Biometric Person Authentication IS A Multiple Classifier Problem*

Samy Bengio[1] and Johnny Mariéthoz[2]

[1] Google Inc, Mountain View, CA, USA, `bengio@google.com`
[2] IDIAP Research Institute, Martigny, Switzerland, `marietho@idiap.ch`

Abstract. Several papers have already shown the interest of using multiple classifiers in order to enhance the performance of biometric person authentication systems. In this paper, we would like to argue that the core task of Biometric Person Authentication is actually a multiple classifier problem as such: indeed, in order to reach state-of-the-art performance, we argue that all current systems , in one way or another, try to solve several tasks simultaneously and that without such joint training (or sharing), they would not succeed as well. We explain hereafter this perspective, and according to it, we propose some ways to take advantage of it, ranging from more parameter sharing to similarity learning.

1 Introduction

Biometric authentication is the task of verifying the identity of someone according to his or her claim, by using some of his or her biometric information (voice record, face photo, fingerprint, etc). A Biometric authentication system is thus trained to accept or reject an access request of one of the registered clients. This can be done efficiently by solving a two-class classification problem for each client separately.

When using more than one biometric information [1–3], the underlying verification system is said to be a *multiple classifier system*, as it merges several data sources coming from various biometric scanning devices, and hence, fits very well the topic of this workshop.

In this paper, we would like to argue that even when trying to solve a biometric authentication system based on a single modality and using a single classifier per client, one still needs to solve several classifier tasks jointly in order to reach state-of-the-art performance. We will argue in the following that there are several ways to solve these tasks jointly, ranging from the so-called world model approach, which is used to share common knowledge among several client models, to the learning of specialized distances or representation spaces, that can then be used for each client to take an accept/reject decision.

* This research was partially funded by the European PASCAL Network of Excellence and the Swiss OFES. Part of this work was done while Samy Bengio was at IDIAP Research Institute, Martigny, Switzerland.

The paper goes as follows: in Section 2 we explain the main argument of the paper, giving several examples of how it is already used in the various state-of-the-art approaches of the literature. In Section 3, we propose yet another way to use this new perspective of the biometric authentication task, and finally, in Section 4, we briefly conclude.

2 A Multiple Classifier Problem

The purpose of this paper is to show that the essence of a biometric authentication task is by nature a **multiple classifier problem**. This is not to be mixed up with the fact that multiple classifier systems often yield better performance for the task of biometric authentication [1–3].

Instead, we would like to advocate that, while biometric authentication can be seen as a two-class classification problem (each access should either be accepted or rejected), it is in fact *several two-class classification problems (one for each client model) that are inter-connected to each other* and one should take this into account in order to better design such systems.

Indeed, the general setup of a biometric authentication task is to be able to recognize whether a legitimate client is or is not who he or she claims to be by showing some biometric information. The expected resulting system should be able to accommodate for a growing number of clients[3], and should be able to enroll a new client with as little as possible of this new client's biometric material (be they voice, face, finger, or other modality prints).

2.1 Global Cost Function

The best way to illustrate that person authentication is a multiple classifier problem is to look at how such systems are evaluated in the research community. The most used measures of performance evaluate not only the performance of a single client model, but that of a large set of client models. Furthermore, this performance is not additive with respect to these apparently separate problems: indeed performance measures in person authentication always involve information such as False Acceptance Rate (FAR), False Rejection Rate (FRR), aggregates of them such as Half Total Error rate (HTER) or Detection Cost function (DCF) [4], and curves summarizing them, such as Detection Error Trade-off (DET) [5] and Expected Performance Curves (EPC) [6]. In all these cases, the global performance of a (set of) system(s) is not simply the sum of the performance of each client model (as the number of accesses per client model, be they legitimate or impostor, varies greatly from one client to another).

Hence, in order to train a good set of client models, one should select the corresponding parameters in order to maximize the joint performance of all models, and not separately the performance of each model. In this sense, it is clear that one needs to solve a **multiple classifier problem** jointly.

[3] It is expected that it should scale at most linearly with the number of clients, in terms of training time, and should be constant in terms of access time.

2.2 Parameter Sharing

Furthermore, given the inherent constraints of biometric authentication systems already discussed (scarce available data for each client, need for efficient access time, etc) most (if not all) of the state-of-the-art approaches in biometric authentication try to make use of a large quantity of previous client information in order to build a generic model, out of which each new client model starts from.

In order to illustrate this phenomenon, we will concentrate on the task of text-independent speaker verification, but bear in mind that the explanation is valid for any biometric authentication system. We can divide most of the current approaches into (apparently) generative approaches and discriminant approaches[4]. Let us review these two broad families of approaches.

The most well-known generative approach and still state-of-the-art method for text-independent speaker verification is based on adapted Gaussian Mixture Models (GMMs)[8]. It starts by training by Expectation-Maximization a generic GMM over a large quantity of voice data, and then adapts this generic model, using for instance Maximum A Posteriori (MAP) techniques, for each new client. Moreover, not only some parameters from the client model are adapted from a generic model (this usually applies to Gaussian means), but several other parameters of client models are simply *copied* from the generic model (this usually applies to Gaussian variances and weights). This approach turns out to be the most efficient way to make the best use of the small amount of each client's information. Furthermore, it is also state-of-the-art for many other biometric authentication problems, including face verification [9].

Several experiments have shown in the past that if one tries to solve the speaker verification problem by training a new GMM for each client instead of starting from a generic model and adapting it, then the performance results are poorer, even when tuning the number of Gaussians for each client separately.

Other evidences of the same phenomenon can be seen in various enhancements of the basic GMM based approach that have been proposed in the biometric authentication literature over the years, including the use of normalization techniques (Z-norm, T-norm, etc) which aim at trying to normalize the obtained score to make it more robust to several kinds of variations (intra-speaker, inter-speaker, inter-session, channel, etc) [10]. Once again, in order to compute efficient normalization parameters, one needs to use a large number of previous client information. This has already been demonstrated empirically.

For instance, one can see in Table 1 and in Figures 1 and 2 the comparative performance of three systems on the NIST database described in appendix A. One system is trained using the classical EM training approach (also called *maximum likelihood approach*, or ML), the second one is trained using the MAP adaptation technique, and the third one is trained using MAP and applying the T-norm normalization technique. Table 1 shows the *a priori* performance of all three systems on the test set in terms of FAR, FRR, and HTER, after selecting

[4] Note that actually, generative approaches that effectively work usually implement several tricks that make the overall system quite discriminant in various aspects [7].

the threshold that minimized the Equal Error Rate (EER) on the development set. Figure 1 shows the DET curves on the development set (the lower to the left the better). Finally, Figure 2 shows the EPC curves on the test set. The latter curves provide unbiased estimates of the HTER performance of all three systems for various expected ratios of FAR and FRR (represented on the X-axis of the graph by γ). The lower part of the graph also shows whether one of the three models was statistically significantly better than another one, according to the statistical test described in [11].

In all cases (table results, DET and EPC curves), it is clear that the more one shares information among client models, the better the expected performance on new client models becomes.

Table 1. Point-wise performance results, in terms of FAR, FRR and HTER (%), on the test set of the NIST database using classical ML training, MAP training and MAP plus T-normalization procedure. These results where obtained by selecting the threshold that minimized the EER on the development set.

	FAR	FRR	HTER
ML	3.23	30.80	17.02
MAP	4.79	16.38	10.59
MAP + T-norm	7.06	10.29	8.68

In all these cases, one could never obtain a good authentication system for a given client if no information was shared among various clients. Hence, while it is never explicitly said, **all successful generative approaches to person authentication systems are built by sharing some information among several classifier systems**.

While generative approaches have been used successfully for many years, there are good reasons to think that direct discriminant approaches should perform better; one of them, advocated by Vapnik [12], is that *one should never try to solve a more difficult task than the target task*. Hence if the task is to decide whether to accept or reject an access, there should be no reason to first train a generic model that describes everything about what is a correct access and what is an incorrect access, as the only thing that matters is the decision boundary between these two kinds of accesses.

More recently thus, several discriminant approaches have started to provide state-of-the-art performance in various person authentication tasks.

For instance, the *Nuisance Attribute Projection* (NAP) approach [13] tries to find a linear transformation of the access data into a space where accesses of the same client are near each others, in terms of the L2-norm. In order to refrain from finding an obvious bad solution, the size of the target space (or more specifically its Co-rank) is controlled by cross-validation. This transformation is learned on a large set of clients (hence, similarly to learning a generic GMM in the generative approach). After this step is performed, a standard linear support vector machine (SVM) [14] is trained for each new client over the transformed

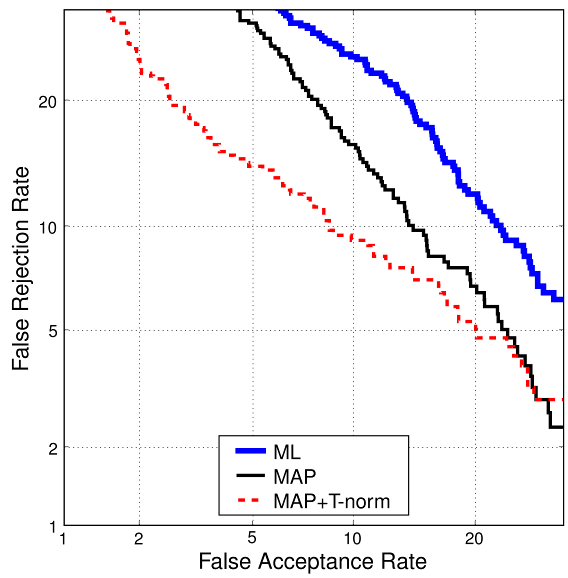

Fig. 1. DET curves on the development set of the NIST database using classical ML training, MAP training and MAP plus T-normalization procedure.

access data. This approach provided very good performance in the recent NIST evaluations.

This shows even more the fact that one has to share some information among many clients in order to obtain good performance. In this case, the shared information is used to learn how to transform the original data into a space which will be invariant to various aspects, such as the channel, and concentrate on the important topic, the client specific information.

Unfortunately, one thing in the NAP approach is somehow disappointing: the transformation function is not learned using the criterion that is directly related to the task; rather, it tries to minimize the mean squared distance between accesses of the same client to get rid of the *channel effect*, but do nothing about accesses from different clients for instance. In other words, we might still try to do more than the expected task, which is not optimal, according to Vapnik.

Another recent approach that also goes in the right direction and that obtains similar and state-of-the-art performance as the NAP approach is the Bayesian Factor Analysis approach [15]. In this case, one assumes that the mean vector of a client model is a linear combination of a generic mean vector, the mean vector of the available training data for that client, and the mean vector of the particular channel used in this training data. Once again, the linear combination parameters are trained on a large amount of access data, involving a large amount

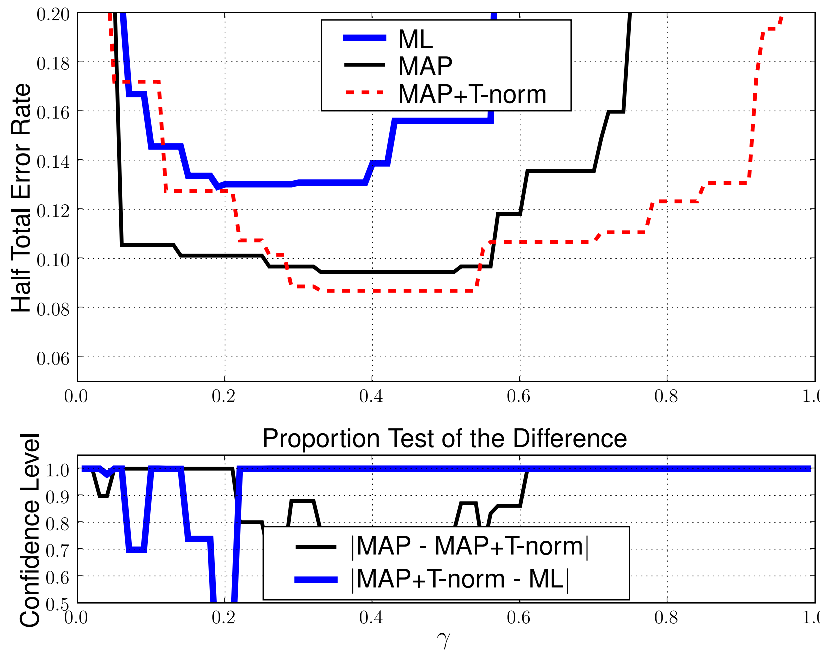

Fig. 2. Expected Performance Curves (EPC, the lower the better) on the test set of the NIST database using classical ML training, MAP training and MAP plus T-normalization procedure. The lower graph shows the confidence level of one model being statistically significantly better than another one in each part of the EPC curve.

of clients. While this approach is nicely presented theoretically (and obtains very good empirical performance), it still does not try to find the optimal parameters of client models and linear combination by taking into account the global cost function.

3 Similarity Learning

In this section, we would like to propose at least one idea that would directly take into account the *multiple classifier problem* dimension of biometric authentication tasks.

We advocated in the previous section that one should try to learn jointly some information about several clients that would directly help in the final task of accepting or rejecting accesses. We also advocated that a promising approach should be discriminant, as the NAP approach is or similarly to Campbell's polynomial expansion for sequence kernel approach based on support vector machines (SVMs)[16].

We also acknowledge that the SVM approach to speaker verification is discriminant but given that each client SVM is trained separately, the currently only parameters that can be shared among clients in this approach are through the transformation of the input space (as it is done in the NAP approach)[5].

We would like to propose here another discriminant method based on SVMs, but that would use a particular kernel (or similarity measure between two accesses) that would be learned on several clients' accesses. Indeed, if we knew in advance a distance or similarity measure that would quantify reasonably well whether two accesses are coming from the same client or not, and even assuming this measure to be noisy, putting it into an SVM and training the SVM to solve the final authentication task would yield a better performance than using the standard Gaussian or polynomial kernel for the same task, as it is done in Campbell's approach for instance.

Hence, our proposed approach is the following. Using a large base of accesses for which one knows the correct identity, train a parametric similarity measure that will assess whether two accesses are coming from the same person or not. That can be done efficiently by stochastic gradient descent using a scheme similar to the so-called *Siamese neural network* [17] and a margin criterion with proximity constraints, as follows.

Let $\phi(\cdot)$ be a mapping of a given access into a space where two accesses of the same client are near while two accesses from different clients are far. More formally, given a triplet (x, x^+, x^-) such that x is a vector representation of a given access, x^+ is a vector representation of an access of the same client as x and x^- a vector representation of an access of a client different from that of x, we would like the scalar product of the similar ones in the $\phi(\cdot)$ space to be higher than that of the dissimilar ones:

$$\phi(x) \cdot \phi(x^+) > \phi(x) \cdot \phi(x^-). \tag{1}$$

Let $\phi(\cdot)$ be a multi-layer perceptron (MLP), the following ranking loss function L [18–20] can be used to search for a good candidate for (1):

$$L = |1 - \left[\phi(x) \cdot \phi(x^+) - \phi(x) \cdot \phi(x^-)\right]|_+ \tag{2}$$

where $|a|_+ = \max(0, a)$.

Finally, let us consider for the moment that a given access can be transformed into a vector representation using a trick such as the one used in Campbell's polynomial expansion approach [16]. That constraint could be relaxed to any other sequence kernel technique that have been proposed in the literature, such as in [7].

We now have all the ingredients to learn efficiently $\phi(\cdot)$ by stochastic gradient descent. One simply needs to prepare, out of a database of several client accesses, a training set of triplets (x, x^+, x^-); one then needs to select a particular form for the parametric function $\phi(\cdot)$ noting that the only constrains here is that it

[5] Actually, another way the SVMs share information among them is through the same list of negative examples, or impostor accesses.

should be positive and differentiable with respect to its parameters (in particular, $\phi(\cdot)$ can be non-linear, which is not the case for the NAP and Bayesian Factor Analysis approaches). One can then train $\phi(\cdot)$ using stochastic gradient descent to optimize (2) on that data. The chosen loss function (2) involves a *security* margin, as not only do we want similar accesses to be nearer each other than dissimilar ones in that space, we also want the difference between the two similarity measures to be at least 1 (or any positive constant, for that matter).

Once $\phi(\cdot)$ is trained on a reasonably large database, one can then use it to create the following kernel k for each client SVM, similarly to the NAP and Campbell's approaches:

$$k(x, y) = \phi(x) \cdot \phi(y) \tag{3}$$

which guaranties Mercer's conditions for proper SVM training [14], as long as we put some mild constraint on $\phi(\cdot)$ such as being continuous and positive, which is straightforward to enforce.

4 Conclusion

In this paper, we have presented a somehow novel view of the task of biometric person authentication, advocating that it should be solved by taking into account that one needs to create simultaneously several two-class classifiers, one for each client, and that parameter sharing of one sort or another during this process is of paramount importance. We have shown that all currently state-of-the-art approaches to several biometric authentication tasks are indeed following this approach while never referring to it specifically. We have then proposed a novel approach, based on learning a similarity measure between two accesses, trained by a margin criterion on a large set of previous client accesses, that can then be plugged in an SVM for each client to replace standard kernels such as the polynomial or the Gaussian kernel. A nice extension of the following framework could be to incorporate the transformation of an access (which is normally a variable size sequence of feature vectors such as MFCCs) into a vector representation. A standard Time Delay Neural Network (TDNN) [21] could be used inside the $\phi(\cdot)$ function to accomplish this.

A The NIST Database Used in This Paper

The NIST database used here is similar to the one described in [7] and its description goes as follows: it is a subset of the database that was used for the *NIST 2002 and 2003 Speaker Recognition Evaluation*, which comes from the second release of the cellular switchboard corpus, Switchboard Cellular - Part 2, of the Linguistic Data Consortium. This data was used as test set while the world model data and the development data comes from previous NIST campaigns. For both development and test clients, there were about 2 minutes of telephone speech used to train the models and each test access was less than 1 minute long. Only female data are used and thus only a female world model is used. The development population consisted of 100 females, while the test set is composed of 191 females. 655 different records are used to compute the world model or as negative examples for the discriminant models. The total number of accesses in the development population is 3931 and 17578 for the test set population with a proportion of 10% of true target accesses.

Table 2 gives a summary of the hyper-parameters used for GMM based experiments after selection based on minimizing EER on the development set.

Table 2. Summary of the hyper-parameters for GMMs based systems on the NIST database

ML Hyper-Parameters		
# of Iterations	# of Gaussians	Variance Flooring (%)
25	75	60

MAP Hyper-Parameters		
# of ML Iterations	# of Gaussians	Variance Flooring (%)
25	100	60
# of MAP Iterations	MAP Factor	Variance Flooring (%)
5	0.5	60

References

1. Hong, L., Jain, A.: Multi-Model Biometrics. In: Biometrics: Person Identification in Networked Society. (1999)
2. Kittler, J., Matas, G., Jonsson, K., Sanchez, M.: Combining Evidence in Personal Identity Verification Systems. Pattern Recognition Letters **18**(9) (1997) 845–852
3. Kittler, J., Messer, K., Czyz, J.: Fusion of Intramodal and Multimodal Experts in Personal Identity Authentication Systems. In: Proc. Cost 275 Workshop, Rome (2002) 17–24
4. Martin, A., Przybocki, M.: The NIST 1999 speaker recognition evaluation - an overview. Digital Signal Processing **10** (2000) 1–18
5. Martin, A., Doddington, G., Kamm, T., Ordowski, M., M.Przybocki: The DET curve in assessment of detection task performance. In: Proceedings of Eurospeech'1997. (1997) 1805–1809
6. Bengio, S., Mariéthoz, J.: The expected performance curve: A new assessment measure for person authentication. In: Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop. (2004)
7. Mariéthoz, J.: Discriminant models for text-independent speaker verification. Technical Report IDIAP-RR 06-70, IDIAP (2006)
8. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted gaussian mixture models. Digital Signal Processing **10**(1–3) (2000)
9. Cardinaux, F., Sanderson, C., Bengio, S.: User authentication via adapted statistical models of face images. IEEE Transactions on Signal Processing **54**(1) (2006) 361–373

10. Auckenthaler, R., Carey, M., Lloyd-Thomas, H.: Score normalization for text-independent speaker verification systems. Digital Signal Processing **10** (2000) 42–54
11. Bengio, S., Marithoz, J.: A statistical significance test for person authentication. In: Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop. (2004)
12. Vapnik, V.N.: The nature of statistical learning theory. second edn. Springer (2000)
13. Solomonoff, A., Campbell, W.M., Quillen, C.: Channel compensation for SVM speaker recognition. In: Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop. (2004)
14. Burges, C.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery **2**(2) (1998) 1–47
15. Kenny, P., Boulianne, G., Dumouchel, P.: Eigenvoice modeling with sparse training data. IEEE Transactions on Speech and Audio Processing **13**(3) (2005)
16. Campbell, W.M.: Generalized linear discriminant sequence kernels for speaker recognition. In: Proceedings of the IEEE International Conference on Audio Speech and Signal Processing, ICASSP. (2002) 161–164
17. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: Proc. of Computer Vision and Pattern Recognition Conference. (2005)
18. Schultz, M., Joachims, T.: Learning a distance metric from relative comparison. In: Advances in Neural Information Processing Systems 16. (2003)
19. Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G.: Learning to rank using gradient descent. In: ICML. (2005) 89–96
20. Grangier, D., Bengio, S.: Exploiting hyperlinks to learn a retrieval model. In: NIPS Workshop on Learning to Rank. (2005)
21. Weibel, A., Hanazawa, T., Hinton, G., Shikano, K., Lang, K.: Phoneme recognition using time-delay neural networks. IEEE Transactions on Accoustics, Speech and Signal Processing **37**(3) (1989) 328–339