(C) In Proc. First International Conference on Audio- and Video-based Biometric Person Authentication, *Crans-Montana, Switzerland, 12-14 March 1997 (Springer, Lecture Notes in Computer Science)*

Expert conciliation for multi modal person authentication systems by Bayesian statistics

E. S. Bigün[1,3], J.Bigün[2], B. Duc[2] and S. Fischer[2]

[1] Stockholm University Dep. of Statistics S-106 91 Stockholm
[2] EPFL Signal Processing Laboratory CH-1015 Lausanne
[3] KTH Center for Safety Research S-100 44 Stockholm

Abstract. 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. These are then used to calibrate and conciliate expert opinions to one opinion. We present a framework for simulating decision strategies using expert opinions whose properties are easily modifiable. By using real data coming from a person authentication system using image and speech data we were able to confirm that the proposed supervisor improves the quality of individual expert decisions by reaching success rates of 99.5 %.

1 Introduction

Automatic access of eligible persons (clients) to services (privileges) is becoming common. A factor hampering the growth of these services is the weakness of security for the clients being authentic. In order to have low false acceptance (FA) and false rejection rates (FR) using different and preferably independent sensors, e.g. picture, video, voice ...etc, is under consideration. Machine experts, referred to as experts below, can deliver "opinions" about the authenticity level of a person's claim being a certain client. This paper addresses the issue of how these opinions should be represented and conciled to a single opinion on the authenticity level of the client pretentions.

Subjective assessments are a natural part of the Bayesian statistics, Bernardo & Smith [1]. There are several papers on how to concile different expert assessments, Lindley et. al. [7–9], French, [5], West [12], Winkler [13]. However, the conciliation procedure in the current paper is built on the ideas of Bigün [2,3]. Bigün [2] deals with aggregation and calibration of the experts' assessments when independency between the assessments are assumed, while, Bigün [3] treats the cases when there are dependencies between the expert assessments.

In this paper we will estimate the posterior expected true authenticity score of multi-sensor data (speech, image...etc). Or to be more precise, expected true authenticity score of a candidate person who arrives to the system in a future time instant, given the earlier miss-identifications (either FA or FR types) and the experts' authenticity scores of the candidate person. We will also estimate the precisions of the true authenticity scores. To obtain these estimations, the

posterior density of the true authenticity scores will be used i.e., we have a model. The basic assumption is that the logarithm of the misidentification score have a normal distribution, given the true authenticity score. The reasons are the following; the logarithms of the observed misidentification scores, exhibit symmetric properties and the normal distribution is mathematically very convenient.

2 System Model

We would like to construct a system consisting of one supervisor and m experts. An expert i consists of a hardware or software module which processes signals originating from one or more sensors in such a way that it can give graded opinions about the authenticity of a candidate person's being a client. A client is someone who is known to the system and has her own privileges, e.g. accessing to a particular office, or billing a certain account. An impostor is a person who falsely claims to have the identity of a client. An expert delivers its opinion on a "package" of data collected by sensors e.g. video camera, microphone, ..etc, at a relatively short instant of time (a few seconds). Such a data package *contains the claim of an identity,* and will be referred to as a *shot.* The supervisor does not interfer with or has access to the computational processes of the experts. Since the world population is the potential set of impostors, 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" priviliges.

Below is a list of notations we use throughout the paper.

i: Index of the experts. $i \in 1 \cdots m$,

j: Index of shots (one or more per candidate), $j \in 1 \cdots n, n+1$. It is equivalent to time since an expert has one shot per evaluation time (period).

X_{ij}: The authenticity score, i.e. the score delivered by expert i on shot j's claim of being a certain client

s_{ij} The variance of Z_{ij} as estimated by expert i

Y_j The true authenticity score of shot j's claim being a certain client. This variable can take only two numerical values corresponding to "True" and "False"

Z_{ij} The mis-identification score, that is $Z_{ij} = Y_j - X_{ij}$

The identity claim contained in a shot is assumed to be the identity of a certain client, since they can be immediately rejected without further processing otherwise. Although we report on when expert i is an expert of recognizing all clients with a particular modality, it is also possible to interpret the expert i as an expert of recognizing a particular client with a particular modality when

enough training shots are available. We assume that the shots $1 \cdots n$ are fresh shots of the clients, i.e. the experts have trained on other shots (with other sensor data). Shot $n+1$ is the shot of a candidate which neither the experts nor the supervisor have trained on. Therefore shot $n+1$ can be considered to belong to a future instant, or an instant when the system is in full use. Consequently we assume that the training phase of the experts (not the supervisor's) is already achieved.

3 Statistical Model

An extensive presentation of the mathematical background of the model can be found in Bigün [2].

We denote the expected value of Z_{ij} by b_i. Assume that Z_{ij} given b_i is normally distributed;

$$(Z_{ij}|b_i) \in N(b_i, \sigma_{ij}^2) \tag{1}$$

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

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

respectively. Since $(Z_{i,n+1}|b_i)$ is normally distributed, (1), 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 $Y_{i,n+1}$ is also normal;

$$(Y_{n+1}|z_{i1}, z_{i2}, ..., z_{in}, x_{i,n+1}) \in N(M_i', V_i') \tag{3}$$

where

$$M_i' = x_{i,n+1} + M_i \quad \text{and} \quad V_i' = V_i + \sigma_{i,n+1}^2. \tag{4}$$

The i:th expert's authenticity score on a future shot will be calibrated by means of the expected value in (2). 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;

$$(Y_{n+1}|z_{11}, ..., z_{1n}, x_{1,n+1}, ..., z_{m1}, ...z_{mn}, x_{m,n+1}) \in N(M_i'', V_i'') \tag{5}$$

where

$$M'' = \frac{\sum\limits_{i=1}^{m} \frac{M_i'}{V_i'}}{\sum\limits_{i=1}^{m} \frac{1}{V_i'}} \qquad \text{and} \qquad V'' = \left(\sum_{i=1}^{m} \frac{1}{V_i'} \right)^{-1} \tag{6}$$

This is the posterior distribution of Y_{n+1} after having observed the assessment errors of the m experts. The prior distributions of b_i and Y_{n+1} are non-informative. Since the variances are not known we have to estimate them. We

suppose that the experts give the precisions correctly except for an individual proportionality constant.

$$s_{ij} = a_i \sigma_{ij}^2 \tag{7}$$

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

$$f(a_i, b_i) = \frac{1}{a_i} \tag{8}$$

Assume also that $(Z_{ij}|a_i, b_i, s_{ij}) \in N(b_i, \sigma_{ij}^2)$ are independent for all i and j. The posterior distribution of a_i may now be computed as the marginal distribution of $f(a_i, b_i|(z_{i1}, s_{i1}), ..., (z_{in}, s_{in}))$

$$f(a_i|(z_{i1}, s_{i1}), ..., (z_{in}, s_{in})) = \frac{\beta^{\frac{n-1}{2}} (a_i)^{\frac{n-3}{2}} \exp(-a_i \beta)}{\Gamma\left(\frac{n-1}{2}\right)} \tag{9}$$

which is a Gamma distribution, $\Gamma\left(\frac{n-1}{2}, \beta\right)$ with $\beta = \frac{1}{2}(G_i - Di)$ where

$$G_i = \sum_{j=1}^{n} \left(\frac{z_{ij}^2}{s_{ij}}\right) \quad \text{and} \quad D_i = \left(\sum_{j=1}^{n} \left(\frac{z_{ij}}{s_{ij}}\right)\right)^2 \left(\sum_{j=1}^{n} \left(\frac{1}{s_{ij}}\right)\right)^{-1} \tag{10}$$

Since $\left(\sigma_{ij}^2 | s_{ij}\right) = \frac{s_{ij}}{a_i}$ we have:

$$E\left(\sigma_{ij}^2 | s_{ij}, (z_{i1}, s_{i1}), ..., (z_{in}, s_{in})\right) = s_{ij} E\left(\frac{1}{a_i}\right) = \frac{s_{ij}\,(G_i - D_i)}{n-3} = s_{ij}\alpha_i, \tag{11}$$

where $\alpha_i = \frac{(G_i - D_i)}{n-3}$ and $n > 3$ The s_{ij} is given by $s_{ij} = \frac{1}{p_{ij}^2}$, where p_{ij} is the quality of score assumed to be provided by the expert on its own score x_{ij}. Consequently we use $\bar{\sigma}_{ij}^2 = E(\sigma_{ij}^2 | s_{ij})$

$$\bar{\sigma}_{ij}^2 = \frac{\alpha_i}{p_{ij}^2} = \frac{(G_i - D_i)}{n-3} \cdot \frac{1}{p_{ij}^2} \tag{12}$$

in (2) and (4) instead of σ_{ij}^2.

Basics of the supervisor algorithm

1. (Supervisor Training) Estimate the bias parameters of each expert, i.e. $\{M_i, V_i, \alpha_i\}$, by using a training set i.e. x_{ij}, y_j, and p_{ij} with j up to n, (2). σ_{ij}^2 is computed according to (12), and (10).
2. (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 $x_{i,n+1}$, and $p_{i,n+1}$, but not access to the true authentication scores, y_{n+1}. The expert opinions are normalized yielding M', and V', (4). $\sigma_{i,n+1}^2$ is computed according to (12). M'' and/or V'' are computed according to (6) (and are ready to be thresholded to yield a definite decision).

Score transformation

Depending on the algorithms used, the scores of the experts, X_{ij}, may or may not be dimensionless (scaled) or in the correct range i.e. $[-\infty, \infty]$. Below, the prime on X and Y variables represent these variables before transformation.

For our purposes, the transformation

$$X_{ij} = \log \frac{X'_{ij}}{1 - X'_{ij}} \tag{13}$$

which is also known as the "odds of X'_{ij}", will be used to map to scores to $[-\infty, \infty]$. It can be shown that the formulas (2,4,6,11) still hold when X_{ij} is substituted by $X'ij$, and Y_j by $Y\prime_j$. The only difference is in the conditional distribution of Y_{n+1}, (5), 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 the 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 $bs_{n+1} = M''$ where bs_{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.

4 Experiments

Even when the experts are machine components, it is quite expensive to obtain a multitude of expert opinions. We have therefore conducted one set of experiments based on simulated expert opinions, and one set of experiments using true expert scores. The purpose of the first set of experiments was to conclude on the general performance of the supervisors, to different settings of parameters such as to the the number of experts, the skills of experts, the size of the training set, ...etc. The second set of experiments, which obviously had to be more limited, 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 quality of the scores, it was not possible to obtain these, such that they consistently fit to our interpretation, from our two real experts. We have therefore, used the same quality of score for all shots, simulated or real. This has the effect that the supervisor calibrates the experts by using their historical success alone.

In the case of simulations the scores of an expert i, X'_{ij}, were drawn from two uniform distributions defined on the open intervals $(0, I_j)$, $(C_j, 1)$. 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, here with probabilities 0.8 and 0.2 respectively. This is because in real conditions, the experts have access to a limited number of shots of the same person which results in that the number

of clients is smaller than the number of impostors, both for training and test sets. In order to simulate the different skills of experts, I_j and C_j were varied. We imposed the conditions $0.5 < I_j < 1$ and $0 < C_j < 0.5$, the non-fulfillment of which would yield perfect experts. We picked I_j and C_j at random (uniformly) from the intervals $(0.5, 0.75)$ and $(0.25, 0.5)$ for systems with many experts. Since our expert scores are in the interval $[0, 1]$ we utilized the odds-transformation. In practice, in order to avoid singularities, we had to force our X'_{ij}'s not to come very close to the interval ends that is we forced them to be in $[\epsilon, 1-\epsilon]$ where we have chosen $\epsilon \approx \exp(-6)$ yielding the odds -6, and 6 for 0 and 1. 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 small choices of ϵ.)

It is crucial to observe that 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 straight forwardly. To obtain a mono modal density, we build two supervisors, one which is specialized on impostors and one specialized on clients. That is step 1 of the algorithm is applied to the impostors and the clients of the training set separetaly, yielding two bias parameter sets, $\{M_i^c, V_i^c, \alpha_i^c\}$ and $\{M_i^I, V_i^I, \alpha_i^I\}$. The step 2 is then applied twice for each of these parameters. The result of the two supervisors can be combined further by using another supervisor, which does not need to be a complex one. In our case we chose the response of the supervisor which came closest to its goal, -6 or 6 as the final M''^c. That is, if

$$|6 - M''_c| - | - 6 - M''_I| > \epsilon_2 \tag{14}$$

$M'' = M''^I$, otherwise $M'' = M''^c$. We used $\epsilon_2 = 0$ but all $\epsilon_2 \in [-0.1, 0]$ gave no significant change of the results. The false acceptance rates of our simulated experts, given by $FA_i = I_i - 0.5$, and those of false rejection, given by $FR_i = 0.5 - C_i$, are varied. Had our experts' opinions been derived from real experts, we would not be able to have this flexibility.

We will work with the histograms of z such as

$$z^{bs}_{n+1} = y_{n+1} - bs_{n+1}$$

where the bs_{n+1} is the output of our Bayesian supervisor, which is the final M'' for a new claim, $n+1$.

For all experiments, the number of shots in the training set, i.e. the pairs, $Y_j, X_{i,j}$, was $n = 2664$ and the number of shots in the test set, i.e. the pairs $Y_{n+1}, X_{i,n+1}$, was $n' = 7992$. As a comparison we also simulated a simple supervisor with its scores consisting of estimates of the mean values of expert scores, $ms'_{n+1} = \frac{1}{m}\sum_{i=1}^{m} X'_{i,n+1}$. The ms_{n+1} used subsequently is the odds of ms_{n+1} computed according to (13), for the sake of comparisons.

bs_{n+1} yielded consistently better success rates, SR=1-(FA+FR), than ms_{n+1}. That is few experts (e.g. 2), or many experts, (4 and more), were always conciled better by bs_{n+1} than ms_{n+1}. In case of equally skilled experts ms_{n+1}

SR_1	SR_2	SR_3	SR_4	SR_{bs}	SR_{ms}
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_{bs}	SR_{ms}
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. Left: Row 1, equal skills; Row 2, random skills. Right: Row 1, 2 experts case; Row 2, 4 experts.

was outperforming significantly the individual experts. The improvement offered by ms_{n+1} was in the average marginal when the experts had unequal skills, meaning that often it performed worse than the best expert. A typical decision result, at the threshold $T = 0$ for bs and ms, illustrating these are given by Table 1 in which 4 simulated expert opinions were used. We note that by using a threshold T one can decide that if for example $bs_{n+1} < T$ then the person is an impostor, client otherwise. Since it represents a realistic scenario, we will report in a more detailed fashion about the case: *Randomly skilled* experts, $FA_i + FR_i \neq FA_{i'} + FR_{i'} \quad i \neq i'$, with *unsymmetric* skills $FA_i \neq FR_i$.

Figure 1 illustrates the density curve (estimated by a 512 bin histogram) of $z = y_{n+1} - ms_{n+1}$ of a simulation. 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 figure for the Bayesian supervisor is given by Figure 1. In both cases 2 experts were employed. When 4 experts were utilized the density curves in Figure 2 correspond to $Z^{ms}_{n+1} = y_{n+1} - ms_{n+1}$, and $Z^{bs}_{n+1} = y_{n+1} - bs_{n+1}$. In all 4 simulations expert skills were random. The success rates corresponding to these 4 figures, at $T = 0$, are given by Table 1 These 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.

Finally in Figure 3 we present the miss-identification densities of the Mean Value and the Bayesian supervisor when scores of two real machine modules are utilized. The number of shots in the training and test sets were the same as before. The machine modules were experts in authentication of frontal faces and speech. The Bayesian supervisor was trained on scores obtained by testing the image and the speech experts on the M2VTS database, see Pigeon and Vandendorpe [11]. The algorithms constituting their expertise are described in Duc et. al., [4]. The success rates at $T = 0$ are given by the table in Figure 4. As the bimodal nature of z curves indicates, the experts must be equally and highly skillful in order to guarantee a good Mean Value supervisor performance which is not necessary for the Bayesian supervisor.

In real experts situations the individual experts' skills are not easy to measure resulting in that the decision threshold is varied yielding FR, and FA curves. Normally at the expert level, by using a threshold T, the inequality $X_{i,n+1} > T$ can be turned to a decision of *accept* when fulfilled, a decision of *reject* otherwise. We apply the same principle to obtain a supervisor decision and rewrite the

inequality by subtracting it from y_{n+1} as $Z_{i,n+1} < y_{n+1} - T$. The supervisor decision represents an acceptance decision when this inequality is fulfilled, rejection otherwise. Therefore an acceptance decision is a false acceptance when the inequality is fulfilled for $y_{n+1} = -6$. Likewise a rejection decision represents a false rejection when the inequality is *not* fulfilled for $y_{n+1} = 6$, i.e. $Z_{i,n+1} \geq y_{n+1} - T$. Consequently, the left masses excluded by a moving window of size 12 applied to a z histogram represents the values of an FA curve of a supervisor (or an expert). Likewise, the excluded right masses represent the values of the FR curve. The corresponding FA and FR curves for both ms and bs supervisor are given by Figure 4. The curves of the two supervisors had originally different T scales (x-axis). In order to allow for comparison the T scale of the mean supervisor is mapped linearly so that its minimum and maximum coincide with those of the Bayesian supervisor. It can be observed that the Bayesian supervisor has a larger interval of threshold where both FA and FR are extremely small, ≈ 0.005, as compared to the Mean Value supervisor. This property is important for robust thresholds.

These and other experiments indicate that the Bayesian supervisor is more successful in decision making due its capability of symmetrizing the miss-identification densities.

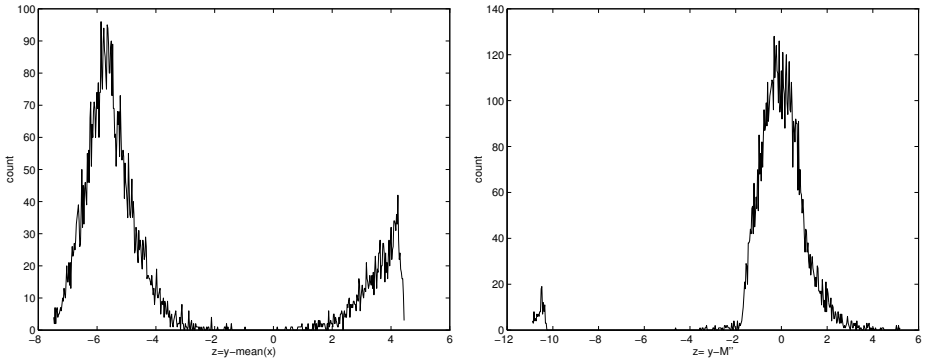


Fig. 1. The miss-identification density of the Mean Value supervisor (Left), and the Bayesian supervisor (Right) in the case of *2 unequally skilled experts*

5 Conclusions

We calibrated the experts' authenticity scores of a future instant about a client only by means of the earlier made identification errors. But we may even calibrate scores about a particular client. This work is straightforward. 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, [3].

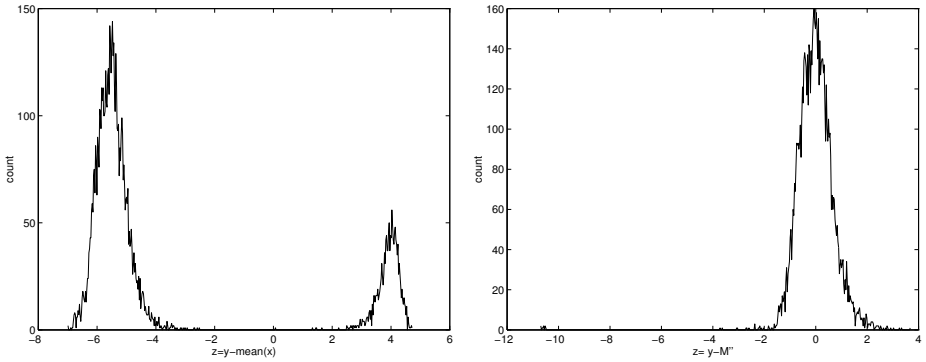


Fig. 2. The misidentification density of the Mean Value supervisor (Left), and the Bayesian supervisor (Right) in the case of *4 unequally skilled experts*

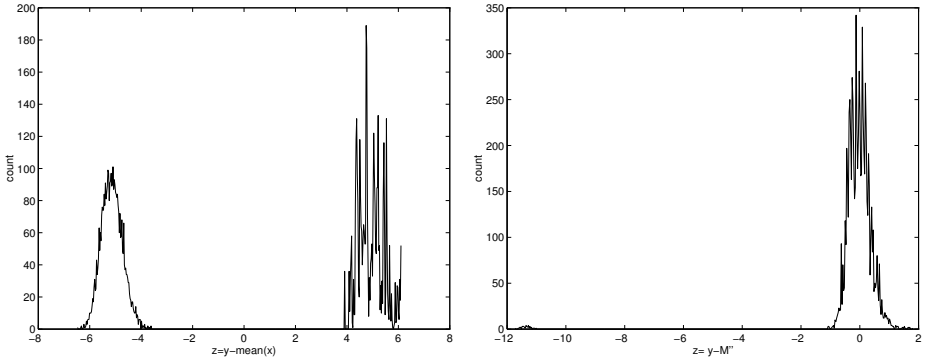
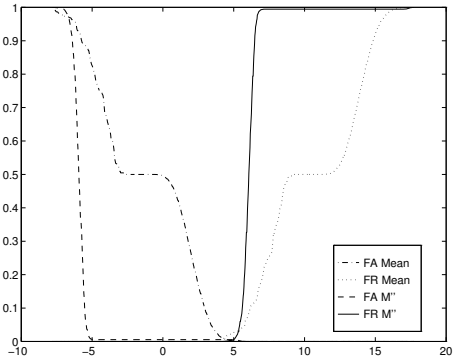


Fig. 3. The miss-identification density of the Mean Value supervisor (Left), and the Bayesian supervisor (Right) in the case of *2 real experts: face and speech.*



SR_1	SR_2	SR_{bs}	SR_{ms}
0.944	0.965	0.995	0.980

Fig. 4. Left: FA and FR curves for Mean Value and Bayesian supervisor using image and speech experts. Right: Success rates at $T = 0$.

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 in (14) for merging the impostor versus client predictions could be investigated.

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.

We presented a framework for simulating expert decisions and shown that it approximates the real conditions reasonably well.

Acknowledgement

We gratefully acknowledge the use of speech expert opinions provided by IDIAP, Switzerland, in particular G. Maitre for his assistance in transfer. We also thank H. Bigün, KTH Sweden, for constructive discussions, J. Wallander and T. Hedelius foundations, as well the European project M2VTS for financial aid.

References

1. J. M. Bernardo and M. F. A Smith. *Bayesian Theory.* Chichester. Wiley, 1994.
2. E. S. Bigün. "Risk analysis of catastrophes using experts' judgements: An empirical study on risk analysis of major civil aircraft accidents in Europe". *European J. Operational research*, Vol. 87, pp. 599–612, 1995.
3. E. S. Bigün. "Bayesian prediction based on few and dependent data". Technical report, Department of Statistics, Stockholm University, 1996.
4. B. Duc, G. Maître, S. Fischer, and J. Bigün. "Person authentication by fusing face and speech information". *Proc. AVBPA*, Springer LNCS, Bigün, et. al., Eds., 1997.
5. S. French. "Updating of belief in the light of some else's opinion". *J. R. Statist. Soc. A*, Vol. 143, pp. 43–48, 1980.
6. S. French. "Group consensus probability distributions: A critical survey". *Bayesian statistics*, Vol. 2, pp. 183–202, 1985.
7. V. D Lindley, A. Tversky, and R. V. Brown. "On the reconciliation of probability assessments". *J. R. Statist. Soc. A*, Vol. 142, pp. 146–180, 1979.
8. V. D Lindley. "The improvement of probability judgments". *J. R. Statist. Soc. A*, Vol. 145, pp. 117–126, 1982.
9. V. D Lindley. "Reconciliation of discrete probability distributions". *Bayesian statistics*, Vol. 2, pp. 375–390, 1985.
10. V. D Lindley and Singpurwalla. "Reliability and fault tree analysis using expert opinions". *ASAS*, pp. 87–90, 1986.
11. S. Pigeon, and L. Vandendorpe *The M2VTS Multimodal Face Database Proc. AVBPA*, Springer LNCS, Bigün, et. al., Eds., 1997.
12. M. West. "Modelling expert opinion". *Bayesian statistics*, Vol. 3, pp. 493–508, 1988.
13. R. L. Winkler *Combining probability distributions* Management Science, vol. 27 pp. 479-488, (1981)