EXPERIMENTAL EVALUATION OF TEXT-INDEPENDENT SPEAKER VERIFICATION ON LABORATORY AND FIELD TEST DATABASES IN THE M2VTS PROJECT

L. Besacier[1], *J. Luettin*[2], *G. Maître*[2,3], *E. Meurville*[1,4]

(1) IMT, Neuchâtel (CH) - laurent.besacier@imt.unine.ch
(2) IDIAP, Martigny (CH) - luettin@idiap.ch
(3) now at EIV, Sion (CH) - Gilbert.Maitre@eiv.ch
(4) now at EPFL, Lausanne (CH) - Eric.Meurville@epfl.ch

ABSTRACT

This paper describes experiments of a text-independent speaker verification method that has been evaluated on two laboratory databases (*M2VTS* and *XM2VTS*) and on one field test database (*LoCoMic*). This work has been performed within the European ACTS-M2VTS project (Multi-Modal Verification for Teleservices and Security Applications) which is concerned with person authentication using multiple modalities. The system achieved good performance on the M2VTS and XM2VTS databases whereas the performance decreased on the LoCoMic database which is more representative of real conditions. In fact, we have shown that in text independent mode, the performance differs significantly according to the test item considered (digits, name, and sentence…). This shows that, even in text independent mode, the access control system should propose a structure for the uttered sentence.

1. INTRODUCTION

1.1 The M2VTS Project

Among the different European ACTS projects, the M2VTS project (Multi Modal Verification for Teleservices and Security applications - Project AC102) deals with the security aspect and considers access control by the use of multimodal biometric authentication. The motivation for using a multimodal recognition scheme is to improve the recognition efficiency by combining single modalities, namely face and voice features [1].
Further information about M2VTS can be obtained by visiting our Web site: *http://www.tele.ucl.ac.be/M2VTS* .

1.2 Databases recorded during the project

M2VTS Multimodal Face Database

This database [2] is made up of 37 different speakers and provides 5 recordings (shots) for each person. These shots were taken at one-week intervals. During each shot, people have been asked to count from '0' to '9' in French, rotate the head from 0 to -90 degrees, again to 0, then to +90 and back to 0 degrees. For each person belonging to the database, the most difficult shot to recognize is labeled as the 5th shot. These shots mainly differ from the others because of face variations (head tilted, eyes closed, different hairstyle, presence of a hat/scarf...), voice variations or shot imperfections (poor focus, different zoom factor, poor voice SNR...). Concerning voice acquisition, the sound track is digitally recorded using a 48 kHz sampling frequency and 16 bit linear encoding. Further information about the M2VTS database can be obtained on: *http://www.tele.ucl.ac.be/M2VTS/m2fdb.html* .

The extended XM2VTS Multimodal Face Database

In acquiring the XM2VTS database [3], 295 persons from the University of Surrey were recorded at four sessions at approximately one-month intervals. On each session, two recordings (shots) were made. Each recording contains a speaking head shot and a rotating head shot. Sets of data taken from this database are available including high quality color images, 32 KHz 16-bit sound files and video sequences. At the third session a high-precision 3D model of the subject's head was built using an active stereo system provided by the Turing Institute. Further information about XM2VTSDB can be obtained on: *http://www.ee.surrey.ac.uk/Research/VSSP/xm2fdb/* .

The LoCoMic Database

This database has been recorded at IDIAP for the evaluation of speaker verification methods under realistic conditions (recording with low-cost microphone). The database includes speech recordings of 22 persons, performed in 10

sessions. Each session contains 9 items spoken in French with the following content: item01 (name), item02 (date of birth), item03 (address and phone number), item04 (the ten digits in ascending order), item05 (the ten digits in a random order, the same for all the sessions of the same speaker, but different for each speaker), item06 (a sentence, the same for all speakers and for all sessions), item07 (a sentence, the same for all the sessions of the same speaker, but different for each speaker), item08 (a sentence, different for each session and for each speaker), item09 (spontaneous speech with comments of a picture, picked at random in each session).

The recorded audio signals are quantized at 16 bits and sampled at 16 kHz. The database has been recorded on a Pentium PC with a Soundblaster board and a low-cost omni-directional microphone placed at a distance of about 50-100 cm from the speaker's mouth. The microphone had an automatic gain control.

This database is difficult for the speaker verification task : speech was collected over several months with a low-cost microphone in a reverberant room. Moreover, no particular advice was given to the speakers during the recording, which resulted in variable speaker position and orientation relative to the microphone.

2. TEXT INDEPENDENT SPEAKER VERIFICATION SYSTEM

2.1 Technology

Speaker verification measure

The speaker verification method used in all the experiments presented in this paper is inspired from second-order statistical tests on covariance matrices, computed from acoustic parameters [4]. The symmetrical sphericity measure we use is easy to implement, computationally efficient, and has shown to give good results in text-independent mode.

Let X and Y denote two covariance matrices of a reference speaker and of a test speaker respectively, corresponding to the covariance of some acoustic vectors computed for the training and test utterance, respectively. Let M and N denote the number of the vectors of the training and test utterance, respectively, and p the dimension of these vectors. The mathematical expression of the symmetrical sphericity measure that we used in our experiments is then:

$$\mu_{Sc}(X,Y) = \frac{M}{M+N} log(tr(YX^{-1})) + \frac{N}{M+N} log(tr(XY^{-1})) - \frac{1}{p}.\frac{M-N}{M+N} log\left(\frac{det\,Y}{det\,X}\right) - log(p)$$

where 'tr' denotes the trace and 'det' the determinant of a matrix.

Signal analysis

First, the silences in the speech utterances are automatically removed. The speech analysis module extracts 12 Linear Frequency Cepstral Coefficients and 1 coefficient corresponding to the average energy of a speech frame. This energy coefficient is then normalized between 0 and 1. Under these analysis conditions, each acoustic vector is 13 dimensional, while covariance matrices are 13x13 dimensional.

The effects on the performance, of cepstral mean subtraction (subtraction of the mean acoustic vector to each instantaneous feature vector - CMS) and of time domain subtraction (subtraction of the mean signal to each sample – ZeroMean) are notably studied for the experiments on LoCoMic and XM2VTS, described in *section 3*.

2.2 Methodology

An important contribution of the M2VTS project was the definition of experimental protocols in order to test speaker verification systems on databases. These protocols are precisely described in this paragraph.

M2VTS Database Protocol

Only four shots of the M2VTS database are used during in the experiments. The protocol chosen follows the « leave one out » principle [2]. Each *experiment* session uses a *training* and a *test* database. The training database is built of 3 shots (4 are available) of 36 persons (37 available). The test database is built of the left-out shot of the left-out person (used as impostor) and the left-out shot of the 36 persons (clients) present in the training database.

The training database is used to build a reference model for each client. The performance of the identification algorithms is evaluated by matching the 37 candidate persons (36 clients and 1 impostor) from the test database with the 36 reference clients. Such an experiment session provides 36 *authentic* and 36 *imposture* tests. An authentic test consists of candidate claims which

are true. An imposture test consists of candidate claims which are false. 4 different training/test configurations are possible for each experiment session of each speaker. We have then 5328 (36x4x37) authentic and 5328 impostor tests for this database.

XM2VTS Database Protocol

The database was divided into three sets: training set, evaluation set, and test set. The training set is used to build client models and the evaluation set is selected to produce client and impostor access scores which are used to find a threshold that determines if a person is accepted or rejected. The threshold can be set to satisfy certain performance levels on the evaluation set. In the case of multi-modal classifiers, the evaluation set might also be used to optimally combine the outputs of several classifiers. The test set is selected to simulate real authentication tests. The three sets can also be classified with respect to subject identities into client set, impostor evaluation set, and impostor test set. For this description, each subject appears only in one set. This ensures the realistic evaluation of imposter claims whose identity is unknown to the system.

The protocol is based on 295 subjects, 4 recording sessions, and two shots (repetitions) per recording sessions. One shot consists of the two audio digit sequences and of one still image. The database was randomly divided into 200 clients, 25 evaluation impostors, and 70 test impostors.

Two different evaluation configurations were defined. They differ in the distribution of client training and client evaluation data. More details can be found in [5].

LoCoMic Database Protocol

The training of the speaker models is either made with the first full session (all of the 9 items) or with *Item6* (*Item 6* corresponds to the longest sentence) of the first session alone, since in most applications, one can afford just a single enrolment session. Tests are performed for each item separately. Thus, 1782 authentic accesses (9 remaining sessions, 22 clients, and 9 items) and 4158 impostor accesses (only 1 session, 21 impostors for each of the 22 clients, and 9 items) are made.

3. ACHIEVED PERFORMANCE

3.1 Performance on M2VTS DB

The Equal Error Rate (EER) obtained on M2VTS with 5328 client accesses and 5328 impostor accesses (*a posteriori* person-dependent threshold) is 3.58%.

3.2 Performance on XM2VTS DB

The EER obtained on the evaluation set of XM2VTS DB are given in *Tab. 1* (signal downsampled to 8kHz – 600 client accesses, 40000 impostor accesses - *a posteriori* person-dependent threshold).

Sampling	Baseline	CMS	CMS + ZeroMean
8kHz	3.30%	1.50%	2.92%

Table 1 : Performance on XM2VTS DB

3.3 Performance on LoCoMic DB

The performance obtained on the field test database, with the configuration *CMS + ZeroMean*, is reported in *Fig. 1* and *Fig. 2* where the False Rejection rate (FR) is represented versus the False Acceptance rate (FA), obtained with a person-independent threshold (198 client accesses, 462 impostor accesses, for each item). For *Fig. 1*, the training of the speaker models is performed with the first full session (approximately 60 s of speech per model) whereas in *Fig. 2*, the training is done on *Item6* (approximately 10 s of speech per model) of the first session alone. Access tests for *item1* and *item2* are not reported since these utterances are too short to achieve a reliable verification.

Moreover, in order to see the influence of speech duration on the performance, the EER obtained on different test items (corresponding to the diagonals of *Fig. 1* and *Fig. 2*) are reported on *Tab. 2* and compared to the average item duration.

Test item	3	4	5	6	7	8	9
test duration (s)	7.86	6.11	6.29	9.44	7.45	7.84	15
EER train. session1	15.6	23.2	15.7	11.6	19.9	19.7	27.7
EER train. item6	24.2	35.4	28.8	10.6	31.6	24.2	27.1

Table 2 : EER for different training and test durations (LoCoMic - 198 client accesses / item, 462 impostor accesses / item).

4. DISCUSSION

The results on XM2VTS DB show that Cepstral

Mean Subtraction improves the performance whereas the effect of Zero Mean is less conclusive. The results obtained on the field-test database (*LoCoMic*) show that speaker verification performance differs significantly according to the kind of speech used : spontaneous / read, duration (error rates from 10% to 30%). The best performance is obtained when the verification is performed with the longest read sentence (*item6*). For test on spontaneous speech (*item9* – description of an image), the performance is very low, although *item9* is the longest item. These studies confirm the known observation that performance increases with the length of the test utterance, but the performance also differs largely between read and spontaneous speech (cf. *item 9*). As expected, the results are better when the training is done on a full session. This fact is not observed when training and test are both made on *item6* since in that case, the task is almost *text-dependent* and performance is thus much higher.

5. CONCLUSION

This paper aims at reporting protocols and speaker verification performance obtained with a well known text independent method on 3 databases. The system achieved good performance on M2VTS and XM2VTS databases whereas the performance decrease on the LoCoMic database which is more representative of real conditions. In fact, we have shown that in text independent mode, the performance differs significantly according to the test item considered (digits, name, and sentence…). This shows that, even in text independent mode, the access control system should propose a structure for the uttered sentence, impose a minimum length for the verification utterance, and control the distance to the microphone (which was not the case for LoCoMic database).

Acknowledgement : this work has been performed within the framework of the M2VTS project granted by the European ACTS program.

6. REFERENCES

[1] G. Richard, Y. Menguy, I. Guis, N. Suaudeau, J. Boudy, P. Lockwood, C. Fernandez, F. Fernández, C. Kotropoulos, I. Pitas, R. Heimgartner, P. Ryser, C. Beumier, S. Pigeon, G. Matas, J. Kittler, J. Bigün, Y. Abdeljaoued, E. Meurville, L. Besacier, G.Maitre, J. Luettin, S. Ben-Yacoub B. Ruiz. "Multi Modal Verification for Teleservices and Security Applications (M2VTS)" *In Proc. IEEE Conference on Multimedia Computing and Systems'99*. Florence, Italy, 7-11 June 1999.

[2] S. Pigeon, L. Vandendorpe. "The M2VTS Multimodal Face Database". In *Proc. Audio and Video-based Biometric Person Authentication (AVBPA),* Springer LNCS, Bigün et al. Eds, 1997.

[3] K. Messer, J. Matas, J. Kittler, J. Luettin and G. Maitre. "XM2VTSDB : the extended M2VTS database". In *Proc. AVBPA 99.* Washington DC, USA. 22-23 March 1999.

[4] F. Bimbot, I. Magrin-Chagnolleau, L. Mathan. "Second-order statistical methods for text-independent speaker identification" *Speech Communication*, n°.17(1-2), August 1995.

[5] J. Luettin and G. Maitre. "Evaluation protocol for the extended M2VTS database (XM2VTS)". *Technical Report IDIAP-COM 98-05.* IDIAP, 1998.

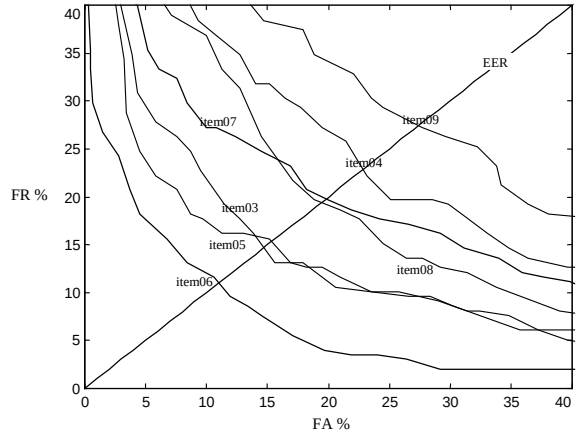

Figure 1: FR (198 accesses) versus FA (462 accesses) rates obtained for test on different items - training made on the first full session – LoCoMic DB

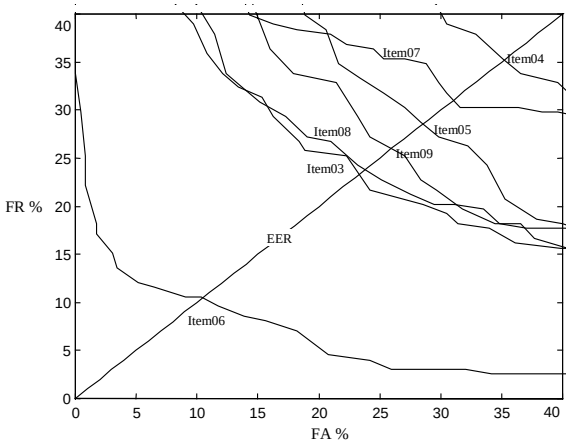

Figure. 2: FR (198 accesses) versus FA (462 accesses) rates obtained for test on different items - training made on item6 of the first session - LoCoMic DB