INTERNATIONAL JOURNAL OF MATHEMATICAL MODELS AND METHODS IN APPLIED SCIENCES

Two factor user authentication via biometry and neural networks

Monica Carfagni, Matteo Nunziati, Matteo Palai

***Abstract*—** A multi-factor authentication framework relays on the usage of two or more sources of data for identity retrieval. Moving from authors' previous experimentation on two factor biometric authentication frameworks, the present paper shows the impact of the usage of neural networks as a way to fuse the outcomes of the different factors. The aim is to experiment neural networks as a way to map the multidimensional probabilistic space of a multi-factor biometric framework into the mono-dimensional binary space required by an authentication system.

***Keywords*—** authentication, neural networks, signature, voice.

I. Introduction

MULTI-FACTOR authentication is the most diffused solution to secure identification systems world wide. Examples of this approach are the two-factor verification systems introduced by banks, in order to authenticate users during online transactions [1]. Additionally, relevant ICT providers such as Google [2] are moving their authentication infrastructure towards this model.

In a multi-factor framework, people is authenticated by means of something they know (STK), something they have (STH) and something they are (STA). From a simplified perspective biometry could be considered as a way to obtain information of STA type.

Biometry is generally referred to as the science (and technology) which provides information about someone's identity by means of his/her biological traits. Generally speaking, two different kinds of biometric traits exist: behavioral and physical. Behavioral biometry (BB) tries to assign unique habits to a given subject. Examples of BB are keystroke rhythm and signature pressure and speed. On the other side, physical biometry involves the analysis of anatomic traits such as fingerprints, face or voice (actually voice indirectly represents vocal tract anatomy [3]).

In a previous work from the authors [4], the feasibility of a two factor biometric authentication system has been investigated, where signature has been used as STK, while voice represented the STA component. Indeed, signature recognition is the ability to assign/verify a unique id to a given signature act, analyzed during its realization (we refer here to on-line signatures, that is, those obtained by means of digital tablets). Speaker recognition, on its turn, is the ability to assign/verify a unique id to a given pass-phrase. Moving from the fact that applying a signature is a behavioral trait, this element can be considered as the unique ability a person has to impress a paper (rather that transmitting a set of impulses at a given speed and pressure [5]). On its turn, a pass-phrase can be considered as both something a person knows (the content of the pass-phrase) and something the person is (the frequency response induced by the vocal tract anatomy).

Every multi-factor framework has to deal to the fusion of scores: each factor is analyzed by a different subsystem, whose outcome defines the possibility to grant access to a facility from the perspective of the single subsystem. How to fuse together the different outcomes generating an unique binary decision (go/no go) is defined as the fusion problem.

In a previous work [4] a naïve solution has been applied in order to fuse the biometric outcomes obtaining an unique probabilistic value, later, by means of a training session, a decision threshold was defined, which allowed to map a probability into a binary space. In this paper we want to verify the impact of more complex fusion/mapping rules on the discrimination capability of a biometric multi-factor system.

The reminder of this paper is organized as follow. Section II formulates the authentication problem in terms of biometric traits; Section III presents the different solutions applicable in a fusion/mapping scenario. Section IV deals with the experimental setup and the results obtained by using the proposed framework. Eventually, section V presents the conclusions derived from this paper.

II. Authentication Problem in Biometry

A. Authentication problem

Biometric techniques involve a two stage procedure for a system to be able to associate an ID to a subject. As first, specific parameters, the features, are extracted from a signature, later, a statistical model is enrolled against such features and stored in a specific facility along with an used ID. The model itself, also referred as a template, is used in order to provide a representation of biometric statistical properties, which are expected to contain all the relevant information required to detect a subject among a cohort of people. A second step is used for the *authentication* itself. Given a set of models with an associated ID, a biometric data set is requested during any new transaction as well as a claimed ID. The features extracted from this data set are then compared to the declared ID. In classical approaches, if a template obtains a score major than a fixed threshold θ, it is considered the winner template and the declared ID as the one actually belonging to the subject under analysis.

Manuscript received July 31, 2012. This work was supported by the Tuscany Region POR FSE 2007-2013 Objective 2 Research Plan.

M. Carfagni is with the Dipartimento di Meccanica e Tecnologia Industriale of University of Florence (monica.carfagni@unifi.it).

M. Nunziati is with the Dipartimento di Meccanica e Tecnologia Industriale of University of Florence (matteo.nunziati@unifi.it).

M. Palai is with the Dipartimento di Meccanica e Tecnologia Industriale of University of Florence (matteo.palai@unifi.it).

Authentication is performed by means of *similarity* and *typicality* comparison. Let $F=\{f_i with\ i=1,\dots,n\}$ be the feature set acquired by a biometric device at a fixed sampling frequency, and Θ_0 be a template, a *similarity* score S_0 is defined as:

$$S_0=\frac{1}{n}\sum_{i=1}^{n} P(f_i|\Theta_0) \tag{1}$$

where P is the probability operator. If an alternative template Θ_1 exists, it is possible to estimate a second *typicality* score S_1 and retrieve the normalized log-score (NLS) as:

$$NLS=\log\left(\frac{S_0}{S_1}\right) \tag{2}$$

Usually, Θ_1 is named Universal Background Model (UBM) and it is generated by pooling together feature sets obtained from a reference database. This model is expected to provide a good estimation of the probability that certain features can occur among different people.

In other terms Θ_0 accounts for the similarity between a template and a person's biometric trait, while Θ_1 assesses for the typicality of a certain feature, that is, its frequency among a wide group of people.

The typicality allows a system to weight the actual similarity between a template and a feature set, evaluating how much original a component of a biometric trait is.

B. Feature sets

On-line signature recognition requires the employment of digitizing tablets. Such tools allow to record several temporal patters, such as: the pen position on the tablet (x,y), its pressure (p).

In [5], the authors have reviewed the signature process from a physical perspective. Briefly, the whole act of signature making can be reduced to the motion of a point in space (the pen tip); therefore, the signature can be described by the classical problem of a material point moving in a bi-dimensional space. According to classical equations of mechanics, a material point moving on a generic path can be represented by a dynamic system, where the state is defined by the vector (x, y, δ, $\dot{x}$, $\dot{y}$, $\dot{\delta}$), that is, point's position and instantaneous velocity (being δ the angular velocity), while the input is defined by the acceleration provided to it by external forces: ($\ddot{x}$, $\ddot{y}$, $\ddot{\delta}$). Moving from this model and by adding the pressure information, authors proposed the following feature vector:

$$f'=[x, y, \delta, p, v, \dot{\delta}, \dot{p}, \dot{v}, \ddot{\delta}, \ddot{p}] \tag{3}$$

This feature vector is employed again in this paper, computing a vector for each data set acquired at 100Hz by a digitizing tablet.

One of the most commonly used features in speaker recognition are Mel Frequency Cepstral Coefficients (MFCC) [6]. In order to increase the discrimination capability of a speaker recognition system, such features are associated with their first and second order derivatives in a manner similar to the procedure described for signatures. Moreover the first and second order derivatives of signal energy are included to incorporate user habits related to loudness modulation in voice. MFCC are used to approximate the voice spectrum via discrete cosine transformation. In other terms by varying the number of MFCCs a more or less accurate representation of the speaker voice can be attained. State of art uses from 13 to 19 MFCC. Thus, each feature vector accounts for 41 to 59 parameters. Each feature vector is extracted from a 20 to 30 ms signal window (commonly overlapped Hamming windows are used). In this paper 20 ms Hamming windows are employed, extracting 13 MFCC plus their derivatives and energy derivatives, leading to 41 parameters for each signal window.

Additionally, the average energy value of each window is used in order to discriminate between silence and actual voice in recordings. Windows are grouped by means of a 2 component GMM [7]. The windows belonging to the GMM component with the lower average energy are discarded as silence, while the others are employed for model training.

C. Models

The so named UBM-GMM model is widely employed in this paper for both signature and voice. This kind of model represents a special case of the Maximum A Posteriori (MAP) estimator for HMM parameters, described in [7].

In order to compute a proper template, Gaussian Mixture Models (GMM) are commonly used (compare [8] among others). Given a number h of multivariate Gaussian distributions $N(x,\mu_j,\Sigma_j)$, with x being a vector of iid random variables, a GMM based template is defined as:

$$\Theta=\sum_{j=1}^{k}\alpha_j N(x,\mu_j,\Sigma_j) \qquad with\ \ \alpha_j\in\mathbb{R},\forall\, j\ \ constrained\ to\ \ \sum_{j=1}^{k}\alpha_j=1 \tag{4}$$

where the covariance matrices are commonly constrained to diagonal form and weight coefficients (α_j) are constrained to satisfy:

$$\int_{-\infty}^{\infty}\sum_{j=1}^{h}\alpha_j N(x,\mu_j,\Sigma_j)\,dx=1 \tag{5}$$

In order to properly compute a GMM based template, the unbiased estimators for each mean μ_j and covariance matrix Σ_j as well as the weights α_j must be retrieved. A straightforward solution to this problem is the application of the well known iterative Expectation-Maximization (EM) algorithm.

Anyway, the classical EM algorithm needs a relevant amount of data for its estimates to be accurate enough. As a matter of fact, common biometric traits do not provide such an amount of data and the overall system accuracy is degraded by this lack. By applying MAP estimation to biometric data, authors of [9] have made their system less sensitive to this issue. The procedure, detailed in [9] and [7], can be synthesized as follow: EM is applied to compute an UBM model - which does not suffer of data lack, being generated by pooled data -, then the MAP algorithm is applied in order to derive templates from the specific subject's features.

The MAP algorithm interpolates between the UBM parameters and the template parameters as computed by

directly applying EM to the subject's features. Specifically, the MAP procedure interpolates at each iteration of the EM algorithm. According to terms defined in eq. (3), template parameters are estimated iteratively as:

$$\begin{cases} \mu_j^+ = \mu_{UBM} + D_\mu \mu_j^- \\ \Sigma_j^+ = \Sigma_{UBM} + D_\Sigma \Sigma_j^- \\ \alpha_j^+ = \alpha_{UBM} + D_\alpha \alpha_j^- \end{cases} \quad (6)$$

where *j* accounts for the iterations of the EM algorithm and $D_{(.)}$ are diagonal relevance matrices. Each entry of $D_{(.)}$ defines a weight to be applied in the sum. Possible values for $D_{(.)}$ are proposed in [9] and [10]; namely in [9] an a priori set of weights is employed, while in [10] a more advanced adaptive method is presented. The a priori set is employed also in this paper and it is defined as:

$$\begin{cases} D_\mu = \lambda_\mu I \\ D_\Sigma = \lambda_\Sigma I \\ D_\alpha = \lambda_\alpha I \end{cases} \quad (7)$$

$$where\ \lambda_{(.),k} = n_k / (n_k + r)$$

n_k and r assume the same meaning defined in [9] and are independently computed for each dimension of the feature space. MAP is widely used in speaker recognition and the authors in [5] have successfully proposed the MAP approach for signature recognition tasks. Therefore, the templates used in this paper are based on the same computational model.

D.Fusion and decision

Once a biometric trait is available, a feature set is extracted, and it is compared against the associated biometric templates. In other terms, given a feature stream for the *i*-th trait, and a set of M templates $T_i = \{\Theta_{0,i,j},\ j=1,...,M\}$, the outcome of a classical authentication problem constrained in the *i*-th trait subspace is defined as:

$$auth = \begin{cases} \Theta_{0,i,j} & if\ NLS_{i,j} > \theta \\ \Theta_{1,i} & otherwise \end{cases}$$
$$where \quad NLS_{i,j} = \log\left(\frac{S_{0,i,j}}{S_{1,i}}\right) \quad (8)$$
$$and\ \theta \in \mathbb{R}$$

where $S_{0,i,j}$ is the S_0 score obtained comparing the feature set with the $\Theta_{0,i,j}$ template and θ is the application-dependant acceptance threshold, fixed during the framework enrollment. In a multi-factor system this is not the case; indeed, the winning template is defined by analyzing the outcome of an identity in the whole multi-factor domain:

$$auth = \begin{cases} \Theta_{0,j} & if\ f(NLS_{i,j}) > \theta \\ \Theta_1 & otherwise \end{cases} \quad (9)$$

$$with\ f(NLS_{i,j}) = \mathbb{R}^M \to \mathbb{R}$$

where $f(NLS_{i,j})$ is a function that maps each set of $NLS_{*,j}$ scores obtained by the *j*-th identity in the $\mathbb{R}^M$ factor space, to a probabilistic value in $\mathbb{R}$.

Such a value contains the fusion of all the scores obtained by a given identity, when tested against the different features streams. The focus of this paper is on the definition of such a fusion function *f(.)*.

III. THE FUSION AND MAPPING PROBLEM

A. Factors outcome refinement

State-of-the-art models and features are usually affected by some problems. Numerical models are either too complex to be perfectly calibrated with common amounts of data or too simple to perfectly fit real situations. Moreover, known features can be far from optimal descriptors of a given problem. In other terms, MAP generated templates can still suffer in terms of discriminatory capabilities. The most common solution to this issues is to adopt a fusion stage which is able to merge *and* refine the outcomes of the subsystems. As first a simple single factor system is analyzed in this paragraph, in order to point out how to attain such a refinement capability.

The underlying idea, proposed in [11], is to filter the NLS by means of Artificial Neural Networks (ANN). ANN have been widely used in order to introduce both prediction and automatic classification capabilities in software ([12][13][14] among others). Another way to employ ANN is to use them as a mapping function, which represent the goal in [11] and, partially, in this paper.

According to [11], an ANN can be employed as *f(.)* in order to reduce the average error of a multi-factor system. Authors in [15] have successfully applied this approach to signature identification. The steps required in order to solve the refinement problem are briefly recalled:

1. a set of NLS is obtained during a controlled session, that is a simulated authentication session where the ID are known a priori. During this session NLS are retrieved from comparisons where the winner is either the actual ID of the tested feature stream, or, in case of error, a wrong ID;
2. according to the fact that the winner is or is not the correct ID, a label is assigned to each NLS. Namely label *1* is assigned to correct ID, label *0* to wrong ones;
3. an ANN is trained by minimizing the following cost function:

$$cost = -\frac{1}{N_T}\sum_{i=1}^{N_T} \log_2\left(1 + \frac{1}{\exp(NLS_i)}\right) - \frac{1}{N_I}\sum_{j=1}^{N_I} \log_2\left(1 + \exp(NLS_j)\right) \quad (10)$$

where N_T is the number of correct IDs – computable as the numerosity *#1* of *1* labels-, that is the comparisons made between a data stream and a template belonging to the same subject, while N_I is the number of impostor comparisons (again *#0*), that is the number of wrongly assigned IDs. The NLS are feed into the ANN as input, while the relative labels

are used to correctly locate the network outcomes in to the two members of the right side of eq.(10).

The output of the trained ANN is expected to be a well-calibrated loglikelihood ratio (llr). According to the Bayesian theorem, llr are related to the probability of template matching by:

$$\log\left(\frac{P_{match}}{1-P_{match}}\right)=LLR+\log\left(\frac{P_{prior}}{1-P_{prior}}\right) \tag{11}$$

where P_{prior} is a prior probability, fixed to 0.5 in [11]. The proposed topology for the ANN is *1-1-1*, with one input, a hidden layer with a single hyperbolic tangent neuron and one linear output neuron. The hidden layer is used to saturate the result in order to limit out layers in scores. The presented ANN hereafter will be referred to as ANN-1.

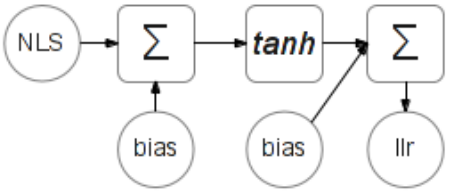

Fig. 1 – topology of the ANN-1 for the refinement problem.

The final result is, therefore, an unbiased predictor of the probability of match between a feature stream and an ID. By substituting the llr to the NLS in equations (7) and (8), an optimal and, hopefully, unbiased decision can be attained.

B. Factors outcome fusion

In the recent past, the fusion stage was resolved by means of *naive fusion*. A naive fusion is obtained by summing all NLS coming from the different factors. As pointed out, among others, by [16], this approach overestimate the real LLR due to unaccounted *correlation* effects between factors. More accurately, each output can be correlated to others by computing their *correlation matrix*. A possible solution is to apply a pre-fusion stage which de-correlates data: as instance, PCA. PCA is a widely applied approach to data decorrelation and dimensionality reduction for machine vision and artificial intelligence ([17][18] among others). Historically, the biometric community has followed a different approach.

In [19], a different fusion approach is presented which is named *logistic regression*. Logistic regression is used extensively in the medical and social sciences as well as marketing applications. The general formulation of logistic regression is:

$$P_{match}=\frac{1}{1+e^{-g(x)}}$$

$$with:$$

$$g(x)=\sum_{i=0}^{n} a_i x_i \quad being \quad x_0=1 \quad by\ definition \tag{12}$$

where x_i are the NLS, a_i are their respective weights. The logistic regressive fusion problem is focused on the definition of correct weights.

In [20] a numerical approach is proposed with the aim to provide a balanced solution. Such solution is responsible to account for all correlation effects, providing the correct weight to each output. According to [20] and [11], equation (11) is convex, thus a gradient descent approach can attain the global minimum of function. Thus the problem is formulated as:

$$\vec{a^{opt}}\equiv(a_{0,}\ldots,a_n)^{opt}=\underset{\vec{a}\in\mathbb{R}^{m+1}}{arg\,min}\, y(\vec{a}) \tag{13}$$

Beside the optimal fusion approach, [11] demonstrated that the obtained solution cost is also close to C_{llr}^{min}. In other terms solving problem (12) leads to simultaneous fusion and refinement of multi-factor systems.

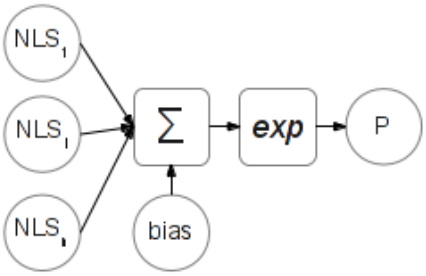

Fig. 2 – topology of the ANN-2 for the fusion problem.

There is a main difference with respect to the refinement topology proposed by the authors: the logistic problem is equivalent to an ANN with topology *n-1*, with a sigmoid output neuron. The exponential term is here used to invert equation (10), where P_{prior} is assumed to be 0.5. In this case the cost function (9) can be filled with the inversion of eq.(11), as the ANN output already provides an estimation for P, therefore the resulting llr must be explicitly recomputed. The presented ANN hereafter will be referred to as ANN-2.

C. Fusion and refinement

It is possible to reorganize the refinement and fusion problems into a common general schema. In order to explain this approach, the two problems have to be reviewed from an operator-based stand point.

1. A linear transformation is applied in the fusion problem. The linear transformation is just the generalized form of the biasing applied as input into the hyperbolic tangent neuron of an ANN-1, as in *A.*. Thus, both problems start with the same operator, here named Σ_1.
2. Σ_1 is mapped by a sigmoid function in both problems. In the former (refinement) this function is applied as basis for the final soft thresholding. In the latter it is used to map llrs to probabilities. In fact, logistic regressive outputs are probabilities and not llrs. For

the refinement problem this operator is here named Ψ, while it is named M for equation (11).

According to the previous description, both problems can be respectively reduced to the following operator chains: Σ_1-M for fusion and Σ_1-Ψ for refinement. The main issue with logistic regression is the missing saturation stage Ψ.

As a possible solution, we propose to plug the fusion stage into the refinement problem, removing the implicit llr-to-P map M. In this paper we will experiment with an ANN topology, which resembles the one of eq. (12), but where the input is the multidimensional vector $NLS_{i,j}$, and the output is a well-calibrated llr_j, for each subject (fig.3). The presented ANN hereafter will be referred to as ANN-3.

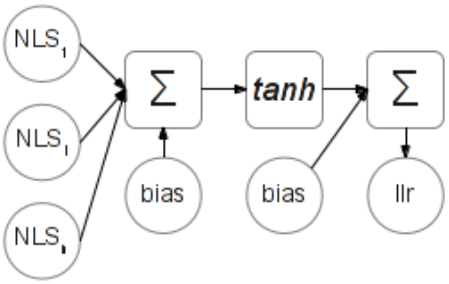

Fig. 3 – topology of the ANN-3 for refinement and fusion.

D. Map from multidimensional probabilistic space to mono-dimensional binary space

The main aim of a multi-factor authentication system is to discriminate between genuine accesses and impostors, that is provide a binary decision to the authentication problem.

The biometric framework, paired with an ANN-3, is a probabilistic system, which do not provide any real solution to the authentication problem; therefore, a post processing stage has to be designed in order to move into the binary space. The goal is to allow an ANN to automatically setup a decision threshold, which will map the biometric framework outputs in to the binary domain, filling the gap between biometry and authentication frameworks requirements.

The underlying logic is as follow: as first, llr coming out from ANN-3 are remapped in probabilities as per eq.(12), later, a new neuron is employed to define a probability $P_{1/0} \in [0,1]$, that is, the specific probability threshold for which outcomes of eq.(12) are either saturated to 1 (true, accept identity) or floored to 0 (false, reject identity). In other terms, we propose to rework the neural network stack as follow:

- the output of ANN-3 is feed in a non-biased sigmoid neuron. This neuron is responsible for mapping llr to P as per eq.(12);
- the output of the sigmoid neuron is feed into a step neuron, which is responsible to set a final hard threshold $P_{1/0} \in [0,1]$.

The final ANN stack has topology n-1-1/1-1-1, and is composed by two networks: ANN-3 and the newly proposed 2-neuron network, hereafter referred to as ANN-4.

In this case, the training is composed by two serial steps:

- ANN-3 is trained by minimizing eq.(10);
- the trained output of ANN-3 is used to enroll ANN-4 by means of any one of the classical cost function ([21] among others).

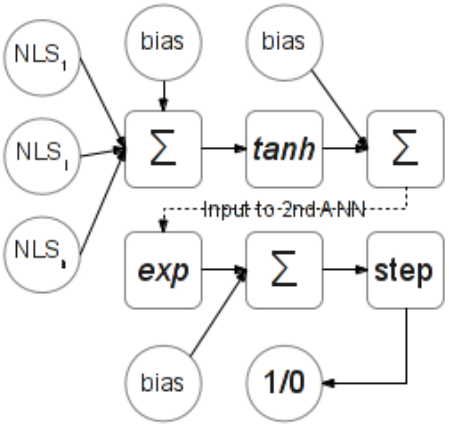

Fig. 4 – topology of the ANN stack for the multidimensional mapping.

IV. Experimental setup and results

A. Data-sets and models

In order to test our hypotheses, the myIDea, the Speaker Recognition and the 22-Language databases have been employed [22][23][24]. The myIDea signature data set is composed by 3537 signatures collected from 73 different subjects. Each subject has been acquired in different sessions, collecting up to five genuine signatures per session. 1173 of these signatures have been employed to train the UBM model. Other samples have been used to generate the attempted accesses to a given facility.

The Speaker Recognition (v.1.1) corpus consists of telephone speech from about 500 participants [23]. Each participant has recorded speech in twelve sessions over a two-year period. All of the data in the corpus were collected over digital telephone lines (8khz 8-bit ulaw files). Some of these speakers have been employed in order to simulate the voice of people attempting to access a biometrically controlled facility.

The 22-Language database (v.1.2) consists of telephone speech from 22 languages (Eastern Arabic, Italian, Russian, English among others), containing fluent continuous speech [24]. All of the data in this corpus were collected over digital telephone lines (8khz 8-bit ulaw files) too. The voices in the 22-Language database have been employed to train an UBM for the speech factor.

The experiment proposed in this paper has been conducted by using chimeric data. A biometric chimera is a set of biodata obtained by grouping biometric traits belonging to different people. In such a way, even if a multi-modal database doesn't exist, it is possible to generate fake authentication sessions. Researchers have warned [25] about the overestimation of discrimination capabilities induced by the usage of chimeric data. The authors are aware of this issue, nonetheless we consider that the provided results still provide insightful information about the strength obtainable by merging different biometric traits.

The chimeric data have been obtained by associating a given identity $ID_{s,i}$ of the signature database with one $ID_{v,j}$ coming from the Speaker Recognition database. Given a pair

(i,j) of identities randomly extracted from the two databases, a new chimera $C_k=\{ID_{s,i}, ID_{v,j}\}$ is introduced into the ID list.

The proposed identification experiment involves the usage of 38 chimeras. 19 of them are used to train the ANN stack described in *III.D*, while the others are used for validation, that is, to simulate the effective performance of the system during its runtime. The chimeras have been generated by randomly selecting a number of ID without repetition, in such a way it is granted that a given random $ID_{(.),(.)}$ is not present in two or more chimeras.

During training, a template $\Theta_c=\{\Theta_s,\Theta_v\}$ composed by two GMM (one for signature and one for voice) has been retrieved from one session of each chimera. Later, 8 additional sessions per user have been employed for training the ANN.

By cross comparing the different sessions against each other, a total of 152 target accesses have been computed, as well as 2682 fraudulent accesses. A target comparison is a comparison where the *h*-th chimera C_h claims an identity C_h, while a fraudulent access is one where C_h claims an identity C_t with $t\neq h$. These comparisons have been performed in order to evaluate both system resistance to attackers and system discriminatory capabilities.

The comparisons have been carried on independently for the two biometric factors, by using eq.(2); the result is a set of 2834 bi-dimensional vectors $\vec{NLS}_i$, with *i* defining the *i*-th attempted access. As described in *III.A*, these vectors have been assigned a label of 1 or 0 according to the rule:

$$label=\begin{cases} 1 \;\; if\; C_h\, claimed\, identity\, C_h \\ 0 \;\; if\; C_h\, claimed\, identity\, C_t, with\, t\neq h \end{cases} \tag{14}$$

the ANN training has been conducted via the well known stochastic gradient descent algorithm [26].

Validation has been obtained by processing the other half of the chimeras. Namely, each factor has been processed independently in order to retrieve new $\vec{NLS}$ vectors, then those vectors have been processed by the trained ANN, obtaining a label (0/1).

B. System training and validation

In order to generate the $\vec{NLS}$ vectors, a template for each chimera must be enrolled. As explained in section *II*, template enrollment is performed via MAP, which, on its turn, requires a given base model, commonly the UBM. As a consequence, a necessary condition to boostrap the process is to define a correct UBM for each factor. The present work requires two UBM: one for speech and one for signatures.

UBM are obtained via the EM algorithm by training a GMM against a set of pooled features, coming respectively from the myIDea database and from the 22-Language database. The only mandatory input required by the EM algorithm is the number of components of the GMM, that is, the model size. In [27] has been shown how model size predictors such as the Minimum Description Length do not apply to diagonally constrained GMM (the kind of GMM used in this paper) and, therefore an iterative a posteriori sizing procedure is required. Usually, this implies the following steps:

1. define a tentative size for the UBM and enroll it via EM;
2. use the tentative UBM to enroll all the templates with MAP, given an adaptation strategy;
3. perform a batch of comparisons, with both target and fraudulent accesses;
4. define an optimal acceptance threshold θ, for which the equal error rate (EER) configuration is attained, for the given UBM size [27];
5. repeat steps from 1 to 4 with different sizes, until the obtained EER is minimized.

Usually progressively increasing or decreasing powers of 2 are used as tentative sizes. In this paragraph we report the different levels of equal error rates reached by different sizing for both the signature and voice UBM. Referring to eq.(6), the MAP adaptation strategy employed in this paper is the following for both speech and signature:

$$\begin{cases} D_\mu=\lambda_\mu I \\ D_\Sigma=0 \\ D_\alpha=0 \end{cases} \tag{15}$$

$$where\, \lambda_{\mu,k}=n_k/(n_k+16)$$

In other terms only UBM means are adapted by using a relevance factor of 16. This very common strategy can be found in a number of papers focusing on MAP enrollment, noticeably it is used in [9].

Tab. 1- most relevant sizing results for the signatures UBM

System	size	EER [%]
signatures	512	5%
signatures	1024	1%
signatures	2048	2.5%

Tab. 2 - most relevant sizing results for the voice UBM

System	size	EER [%]
voice	256	12%
voice	512	11%
voice	1024	16%

According to Tab.1 and 2, which report the most relevant results of the sizing procedure, the optimal size for the signature UBM is 1024, with an EER of 1%, while the optimal sizing of the voice UBM is here fixed at 512, with an EER of 11%. The higher EER in the voice system with respect to the signature one is caused by the highest variability of the voice features, which generate a more complex scenario for any biometric technology.

Once the optimal UBM size is defined, along with the associated models, a template is trained for each subject, which is composed by two GMM: one MAP enrolled from the voice UBM and the other one MAP enrolled from the signature UBM.

The given models are tested against feature stream coming from each session of each chimera, generating both target and fraudulent accesses. The result of these comparisons is plotted in fig.(5).

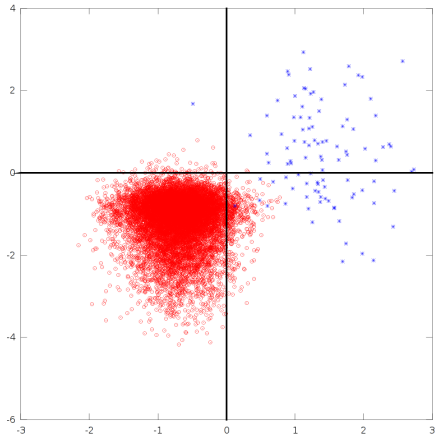

Fig. 5 – NLS vectors coming from signature (x axis) and speech (y axis) comparisons. Blu dots represent target accesses while red ones represent fraudulent accesses.

The horizontal and vertical zero crossing lines of fig.5 define the boundaries of log-likelihood values (llr). A negative llr implies a mismatch between the claimed ID and the feature stream, on the opposite, a positive llr confirms a match between the ID and the features. Analyzing raw NLS vectors, where ANN-1 is not applied yet, the most of the comparisons show a coherent behavior between the two factors: almost all the fraudulent trials (red dots) belong to the negative domain $\mathbb{R}^- x \mathbb{R}^-$, while the majority of target trials (blu dots) lies in $\mathbb{R}^+ x \mathbb{R}^+$. In some cases the two factors provide inconsistent behavior, generating NLS vectors in the second and fourth quadrant. The highest percentage of inconsistencies is noticeable in the fourth quadrant, where a lot of voice comparisons (y axis) show a mismatch where a match should be proved instead. This is in line with the predicted higher EER of the voice factor compared to the signature one. By training ANN-3, a new set of values is computed, which should be considered a set of well-calibrated llr: results are shown in fig. 6.

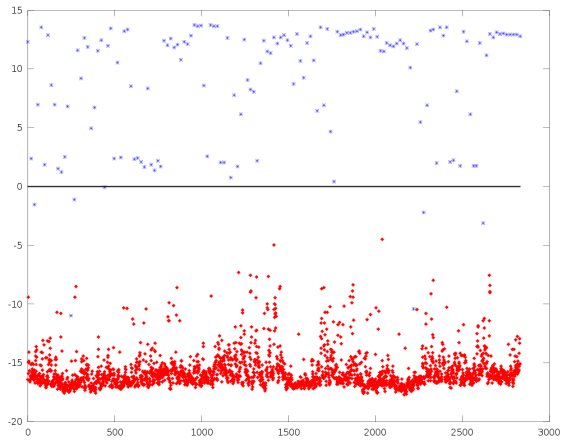

Fig. 6 – well-calibrated llr obtained with ANN-1 during the training session. The outcomes have been obtained by feeding raw NLS vectors of fig.5 into an network, whose topology resembles the one described in section *III.C*.

As shown in fig.6, a good separation between fraudulent and target trials is attained by means of ANN-3, leading to a limited amount of errors. Moreover the effect of the hyperbolic tangent and the linear out layer are depicted, showing a relevant number of trail clustered in proximity of the boundaries automatically defined by ANN-1 (-17 and 14).
ANN-3 results are a good predictor of what can be expected training the actual ANN stack required by the framework: indeed, ANN-4 simply applies the Bayesian theorem and defines the optimal separation threshold required to saturate the outputs to 0 and 1.
First row of table 3 shows the training errors obtained by training the actual network stack of section *III.D* and mapping llr into binary values.
Analogously, by re-applying both template generation procedure and the cross comparison to the validation set and by processing it by means of the trained stack, the errors shown in the second row of table 3 are attained. This last row defines the actual reliability of the proposed system as meant during its usage in a *real scenario*.

Tab. 3 – training and validation errors for the proposed framework

Phase	**False rejection [%]**	**False acceptance [%]**
training	0,25%	0%
validation	0,35%	2%

V. Conclusion

The following paper deals with features and models for a biometry-based multi-factor authentication framework. The proposed solution approaches biometry both as something users know (signature make process) and something they are (voice as correlate of vocal tract anatomy) and approaches probabilistic biometric models as tools for strong authentication. The underlying idea is to enforce authentication without the usage passwords and/or other authentication gadgets, while the aim is to experiment neural networks as a way to map the multidimensional probabilistic space generated by the multi-factor biometric system into the mono-dimensional binary space required by the authentication process.

An ANN stack has been proposed, which is composed by two stages: the first one attempts to correctly fuse the probabilistic information coming from the two subsystems (signature and voice), mapping it into a set of well-calibrated log-likelihoods (llr), while the second maps llr into probabilities (according to he Bayesian theorem) and saturates them in order to generate a binary outcome. By using this stack it is possible to solve the map problem and obtain an automatic authentication system from biometric data. The presented experiment shows a potential error of 0.35% in terms of false rejection and a potential error of 2% in false acceptance.

One relevant aspect which deserves attention and further investigation is the usage of chimeric data: as pointed out in [25] they tend to underestimate actual errors. Therefore additional commitment must be devoted in retesting the framework with real multi-modal data.

REFERENCES

[1] Federal Financial Institutions Examination Council, Guidance on Authentication in Internet Banking Environment, "http://www.ffiec.gov/pdf/authentication_guidance.pdf", 2005.

[2] S. Nishit Shah, "http://googleblog.blogspot.com/2011/02/advanced-sign-in-security-for-your.html", 2011.

[3] J. P. jr. Campbell, "Speaker recognition: a tutorial", *proceedings of the IEEE*, vol. 85, n. 9, pp. 1437-1462 , 1997

[4] M. Carfagni, M. Nunziati, M. Palai, "Two factor biometric framework for user authentication", *Recent Researches in Communications and Computers*, vol.1, pp. 205-208, 2012.

[5] M. Carfagni, M. Nunziati, "An Improved Model and Feature Set for Signature Recognition", *Proceedings of the International Conference on COMPUTERS and COMPUTING*, vol .1, pp.75-80, 2011.

[6] T Shurer, "An experimental comparison of different feature extraction and classification methods for telephone speech", *2nd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications*, pp.93-96, 1994.

[7] J. L. Gauvain, C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", *IEEE Trans. Speech Audio Process.*, Vol.2, pp. 291–298, 1994.

[8] J. Richiardi, A. Drygajlo, "Gaussian Mixture Models for On-line Signature Verification", *Proceedings of the 2003 ACM SIGMM workshop on Biometrics*, 2003.

[9] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing*, Vol.10, pp. 19–41, 2000.

[10] P. Kenny, G. Boulianne, P. Dumouchel, "Eigenvoice modeling with sparse training data", *IEEE Trans. Speech Audio Processing*, Vol. 13, No. 3, 2005.

[11] Brummer N., du Preez J., "Application-independent evaluation of speaker detection", C*omputer speech and language*, vol. 20, pp.230-275, 2006.

[12] R. Furferi, L. Governi, Y. Volpe, "ANN-based method for olive Ripening Index automatic prediction", *Journal of Food Engineering*, vol.101, n. 3, pp. 318-328, 2010.

[13] R. Furferi, L. Governi, Y. Volpe, "Neural network based classification of car seat fabrics", *International Journal of Mathematical Models and Methods in Applied Sciences*, vol.5, n.3, pp. 696-703, 201.

[14] R. Furferi, L. Governi, "The recycling of wool clothes: An artificial neural network colour classification tool", *International Journal of Advanced Manufacturing Technology*, vol.37, n.7-8, pp. 722-731, 2008.

[15] M. Carfagni, L. Governi, M. Nunziati, "Offline continuous adaptation of templates for signature identification", *International Journal of Mathematical Models and Methods in Applied Sciences*, Vol 5. n. 6, pp.1003-1010, 2011.

[16] C. Aitken, D. Lucy, "Evaluation of trace evidence in the form of multivariate data ", *Applied Statistics*, vol 53, n.1, pp.109-122, 2004.

[17] M. Carfagni, R. Furferi, L. Governi, M. Palai, Y. Volpe, "3D reconstruction problem: An automated procedure", *Applications of Mathematics and Computer Engineering - American Conference on Applied Mathematics, AMERICAN-MATH'11, 5th WSEAS International Conference on Computer Engineering and Applications, CEA'11*, pp. 99-104, 2011.

[18] R. Furferi., L. Governi, M. Palai, Y. Volpe, "From unordered point cloud to weighted B-spline - A novel PCA-based method", *Applications of Mathematics and Computer Engineering - American Conference on Applied Mathematics, AMERICAN-MATH'11, 5th WSEAS International Conference on Computer Engineering and Applications, CEA'11*, pp. 146-151, 2011.

[19] S. Pigeon, P. Druyts, P. Verlinde P., "Applying Logistic Regression to the Fusion of the NIST'99 1-Speaker Submissions ", *Digital Signal Processing*, vol. 10, pp. 237–248, 2000.

[20] TP. Minka, "A comparison of numerical optimizers for logistic regression", available on line at *http://research.microsoft.com/~minka/*, 2007.

[21] D.Wackerly, W. Scheaffer, "Mathematical Statistics with Applications", Thomson Higher Education, 2008.

[22] B. Dumas, C. Pugin, J. Hennebert, D. Petrovska-Delacrétaz, A. Humm, F. Evéquoz, R. Ingold, D. Von Rotz, "MyIdea - Multimodal Biometrics Database, Description of Acquisition Protocols"**,** *proc. of Third COST 275 Workshop*, pp. 59-62, 2005.

[23] R. Cole , M. Noel , V. Noel, "The CSLU Speaker Recognition Corpus"*, Proceedings of the International Conference on Spoken Language Processing*, 1998.

[24] T. Lander, R. A. Cole, B. T. Oshika and M. Noel, "The OGI 22 language telephone speech corpus", *Proc. of the Fourth European Conference on Speech Communication and Technology*, 1995.

[25] N. Poh and S. Bengio, "Can Chimeric Persons Be Used in Multimodal Biometric Authentication Experiments?", *Machine learning for multimodal interaction*, Springer, pp.87-100, 2006.

[26] TS Ferguson, "An inconsistent maximum likelihood estimate". *J. Am. Stat. Assoc.,* vol. 77, n. 380, 831–834, 1982.

[27] J. Richiardi, A. Drygajlo, "Gaussian Mixture Models for On-line Signature Verification", *Proceedings of the 2003 ACM SIGMM workshop on Biometrics*, 2003.