A MULTI-SAMPLE MULTI-SOURCE MODEL FOR BIOMETRIC AUTHENTICATION

Norman Poh, Samy Bengio and Jerzy Korczak
LSIIT, ULP-CNRS, Bld S. Brant, 67400 Illkirch, France
Dalle Molle Institue for Perceptual Artificial Intelligence, Martigny, Switzerland
Phone: +33 3 9024 4459
Fax: +33 3 9024 4455
E-mail: norman,bengio@idiap.ch, jjk@dpt-info.u-strasbg.fr
Web: www.idiap.ch

Abstract. In this study, two techniques that can improve the authentication process are examined: (i) multiple samples and (ii) multiple biometric sources. We propose the fusion of multiple samples obtained from multiple biometric sources at the score level. By using the average operator, both the theoretical and empirical results show that integrating as many samples and as many biometric sources as possible can improve the overall reliability of the system. This strategy is called multi-sample multi-source approach. This strategy was tested on a real-life database using neural networks trained in one-versus-all configuration.

INTRODUCTION

Biometric authentication is the problem of verifying an identity claim using a person's behavioural and physiological characteristics. Biometric authentication is becoming an important alternative to traditional authentication methods such as keys ("something one has", i.e., by possession) or PIN numbers ("something one knows", i.e., by knowledge) because it is essentially "who one is", i.e., by biometric information. Therefore, it is not susceptible to misplacement, forgetfulness or reproduction. Examples of biometric sources are fingerprint, face, voice, hand-geometry and retina scans.

However, to date, biometric-based security systems (devices, algorithms, architectures) still have room for improvement, particularly in their accuracy, tolerance to various noisy environments and scalability as the number of individuals increases. The focus of this study is on minimising the noise by using multiple biometric sources and multiple samples.

Biometric data is often noisy because of deformable templates, corruption by environmental noise, variability over time and occlusion by the user's accessories. The higher the noise, the less reliable the biometric system becomes.

Advancements in biometrics show two emerging solutions: combining several biometric sources [1, 5, 6] and combining several samples of a single biometric modality [3]. Combining several biometric sources can further be divided into a loosely coupled solution and a tightly coupled solution. A loosely coupled solution assumes very little or no interaction among the inputs. It integrates biometric data output of a relatively autonomous agent. An example of a loosely coupled system is the integration of audio and visual biometric data in an asynchronous manner. On the other hand, a tightly coupled solution assumes a strong interaction among the input measurements. It integrates biometric data at the sensor or representation level. An example of a tightly coupled system is the integration of audio and visual biometric data in a synchronous manner. In our opinion, combining several samples of a single biometric source can be considered a very tightly coupled solution because taking several life-scans of the same source of biometric data implies that the samples must be strongly correlated. Another category of solutions is to combine a biometric system with a non-biometric system.

Combining several biometric sources offers the advantage of relaxing the assumption of universality (the fact that each user should possess the biometric information), collectability (the extent to which the biometric information is measurable and adequately represented for the matching purpose), acceptability (the fact that each user agrees to have his/her biometric information scanned) and integrity (the degree of trustworthiness of the biometric system) of a target population in a given application.

Several studies have shown that a multi-model biometric system can improve the incompleteness of any single-model biometric system [1, 6]. In particular, Hong et al. have proven both theoretically and empirically that integrating multiple biometric models at score level and decision level can improve the overall system accuracy [5]. Kittler et al. have shown that combining several samples of a single biometric source can also improve the accuracy of the overall system [3].

The purpose of this paper is to examine how noise can be suppressed by using two separate approaches: multi-sample and multi-model. Section 2 gives an overview of a generic biometric framework and proposes a theoretical model to justify these two approaches. Section 3 shows some empirical results and is followed by our conclusions.

TOWARDS A MULTI-SAMPLE MULTI-SOURCE BIOMETRIC SOLUTION

A generic biometric integration model

In order to study the reliability [1] of a biometric system, a biometric-independent framework (see Figure 1) is proposed. Based on this framework, it will be

[1]The reliability of a system is defined as the probability that the system works correctly

shown that a system arranged in a serial manner can weaken the overall reliability of the system. By proposing a noise model, it will be shown that the reliability of the system can be improved by simply using multiple samples. This is due to the fact that when multiple samples are used, noise will be reduced.

Furthermore, the higher the level of noise is, the lower the reliability of the system becomes. By using the concept of the “committee of classifiers” [2], it will be justified that by combining several biometric sources via averaging, the reliability of the combined system is higher than the average reliability of its subsystem. Finally, an average operator is proposed to combine several samples of different biometric sources.

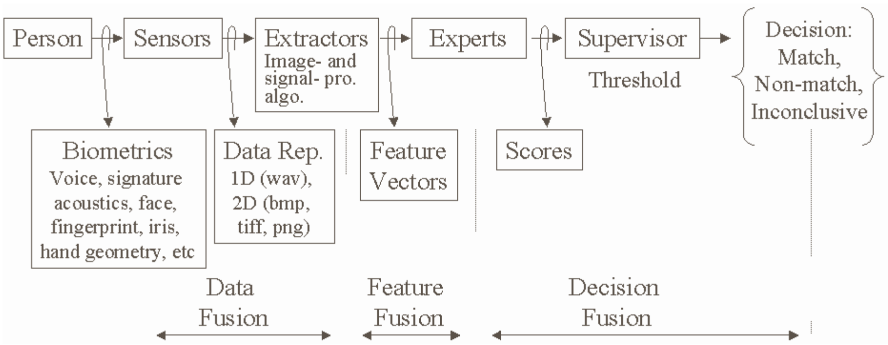

Figure 1: A generic biometric taxonomy and fusion scheme

In a biometric-independent framework (see Figure 1), a user’s biometric data is captured using sensors. Examples of sensors are Charged Couple Device (CCD) cameras, Infrared-Red (IR) cameras, fingerprint scanners and microphones. Each sensors has its own standard data representation. A set of operations, often founded on signal- and image-processing algorithms, constitute the building blocks of extractors. Extractors have two functions: to detect and to extract user-discriminant information. Each extractor produces its own type of vectors or feature vectors, also called templates in a more generic setting. Experts or classifiers are used to recognise these produced vectors. Classifiers are a set of pattern-matching algorithms, which may be learning-based (e.g. Multi-Layer Perceptron, Support Vector Machine, etc) or template-based (dynamic time wrapping, Euclidean distance, normalised correlation, etc). Classifiers map a vector belonging to an associated identity. They do so with a certain degree of confidence commonly called a score or a confidence measure. It could be a scalar value or a vector when more information is supplied. A score could be interpreted as the estimated a posteriori probability that a given feature belongs to the claimed class label. When there are several classifiers, a supervisor merges different scores to obtain the final decision. If the final decision is a match, then the system accepts the identity claim. If the decision is a non-match, then the system rejects the identity claim. Finally, if the decision is inconclusive, a fallback procedure should be activated.

Reliability within a single-model biometric system

The whole process from biometric acquisition to supervisor decision can be viewed as a serial system. Errors in each sub-module accumulate along the way. To analyse how the error affects the score, it is necessary to introduce the notion of reliability. The reliability $R_s$of a system (or sub-system) $s$is defined as the probability that the system works correctly, i.e., $R_s = P(correct)$. Conversely, the probability that a system c commits an error, i.e., $E_s = P(error)$ is defined by:

$$E_s = 1 - R_s. \tag{1}$$

Let R_i be the reliability of the subcomponent i, i.e., R_1 is the reliability of the sensor; R_2, the reliability of the extractor; R_3, the reliability of the classifier; and R_4, the reliability of the supervisor. Note that the error of each of the subcomponent can be calculated by using Equation 1.

Where do these errors come from? The very first error introduced is during biometric acquisition by the sensor. This could be due to errors in localisation, environmental interference, etc. In the next step, the biometric data may not be adequately represented. This is most evident during data sampling. Information is further lost during the extraction process. Finally, each classifier and supervisor introduces certain errors. For example, in face verification system using principle component analysis as an extractor, probability of the system error E (from sensor up to extractor) is the probability that it makes a false decision (false acceptance or false rejection). In this case, E is a function of a number of selected components. In reality, very often, one cannot measure the error in each subcomponent. However, intuitively, we know that good extractors (i.e. with high user discriminant power and produce decorrelated feature vectors) can increase the classifier performance. Here, we only provide a theoretical model to illustrate this interdependency nature of the serial process. Understanding this interdependency will eventually lead us to proposing a "parallel process" to be discussed in the next section.

According to the product law of reliability [7], the reliability of the whole chain process, R_s can be calculated as:

$$R_s = \prod_{i=1} R_i. \tag{2}$$

The multiplication rule is used because the reliability of each component is assumed to be independent. Equation 2 implies that $\forall R_i (R_i \leq R_s)$. Consequently, the reliability of a serial system is always lower than its subcomponents.

Since the theory of reliability assumes that there is a "true score", any real observed score are considered corrupted by a certain amount of additive noise.

In reality, biometric data changes with time due to ageing. In this analysis, so far, nothing is mentioned about the interval between two biometric samples. Our analysis does not consider samples taken so far apart in time that the effect of ageing could give raise to significant error. The theory of reliability will not work because the "true score" now is a function that moves with time. This, however, is true when the interval between sampling is short (say within a month).

Reliability of a single-modality biometric

This section establishes a method to calculate the reliability of a single modality biometric system. In general, a single model biometric system can be regarded as a function f that receives a vector feature $\mathbf{x}$ and outputs a score y: $y = f(\mathbf{x})$. The function f could be linear (e.g. Fisher discriminant analysis, Support Vector Machine with linear kernel) or non-linear (Multi-layer Perceptron, Support Vector Machines with non-linear kernel). y represents a measurement. It could be a score $\in [-1, +1]$, a confidence score (*a posteriori* probability) $\in [0, 1]$ or a distance metric $\in R^+$.

In the following section, y will be a confidence score showing a posteriori probability (e.g. an MLP having a single output neuron using a sigmoid activation function). Let $p(w_1|y)$ be the probabilistic distribution function (pdf) of client confidence scores and $p(w_2|y)$ be the pdf of impostor confidence scores.

The shaded area in Figure 2 then shows the mistakes (both false acceptance and false rejection errors) committed by the system at the threshold s. The bounded box in Figure 2 shows what the valid values of $p(w_1|y)$ and $p(w_2|y)$ which could have been otherwise normal if not bounded by the constraint that probability $\in [0, 1]$.

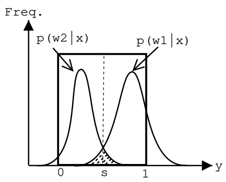

Figure 2: A schematic diagram of genuine and impostor distribution score

In other words, the shaded area gives the probability that the system commits error given a threshold s, which we denote as $E(s)$. Note that this probability is a function of s. It can be calculated using:

$$E(s) = \int P(false\ \ rejection) + P(false\ \ acceptance) dx \quad (3)$$

$$= 1 - \int_{-\infty}^{s} p(w_1|y) dx + \int_{s}^{+\infty} p(w_2|y) dx. \quad (4)$$

With a rewritten form of Equation 3 into Equation 1, we obtain reliability as:

$$R(s) = 1 - E(s) \tag{5}$$

$$= \int_{-\infty}^{s} p(w_1|y)dx - \int_{s}^{+\infty} p(w_2|y)dx. \tag{6}$$

One seeks to minimise E such that $E_{min} = min_s E(s)$. This is the same as to maximise R such that $R_{max} = max_s R(s)$. The optimum threshold s is at the point called Equal Error and E_{min} is called Equal Error Rate (EER).

It is obvious that if the two distributions completely overlap each other, R=0 and if they do not overlap at all, R=1. Note that in biometric applications, there are three categories of scores: genuine, "inter-template" (other clients) and impostors (also called "background database" [8]). There are also informed and uninformed impostors. In real-life hacking, impostors are informed, i.e., they possess certain amount of information about the identity to be faked. Among these three major categories of scores stated earlier, the genuine user scores are often the smallest data set. This study considers only genuine users and impostors. From now onwards, impostors are taken as a union of "inter-template" and "background database".

Classification of biometric system models

In the interest of avoiding error accumulation through such a serial process, one is led to study available types of biometric system models. In our opinion, biometric systems can be classified according to the number of samples per access and the number of biometric sources. The term "source" is used here to signify a particular class of biometric modality such as face, voice and so on. This is to distinguish it from the term "model" to be introduced later to signify different architecture. A "sample" is therefore a life-scan or shot of a biometric source. Using these two definitions, we propose four biometric "models". Figure 3(i) shows the typical serial process consisting of sensor, extractor and classifier. It is called a single-sample single-source (SSSS) biometric model. Figures 3(ii)-(iv) show a multi-sample single-source (MSSS), a single-sample multi-source (SSMS) and a multi-sample multi-source (MSMS) biometric model respectively. With this categorisation, Kittler et al.'s work [3] falls into MSSS model because several face samples are used during authentication. Hong et al.'s work [5] falls into SSMS model because they used a face sample and a fingerprint sample during authentication. Ross et al.'s work [1] that used face, fingerprint and hand geometry also falls into SSMS model.

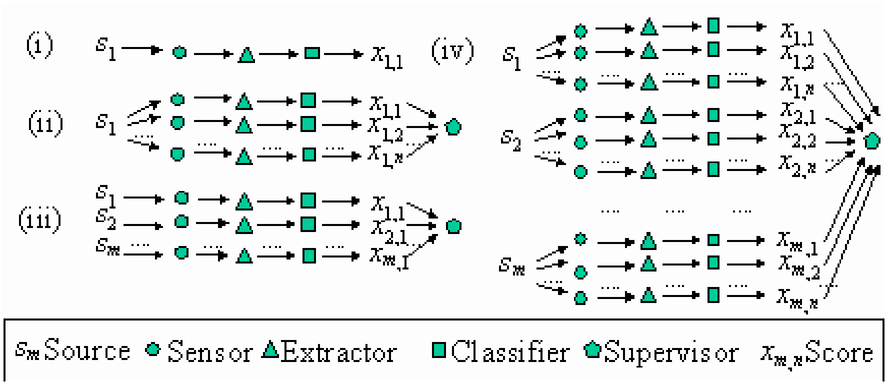

Figure 3: Single-sample single-source, multi-sample single-source, single-sample multi-source multi-sample multi-source biometric model

In Figure 3, S_i is the i-th biometric source (or modality) and $i = 1, \ldots, m$. is the score obtained from j-th sample of the i-th biometric source and . A MSSS model differs from a SSSS model in that a MSSS model uses several samples (therefore several serial processes) originating from the same biometric source. Note that a SSSS model does not require any supervisor. On the other hand, a SSMS model differs from a SSSS model in that a SSMS model uses several independent biometric sources. A MSSS model is more tightly coupled than a SSMS model. This means that if a biometric source is corrupted by the same noise (for instance, a cut on finger, a soar throat, a sun-burnt face), a MSSS model will probably fail to verify the identity of the person. On the other hand, multi-source biometric models (SSMS and MSMS) will probably be more robust against this kind of noise because their sources are not corrupted by the same noise.

Multi-sample single-source approach

Kittler et al. have shown that a MSSS model can be an effective way to boost a biometric system [3]. In their experiment setting with face images, several face samples are used. The optimal final decision score is found by averaging classifier scores. This can be justified by modelling a score corrupted by noise.

Let y_i be the "observed" measure (i.e. a score, $y_i = f(\mathbf{x}^i)$ of a given sample i out of a total of N samples and η_i be a noise drawn from a random zero-mean additive distribution. The "observed" measure $\hat{y}_i$ can be written as:

$$y_i = \hat{y} + \eta_i. \tag{7}$$

The mean of y_i, denoted as $\bar{y}$ is:

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i. \tag{8}$$

With enough samples, the expected value of y_i, denoted as $E\{y_i\}$ approximates the "true" measure:

$$E\{y_i\} = \hat{y}. \tag{9}$$

The expected value of random noise η_i, i.e., $E(\eta_i)$ is always zero. The variance of the observed y can be written as:

$$\sigma_y^2 = \frac{1}{N}\sigma_\eta^2. \tag{10}$$

Therefore, it can be concluded that when two or more scores of a single modality biometric are averaged, noise that occurs due to classification can be reduced.

Single-sample multi-source approach

In this section, we would like to show that the reliability of the joint (distributed) system, R_d, is better than the reliability of its sub-components, R_j, for a certain number of j biometric sources. Note that the index j is used to signify a biometric source so as to distinguish it from the index i that was used to signify one of N samples taken from a given biometric modality. This can be represented as: $\forall_j(R_d \geq R_j), j = 1, \ldots, M$, which is the same as $\forall_j(E_d \leq E_j)$, where E_j is the error of one of M sub-components j and E_d is the error of the joint system.

It is desirable that the following relationship holds:

$$E_d = \prod_{j=1}^{M} E_j \tag{11}$$

Equation 11 can be interpreted like this: in a system with M biometric sources, the whole system will always select the best sub-system. In other words, this is the optimal decision, called Oracle in [4]. Such a supervisor (one that merges the scores, as defined earlier) is the best result that one can get out of scores combination. From Equation 11, it is obvious that $\forall_j(E_d \leq E_j)$. Using normal and uniform distribution to model the probability of false rejection, Kuncheva [4] studied six supervisors: minimum, maximum, average, median, majority vote and Oracle. In practice, the Oracle supervisor does not exist because one does not know in advance the true identity during verification. Therefore, it is singled out in this discussion. Among the five classifiers mentioned, Kuncheva found that the average supervisor works the best when the error comes from the two distributions mentioned above.

By considering each sub-component of the multi-source system as an independent classifier, we can use the proof discussed by Bishop [2] (in Chap. 9) to show that the average supervisor satisfies $E_d \leq mean_j(E_j)$, instead of $\forall_j(E_d \leq E_j)$ the "perfect" requirement established earlier. He has shown that a committee of average and weighted average classifiers could perform better than a single classifier. The assumptions here are that each biometric single-modality subsystem is not correlated and that the error has zero mean. The result of proof is shown here:

$$E_d = \frac{1}{M^2}\sum_{j=1}^{M} E_j \quad (12)$$

$$= \frac{1}{M} mean_j(E_j) \quad (13)$$

Note that the difference between the context in [2] and our context here is that the independency of each sub-component is true and not an assumption. This is because each biometric sub-component operates on different biometric sources.

Multi-sample multi-source approach

It has already been shown that not only averaging scores of multiple sample can reduce noise in the serial process that is made of the sensor-extractor-classifier chain but also that averaging scores of multiple sources can achieve lower error by a factor of the number of biometric sources.

In this section, we wish to combine the two findings above using the strategy that we call multi-sample multi-source approach (MSMS). In such a system, we assume that there are M biometric modalities and for each modality, N samples are available. Scores made available to this system is denoted as $y_{i,j}$, where $i \in 1, \ldots, N$ and $j \in 1, \ldots, M$. By taking each score $y_{i,j}$ as a channel of the serial process of sensor-extractor-classifier, we can also associate the correspoding error involved, which we denote as $E_{i,j}$.

We argue that Equation 13 used in multi-source biometric holds as well for multi-sample biometric, with the weak assumption that the errors ($\forall_i E_{i,j}$ for a given j) are independent and have a zero mean. We can therefore write Equation 13 by changing the index from j to i, as follows:

$$E_d = \frac{1}{N} mean_i(E_i) \quad (14)$$

Violation of such assumption (in our case) results in increase of performance not by a factor of N but less [2]. However, we hope that in practice $E_d \leq mean_i(E_i)$ holds. (Empirical results by the work of Kittler [3] and our result in the later section also support this argument.) We will use the inequality 14 to deduce the inequality of the MSSS model:

$$E_{msss_j} \leq mean_i(E_{i,j}) \quad (15)$$

In the same way, from Equation 13, we can deduce that $E_d \leq mean_j(E_j)$. This inequality can be applied to the SSMS model as follows:

$$E_{ssms_i} \leq mean_j(E_{i,j}) \quad (16)$$

An MSMS model is by its definition a multi-model version of the MSSS model. So, the inequality $E_d \leq mean_j(E_j)$ that applies to multi-model holds for the MSMS model. Therefore, it is valid to write:

$$E_{msms} \leq mean_j(E_{msss_j}) \tag{17}$$

By replacing inequality 16 into inequality 17, we can write:

$$\begin{aligned}
E_{msms} &\leq mean_j(mean_i(E_{i,j})) && (18)\\
E_{msms} &\leq \frac{1}{M}\sum_j^M(\frac{1}{N}\sum_i^N(E_{i,j})) && (19)\\
E_{msms} &\leq \frac{1}{NM}\sum_j^M\sum_i^N(E_{i,j}) && (20)\\
E_{msms} &\leq \frac{1}{NM}\sum_{(i,j)}^{(N,M)}(E_{i,j}) && (21)
\end{aligned}$$

An intuitive way to combine scores in MSMS model is to introduce a mean operation similar to the inequality 18 as follows:

$$\begin{aligned}
y_{msms} &= mean_j(mean_i(y_{i,j})) && (22)\\
&= \frac{1}{M}\sum_j^M(\frac{1}{N}\sum_i^N(y_{i,j})) && (23)\\
&= \frac{1}{NM}\sum_j^M\sum_i^N(y_{i,j}) && (24)\\
&= \frac{1}{NM}\sum_{(i,j)}^{(N,M)}(y_{i,j}) && (25)
\end{aligned}$$

EXPERIMENTS AND RESULTS

Database

Briefly, there are 30 persons in the publicly available LSIIT database [2]. It has two types of biometric modalities: face and voice. For each type of biometric modality, 5 out of 10 samples of each person are used for training and the other 5 samples are used for testing. Both the training and test sets are mutually exclusive. For each biometric model, a set of one-versus-all configuration Multi-Layer Perceptrons (MLPs) are used, meaning to say that each MLP is associated with the biometric sample of a single client. In other

[2] The LSIIT BAS database is available at http://hydria.u-strasbg.fr

words, the samples of an associated client is treated as positive examples while all other samples belonging to other clients are treated as negative examples. Each MLP has an output neuron with a sigmoid activation function so that it can estimate the *a posteriori* probability $pw_1|\mathbf{x}$ when given a feature $\mathbf{x}$. For a database of 30 persons, 30 MLPs are needed to distinguish face features and the other 30 MLPs are needed to distinguish voice features. More details on feature pre-processing and extraction can be found in [6]. During testing with the previously unseen 5 samples, the experiment is first carried out with 1 sample, then 2 samples, and so on up to 5 samples.

Experiments and Results

Figure 4 and Figure 5 show the DET curves of face and voice biometric sources on test sets. Table 1 shows the corresponding min HTER (minimum point of the half total error rate) in percentage and approximated EER in percentage (a point nearest to FAR=FRR) of DET curves in Figure 4 and 5.

Table 1: The HTERs and EERs of face and voice biometric features

No. of samples used	Face min HTER	Face EER	Voice min HTER	Voice EER	Combined min HTER	Combined EER
1	7.184	10.000	6.897	6.897	0.805	0.805
2	2.701	3.333	4.828	6.494	0.690	0.690
3	1.207	2.874	4.540	6.667	0.000	0.000
4	0.000	0.000	2.126	3.333	0.000	0.000
5	0.000	0.000	2.414	3.333	0.000	0.000

The experiment shows that the multi-sample multi-source approach indeed increase in performance as more and more samples of each source are available. Indeed, using the current database, it reaches perfect performance. This is a very encouraging result but in reality, one does not have the luxury of such large pool of data. In our opinion, to further verify this approach, more experiments should be tested with different biometric features and larger databases that have several samples for each source of biometrics.

CONCLUSIONS

Biometric authentication can be viewed as a serial process involving a sensor, an extractor, a classifier and a supervisor. By using the product law of reliability, such a serial process could accumulate errors and reduce the overall reliability. According to this law, one can increase the overall system reliability by using several serial processes arranged in a parallel manner. Two techniques to create such processes are using multiple samples and multiple biometric sources. By assuming noise at the score level, it is proven that

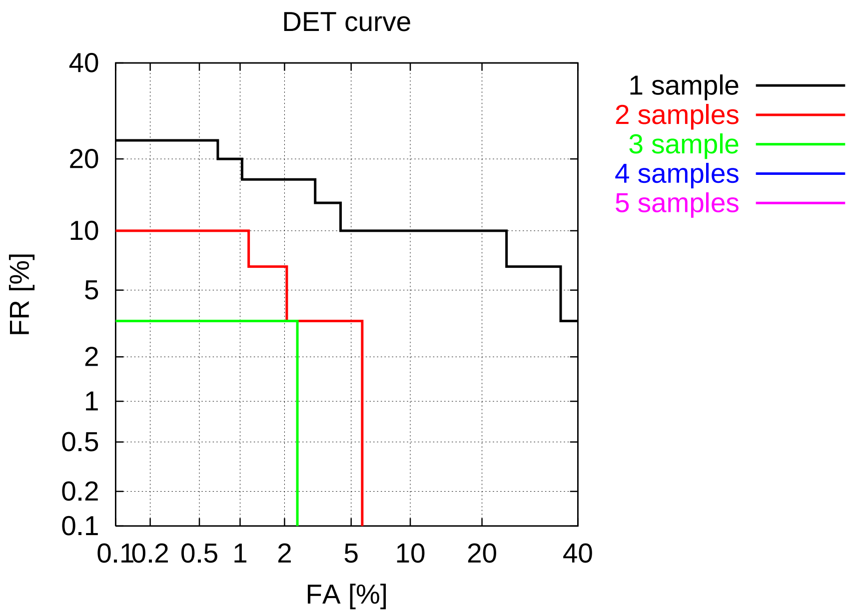

Figure 4: DET curves plotted using 1-5 face samples

averaging classifier scores from multiple biometric samples can reduce noise. However, if multiple biometric sources are available, by using the product law of reliability, it is proven that the reliability of the joint system can be further increased via averaging. By averaging N samples, our analysis shows that the joint system will have a reduction of error by a factor of N but less due to correlation between samples. However, by averaging M sources, one can achieve a reduction of error approaching a factor of M. Combining these two approaches can lead very good verification rate.

This hybrid approach is implemented with a set of neural network classifiers and is tested on a face and voice biometric database of 30 individuals. Using this small database, a perfect verification is recorded. This result is certainly promising but most importantly, it shows that one can use multiple samples or multiple biometric sources to boost the reliability of the whole system. An interesting application using this approach is in the inconclusive situation, i.e., the final decision score is marginal for acceptance. Under such situation, multi-sample multiple- source approach can be taken immediately. This will definitely increase the fault tolerance of intrusion. Furthermore, this approach suggests that it is always beneficial to life-scan longer features (i.e., longer speech signal) and more frames of facial features to increase robustness without adding much cost to the exisiting system.

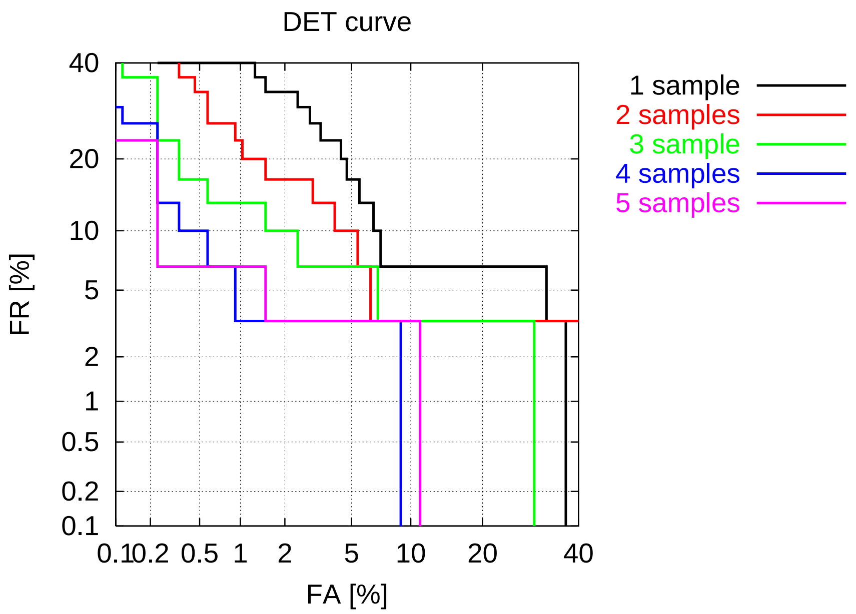

Figure 5: DET curves plotted using 1-5 voice samples

REFERENCES

[1] J.-Z. Q. A. Ross, A. Jain, "Information Fusion in Biometrics," in **The 3rd International Conference on Audio-Visual Biometric Person Authentication (AVBPA)**, 2001, pp. 354–359.

[2] C. M. Bishop, **Neural Networks for Pattern Recognition**, Oxford University Press, 1999.

[3] J. Kittler, G. Matas, K. Jonsson and M. Sanchez, "Combining evidence in personal identity verification systems," 1997.

[4] L. I. Kuncheva, "A Theoretical Study on Six Classifier Fusion Strategies," in **IEEE Transaction on Pattern Analysis and Machine Intelligence**, February 2002, vol. 24, pp. 281–286.

[5] S. P. Lin Hong, Anil K. Jain, "Can Multibiometrics Improve Performance?" Technical report msu-cse-99-39, **MSU-CSE**, 12 1999.

[6] N. Poh and J. Korczak, "Hybrid Biometric Authentication System Using Face and Voice Features," in **The 3rd International Conference on AVBPA**, 2001, pp. 348–353.

[7] W. M. Trochim, **The Research Methods Knowledge Base**, Internet WWW at http://trochim.human.cornell.edu/kb/index.htm, 2nd edn., August 2000.

[8] J. L. Wayman, **BIOMETRICS: Person Identification in Networked Society**, Kluwer Publishers, chap. 7, 1999.