IDIAP RESEARCH REPORT

The BANCA Database and Experimental Protocol for Speaker Verification

Fabienne Porée [5] Johnny Mariéthoz [3]
Samy Bengio [1] Frédéric Bimbot [2]

IDIAP–RR 02-13

April 18, 2002

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[5] IRISA - Campus Beaulieu, 35042 Rennes, France, fporee@irisa.fr
[3] IDIAP, CP 592, 1920 Martigny, Switzerland, marietho@idiap.ch
[1] IDIAP, CP 592, 1920 Martigny, Switzerland, bengio@idiap.ch
[2] IRISA - Campus Beaulieu, 35042 Rennes, France, bimbot@irisa.fr

IDIAP Research Report 02-13

The BANCA Database and Experimental Protocol for Speaker Verification

Fabienne Porée Johnny Mariéthoz Samy Bengio Frédéric Bimbot

April 18, 2002

Abstract. Identity verification has become a very important research topic recently, particularly using methods based on the face or the voice of the individuals. In the context of the BANCA european project, a novel multi-modal database was recently recorded, spanning 5 european languages, 2 modalities (face and voice), 2 microphones, 2 cameras and almost 300 individuals. As we believe that this database offers many advantages for this research community, this paper essentially presents the database and its associated experimental protocol, as well as a baseline state-of-the-art system using the voice data for a text-independent speaker verification task.

1 Introduction: The BANCA Project

The objective of the BANCA European project is to develop and implement a secure system with enhanced identification, authentication and access control schemes for applications over the Internet such as tele-working and Web - or remote - banking services. One of the major innovations targeted by this project is to obtain an enhanced security system by combining classical security protocols with robust multi-modal verification schemes based on speech and videos.

In this context, there was a need for a multi-modal database that would include various but realistic recording scenarios, using different kinds of material, in different European languages. Existing multi-modal databases such as XM2VTS [4] could not be used as they were not realistic enough (for instance, the subjects were recorded in a controlled environment only, with blue background, which cannot be compared to a real situation when one makes a transaction at home or through an ATM).

A new database was thus recorded recently, and an experimental protocol using the database was developed [1]. The aim of this paper is thus to present the database and its associated protocol, as well as a first series of experimental results obtained on it for the unimodal task of text-independent speaker verification.

In the following, we first present the task of identity verification, then the BANCA database is described followed by an associated experimental protocol. Afterwards, the experiment section presents the results of a first state-of-the-art system tested on this database.

2 Notation and Formalism

Identity Verification (IV) can be defined as the task that consists in verifying the identity X claimed (explicitly or implicitly) by a person U, using a *sample* y from this person, for instance an image of the face of U, a speech signal produced by U, etc... By comparing the sample to some *template* (or *model*) of the claimed identity X, the IV system outputs a decision of *acceptance* or *rejection*. The process can be viewed as a hypothesis testing scheme, where the system has to decide within the following alternative:

- U is the *true client* (acceptance, denoted $\hat{X}$),
- U is an *impostor* (rejection, denoted $\hat{\bar{X}}$).

In practice, an IV system can produce 2 types of errors:

- False Acceptance (FA) if the system has wrongly accepted an impostor,
- False Rejection (FR) if a true client has been rejected by the system.

In practical applications, these 2 types of error have an associated cost, which will be denoted as C_{FA} and C_{FR} respectively. Moreover, in order to measure the quality of the system independently from the distribution of the accesses, we define the following quantities:

- the False Acceptance Rate (P_{FA}) is the ratio between the number of FA and the number of impostor accesses,
- the False Rejection Rate (P_{FR}) is the ratio between the number of FR and the number of client accesses.

IV approaches are usually based on the characterization of hypotheses X and $\bar{X}$ by a client template and a non-client template respectively, which are learned during a *training* (or enrollment) phase (the non-client model may even be trained during a preliminary phase, also called *installation* phase, and is often the same for every client, in which case it is called the *world model*). Once the template for client X has been created, the system becomes operational for verifying identity claims on X. In the context of performance evaluation, this is referred to as the *test* phase. Conventionally, the procedure used by an IV system during the test phase can be decomposed as follows:

- *feature* extraction, i.e. transformation of the raw sample into a (usually) more compact representation,
- *score* computation, i.e. output of a numerical value $S_X(y)$ based on a (normalized) distance between y and the templates for X (and $\bar{X}$),
- *decision* by comparing the score $S_X(y)$ to a threshold Θ, independent of X.

3 The BANCA database

The BANCA database was designed in order to test multi-modal IV with various acquisition devices (2 cameras and 2 microphones) and under several scenarios (controlled, degraded and adverse). For 5 different languages (English, French, German, Italian and Spanish), video and speech data were collected for 52 subjects (26 males and 26 females), i.e. a total of 260 subjects. Each language - and gender - specific population was itself subdivided into 2 groups of 13 subjects, denoted in the following $g1$ and $g2$. Each subject recorded 12 sessions, each of these sessions containing 2 records: 1 true *client access* and 1 informed (the actual subject knew the text that the claimed identity subject was supposed to utter) *impostor attack*. The 12 sessions were separated into 3 different scenarios:

- *controlled* (c) for sessions 1-4,
- *deraded* (d) for sessions 5-8,
- *adverse* (a) for sessions 9-12.

During a record, the subject uttered his name, address and phone number (these were faked information), which took an average of 15 seconds. Two cameras were used, a cheap one and an expensive one. The cheap camera was used in the degraded scenario, while the expensive camera was used for controlled and adverse scenarios. In parallel, two microphones, a poor quality one and a good quality one, were used simultaneously in each of the three scenarios.

In a given session, the impostor accesses by subject X were successively made with a claimed identity corresponding to each other subject from the *same group* (as X). In other words, all the subjects in group g recorded one (and only one) impostor attempt against each other subject in g and each subject in group g was attacked once (and only once) by each other subject in g. Moreover, the sequence of impostor attacks was designed so as to make sure that each identity was attacked exactly 4 times in the 3 different conditions (hence 12 attacks in total).

For each language, an additional set of 30 other subjects, 15 males and 15 females, recorded one session (audio and video). This set of data is referred to as *world data*. These individuals claimed two different identities, recorded by both microphones. Finally, any data outside the BANCA database will be referred to as *external data*.

4 Experimental Protocol

In this paper, we present a first protocol using only one language [1]; other protocols taking into account all 5 languages will be presented later.

4.1 A Monolingual Protocol

In order to define an experimental protocol, it is necessary to define a set of evaluation data (or *evaluation set*), and to specify, within this set, which are to be used for the training phase (enrollment) and which are to be used for the test phase (test accesses).

Moreover, before becoming operational, the development of an IV system requires usually the adjustment of a number of configuration parameters (model size, normalization parameters, decision thresholds, etc.). It is therefore necessary to define a *development set*, on which the system can be calibrated and adjusted, and for which it is permitted to use the knowledge of the actual subject identity during the test phase. Once the development phase is finished, the system performance can then be assessed on the evaluation set (without using the knowledge of the actual subject identity during the test phase).

To avoid any methodological flaw, it is essential that the development set is composed of a distinct subject population as the one of the evaluation set. In order to carry realistic (and unbiased experiments), it is necessary to use different populations and data sets for development and for evaluation. We distinguish further between 2 circumstances: single-modality evaluation experiments and multi-modality evaluation experiments. In the case of single-modality experiments, we need to distinguish only between two data sets: the development set, and the evaluation set. In that case, $g1$ and $g2$ are used alternatively as development set and evaluation set (when $g1$ is used as development set, $g2$ is used as evaluation set, and vice versa).

In the case of multi-modality experiments, it is necessary to introduce a third set of data: the *(fusion) tuning set* used for tuning the fusion parameters, i.e. the way to combine the outputs of each modality. If the tuning set is identical to the development set, this may introduce a bias in the estimation of the tuning parameters (*biased* case). An other solution is to use three distinct sets for development, tuning and evaluation (*unbiased* case). In that case, we expect the experimenters to use data from the other languages as development set, while $g1$ and $g2$ are used alternatively for tuning and evaluation.

In the BANCA protocol, we consider that the true client records for the first session of each condition is reserved as training material, i.e. the true client record from sessions 1, 5 and 9. In all our experiments, the client model training (or template learning) is done on at most these 3 records.

We then consider 7 distinct training-test configurations, depending on the actual conditions corresponding to the training and to the testing conditions, as exposed in Table 1.

Test Sessions	Train Sessions			
	1	5	9	1,5,9
C: 2-4 I: 1-4	Mc			
C: 6-8 I: 5-8	Ud	Md		
C: 10-12 I: 9-12	Ua		Ma	
C: 2-4,6-8,10-12 I: 1-12	P			G

Table 1: Description of all protocols as a function of train and test (C: client, I: impostor) session numbers. The configurations obtained are: Matched controlled (Mc), Matched degraded (Md), Matched adverse (Ma), Unmatched degraded (Ud), Unmatched adverse (Ua), Pooled test (P), Grand test (G).

From the comparison of these various performances, it is possible to measure:

- the intrinsic performance in a given condition,
- the degradation from a mismatch between controlled training and uncontrolled test,
- the performance in varied conditions with only one (controlled) training session,
- the potential gain that can be expected from more representative training conditions.

4.2 Performance Measure

In order to visualize the performance of the system, irrespective of its operating condition, we use the conventional DET curve [5], which plots on a log-deviate scale the *False Rejection Rate* P_{FR} as a function of the *False Acceptance Rate* P_{FA}. Traditionally, the point on the DET curve corresponding to $P_{FR} = P_{FA}$ is called EER (Equal Error Rate) and is used to measure the closeness of the DET curve to the origin. The EER value of an experiment is reported on the DET curve, to comply with this tradition. We also recommend to measure the performance of the system for 3 specific operating conditions, corresponding to 3 different values of the Cost Ratio $R = C_{FA}/C_{FR}$, namely $R = 0.1, R = 1, R = 10$. Assuming equal *a priori* probabilities of genuine clients and impostor, these situations correspond to 3 quite distinct cases:

$R = 0.1$	$\rightarrow$	a FA is an order of magnitude less harmful than a FR,
$R = 1$	$\rightarrow$	a FA and a FR are equally harmful,
$R = 10$	$\rightarrow$	a FA is an order of magnitude more harmful than a FR.

When R is fixed and when P_{FR} and P_{FA} are given, we define the Weighted Error Rate (WER) as:

$$WER(R) = \frac{P_{FR} + R\, P_{FA}}{1 + R} \; . \tag{1}$$

P_{FR} and P_{FA} (and thus WER) vary with the value of the decision threshold Θ, and Θ is usually optimized so as to minimize WER on the development set D:

$$\hat{\Theta}_R = \arg\min_{D} \; WER(R) \; . \tag{2}$$

The *a priori threshold* thus obtained is always less efficient than the *a posteriori threshold* that optimizes WER on the evaluation set E itself:

$$\Theta^*_R = \arg\min_{E} \; WER(R) \; . \tag{3}$$

The latter case does not correspond to a realistic situation, as the system is being optimized with the knowledge of the actual test subject identities on the evaluation set. However, it is interesting to compare the performance obtained with *a priori* and *a posteriori* thresholds in order to assess the reliability of the threshold setting procedure.

5 Experiments, Results and Comments

In this section, we present a first series of experimental results using the English data from the BANCA database applied to the task of text-independent speaker verification.

5.1 General Methodology

All the experiments described here have followed the same methodology. First of all, the original waveforms have been sampled every 10ms and then parameterized into 16 LFCC coefficients and their first derivative, as well as the energy together with its first derivative, for a total of 34 features. The feature warping technique proposed by Queensland University of Technology [6] was also performed.

Afterward, a *bi-Gaussian method* has been used in order to remove the silence frames from the data. We trained a Gaussian Mixture Model (GMM) with two Gaussians in an unsupervised mode, with the hope that one Gaussian would capture the speech frames while the second would capture the silence frames, since they have quite different characteristics. We then simply removed the frames for which the maximum likelihood was given by the Gaussian corresponding to low energy frames.

While the energy is important in order to remove the silence frames, it is not adapted to the task of discrimination between clients and impostors, and they were thus removed from the features after silence removal. Hence, the world and client models were trained with 33 features (instead of 34).

The next step consisted in the selection of the *hyper-parameters* of the system, such as the number of Gaussians of the model, the *v-floor* which represents the minimal proportion of the global variance that a Gaussian can take, the MAP adaptation factor between the world model corresponding to the gender of the client and the client model. We used the following methodology: we trained two gender-specific world models using the world model data and trained client models using the development set by adapting their corresponding gender world model. We then selected the value of the hyper-parameters that optimized the Equal Error Rate (EER) on the test accesses of the development set. Finally, we trained the models of the evaluation set using these hyper-parameters and report the results obtained on the test accesses of the evaluation set. Hence, these results are unbiased as the corresponding data have not been used for any purpose during the development of the models.

5.2 Baseline Experiment

The following experiments have been performed only on the speech modality with both microphones.

In both cases, the hyper-parameters have been selected for the P sub-protocol (training in controlled condition, testing in all conditions) as it is probably the most representative scenario for a real-life application.

With the good quality microphone (mic1), the values of the hyper-parameters found on the development set are the following: 200 Gaussians for world and client models, $\alpha = 0.25$ for MAP, *v-floor* $= 0.001$ to constraint the variance and $\tau = 3s$ for the window size of the feature warping method. Figure 1 presents the DET curves obtained with the protocols P and G for groups $g1$ and $g2$.

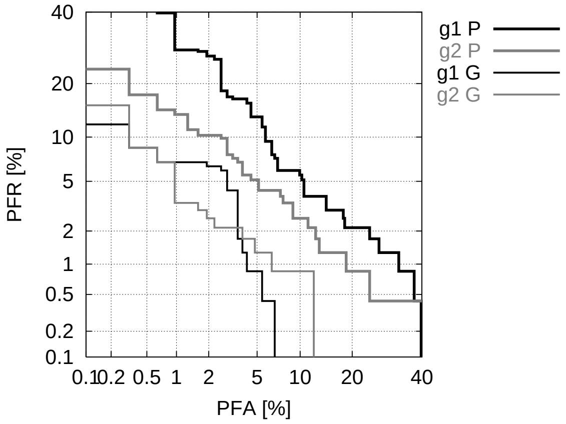

Figure 1: DET curves for protocols P and G with good quality microphone.

Tables 2 and 3 present the results for the 4 operating points as defined in section 4.2, both using the *a priori* selected threshold (realistic) and the *a posteriori* selected threshold (optimistically biased), for both groups $g1$ and $g2$, with protocol P.

	$R = 0.1$		$R = 1$		$R = 10$		EER	
A post.	2.1	18.3	6.0	7.0	28.6	1.0	6.8	7.0
threshold	3.6		6.5		3.5		6.9	
A priori	2.1	23.1	6.4	7.0	25.6	2.6	6.0	8.6
threshold	4.0		6.7		4.7		7.3	

Table 2: Results for $g1$, protocol P, good quality microphone ($\frac{PFR|PFA}{WER}$).

	$R = 0.1$		$R = 1$		$R = 10$		EER	
A post. threshold	1.3	13.1	5.6	3.8	17.5	0.3	5.1	5.1
	2.4		4.7		1.9		5.1	
A priori threshold	2.6	9.6	5.6	4.2	23.9	0.0	6.8	3.8
	3.2		4.9		2.2		5.3	

Table 3: Results for $g2$, protocol P, good quality microphone.

	$R = 0.1$		$R = 1$		$R = 10$		EER	
A post. threshold	0.0	6.7	0.9	4.2	8.9	0.3	3.4	3.5
	0.6		2.5		1.1		3.5	
A priori threshold	0.0	18.3	4.3	3.5	8.6	0.3	1.3	4.2
	1.7		3.9		1.1		2.7	

Table 4: Results for $g1$, protocol G, good quality microphone.

Tables 4 and 5 present the results obtained with protocol G.

The values of the hyper-parameters found on the development set and protocol P with the poor quality microphone (mic2) are the following: 400 Gaussians, $\alpha = 0.2$, *v-floor* $= 0.4$ and $\tau = 3s$. Table 6 presents the summary of results in term of *a posteriori* EER and *a priori* HTER (Half Total Error Rate, i.e. WER for $R = 1$), obtained by averaging $g1$ and $g2$ results, for protocols P and G, and both microphones.

The use of the QUT feature warping technique brought an average improvement of approximately 1% (absolute) error rate over conventional Cepstral Mean Substraction. It is also important to mention that the same system yielded an EER of 1.7% in speaker verification on the XM2VTS database and less than 0.5% EER with multi-modal fusion [2].

This first set of calibration results shows that the BANCA database is clearly more realistic that the XM2VTS, especially when tested under protocol P (1 single enrollment session). While the low quality microphone yields worse results that the good quality one, the degradation does not seem to be excessive. The large improvement obtained with protocol G (error rates divided by 2) shows the importance of introducing more representative enrollment material to improve the performance. In this respect an incremental enrollment scheme [3] can be expected to be very beneficial in a practical application. Finally, the level of performance obtained with speech only leaves room for a significant improvement using the fusion of speech and vision modalities.

	$R = 0.1$		$R = 1$		$R = 10$		EER	
A post. threshold	0.0	12.2	3.4	1.0	8.6	0.3	2.1	2.2
	1.1		2.2		1.1		2.2	
A priori threshold	1.3	4.8	2.1	3.2	8.6	0.6	3.4	1.6
	1.6		2.7		1.4		2.5	

Table 5: Results for $g2$, protocol G, good quality microphone.

		EER a posteriori	HTER a priori
mic1	P	6.0	5.8
	G	2.8	3.3
mic2	P	8.1	8.7
	G	3.8	3.7

Table 6: Summary of results with both microphones.

6 Conclusion

In this paper, we have presented a new multi-modal database and its associated protocol that can be used for realistic identity verification tasks using up to two modalities: video sequences or images of the subject as well as a speech utterance of the same subject. We have presented a first series of experiments using a state-of-the-art text-independent speaker verification system, which exposed the advantages of such database, in particular its realistic character presenting a good scientific challenge. We hope that other researchers interested in multi-modal identity verification will use the database in order to compare their future research solutions.

7 Acknowledgments

This research has been partly carried out in the framework of the European BANCA project, which was also funded by the Swiss OFES project number 99-0563-1 for IDIAP.

References

[1] S. Bengio, F. Bimbot, J. Mariéthoz, V. Popovici, F. Porée, E. Bailly-Baillière, G. Matas, and B. Ruiz. Experimental protocol on the BANCA database. Technical Report IDIAP-RR 02-05, IDIAP, 2002.

[2] S. Bengio, C. Marcel, S. Marcel, and J. Mariéthoz. Confidence measures for multimodal identity verification. Technical Report IDIAP-RR 01-38, IDIAP, 2001.

[3] C. Fredouille, J. Mariéthoz, C. Jaboulet, J. Hennebert, C. Mokbel, and F. Bimbot. Behavior of a bayesian adaptation method for incremental enrollment in speaker verification. In *ICASSP2000*, Istanbul, Turkey, June 5–9 2000.

[4] J. Lüttin and G. Maître. Evaluation protocol for the extended M2VTS database (XM2VTSDB). Technical Report RR-21, IDIAP, 1998.

[5] A. Martin et al. The DET curve in assessment of detection task performance. In *Eurospeech'97*, volume 4, pages 1895–1898, 1997.

[6] J. Pelecanos and S. Sridharan. Feature warping for robust speaker verification. In *Proc. Speaker Odyssey 2001 conference*, June 2001.