A Generic Protocol for Multibiometric Systems Evaluation on Virtual and Real Subjects

Sonia Garcia-Salicetti, Mohamed Anouar Mellakh,
Lorène Allano, and Bernadette Dorizzi

Département Electronique et Physique, Institut National des Télécommunications
9 Rue Charles Fourier, 91011 Evry France
{Sonia.Salicetti,Mohamed.Anouar_mellakh,Lorene.Allano,
Bernadette.Dorizzi}@int-evry.fr

Abstract. We propose in this paper a methodology for multibiometric systems evaluation on databases of virtual and real subjects of limited size (about 100 persons). Our study is limited to two biometric traits (modalities) that are a priori mutually independent, namely on-line signature and voice. Experiments are conducted on bimodal data of real subjects of the BIOMET database [9] and on several databases of virtual subjects constructed from BIOMET.

1 Introduction

The evaluation of a multibiometric system is not an easy task: indeed, there are very few available multimodal databases (M2VTS [1,2], XM2VTS [3,4], BANCA [5,6], DAVID [7], SMARTKOM [8]), most of which contain only two biometric modalities, usually face and voice. Also, multimodal databases available nowadays contain only about a hundred subjects, which makes difficult to extrapolate the success of a multimodal algorithm or method when being tested on a large population (thousands or millions of people). Moreover, multimodal databases more recently constructed as BIOMET [9], or under construction [10] have the tendency to contain more modalities (4 or 5) but not more subjects. In this precise matter, the order of magnitude of such databases remains indeed in about one hundred subjects. This can be explained by the fact that acquiring multimodal data is more time consuming and expensive than acquiring data from a single modality, and rises some other problems as higher acquisition failure and critical personal data protection. Indeed, acquisition failure is generated because the more modalities there are, the more it is likely that a data sample cannot be acquired in a given modality, thus generating the loss of a complete multimodal sample. This phenomenon is of course amplified whenever several sessions are recorded. Also, regarding personal data protection, the fact that a data collection may contain together fingerprints, signature, iris, and face, among others, of a given person, is obviously critical and not easily acceptable for donators which can be afraid of misuse or forgeries.

Many works in the multimodal fusion literature give results on about 100 real subjects, with no insight in the fact that such results may be in fact very biased. We address this problem in the present work and propose a new protocol for multibiometric systems evaluation on limited size databases of real subjects.

Moreover, it is also natural to study the possibility of using databases of virtual subjects, that is an individual generated by combining different biometric traits (mo-

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 494–502, 2005.

dalities) that belong to different persons. This procedure would simplify multimodal data construction because it would be sufficient to merge two or more databases of approximately the same number of subjects, containing each different modalities, to generate a multimodal data corpus containing more modalities. Although this question is crucial for the progress of research in multimodal fusion, few works have exploited the creation of virtual subjects for multimodal fusion [10,11]. The first question that arises is: which is the validity of this procedure? Then the next question is: if it is valid, which methodology should be used to evaluate multimodal systems on a given corpus of virtual subjects? Our aim in this work is also to answer to such crucial questions.

To that end, our methodology has been to create virtual subjects with data coming from a multimodal database of real subjects, that is the BIOMET database [9]. This permits us to do a comparative study of the behaviour of a bimodal fusion system (on-line signature and voice) on the real subjects and on several databases of virtual subjects generated from BIOMET. Indeed, the originality of this work is that we set the problem of using virtual subjects for systems evaluation relatively to the use of real subjects in multimodal databases. This gives more insight into what is in fact a real subjects database relatively to a virtual subjects one, and how evaluation should be performed in both cases.

As mentioned, our work is limited to two modalities, voice and on-line signature, combined in a previous work [12]. The choice of the modalities is a delicate question since it rises the problem of their mutual dependence/independence. We focus here in the combination of modalities that are a priori mutually independent, since it is only in this framework that we may consider building a virtual subject.

We combine such two modalities by a Support Vector Machine classifier with a linear kernel [13], a statistical technique that allows to learn the coefficients of a hyperplane and does not necessitate a priori scores normalisation. Actually, the objective of the present work is not to compare different classifiers. We show in this framework that a bimodal (voice, signature) database of real subjects of limited size (around 100 persons) introduces a bias when evaluating the fusion system, because the size of the database does not permit to represent all the possible data variability in the bimodal sense. Moreover, we show that using databases of virtual subjects is equivalent in certain conditions (with a given protocol) to the use of a database of real subjects of limited size. We provide here an evaluation protocol on both types of databases.

In the following, both experts (voice and signature) and the fusion method are first described (section 2), the experimental setup on BIOMET bimodal data is given in section 3, section 4 details the creation of virtual subjects from BIOMET bimodal data and section 5 focuses on comparative fusion experiences on real and on virtual subjects. Finally, conclusions and perspectives of this work are given in section 6.

2 Fusion of On-Line Signature and Voice

This study is carried out on a bimodal fusion system composed of two mono-modal biometric systems: a signature verification system described in [14] and a text-independent Speaker Verification system already described in [12]. We briefly describe in the following the two systems and the fusion method.

2.1 The Signature Verification System

As described in [14], each writer's signature is modelled by a continuous left-to-right HMM [15], characterised by a given number of states with an associated set of transition probabilities among them, and, in each of such states, a continuous density multivariate Gaussian mixture. The topology of the HMM only authorises transitions from each state to itself and to its immediate right-hand neighbours. An optimal number of states is computed for each writer and a personalised feature normalisation (of 25 features) is carried out to improve the quality of the modelling. The system exploits a fusion strategy of two complementary information provided by both the HMM likelihood and a "segmentation vector" obtained from the Viterbi path of the HMM modelling a given writer. As shown in [14], the combination of such two information permits to better separate the genuine and impostor distributions, thus improving significantly writer verification results.

2.2 The Text-Independent Speaker Verification System

This system is detailed in [12]. Considering a simple hypothesis test between two hypotheses H_λ (X has been uttered by λ) and H_{λ^*} (X has been uttered by another speaker), the system's output score is: [log ($P_\lambda(X)$) – log ($P_{\lambda^*}(X)$)] where $P_\lambda(X)$ and $P_{\lambda^*}(X)$ are the probability density functions associated to the densities of H_λ and H_{λ^*} given X. A single speaker-independent model is used to represent $P_{\lambda^*}(X)$. This model, also called Universal Background Model (UBM) [16], corresponds to a 256 components GMM with diagonal covariance matrices. Each client model is obtained by a mean-only Bayesian adaptation of the UBM using associated training speech data. The decision score for a test sequence corresponds to the mean log-likelihood ratio computed on the whole test utterance.

2.3 The Fusion Method

In this work, we have performed the fusion of two scores, respectively the outputs of the On-line Signature Verification System and the Text-independent Speaker Verification System, by means of a Support Vector Machine (SVM) [13]. In a few words, SVM's goal is to compute a hyperplane in a large dimension feature space which is considered because the input data are not linearly separable in the original space. The distance between the decision surface and the data is maximized, which leads to good generalization performance [13]. Let $X=(x_i)$ be the data with labels $Y=(y_i)$ where $y_i = 0$ *or* 1 represents the class of each person, and Φ is the function which sends the input data X in the feature space F. The distance between the hyperplane $H(w,b) = \{x \in F: \langle w, x \rangle + b = 0\}$ and X, is called the margin Δ. Following the Structural Risk Minimization (SRM) principle, Vapnik [13] has shown that maximizing the margin (or minimizing $||w||$) leads to an efficient generalization criterion. One defines in F the kernel K as: $K(x,y) = \langle \Phi(x), \Phi(y) \rangle$, that avoids handling directly elements in F. The optimal hyperplane is found by solving a quadratic convex problem and, from the optimality conditions of Karush-Kuhn-Tucker [13], one can rewrite w in the following condensed manner:

$$w = \Sigma_{i \in SV} \alpha_i y_i \Phi(x_i)$$

where $SV = \{i: \alpha_i > 0\}$ denotes the set of support vectors.

We have chosen here, as in [12], $K(x,y)=<\Phi(x),\Phi(y)>^d$ with $d = 1$, that is a linear kernel. We fuse the scores of the two experts, each designed for the same person. We thus give as input to the SVM two scores, one per expert.

The optimization of the SVM was carried out on a database considered for training. During this training step, the optimal hyperplane $H(w^*,b^*)$ is computed. This optimal hyperplane generates a given False Rejection Rate (FRR) and a given False Acceptance Rate (FAR). In order to generate a DET (Detection Error Tradeoff) curve [17] during the test phase, the position of the optimal hyperplane is varied. This means that w^* remains constant but that b varies. This corresponds indeed to the variation of a decision threshold.

3 Experimental Setup on BIOMET Bimodal Data

3.1 BIOMET's Signature and Voice Data in Brief

BIOMET is a multimodal biometric database including face, fingerprint, on-line signature, hand shape and voice. We exploit signature and voice data from 77 people with time variability, captured in the two last BIOMET acquisition campaigns, which have a five months spacing between them. More details on the BIOMET database can be found in [9].

Signature data was captured on a digitizer at a rate of 100 samples per second. Each sample contains 5 information: the coordinates $(x(t),y(t))$ of each point sampled on the trajectory, the axial pen pressure $p(t)$ in such a point, and the position of the pen in space (the standard azimuth and altitude angles in the literature). The total number of signatures available per person is 15 genuine and 12 forgeries, made by four different impostors.

Speech data was recorded in quiet environment and using the same kind of microphone. Sampling rate is 16 kHz and sample size is 16 bits. In each session, each speaker uttered twice the 10 digits in ascending and descending order before reading sentences. The amount of available speech for each speaker is about 90 seconds per session.

3.2 Training Protocols per Modality

The Signature Verification expert is trained on 5 signatures randomly chosen among the 15 genuine signatures available.

As for the Text-independent Speaker Verification expert, each client model is adapted using the 10 digits utterance (about 15s of speech). Test data is composed of a segment of speech of approximately 15s, taken from read utterances. For more details, the reader should refer to [12].

3.3 Building the Bimodal Database of Real Subjects

To build the bimodal database, we associate the input data of the two experts (Signature and Voice). We consider for the voice expert two configurations: one without noise, and another with 0db noise.

This bimodal database is then split in 2 subsets: one of 39 persons devoted to training the Support Vector Classifier, named *FLB* (Fusion Learning Base), and the other of 38 persons for testing purposes, named *FTB* (Fusion Test Base). In order to reduce the bias related to the small number of persons in the database, we consider 50 different couples of training and test databases *(FLB,FTB)*, selected randomly, and compute average Errors Rates on the 50 generated *FTBs*. This choice corresponds to the "Trained-Boot" protocol reported in [18], that corresponds to a variant of the Bootstrap sampling principle [19].

For each person in *FLB* and *FTB*, we have at disposal 5 bimodal client accesses and in average 10 bimodal impostor accesses (this number varies across persons from 6 to 12 impostor accesses).

Figure 1 shows the bimodal scores distribution for the 77 persons of the database, in both voice expert's configurations: without noise ("database1") and with 0db noise ("database2"). We notice that discriminating clients from impostors will be more difficult in the case shown in Figure 1 (right), case in which the voice score is very noisy.

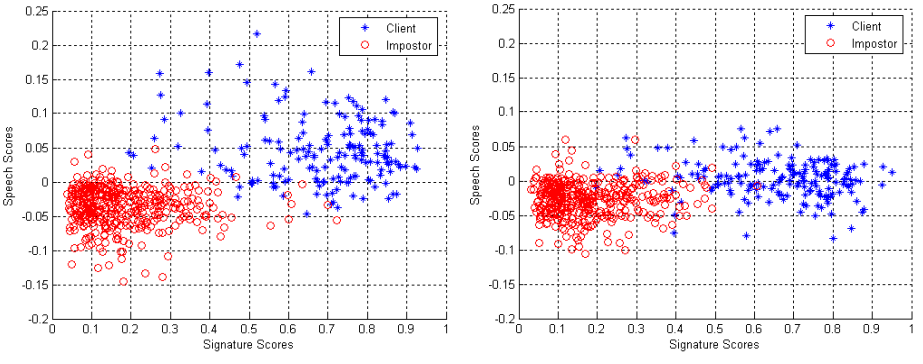

Fig. 1. Bimodal scores' distribution: Signature and Voice without noise (left), Signature and Voice with additive 0db noise (right)

4 Creating Virtual Subjects from BIOMET Bimodal Data

We create a virtual subject by pairing randomly signature data of a given subject to the speech data of another subject. In theory, for two modalities, we can create this way up to *(k-1)!* data sets of virtual subjects, where *k* is the total number of clients in the database. We chose to create 1000 data sets of virtual subjects as in [11].

Every database of virtual subjects is split as described in section 3.3 into a Fusion Learning Base (*FLB*) and a Fusion Test Base (*FTB*). For each possible value of *b* in the equation of the hyperplane *H(w*,b),* where *w** denotes the normal vector to the optimal hyperplane, we compute the mean False Acceptance Rate $\overline{FA}$ and the mean False Rejection Rate $\overline{FR}$ for the 1000 databases of virtual subjects, to obtain a "Virtual Mean DET Curve".

5 Comparative Fusion Experiences on Real and Virtual Subjects

As a first step, we compare the DET curve obtained on the BIOMET database to the 1000 DET curves corresponding to the 1000 databases of virtual subjects. Let's recall that the first curve represents average error rates over 50 different couples *(FLB,FTB)*. Figure 2 (left and right corresponding respectively to the fusion experience without and with additive noise on the voice expert score) shows that the average DET curve on the BIOMET database is inside the band generated by the 1000 DET curves corresponding to virtual subjects sets. This first result permits to conclude that the system behaves on the database of real subjects (when averaging error rates on 50 partitions of the Fusion Learning and Test databases) as on any of the databases of virtual subjects. This also supports the mutual independence assumption between the two modalities that we consider, on-line signature and voice. Moreover, the use of virtual subjects data sets permits to have an estimation of performance variability, providing in fact a "confidence interval" for performance obtained on a real subjects data set of limited size (100 persons). In other words, the database of real subjects is a data set with an inherent bias because of the small number of clients it contains. This bias is greatly increased if a single partition in a Fusion Learning and Testing Databases *(FLB,FTB)* is considered like widely done in the literature. Indeed, the statistics of bimodal data found in the test set (represented by the real subjects present in such set) may be very different of that present in the training set, leading this way to a unreliable and misleading evaluation of the fusion system. It is thus necessary to generate different couples *(FLB,FTB)* that correspond to different distributions of individuals in *FLB* and *FTB* respectively, and to average error rates over those trials. Next experience supports our assumption that considering different partitions or couples *(FLB,FTB)* reduce the bias related to the small number of real subjects in the database. If a large database would be available, this procedure would not be necessary.

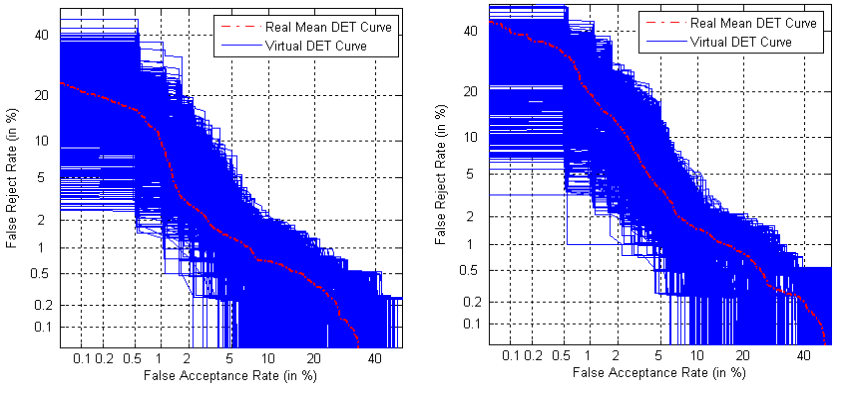

Fig. 2. DET curve for database 1 (left), and for database 2 (right)

Indeed, we now compare, in a second step, the Virtual Mean DET Curve of the 1000 databases of virtual subjects with the mean DET curve on the BIOMET database. In both cases, on database1 (voice without noise) and database2 (voice with 100% noise) shown in Figure 3, we notice that the curves have the same behaviour.

This shows that evaluating this way the fusion system on 1000 virtual data sets is equivalent to evaluating the fusion system on the database of real subjects by averaging results over 50 partitions *(FLB,FTB)* of such database.

The difference between results in Figure 3 left and right can be explained by the fact that the 50 partitions *(FLB, FTB)* are randomly chosen, and therefore are not the same in both cases. This shows that even if 50 partitions or couples *(FLB,FTB)* reduce the bias related to the small number of real subjects in the database, it is still not enough to lead to stable results. Indeed there are C_{77}^{39} possible couples *(FLB, FTB)* and the number of couples that should be considered to "cancel" the bias related to the small number of real subjects in the database is to be studied.

For more insight, we represent in Figure 4 the standard deviation of the errors (False Acceptance Rate and False Rejection Rate) obtained for each value of the decision threshold, on BIOMET data and on 1000 virtual subjects data sets.

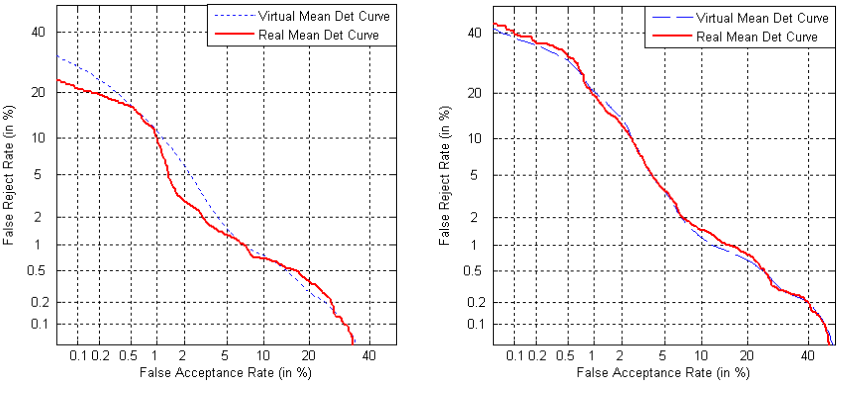

Fig. 3. Virtual Mean DET Curve vs. average Error Rates on the real database for database 1 (left), and for database 2 (right)

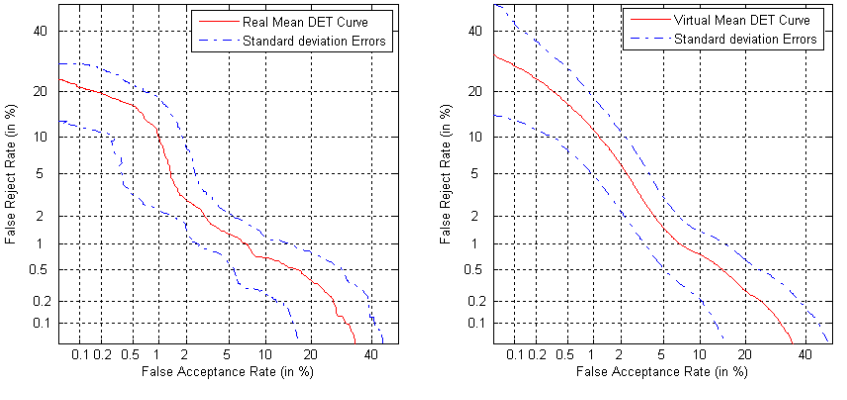

Fig. 4. Mean DET Curve on 50 couples *(FLB, FTB)* on BIOMET data (left) and Virtual Mean DET Curve on 1000 virtual data sets (right), both with associated standard deviation

For this experiment, we chose data without noise. When comparing the standard deviation of errors on real data and on 1000 virtual subjects data sets, we observed that both are comparable when only 50 couples *(FLB, FTB)* are considered. This means that the difference between two samples *(FLB,FTB)* on real data is of the same order of magnitude than the one between two virtual data sets chosen among 1000.

6 Conclusions

We have studied in the present work the possibility of valid multibiometric systems evaluation on limited size databases (about 100 subjects) of real subjects, and also on databases of virtual subjects. Our study focuses on two modalities which are a priori mutually independent, on-line signature and voice, and exploits bimodal data from 77 subjects of the BIOMET database. Several databases of virtual subjects were constructed from BIOMET bimodal data. Our first conclusion is that a limited size database (about 100 subjects) of real subjects behaves exactly as a virtual subjects set of the same size when evaluating the multibiometric system. This of course supports the mutual independence assumption of the two biometric traits that we consider. In other words, this confirms a natural intuition that a database of real subjects has an inherent bias, since each subject represents a specific combination of the modalities considered, and about 100 instances are not enough to cover all the possible variance of such combination, not even for two modalities. This bias is of course stronger if more than two modalities are considered. To cope with this fact, we propose a protocol for multibiometric systems evaluation on limited size databases (about 100 subjects) of real subjects, consisting in creating several partitions (we have shown that 50 partitions is an acceptable compromise) of the data set in a Fusion Learning Base and a Fusion Test Base *(FLB,FTB)* and in averaging error rates over such 50 trials for each value of the threshold. Indeed, evaluating a fusion system on only one partition *(FLB,FTB)* like usually done in the literature, gives biased and thus unreliable results, even if the subjects that are in the database are real!

Moreover, we have shown that it is equivalent to evaluate a fusion system on the database of real subjects by averaging error rates over 50 partitions *(FLB,FTB)*, and on 1000 virtual subjects data sets if a mean False Acceptance Rate and a mean False Rejection Rate are computed on the 1000 data sets for each value of the decision threshold. As a conclusion, we have also proposed a protocol for evaluating a multibiometric system on virtual subjects data sets.

Finally, we can conclude that, in the case of mutual independence of the modalities that are considered, the use of virtual subjects with the protocol above given is a powerful tool to estimate the performance variability, providing a “confidence interval” for performance obtained on a real subjects data set of limited size (100 persons). It is thus recommended for a complete and reliable evaluation of multibiometric systems.

Acknowledgements

This work has been carried out in the framework of GET’s (Groupe des Ecoles des Télécommunications) research project Bio-Identity, to which GET-ENST (Ecole Nationale Supérieure des Télécommunications) participates. The authors thank particularly Gérard Chollet and his team for results of speaker verification.

References

1. S. Pigeon & L. Vandendorpe, "The M2VTS Multimodal Face Database", In Proceedings of AVBPA 97, Springer LNCS, Bigün et al. Eds, 1997.
2. http://www.tele.ucl.ac.be/PROJECTS/M2VTS/
3. K Messer J Matas J Kittler, J Luettin and G Maître, “XM2VTSDB: The Extended M2VTS Database”, Second International Conference on Audio and Video-based Biometric Person Authentication, 1999.
4. http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/results/
5. E. Bailly-Bailliére, S. Bengio, F. Bimbot, M. Hamouz, Jo. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruíz, and J.-P. Thiran. The BANCA Database and Evaluation Protocol. Audio- and Video-Based Biometric Person Authentication (AVBPA), Guilford, 2003.
6. http://www.ee.surrey.ac.uk/Research/VSSP/banca/
7. J.S.D Mason, F. Deravi, C. Chibelushi & S. Gandon BT DAVID – Final Report, Speech and Image PDETessing Research Group, Dept. of Electrical and Electronic Engineering, University of Wales Swansea, UK.
8. http://www.phonetik.uni-muenchen.de/Bas/BasHomeeng.html
9. S. Garcia-Salicetti, C. Beumier, G. Chollet, B. Dorizzi, J. Leroux-Les Jardins, J. Lunter, Y. Ni, D. Petrovska-Delacretaz, "BIOMET: a Multimodal Person Authentication Database Including Face, Voice, Fingerprint, Hand and Signature Modalities", Proc. of 4th International Conference on Audio and Video-Based Biometric Person Authentication, pp. 845-853, Guildford, UK, July 2003.
10. A. Ross, A. Jain, "Information Fusion in Biometrics", Pattern Recognition Letters 24, pp. 2115-2125, 2003.
11. M. Indovina, U. Uludag, R. Snelick, A. Mink, A. Jain, « Multimodal Biometric Authentication Methods: A COTS Approach », Workshop on Multimodal User Authentification (MMUA), pp. 99-106, Santa Barbara, California, USA, Dec. 2003.
12. B. Ly Van, R. Blouet, S. Renouard, S. Garcia-Salicetti, B. Dorizzi, G. Chollet: « Signature with text-dependent and text-independent speech for robust identity verification », Workshop on Multimodal User Authentification (MMUA), pp. 13-18, Santa Barbara, California, USA, Dec. 2003.
13. V. Vapnik, “The Nature of Statistical Learning Theory”, *Statistics for Engineering and Information Science,* Second Edition, Springer, 1999.
14. B. Ly Van, S. Garcia-Salicetti, B. Dorizzi, “Fusion of HMM’s Likelihood and Viterbi Path for On-line Signature Verification”, Biometric Authentication Workshop (BioAW), Lecture Notes in Computer Science (LNCS) 3087, pp. 318-331, Prague, Czech Republic, May 2004.
15. L. Rabiner, B.H. Juang, "Fundamentals of Speech Recognition", *Prentice Hall Signal Processing Series*, 1993.
16. D.A Reynolds, T.F. Quatieri, R.B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models”, Digital Signal Processing 10, Special Issue on the NIST'99 evaluations, pp. 19-41, 2000.
17. A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, "The DET curve in Assessment of Detection Task Performance", Proc. of Eurospeech 1997, 4, pp. 1895-1898, 1997.
18. J. Bigun, J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, "Multimodal Biometric Authentication using Quality Signals in Mobile Communications", in Proc. Of the 12th International Conference on Image Analysis and Processing (ICIAP), 2003.
19. S. Shamsunder, “Signal Processing Applications of the Bootstrap”, IEEE Signal Processing Magazine, pp. 38-55, January 1998.