Spoken Handwriting for User Authentication using Joint Modelling Systems

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

Rolf Ingold
Université de Fribourg
Boulevard de Pérolles 90
1700 Fribourg, Switzerland
rolf.ingold@unifr.ch

Jean Hennebert*
University of Applied Sciences
Western Switzerland, HES-SO // Wallis
3960 Sierre, Switzerland
jean.hennebert@hevs.ch

Abstract

We report on results obtained with a new user authentication system based on a combined acquisition of online pen and speech signals. In our approach, the two modalities are recorded by simply asking the user to say what she or he is simultaneously writing. The main benefit of this methodology lies in the simultaneous acquisition of two sources of biometric information with a better accuracy at no extra cost in terms of time or inconvenience. Another benefit comes from an increased difficulty for forgers willing to perform imitation attacks as two signals need to be reproduced. Our first strategy was to model independently both streams of data and to perform a fusion at the score level using state-of-the-art modelling tools and training algorithms. We report here on a second strategy, complementing the first one and aiming at modelling both streams of data jointly. This approach uses a recognition system to compute the forced alignment of Hidden Markov Models (HMMs). The system then tries to determine synchronization patterns using these two alignments of handwriting and speech and computes a new score according to these patterns. In this paper, we present these authentication systems with the focus on the joint modelling. The evaluation is performed on MyIDea, a realistic multimodal biometric database. Results show that a combination of the different modelling strategies (independent and joint) can improve the system performance on spoken handwriting data.

1. Introduction

Speech and handwritings are two major modalities used by humans in their daily transactions and interactions. In addition, both modalities are well accepted and very natural and non-intrusive. Many automated biometric systems based on speech alone have been studied and developed in the past. See, for example, [16] for a review of such systems. Biometric systems based on online handwriting were not so numerous, however, we can mention [12] or [15] as examples of state-of-the-art systems.

However, we still see few deployments of such authentication systems in commercial applications. Four reasons can be given to explain this: (1) handwriting production is behavioral, therefore variable by nature. A user does not write two times in the exact same way, especially when time is spent between two acquisitions; (2) uniqueness is weak, the handwriting signals are specifically produced to be understandable by all; (3) a handwriting system can be easily attacked by intentional forgers trained to reproduce the handwriting of a genuine user [19][18][4]; (4) the performances are dependent on the acquisition context and on the sensor where mismatched conditions usually decrease performances [18].

Several attempts have already been reported to overcome these difficulties by combining the pen and speech signals. For example, in [11], a tablet PC system is using a combination of online signature and speech to ensure the security of electronic medical records. In [13], a similar system using signature and speech is also proposed to reach better authentication performances. The main difference between these works and our approach lies in the acquisition procedure that is, in their case serial and in our case simultaneous. 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. It is also worth mentioning the work presented in [20], where a similar approach is used, not for biometric aspects but to enhance the recognition of spoken content for noisy mobile environment. In this approach, the user simultaneously writes the first characters of a spoken utterance. The recognition of the first characters is injected

*Jean Hennebert is also affiliated with the University of Fribourg.

6th International Symposium on Image and Signal Processing and Analysis (ISPA 09), Salzburg, Austria, September 16 - 18. 2009. pp. 505-510

in the HMM decoding of the speech part and allows to enhance the speech detection while eliminating less probable hypothesis.

Our proposal is indeed to record the speech and handwriting signals where the user reads what she or he is writing. Such acquisitions are referred here and in our related works as CHASM handwritings for **c**ombined **h**andwriting **a**nd **s**peech **m**odalities handwritings[1], or more simply referred as, **spoken handwriting**. Our motivation to perform a synchronized acquisition is multiple. Firstly, it avoids doubling the acquisition time. Secondly, the synchronized acquisition will probably give better robustness against intentional imposture. Indeed, imitating simultaneously the voice and the writing of somebody is more difficult than imitating each modality taken separately. This double task is also studied in the psychological research. For such a particular issue of resource sharing for simultaneous speech and fine motor movement of the right hand [10]. Finally, the synchronization patterns (i.e. where do users synchronize) or the intrinsic deformation of the inputs (mainly the slow-down of the speech signal) may be dependent on the user, therefore bringing an extra piece of useful biometrics information. The work reported here is specifically focusing on this last statement.

Our previous works on spoken handwriting have been dedicated to data acquisition [2], survey and definition of realistic scenario [7] and experiments with spoken handwriting using a independent modelling of both data streams [9][8]. In this paper, we report on the development of a novel approach to perform joint modelling and on its evaluation using a realistic database.

The remainder of this paper is organized as follows. In section 2, we give an overview of MyIDea, the database used for this work and of the evaluation protocols. In section 3, we describe our systems using spoken handwriting data. Section 4 presents the experimental results. Finally, conclusions, discussions and future work are provided.

2. Spoken Handwriting Database

2.1. MyIDea Database

Spoken handwriting data was acquired in the framework of the MyIDea biometric data collection [2][3]. MyIDea is a multimodal database that also contains other modalities such as fingerprints, talking faces, etc. About 70 users were recorded over three sessions spaced in time. The data set used to perform the experiments reported in this article has been given the reference MYIDEA-CHASM-SET1 by the distributors of MyIDea. This set should be considered as a development set. A second set of data containing 66 users was also recorded and will be used as evaluation set.

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

In MyIDea, spoken handwriting was acquired with a WACOM Intuos2 graphical tablet and a standard computer headset microphone (Creative HS-300). For the tablet stream, x,y-coordinates, pressure, azimuth and elevation angles of the pen are sampled at 100 Hz. The speech waveform is recorded at 16 kHz and coded linearly on 16 bits. The data samples are also provided with timestamps to allow a precise synchronization of both streams. The timestamps are especially important for the handwriting streams as the graphical tablet does not send data samples when the pen is out of range.

Fig. 1 shows an example of spoken handwriting. All signals are synchronized thanks to the timestamps. The upper part of the figure shows the time evolution of the tablet data streams including the x and y coordinates, and pressure p. The azimuth and elevation angles are not represented for sake of clarity. The bottom part of Fig. 1 presents the speech signal evolution. The grey area on the figure corresponds to empty stretches of the pen signal occurring when the user lifts the pen out of the range of the tablet.

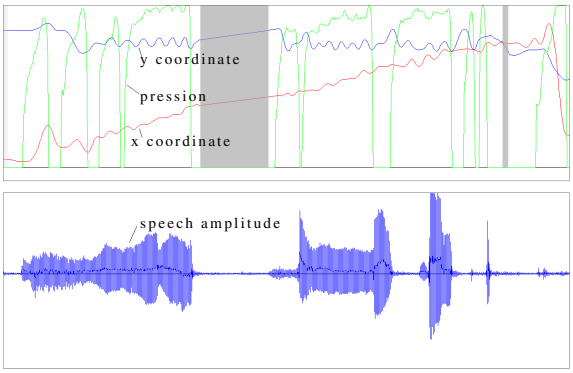

Figure 1. Synchronized visualization of handwriting (upper part including x,y-coordinates, pressure, not including angles for sake of clarity) and speech signals (bottom part).

As a general observation on the acquisition, we have noticed that the voice is slightly de-synchronized with the writing. Re-synchronization is happening and corresponds roughly to syllables or end of words.

In [5], we report on a usability survey conducted on the subjects of MyIDea. The main conclusions of the survey are the following. First, all recorded users were able to perform the CHASM acquisition. Speaking and writing at the same time did not prevent any acquisition to happen. Second, the survey shows that such acquisitions are acceptable from a usability point of view.

2.2. Recording and Evaluation Protocols

A spoken handwriting assessment protocol has been defined on MyIDea [5] 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 to make the system more secure against spoofing attacks where the forger plays back a pre-recorded version of the genuine data. This scenario has also the advantage to be very convenient for the subject who does not need to remember any password phrase.

In MyIDea, for each of the three sessions, the subject is asked to read and write a random text fragment of about 5 lines for a total of 50 to 100 words. The subject is allowed to train for a few lines on a separated sheet in order to be accustomed to with the procedure of talking and writing at the same time. The data from the first session is used to train the system. Each genuine test uses the data available from session two and three. Therefore, 2 genuine tests can be performed per user, giving a total of 70 users $\times$ 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 3 available imitations for a total of 70 users $\times$ 3 accesses = 210 skilled forgeries. We also consider random (zero-efforts) forgeries that are performed using one recording from the remaining subjects, giving 70 users $\times$ 69 accesses = 4830 random forgeries.

3. System Description

3.1. Feature Extraction

For each point of the handwriting, we extract 25 dynamic features based on the x and y coordinates, the pressure and angles of the pen in a similar way as in [14] and [6]. This feature extraction was actually proposed to model signatures, however it can be used without modification in our case as the signals are similar (coming from a graphical tablet) and as nothing specific to signature is included in the computation of the features. Here, an additional segmentation of the handwritten text into individual text lines is performed before extracting the features. The features are mean and standard deviation normalized on a per user basis.

For the speech signal, we compute 12 Mel Frequency Cepstral Coefficients (MFCCs) and the energy every 10 ms on a window of 25.6 ms. MFCC coefficients are mean and standard deviation normalized.

3.2. Independent Modelling and Score Fusion System

In our first strategy, we have opted for a simple system architecture where both streams of data are modelled independently [8]. In short, the system uses state-of-the-art modelling tools, like Gaussian mixture models (GMMs) and training algorithms, like Expectation-Maximization (EM) or Maximum A Posteriori criterion (MAP) for handwriting and speech [16]. The scores of each stream are then simply fused, for example with a sum fusion, to obtain a global verification score as illustrated in Fig. 2.

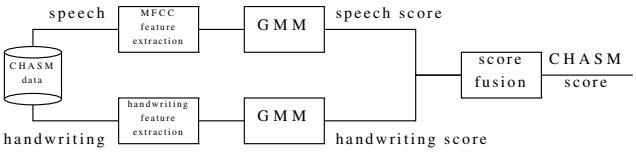

Figure 2. Spoken handwriting verification system (independent modelling).

The underlying assumption is that the voice and pen signals are produced according to two independent systems. This is of course not the case in reality. However, if we look at the granular pieces of biometric information brought by both streams, the assumption of independence is maybe not so inappropriate. Indeed, MFCC features model information linked to the vocal tract configuration and the feature extraction applied to the pen signal is also based on local features, dependent to the position and dynamics of the hand. At such an "atomic" level, there is probably no physiological evidence that such low-level features are dependent.

3.3. Joint Modelling System

Generally speaking, multimodal approaches try to combine somehow the information of two or more data streams. Evidence in a multimodal system can be integrated with a fusion at different levels [17]. In our second strategy, we would like to investigate systems that are based on higher levels of modelling.

We have actually developed and evaluated different approaches that were not successful. These approaches were based on attempting to model synchronizations at a rather granular level. A possible explanation of these attempts is to be found in the fact that they did not allow any desynchronization of the two streams while actually the desynchronization of the two streams goes up to one second, i.e. 100 feature vectors. Another factor that we did not take into account in our preliminary experiments is linked to the high variability of such synchronization patterns. For example, if a user starts to speak after writing a word, with a delay of, let's say 0.2 seconds, the observation for the exact same content could be 0.4 seconds in the next acquisition. It appears from our attempts that allowing such de-synchronization mechanism is necessary as the synchronization of speech and handwriting is variable[2].

[2]In a similar way as for spoken handwriting, we also developed systems using spoken signature. All these systems were also evaluated and we reached to the same conclusions that we should allow such desynchronization in the modelling

Knowing that the synchronization patterns are variable by nature, we have designed a modelling method that has proven to be successful. The idea is to measure general statistics about large categories of synchronizations. It tries to find the synchronization patterns occurring between each stream. This system is exploiting the outputs of a recognition system. More specifically, we use Viterbi forced alignments of HMMs to measure statistics about respective start and end of tokens (word or phoneme levels) in both streams.

Fig. 3 shows such an example with the French sentence "`Portez ce vieux whisky au juge blond qui fume`". The upper part of the graph shows the speech amplitude as a function of time with segmentation at the word level. The lower part presents the handwritten words. This is a symbolic representation showing the start and end of the written words as a function of time. Looking at one word, for example the word "whisky", we see that the user begins first to speak and then begins to write the word, i.e., the speech data corresponding to the word "whisky" begins temporally before the handwriting stream begins. For the end part, the user speaks the end of the word "whisky" before terminating to write the word. We are now interested to measure this time delta for all words and to build a histogram for every user. In the following, the system is described in more technical details.

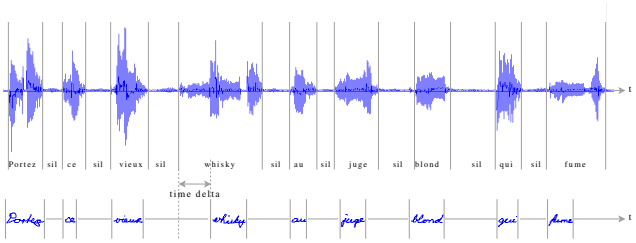

Figure 3. Forced alignment of spoken handwriting.

In the training phase, we have to train letter and phoneme models with the available data. In the case of handwriting, we have built for all upper and lower case letters of the alphabet an individual model. The models are represented by a single state continuous HMM modelling the emission probability using a mixture of Gaussians. Similarly, in the speech part, we train the models corresponding to every phoneme. The mapping of a word to phonemes can be found in dictionaries like BREF for French and CMUdict for English words. The example sentence "`Portez ce vieux whisky au juge blond qui fume`" is represented in the two dictionaries as illustrated in table 1. For the training, we use an embedded HMM training: we iterate over all users and take all the training data to train the different models. As we follow a text-prompted scenario, we know the content (transcription) of the written and spoken text and we can build the HMM topology. In more details, for every written line, we take the

Table 1. Handwriting dictionary on the left side, speech dictionary on the right side.

word	characters	word	phonemes
Portez	P o r t e z	Portez	pp oo rr tt ei
ce	c e	ce	ss ee
vieux	v i e u x	vieux	vv yy eu
whisky	w h i s k y	whisky	ww ii ss kk ii
au	a u	au	au
juge	j u g e	juge	jj uu jj
blond	b l o n d	blond	bb ll on
qui	q u i	qui	kk ii
fume	f u m e	fume	ff uu mm

words and split it according to the dictionary into individual characters. The same thing for speech, we split the whole spoken text into phonemes. As initialization, we can either use a linear segmentation of the data or we can use a segmentation file describing the words or silences with labels. The different letter and phoneme models are then trained independently with a simple EM training procedure. This procedure is iterated until a given stopping criterion is fulfilled. The forced alignment of the training data is then recovered with the Viterbi decoding. Once we have the forced alignment of the handwriting part and the speech part, we apply a procedure to compute a histogram of start time deltas between handwriting and speech on the word level. We simply take the handwriting forced alignment on word level as reference and search in a given time interval around the start timestamp in the speech alignment for the corresponding word. If available, we compute the time delta between the two start timestamps and build a histogram with these values. A negative value means that the user began to speak the word first and then he began to write. A positive value means the opposite. The number of bins in the histogram can be set according to the size of the bin.

In the testing phase, we do a similar procedure as in the training phase. We build an HMM according to the transcription (again, as we follow to a text-prompted scenario, we know the content of the written or spoken text) to get the forced alignment. We then compute the start time deltas between handwriting and speech on the word level as in the training phase and compute a score according to the values of the histogram by locking up the corresponding value.

4. Experimental Results

As visually observed in the data acquisition, not all users synchronize the speech and the handwriting at the same level (e.g. at syllable level). We classified manually all spoken handwriting data according to defined classes to get an idea of the distribution. Fig. 4 summarizes the evaluation of the classification into synchronization levels. We observed that more than 50% of the users did a global synchronization (upper part in Fig. 4) of speech and handwriting at syllable level while only 16% did it at the word level. About a quarter did a mix over the three sessions. These users

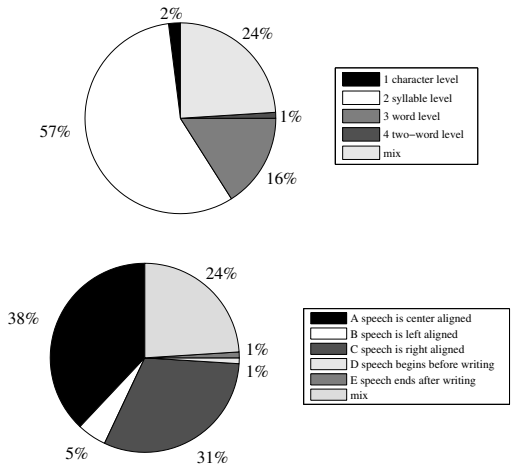

Figure 4. Distribution of global and local synchronization categories in the upper part, respectively in the bottom part.

did not always synchronize at the same global level. On the detailed local synchronization (bottom part in Fig. 4), the speech of about 40% of the users is centered with the handwriting, meaning that the user began to write first then started speaking and terminated speaking before the end of the handwriting. Nearly one third did a right alignment of the speech and one quarter did again a mix of different local synchronizations. A possible interpretation can be found in the neuropsychological. Users could potentially be divided into those that see writing as a primary task and speaking as a secondary task, and those that have reverse priority.

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

First, we performed a set of experiments to tune the alignment system described in subsection 3.3. We found out that the best settings for the time delta and the number of bins in the histogram is 4000 ms, respectively 21 bins. Therefore, we will report all results obtained in this paper using these settings.

Second, we performed a set of experiments to compare different computation methods for the time delta. In details, we computed the time delta firstly, with the start timestamps of the two words and secondly, with the end timestamps of the words. The latter computation is potentially less characteristic for users but still worth to investigate. Table 2 shows a comparison of results using the start versus end timestamps. The fusion used in our approach is a simple summation fusion. The first two lines "handwriting" and "speech" represents the EER obtained using respectively the handwriting or speech signals alone. The result "sum fusion (hw + sp)" represents the simple fusion of the scores of handwriting and speech signals modelled independently. Several interesting conclusions can be drawn from these results.

Table 2. Comparison of start and end timestamps for the time delta computation, co-occurrence at the word level, independent and joint modelling, EM algorithm.

forgeries	random	skilled
handwriting	6.8	6.9
speech	7.5	9.7
co-occurrence (start)	45.3	17.1
co-occurrence (end)	47.0	19.4
sum fusion (hw + sp)	**2.3**	**4.6**
sum fusion (hw + sp + start)	**1.7**	**4.6**
sum fusion (hw + sp + end)	**1.9**	**4.6**

1. As expected, the EER for the co-occurrence, taking the start timestamp, leads to better results than taking the end timestamp of the written and spoken words as synchronization boundaries are by nature at the begin of a word and not at the end. This conclusion can also be drawn for the sum fusion.

2. The joint modelling performs better for skilled than for random forgeries (see lines 3 and 4). This can be explained by the fact that random forgeries are other genuine users who did not perform an imitation. They synchronize the speech and handwriting in a similar way as genuine users. Skilled forgeries are more stressed by the forgery procedure, and therefore disrupt synchronization patterns and are therefore more easily rejected.

3. The fusion of speech and handwriting can really improve the performance of the biometric system ("sum fusion hw + sp"). This gain of performance can be obtained at no extra cost for the user as both streams of data are recorded simultaneously. This result is achieved by the independent modelling system.

4. The fusion of the independent and joint modelling can improve the system in terms of performance for random forgeries (see last three lines). Regarding skilled forgeries, there is no improvement. We can conclude that co-occurrences at the word level regarding to the start and end timestamps do include biometric information that can be modelled using our approach.

We can also mention a second set of experiments, where we attempted to move down the scope of analysis from word to phonemes. The difference of the phoneme level results compared to the word level results in terms of performance is not significant. Similar conclusions as for table 2 can also be drawn.

We can conclude that the alignment system alone was not always conclusive in terms of gain of performances. A reason for this can be the variability of these joint biometric patterns that is potentially large and, as we have relatively

few data per user, statistics of such patterns are difficult to estimate reliably. However, results of the alignment system have shown that a combination of the two modelling techniques (independent and joint) could improve the system performance.

5. Conclusions and Future Work

A new user authentication system based on combined acquisition of online pen and speech signals using joint modelling has been presented and evaluated. The presented approach uses a recognition system to compute the forced alignment of HMMs. It has been shown that the performance of the joint modelling approach alone did not provide good verification rates. However, results have shown that a combination of the two modelling techniques (independent and joint) could improve the system performance. We can reasonably conclude that there are synchronization patterns between pen and speech data streams that can be exploited to perform biometric verification.

In our future work, we plan to investigate other handwriting features which would potentially bring lower error rates [12] or the use of Asynchronous Hidden Markov Models (A-HMMs) that would apply especially well to model data that are consisting of two slightly de-synchronized streams [1]. Another perspective is to use spoken handwriting data not for personal authentication purposes but rather for text recognition. Some researches have already begun to go in such directions [20].

6. Acknowledgments

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] S. Bengio, "An asynchronous hidden markov model for audio-visual speech recognition", In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems, NIPS 15*, Vancouver, Canada, 2003. MIT Press.

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

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

[4] J. Hennebert, R. Loeffel, A. Humm, and R. Ingold, "A new forgery scenario based on regaining dynamics of signature", In *Proceedings of 2nd International Conference on Biometrics (ICB'07)*, pp. 366–375, Seoul (Korea), August 27 - 29 2007.

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

[6] A. Humm, J. Hennebert, and R. Ingold, "Gaussian mixture models for chasm signature verification", In *3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, Washington, 2006.

[7] A. Humm, J. Hennebert, and R. Ingold, "Scenario and survey of combined handwriting and speech modalities for user authentication", In *6th Int'l Conf. on Recent Advances in Soft Computing (RASC 2006)*, pp. 496–501, Canterburry, Kent, United Kingdom, 2006.

[8] A. Humm, J. Hennebert, and R. Ingold, "Combined handwriting and speech modalities for user authentication", *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 39(1), January 2009, pp. 25–35.

[9] A. Humm, R. Ingold, and J. Hennebert, "Spoken handwriting verification using statistical models", In *Proceedings of 9th International Conference on Document Analysis and Recognition (ICDAR'07)*, pp. 999–1003, Curitiba (Brazil), September 23 - 26 2007.

[10] M. Kinsbourne and J. Cook, "Generalized and lateralized effects of concurrent verbalization on a unimanual skill", *Quarterly Journal of Experimental Psychology*, 23(3), 1971, pp. 341–345.

[11] S. Krawczyk and A. K. Jain, "Securing electronic medical records using biometric authentication", In *Audio- and Video-based Biometric Person Authentication (AVBPA)*, pp. 1110–1119, Rye Brook, NY, 2005.

[12] M. Liwicki, A. Schlapbach, H. Bunke, S. Bengio, J. Mariéthoz, and J. Richiardi, "Writer identification for smart meeting room systems", In *Proceedings of the 7th International Workshop on Document Analysis Systems*, volume 3872 of *LNCS*, pp. 186–195. Springer, 2006.

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

[14] B. Ly Van, S. Garcia-Salicetti, and B. Dorizzi, "Fusion of HMM's Likelihood and Viterbi Path for On-Line Signature Verification", In *Biometrics Authentication Workshop*, pp. 318–331, May 15th 2004. Prague.

[15] Y. Nakamura and M. Kidode, "Online writer verification using kanji handwriting", In B. G"unsel, A. K. Jain, A. M. Tekalp, and B. Sankur, editors, *Int'l Workshop on Multimedia Content Representation, Classification and Security (MRCS'06)*, volume 4105 of *Lecture Notes in Computer Science*, pp. 207–214, Istanbul (Turkey), 2006. Springer.

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

[17] A. Ross and R. Govindarajan, "Feature level fusion of hand and face biometrics", In *SPIE Conference on Biometric Technology for Human Identification II*, volume 5779, pp. 196–204, Orlando, USA, 2005.

[18] C. Vielhauer, *Biometric User Authentication for IT Security*, Springer, 2006.

[19] A. Wahl, J. Hennebert, A. Humm, and R. Ingold, "Generation and evaluation of brute-force signature forgeries", In *Int'l Workshop on Multimedia Content Representation, Classification and Security (MRCS'06)*, pp. 2–9, Istanbul, September 2006.

[20] Y. Watanabe, K. Iwata, R. Nakagawa, K. Shinoda, and S. Furui, "Semi-synchronous speech and pen input", In *32nd International Conference on Acoustics, Speech, and Signal Processing (ICASSP'07)*, Honolulu, 2007.