzerland
E-mail: andreas.humm@unifr.ch

Jean Hennebert
Université de Fribourg
Boulevard de Pérolles 90
1700 Fribourg, Switzerland
and
Institute of Business Information Systems HES-SO Valais
TechnoArk 3
3960 Sierre, Switzerland

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

Abstract. *We propose a new user authentication system based on spoken signatures, where online signature and speech signals are acquired simultaneously. The main benefit of this multimodal approach is better accuracy at no extra cost for the user in terms of access time or inconvenience. Another benefit lies in a better robustness against intentional forgeries due to the extra difficulty for the forger to produce both signals. We set up an experimental framework to measure these benefits on MyIDea, a realistic multimodal biometric database publicly available. More specifically, we evaluate the performance of state of the art modeling systems based on Gaussian mixture models (GMM) and hidden Markov models (HMM) applied independently to the pen and voice signal, where a simple rule-based score fusion procedure is used. We conclude that the best performance is achieved by the HMMs, provided that their topology is optimized on a per user basis. Furthermore, we show that more precise models can be obtained through the use of maximum a posteriori probability (MAP) training instead of the classically used expectation maximization (EM). We also measure the impact of multisession scenarios versus monosession scenarios, and the impact of skilled versus unskilled signature forgeries attacks. © 2008 SPIE and IS&T.* [DOI: 10.1117/1.2898526]

1 Introduction

Signatures are widely used by humans in their daily transactions and interactions. In past decades, many automated authentication systems based on signature have been proposed (see Sec. 2 for a survey). However, we still see few deployments of signature systems in commercial applications, while iris or fingerprint systems are currently numerous. Four factors can be identified to explain this gap. First, signature production is behavioral, therefore variable by nature. A user does not sign two times in the exact same way, especially when time is spent between two signature samples.[1] Second, uniqueness is not guaranteed, as most signatures are based on the characters included in the name of the user. Third, a signature can be easily reproduced by intentional forgers under certain assumptions (such as using dedicated training software), even when online signature systems are used.[2–4] Finally, the signature signal is dependent to the acquisition context and sensor, where mismatched conditions usually decrease performance.[3]

A potential direction to compensate for these factors is to augment the signature signal with other biometric modalities. Such multimodal systems have recently aroused a growing interest among the industrial and scientific communities thanks to the potential increase in accuracy and better robustness against forgeries. Several works have taken this direction using a speech signal to complement the signature (see Sec. 2 for a survey of such systems). This combination is further motivated because speech and signature are two well accepted modalities that are nonintrusive and natural to produce. While there is a clear gain in terms of accuracy, all these approaches suffer from an additional cost in terms of acquisition time, as these modalities are acquired sequentially.

The novelty of our proposal is to record simultaneously a signature signal with a speech signal. These so-called spoken signatures can be acquired simply by asking the user to speak the content of the signature. Our motivations for carrying out such a synchronized acquisition can be summarized as follows. First, we leverage the advantages of multimodal biometric systems while keeping the acqui-

Paper 07108SSRR received Jun. 15, 2007; revised manuscript received Nov. 14, 2007; accepted for publication Nov. 15, 2007; published online Apr. 2, 2008.
1017-9909/2008/17(1)/011013/11/$25.00 © 2008 SPIE and IS&T.

sition time equivalent. Second, the synchronized acquisition will probably give better robustness against intentional imposture. Indeed, imitating simultaneously the voice and the writing of somebody imposes a larger cognitive load than for each modality taken separately. Finally, the synchronization patterns or the intrinsic deformation of the inputs (mainly the slow down in speech) may be dependent on the user, therefore potentially adding an extra piece of useful biometric information.

We address the following questions.

- Are signatures readable? In other words, is it possible, from a practical and cognitive point of view, to ask the user to speak and sign at the same time?
- What is the gain of performance using a spoken signature instead of a signature alone?
- Does the process of speaking while signing add variability to the signature signal?
- Does a spoken signature decrease the ability of the forger to produce good imitations?

We are also interested in building a simple and efficient approach to model these multimodal signals using state of the art modeling strategies such as Gaussian mixture models (GMMs) or hidden Markov models (HMMs).

The rest of this work is organized as follows. A survey of automatic signature verification systems and related approaches based on multimodal biometric systems using speech and signature is presented in Sec. 2. We give in Sec. 3 an overview of MyIDea, the database used for this work. The data acquisition procedure, evaluation protocols, and the feedback collected from a usability survey are presented. In Sec. 4, we present our modeling system based on a score-level fusion of GMMs or HMMs. More details are provided on the training procedures and on the selection of the HMM topologies. Section 5 presents the experimental results and the related discussions. Finally, conclusions and future work are presented.

2 Related Work

Numerous automatic signature verification systems have been investigated in the past. General reviews can be found in Refs. 5–8. Speaker verification systems have also raised a great deal of interest, essentially due to the wide proliferation of mobile phones and automated telephony services.[9,10]

Regarding the signature signal, a verification can be performed off-line, where only a scanned image of the signature is available, or online, where temporal and spatial information about the handwriting is available. Various approaches have been investigated to extract signature features and to model them. For example, methods based on dynamic time warping, neural networks, GMMs, or HMMs have been presented. Without being exhaustive, we review some of the key approaches in the next paragraphs.

In Ref. 11, a dynamic time warping approach is described, where global and local features are extracted from the slope of the signature and stored in a string representation. The similarity between an input signature and the reference set is then computed by using a string matching measure.

Neural network approaches have been and still are popular techniques in machine learning, mainly thanks to their ability to model nonlinear phenomena and to cope with complex high dimensional feature input space. In Ref. 12, an architecture based on multilayer perceptrons trained with cepstral coefficients derived from linear predictor coefficients of the writing trajectories is also presented. In Ref. 13, a signature verification system interestingly couples wavelet-based features and back-propagation neural networks to reach good verification performances.

HMMs have been applied to signature verification for quite a long time.[14–18] The large interest brought to HMM systems is due to the good correspondence between the stroke-based nature of signatures and the modeling of these strokes through a sequence of HMM states. An interesting issue with HMMs is related to the choice of the topology and to the number of states that should be in principle dependent to each user. For example, in Refs. 19 and 20, a method is proposed to map trajectory angles to HMM states. Another issue is related to the modeling of the features in each states of the HMM. Usually, a mixture of Gaussians is used as estimators of the continuous probability density function of the features associated to a given state. The best strategy to determine state model orders, i.e., the number of Gaussian mixtures in each state, has been also addressed in different works (see for example Ref. 21).

Recently, a regain of interest have been brought to GMMs that are actually a degraded version of HMMs, where there would be only one state. While GMMs are not anymore implicitly modeling the sequence of stroke they are simpler to use and still present robust modeling capabilities of the features. In Ref. 21, GMMs are reporting good performance for online signature verification and compare fairly to a more complex HMM-based system. In Ref. 22, another approach attempts to mix HMM and GMM to model respectively local and global features.

In 2004, the Signature Verification Competition (SVC2004) was organized with the objective to evaluate and compare the performances of the different signature verification algorithms.[23,24] This initiative was taking motivation toward establishing common benchmark databases and evaluation rules.

As can be seen in Table 1, state of the art performance of available signature verification systems lies generally between 1 and 6% equal error rate. Note that a comparison of these different signature verification systems is a difficult task, since datasets and testing conditions may vary dramatically. There are impacts on the performance due to the number of enrollment sessions, the quality of acquisition platform, the type of forgery, the modeling strategy, and the setting of the rejection threshold.

Now regarding multimodal approaches, several related works have already shown that using speech and signature modalities together improves significantly the authentication performances in comparison to systems based on signature alone.

In Ref. 25, a tablet PC system based on online signature and voice modalities is proposed to ensure the security of electronic medical records. Tablet PCs are already used by many health care professionals to have a patient's record readily available when prescribing or administering treatment. In this system, the user claims his identity by saying his first and last name, which are recognized using speech recognition. The same waveform is then used with a

Table 1 State of the art performance of signature verification systems.

Reference	Performance
11	2.8% (FRR) and 1.6% (FAR)
12	4% (EER)
13	0.0% (FRR) and 0.1% (FAR)
14	1.0 to 1.9% (EER)
16	2.5% (EER)
17	0.95% (EER)
19	2.78% (EER)
21	1.7% (EER)
22	6.7% (EER)
23 (task 1)	2.84% (EER)
23 (task 2)	2.89% (EER)

speaker verification system based on GMMs to produce a score. In this way, the identification and verification steps are performed simultaneously. A signature is then acquired and a dynamic time warping verification system is used to produce a score. Speech and signature scores are then normalized and fused.

In Ref. 26, an online signature verification system and a speaker verification system are also combined. Both subsystems use HMMs to produce independent scores that are then fused together. Results are reported for the two subsystems evaluated separately and for the global system. Better accuracy is reported for the fused bimodal system. For this test, fictitious users are built by randomly associating signature and speech samples from two independent databases, namely, Philips' online signature database, Polyphone, and Polyvar.

In Ref. 1, 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. This last work uses the BIOMET database,[27] where the speech and signature data are recorded from the same user.

In Ref. 28, 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 approach lies in the acquisition procedure. In our case, the speech and signature data streams are recorded simultaneously, asking the user to actually say the content of the signature. 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.

3 MyIDea Database

Spoken signatures have been acquired in the framework of the MyIDea biometric data collection,[29,30] lead at the University of Fribourg, Switzerland. MyIDea is a multimodal database that contains many other modalities such as fingerprints, talking faces, etc. The "set 1" of MyIDea is publicly available for research institutions. (The dataset used to perform the experiments reported in this work has been given the reference MYIDEA-CHASM-SET1 by the distributors of MyIDea.)

3.1 *Data Acquisition*

About 70 users have been recorded over three sessions, with an interval of time between sessions ranging from one week to several months. This procedure eased the planning of recordings and actually corresponds to real-life situations where users get authenticated at random frequencies. Signature alone and spoken signatures were recorded to assess the gain of the proposed procedure.

Regarding the equipment, signatures are acquired with a WACOM Intuos2 graphical tablet (WACOM Technology Corporation, Vancouver, Washington). A WACOM InkPen, similar to a regular pen that produces ink, is used to write on regular sheets of paper positioned on the tablet. The writing feeling is therefore close to the one of writing on a standard sheet of paper. This hardware is similar to the one used for other databases such as BIOMET,[27] MCYT,[31] and IAM[32] databases. The speech waveform is recorded with a standard computer headset microphone (Creative HS-300) at 16 kHz and coded linearly on 16 bits. A dedicated software has been developed to perform the synchronized acquisition of pen and speech data. For each sampled text point, the tablet records five parameters at a frequency of 100 Hz: x,y coordinates, pressure, and the two angles, azimuth and altitude. We further record time stamps for each data packet sent by the tablet. Time stamps are also recorded for the beginning and end of speech acquisition. This procedure allows us to precisely synchronize speech and pen streams, even when pens-up is occurring while signing.

Figure 1 illustrates the pen (upper part) and speech signals (bottom part) of a spoken signature. For sake of clarity, the azimuth and elevation angles are not represented. The gray area on the figure corresponds to pens-up, i.e., moments when the user lifts the pen out of the range of the tablet. During pens-up, the tablet does not send out any packets, and a simple linear interpolation of the text points is applied. One can observe on this figure that some speech events have synchronized starting times, with some sets of strokes located where the pressure is increasing. The signature signals correspond to the signature displayed in Fig. 2. In this specific example, the signature is composed of the first and last names of a user, but we must note that, for most other users, only the last name is available.

In the first step, single signatures are recorded without the speech part. As illustrated in Fig. 3, template papers are used for recording signatures. During each session, subjects sign six times using the cells on the template. The two remaining cells are used in the case of missed signatures that need to be redone. During each session, the subject is also asked to imitate the signature of another subject. To do this, the subject has access, during a limited time of two

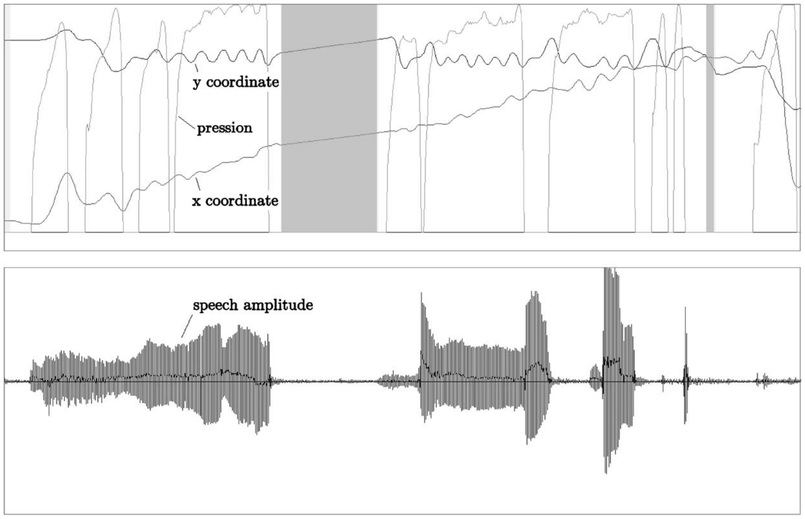


Fig. 1 Synchronized representation of signature and speech signals. The upper part of the graph shows the evolution of *x* and *y* coordinates and the pression *p* as a function of time. Azimuth and elevation angles are not displayed for sake of clarity. The bottom part shows the speech amplitude as a function of time. Signature and speech signals are synchronized thanks to the time stamps.

minutes, to the static image of the signature to imitate. Six imitations of the signature of another subject are performed per session for a total of 18 impostor signatures after the three sessions.

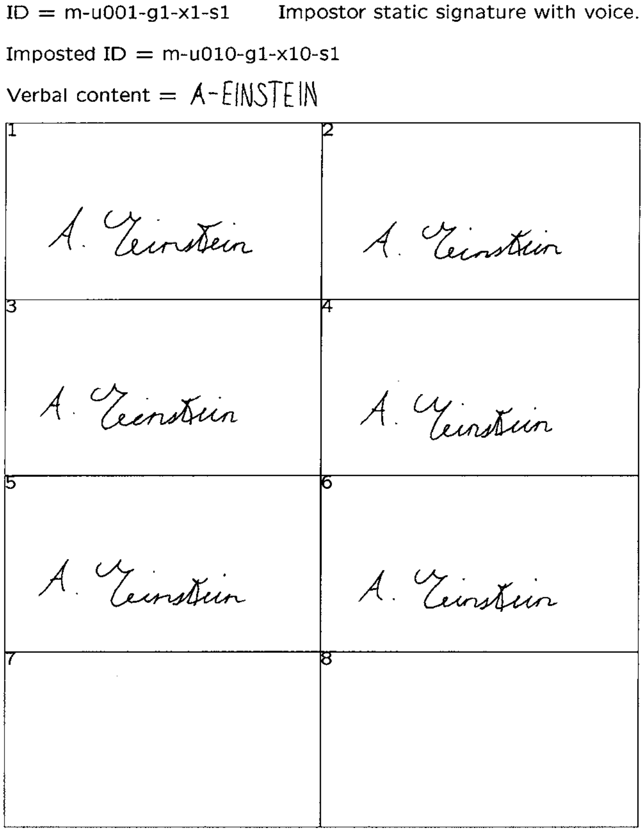
In the second step, spoken signatures are recorded by asking the user to say, in a synchronized manner, the actual content of the written symbols. Prior to the recording, the subject is allowed to train for a few spoken signatures to get used to the procedure. Six signatures are required using similar templates as explained before. The subject is also asked to imitate the spoken signature of another subject, different to the one imitated in the first step. The subject has access to the static version of the signature and to the verbal content of the signature to imitate. In other words, access to the voice recording is not given to the impostor and there is no intention to imitate the voice.

3.2 *Comments on the Acquisition and Usability Survey*

During the acquisition campaign, all 70 users without exception were able to perform the spoken signature acquisition. The fact that they had to speak and sign at the same time did not prevent any acquisition from happening. We also observed that the speech production is generally faster than the signature. The speech signal is therefore deformed due to its slow down, and resynchronization occurs at specific times. A visual inspection showed that most of the users synchronized the written symbols with syllables. While the deformation of the speech signal was clearly identified, we did not visually observe any deformation of the signature signal. Many signatures contained some pre- or postflourishes that were spontaneously not said by the user. Very few users had signatures containing only flourishes or nonreadable signs. These users were then asked to simply utter their name while signing.

Fig. 2 Sample of a signature including first and last names.

A simple usability survey was organized, where each subject was asked to answer some questions about the acquisition of spoken signatures. For each question, subjects



ID = m-u001-g1-x1-s1 Impostor static signature with voice.

Imposted ID = m-u010-g1-x10-s1

Verbal content = A-EINSTEIN

1
A. Einstein
2
A. Einstein
3
A. Einstein
4
A. Einstein
5
A. Einstein
6
A. Einstein
7
8

Fig. 3 Example of a signature acquisition page.

Table 2 Questions and results of the usability survey.

	Question	Result
1	Did you find it simple/difficult to write on a tablet?	s●○○○○○d
2	Did you find it simple/difficult to speak and sign at the same time?	s○○○●○○d
3	Do you think that the act of speaking and writing at the same time affected your capacities to imitate the writing?	y○●●○○○n

were asked to answer according to a predefined scale. The questions and the respective average answers are listed in Table 2. The main conclusions of the survey are the following.

- A large majority of users found it easy to write on a tablet. The use of regular pen and paper may have help to reach this result.
- Users ranked as average the difficulty of signing and speaking at the same time. This is most probably due to the extra level of concentration needed to perform such acquisitions. Also, a signature usually contains pre- or postflourishes on which users cannot utter anything that is potentially disturbing. However, all users were able to sign and utter the content of their signature.
- Interestingly, users felt that the act of speaking and signing at the same time affected their capacities to imitate signatures. While this feeling is of course not related to the real capacity of the system to reject forgers, the perceived security of the procedure is potentially higher than for monomodal systems.

According to the fact that all users were able to perform the acquisitions, and considering the prior answers given to the survey, our current conclusion is that such bimodal acquisitions are acceptable from a usability point of view.

3.3 *Evaluation Protocols*

The scenario of spoken signatures is similar, in essence, to password-based systems, where the signature and speech content remains the same from access to access. Two assessment protocols have been defined on MyIDea with the objective of being as realistic as possible.[33] The first one is called "without time variability" (monosession scenario), where user templates are built using five spoken signatures of the first session. For testing, the remaining signature of the first session is used. The same procedure is repeated for the other five signatures and for sessions two and three, leading to a total of 70 users × 1 access × 6 repetitions × 3 sessions = 1260 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 × 69 accesses × 6 repetitions × 3 sessions = 86,940 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 × 18 accesses × 6 repetitions × 3 sessions = 22,680 skilled forgeries. The second protocol is called "with time variability" (multisession scenario), 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 × 12 accesses = 840 genuine tests. Random and skilled impostor attempts are performed in the similar manner as for the monosession protocol, with the distinction that models are here trained on the first session only, giving a total of 70 users × 69 accesses = 4830 random forgeries, and 70 users × 18 accesses = 1260 skilled forgeries. The amounts of tests mentioned before are approximate, as some users did not complete all sessions.

4 System Description

As illustrated in Fig. 4, at the current stage our system models independently the speech and signature signals. Features are first extracted from both stream of data using standard feature extraction front-ends. Feature vectors are then modeled using state of the art statistical models based on either GMMs or HMMs. Finally, the speech score and the signature score are fused to obtain the spoken signature score.

4.1 *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 in Ref. 17. Each feature vector includes:

- the absolute speed and acceleration, the speed and acceleration in x and y directions, and the tangential acceleration;
- the angle α of the absolute speed vector, its cosine and sine, the derivative of α, and its cosine and sine;
- the pressure and the pressure derivative;
- the azimuth and elevation angles of the pen and their derivatives;
- the curvature radius;

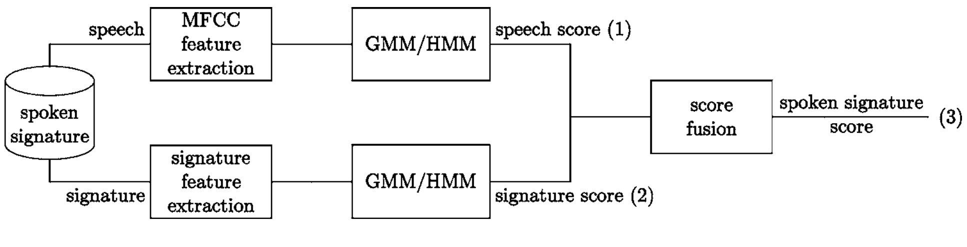


Fig. 4 Spoken signature system.

- the normalized coordinates $[x(n)-x_g, y(n)-y_g]$ relative to the gravity center (x_g, y_g) of the signature;
- the length-to-width ratio of windows of five and seven points centered on the current point, and the ratio of the minimum over the maximum speed on a window of five points centered on the current point.

The signature features are further 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.[34] 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 signal. This silence removal component uses a classical energy-based speech detection module based on a bi-Gaussian model.[35] MFCC coefficients are mean and standard deviation normalized using normalization values computed on the speech part of the data. Delta features were not used, as they did not lead to improvement of the results.

4.2 *Gaussian Mixture Models System*

GMMs are used to model the likelihoods of the features extracted from the signature and from the speech signal. One could argue that in this case, GMMs are actually not the most appropriate models, as they are intrinsically not capturing the time-dependent specificities of speech and signature. However, GMMs have been reported to compare reasonably well to HMMs in terms of signature verification,[21] and are often considered as baseline systems in speaker verification. Furthermore, GMMs are well-known flexible modeling tools able to approximate any probability density function. With GMMs, the probability density function $p(x_n|M_{\text{client}})$ 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:

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

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. The Gaussian weights w_i satisfy the constraint $\Sigma_{i=1}^{M} w_i = 1$. The Gaussian densities $\mathcal{N}$ have the form:

$$\mathcal{N}(x_n, \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \times \exp\left[-\frac{1}{2}(x_n-\mu_i)'\Sigma_i^{-1}(x_n-\mu_i)\right]. \tag{2}$$

In our case, we use diagonal covariance matrices as approximations of the full covariance matrices. This approximation is classically done when using GMMs for two reasons. First, it allows reduction of 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 S_{client} for the sequence of feature vectors $X=\{x_1, x_2, \ldots, x_N\}$ is computed with:

$$S_{\text{client}} = p(X|M_{\text{client}}) = \prod_{n=1}^{N} p(x_n|M_{\text{client}}). \tag{3}$$

Here, the likelihood score S_{world} of the hypothesis that X is not from the given client is estimated using a world GMM model M_{world} trained by pooling the data of many other users.[36] The decision whether to accept or to reject the claimed user is performed comparing the ratio R_{client} of client and world score against a global threshold value T. The ratio is here computed in the log-domain with:

$$R_{\text{client}} = \log(S_{\text{client}}) - \log(S_{\text{world}}). \tag{4}$$

The training of the client and world models is usually performed with the expectation-maximization (EM) algorithm[37] 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).[38] The MAP training procedure is known to perform well in the case of few training data, which is the case in our approach.

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 iterative process of the EM training is stopped when the relative increase of the accumulated likelihood is below a threshold value (0.1% in our case). As it is classically applied when training GMMs with few data, we also prevent the variances to converge below a given floor value (0.01 in our settings). The world model is trained by pooling half of 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). 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.

4.3 *Hidden Markov Models System*

HMMs have been extensively used to model the likelihoods of the features extracted from signatures[14,17] and from the speech[34] signals. Our motivations to use HMMs are multiple. First, they are the natural extension of the previously presented GMM-based systems. Second, they allow more detailed modeling of the data, incorporating sequential information of the strokes for signature and of the phonemes for speech (time-dependent specificities of speech and signatures).

As for the GMMs, 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

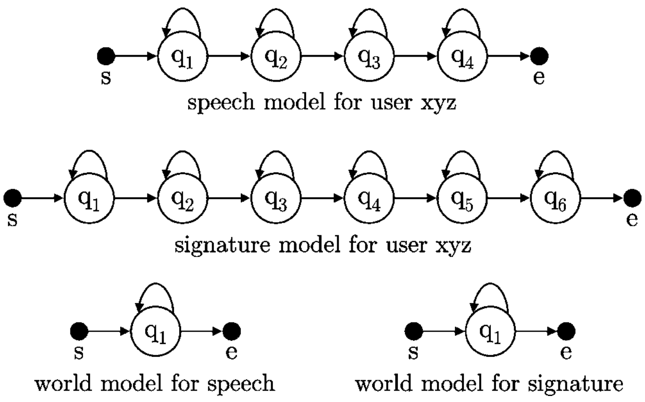

s
q1
q2
q3
q4
e
speech model for user xyz
s
q1
q2
q3
q4
q5
q6
e
signature model for user xyz
s
q1
e
world model for speech
s
q1
e
world model for signature

Fig. 5 HMM topology.

simplifying assumption of HMM-based modeling (see for example Ref. 34), the likelihood of X given the model M_{client} can be written

$$S_{\text{client}} = P(X|M_{\text{client}}) = \sum_{\text{all paths}} \prod_{n=1}^{N} \underbrace{P(x_n|q_n,M_{\text{client}})}_{\text{em. probs}} \underbrace{P(q_n|q_{n-1},M_{\text{client}})}_{\text{trans. probs}}, \quad (5)$$

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 $P(x_n|q_n,M_{\text{client}})$ is the so-called emission probability and represents the probability to observe a feature vector x_n when visiting state q_n. The value $P(q_n|q_{n-1},M_{\text{client}})$ is the transition probability and represents the probability to go from state q_{n-1} to state q_n in the HMM. The likelihood of Eq. (5) can be efficiently computed using a forward-backward algorithm. Alternatively to Eq. (5), 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. In the work reported here, we have decided to use the Viterbi criterion to compute the likelihoods. We have also chosen to use continuous HMMs, where the emission probability is modeled using a probability density function computed with weighted Gaussian mixtures as expressed in Eqs. (1) and (2). We also use diagonal covariance matrices as approximations of the full covariance matrices.

In a similar way as in the GMMs system, we also compute the likelihood score of the hypothesis that X is not from the given client using a world model M_{world}. As this world model is trained by pooling the data of many other users, there is no reason to attempt to model any sequence of strokes or phonemes. We therefore use a single-state HMM as illustrated on Fig. 5, which is actually nothing more than a GMM system. The decision whether to reject or accept the claimed user is performed as before, comparing the ratio of client and world score against a global threshold value T.

The training of each HMM is done in several iterations. In each iteration, two steps are performed. In the first step, a Viterbi forced alignment is performed[34] to find the most likely sequence of states given the parameters of the HMM. In the second step, the Gaussian densities of each state are re-estimated with the expectation-maximization (EM) algorithm. The number of Gaussians in the mixture is also increased using a simple binary splitting procedure during training. Transition probabilities are also updated by simply counting the accumulated number of passages on a given transition. As for the GMMs, the world model is trained by pooling half of the available genuine accesses in the database, and the skilled forgeries attempts are excluded from this set. Alternative to the EM-based training, we also investigated the use of the MAP criterion that is here applied on a client model, where the HMM states are duplicated from the single-state world model.

As illustrated in Fig. 5, 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 modeling speech, as states will naturally correspond to the sequence of phonemes. For the signature, the state sequence is modeling the sequence of strokes. We investigated different strategies regarding the number of states used in the HMM, and concluded that the best configuration was obtained when using a variable number of states for each user. This result is comprehensible, as users have different sizes of signatures and also different numbers of phonemes in their name. Also, we measured that the optimal number of states for the speech part and for the signature part are different. This is again understandable, as the respective signal production processes are different.

As we do not know *a priori* what is actually pronounced and written in spoken signature, we have chosen to compute the number of states proportionally to the number of signature and speech feature vectors. For the signature part (si), the number of states K_{si} in the HMM is computed proportionally to the average number N_a of feature vectors in the available genuine signatures:

$$K_{\text{si}} = \frac{N_a}{\alpha}, \quad (6)$$

where α is a dividing factor that needs to be tuned by simply varying the alpha and choosing the optimal one that leads to the best performance. In a similar way for the speech part (sp), we compute the number of states K_{sp} in the HMM, taking into account the number of speech frames instead of the number of signature points.

4.4 *Score Fusion*

We opted to use a rule-based score fusion[16,18] using a simple summation of the signature and speech log-likelihood ratios with

$$R_c = R_{c,\text{sp}} + R_{c,\text{si}}. \quad (7)$$

This choice was motivated by the fact that both subsystems are very similar. They are indeed taking as input approximately the same amount of feature vectors at a frequency of 100 Hz for both modalities, and they use the same modeling techniques based on GMMs or HMMs. Furthermore, Eq. (7) is a reasonable approximation if we assume that the local observations of both subsystems are

Table 3 Summary of signature-alone and spoken signature (SS) results with GMMs. Multisession and monosession protocols are used and results are reported for random and skilled forgeries.

Training	EM (% EER)				MAP (%EER)			
Protocol	Monosession		Multisession		Monosession		Multisession	
Forgeries	Random	Skilled	Random	Skilled	Random	Skilled	Random	Skilled
Signature alone	2.0 %	4.4 %	4.6 %	9.8%	0.6 %	3.1 %	2.1 %	9.4%
SS—signature part	1.7 %	4.3 %	5.5 %	9.0%	0.4 %	2.9 %	2.6 %	7.4%
SS—speech part	2.8 %	3.7 %	7.6 %	12.7%	1.9 %	5.3 %	5.2 %	13.5%
Spoken signature	0.8 %	1.6 %	3.7 %	5.8%	0.1 %	1.0 %	1.8 %	5.6%

independent. This assumption is probably correct, as both modalities are produced by very distinct processes. However, as users are intentionally trying to synchronize their speech with the signature signal, the observations taken globally as a sequence should show some dependence. For this reason, it would be worthwhile to investigate a more advanced joint modeling such as, for example, asynchronous HMM,[39] but this is out of the scope of this work. A personalized normalization has been done in Ref. 17. Classifier score fusion with multilayer perceptron or a support vector machine could also be applied, or a trained-based score recombination such as weighted sum fusion. While improving further the performances of our system, we also tried applying a weighted sum.[40,41] But an optimization of the weights is optimistic, as it was done *a posteriori* on the scores. However, such approaches need more parameters to estimate using typically an independent dataset, and the limited size of the database did not allow this.

5 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 for the testing set. Tables 3 and 4 summarize the EER for the GMM and HMM systems. In these tables, the columns correspond to the different configurations of the training algorithm (EM versus MAP), of the protocols (multisession or monosession) and of the forgeries (random or skilled). The first line entitled "signature alone" corresponds to our baseline signature verification system, where the user is not speaking at the same time. The three remaining lines correspond to results obtained using spoken signatures, where results are detailed per modality on lines SS—signature part, and SS—speech part.

Most of the system parameters have been optimized to obtain the results reported in these tables. For the GMMs, we mainly investigated the model orders trying 8, 16, 32, 64, and 128 Gaussians. The best configuration for the EM training procedure was 16 Gaussians for the client and 16 Gaussians for the world model. For the MAP adaptation training procedure, the best configuration was with 128 Gaussians for the world model, from which the client models were adapted. For the HMM, the best configuration was obtained using 16 Gaussians in each state of the client and world models, for both EM and MAP. We also computed the best α value from Eq. (6) that condition the number of states used in the HMMs. We found out that the optimal value of α for our best results using MAP adaptation is 25 for SS—signature part, respectively, and 75 for SS—speech part. Note that the optimal value of α is different for the speech and signature parts leading to HMMs, with more states for the signature part than for the speech part. This

Table 4 Summary of signature-alone and spoken signature (SS) results with HMMs. Multisession and monosession protocols are used and results are reported for random and skilled forgeries.

Training	EM (% EER)				MAP (%EER)			
Protocol	Monosession		Multisession		Monosession		Multisession	
Forgeries	Random	Skilled	Random	Skilled	Random	Skilled	Random	Skilled
Signature alone	1.9 %	4.0 %	4.3 %	9.0%	1.2 %	3.2 %	2.2 %	5.9%
SS—signature part	1.4 %	2.8 %	4.2 %	7.8%	0.5 %	2.0 %	2.8 %	5.6%
SS—speech part	1.9 %	2.0 %	9.2 %	14.5%	1.3 %	4.1 %	5.4 %	12.6%
Spoken signature	1.0 %	1.6 %	3.5 %	5.6%	0.1 %	0.8 %	1.5 %	4.2%

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.

The most important results can be summarized in the following list.

1. **Comparison of signature and spoken signature**: as a general comment on the approach, spoken signatures brings systematically a clear and significant improvement in comparison to the results obtained with signature alone. Also, modeling independently the signature and speech signal proves to be a simple and efficient strategy, even though the sum-based-fusion procedure is very straightforward.
2. **Impact of speaking while signing**: for all configurations, the results show no significant difference of performance between signature alone and the signature part of spoken signatures. We can reasonably conclude from this that the process of speaking while signing does not seem to degrade the signature signal by adding extra variabilities.
3. **Impact of spoken signatures on intentional forgeries**: the skilled forgeries that are produced for the signature-alone system degrade the results more than the skilled forgeries produced with the spoken signature procedure. Even if the difference is not so large, the increase of the EER is systematic for all configurations. We can then conclude that spoken signatures seem to decrease, to some extent, the ability of forgers to produce stronger imitations.
4. **Impact of multisession accesses**: the multisession protocol (with time variability) shows a systematic and significant drop of performance in comparison to the monosession protocol (without time variability). This conclusion is valid for both signature and speech modalities, and is therefore also reflected in the spoken signature results. This result is the direct consequence of using behavioral biometrics such as signature or speech, where the production process is dependent to the context and state of the user.
5. **Comparison of random and 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.
6. **Comparison of GMM to HMM**: for most of the configuration, HMM modeling leads to slightly better accuracy than GMM modeling. However, this gain has to be balanced with the fact that HMMs are more complex to set up and require tuning of an extra parameter linked to the number of states.
7. **Comparison of EM to MAP**: as it was already reported in many previous works, GMMs benefit significantly from a MAP adaptation instead of a full EM training. Interestingly, we see the same tendency for 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.
8. **Comparison of signature and speech**: for most configurations, the signature modality performs better than speech. Signatures are probably more discriminative and more stable than speech for the protocols and quantity of data used in these experiments.

6 Conclusions and Future Work

We present a new user authentication system based on spoken signatures, where online signature and speech signals are acquired simultaneously. A spoken signature verification system using independent modeling of both streams of data is presented and evaluated on MyIDea, a realistic multimodal biometric database. The system is composed of feature extraction modules dedicated to speech and signature signals, followed by statistical modeling tools based on GMMs or HMMs, and terminated by a simple sum-based score fusion module.

The results of the acquisition campaign of spoken signatures and the results of a survey conducted on the subjects clearly indicate that the procedure is viable from a cognitive and usability point of view. In other words, it is possible to ask the user to speak and sign at the same time. The results of the evaluation of the spoken signature system show that this multimodal approach leads to significantly better accuracy than signature-alone systems, at no extra cost for the user in terms of access time or inconvenience. The results also show that the process of speaking while signing does not add variability to the signature signal. Another measured benefit lies in better robustness against intentional forgeries, probably due to the extra cognitive load for the forger to reproduce both signals. Considering all these results, we can conclude that the proposed bimodal speech and signature approach seems then to be a viable alternative to systems using single modalities.

From a more technical point of view, we measure from our experiments that HMMs lead to slightly better models than GMMs, provided that their topology is optimized on a per user basis. Furthermore, we show that more precise models can be obtained through the use of MAP adaptation instead of the classically used EM training. Results also show that there is a clear impact of multisession accesses (time variability) and skilled forgeries on the performances.

In our future work, we plan to investigate the use of more robust modeling techniques against multisession accesses and forgeries. In this direction, we have identified potential modeling techniques such as time-dependent score fusion, joint modeling using asynchronous HMMs, etc. We also intend to compare our work with the most similar preceding work, namely, authentication based on the combination of nonsimultaneously-elicited speech and signature.

Acknowledgments

We warmly thank Asmaa El Hannani for her precious help with GMM-based systems. This work was partly supported by the Swiss NSF program “Interactive Multimodal Information Management (IM2)," as part of NCCR, by the EU BioSecure IST-2002-507634 NoE project and by the University of Fribourg.

References

1. B. Ly-Van, R. Blouet, S. Renouard, S. Garcia-Salicetti, B. Dorizzi, and G. Chollet, "Signature with text-dependent and text-independent speech for robust identity verification," in *Proc. Workshop Multimodal User Authentication (MMUA)* (unpublished), pp. 13–18 (2003).
2. A. Wahl, J. Hennebert, A. Humm, and R. Ingold, "Generation and evaluation of brute-force signature forgeries," in *Int. Workshop Multimedia Content Representation, Classification Security (MRCS)* (unpublished), pp. 2–9 (2006).
3. C. Vielhauer, *Biometric User Authentication for IT Security*, Springer, Berlin (2006).
4. J. Hennebert, R. Loeffel, A. Humm, and R. Ingold, "A new forgery scenario based on regaining dynamics of signature," in *Intl. Conf. Biometrics (ICB)* (2007) (unpublished), pp. 366–375.
5. R. Plamondon and G. Lorette, "Automatic signature verification and writer identification—the state of the art," *Pattern Recogn.* **22**(2), 107–131 (1989).
6. F. Leclerc and R. Plamondon, "Automatic signature verification: the state of the art—1989–1993," *Int. J. Pattern Recognit. Artif. Intell.* **8**(3), 643–660 (1994).
7. J. Gupta and A. McCabe, "A review of dynamic handwritten signature verification," James Cook Univ., Australia (1997).
8. G. Lorette and R. Plamondon, "Dynamic approaches to handwritten signature verification," in *Computer Processing of Handwriting*, World Scientific, Singapore (1990).
9. D. Reynolds, "An overview of automatic speaker recognition technology," in *Proc. IEEE Intl. Conf. Acoustics, Speech Signal Process.* (IEEE, 2002), Vol. 4, pp. 4072–4075.
10. F. Bimbot *et al.*, "A tutorial on text-independent speaker verification," *EURASIP J. Appl. Signal Process.* **4**, 430–451 (2004).
11. A. Jain, F. Griess, and S. Connell, "Online signature verification," *Pattern Recogn.* **35**(1), 2963–2972 (2002).
12. Q. Z. Wu, I. C. Jou, and S. Y. Lee, "On-line signature verification using lpc cepstrum and neural networks," *IEEE Trans. Syst. Sci. Cybern.* **27**(B), 148–153 (1997).
13. D. Z. Lejtman, "On-line handwritten signature verification using wavelets and back-propagation neural networks," in *ICDAR'01: Proc. IEEE Document Anal. Recog.* (IEEE, 2001), pp. 992–996.
14. J. Dolfing, E. Aarts, and J. van Oosterhout, "On-line signature verification with hidden markov models," in *Proc. IEEE Patt. Recog. ICPR* (IEEE, 1998), Vol. 2, pp. 1309.
15. L. Yang, B. K. Widjaja, and R. Prasad, "Application of hidden markov models for signature verification," *Pattern Recogn.* **28**(2), 161–170 (1995).
16. R. S. Kashi, J. Hu, W. L. Nelson, and W. Turin, "A hidden Markov model approach to online handwritten signature verification," *Int. J. Doc. Anal. Recg.* **1**(2), 102–109 (1998).
17. B. Ly Van, S. Garcia-Salicetti, and B. Dorizzi, "Fusion of HMMs likelihood and viterbi path for on-line signature verification," in *Biometrics Authentication Workshop*, pp. 318–331 (2004).
18. J. Firrez-Aguilar, L. Nanni, J. Lopez-Pealba, J. Ortega-Garcia, and D. Maltoni, "An on-line signature verication system based on fusion of local and global information," *Lect. Notes Comput. Sci.* **3546**, 523–532 2005).
19. D. Muramatsu and T. Matsumoto, "An hmm on-line signature verification algorithm," in *AVBPA03* (2003) (unpublished), pp. 233–241.
20. D. Muramatsu and T. Matsumoto, "An hmm on-line signature verification with pen position trajectories," in *IEEE Conf. on Artificial Intelligence (IC-AI'03)* (2003) (unpublished), pp. 299–303.
21. J. Richiardi and A. Drygajlo, "Gaussian mixture models for on-line signature verification," in *Proc. 2003 ACM SIGMM Workshop Biometrics meth. Appl.* (2003) (unpublished),, pp. 115–122.
22. L. Wan and B. Wan, "On-line signature verification with two-stage statistical models," in *Proc. IEEE Document Anal. Recog. (ICDAR'05)* (2005) (unpublished), pp. 282–286.
23. A. Kholmatov and B. Yanikoglu, "Identity authentication using improved online signature verification method," *Pattern Recogn. Lett.* **26**(15), 2400–2408 (2005).
24. D. Y. Yeung, H. Chang, Y. Xiong, S. George, R. Kashi, T. Matsumoto, and G. Rigoll, "SVC2004: first international signature verification competition," in *Proc. ICBA* (2004) (unpublished), pp. 16–22.
25. S. Krawczyk and A. K. Jain, "Securing electronic medical records using biometric authentication," in *Audio-Video-Based Biometric Person Authentication (AVBPA)*, Springer, Berlin, pp. 1110–1119 (2005).
26. M. Fuentes, D. Mostefa, J. Kharroubi, S. Garcia-Salicetti, B. Dorizzi, and G. Chollet, "Identity verification by fusion of biometric data: On-line signature and speech," in *Proc. COST 275 Workshop Advent Biometrics Internet* (2002) (unpublished), pp. 83–86.
27. S. Garcia-Salicetti, C. Beumier, G. Chollet, B. Dorizzi, J. L. les Jardins, J. Lunter, Y. Ni, and D. Petrovska-Delacrétaz, "Biomet: a multimodal person authentication database including face, voice, fingerprint, hand and signature modalities," in *Proc. 4th Intl. Conf. Audio Video-Based Biometric Person Authentication (AVBPA)*, Springer-Verlag, Berlin, pp. 845–853 (2003).
28. J. Koreman, A. Morris, D. Wu, S. Jassim, H. Sellahewa, J. Ehlers, G. Chollet, G. Aversano, H. Bredin, S. Garcia-Salicetti, L. Allano, B. L. Van, and B. Dorizzi, "Multi-modal biometric authentication on the securephone pda," in Second International Workshop on Multimodal User Authentication, Toulouse, 2006 (unpublished).
29. B. Dumas, C. Pugin, J. Hennebert, D. Petrovska-Delacrétaz, A. Humm, F. Evéquoz, R. Ingold, and D. Von Rotz, "Myidea—multimodal biometrics database, description of acquisition protocols," in *Proc. 3rd COST 275 Workshop (COST 275)* (2005) (unpublished), pp. 59–62.
30. J. Hennebert, B. Dumas, C. Pugin, D. Petrovska-Delacrétaz, A. Humm, F. Evéquoz, R. Ingold, and D. Von Rotz, "Myidea multimodal database," see http://diuf.unifr.ch/go/myidea (2005).
31. J. Ortega-Garcia, J. Fierrez-Aguilar, D. Simon, J. Gonzalez, M. Faundez-Zanuy, V. Espinosa, A. Satue, I. Hernaez, J. J. Igarza, C. Vivaracho, D. Escudero, and Q. I. Moro, "Mcyt baseline corpus: a bimodal biometric database," *IEE Proc. Vision Image Signal Process.* **150**, 395–401 (2003).
32. U. V. Marti and H. Bunke, "A full english sentence database for off-line handwriting recognition," in *Proc. 5th Intl. Conf. Document Anal. Recog. (ICDAR)* (1999) (unpublished), pp. 705–708.
33. A. Humm, J. Hennebert, and R. Ingold, "Combined handwriting and speech modalities for user authentication," Tech. Rep. 06-05, Univ. of Fribourg, Dept. of Informatics (2006).
34. L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ (1993).
35. J. Mariethoz and S. Bengio, "A comparative study of adaptation methods for speaker verification," in *Intl. Conf. Spoken Language Process.* (2002) (unpublished), pp. 581–584.
36. M. Martinez-Diaz, J. Fierrez, and J. Ortega-Garcia, "Universal background models for dynamic signature verification," in *IEEE Conf. Biometrics: Theory, Appl. Syst. (BTAS'07)*, pp. 1–6 (IEEE, 2007).
37. A. P. Dempster, N. M. Laird, and D. B. Rubin, *J. R. Stat. Soc.* **39**, 1 (1977).
38. D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digit. Signal Process.* **10**, 19–41 (2000).
39. S. Bengio, "Multimodal speech processing using asynchronous hidden markov models," in *Inf. Fusion* **5**(2), 81–89 (2004).
40. A. Ross, K. Nandakumar, and A. Jain, *Handbook of Multibiometrics*, International Series on Biometrics (Springer-Verlag, New York, 2006).
41. A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recogn.* **38**, 2270–2285 (2005).

Andreas Humm received a Master of Science degree in computer science from the University of Bern, Switzerland, in February 2005. He is currently working toward the PhD degree at the University of Fribourg, Switzerland. His research interests are in the fields of biometrics, statistical modeling applied to speaker verification and writer verification, and signature modeling.

Jean Hennebert received an electrical engineering degree from the Faculté Polytechnique de Mons, Belgium, in June 1993. In October 1998 he received the PhD degree in computer science from the Swiss Federal Institute of Technology, Switzerland. He then worked for 6 years as an IT system architect and independent consultant for different private companies. In 2004, he joined the DIVA group of the computer science department of the University of Fribourg, Switzerland, where he is in charge of teaching and research activities. His research interests are in the areas of statistical modeling applied to speech recognition, speaker verification, handwriting recognition, signature modeling, and image identification.

Rolf Ingold received a diploma degree of mathematical engineer from the Swiss Federal Institute of Technology, Switzerland, in January 1983. In February 1989, he received the PhD degree in sciences from the Swiss Federal Institute of Technology, Switzerland. Since 1989, he has worked as a professor in the DIVA group of the computer science department of the University of Fribourg, Switzerland, where he is in charge of teaching and research activities. His research interests include image processing and analysis, pattern recognition, document analysis and recognition, speech processing, multimodality, and user interfaces.