Fusion of Audio and Video Information for Multi Modal Person Authentication

Benoît Duc [a] Elizabeth Saers Bigün [b,c] Josef Bigün [d]

Gilbert Maître [e] Stefan Fischer [a]

[a] *Signal Processing Laboratory, Swiss Federal Institute of Technology, Lausanne, Switzerland*

[b] *Stockholm University, Dep. of Statistics, Stockholm, Sweden*

[c] *Centre for Safety Research, Royal Institute of Technology, Stockholm, Sweden*

[d] *Microprocessor and Interface Laboratory, Swiss Federal Institute of Technology, Lausanne, Switzerland*

[e] *IDIAP, Martigny, Switzerland*

Preprint submitted to Elsevier Preprint *24 April 1997*

We present an algorithm functioning as a supervisor module in a multi-expert decision making machine. It uses the Bayes theory in order to estimate the biases of individual expert opinions. The biases are used to calibrate and conciliate expert opinions to a single decision. This supervision technique is applied to the real case of a person authentication technique using two modalities, face and speech. The visual part involves the matching of a coarse grid containing Gabor phase information from face images. The acoustic part is performed by a text-dependent speaker verification system based on Hidden Markov Models. Experimental results show that the proposed fusion method improves the quality of individual expert decisions by reaching success rates of 99.5%.

Key words: multimedia data analysis, decision fusion, Bayesian statistics, elastic graph matching, hidden Markov models

1 Introduction

There is an increasing interest in biometric techniques for person authentication, in particular for those where the user is not involved in complicated and intrusive procedures. The applications of such techniques prevent misuses of services with automatic access of clients. Unfortunately, mono-modal recognition techniques are likely to reach in a close future a saturation in performance. A potential way of overcoming such limitations, consists in combining results from several modalities.

Machine experts, referred to as experts below, can deliver opinions, more precisely *scores*, about the authenticity level of a person's claim being a certain client. This paper addresses the issue of how these scores can be represented and conciliated to a single opinion on the authenticity level of the user's identity claim.

Subjective assessments are a natural part of the Bayesian statistics (Bernardo and Smith, 1994). The conciliation procedure in the current paper is built on the ideas of Bigün (1995), who deals with aggregation and calibration of the experts' assessments when independency between the assessments is assumed. We estimate the expected true authenticity score of a candidate person who arrives to the system in a future time instant, given earlier authentication experiments used as supervised training. To obtain these estimations, the posterior density of the true authenticity scores will be used. The basic assumption is that the logarithm of the misidentification score has a normal distribution. A work with a similar interest in modality fusion for person identification has been carried out by Brunelli and Falavigna (1995). From face images and speech, they extract five measurements that are normalised into scores. They propose two different ways to combine them: a weighted geometric averaging, and the integration of ranks and scores by means of an artificial neural network. Here, we do authentication instead of recognition. Another difference is that we use a Bayesian approach for decision fusion. More recently, fusion of lip motion and speech by *synergetic computer* has been proposed for identification and authentication (Dieckmann et al., 1997). Jourlin et al. (1997) achieve the fusion of labial and speech information by a weighted sum of scores. The weighting factors are determined so as to minimise the total error rate on an evaluation set.

The paper is organised as follows: the framework of multi-modal authentication is presented in Section 2, while the fusion method itself is described in Section 3. In Section 4, the authentication experts based on speech, respectively face, are presented. Section 5 is devoted to experiments. Finally, some conclusions are drawn.

2 Multi-Modal Authentication System

For an authentication system, the world population is divided into two categories. A *client* is someone who is known to the system and is entitled to privileges, while an *impostor* is a person who falsely claims to have the identity of a client. Usually, the set of clients is small compared to the set of impostors. Since the potential set of impostors is almost the world population, the experts are not assumed to train on impostors, except possibly if the same impostors show up repetitively in which case they can be considered as clients with "special" privileges. In the identity verification scenario considered here, the person willing to access a service is cooperative, i.e. an identity is claimed. The authentication device then takes the decision to accept or reject the person according to the claimed identity. Two types of error are possible: false acceptance of an impostor (FA), and false rejection of a client (FR).

The information provided by sensors in a multi-modal authentication system can be highly heterogeneous, like sound (speech) and images (face, iris, fingerprints). Therefore, one cannot expect to merge them at the sensor level. We rather fuse partial decision results obtained by each expert, which is known as *decision fusion* (Dasarathy, 1994). The supervisor does not interfere with and does not have access to the computational processes of the experts. Thus, the whole system is modular, so that modalities may be added or removed depending on the requirements of the application at hand. This results in a general system where m independent modules (experts) perform identity verification on their own type of data and provide their decision as a number, called *score*. Furthermore, they may give an estimation of the variance of their score. By convention, the scores of the experts are in the $[0, 1]$ interval.

3 Expert Decision Fusion

3.1 Statistical Model

An extensive presentation of the mathematical background of the model can be found in Bigün (1995). The set of data together with an identity claim is called *shot*. The authenticity score of expert i on shot j will be denoted X_{ij}. The true authenticity score of shot j's claim, Y_j, which is independent of the expert, takes only two numerical values corresponding to "True" and "False". The *miss-identification score* Z_{ij} is defined by $Z_{ij} = Y_j - X_{ij}$. Finally, the variance of Z_{ij} as estimated by expert i is denoted by s_{ij}.

We denote the expected value of Z_{ij} by b_i. Assume that Z_{ij} given b_i is normally distributed, i.e. $(Z_{ij}|b_i) \in N(b_i, \sigma_{ij}^2)$. We assume first that the variances σ_{ij}^2 are known and that Z_{ij} are independent for all i and j. We suppose that b_i has a non-informative prior distribution, i.e. $b_i \in N(0, \infty)$. Then the posterior distribution of b_i is normal with mean and variance:

$$M_i = \frac{\sum_{j=1}^n \frac{z_{ij}}{\sigma_{ij}^2}}{\sum_{j=1}^n \frac{1}{\sigma_{ij}^2}} \quad \text{and} \quad V_i = \frac{1}{\sum_j^n \frac{1}{\sigma_{ij}^2}}, \tag{1}$$

respectively. Since $(Z_{i,n+1}|b_i)$ is normally distributed, the predictive distribution of $Z_{i,n+1}$ given $z_{i1}, z_{i2}, \cdots, z_{in}$ is also normal with mean M_i and variance $(V_i + \sigma_{i,n+1}^2)$. If we know $X_{i,n+1}$ the predictive distribution of Y_{n+1} is also normal:

$$(Y_{n+1}|z_{i1}, z_{i2}, ..., z_{in}, x_{i,n+1}) \in N(M_i', V_i'), \tag{2}$$

where

$$M_i' = x_{i,n+1} + M_i \quad \text{and} \quad V_i' = V_i + \sigma_{i,n+1}^2. \tag{3}$$

The ith expert's authenticity score on a future shot will be calibrated by means of the expected value in (1). Assume that the m independent experts are asked to give their authenticity scores on all shots $(j = 1, 2, ..., n, n+1)$ being the shots of a client. Then, the posterior distribution of b_i, given all these scores and the earlier identification errors, is normal:

$$(Y_{n+1}|z_{11}, ..., z_{1n}, x_{1,n+1}, ..., z_{m1}, ..., z_{mn}, x_{m,n+1}) \in N(M, V), \tag{4}$$

where

$$M = \frac{\sum\limits_{i=1}^{m} \frac{M_i'}{V_i'}}{\sum\limits_{i=1}^{m} \frac{1}{V_i'}} \quad \text{and} \quad V = \left(\sum_{i=1}^{m} \frac{1}{V_i'}\right)^{-1}. \tag{5}$$

This is the posterior distribution of Y_{n+1} after having observed the assessment errors of the m experts. Since the variances σ_{ij} are not known we have to estimate them. We suppose that the experts give the precisions correctly except for an individual proportionality constant:

$$s_{ij} = a_i \sigma_{ij}^2. \tag{6}$$

The constant a_i may be interpreted as a factor representing over- or under-confidence. Assume, a priori, the following non-informative joint distribution of a_i and b_i :

$$f(a_i, b_i) = \frac{1}{a_i}. \tag{7}$$

Under the assumption that $(Z_{ij}|a_i, b_i, s_{ij}) \in N(b_i, \sigma_{ij}^2)$ are independent for all i and j, it can be shown that the estimate $\bar{\sigma}_{ij}^2$ of σ_{ij}^2 is given by:

$$\bar{\sigma}_{ij}^2 = \alpha_i s_{ij} = \frac{(G_i - D_i)}{n-3} \cdot s_{ij}, \tag{8}$$

where

$$G_i = \sum_{j=1}^{n} \left(\frac{z_{ij}^2}{s_{ij}} \right), \quad D_i = \left(\sum_{j=1}^{n} \left(\frac{z_{ij}}{s_{ij}} \right) \right)^2 \left(\sum_{j=1}^{n} \left(\frac{1}{s_{ij}} \right) \right)^{-1}, \quad \alpha_i = \frac{G_i - D_i}{n-3} \tag{9}$$

Consequently we use $\bar{\sigma}_{ij}^2$ in (1) and (3) instead of σ_{ij}^2.

3.2 Fusion Algorithm

From this mathematical description of the Bayesian conciliation of experts, one can obtain a fusion algorithm, which is summarised as follows:

(i) **Training phase**: by using a training set $\{(x_{ij}, y_j, s_{ij}), \quad j = 1...n\}$, estimate the bias parameters of each expert, i.e. $\{M_i, V_i, \alpha_i\}$, according to (1); σ_{ij}^2 is computed according to (8), and (9).

(ii) **Authentication phase**: at this step, the supervisor is operational, meaning that the time instant is *always* $n+1$ and that the supervisor has access to expert opinions $(x_{i,n+1}, s_{i,n+1})$, but not access to the true authentication scores, y_{n+1}; $\sigma_{i,n+1}^2$ is computed according to (8). The expert opinions are normalised yielding M_i', and V_i', (3). M and V are computed according to (5), and may be thresholded to yield a definite decision.

3.3 Score transformation

By convention, the scores of the experts are in the $[0, 1]$ interval. For our purposes, the transformation

$$X_{ij} = \log \frac{X_{ij}^*}{1 - X_{ij}^*}, \tag{10}$$

which is also known as the "odds of X_{ij}^*" (Lindley, 1965), will be used to map the scores to $[-\infty, \infty]$. Here, the superscript * denotes the scores before transformation.

It can be shown that the formulas (1,3,5) still hold when X_{ij} is substituted by X^*_{ij}, and Y_j by Y^*_j. The only difference is in the conditional distribution of Y_{n+1}, (4), which is log normal with the expected value $\exp(M + V/2)$ and variance $(\exp(2M + 2V) - \exp(2M + V))$.

Finding the expectation value of Y^*, given the knowledge of the expert estimations of it, is what would be ideal for applications. However, obtaining an analytical expression of it from the expected values and variances above is, to the best knowledge of the authors, not possible. Instead we have used $S^{\text{Bayes}}_{n+1} = M$ where S^{Bayes}_{n+1} represents an authenticity score of a candidate person's being client in the future, given the past experience from the experts. We do not make use of V in our supervisor currently.

It is crucial to observe that in our application, $Y^* = 0$ in case of an impostor, and $Y^* = 1$ in case of a client. In order to avoid singularities with transformation (10), one chooses to work on the $[\epsilon, 1 - \epsilon]$ interval instead of $[0, 1]$. In our experiments, we chose $\epsilon \approx \exp(-6)$, so that $Y = \pm 6$. However, possible values of Y^* are still at the interval extremities. As a consequence, if $y_j = -6$, $z_{ij} = y_j - x_{ij}$ is always negative. Likewise z_{ij} is always positive if $y_j = 6$. This means that we do not have a mono-modal density but a bimodal density for z_{ij}, if we interpret the scores of experts straightforwardly. To obtain a mono-modal density, we build and train two supervisors, one which is specialised on impostors and one specialised on clients. Step (i) of the algorithm is applied to the impostors and the clients of the training set separately, yielding two bias parameter sets, $\{M_i^I, V_i^I, \alpha_i^I\}$ and $\{M_i^C, V_i^C, \alpha_i^C\}$. The second step is then applied twice for each of these parameters. The result of the two supervisors can be combined further by using another supervisor (see Section 5.1).

4 Authentication Experts

In the authentication experiments, two experts were used, processing two different signals: speech audio data and face images. They provide scores in the $[0, 1]$ interval.

4.1 Speech-based Authentication

Speech-based authentication is performed by a text-dependent speaker verification system which accepts as text only a known sequence of digits. The system output is a score as specified by the chosen multi-modal architecture (Section 2). This authentication modality is an adaptation of the work of Genoud et al. (1996).

The audio input signal is first segmented into the given sequence of digits by using a recognition module based on *Hidden Markov Models* (HMMs). Each segment is then transformed into the *Linear Predictive Cepstral Coefficients*

(LPCC) representation. In the learning phase digit HMMs are trained with these segments. In the verification phase, the likelihood of the segments to be produced by the corresponding digit models is estimated.

All digit models have the same left-right structure. The number of states depends on the digit and has been chosen so that there is one state per phoneme and one state per transition between phonemes. In all cases, a Gaussian distribution with diagonal variance vector is used to model the feature distribution within one state. Training is performed with the help of the Baum-Welch algorithm and the likelihood during the verification phase is computed with the Viterbi algorithm.

The verification of one person uses two models for each digit: the client model, which models the person whose identity is claimed, and a world model, which represents an average of speakers. The client model is built from the recorded voice data of that client, while the world model is built from the voice recordings of a large set of persons, including or, preferably, not including that client. In the implementation of the system used for the experiments, the digit world models have been trained on the *Polycode* database (Genoud and Chollet, 1995) with 300 examples, each sample being from a different speaker.

The verification score x_{speech} is computed as follows:

$$x_{\text{speech}} = \sum_{digits} \frac{log(L_p) - log(L_w)}{N} \tag{11}$$

where L_p is the likelihood of the person model, L_w the likelihood of the world model, and N the number of LPCC frames. The score is mapped into the interval $[0, 1]$ with the use of a sigmoid function.

4.2 Face-based authentication

Each face is represented by a set of feature vectors positioned on nodes of a coarse grid placed on the image. Comparing two face images is accomplished by elastic graph matching, i.e. by adapting a grid taken from one image to the features of the other image (Lades et al., 1993). We use the modulus of complex Gabor responses as feature vectors from a set of filters with 6 orientations and 3 resolutions. These are sets of features that describe local properties of points in the image, similar to those in Bigün and du Buf (1994).

The grid matching aims at normalising the input face, in order to make the subsequent comparison invariant with respect to translation and a reasonable amount of deformation. The residual error accounts for the difference between the normalised input and the reference pattern. The grid matching consists here of two consecutive steps. The first minimises an objective function that measures the difference of the reference and the test feature vectors, by translating an undeformed grid on the test image and computing test feature vectors at current node locations. The second step translates every node incrementally to find local minima, resulting in a deformed grid with a lower objective value.

The residual grid matching error could be used as a simple discriminant measure. Our experiments showed that it is not powerful enough. Better results can be obtained by weighting the node contributions according to their significance for authentication. This is achieved by designing at each node a local discriminant measure that minimises a criterion for all views of a given person while maximising it for the average of the other people in the training set (Duc et al., 1997). All local discriminant responses, which are actually projections on relevant subspaces, are added to provide a unique, global response R for a test face. The score $x_{\text{face}} \in [0,1]$ is obtained from this response by applying a sigmoid mapping.

5 Experiments

We have conducted a set of experiments based on simulated expert opinions and a set of experiments using true expert scores. The purpose of the first set of experiments was to study the general performance of the supervisors, according to different settings of parameters such as the number of experts and the skills of experts. The second set of experiments had the purpose of checking the validity of the conclusions as predicted by simulations.

Although the presented theory is capable of making use of the experts' self-estimation of variance s_{ij}, it was not possible to obtain these numbers from our two real experts. We have therefore used $s_{ij} = 1$ for all shots, simulated or real. As a consequence, the supervisor calibrates the experts by using their historical success alone.

5.1 Simulated Experts

In the case of simulations, the scores X^*_{ij} of an expert i were drawn from two uniform distributions defined on the open intervals $(0, I_i)$, $(C_i, 1)$. The first interval was used when the identity claim of the candidate was truly false, i.e. $Y^*_j = 0$, and the second interval was used when the claim was authentic $((Y^*_j = 1)$. The Y^*_j's were picked at random to be either 0 or 1, with probabilities 0.8 and 0.2 respectively, so as to reflect real applications, where the number of impostors is larger than the number of clients.

In order to simulate the different skills of experts, I_i and C_i were varied. In order to obtain error-prone experts, which is a necessary condition for the usefulness of a supervisor, the conditions $0.5 < I_i < 1$ and $0 < C_i < 0.5$ were imposed. We picked I_i and C_i uniformly at random from the intervals $(0.5, 0.75)$ and $(0.25, 0.5)$. Thus, the false acceptance rates of our simulated experts are given by $FA_i = I_i - 0.5$, and the false rejection rates by $FR_i = 0.5 - C_i$. Since our expert scores are in the interval $[0, 1]$ we used the odds-transformation.

In practice, in order to avoid singularities, we had to restrict X^*_{ij}'s to be in $[\epsilon, 1-\epsilon]$. We have chosen $\epsilon \approx \exp(-6)$, yielding the odds -6 and 6 for 0 and 1, respectively. Consequently, when applying the odds-transformation to Y_j^*'s, we used ϵ and $1-\epsilon$ instead of 0 and 1. In simulations, the results did not show a change with various choices of small ϵ.

According to Section 3.3, two supervisors were trained, based on the "client" and "impostor" categories of the training set. The combination of the two supervisors was achieved by choosing the response of the supervisor which comes closest to its goal, i.e. -6 or 6, as the final S^{Bayes}. Formally, if $|6 - M^C| - |-6 - M^I| > 0$, then $S^{\text{Bayes}} = M^I$, otherwise $S^{\text{Bayes}} = M^C$.

For all experiments, the number of shots in the training set was $n = 2664$ and the number of shots in the test set was 7992. As a comparison we also simulated a simple supervisor, with its scores consisting of the mean values of expert scores, $S_{n+1}^{\text{mean}^*} = \frac{1}{m}\sum_{i=1}^{m} X_{i,n+1}^*$. The S_{n+1}^{mean} used subsequently is the odds of $S_{n+1}^{\text{mean}^*}$ computed according to (10), for the sake of comparisons with the Bayesian supervisor.

We compared the performance in terms of the *success rate SR*, defined by $SR = 1 - (FA + FR)$. The Bayesian supervisor yielded consistently better results than the average of scores. Few experts (e.g. 2), or many experts, (4 and more), were always conciliated better by S_{n+1}^{Bayes} than S_{n+1}^{mean}. Hard decisions were obtained by thresholding S_{n+1}^{mean} and S_{n+1}^{Bayes} at the zero threshold. In case of equally skilled experts, S_{n+1}^{mean} was outperforming significantly the individual

SR_1	SR_2	SR_3	SR_4	SR_{Bayes}	SR_{mean}
0.814	0.816	0.819	0.825	1.00	0.974
0.649	0.952	0.640	0.608	0.991	0.801

SR_1	SR_2	SR_3	SR_4	SR_{Bayes}	SR_{mean}
0.664	0.821	-	-	0.985	0.815
0.763	0.732	0.701	0.681	0.999	0.872

Table 1
Success rate simulations. $SR_1 \cdots SR_4$ are expert success rates, SR_{Bayes} is the success rate of the Bayesian supervisor and SR_{mean} is the success rate of the arithmetic mean of the scores of the experts. First table: Row 1, equal skills; Row 2, random skills. Second table: Row 1, 2 experts case; Row 2, 4 experts.

experts. However, in case of unequally skilled experts, it often performed worse than the best expert. Typical results with four experts are given in Table 1.

Figure 1, left, illustrates the density curve (estimated by a 512 bin histogram) of $z_{n+1} = y_{n+1} - S_{n+1}^{\text{mean}}$ of a simulation with randomly skilled experts with unsymmetric skills, which happens most likely in practice. The larger mass around -6 corresponds to the over representation of impostors due to our deliberate choice. The bimodal nature of the errors are clearly visible. The corresponding plot for the Bayesian supervisor is given by Figure 1, right. In both cases two experts were employed.

The results suggest that an increase in the number of experts, even if the experts are not highly skillful, improves the supervisor decisions. While the improvement is consistently significant in the case of Bayesian estimator, it is not significant when the mean supervisor has experts with high variations in their skills.

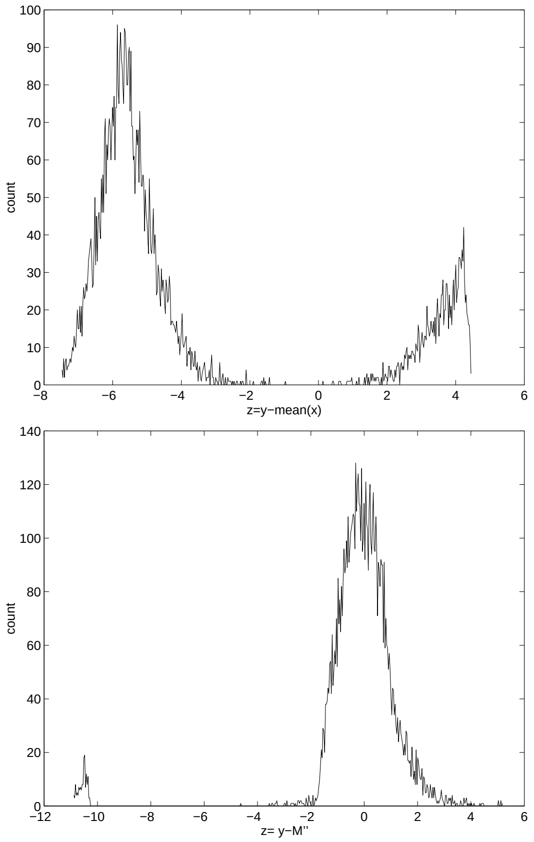

Fig. 1. The miss-identification density of the mean value supervisor (up), and the Bayesian supervisor (down) in the case of two unequally skilled experts

5.2 Speech and Face Based Authentication

5.2.1 Experimental Database and Protocol

The database and the protocol of experiments are those of the European M2VTS project (Pigeon and Vandendorpe, 1997). The database includes 4 shots of 37 persons, taken at one week intervals. Each shot is built of the video recording of the person rotating his head and of the synchronised audio and video recording of the person spelling the sequence of digits from '0' to '9'. The speech-based authentication uses the audio recording of the digits sequence while the face-based authentication uses a set of frontal, grey-level images selected from the head-rotating video sequence in a semi-automatic manner at QCIF format (144 × 176).

The experiments were conducted following a combination of the left-one-out and the rotation estimates (Devijver and Kittler, 1982). Each person in turn is labelled as an *impostor*, while the 36 others are considered as *clients*. Three shots of the 36 clients constitute the training set while the fourth shot is used as evaluation set in the following way: each client tries to access under its own identity, and the impostor tries to access under the identity of the 36 clients. This results in 36 authentic tests and 36 impostor tests. This procedure is repeated four times, by considering each shot as the evaluation series successively. In total, the client and impostor tests amount each to $37 \times 4 \times (37 - 1) = 5328$.

For both single modalities, experiments have been conducted on the M2VTS database according to that protocol. Results at the a priori threshold are given in the first two raws of Table 2, and total error curves in Figure 2. Clearly, the experts are unequally skilled.

5.2.2 Experimental results

Experiments on fusion were conducted with the Bayesian supervisor as well as with the arithmetic mean of scores (see Section 5.1). For training the Bayesian

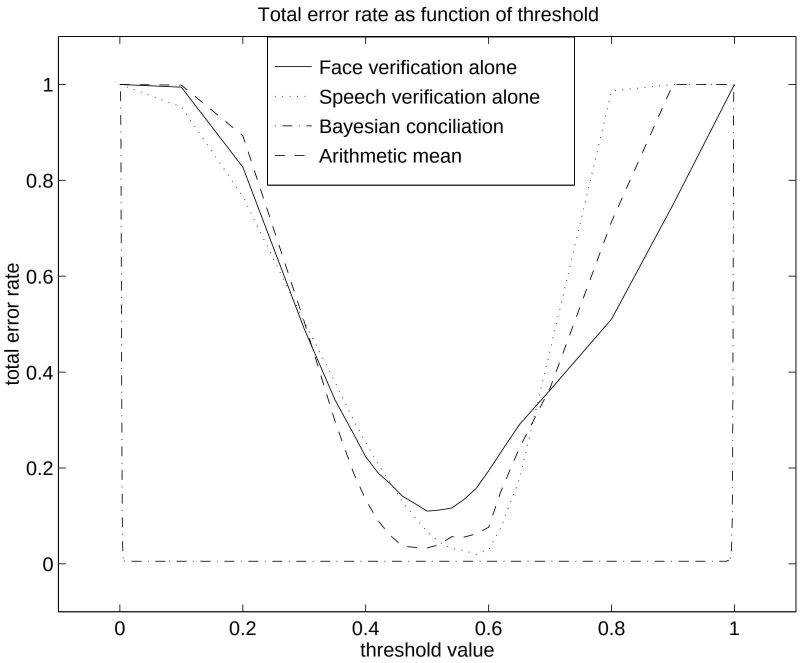

Fig. 2. Total error ($TE = FA + FR$) curves for single modalities and fusion of modality scores.

supervisor, 2664 samples from the test experiments were selected, with equal proportion of impostor and authentic accesses. The 7992 remaining experiments were used for test. For the arithmetic mean of scores, as no training is required, all $2 \cdot 5328 = 10656$ experiments were used for testing.

Figure 2 shows the total error $TE = FA + FR$, as a function of the threshold value. One can see that single modalities as well as the fusion by arithmetic mean are sensitive to the threshold: their TE curves show a narrow minimum near 0.5. For Bayesian conciliation, the total error remains stable at a very low value of $5.4 \cdot 10^{-3}$ over almost the whole threshold range. As a consequence, the choice of the threshold is less critical for the Bayesian conciliation than for other alternatives. This appears as a major advantage of the Bayesian approach.

fusion method	*FA* rate (%)	*FR* rate (%)	*TE* rate (%)	*SR* (%)
face	3.6	7.4	11.0	89.0
speech	6.7	0.0	6.7	93.3
arithmetic mean	1.2	2.1	3.3	96.7
Bayesian conciliation	0.54	0.0	0.54	99.46

Table 2
Verification error rates with face, speech and two combinations of face and speech scores. The sum of the error rates (total error, *TE*) and the success rate ($SR = 1 - TE$) are given in the last two columns.

Table 2 shows the error rates for the 0.5 threshold. The choice of this value is motivated by the fact that it is the a priori optimal threshold for single modalities. If, as this was the case here, no validation set is available for choosing an experimentally optimal threshold, the a priori threshold is used. Clearly, the Bayesian supervisor provides the best performance.

6 Conclusion

We have presented a Bayesian model in order to construct a supervisor conciliating machine experts. We verified that the score fusion we propose improves the decision making by using simulated as well as real data. In particular, the Bayesian supervisor shows robustness with respect to the choice of the decision threshold.

We calibrated the experts' authenticity scores of a future instant about a client only by means of earlier, supervised experiments. We may also calibrate scores about a particular client. In this paper we treated the case where the assumption of independency was made about the identification errors. It is possible to omit this assumption with some a priori knowledge about the covariance matrices, (Bigün, 1997).

It has proven to be non-trivial to request from expert designers to provide experts delivering a consistent quality of score. Since expert designers had different comprehensions of what the quality of scores should be like, it was not possible to evaluate the performance had we chosen to simulate them. Independently, however, the use of other rules than the one described in Section 5.1 for merging the impostor versus client predictions could be investigated.

As future work, more modalities will be added, like profile authentication and speech-lip synchronisation. This will allow to study the behaviour of the supervisor with respect to the number of modalities experimentally.

Acknowledgement

This work has been carried out within the framework of the European ACTS-M2VTS project, and has been funded by the Swiss Federal Office for Education and Science (OFES). We also thank J. Wallander and T. Hedelius foundations (Sweden) for financial support. The authors are grateful to Dominique Genoud, Pierre Jourlin and Hubert Wassner from IDIAP for their collaboration, and also to Gérard Chollet (IDIAP) and Helen Bigün (KTH, Sweden) for fruitful technical discussions.

References

Bernardo, J. M. and Smith, M. F. A. (1994). *Bayesian Theory.* Wiley, Chichester.

Bigün, E. (1997). Bayesian prediction based on few and dependent data. In Guedes Soares, C. and et. al., editors, *Safety and Reliability*, page 146ff. Elsevier, Oxford.

Bigün, E. S. (1995). Risk analysis of catastrophes using experts' judgements: An empirical study on risk analysis of major civil aircraft accidents in Europe. *European J. Operational research*, 87, 599–612.

Bigün, J. and du Buf, J. M. H. (1994). N-folded symmetries by complex moments in Gabor space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(1), 80–87.

Brunelli, R. and Falavigna, D. (1995). Person identification using multiple cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(10), 955–966.

Dasarathy, B. V. (1994). *Decision Fusion.* IEEE Computer Society Press.

Devijver, P. and Kittler, J. (1982). *Pattern Recognition: a Statistical Approach.* Prentice-Hall International, London.

Dieckmann, U., Plankensteiner, P., Schamburger, R., Fröba, B., and Meller, S. (1997). Sesam: A biometric person identification system using sensor fusion. In Bigün, J., Chollet, G., and Borgefors, G., editors, *First International Conference on Audio- and Video-based Biometric Person Authenti-*

cation (AVBPA'97), volume 1206 of *LNCS*, pages 301–310, Crans-Montana, Switzerland. Springer.

Duc, B., Fischer, S., and Bigün, J. (1997). Face authentication with sparse grid Gabor information. In *1997 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97)*, Munich, Germany.

Genoud, D. and Chollet, G. (1995). Polycode: a verification database. Technical report, IDIAP, CH-1920 Martigny.

Genoud, D., Gravier, G., Bimbot, F., and Chollet, G. (1996). Combining methods to improve the phone based speaker verification decision. In *ICSLP'96*, volume 3, pages 1756–1760, Philadelphia.

Jourlin, P., Luettin, J., Genoud, D., and Wassner, H. (1997). Acoustic-labial speaker verification. In Bigün, J., Chollet, G., and Borgefors, G., editors, *First International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA'97)*, volume 1206 of *LNCS*, pages 319–326, Crans-Montana, Switzerland. Springer.

Lades, M., J. C. Vorbrüggen, J. B., Lange, J., v.d. Malsburg, C., Würtz, R. P., and Konen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3), 300–311.

Lindley, D. (1965). *Introduction to Probability and Statistics.* Cambridge University Press, Cambridge, UK.

Pigeon, S. and Vandendorpe, L. (1997). The M2VTS multimodal face database. In Bigün, J., Chollet, G., and Borgefors, G., editors, *First In-*

ternational Conference on Audio- and Video-based Biometric Person Authentication (AVBPA'97), volume 1206 of *LNCS*, pages 403–409. Springer.