MULTIMODAL BIOMETRIC SCORE FUSION: THE MEAN RULE VS. SUPPORT VECTOR CLASSIFIERS

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

DEPARTEMENT ELECTRONIQUE ET PHYSIQUE

INSTITUT NATIONAL DES TELECOMMUNICATIONS

9 RUE CHARLES FOURIER, 91011 EVRY FRANCE

Telephones: (33-1) 60.76.44.30 , (33-1) 60.76.46.73

Fax: (33-1) 60.76.42.84

{ Sonia.Salicetti, Mohamed.Anouar_mellakh, Lorene.Allano, Bernadette.Dorizzi}@int-evry.fr

ABSTRACT

Recently, a discrepancy in results has appeared in the literature concerning score fusion methods, classified in "combination methods" and "classification methods" [1]. Some works suggest that a simple Arithmetic Mean Rule (AMR) can outperform some training-based methods on multimodal data [2], while others favour, among other trained classifiers, a Support Vector Machine [3]. This paper makes a comparative study of the Arithmetic Mean Rule (AMR) coupled with different state-of-the-art normalization techniques [4, 5] and a linear Support Vector Machine (SVM), in the framework of voice and on-line signature scores fusion. Two experiments differing in the difficulty to discriminate genuine from impostor accesses are carried out on the BIOMET database [6].

1. INTRODUCTION

Multibiometrics, that is the verification of the identity of a person by more than one biometric trait, is nowadays a promising research area, that has generated expectations as an alternative in solving, among other problems, performance requirements in real applications. In general, it is not easy to combine biometric traits at the feature level, which explains why so often score fusion is performed. In this framework, many fusion techniques have so far been compared; they can be classified as score combination rules opposed to statistical learning techniques. On one hand, recent works in multimodal biometrics show experiments in which a simple mean rule performs better than two other classification schemes, a decision tree and a Linear Discriminant Analysis (LDA) coupled with a minimum distance rule [2]. On the other hand, some other works in the literature have compared the mean or a weighted mean rule to a Support Vector Machine: one in the framework of fusion (fingerprint with voice) based on signal quality for a mobile application [3]; another in different conditions of noise in the speech signal for audio-visual fusion [7]. Both works report that the trained classifier outperforms the mean or weighted mean rule [3,7].

Our aim in this work is to study further this apparent contradiction in the literature, and also to open the discussion about the limits and advantages of both methods. Indeed, the Arithmetic Mean Rule (AMR) certainly requires no training, but at the price of a previous normalization of scores. Such normalization is not straightforward; indeed, according to which normalization method is used, results can vary a lot. Many normalization schemes have so far been studied [4,5]; they rely on the genuine and impostor distributions, that is on a statistics on the used database. Among the most efficient, the so-called "adaptive normalizations", aim at reducing the overlap between the genuine and impostor distributions. To this aim, normalizing output scores of different biometric systems requires that part of the data is devoted to tuning normalization, and the rest to testing the method, exactly as training-based methods. This is in general not mentioned in the literature.

Our statement is that the crucial factor is the relative position of the impostor and genuine distributions, sometimes leading to simple configurations in which genuine and impostor data are easy to discriminate, or to more complex configurations with an important overlap of both distributions. In the former case, the AMR with a "standard" normalization, that is a transformation performing a rescaling of each expert's scores to a given interval, may be sufficient. In the complex configuration, a statistically trained classifier as a SVM performs much better; besides, it requires no normalization of scores. Also, in the same conditions, the AMR with score normalization via a posteriori probabilities, that is a normalization based namely on the distributions of client and impostor scores, performs as well as the SVM.

It is clear that the more modalities are fused, the more it will be easy to discriminate impostors from clients; but using a lot of modalities is a burden in an applicative context. Thus, a multimodal system should be able to face complex situations, which is the case in the presence of noise. This aspect will actually become crucial in mobile applications.

The former analysis explains the methodology that we have chosen in this paper: we perform fusion of voice and on-line signature data from the BIOMET database [6]; then we introduce noise in the speech data to generate two configurations of the clients/impostors bimodal distributions, a simple one and a more complex one. We compare in both cases the AMR and a SVM. We demonstrate the superiority of the SVM and of the AMR scheme with a posteriori probabilities normalization in the complex configuration.

2. FUSION OF ON-LINE SIGNATURE AND VOICE

The bimodal fusion system is composed of a signature verification system described in [8] and a text-independent Speaker Verification system described in [9]. We briefly introduce in the following such systems and two fusion methods: AMR and SVM.

2.1 The Signature Verification System

Each writer's signature is modelled by a continuous left-to-right HMM. 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.

2.2 The Text-independent Speaker Verification System

This system is detailed in [9]. Considering a simple hypothesis test between two hypotheses H_λ (X has been uttered by λ) and $H_{\lambda*}$ (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 $H_{\lambda*}$ given X. A single speaker-independent model is used to represent $P_{\lambda*}(X)$. This model, also called Universal Background Model (UBM) [10], corresponds to a 256 components Gaussian Mixture Model (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 Methods: AMR and SVM

2.3.1 Fusion by AMR with associated normalizations

We combine the two monomodal scores by means of a simple AMR after performing a normalization of these scores. We studied 3 types of normalization: the first one is based on the Min-Max normalization [4], the second one is referred in the state of the art as "Tanh estimator" [4], and has given the best results in previous experiments reported in [5], and the last uses a posteriori class probabilities.

We define the "Min-Max" normalization of score s of one monomodal expert as $n=(s-m)/(M-m)$ where M is the maximum and m is the minimum of all scores. We consider the mean (μ) and standard deviations (σ) of both the client and impostor distributions in the training database, and set: $m=\mu_{imp}-2\sigma_{imp}$ and $M=\mu_{cl}+2\sigma_{cl}$. Indeed, assuming that genuine and impostor scores follow Gaussian distributions, 95% of the values lie in the $[\mu-2\sigma, \mu+2\sigma]$ interval; following this model, our choice of m and M permits to cover most of the scores. The values higher than M or lower than m are thresholded. This linear normalization maps the score in the [0,1] interval.

The "Tanh Estimator" normalization of score s is given by: $n=0.5[\tanh(0.01*(s-\mu)/\sigma)+1]$, where μ and σ are chosen as μ_{cl} and σ_{cl} since it gives the best results, as reported in [5].

Finally, the last normalization (called Bayes normalization in the following) uses the a posteriori client class probability $P(C/s)$ given score s, as a normalized score. The estimation of the a posteriori probabilities is done via the Bayes rule:

$$P(C/s) = \frac{P(s/C)P(C)}{P(s/I)P(I) + P(s/C)P(C)},$$

where $P(C)$ et $P(I)$ are respectively the relative frequencies of the client and impostor classes, and $P(s/C)$ and $P(s/I)$ are the conditional probability densities. We estimate the conditional probability densities by assuming Gaussian distributed scores, and using the empirical means (μ) and standard deviations (σ) of both the client and impostor distributions computed on the training database. Assuming independency between the two scores s_1 and s_2, and following [13], we compute the arithmetic mean of $P(C/s_1)$ and $P(C/s_2)$.

2.3.3 Fusion by a Support Vector classifier

In a few words, a SVM [11] looks for the optimal hyperplane in the sense of the minimum of the Total Error Rate (TER) in a high dimensional space. Nevertheless, in this work, we focus on a linear SVM simply to compare the AMR in which the weight given to each score is fixed, to a method in which such weights are learned by a statistical technique. Therefore, the decision surface obtained is a straight line in our initial bidimensional scores space.

The optimization of the SVM was carried out on a specific database dedicated to training. In order to generate a DET curve [12] during the test phase, the position of the optimal hyperplane is varied. This corresponds indeed to the variation of a decision threshold.

3. EXPERIMENTAL SETUP

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 [6].

Signature data was captured on a digitizer at a rate of 100 samples per second. Each sample contains 5 features: 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 genuines and 12 forgeries, made by four different impostors.

Speech data was recorded in quiet environment. 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 [9].

3.3 Building the Bimodal database of real subjects

To build the bimodal database, we associate the scores of the two experts (Signature and Voice). We consider for the voice expert two configurations: one without noise ("database1"), and another with 0dB noise ("database2"). The noise introduced is Gaussian additive white noise, added to the raw speech signal. To obtain a Signal-to-Noise Ratio of 0dB, the power of the raw signal is computed excluding silences.
Figure 1 below shows the effect of noise on client and impostor scores' distributions.

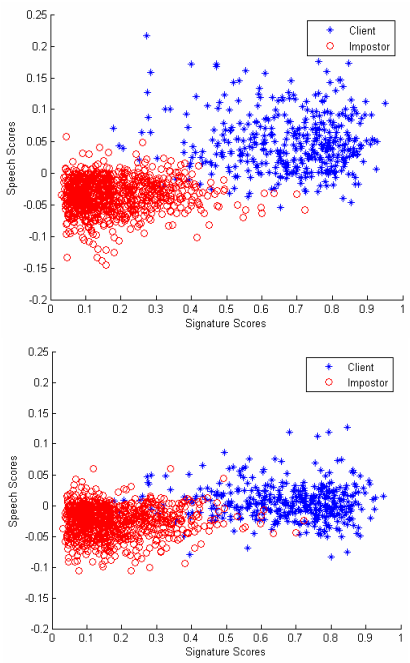

Fig. 1. Bimodal scores' distribution on scores without (up) and with noise (down).

Clearly, the overlap between client and impostor scores is much more important in the noisy case, the classification task will therefore be more difficult.
This bimodal database is then split into a Fusion Learning Base (*FLB*) and a Fusion Test Base (*FTB*) according to a 5-fold Cross-Validation. In *k*-fold Cross-Validation, the data is divided into *k* subsets of approximately equal size. Then, *(k-1)* subsets are used for training and the remaining one for test. This is repeated *k* times while changing the training and test subsets, in order that every subset has been devoted to test once. The Fusion Learning Base is devoted either to normalization in the case of fusion by AMR, and to training the Support Vector Classifier. In order to reduce the bias related to the small number of persons in the database, we consider 10 different samplings of the 5 initial subsets, and compute average error rates.

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).

4. EXPERIMENTS

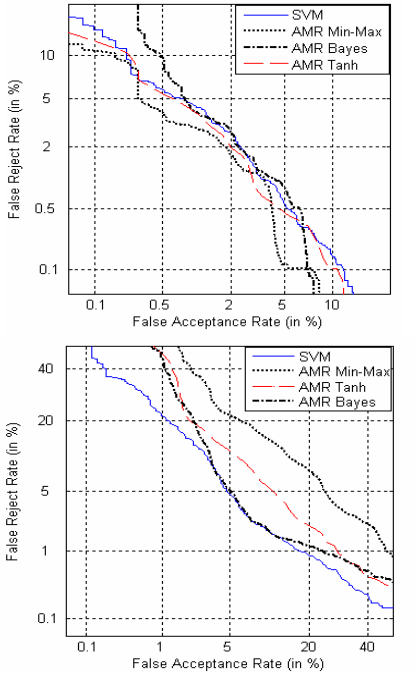

Fig. 2. DET curves of AMR with "Min-Max", "Tanh Estimator" and Bayes normalizations and SVM fusion paradigms on clean (up) and noisy data (down).

The AMR and SVM fusion paradigms were compared following the protocol given in section 3.3. The linear SVM receives as input the on-line signature score that belongs to the [0,1] interval, and a speech score that is not normalized, as shown in Figure 1. Indeed, the signature expert already gives as output a score in [0,1], but such normalization does not depend on the client and impostor distributions of all signature scores of the database, it is in fact a personalized normalization depending only on each client's scores [8]. Figure 2 shows the DET curves [12] of the two fusion paradigms in the simple configuration (clean data), and in the complex one (noisy data). We first notice that in the simple configuration (up), for values of the threshold close to the one corresponding to the Equal Error Rate, AMR with "Min-Max" normalization gives the best results (1.88%), followed by "Tanh Estimator" (1.99%), while SVM and AMR with Bayes normalization give equivalent performance (2.16% and 2.22%). In the case of the SVM, an optimal decision surface is searched on a small training set; we observe in fact an overfitting. For more insight, we performed an experience with training and testing on the whole database1. The result is shown in Table 1; the SVM gives the best result at the Total Error Rate point as expected, but two other methods ("Min-Max" and "Tanh Estimator") reach the same performance at this point.

	SVM	Min-Max	Tanh	Bayes
EER	1.96%	**1.73%**	2.08%	2.02%
TER	**1.41%**	**1.41%**	**1.41%**	1.64%

Table 1. Results on the whole clean database at the Equal Error Rate (EER) and Total Error Rate (TER) points.

Figure 2 (down) shows results on database2: the SVM and AMR with Bayes normalization perform much better than AMR with "Min-Max" and "Tanh Estimator" normalizations, and for most values of the threshold. Indeed, Bayes normalization is directly based on the client and impostor distributions of each expert's scores. It is not only a rescaling of each expert's score as it is the case for the "Tanh Estimator" and "Min-Max" normalizations. We also notice that the results obtained by AMR after Bayes normalization are equivalent to those given by the SVM. Both methods take into account the client and impostor distributions: for the AMR, it is the case on the distributions of each expert's scores separately, while the SVM considers bimodal scores distributions. Nevertheless, in the present case, we have not seen any improvement in using the SVM instead of AMR after Bayes normalization of scores. This can be explained by the fact that we consider uncorrelated modalities. Future prospective work on correlated modalities would be interesting; in this framework, one can indeed expect a significant improvement of results when using the SVM. On the other hand, our results also seem to assess the Gaussian assumption for class conditional densities.

5. CONCLUSIONS

In this work, a compared analysis of two fusion paradigms, the AMR and a Support Vector classifier, is carried out. Two different experimental conditions are considered. The first one corresponds to a situation where it is quite easy to discriminate genuine from impostor accesses. It results from the use of two modalities, namely clean speech and on-line signatures. The second is a difficult configuration, showing an important overlap of the genuine and impostor distributions, obtained by adding noise to the speech data. Experiments involve three normalizations for fusion by the AMR, and no normalization for the SVM. Results show that in the non noisy case, the AMR with a normalization performing only a rescaling of each expert's scores to a given interval as "Min-Max", gives the best results. In the noisy case, the SVM gives equivalent results to those obtained with AMR after scores' normalization via a posteriori class probabilities; such results are much better than those given by the AMR paradigm with "Min-Max" and "Tanh Estimator" normalizations, for most values of the threshold. We conclude that, in noisy conditions, only the methods that take into account the scores' distributions are the more efficient.

Acknowledgements

The authors thank Gérard Chollet and his team for providing speech verification scores, and the referees for their comments which helped to improve the quality of the paper.

This work was partially funded by IST-2002-506883 Project "SecurePhone" (Secure Contracts Signed by Mobile Phone), and the IST-FP6 project BioSecure.

REFERENCES

[1] A.K. Jain, A. Ross, "Multibiometric Systems", Communications of the ACM, vol. 47, N° 1, pp. 34–40, Jan. 2004.

[2] A. Ross, A.K. Jain, "Information Fusion in Biometrics", Pattern Recognition Letters, vol. 24, N° 1, pp. 2115–2125, 2003.

[3] J. Bigun, J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, "Multimodal Biometric Authentication using Quality Signals in Mobile Communications", in Proc. ICIAP 2003.

[4] M. Indovina, U. Uludag, R. Snelick, A. Mink, A. Jain, "Multimodal Biometric Authentication Methods : A COTS Approach", in Proc. MMUA 2003, Santa Barbara, California, USA, Dec. 2003, pp. 99-106.

[5] A. Jain, K. Nandakumar, A. Ross, "Score Normalization in Multimodal Biometric Systems", to appear in Pattern Recognition, 2005.

[6] 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", in Proc. AVBPA 2003, Guildford, UK, July 2003, pp. 845-853.

[7] C. Anderson, K.K. Paliwal, "Information Fusion and Person Verification Using Speech & Face Information", IDIAP Research Report 02-33, September 2002.

[8] B. Ly Van, S. Garcia-Salicetti, B. Dorizzi, "Fusion of HMM's Likelihood and Viterbi Path for On-line Signature Verification", in Proc. BioAW 2004, Prague, Czech Republic, May 2004, Lecture Notes in Computer Science (LNCS) 3087, pp. 318-331.

[9] 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", in Proc. MMUA 2003, Santa Barbara, California, USA, Dec. 2003, pp. 13-18.

[10] 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.

[11] V. Vapnik, The Nature of Statistical Learning Theory: Statistics for Engineering and Information Science, Second Edition, Springer, 1999.

[12] A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, "The DET Curve in Assessment of Detection Task Performance", in Proc. EUROSPEECH'97, Vol. 4, pp. 1895-1898, Rhodes Greece, 1997.

[13] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, "On Combining Classifiers", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, N°3, pp. 226-239, March 1998.