Fusion Strategies for Speech and Handwriting Modalities in HCI

Claus Vielhauer[a], Sascha Schimke[a], Valsamakis Thanassis[b] , Yannis Stylianou[b]
[a]Otto-von-Guericke University Magdeburg, Universitaetsplatz 2, D-39106, Magdeburg, Germany
[b]University of Crete, Department of Computer Science, Heraklion, Crete, Greece

ABSTRACT

In this paper we present a strategy for handling of multimodal signals from pen-based mobile devices for Human to Computer Interaction (HCI), where our focus is on the modalities of spoken and handwritten inputs. Each modality for itself is quite well understood, as the exhaustive literature demonstrates, although still a number of challenges exist, like recognition result improvements. Among the potentials in multimodal HCI are improvements in recognition and robustness as well as seamless men-machine communication based on fusion of different modalities by exploiting redundancies among these modalities.[16] However, such valuable fusion of both modalities still poses some problems. Open problems today include design approaches for fusion strategies and with the increasing number of mobile and pen-based computers, particularly techniques for fusion of hand-writing and speech appear to have a great potential. But today few publications can be found that addresses this potential.

In this work we introduce a conceptional approach based on a model to describe a bimodal HCI process. We analyze four exemplary applications with respect to the structure of this model, and highlight the open problems within these applications. Further, we will outline possible solutions to these challenges. Having such fusion model for HCI may simplify the development of seamless and intuitive to user interfaces on pen-based mobile devices. For one of our application scenarios, a bimodal system for form data recording and recognition in medical or financial environment, we will present some first experimental results.

Keywords: Biometrics, handwriting recognition, speaker recognition, fusion, human computer interaction, multimodality, mobile multimedia

1. Introduction

Mobile computer devices have undergone an enormous technical development in the recent years. Today, many portable computers are audio enabled, have displays allowing for a perceptible video presentation, have integrated digitizer displays for pen-based input and possess sufficient computing power for multimedia applications, which a few years ago were imaginable only on stationary computers. This development imposes a number of challenges for scientists and developers in the area of HCI, as it allows not only the exploitation of different singular modalities for man-machine communication, but also combination of different modalities. International research projects such as SIMILAR currently address these challenges and attempt to open perspectives towards interface concepts, which are in analogy to the common human-human interaction.

Undoubtedly, two important means for the interaction between two human beings in daily life are handwriting and speech. Human-friendly interfaces are expected to support human handwriting as a major input modality. In practice, an interface based on natural handwriting has to support textual/symbolic recognition of human handwriting and signature identification and recognition, as hand-written signature can be considered as an accepted biometrics for document management. On the other hand, speech is another meaningful modality and is highly desirable to be supported by interfaces. Users prefer entering descriptive information via speech while they prefer writing for digits and symbols. Also, characteristic features from voice samples may be used for biometric identity recognition of users, allowing for automated user-awareness for the computer. To make use of both advances in pen-based hardware and natural language processing, speech and writing modes should provide parallel or duplicate (mutual disambiguation is then possible) functionality which means that users can accomplish their goals using either mode.

The goal of this paper is to outline the potential of biometric techniques based on speech and handwriting in the scenario of mobile devices. In order to do so, first an introduction to the goals of biometric techniques for both modalities is given, followed by considerations for the design of a fusion model. We then analyze four exemplary applications with respect to the structure of this model, and highlight the open problems within these applications. Further, we will outline possible solutions to these challenges. We will then suggest a bimodal fusion model for user authentication un-

Multimedia on Mobile Devices, edited by Reiner Creutzburg,
Jarmo H. Takala, Proc. of SPIE-IS&T Electronic Imaging,
Vol. 5684 © 2005 SPIE and IS&T · 0277-786X/05/$15

der noisy conditions in one of our scenarios, which has been implemented and experimentally evaluated, for which we will present some first results.

1.1. Overview of Speech Biometrics

Speech has a unique advantage over other biometrics by relying on the modality, which is the primer way of communication and is especially important in applications such as telephony. By extracting appropriate features from a person's voice the uniqueness of the physiology of the vocal tract and the articulatory properties can be captured to a high degree and can serve the purpose of authentication. Similar features are also used for speech recognition. Speech and speaker recognition modules have been developed for text-dependent (easier task) as well as for text-independent (more challenging) systems. Despite impressive results in recognition scores many unknown factors in speech and speaker recognition still exist: uniqueness (in speaker recognition case), speakers behavior, robustness in ad-verse acoustic conditions, etc.

1.2. Overview of Handwriting Biometrics

Processing of handwritten input nowadays has two points of origin and at least three goals. Two kinds of data acquisition are differentiated – off-line and on-line. The off-line method has 2 dimensional images or pictures of text as input, whereas in on-line processing the input data is available as a set of signals, representing the pen movement. The three aforesaid goals are biometric authentication, textual recognition and data retrieval in handwritten documents. Biometric authentication tries to answer the question, "Who is the writer?", textual recognition asks for the content ("What was written?") and in retrieval a set of handwritten documents is searched for a text with special properties (e.g. containing of a special piece of information).

2. Fusion Model

We introduce a new conceptional model for fusion of the two modalities of speech and handwriting in HCI. The model on the one hand consists of three phases of a process view: User Authentication, Textual Recognition and Semantic Analysis. On the other hand, we differentiate two temporal aspects of multimodal signal fusion: synchronous and asynchronous modes. The two dimensions and the structure of of this model view are outlined in Figure 1 whereas Figure 2 illustrates the process phases. Our model will be explained in detail in the remaining part of this section.

Dimension A:
3 Phases:

1. Authentication
   - Implicit (system recognizes the user by biometric techniques)
   - Explicit (user declares her identity)
2. Textual Recognition
3. Semantic level

Dimension B:

- multimodal synchronous
- multimodal asynchronous

Figure 1 – Two dimensions of multimodality of speech and handwriting

- With respect to the first process step (*authentication phase*), the identity of the origin of the input signals is verified or determined by using biometric algorithms for both modalities of handwriting and speech. In writer authentication, either text-dependent (e.g. by signature verification[6, 8, 13]) or text independent approaches[14] can be considered to identify the writer. For the modality of speech, also two basic concepts of text-dependent and text independent approaches for speaker identification can be found.[1, 11] Having information about the origin can be beneficiary for purposes of security (i.e. non-repudiation) as well as for convenience e.g. selecting a special, user trained recognition algorithm.

- The second process phase (*textual recognition*) of a combined handwriting-speech interface attempts to recognize textual content from the handwritten or spoken input. In the respective discipline of handwriting and speech recognition, a variety of approaches have been introduced in the recent years, with a strong tendency of improvements in recognition accuracy.[9, 10, 17]

- In the last phase of a human-to-computer interaction (*semantic analysis*), semantic knowledge may be retrieved from the textual content, for example to improve recognition accuracy on a word or sentence level. With Semantic Analysis the *domain* of the content can be identified, i.e. if the system detects special catchwords which are characteristic for a domain, then the textual recognition algorithms may be tuned by using special domain/purpose dictionaries. Recognition applications use semantic knowledge for example to select appropriate data sets in dictionary based approaches.[2, 3, 7]

With respect to the temporal classification, we differentiate into synchronous and asynchronous fusion. Humans may produce the two different physical phenomena belonging to the same semantic information either simultaneously or sequentially. One example for simultaneous input is, if the presenter is speaking to an audience and illustrating at the same time, whereas sequentially input for example is the seamless switching of the interface modality from writing to speaking.

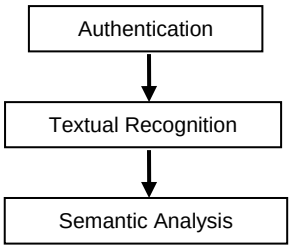

Figure 2 – Three-phase model

In a practical bimodal user interface scenario, each of the process steps can be modeled either synchronously or asynchronously, depending on the application. The following section will discuss aspects of these scenarios by focussing on three hypothetical scenarios.

Considering the example of biometric authentication, the actual fusion of the different modality signals can be performed mainly at three different levels. The process of biometric user authentication can be outlined by the following steps: a) acquisition of raw data, b) extraction of features from these raw data, c) computing a score for the similarity or dissimilarity between these features and a previously given set of reference features and d) classification with respect to the score, using a threshold. This process of a biometric authentication is illustrated in Figure 3. The results of the decision processing steps are *true* or *false* (or *accept/reject*) for verification purposes or the user identity for identification scenarios.

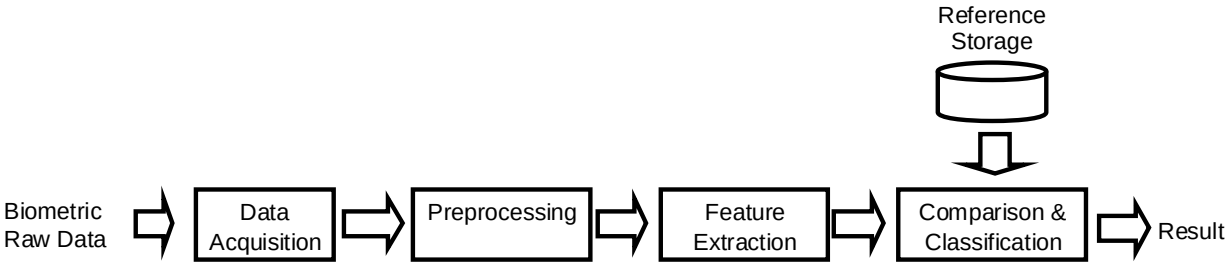

Figure 3 – Process of a biometric authentication

The fusion of different signals can be performed 1) at the raw data or the feature level, 2) at the score level or 3) at the decision level. These different approaches have advantages and disadvantages.[12]

For *raw data* or *feature level* fusion, the basis data have to be compatible for all modalities and a common matching algorithm (processing step c) must be used. If these conditions are met, the separate feature vectors of the modalities easily could be concatenated into a single new vector. This level of fusion has the advantage that only one algorithm for further processing steps is necessary instead of one for each modality. Another advantage of fusing at this early stage of processing is that no information is lost by previous processing steps. The main disadvantage is the demand of compatibility of the different raw data of features. For example, fusing of speech spectrum features and fingerprint minutiae is not possible. In commercial products, often no access to the raw data or the extracted features is possible, so in that case fusion at this level is impossible.

The fusion at *score level* is performed by computing a similarity or dissimilarity (distance) score for each single modality. For joining of these different scores, normalization should be done. For example, a simple summation of the different scores will probably not yield the best results, if one score is a similarity measure while the other is a distance value.

The straightforward and most rigid approach for fusion is the decision level.[12] Here, each biometric modality results in its own decision; in case of a verification scenario this is a set of *trues* and *falses*. From this set a kind of voting[19] (majority decision) or a logical *AND* or *OR* decision can be computed. This level of fusion is the least powerful, due to the absence of much information.[12] On the other hand, the advantage of this fusion strategy is the easiness and the guaranteed availability of all single modality decision results.

In practice, score level fusion is the best-researched approach, which appears to result in better improvements of recognition accuracy as compared to the other strategies.

3. Application Scenarios

Generally the goals of HCI applications can be classified in “Who speaks or who writes?” (speaker/writer identification), “What is spoken and what is written?” (textual recognition) and “What is the relation between both signals?”. With regard to these differentiations, we will describe four application scenarios in this section.

3.1. Conference Talk / Lecture

In situations like a conference talk or a lecture, one single person is the origin of speech. In some cases, that person also produces an additional writing signal, e.g. if she writes or illustrates on a digital white board (e.g. Xerox LiveBoard) or on a tablet PC device connected to a projector. Further, besides to the orator, persons from the auditory could write down notes about the talk or lecture content.[5] Thus, one speech and one or additional writing signal can be produced and recorded synchronously while talk. Due to the nature of these signals, we may assume that these are correlated by the time of their occurrence. However, since in the very rare cases notes of the auditory are word by word transcriptions of the speech signal, an enhancement of recognition of one signal on a word level by using the other signals seems hardly feasible, although these signals are related on a semantic level. Nevertheless these different signals can be interpreted and recognized.

One of the possible goals for such multimodal recording can be the following; from the textual representation, we may determine similar or equal words or semantics in both modalities, which are useful for indexing. This indexing can be the basis for future user interfaces allowing context sensitive and modality independent retrieval.

Aspects of security by biometric authentication are of marginal relevance in conference scenarios, since the identity of the speaker in most cases is known and the identity of auditory is not important. Of course, for enhancing recognition results, signal origin identity should be used.

3.2. Interrogation / Interview

In interrogation situations, primary one person speaks while another person transcribes the textual content of what is said. This can be done using only catchwords or word by word. The latter case mostly requires a stenographer. The recognition results of these input signals (speech and handwriting) could be disambiguated by each other. For example, if the speech recognizing algorithm has an output of two resulting words A and B with the same matching probability, while the handwriting recognizer outputs B, C and D. As one possible result, the combination of both output sets can conclude, B is the correct value.

In the scenario of interrogations the identities of the origins of the different modality signals are important for reasons of non-repudiation.

3.3. Form Filling / Completion

Other scenarios for use of multimodal speech and handwriting signals are user interfaces for computers. For example, if the user needs to fill in a form, she has the choice to use a pen or her voice to fill in the fields of the form. In this case the origin of both signals is the same person. Since in most cases, the user would choose just one of the two modalities for a field, it is hardly possible to disambiguate one signal by using the other one. On the other hand, because the type of fields in most forms is known, the recognition rate can be quite high. For example, if a field is a German or a Greek ZIP code, the input can only consist of five digits.[14] So even if the letter *I* looks like the digit *1*, the algorithm may recognize the *1* correctly. Using both, speech and handwriting input signals, the user can freely choose, which one of them she will use for different form field types. Perhaps, some fields are easier using writing (names or words in foreign languages) while other ones may be easier and faster to provide by voice (e.g. numbers). Another advantage of multimodal interfaces over mono-modal systems is that, depending on the situation, the user might not be able to use one modality so she simple can switch to the other. An example for this is a medical ward round. In this case, the physician may have no hand free to fill in an anamnesis form; however, using a multimedia enabled mobile computer, he could use his voice.

3.4. Multimodal Authentication in Noisy Situations

As in the previous scenario, the authentication of the origin of the input is in form filling scenario in many cases of importance. As in the analogous world of paper forms, for example a physician should sign her forms, to eliminate chances for repudiation. Using handwriting biometrics, with or without combination by speech, on mobile multimedia devices such as tablet PCs, may allow transferring this intuitive process to the digital domain.

The scenario is for authentication purposes. Performance of the speaker identification system may sometimes be unacceptably low due to noisy conditions during the acquisition of the input signal (speech). For this reason an aid from a second system can come in order to compensate for this inadequacy. In the case discussed herein the aid comes from a handwriting (HW) recognition system. Especially dynamic (in contrast to static) biometrics are strongly influenced by noise. In biometric speaker authentication, for example loud traffic noisiness or office sounds have a negative impact on the authentication error rates. This is true for biometric handwriting authentication, too. Here, for example physical agitation, like appearing in moving cars, lifts or trams, affect the quality if the handwriting signal and so the authentication rates.

To solve this problem, an approach could be to combine the authentication results of different input signals. The user should be able to choose her preferred modality in each situation and even combinations of modalities should be possible. In this case, for example a verification test succeeds, if verification using of at least one of the signals succeeds or if each partial verification succeeds with a less strict threshold. This can be formalized as following. Let be $V_r(s)$ a monomodal verification process, where r is the reference data set, s is the test sample, D is a distance measure or r and s and τ is a threshold:

$$V_r(s) = \begin{cases} \textit{true} & \text{if } D(r, s) \le \tau, \\ \textit{false} & \text{else.} \end{cases} \tag{1}$$

Using the same notation, then the multimodal verification using speech (*SP*) and handwriting (*HW*) data can be defined as follows:

$$V_{r_{SP}, r_{HW}}(s_{SP}, s_{HW}) = \begin{cases} \textit{true} & \text{if} \quad D(r_{SP}, s_{SP}) \le \tau_{SP}, \\ & \text{or} \quad D(r_{HW}, s_{HW}) \le \tau_{HW}, \\ & \text{or} \quad D(r_{SP}, s_{SP}) \le \tau'_{SP} \text{ and } D(r_{HW}, s_{HW}) \le \tau'_{HW}, \\ \textit{false} & \text{else.} \end{cases} \tag{2}$$

Here r_{SP}, r_{HW}, s_{SP}, s_{HW}, τ_{SP} and τ_{HW} are the respective speech or handwriting references, samples and thresholds, analogous to formula 1. The multimodal thresholds τ'_{SP} and τ'_{HW} can be less strict than τ_{SP} and τ_{HW}: $\tau'_{SP} > \tau_{SP}$ and $\tau'_{HW} > \tau_{HW}$.

For identification instead of verification scenarios, as shown before, the following approach could be used. In HW recognition system we compute distances of the input signature to reference signatures form all signees. So we have ten values i.e. the distances of the input to all other signees. This is also true for the speaker recognition system, that is, we compute some sort of distance from the input speaker to all others. We have to merge these to results in order to be able to produce high accuracy. The fusion procedure suggested can be described as follows. Because the results from the two systems are in different scale we have to transform them in order to compare and merge. The transformation function used is a modified z-score, that is

$$z = \frac{x - \min(x)}{\sigma}$$

where x is the vector with scores (results) from all individuals, $\min(x)$ is the minimum value and σ is the standard deviation of all scores. This transformation produces a new vector of scores and this is done for both speaker recognition system and HW recognition system. Thus, we conclude with two score vectors coming from the two recognition systems denoting by z_{SR} and z_{HW}. In order to merge the two outputs we adopt the following very simple formula

$$z_{final} = z_{SR} + z_{HW}$$

4. First Experimental Setup

For testing, we used speech and handwriting data of ten subjects. For speech authentication, each subject had to read 15 sentences for training and one different sentence for testing. The spoken inputs are German. For signature authentication, each subject had to write down her or his signature five to ten times as reference data. One further signature sample per subject for testing was acquired.

The audio data were recorded in a special soundproof cabin. For simulating a mobile setup, we superimposed the clean audio data with two kinds of noise with different SNR to the raw audio signal; generated white Gausian noise, and recorded laptop fan noise.

For handwriting data acquisition, a set of actual tablet devices was used, which use the same digitizing technology as those digitizers integrated in most common tablet PCs[17] and yields the same kind of signals. Consequently, although samples were recorded under laboratory conditions, we assume our results are relevant for applications on handheld mobile devices as well.

The tests have been done in identification mode. Therefore for each test sample (test sentence or handwritten signature) the distance score against all references (as described in 4.1 and 4.2). The identification result is the identity of that reference with the least distance score. (See section 3.4)

4.1. Speech Biometrics

In this section it is briefly described how the speech signal can be used in speaker identification tasks. Our goal is to decide who from a known group of individuals is speaking an utterance. Over the past few years many methods have been suggested in the literature. In this work, the discussion will focus on the method described in.[11]

The system described herein for speaker identification is text-independent i.e. the speech used to train and test the system is completely unconstrained. This means that the speaker is not restricted to say any particular phrase in contrast to text-dependent systems where a particular phrase (e.g. a digit string) must be uttered.

In what follows, it is portrayed the use of a Gaussian mixture model (GMM) as a robust representation of speaker identity and a maximum likelihood classifier. One way to represent the probabilistic variability of speech production is through a mixture of different Gaussian probability density functions. More specifically, the distribution of feature vectors extracted from a person's speech is modeled by a Gaussian mixture density. A GMM is a weighted sum of M component densities and is given from

$$p(x \mid \lambda) = \sum_{i=1}^{M} a_i f_i(x)$$

where a_i are the mixture weights and x is a d-dimensional feature vector. For speaker identification, each speaker is represented by a GMM and it is commonly referred to by his model λ. GMM parameters are estimated using the standard maximum likelihood estimation (MLE) method via the Expectation-Maximization (EM) iterative algorithm. The initialization is accomplished using vector quantization (VQ).

There are a variety of attributes that characterize a particular speaker. The goal is to choose a kind of attributes that are easily extracted by machine for automatic speaker recognition. In other words, we want our feature set to reflect the unique characteristics of a speaker. There are many studies which have directly addressed the feature selection problem for speaker recognition. Spectral measurements were found to be practically good for our purpose. In our study, we have chosen to use the mel-scale cepstral coefficients which exploit auditory principles as well as decorrelating properties of the cepstrum. In addition, the mel-cepstrum coefficients are amenable to compensation for convolutional channel distortion. As such, the mel-scale cepstral coefficients have proven to be one of the most successful feature representations in speaker recognition tasks. The feature extraction consists of the following steps. Every 10ms the speech signal is multiplied by a Hamming window with duration 20ms to produce a short time segment for analysis. The magnitude spectrum from the 20ms short-time segment of speech is pre-emphasized and processed by a simulated mel-scale filterbank. The log-energy filter outputs are then cosine transformed to produce the cepstral coefficients. The zeroth cepstral coefficient is not used in the cepstral feature vector because it actually represents the energy level of speech segment and has nothing to do with speaker identity. Most importantly, to take best advantage of the information content of the speech utterances, it is used a speech activity detector (SAD) prior to feature extraction in order to avoid modeling the environment rather than the speaker.

Given a sample of speech utterance the goal is to extract the identity of the person speaking the utterance. This is accomplished by using a maximum likelihood classifier. It is assumed that a reference group of N speakers is available i.e. $S = \{\lambda_1, \lambda_2 \ldots \lambda_N\}$. The objective is to find the speaker that is most probable of having the input feature vector sequence $X = \{x_1, x_2 \ldots x_T\}$. Put another way, the maximum value is picked from the following sequence of probabilities $\{ P(\lambda_1 \mid X), P(\lambda_2 \mid X) \ldots P(\lambda_N \mid X) \}$. And utilizing the Bayes' formula the aforementioned probabilities are transformed into

$$P(\lambda_i \mid X) = \frac{p(X \mid \lambda_i) P(\lambda_i)}{P(X)}, \quad i = 1,2 \ldots N$$

This formula can be further simplified assuming equal prior probabilities $P(\lambda_i)$ and observing that the denominator is constant for all speakers, so it can be left out from calculations. Furthermore, if we take the assumption of independent observation, the formula becomes

$$P(\lambda_i \mid X) = k \prod_{t=1}^{T} p(x_t \mid \lambda_i)$$

where k is a constant for the equality to hold.

4.2. Handwriting Biometrics

The signature authentication algorithm for the test is described by Schimke et al.[13] The basic idea of the matching is to use a distance measure for string-like sequence data, as even used in the domain of text processing (fuzzy search for substrings), in bioinformatics for searching in gene sequences or in handwritten text recognition. The string distance measure we used is called *edit distance* and the main idea is to count the character wise operations to transform one string sequence into another one. The allowed character operations are *insert*, *delete* and *replace*. The higher is the similarity of two strings, the less is the minimal number of operations for that transformation.

To derive feasible string sequence from signature samples, the handwriting signals over the time (x- and y-position, pressure, velocities in x- and y-direction, pen tip track speed, …) where analyzed. Special points which describe these signal functions – the local minima and maxima – are interpreted as symbols from an alphabet and are put together in a sequence, ordered by the time of their respective occurrence. The different symbols represent the kind of special point – type of signal and minimum or maximum.

To compare the distance of two handwritten samples, the string sequences of these samples are derived. For these two strings the edit distance is calculated and the resulting value is normalized regarding to the length of the two string sequences.

4.3. Fusion Results

The results of our fusion experiments base on speech and handwriting data, which were acquired and used as described in section 4. Experimental results for the suggested procedure are depicted in Figure 4. The performance of the speaker

recognition system, handwriting authentication system and the fusion of the two systems can be seen. It is apparent that albeit the low performance of the HW recognition system the fusion performs equally well to the speaker recognition system and sometime better. The left part of Figure 4 shows the performance of experiments, using white Gausian noise, while the right one shows the results of using laptop fan noise. At the abscissa of both diagrams, the SNR can be seen. The ordinate displays the identification rate at the respective noise ratio in %.

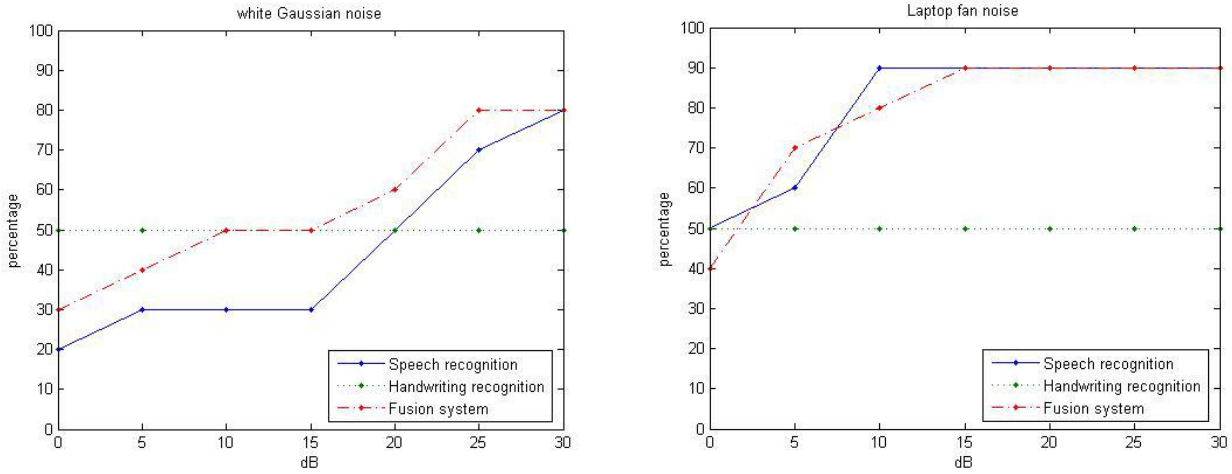

Figure 4 – Identification rates (in %) for speech, handwriting and fused data, in different noisy situations.

As can be seen in the fusion curves in figure 4, in noisy environments the using of handwriting authentication additionally to the speech signal can enhance the identification rate – the fusion curve is nearly every time better, when the SNR is low. It has to be mentioned, that only the audio signal is affected by a noise, in our experimental test setting. For mobile applications, the observation could be interesting, that laptop fan noise has less affect to the identification than artificial white Gausian noise.

5. CONCLUSIONS AND FUTURE WORK

We have discussed different dimensions of fusion of speech and handwriting data in the domain of human computer interaction. Furthermore we presented various strategies of signal fusion especially for the authentication step in the HCI. Our first results show, that using more than one modality indeed can get better results than only one modality.

We did not investigate the other dimensions of multimodality like recognition of spoken or written contents. Furthermore we concentrated only on audio noise. Both aspects, the recognition of contents as well as noisy handwriting data, will be focused in future work.

Acknowledgements

This work has been partly supported by the EU Network of Excellence SIMILAR (Proposal Reference Number: FP6−507609). The contents of this publication are the sole responsibility of the authors and can in no way be taken to reflect the views of the European Union.

REFERENCES

1. J. P. Campbell, “Speaker recognition: A tutorial”, in *Proc. IEEE*, Vol. 85, pp. 1437–1462, Sept. 1997.
2. D. Dori, “Syntactic and Semantic Graphics Recognition: The Role of the Object-Process Methodology”, in *GREC’99*, pp. 277–287, 2000.
3. F. Grandidier, R. Sabourin, and C. Y. Suen, “Integration of Contextual Information in Handwriting Recognition Systems”, in *Proceedings of ICDAR*, 2003.
4. A. K. Jain, and A. M. Namboodiri, “Indexing and Retrieval of On-line Handwritten Documents”, in *Proceedings of ICDAR*, 2003.

5. J. A. Landay, and R. C. Davis, “Making sharing pervasive: Ubiquitous computing for shared note taking”, *IBM Systems Journal*, Vol. 38, No. 4, 1999.
6. F. Leclerc, and R. Plamondon, “Automatic Verification and Writer Identification: The State of the Art 1989–1993”, in *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 8, pp. 643–660, 1994.
7. J. Navratil, J. Kleindienst, and S. H. Maes, “An instantiable speech biometrics module with natural language interface: Implementation in the telephony environment”, in *IEEE ICASSP 2000*, Istanbul, Turkey, June 2000.
8. R. Plamondon, and G. Lorette, “Automatic Signature Verification and Writer Identification - the State of the Art”, in *Pattern Recognition*, Vol. 2, pp. 107–131, 1989.
9. R. Plamondon, and S. N. Srihari, “On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey”, in *Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 1, pp. 63–84, 2000.
10. L. R. Rabiner, and B. H. Juang, “*Fundamentals of Speech Recognition”*, Prentice Hall, 1993.
11. D. A. Reynolds, "*Speaker Identification and verification using Gaussian mixture speaker models"*, Speech Communications, pp.91-108, 1995.
12. A. Ross, and A. K. Jain, “Multimodal Biometrics: An Overview*”, in Proceedings of 12th Signal Processing Conference (EUSIPCO)*, pp. 1221–1224, 2004.
13. S. Schimke, C. Vielhauer, J. Dittmann, “*Using Adapted Levenshtein Distance for On-Line Signature Authentication*”, ICPR, 2004
14. L. Schomaker, M. Bulacu, M. van Erp, “*Sparse-Parametric Writer Identification using Heterogeneous Feature Groups*”, International Conference on Image Processing. Vol. 1, pp. 545-548, 2003
15. G. Seni, K. Rice, E. Mayoraz, “*Online Handwriting Recognition in a Form Filling Task – Evaluating the Impact of Context-Awareness*”, SPIE-IS&T 2003, pp.109-115
16. SIMILAR Network of Excellence: **–** The European taskforce creating human-machine interfaces SIMILAR to human-human communication, http://www.similar.cc/
17. T. A. Stephenson, M.M.Doss, H.Bourlard, "*Speech recognition with auxiliary information,"* IEEE Trans. on Speech and Audio Processing, May 2004, pp. 189-203, Vol.12(3)
18. Wacom Technologies Co., http://www.wacom.com/tabletpc/index.cfm, 2004.
19. Y. Zuev, and S. Ivanov, “The voting as a way to increase the decision reliability”, in *Proceedings of Foundations of Information/Decision fusion with applications to engineering problems*, pp. 206–210, 1996.