# Applying Logistic Regression to the Fusion of the NIST'99 1-Speaker Submissions

Stéphane Pigeon, Pascal Druyts, Patrick Verlinde
*SIC Laboratory, Royal Military Academy, Av. Renaissance 30, 1000 Brussels, Belgium.*
E-mail: {stephane.pigeon,pascal.druyts}@elec.rma.ac.be; patrick.verlinde@tele.rma.ac.be

This contribution formulates the decision fusion problem encountered in the design of a multi-expert identity verification system. Logistic regression is introduced as a particular case of the naive Bayesian classifier and is applied to the fusion of all submitted data of the 1-speaker recognition task, part of the NIST'99 campaign.

*Key Words:* decision fusion; Bayes; logistic regression.

## 0. INTRODUCTION

The automatic verification of a person's identity is becoming an important task in several applications, especially in the field of automatic access to restricted physical or virtual environments. Passwords, personal magnetic cards and PIN-numbers are already widely used in this context. Although they are quite convenient to use, they can be forgotten, lost or stolen. Therefore, a new kind of method is emerging, based on so-called *biometric* measures such as vocal (speech), visual (face, profile), fingerprint or any other information that physically characterizes the person to be identified. A system based on a particular set of biometric features is referred to as an *expert* for that biometric feature set. In order to gain in robustness, multimodal systems tend to combine several experts together, like the frontal, facial and speech experts used in [10]. In our contribution, all experts which will be used are taken among the NIST'99 evaluation campaign submissions, and thus refer to the same speech modality. Although we cannot expect as much improvement by fusing speech-related experts as compared to the fusion of independent biometrics, it is interesting to see how fusion improves the overall system performance, even in such a particular case as the one we are working with in this paper.

The first Section introduces the decision fusion framework adopted in this work. Section 2 reminds the reader about the general Bayesian theory. Section 3 introduces the independence hypothesis that will be used in order to derive the logistic regression model presented in Section 4. Section 5 illustrates the influence of the a priori probabilities that are inherent to the Bayesian approach. Section 6 presents

the results of fusing all NIST'99 speech experts and, finally, Section 7 concludes this work.

## 1. DECISION FUSION IN AN IDENTITY VERIFICATION SYSTEM

The purpose of an identity verification system is to decide whether someone claiming the identity of a registered user is indeed that client, or an impostor. In a mono-modal system, this is done by comparing the score obtained for that person with a decision threshold. Such a system can make two types of errors: (1) reject a client (i.e. *False Rejection* – FR – or *Miss*) and (2) accept an impostor (i.e. *False Acceptance* – FA – or *False Alarm*). The performance of a speaker verification system is usually given in terms of global error rates computed during tests, namely the *False Rejection Rate* (FRR – the number of FR divided by the number of client claims) and the *False Acceptance Rate* (FAR – the number of FA divided by the number of impostor claims) [1]. The *Equal Error Rate* (EER) stands for the operating point in which the FAR and FRR are the same.

One possible and straightforward way of building a multi-modal verification system from $n$ such mono-modal systems is to input all $n$ scores provided in parallel into a fusion module which has to take the decision to *accept* or *reject* the claim. This is a typical *decision fusion* approach, in which the fusion module receives as input the *decisions* issued by the several individual experts, and typically has neither access to the input feature vectors of these experts (*feature fusion*), nor to the original raw data streams (*data fusion*) [3]. Once the choice of a particular fusion scheme has been made, two main alternatives still remain for the fusion module: a global (i.e. the same for all persons) or a personal (i.e. tailored to the specific characteristics of each authorized person) approach. For the sake of simplicity and since the personal approach requires much more training data, we opted for a global fusion module. As, in a verification system dealing with $n$ modalities, the fusion module has to realize a mapping from $\Re^n$ into the binary set $\{reject, accept\}$, this can be seen as a multi-dimensional classification problem, splitting a n-dimensional space into two classes. Bayesian classifiers will be introduced in the next section.

## 2. BAYESIAN FRAMEWORK FOR DECISION FUSION

In a number of references such as [4], a general overview of Bayesian decision theory is presented in the case of the classification problem. We will give here only a brief overview of the most important results in the specific case of a two-class problem. These two classes will be denoted by $C_i$, i=1,2; with $C_1$ and $C_2$ respectively denoting the clients (targets) and the impostors (non-targets).

Let $X$ be a random observation coming from either classes. In the most general case $X$ will be a multi-dimensional feature vector constructed by the concatenation of all feature vectors $\vec{M}_k$, given to all $n$ experts ($k = 1, \ldots, n$). The decision problem is to correctly classify each observation in its respective class. To measure the performance of a classifier we define a *loss function* $l_{ji}$, which gives the cost of classifying a class $i$ observation into a class $j$ event. As an example, we may opt

for the zero-one loss function defined by:

$$l_{ji} = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{if } i \neq j \end{cases} \tag{1}$$

which assigns no loss to correct classification and a unit loss to any error, regardless of the class. Under such assumptions, it can be shown that the optimal classifier, defined as the one that achieves the *minimum classification error probability*, is the classifier that implements the *maximum a posteriori* (MAP) decision rule, i.e. maximizing $P(C_i|X)$, the conditional probability of $C_i$ given $X$. The minimum error rate achieved by this optimal classifier is then called the *Bayes risk*. Using Bayes rule, the a posteriori probability can be rewritten as:

$$P(C_i|X) = \frac{P(X|C_i).P(C_i)}{P(X)} \tag{2}$$

where $P(C_i)$ and $P(X)$ are the *a priori* probabilities of $C_i$ and $X$ respectively. Since $P(X)$ does not depend on the class index, the MAP decision only depends on the numerator of the right-hand side of the previous equation:

$$MAP = \max_i P(X|C_i).P(C_i) \tag{3}$$

By assuming that the a priori probabilities are equal for both classes (this is a strong assumption, which is going to be discussed later), the MAP decision rule reduces to a maximum conditional probability (MCP) rule. $P(X|C_i)$ is often called the *likelihood* of $X$ given $C_i$ and a decision that maximizes $P(X|C_i)$ is hence also called a *maximum likelihood* (ML) decision.

$$MCP = ML = \max_i P(X|C_i) \tag{4}$$

In a multi-expert decision fusion context, each expert $k$ has access to a feature vector $\vec{M}_k$. As developed above, the final decision should be based on $P(C_i|X)$ or, by expanding explicitly $X$, on $P(C_i|\vec{M}_1, \ldots, \vec{M}_n)$ which implies the direct use of the feature vectors. This might be burdensome to deal with or even impossible to implement in some practical cases. This also means that we deny the pertinence of the experts, by bypassing their opinion. As the theory states that the optimal classification should be based on $P(C_i|\vec{M}_1, \ldots, \vec{M}_n)$, we nevertheless need a way to obtain the best estimate of these probabilities. A brute force approach, in which one tries to estimate directly the above probabilities using for instance Multi Layer Perceptrons (MLPs), might be appealing because it could lead to the optimum decision and it does not rely on any of the hypotheses that we will introduce in the next Sections. However, due to the limited amount of data available at training time, MPLs (and other methods as well) often result in *rough estimates* of the *real* probabilities. These estimates could be so bad that they would become useless. Therefore, one usually prefers to *approximate the real probability functions* first (e.g. the logistic regression approximation introduced in Section 4), then find *correct estimates* for these approximations. In other words, correctly estimating approximate probability distributions often yields to better results than approximating exact distributions.

As mentioned in Section 1, the objective of this paper is to issue the best decision, based on the (scalar) output *scores* $s_k$ of the available experts, and not on their input *feature vectors*. If expert input measures are conditionally independent of the class, the probabilities $P(C_i|\vec{M}_1,\ldots,\vec{M}_n)$ needed may be computed by combining the $n$ different $P(C_i|\vec{M}_k)$. This implies that if the expert returns $s_k = f\{P(C|\vec{M}_k)\}$ where $f\{\cdot\}$ is any monotonic function, then no loss of information is introduced by using the score $s_k$ instead of the feature vector $\vec{M}_k$ [5]. The reason why an expert should return $f\{P(C|\vec{M}_k)\}$ may be explained by considering the design goal of the expert designer. Indeed, even if the development is not probabilistically driven, the objective is usually to output higher degrees of confidence for higher probabilities. If the above conditions are fulfilled, $f\{\cdot\}$ can be inverted to provide the desired probabilities. This transformation is called the *calibration* of the expert [6]. In fact, reality is more complicated than this, because the output of an expert is in general corrupted by an estimation error.

The main advantage of the Bayesian approach is that it leads to the optimal classifier, in the sense that it implements the lowest Bayes risk. There are however a number of problems with this approach. The most important problem is that the probability density functions (pdfs) have to be estimated correctly. This usually implies the selection of a structure (i.e. a class of functions) for the approximator first, and then the optimization of the free parameters to best fit the pdf. This optimization is performed on a training set. The plasticity (i.e. degree of freedom) of the approximator has to be chosen carefully. For highly plastic approximators, quite general pdfs may be approached, but an important (often impossible to obtain) number of samples is needed for performing the training. Furthermore, the training set should be representative (which in general does not correspond to the equal a priori probability hypothesis previously made) and over-training has to be avoided to reach good generalization [2]. On the other hand, by using an approximator with limited plasticity (few parameters), fewer examples are needed but more a priori knowledge is intrinsically encoded by limiting the possible solutions. Poor a priori knowledge will lead to bad results. In practice, the best compromise should be sought, but the true MAP or MCP decision rules can not be implemented most of the time and the theoretical minimal Bayes risk remains an unachievable lower bound. In Section 4, we will suppose that the probability distributions involved are members of the exponential family with equal dispersion parameters (the logistic regression model). The first step towards deriving this specific case is to introduce the hypothesis of independence between experts. This transforms the general Bayesian approach presented above into the so-called *naive Bayes classifier* [9], introduced in the next Section.

## 3. THE NAIVE BAYES CLASSIFIER

From now on, we will suppose that the different experts are independent from each other. This can be formalized by the following hypotheses:

$$h1 : P(s_1,\ldots,s_n|C) = \prod_{k=1}^{n} P(s_k|C) \tag{5}$$

$$h2 : P(s_1, \ldots, s_n|I) = \prod_{k=1}^{n} P(s_k|I) \tag{6}$$

where $C$ and $I$ stand for the client and the impostor classes respectively.

The first hypothesis may seem a bit counter-intuitive and indeed scores are correlated, which implies that $P(s_1, \ldots, s_n) \neq \prod_{k=1}^{n} P(s_k)$. This correlation, however, is reduced when *conditional* probabilities are considered, like in $h1$. To illustrate this, one may imagine a case where the score $s_k$ varies between 0 (unconditional rejection) and 1 (unconditional acceptance). Suppose the score obtained by a given expert is close to one. If this first expert is a good expert, this means that the identity claim has already a high probability of being true. Therefore the scores coming from the other experts will also be situated also in the vicinity of one. This leads to a significant correlation. However, it is more reasonable to believe that the *deviations* from one prototype ($s_k$ for an impostor, $1 - s_k$ for a client) are uncorrelated for the various experts.

The justification made for hypothesis $h1$ is not sufficient for $h2$. Indeed, for a specific identity claim the class of impostors contains many persons whereas the class of clients contains only one person (the client whom the person under test claims to be). The scores provided by the experts could give insight into the identity of the impostor. In such a case, the scores would be highly correlated to the identity of the impostor and therefore correlated for the different experts. However, experts are usually designed to decide whether the person under test is a client and don't care about the identity of a possible impostor. This makes $h2$ more reasonable.

Under these two hypotheses, it can be shown that [12]

$$P(C|s_1, \ldots, s_n) = \frac{1}{1 + e^{-\{(\sum_{k=1}^{n} x_k) + x_0\}}} \tag{7}$$

where

$$x_0 = ln\frac{P(C)}{P(I)} \tag{8}$$

$$x_k = ln\frac{P(s_k|C)}{P(s_k|I)} \tag{9}$$

with $s_k$ being the scalar score given by the $k-$th expert.

## 4. THE LOGISTIC REGRESSION MODEL

A particular instance of the general framework of Section 2, can be obtained by assuming that for each expert, probabilities are members of the exponential family:

$$P(s_k|C) = f(s_k).e^{(C_k.s_k + C_{k0})} \tag{10}$$

$$P(s_k|I) = f(s_k).e^{(I_k.s_k + I_{k0})} \tag{11}$$

Using this, it is easy to see that equations (7), (8) and (9) reduce to

$$P(C|s_1, \ldots, s_n) = \frac{1}{1 + e^{-g(s)}} = \pi(s) \tag{12}$$

where

$$g(s) = \beta_0 + \beta_1.s_1 + ... + \beta_n.s_n, \quad (13)$$

$$\beta_0 = \sum_{k=1}^{n}(C_{k0} - I_{k0}) + ln\frac{P(C)}{P(I)}, \quad (14)$$

$$\beta_k = C_k - I_k\,. \quad (15)$$

This function is known as the logistic regression (LR) model or logistic distribution function [7, 13]. Note that the class of conditional probabilities, as defined by equations (10) and (11), is known as the exponential family *with equal dispersion parameters* for the clients and impostors [8]. One particular case of this family is the well-known Gaussian distribution with equal variance (the quadratic exponent is generated thanks to the $f(s_k)$ multiplier), which transforms equations (10) and (11) into:

$$P(s_k|C) = \frac{1}{\sqrt{2\pi}.\sigma_k}.e^{-\frac{(s_k-\mu_k^C)^2}{2\sigma_k^2}} \quad (16)$$

$$P(s_k|I) = \frac{1}{\sqrt{2\pi}.\sigma_k}.e^{-\frac{(s_k-\mu_k^I)^2}{2\sigma_k^2}}, \quad (17)$$

where $\mu_k^C$ and $\mu_k^I$ represent the mean of respectively the client and impostor classes and $\sigma_k^2$ their common variance. In this particular case, equations (14) and (15) may be rewritten as:

$$\beta_0 = \sum_{k=1}^{n}\frac{(\mu_k^I)^2 - (\mu_k^C)^2}{2\sigma_k^2} + ln\frac{P(C)}{P(I)}, \quad (18)$$

$$\beta_k = \frac{\mu_k^C - \mu_k^I}{\sigma_k^2}, \quad (19)$$

$\beta_k$, the weight given to the $k$-th expert, is nothing else than the difference of the means of the distributions for the two classes for the $k$-th expert, divided by their common variance. This is in accordance with our intuition which says that an expert performs better when the distributions relative to the clients and impostors are more separated and when their variance is smaller.

It is also interesting to contrast the exponential family with shared variances to the Gaussian case with different client and impostor variances. Strictly speaking, imposing the same variance for both classes might be seen as a restriction compared to the Gaussian approach. However, a fusion scheme based on the exponential model appears to be more flexible and robust [8]. This improved flexibility may be explained by the fact that one does not specify a *particular* distribution when assuming that the class-conditional pdfs are members of the (same) exponential family. The increase in robustness results from the fact that the approach based on members of the exponential family with equal dispersion parameters requires fewer parameters to be estimated compared to the Gaussian case: $n+1$ parameters instead of $4n$ ($\mu_k^i$ and $\sigma_k^i$ for each class $i$ and for each expert $k$) or $3n$, if variances are equal.

By minimizing the classification errors on the training data, one can estimate the various $\beta_k$ parameters. During the test phase, an unknown test pattern will be classified among the clients if $\pi(s)$ is greater than the optimal theoretical threshold (0.5 or any other threshold if one wants to take into account the effects of $l_{ij}$, $P(C)$ and $P(I)$).

One of the advantages of the LR is that the parameters $\beta_k$ are a direct measure of the relative importance of the expert $k$ (assuming that all expert scores have been previously normalized to the same scale). This interesting property allows the designer to identify the most relevant experts very easily, without using Principal Component Analysis or other less convenient methods as suggested in [11].

## 5. THE ISSUE OF THE A PRIORI PROBABILITIES

We have shown in Section 2 that in the Bayesian framework, the optimal decision is a function of the a priori probabilities. These probabilities may be fed in explicitly (e.g. the MAP rule) or may be learned from a training database (e.g. LR model). The client/impostor frequencies of occurrence faced during a system's operational deployment should ideally be used, but is often unknown. The frequency estimated from the training database will often be biased as impostor claims are more numerous than client claims (for a $p$-person database, one usually deals with $p$ client tests and $p(p-1)$ impostor tests). Depending on the learning scheme, the poorly represented class will only have a week contribution to the error function at training time. This may seriously bias the system if the a priori probabilities relative to the training set are very different from the operational ones.

This argument is often used to criticize the Bayesian approach and to promote other methods which do not require these probabilities. In our opinion, however, the optimal decision *does* depend to a greater or lesser extent on the a priori probability. This may be illustrated by the following example. If a man has to classify a person moving in the dark in his home using only visual information, he will probably classify the person as his partner, although no detail is visible. The a priori probability for a person moving in the house was high for the partner. If the person was in reality a burglar, one could easily be misled. On the other hand, that same man will have no difficulties using only visual information by daylight to classify a person moving in the house as a burglar, even if the a priori probability for the partner is higher. This hypothetical example shows that if the measurements are sufficiently discriminative, we do not need to rely on a priori probabilities. On the other hand, if the measurements are not discriminative enough, we do need the a priori probabilities. This effect may be understood mathematically by analyzing equation (7). The discrimination power of the measures is represented by $\sum_k x_k$, whereas the a priori probabilities are represented by $x_0$. It is only in the case that the first term is significantly bigger than the second one that the a priori probabilities have no effect on the global sum.

## 6. NIST'99 FUSION EXPERIMENTS

LR has been applied to the fusion of all NIST'99 1-Speaker submissions (12 experts in total, all conditions mixed). In order to ensure distinctive training and testing sets, LR has been trained on the male subset of the expert test data (1311

client claims, 14617 impostor claims), while testing the logistic regression has been performed on the female subset (1846 client claims, 19846 impostor claims). At training time, optimal $\beta$ coefficients (equation 14 and 15) are found by minimizing the classification errors. This reverts to maximizing a log-likelihood function (llk) as the one described below

$$llk = \sum_{clients} log(\pi(s)) + \sum_{impostors} log(1 - \pi(s)) \tag{20}$$

where $\pi(s)$ is given by equation (12). Once the optimal $\beta$ coefficients are found, test samples are classified by comparing $\pi(s)$ – the a posteriori probability of being a client – with a given threshold. By continuously varying this threshold, one obtains a DET curve as represented in Figure 1. This figure provides all achievable tradeoffs between the FAR and FRR for both individual experts (dashed lines) and the LR system (continuous line), in test conditions. The operating points relative to the theoretical 0.5 threshold are marked by a "+" sign. These points are located relatively low in FA and high in FR due to the nature of equation (20), which does not compensate for the (one magnitude) higher number of impostors compared to the clients. In other words, equation (20) implicitly optimizes the system taking into account the nearly ten times higher impostor a priori probability. This effect may be overcome by decreasing the decision threshold below 0.5. As seen in Figure 1, even under these very unfavorable circumstances (12 vocal experts that are probably highly correlated), LR improves the performance of the system by 30% around the EER operating point compared to the performance achieved by the best expert (EER=7% instead of EER=10% for the best expert).

Moreover, Figure 2 provides the DET curves obtained when fusing the three most uncorrelated experts found among the twelve available (the correlation is computed from the training data), while Figure 3 shows the performance achieved by the three best experts (i.e. the experts which achieved the best performance at training time). As one can see, fusing the three best experts performs better than fusing the most uncorrelated ones. This can be explained by the fact that the most uncorrelated experts are made from the very best contribution... and the two worst ones! Although being uncorrelated, the worst contributions cannot add much (pertinent) information to the best one. To some extent, this may be similar to the combination of a good expert with random score generators. Although being independent from each other, no fusion gain can be foreseen in such a case. In other respects, no one will ever get a good system by fusing random score generators...

This leads to the following remark: although correlation between experts should be as small as possible, it only makes sense when experts are characterized by the same performance level. Then, the lower the correlation, the higher the fusion gain. On the other hand, when dealing with highly heterogeneous experts, fusion results will much more depend on the performance level achieved by the best expert, rather than on correlation issues.

## 7. CONCLUSION

This paper showed how to derive the logistic regression model from the general Bayesian theory and discussed the influence of the a priori class probabilities on

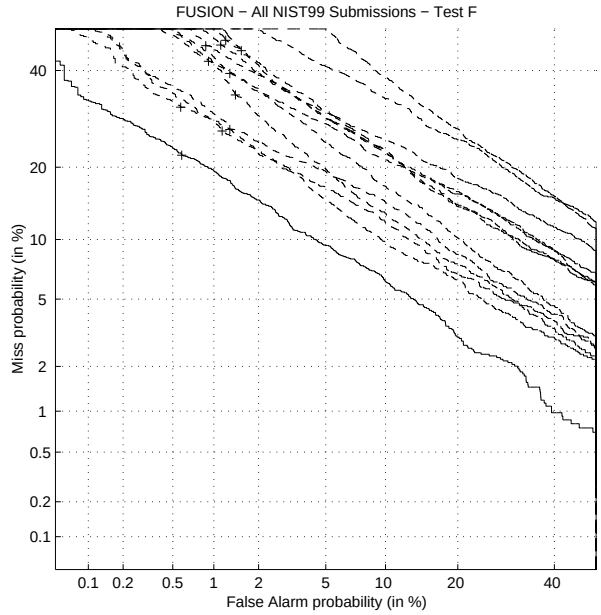


**FIG. 1.** DET curve relative to the TEST condition (Individual experts in dashed lines, fusion in solid, "+" signs refer to the 0.5 decision threshold).

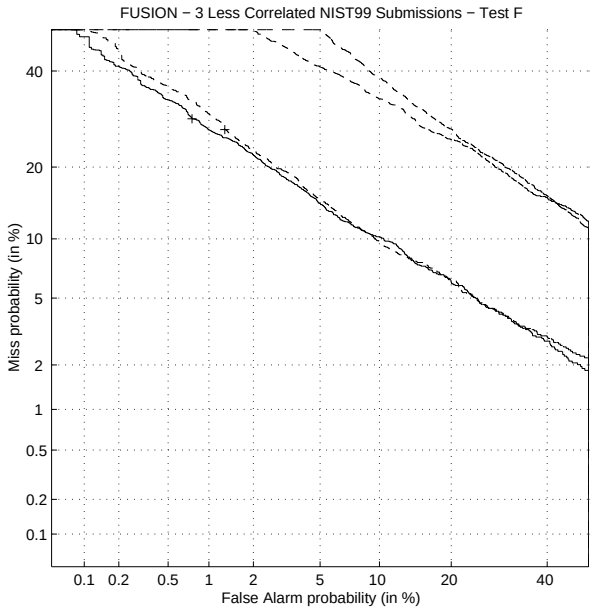


**FIG. 2.** DET curve relative to the fusion of the 3 less correlated experts (Test condition, Individual experts in dashed lines, fusion in solid, "+" signs refer to the 0.5 decision threshold).

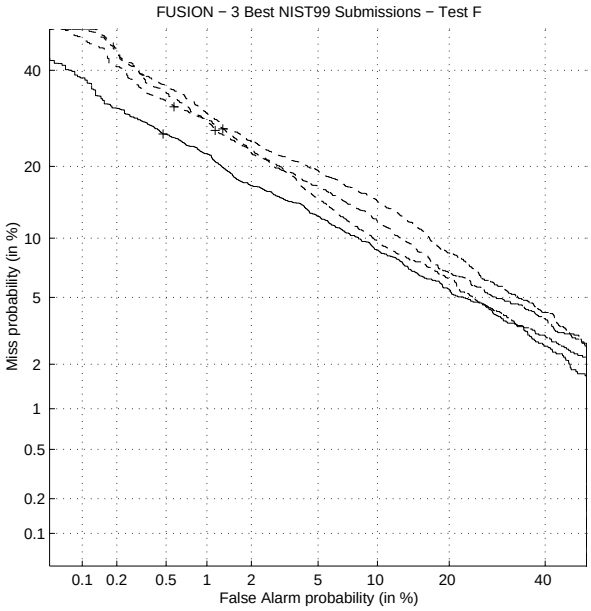


**FIG. 3.** DET curve relative to the fusion of the 3 best experts (Test condition, Individual experts in dashed lines, fusion in solid, "+" signs refer to the 0.5 decision threshold).

the decision threshold. The logistic regression approach is based on an independence hypothesis between experts, and has been successfully used in the past for fusing several image and speech experts together [13]. Since the NIST submissions used here are all speech related and thus strongly correlated, one could fear that the suggested method would not fit this particular case. Nevertheless, logistic regression showed a non-negligible improvement compared to the best individual speech expert, despite the non-optimal working conditions. Thanks to a rather straightforward implementation, logistic regression may be thus recommended as an easy-to-implement and yet efficient method for fusing independent as well as dependent experts.

## 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. C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford UK, 1995.
3. B. V. Dasarathy. *Decision Fusion*. IEEE Computer Society Press, 1994.
4. P. A. Devijver and J. Kittler. *Pattern Recognition: A Statistical Approach*. Prentice-Hall Inc., Englewood Cliffs NJ, 1982.
5. P. Druyts and M. Acheroy. A modular multi-layer perceptron (MMLP) to identify objects composed of characteristic sub-parts. In Dagli, Akay, Fernandez, Erosy, and Smith, editors, *ANNIE'97*, University of Missouri, USA, 1997. ASME Press.
6. S. French and J. Q. Smith. *The Practice of Bayesian Analysis*, chapter Bayesian analysis. Edward Arnold, London, 1997.
7. D. W. Hosner and S. Lemeshow. *Applied Logistic Regression*. John Wiley & Sons, 1989.

8. M. I. Jordan. Why the logistic function? A tutorial discussion on probabilities and neural networks. Computational Cognitive Science 9503, Massachusetts Institute of Technology, Cambridge MA, August 1995.
9. T. M. Mitchell. *Machine learning*. Mc Graw-Hill, 1997.
10. S. Pigeon. *Authentification multimodale d'identité*. PhD thesis, Université Catholique de Louvain, February 1999.
11. B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.
12. P. Verlinde. *A contribution to multi-modal identity verification using decision fusion*. PhD thesis, Ecole Nationale Supérieure de Télécommunications, Paris, France, 1999.
13. 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 *Proceedings of the Second International Conference on Audio- and Video-based Biometric Person Authentication*, pages 188–193, Washington D. C., USA, March 1999.