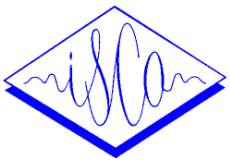

ESCA Workshop on Interactive
Dialogue in Multi-Modal Systems (IDS-99)
Kloster Irsee, Germany
June 22-25, 1999

About Multi-Modal Identity Verification in Interactive Dialogue Systems

Patrick Verlinde *Gérard Chollet*[*] *Marc Acheroy*

Royal Military Academy, Signal and Image Centre
Renaissancelaan 30, B-1000 Brussels, Belgium

patrick.verlinde@tele.rma.ac.be, marc.acheroy@elec.rma.ac.be
{verlinde,chollet}@tsi.enst.fr

ABSTRACT

The aim of this paper is to introduce multi-modal identity verification techniques to the Interactive Dialogue Systems society, and to identify possible applications. The multi-modal identity verification system presented here is based on two biometric modalities (speech and vision) and it uses 3 experts (using voice, frontal face and profile images), based on these two modalities, in parallel. Each expert delivers as output a scalar number, called score, stating how well the claimed identity is verified. A fusion module receiving as input the 3 scores has to take a binary decision: *accept* or *reject* identity. We have solved this fusion problem using a wide range of statistical pattern recognition techniques. The performances of the different fusion modules have been evaluated and compared on a multi-modal database, containing both vocal and visual modalities.

Keywords: *Multi-modal identity verification, biometrics, decision fusion*.

1 INTRODUCTION

One of the consequences of the booming of the *Information Society* is that accessing a computer, and from there on computer networks with their distributed knowledge, has become trivial, even for a layman. However, there are still a number of *sensitive* interactive dialogue applications between a human and a computer, for which one would like to ensure that the person using the dialogue system is really the one he or she claims to be. This can be done using an automatic *verification*[1] system. This verification requirement needs to be met, not only at the beginning of an interactive dialogue session, but throughout the entire duration of the interactive dialogue. If the dialogue system under consideration is a multi-modal system, then it is only logical to take advantage of this multi-modality to improve the performances of this *dynamic* verification process.

In this paper, the term *modality* is reserved for a biometric measure, such as voice, face (including profile), eye (iris-pattern, retina-scan), fingerprint, hand-shape or some other measurable physiological or behavioral characteristic information of the person to be verified. An *expert* is each algorithm or method using characteristic features coming from a particular biometric measure to verify the identity of a person under test.

If one uses only a single biometric measure, the results obtained may be found to be insufficient. This is due to the fact that some (user-friendly) biometric measures tend to *vary with time*. This especially is true for the vocal (speech) modality, which shows an important *intra-speaker variability*. One possible solution to cope with this problem is *to use more than one (user-friendly) biometric measure*. In this new *multi-modal* context, it is thus becoming important to be able to combine (or *fuse*) the outcomes of different modalities or of different experts.

The organization of this paper is as follows. Biometric modalities are dealt with in section 2. The characteristics of an automatic verification system are given in section 3. The multi-modal database, and the experimental protocol used in this study are detailed in section 4, and the individual experts are commented upon in section 5. Section 6 introduces data fusion and section 7 cites the different methods used to solve our particular decision fusion problem. Section 8 compares the results of all the methods used and section 9 concludes this work and formulates some recommendations.

2 BIOMETRIC MODALITIES

Automatic mono-modal biometric verification systems are usually built arranging two main modules in series: (1) a module which compares the measured features from the person under test with a reference client model and gives a scalar number[2] as output, followed by (2) a decision module realized by a thresholding operation.

[*] CNRS URA-820, ENST/TSI Department, Paris, France

[1] In a verification process, the identity claimed by the person under test is verified in a binary {accept,reject} hypothesis test

[2] This scalar number will be called a *score*, which states how well the claimed identity has been verified

Generally, the criterion to choose between mono- or multi-modal systems will be *system performance*. And although there exist mono-modal biometric verification techniques that do offer very small classification errors, the main problem with this category of biometrics is that they are either too *expensive* to be used in a general purpose context (there is for instance not a fingerprint scanner on every mouse) or perceived by the user as too *invasive*. So very often one is confronted with the obligation of using inexpensive hardware and non-invasive user-friendly biometrics. Two of the most popular biometrics that can conform to these constraints are faces and voices. However, the drawback of using inexpensive hardware (cheap black and white CCD-cameras and low-cost microphones) to obtain the raw data measurements of these biometrics, has as a direct consequence that the measurements generally will be corrupted with noise and/or distorted. This makes the use of a mono-modal biometric verification system based solely either on the facial or on the vocal modality a very big challenge, especially since it is usually not possible to update the database references of the authorized users on a regular basis.

A possible and straightforward way of building a multi-modal verification system from d mono-modal systems is to input the d scores provided in parallel into a fusion module, which combines the d scores and passes the fused score on to the decision forming module. This module then has to take the decision *accept* or *reject*, based on a threshold. The formal presentation of this general data fusion problem will be given in section 6.

3 CHARACTERIZATION OF VERIFICATION SYSTEMS

When dealing with binary {accept, reject} hypothesis testing, it is trivial to understand that the decision module can make two kind of errors. Applied to this problem of the verification of the identity of a person, these two errors are called False Rejection (FR), *i.e.* when an actual *client* is rejected as being an *impostor*, and False Acceptance (FA), *i.e.* when an actual *impostor* is verified as being a *client*. The performances of a speaker verification system are usually given in terms of the global error rates computed during tests: the False Rejection Rate [FRR = (number of FR)/(number of client claims)] and the False Acceptance Rate [FAR = (number of FA)/(number of impostor claims)] [1]. A unique measure can be obtained by combining these two errors into the Total Error Rate [TER = (number of FA + number of FR)/(total number of accesses)] or its complimentary, the Total Success Rate [TSR = 1 - TER]. The overall performance of an identity verification system is better characterized by it's so-called *Receiver Operating Characteristic (ROC)*, which represents the FAR as a function of the FRR [9]. Each point on such a ROC characteristic corresponds with a particular decision threshold. The Equal Error Rate (EER: *i.e.* when FAR = FRR), is often used as the only performance measure of an identity verification method, although this measure gives just one point of the ROC and comparing different systems solely based on this single number can be very misleading [7]. Unfortunately in practice it is not always possible to explicitly identify a continuous decision threshold in a certain paradigm. In these specific cases only a single operating point on the ROC will be found.

All verification results in this paper will be given in terms of FRR, FAR, and TER. For each error the 95 % level confidence interval will be given between square brackets. These confidence levels will be calculated assuming that the probability distribution for the number of errors is binomial. But since the binomial law can not be easily analyzed in an analytical way, the calculation of confidence intervals can not be done directly in an analytical way. Therefore we have used approximations of the binomial law that do allow an analytical analysis. Two of these approximations are the Poisson law (in the case of small samples) and the Normal law (in the case of large samples). The large sample approach is already statistically allowed starting from 30 samples. More detailed information can be found in [3].

4 EXPERIMENTAL PROTOCOL

The M2VTS multi-modal database comprises 37 different persons and provides 4 regular enrollment sessions (shots) for each person. These shots were taken at intervals of at least one week. During each shot, people were asked (1) to count from "0" to "9" in French (which was the native language for most of the people) and (2) to rotate their head from 0 to -90 degrees, back to 0 and further to +90 degrees, and finally back again to 0 degrees. The first enrollment session has been used for training the individual experts. The second enrollment session containing the same 37 subjects has been used for validating the individual experts (calculating thresholds). The third enrollment session has been used to test these validated experts and this data set has also been used to train the fusion modules, which gives us 37 client and $36 \times 37 = 1332$ impostor reference points. The fourth enrollment session has been used to test the fusion modules, which gives again 37 client and 1332 impostor claims.

5 INDIVIDUAL EXPERTS

All the experiments have been performed using three experts: a profile image expert based on a template matching method [8], a frontal image expert based on a robust correlation method [6], and a vocal expert based on a sphericity measure calculated on second order statistics [2]. The performances achieved by these three mono-modal identity verification systems are given in Table 1. The results have been obtained by adjusting the threshold

at the EER on the validation set and applying this threshold as an a priori threshold on the test set.

Table 1: Verification results for individual experts

Expert	FRR (%) (37 tests)	FAR (%) (1332 tests)	TER (%) (1369 tests)
Profile	21.6 [11.4,37.2]	8.5 [7.1,10.1]	8.9 [7.5,10.5]
Face	21.6 [11.4,37.2]	8.3 [6.9, 9.9]	8.7 [7.3,10.3]
Vocal	5.4 [1.5,17.7]	3.6 [2.7, 4.7]	3.7 [2.8, 4.8]

6 DATA FUSION

6.1 Taxonomy of data fusion

Data fusion in the broad sense can be performed at different hierarchical levels or processing stages. A very commonly encountered taxonomy of data fusion in the broad sense is given by a three-stage hierarchy [4]. *Data fusion* in the strict sense is the process of combining directly the data streams of raw measurements coming out of the different sensors,*feature fusion* is the process of combining features extracted from the raw measurements, and *decision fusion* is the process of combining partial soft (for instance a continuous score between 0 and 1) or hard (0 for a reject and 1 for an acceptance) decisions, given by the different experts. This three-level hierarchy has become fairly accepted, although it stays a matter of subjective choice. One could add an additional, orthogonal, dimensionality to the fusion process, representing *temporal* fusion. In this paper, we limit ourselves to soft *decision* fusion techniques.

6.2 Decision fusion architectures

Combining the partial decisions from the d different experts in a decision fusion strategy without considering the temporal fusion aspect, could be done using either a *serial* expert architecture, or a *parallel* expert architecture. In the latter case, the decisions derived from these experts are combined in parallel by the fusion module. This architecture is particularly well suited to combine the decisions or scores from experts that are capable of operating simultaneously and independently of one another. Since this is the case of our three experts, this is the architecture chosen for the decision fusion modules in this work.

6.3 Decision fusion as a particular classification problem

In a verification system with d experts in parallel, the decision fusion module using a binary decision scheme has to realize a mapping from the unitary hypercube of $\mathbb{R}^d$ into the set $\{\text{rejected}, \text{accepted}\}$. A classifier having a d-dimensional input vector and two classes $\{\text{rejected}\}, \{\text{accepted}\}$ is characterized by such a mapping. The *multi-expert* fusion module can therefore be considered as a *multi-dimensional* classifier. This particular classification case will be our standard fusion approach, since it allows to fall back immediately onto techniques available in the vast field of Pattern Recognition. In the next section, we cite some of the methods for solving this decision fusion problem are presented.

7 METHODS

To solve our particular classification problem, classic parametric and non-parametric statistical pattern recognition techniques have been adapted. In the parametric class, classifiers based on the general Bayesian decision theory (Maximum A-posteriori Probability and Maximum Likelihood), and on a simplified version of it (the Naive Bayesian classifier, which has been applied in the case of simple Gaussians and in the case of a logistic regression model), were experimented [12]. Furthermore experiments have been done using Linear, Quadratic, and Multi-Layer Perceptron classifiers [11]. Also, several non-parametric classifiers have been used in this work [10]. A first representative has been the family of the (very simple) voting methods. Thereafter the k-NN classifier and some of its possible variants have been studied. And to finish with this class, another popular paradigm, the binary decision tree, has also been used.

8 COMPARISON OF THE DIFFERENT METHODS

Table 2 gives an overall view of the best verification results obtained with each classifier. The first observation we can make when looking at the results (including the confidence intervals) in Table 2, is that fusion always improves the system performances beyond those of even the best single expert. The second observation is that these results seem to indicate that the logistic regression method gives

Table 2: Summary table of verification results of different fusion methods

Method	FRR (%) (37 tests)	FAR (%) (1332 tests)	TER (%) (1369 tests)
Gaussians	5.4 [1.5,17.7]	0.0 [0.0,0.3]	0.2 [0.1,0.6]
Log. regression	2.7 [0.5,13.8]	0.0 [0.0,0.3]	0.1 [0.0,0.5]
Lin. classifier	0.0 [0.0, 9.4]	3.1 [2.3,4.2]	3.0 [2.2,4.0]
Quad. classifier	0.0 [0.0, 9.4]	2.4 [1.7,3.4]	2.3 [1.6,3.2]
MLP	0.0 [0.0, 9.4]	0.4 [0.2,0.9]	0.4 [0.2,0.9]
Majority vote	0.0 [0.0, 9.4]	3.2 [2.4,4.3]	3.1 [2.3,4.2]
k-NN	8.0 [2.7,21.2]	0.0 [0.0,0.3]	0.2 [0.1,0.6]
k-NN + VQ	0.0 [0.0, 9.4]	0.5 [0.2,1.0]	0.5 [0.2,1.0]
Decision tree	7.7 [2.6,20.8]	0.3 [0.1,0.8]	0.5 [0.2,1.0]

the best overall results in our application. To verify that, we did a validation test using the logistic regression on the same M2VTS database, but this time using a so-called *leave-one-out* protocol [5]. This strategy has led to 2.592 client and an equal number of impostor tests. The results obtained in this manner are given in Table 3.

Table 3: Validation of the logistic regression using a leave-one-out protocol on the M2VTS database

Method	FRR (%) (2.592 tests)	FAR (%) (2.592 tests)	TER (%) (5.184 tests)
Log. regression	0.0 [0.0,0.2]	0.0 [0.0,0.2]	0.0 [0.0,0.2]

This validation experiment shows indeed that the logistic regression does perform as well as predicted by the tests done using the original, more limited, test protocol.

9 CONCLUSIONS

One of the most important conclusions of this paper is that, in our application, fusion always improves the system performances beyond those of even the best single expert. Although the presented multi-modal verification systems based on logistic regression seems to perform almost perfectly, these results have to be seen in their correct perspective, taking into account the very limited database. In any case, these performances are much better than those of verification systems using only one of the presented modalities. This justifies the (dynamic) use of multi-modal verification systems, especially in intrinsically multi-modal applications such as interactive dialogue systems.

The question of which fusion method to choose, is a much more difficult one to answer. A lot depends on the application. To be able to choose a number of potentially powerful fusion paradigms, it is important to have a (large) representative database of your application, which can be used for training purposes. It also helps if one is able to visualize the different populations (clients versus impostors), because in that specific case the choice of the fusion methods could be guided by the shape of the separation frontier between the two populations one wants to obtain.

References

[1] F. Bimbot and G. Chollet. Assessment of speaker verification systems. In *Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter, 1997.

[2] F. Bimbot and L. Nathan. Second-order statistical measures for text-independent speaker identification. In *ESCA workshop on automatic speaker recognition, identification and verification*, Martigny, Switzerland, April 1994.

[3] G. Chollet and C. Montacie. Evaluating speech recognizers and data bases. In H. Niemann, M. Lang, and G. Sagerer, editors, *Recent advances in speech understanding and dialog systems*, volume 46 of *NATO ASI F: Computer and Systems Sciences*, pages 345–348. Springer-Verlag, 1988.

[4] B. V. Dasarathy. *Decision Fusion*. IEEE Computer Society Press, 1994.

[5] P. A. Devijver and J. Kittler. *Pattern Recognition: A Statistical Approach*. Prentice-Hall Inc., Englewood Cliffs NJ, 1982.

[6] J. Matas, K. Jonsson, and J. Kittler. Fast face localization and verification. In A. Clark, editor, *British Machine Vision Conference*, pages 152–161. BMVA Press, 1997.

[7] J. Oglesby. What's in a number?: Moving beyond the equal error rate. In *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pages 87–90, Martigny, April 1994.

[8] S. Pigeon and L. Vandendorpe. Profile authentication using an optimized Chamfer matching algorithm. In *Proceedings of the First International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA'97)*, Lecture Notes in Computer Science. Springer Verlag, 1997.

[9] H. L. Van Trees. *Detection, Estimation and Modulation Theory*, volume 1. John Wiley & Sons, New York, 1968.

[10] P. Verlinde and G. Chollet. Comparing decision fusion paradigms using k-NN based classifiers, decision trees and logistic regression in a multi-modal identity verification application. In *Second International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA)*, Washington D. C., USA, March 1999.

[11] P. Verlinde, P. Druyts, G. Chollet, and M. Acheroy. A multi-level data fusion approach for gradually upgrading the performances of identity verification systems. In B. Dasarathy, editor, *Sensor Fusion: Architectures, Algorithms, and Applications III*, volume 3719, Orlando, FL, USA, April 1999. SPIE Press.

[12] P. Verlinde, P. Druyts, G. Chollet, and M. Acheroy. Applying Bayes based classifiers for decision fusion in a multi-modal identity verification system. In *International Symposium on Pattern Recognition "In Memoriam Pierre Devijver"*, Brussels, Belgium, February 1999.