Combined Handwriting and Speech Modalities for User Authentication

Andreas Humm, *Student Member, IEEE*, Jean Hennebert, *Member, IEEE*, and Rolf Ingold, *Member, IEEE*

***Abstract*—In this paper, we report on the development of an efficient user authentication system based on a combined acquisition of online pen and speech signals. The novelty of our approach is in the simultaneous recording of these two modalities, simply 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. Another benefit comes from an increased difficulty for forgers willing to perform imitation attacks as two signals need to be reproduced. We are comparing here two potential scenarios of use. The first one is called *spoken signatures* where the user signs and says the content of the signature. The second scenario is based on *spoken handwriting* where the user is prompted to write and read the content of sentences randomly extracted from a text. Data according to these two scenarios have been recorded from a set of 70 users. In the first part of this paper, we describe the acquisition procedure, and we comment on the viability and usability of such simultaneous recordings. Our conclusions are supported by a short survey performed with the users. In the second part, we present the authentication systems that we have developed for both scenarios. More specifically, our strategy was to model independently both streams of data and to perform a fusion at the score level. Starting from a state-of-the-art-modeling algorithm based on Gaussian Mixture Models trained with an Expectation–Maximization procedure, we report on several significant improvements that are brought. As a general observation, the use of both modalities outperforms significantly the modalities used alone.**

***Index Terms*—Handwritten signal, multimodal biometrics, speech signal.**

I. Introduction

SIGNATURES and handwritings are natural gestures widely used by humans in their daily transactions and interactions. In the past decades, many automated authentication systems based on signature or handwriting have been proposed [1]–[3]. However, we still see few deployments of such authentication systems in commercial applications. On the other hand, iris or fingerprint systems are nowadays widely deployed. Four factors can be identified to explain this gap. First, signature or handwriting production is behavioral, therefore variable by nature. A user does not sign or write two times in the exact same way, particularly when time is spent between two acquisitions. Second, uniqueness is weak, particularly for handwriting signals that are specifically produced to be understandable by all. Third, a signature or handwriting system can easily be attacked by intentional forgers trained to reproduce the signature or handwriting of a genuine user [4]–[6]. Finally, the performances are dependent on the acquisition context and on the sensor where mismatched conditions usually decrease performances [5].

A way to overcome these difficulties is to complement the pen-based signals with another biometric modality. Such similar multimodal biometric systems have recently raised a growing interest in the industrial and scientific communities. Some approaches are combining a face image and the speech signal [7]–[9], some other approaches are combining face and fingerprint [10]. As proposed in this paper and also suggested in other works (see Section II), we can also combine the pen and the speech signals. This combination is further motivated as these modalities are well accepted, nonintrusive, and natural to produce. While all these multimodal approaches report clear gains in terms of accuracy, they generally suffer from an additional cost in terms of acquisition time as the modalities are acquired sequentially.

We are here proposing a novel approach where the pen and speech modalities are simultaneously recorded. In order to do this, we simply ask the user to utter what she/he is writing. Our motivations for carrying out such a synchronized acquisition can be summarized as follows. First, we leverage on the advantages of multimodal biometric systems while keeping the acquisition time equivalent. Second, the synchronized acquisition will probably give better robustness against an intentional imposter. 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 slowdown in speech) may be dependent on the user, therefore potentially adding an extra piece of useful biometric information.

We are investigating two potential scenarios of use. The first one is called *spoken signatures* where the user signs and says the content of the signature. The second scenario is based on *spoken handwriting* where the user is prompted to write and read the content of sentences randomly extracted from a text. In regard to these two scenarios, we address two important questions. First, are these scenarios practicable? In other words, can we ask a user, from a practical and cognitive point of view, to read and write at the same time in an authentication framework? Second, what is the gain of performance using spoken signature or spoken handwriting instead of the modalities used

Manuscript received August 20, 2007; revised March 20, 2008. Current version published December 17, 2008. This work was supported in part 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. This paper was recommended by Guest Editor K. W. Bowyer.

A. Humm and R. Ingold are with University of Fribourg, 1700 Fribourg, Switzerland (e-mail: andreas.humm@unifr.ch; rolf.ingold@unifr.ch).

J. Hennebert is with University of Fribourg, 1700 Fribourg, Switzerland, and also with the University of Applied Sciences Western Switzerland, HES-SO // Wallis, 3960 Sierre, Switzerland (e-mail: jean.hennebert@unifr.ch; jean.hennebert@hevs.ch).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCA.2008.2007978

alone? Moreover, from a more technical point of view, we are interested to build a simple and efficient system to model these multimodal signals.

The rest of this paper is organized as follows. A survey of related approaches based on multimodal biometric systems using the pen and speech signals is presented in Section II. We give in Section III an overview of MyIDea, the database used for this paper. The data acquisition procedure, evaluation protocols, and the feedback collected from a usability survey are presented. In Section IV, we present our modeling system based on a score-level fusion of Gaussian Mixture Models (GMMs) or Hidden Markov Models (HMMs). More details are provided on the training procedures and on the selection of the HMM topologies. Section V presents the experimental results and the related discussions. Finally, conclusions and future works are presented.

II. Related Works

Many automated biometric systems based on speech alone have been studied and developed in the past [11], [12]. This profusion of systems is mainly due to a large number of potential applications on the telephone network. Similarly, numerous signature systems have also been investigated [1], [2]. For handwriting, fewer systems have been investigated, probably due to the lack of commercial interest. However, we can refer to [3] or [13] as examples of handwriting-based authentication systems.

When considering multimodal systems based on the pen and speech signals, several attempts have already been reported. In [14], an online signature verification system and a speaker verification system are combined to reach better authentication performances. In [15], a tablet PC system based on online signature and speech is proposed to ensure the security of electronic medical records. In [16], the SecurePhone project is presented where a multimodal biometric system using face, signature and speech signals is used to secure access and authenticate transactions on mobile devices.

The main difference between these works and our approach lies in the acquisition procedure. In our case, the speech and pen data streams are recorded simultaneously, asking the user to actually say the content of what she/he is writing.

It is also worth mentioning the work presented in [17], 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 written characters is used to enhance the recognition of the spoken part.

III. Spoken Signature and Spoken Handwriting Signals

A. MyIDea Database

Spoken signature and spoken handwriting data have been acquired in the framework of the MyIDea biometric data collection [18], [19]. 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 paper has been given the reference MYIDEA-CHASM-SET1 by the distributors of MyIDea. The signals 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 data is recorded at 16 kHz and coded linearly on 16 bits. To make the use of the tablet as natural as possible, we used a standard sheet of paper attached on the tablet as well as a pen (Intuos InkPen) that produces ink as a regular pen.

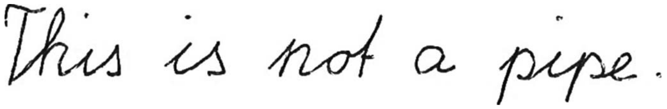


Fig. 1. Handwriting including one sentence.

Figs. 1 and 2 show the recording of the phrase "this is not a pipe," simultaneously written and said by the user. The three first diagrams of Fig. 1 show the time evolution of the tablet data streams coordinates x, y, and pressure p. The azimuth and elevation angles are not represented for sake of clarity. The last diagram represents the waveform of the content read by the user.

In our acquisition system, we are also recording the timestamps associated with each data sample. These timestamps allow us to precisely synchronize both streams of data. This is specially important as pens-ups are sometimes occurring. A pen-up happens when the user lifts the pen out of the range of the tablet that, in turns, do not send anymore sample. Such pens-up are shown by the gray areas on Fig. 2. We have to note that these kinds of events are not very frequent for signatures and are more frequent for handwriting as they refer to pens-up between two written words. Signatures are composed of the equivalent of one word, and therefore, such events are less frequent for signatures.

B. Comments on the Acquisition and Usability Survey

During the acquisition campaign, all 70 users without exception were able to perform the spoken signature and spoken handwriting acquisition. The fact that they had to read and sign at the same time did not prevent any acquisition to happen. We also observed that the speech production is generally faster than the signature. The speech signal is therefore deformed due to its slowdown 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 or handwriting signals. Many signatures contain some pre- or postflourishes that are spontaneously not said by the user. In our database, very few users are having signatures containing only flourishes or nonreadable signs. These users were then asked to simply utter their name while signing.

A simple usability survey was organized where each subject was asked to answer some questions about the acquisition procedure [20]. For each question, subjects were asked to answer according to a predefined scale. The questions and the

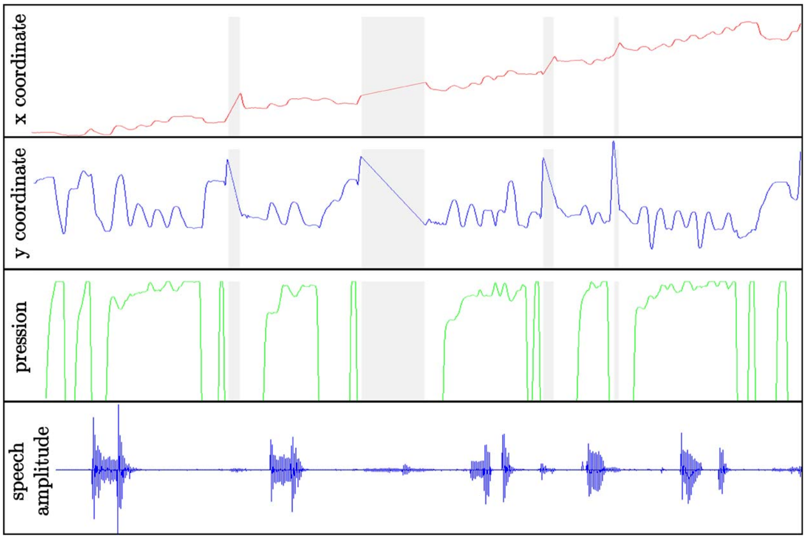


Fig. 2. Synchronized visualization of handwriting and speech signals.

TABLE I
QUESTIONS AND RESULTS OF THE USABILITY SURVEY

1	Did you find it simple/difficult to write on a tablet?	s ●◍○○○○ d
2	Do you think that you wrote faster, at the same speed or more slowly than usual (without simultaneous speaking)?	s ○ ● ◍ ○ ○ f
3	Did you find it simple/difficult to speak and write at the same time?	s ○○●●○○ d
4	Did you find it simple/difficult to speak and sign at the same time?	s ○○◍●○○ d
5	How many lines of text would you accept to say and write in order to perform your own identification in a banking environment?	[0.5 – 10] **2.0**
6	Do you think that the act of speaking and writing at the same time affected your capacities to imitate the writing?	y ○●●○○○ n

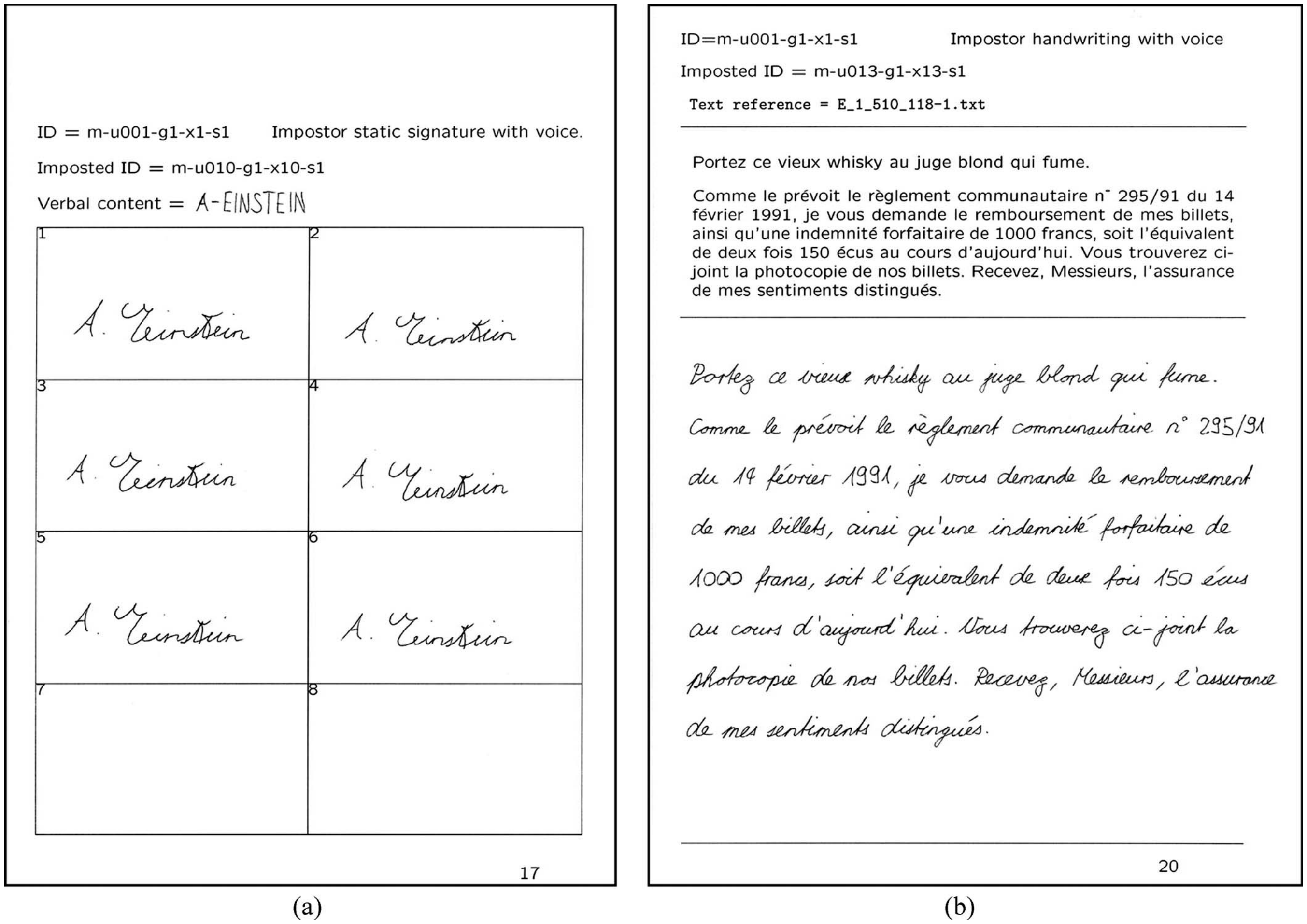
respective average answers are listed in Table I. The main conclusions of the survey are the following.

1) A large majority of users found it easy to write on a tablet. The use of regular pen and paper may have helped to reach this result.
2) Users ranked as average the difficulty of writing and speaking at the same time. This is most probably due to the extra level of concentration needed to perform such acquisitions.
3) Similarly and probably for the same reason, users ranked as average the difficulty of signing and speaking at the same time. There is potentially an extra difficulty associated to spoken signatures coming from the pre- or post-flourishes often available in signatures. Such flourishes do not represent any content and then cannot be said by the users, which is potentially disturbing.
4) Users feel they are writing at a slower speed when they are speaking in the same time. This feeling is potentially due to the fact that people need to slow down their speech production to catch up with the writing production.
5) Users would accept to write up to two lines of text to perform their authentication.
6) 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 answers of the survey, our current conclusion is that such bimodal acquisitions are acceptable from a usability point of view.

C. Evaluation Protocols

1) Spoken Signatures: As shown in Fig. 3(a), template papers are used for recording signatures. 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 have access to the static image and to the textual content of the signature to be forged. Due to time constraints, we established a strict procedure to let the forger train with a training time limited to few minutes per forgery. The forgers are therefore not highly skilled and such forgeries cannot be claimed to be corresponding to fully realistic scenarios (a real forger would probably train for hours). However, a forgery training time of 2 min was still practicable within our budget constraints and still provides much better estimate of the forgery rejection capacities of the system than random forgeries (see results in Section V). Moreover, this procedure corresponds to what is implemented in other standard databases [21], [22]. This procedure leads to a total of 18 *skilled forgeries* after the three sessions, i.e., six impostor signatures on three different 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

17

(a)

ID=m-u001-g1-x1-s1 Impostor handwriting with voice

Imposted ID = m-u013-g1-x13-s1

Text reference = E_1_510_118-1.txt

Portez ce vieux whisky au juge blond qui fume.

Comme le prévoit le règlement communautaire n° 295/91 du 14 février 1991, je vous demande le remboursement de mes billets, ainsi qu'une indemnité forfaitaire de 1000 francs, soit l'équivalent de deux fois 150 écus au cours d'aujourd'hui. Vous trouverez ci-joint la photocopie de nos billets. Recevez, Messieurs, l'assurance de mes sentiments distingués.

Portez ce vieux whisky au juge blond qui fume.
Comme le prévoit le règlement communautaire n° 295/91
du 14 février 1991, je vous demande le remboursement
de mes billets, ainsi qu'une indemnité forfaitaire de
1000 francs, soit l'équivalent de deux fois 150 écus
au cours d'aujourd'hui. Vous trouverez ci-joint la
photocopie de nos billets. Recevez, Messieurs, l'assurance
de mes sentiments distingués.

20

(b)

Fig. 3. Example of a (faked) (a) signature and (b) handwriting acquisition page.

For our tests, we are using a protocol *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.[1] This protocol corresponds to a realistic situation where users enroll one day and make access another day. 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. 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 × 3 sessions = 14490 random forgeries. Impostor tests are also performed using the available skilled forgeries, giving a total of 70 users × 18 accesses × 3 sessions = 3780 skilled forgeries. The numbers of tests previously mentioned are approximate, as some users did not complete all sessions.

2) Spoken Handwriting: A spoken handwriting assessment protocol has already been defined on MyIDea [20] and will be followed for the realization of the tests in this paper. In short, this protocol corresponds to a **text-prompted scenario** where we assume that the system prompts the subject to write and say a random piece of text each time an access is performed. This kind of scenario allows one to make the system more secure against spoofing attacks where the forger plays back a prerecorded version of the genuine data. This scenario has also the advantage to be very convenient for the subject who does not need to remember any password phrase. In MyIDea, for each of the three sessions, the subject is asked to read and write a random text fragment of about five lines for a total of 50 to 100 words. An example of the layout of the forms used for guiding the acquisitions is shown in Fig. 3(b). The data from the first session is used to train the system. Each genuine test uses the data available from session two and session three. Therefore, two genuine tests can be performed per user, giving a total of 70 users × 2 accesses = 140 genuine tests. After acquiring the genuine handwriting, the subject is also asked to imitate the handwriting of another subject and to synchronously utter the content of the text. In order to do this, the imitator has access to a static image of the handwriting to imitate. The access to the voice recording is not given for imitation as this would lead to a too difficult cognitive load, practically infeasible in the limited time frame of the acquisition. Skilled forgeries tests are performed using the three available imitations for a total of 70 users × 3 accesses = 210 skilled forgeries. We also consider random forgeries that are performed using one recording from the remaining subjects, giving 70 users × 69 accesses = 4830 random forgeries.

[1] Other types of protocols have been defined and evaluated such as a protocol *without time variability* where user models are built and tested using signatures from the same session [23], [24].

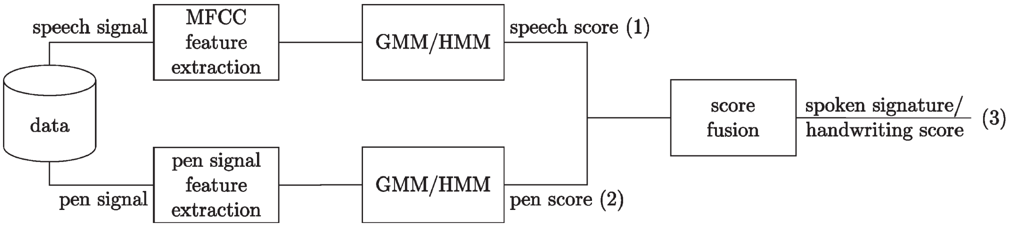


Fig. 4. Verification system.

IV. Systems Description

As shown in Fig. 4, we have opted for a simple system architecture where both streams of data are modeled independently using state-of-the-art feature extraction and modeling tools. The scores of each stream are then simply fused to obtain a global verification score. While this approach is clearly not leveraging on the potential synchronization patterns that may occur between each stream, it still brings an advantage in the case of impaired or temporarily injured users who cannot acquire one of the modalities. In such a case, the score-level fusion approach still allows one to produce a verification score by relying on the available modality.

A. Feature Extraction

1) Pen Signal: For each point of the pen signal, 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 [25] and [23]. Each feature vector includes:

1) the absolute speed and acceleration, the speed and acceleration in x and y directions and the tangential acceleration;
2) the angle α of the absolute speed vector, its cosine and sine, the derivative of α, and its cosine and sine;
3) the pressure and the pressure derivative;
4) the azimuth and elevation angles of the pen and their derivatives;
5) the curvature radius;
6) the normalized coordinates $(x(n) - x_g, y(n) - y_g)$ relative to the center of mass (x_g, y_g) of the pen input;
7) 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 features are further mean and standard deviation normalized on a per user basis. As all computed features are neither specific to signatures nor handwriting, we could apply the same feature extraction to both scenarios.

2) Speech Signal: 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 [26]. We realized that the speech signal contains many silence which is due to the fact that writing is usually slower than speaking. As silence parts usually impair the estimation of reliable models, we therefore implemented a procedure to remove all the silence parts from the speech signal. This silence removal component is using a classical energy-based speech detection module based on a bi-Gaussian model [27]. MFCCs 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 improvements of the results.

B. GMMs System

GMMs can be used to model the likelihoods of the features extracted from the pen and from the speech signal. GMMs have been reported to compare reasonably well to HMMs in terms of signature verification [28] and are often considered as baseline systems in handwriting verification [3] and speaker verification [11], [12]. Furthermore, GMMs are well-known flexible modeling tools able to approximate any probability density function.

In our case, GMMs apply very well for modeling spoken handwriting as our scenario is text independent, i.e., we cannot rely on *a priori* knowledge of the content of the signal. For modeling spoken signature, one could argue that GMMs are actually not the most appropriate models as they are intrinsically not capturing the time-dependent specificities of speech and signature. For this reason and as a competing approach, we are also investigating the use of HMMs to model spoken signature (see Section IV-C).

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

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

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 $\sum_{i=1}^{I} w_i = 1$ and the Gaussian densities $\mathcal{N}$ have the form

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

In our case, we use diagonal covariance matrices as an approximation of the full covariance matrices. This approximation is classically done when using GMMs for two reasons. First, it allows one to reduce the number 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 the CPU time needed in the computation involved in the matrix inversion and multiplication. 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

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

In a similar way, we also compute the likelihood score S_{world} of the hypothesis that X is **not** from the given client but from a broad nonclient category called *world*. This world likelihood is also estimated using a GMM model M_{world} trained by pooling the data of many other users. The decision whether to accept or to reject the claimed user is then 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

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

The training of the client and world models is usually performed with the Expectation–Maximization (EM) algorithm [29] 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 (MAP) criterion [30]. 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 impeach 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.[2] 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.

C. HMMs System

For spoken signatures, we have been investigating HMMs as an alternative to GMMs. First, HMMs have been extensively used to model the likelihoods of the features extracted from signatures [25], [31], handwriting [3], or speech [26]. Second, they are the natural extension of the GMMs, and they allow more detailed modeling of the data, incorporating sequential information of the strokes for handwriting and of the phonemes for speech. While HMMs are potentially richer than GMMs in terms of modeling capabilities, they have more parameters to tune such as the choice of the topology and the number of states.

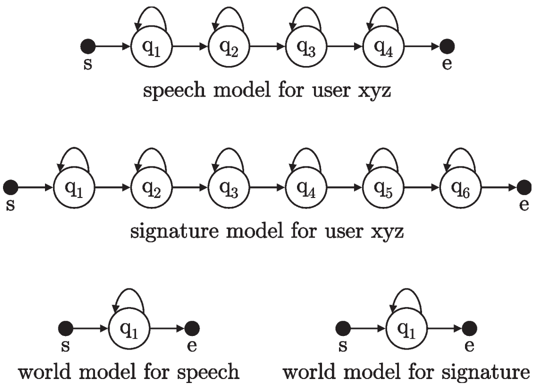

s
q1
q2
q3
q4
e
speech model for user xyz
s
q1
q2
q3
q4
q5
q6
e
signature model for user xyz
s
q1
e
world model for speech
s
q1
e
world model for signature

Fig. 5. HMM topology.

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 simplifying assumption of HMM-based modeling (see, for example, [26]), the likelihood of X given the model M_{client} can be written

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

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. Alternatively to (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. We have chosen to use continuous HMMs where the emission probability is modeled using a probability density function computed with weighted Gaussian mixtures as expressed in (1) and (2). We also use diagonal covariance matrices as approximation 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 shown in Fig. 5, which is actually nothing else than a GMM system. The decision whether to reject or to accept the claimed user is taken as above comparing the ratio of client and world score against a global threshold value T.

The training of each HMM is done in several iterations where two steps are performed. In the first step, a Viterbi forced alignment is computed [26] to find the most likely sequence of states given the parameters of the HMM. In the second

[2] 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.

step, the Gaussian densities of each state are re-estimated with the 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. Alternatively 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 shown 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 neighbor. 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 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.

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. 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. Moreover, 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. Our proposal is therefore the following. 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

$$K_{\mathrm{si}} = \frac{N_{\mathrm{a}}}{\alpha} \tag{6}$$

where α is a dividing factor that needs to be tuned. 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.

D. Score Fusion

We obtain the global score by applying a weighted sum of the signature (si) and speech (sp) log-likelihood ratios as expressed in

$$R_{\mathrm{client}} = W_{\mathrm{si}} R_{\mathrm{client,si}} + W_{\mathrm{sp}} R_{\mathrm{client,sp}}. \tag{7}$$

This approach is reasonable if we assume that the local observations of both subsystems 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 modeling could be more precise than the approach taken here and will be investigated in future work. More advanced score recombination could also be applied such as, for example, using classifier-based score fusion.

An optimization of the weight values W_{si} and W_{sp} is of course possible by tuning their values on a given development data set. However, such optimal values would actually be dependent to the context of use and more specifically to the frequency and the quality of the forgeries [24]. For example, in a given context, the signature modality could be easier to forge than the speech one and the optimal weights will have to give more importance to the speech modality. In another context, it could be the reverse situation. For this reason, we decided to use equal values of the weights for all results we are reporting below.

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.

V. Experimental Results

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

A. Spoken Signature

In a first set of experiments [24], we investigated what is the best model order for our different systems. The GMM system trained with EM or MAP has been evaluated using 8, 16, 32, 64, and 128 Gaussians in the client and world model. The best configuration is obtained with 16 Gaussians for the client and world model when using the EM algorithm. For the MAP adaptation algorithm, the optimum is to use 128 Gaussians in the world model and to adapt it toward the client model. Such a difference is explainable by the nature of the algorithm. With EM, a model is built from scratch using only the limited amount of data available from the genuine user. With MAP adaptation, the world model is used as a starting point for the training and only the relevant parameters of the model are modified by the adaptation procedure. As the world model can be trained using much more data pooled from a large set of users, it is then normal that a larger number of Gaussians can be reached. A similar set of experiments has been performed to determine the best number of Gaussians for the HMM system, using also the EM or MAP training procedure and using a fixed number of states for all users. The best configuration for the EM algorithm is lying between 8 and 16 Gaussians per state for the client and world model, with no significant difference between both values. The best configuration for the MAP algorithm is obtained using 16 Gaussian in the client and world model. Table II summarizes the best configurations.

The second set of experiments aimed at finding the optimal value of the α parameter of (6). We used the MAP adaptation algorithm with 16 Gaussians in each state and 16 Gaussians in

TABLE II
OPTIMAL NUMBER OF GAUSSIANS FOR CLIENT/WORLD MODEL FOR THE DIFFERENT SYSTEMS AND ALGORITHMS USED FOR SPOKEN SIGNATURES

	GMM	HMM
EM	16/16	8/8-16/16
MAP	128/128	16/16

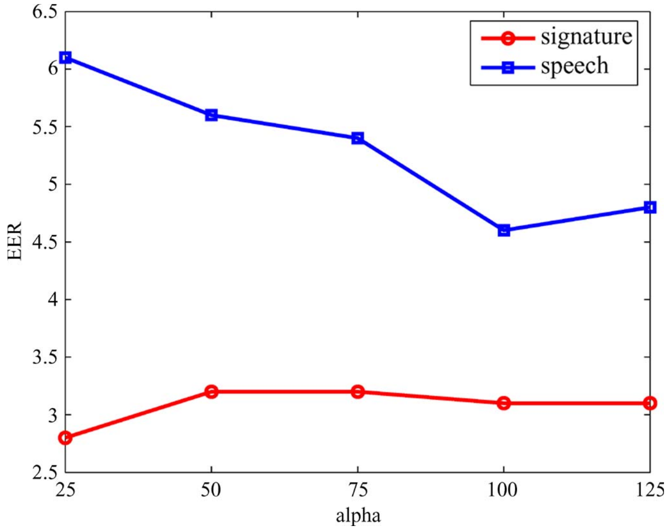


Fig. 6. EERs (random forgeries, MAP) as a function of the dividing factor α, 16 Gaussian mixtures in each state of the HMM and 16 Gaussian mixtures in the world model.

the world model. As it can be observed from Fig. 6, 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, 16 Gaussian per states. All results were in favor of using a variable number of states as described above.

In a third set of experiments, our objective was to perform an extensive comparison of GMMs with HMMs for different training algorithms (EM versus MAP) and for different strengths of forgeries (random versus 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 state and with a variable number of states for each client such as described above. According to Fig. 6, the values of α were, respectively, set to 25 and 100 for the signature and speech parts using MAP. For the EM algorithm, the best α values were slightly different, both equal to 100 for the signature and speech part.

Table III summarizes the results in terms of ERR and the following conclusions can be drawn.

1) **Comparison GMM—HMM.** When considering the fusion of both modalities, the HMM modeling is leading consistently to better accuracy than the GMM modeling. When the signature and speech modalities are considered separately, the HMM modeling is, in most of the cases, leading to better results than the GMM modeling. Only two of the configurations show a slight advantage for GMMs but probably, the difference is not significant in these cases.
2) **Comparison EM—MAP.** As it was already reported in many previous works (including this paper) [24], [30], [32], [33], 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.[3] 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.

An alternative representation of system performance can be given using Detection Error Tradeoff (DET) curves [34]. Such curves are plotted computing the operating points of false acceptation and false rejection rates for different values of T. With DET curves, normal deviate scales are used for the x and y axis. DET curves have then the property to be close to a straight line if the scores are normally distributed, which is generally the case for likelihood-based biometric systems. Such DET curves are shown in Fig. 7 for our best configuration of the spoken signature system evaluated using random forgeries. The DET curves are showing the performance for the individual speech and signature modalities, and the gain that can be obtained by combining both. We can observe that for

[3] 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.

TABLE III
SPOKEN SIGNATURES RESULTS IN TERMS OF EERS WITH GMM AND HMM MODELING FOR EM AND MAP ADAPTATION

modelling	GMM				HMM			
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 %**

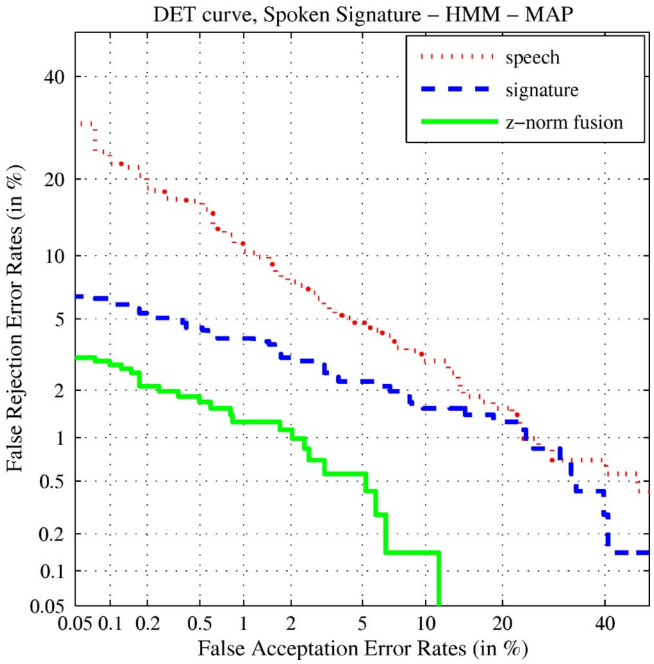


Fig. 7. DET curve—fusion of the signature and speech HMM systems, random forgeries.

all operating values of T, 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.

B. Spoken Handwriting

As spoken handwritings are produced in a text-independent way, only the GMM system is here investigated. In a similar way as for spoken signatures, we first performed a set of tests to find out the best model order. We measured that the optimal model size seems to lie around 256 Gaussians for both client and world models. In comparison to spoken signatures, the optimal model size is larger which is actually reasonable as the quantity of data is here much larger.

Table IV shows a comparison of results obtained using the EM versus the MAP algorithm and random forgeries for testing. The fusion is, in this case, the simple summation fusion, without any z-norm. Similar conclusions as for spoken signatures can be drawn from these results.

1) The fusion of speech and handwriting significantly improves the performance of the biometric system. This gain of performance can be obtained at no extra cost for the user as both streams of data are recorded simultaneously.

TABLE IV
COMPARISON EM/MAP ALGORITHMS ON RANDOM FORGERIES

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

TABLE V
SPOKEN HANDWRITING RESULTS IN TERMS OF TERMS OF EERS. COMPARISON OF RANDOM VERSUS SKILLED FORGERIES

forgeries	random	skilled
handwriting	4.0 %	13.7 %
speech	1.8 %	6.9 %
sum fusion	**0.7 %**	**6.9 %**
z-norm fusion	**0.3 %**	**4.0 %**

2) The sum fusion that is applied here is extremely simple and requires actually no further estimation of parameters. This result can be explained considering that the models used for speech and handwriting are very much similar in architecture and order.
3) The MAP adaptation algorithm is leading to better results than the EM algorithm. The reasoning is probably similar as for spoken signature, i.e., it is better to adapt from a well-trained world model than to build from scratch a GMM using few data.

Table V compares random and skilled forgeries performances using our best GMM system trained with MAP adaptation. The conclusions are again similar as for spoken signatures. Considering the handwriting part, 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 does 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 V, showing an advantage against the sum fusion.

As a conclusion of these experiments with spoken handwriting, we can reasonably say that the speech modelization performs on average better than the handwriting. Intuitively, one could advance that this is understandable as the handwriting is a gesture more or less fully learned (behavioral biometric), while speech contains information that are dependent on learned and physiological features (behavioral and physiological biometric).

C. 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 V can be compared with results of spoken signatures in Table III. 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 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 Tables III and V 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 s for signature, spoken handwriting provides about two min of speech. The speech model is therefore more precise for spoken handwriting than for spoken signature.

Now, considering both modalities simultaneously, if we compare the z-norm fusion results of Tables III and V, 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 used at training and testing time. If we would have less handwriting data for training or testing, the conclusions could also be reversed.

VI. Conclusion and Future Work

We presented a novel user authentication system based on a combined acquisition of online pen and speech signals. We introduced two potential scenarios of use basing the approach on the use of signatures or handwriting. So-called spoken signatures are recorded asking the user to sign and say the content of the signature. Spoken handwriting signals are recorded prompting the user to write and read the content of sentences randomly extracted from a text.

A simple system architecture has been introduced where both streams of data are modeled independently using state-of-the-art feature extraction and statistical modeling tools. The scores of each stream are then simply fused using a summation procedure to obtain a global verification score. The modeling tools are based on GMMs or HMMs for which different configurations and training algorithms have been evaluated on a realistic multisession database. More specifically, it has been shown that the modeling of the spoken signatures can be performed advantageously using HMMs where the number of states is optimized on a per user basis for both modalities. For spoken handwriting, we have shown that a simple GMM system can be used. Consistently for the GMM and HMM architectures and for both scenarios, we observed that a MAP adaptation procedure leads to better results than the classically used EM training algorithm.

As a general observation for both scenarios, we can conclude that 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 either modality 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 bimodal speech and handwriting approach seems then to be a viable alternative to systems using single modalities.

The current best performance of our verification system measured on random forgeries is 1.1% EER for spoken signatures and 0.3% for spoken handwriting. While the best overall performance is obtained using spoken handwriting, the spoken signature approach presents the advantage of using much less data allowing a shorter authentication time.

Future works could go in the direction of using more robust modeling techniques against forgeries. We have identified potential directions such as time-dependent score fusion or joint modeling of both data streams.

Acknowledgment

The authors would like to thank A. El Hannani for her precious help with GMMs-based systems.

References

[1] R. Plamondon and G. Lorette, "Automatic signature verification and writer identification—The state of the art," *Pattern Recognit.*, vol. 22, no. 2, pp. 107–131, 1989.

[2] F. Leclerc and R. Plamondon, "Automatic signature verification: The state of the art-1989–1993," *Int. J. Pattern Recogn. Artif. Intell.*, vol. 8, no. 3, pp. 643–660, 1994.

[3] M. Liwicki, A. Schlapbach, H. Bunke, S. Bengio, J. Mariéthoz, and J. Richiardi, "Writer identification for smart meeting room systems," in *Proc. 7th Int. Workshop Document Anal. Syst.*. New York: Springer-Verlag, 2006, vol. 3872, pp. 186–195.

[4] A. Wahl, J. Hennebert, A. Humm, and R. Ingold, "Generation and evaluation of brute-force signature forgeries," in *Proc. Int. Workshop MRCS*, Istanbul, Turkey, Sep. 2006, pp. 2–9.

[5] C. Vielhauer, *Biometric User Authentication for IT Security*. New York: Springer-Verlag, 2006.

[6] J. Hennebert, R. Loeffel, A. Humm, and R. Ingold, "A new forgery scenario based on regaining dynamics of signature," in *Proc. 2nd ICB*, Seoul, Korea, Aug. 27–29, 2007, pp. 366–375.

[7] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, *Audiovisual Automatic Speech Recognition: An Overview*. Cambridge, MA: MIT Press, 2004.

[8] H. Bredin and G. Chollet, "Audio-visual speech synchrony measure for talking-face identity verification," in *Proc. IEEE ICASSP*, Honolulu, HI, 2007, pp. II-233–II-236.

[9] R. Landais, H. Bredin, L. Zouari, and G. Chollet, "Vérification audiovisuelle de l'identité," in *Proc. Traitement et Analyse de l'Information: Méthodes et Applications (TAIMA)*, Hammamet, Tunisia, 2007.

[10] L. Hong and A. Jain, "Integrating faces and fingerprints for personal identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 12, pp. 1295–1307, Dec. 1998.

[11] D. Reynolds, "An overview of automatic speaker recognition technology," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002, vol. 4, pp. 4072–4075.

[12] F. Bimbot *et al.*, "A tutorial on text-independent speaker verification," *EURASIP J. Appl. Signal Process.*, vol. 2004, no. 4, pp. 430–451, 2004.

[13] Y. Nakamura and M. Kidode, "Online writer verification using kanji handwriting," in *Proc. Int. Workshop MRCS*, B. Gnsel, A. K. Jain, A. M. Tekalp, and B. Sankur, Eds. Istanbul, Turkey: Springer-Verlag, 2006, vol. 4105, pp. 207–214.

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

[15] S. Krawczyk and A. K. Jain, "Securing electronic medical records using biometric authentication," in *Proc. AVBPA*, Rye Brook, NY, 2005, pp. 1110–1119.

[16] J. Koreman, A. C. Morris, D. Wu, S. Jassim, H. Sellahewa, J. Ehlers, G. Chollet, G. Aversano, H. Bredin, S. Garcia-Salicetti, L. Allano,

B. Ly Van, and B. Dorizzi, “Multi-modal biometric authentication on the SecurePhone PDA,” in *Proc. 2nd Int. Workshop Multimodal User Authentication*, Toulouse, France, 2006.

[17] Y. Watanabe, K. Iwata, R. Nakagawa, K. Shinoda, and S. Furui, “Semi-synchronous speech and pen input,” in *Proc. 32nd ICASSP*, Honolulu, HI, 2007, pp. IV-409–IV-412.

[18] B. Dumas *et al.*, “Myidea—Multimodal biometrics database, description of acquisition protocols,” in *Proc. 3rd COST 275 Workshop*, Hatfield, U.K., Oct. 27–28, 2005, pp. 59–62.

[19] J. Hennebert *et al.*, *Myidea Multimodal Database*, 2005. [Online]. Available: http://diuf.unifr.ch/go/myidea

[20] A. Humm, J. Hennebert, and R. Ingold, “Combined handwriting and speech modalities for user authentication,” Dept. Informatics, Univ. Fribourg, Fribourg, Switzerland, Tech. Rep. 06–05, 2006.

[21] 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,” *Proc. Inst. Elect. Eng.—Vis. Image Signal Process.*, vol. 150, no. 6, pp. 395–401, Dec. 2003.

[22] S. Garcia-Salicetti, C. Beumier, G. Chollet, B. Dorizzi, J. Leroux 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 Int. Conf. AVBPA*, 2003, pp. 845–853.

[23] A. Humm, J. Hennebert, and R. Ingold, “Gaussian mixture models for chasm signature verification,” in *Proc. 3rd Joint Workshop Multimodal Interaction Related Mach. Learn. Algorithms*, Washington, DC, 2006, pp. 102–113.

[24] J. Hennebert, A. Humm, and R. Ingold, “Modelling spoken signatures with Gaussian mixture model adaptation,” in *Proc. 32nd ICASSP*, Honolulu, HI, 2007, pp. II-229–II-232.

[25] B. Ly Van, S. Garcia-Salicetti, and B. Dorizzi, “Fusion of HMM’s likelihood and Viterbi path for on-line signature verification,” in *Proc. Biometrics Authentication Workshop*, Prague, Czech Republic, May 15, 2004.

[26] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[27] J. Mariethoz and S. Bengio, “A comparative study of adaptation methods for speaker verification,” in *Proc. Int. Conf. Spoken Language Process.*, Denver, CO, 2002, pp. 581–584.

[28] J. Richiardi and A. Drygajlo, “Gaussian mixture models for on-line signature verification,” in *Proc. ACM SIGMM Workshop Biometrics Methods Appl.*, 2003, pp. 115–122.

[29] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *J. R. Stat. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.

[30] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digit. Signal Process.*, vol. 10, no. 1, pp. 19–41, Jan. 2000.

[31] J. G. A. Dolfing, “Handwriting recognition and verification, a hidden Markov approach,” Ph.D. dissertation, Philips Electron. N.V., Eindhoven, The Netherlands, 1998.

[32] A. Humm, R. Ingold, and J. Hennebert, “Spoken handwriting verification using statistical models,” in *Proc. 9th ICDAR*, Curitiba, Brazil, Sep. 23–26, 2007, pp. 999–1003.

[33] A. Humm, J. Hennebert, and R. Ingold, “Hidden Markov models for spoken signature verification,” in *Proc. IEEE Conf. BTAS*, Washington, DC, Sep. 27–29, 2007, pp. 1–6.

[34] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, “The DET curve in assessment of detection task performance,” in *Proc. Eurospeech*, Rhodes, Greece, 1997, pp. 1895–1898.

Andreas Humm (S’08) received the Master of Science degree “Master of Science in Computer Science” from the University of Bern, Bern, Switzerland, in 2005. He is currently working toward the Ph.D. degree in the multimedia engineering DIVA group of the Computer Science Department, University of Fribourg, 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 (M’06) received the Electrical Engineering degree from the Faculté Polytechnique de Mons, Hainaut, Belgium, in 1993 and the Ph.D. degree in computer science from the Swiss Federal Institute of Technology, Lausanne, Switzerland, in 1998.

He then worked for six years as IT System Architect and Independent Consultant for different private companies. In 2004, he was with the multimedia engineering DIVA group of the Computer Science Department, University of Fribourg, Fribourg, Switzerland, where he is in charge of teaching and research activities. Since 2007, he has been a Professor with the institute of Business Information Systems, University of Applied Sciences Western Switzerland, HES-SO // Wallis, Sierre, Switzerland, and is affiliated as Lecturer with the University of Fribourg. 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 (M’93) received the Diploma degree of “mathematical engineer” and the Ph.D. degree in sciences from the Swiss Federal Institute of Technology, Lausanne, Switzerland, in 1983 and 1989, respectively.

Since 1989, he has been a Professor in the multimedia engineering DIVA group of the Computer Science Department, University of Fribourg, 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.