The BioSecure Talking-Face Reference System

Hervé Bredin [1], Guido Aversano [1], Chafic Mokbel [2] and Gérard Chollet [1]
[1] CNRS-LTCI, GET-ENST (TSI Department), 46 rue Barrault, 75013 Paris, France
[2] University of Balamand, El Koura, BP 100, Tripoli, Lebanon
{bredin, aversano, chollet}@tsi.enst.fr, chafic.mokbel@balamand.edu.lb

Abstract

In the framework of the BioSecure Network of Excellence (http://www.biosecure.info), a talking-face identity verification reference system was developed: it is open-source and made of replaceable modules. This is an extension of the BECARS speaker verification toolkit, implementing the GMM approach. In this paper, the audio and visual features extraction front-ends are presented. The performance of the system on the Pooled protocol of the BANCA database are described.

1. Introduction

In the framework of identity verification, it has been noticed that it is very difficult (if not impossible) to compare two different methods from two different articles in the literature, even though they deal with the very same task. It poses a real problem when one wants to know if a new original method performs better than the current state of the art, for example. This can be explained by the fact that a lot of research laboratories own their own test database and are the only one performing experiments on it, which are subsequently impossible to reproduce. Reference systems bring an easy yet efficient answer to this problem. Since they are open-source and freely available for everybody, when publishing results on a specific database, experiments using the reference system can be added as a way of calibrating the difficulty of this particular database.

Developing a reference system made of replaceable modules is also of great interest. Researchers often work on a specific part of the system and do not have the time nor the interest in building a complete system from A to Z. A researcher could show the improvement of his new features extraction algorithm simply by replacing the corresponding module in the reference system and without having to bother about the pattern recognition algorithm.

On a pragmatic side, using a reference system as a basis for researching a specific area is also a good way to save time, human resources and money and therefore to facilitate advances of the state of the art.

This reference system addresses the relatively new area of identity verification based on talking-faces. This biometric modality is intrinsically multimodal. Indeed, not only does it contain both voice and face modalities, but it also integrates the combined dynamics of voice and lips motion. Identity verification based on talking-faces is a growing subject of research in the recent literature. In [8], fusion of speech, face and visual speech information for text-dependent identification is presented. In this purpose, the authors use the HTK Speech Recognition Toolkit[1] for speech features extraction and Hidden Markov Model (HMM) modeling.

Since our system is designed to perform text-independent identity verification, it uses the Gaussian Mixture Model (GMM) approach for each of the three modalities. GMM for speaker verification is well-known as being very efficient [12]. However, GMM for video-based face recognition is relatively new. It aims at improving robustness of face recognition against light changes, pose variations, etc. In [8], the GMM approach for mouth-based identity verification was concluded to be sufficient (compared to HMM) but not tested.

Therefore, our system mainly consists in an audiovisual front-end extension of the existing open-source BECARS speaker verification GMM toolkit [2]. It also includes a module allowing the detection of basic replay attacks using the synchronization between voice dynamics and lip motion [3].

Section 2 quickly overviews the BECARS toolkit. The new open-source GET-ENST Online Speech Processing Evaluation Library (GOSPEL) is introduced in section 3. It was developed in the aim of being portable to embedded devices such as PDA and SmartPhones. The face and lips visual front-end is described in section 4. It is made of modules allowing face and mouth tracking, eigenfaces features and lips features extraction. A simple yet efficient algorithm tackling replay attacks is quickly described in section

[1] `http://htk.eng.cam.ac.uk/`

5. A more detailed description of the algorithm and its performance in simple replay attacks scenarios is available in [3]. In section 6, the question of the fusion of these different modalities is discussed. Sections 7 and 8 report about the experiments and corresponding results performed on the widely available audiovisual BANCA database. Section 9 draws conclusions and presents our plan and perspective to improve the Biosecure Talking-Face Reference System.

2. Speaker Verification Algorithm

Speech is a biometric modality that may be used to verify the identity of a speaker. The speech signal represents the amplitude of an audio waveform as captured by a microphone. To process this signal a feature extraction module calculates relevant feature vectors on a signal window that is shifted at a regular rate. In order to verify the identity of the claimed speaker a stochastic model for the speech generated by the speaker is generally constructed. New utterance feature vectors are generally matched against the claimed speaker model and against a general model of speech that may be uttered by any speaker called the world model. The most likely model identifies if the claimed speaker has uttered the signal or not. In text independent speaker recognition, the model should not reflect a specific speech structure, i.e. a specific sequence of words. Therefore in state-of-the-art systems, Gaussian Mixture Models (GMM) are used as stochastic models.

Given a feature vector **x**, the GMM defines its probability distribution function as follows:

$$\sum_{i=1}^{N} w_i \frac{1}{\sqrt{(2\pi)^d \|\Gamma_i\|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_i)^T \Gamma_i^{-1} (\mathbf{x}-\mu_i)\right) \tag{1}$$

This distribution can be seen as the realizations of two successive processes. In the first process, the mixture component is selected and based on the selected component the corresponding Gaussian distribution defines the realization of the feature vector. The GMM model is defined by the set of parameters $\lambda = (\{w_i\}, \{\mu_i\}, \{\Gamma_i\})$. To estimate the GMM parameters, speech signals are generally collected. The unique observation of the feature vectors provides incomplete data insufficient to allow analytic estimation, following the maximum likelihood criterion, of the model parameters, i.e. the Gaussian distributions weights, mean vectors and covariance matrices. The Estimation Maximization (EM) algorithm offers a solution to the problem of incomplete data [7]. The EM algorithm is an iterative algorithm, an iteration being formed of two phases: the Estimation (E) phase and the Maximization (M) phase. In the E phase the likelihood function of the complete data given the previous iteration model parameters is estimated. In the M phase new values of the model parameters are determined by maximizing the estimated likelihood. The EM algorithm ensures that the likelihood on the training data does not decrease with the iterations and therefore converges towards a local optimum. This local optimum depends on the initial values given to the model parameters before training. Thus, the initialization of the model parameters is a crucial step. The LBG algorithm is used to initialize the model parameters.

The direct estimation of the GMM parameters using the EM algorithm requires a large amount of speech feature vectors. This causes no problem for the world model where several minutes from several speakers may be collected for this purpose. For the speaker model, this would constrain the speaker to talk for a long duration and may not be acceptable. To overcome this, speaker adaptation techniques may be used [2]. In the current work, BECARS [2] software has been used for speaker recognition. BECARS implements GMM and includes several adaptation techniques, i.e. Bayesian adaptation, maximum likelihood linear regression (MLLR), and the unified adaptation technique defined in [10]. Using the adaptation techniques few minutes of speech become sufficient to determine the speaker model parameters.

During recognition, feature vectors are extracted from a speech utterance. The log likelihood ratio between the speaker and world models is computed and compared to a threshold. This allows to verify the identity of the claimed speaker.

3. Audio Front-End

One of the goals of our research is the realization of a real-time talking-face verification interface that can also run on limited resource platforms such as PDAs or SmartPhones. To this purpose, we have developed and validated a new software module for speech parameterization, named “GOSPEL” (GET-ENST Online Speech Processing Embedded Library for prototyping and evaluation).

This library, written in ANSI C, is compatible with the UNIX, Windows and Windows CE platforms. The GOSPEL programming interface has been specially designed for being used within a reference system for scientific evaluation and for being integrated, without extra efforts, into demonstrators and commercial prototypes.

The current version of GOSPEL allows for MFCC feature extraction (including standard options such as delta calculation, cepstral mean subtraction, or liftering) from online or offline audio. It also provides buffering mechanisms suitable for multithreaded applications. A diagram representing typical operations performed by the GOSPEL audio front-end is shown in figure 1.

The “running-CMS” option of GOSPEL makes possible to perform cepstral mean subtraction (CMS) without requiring that a whole utterance is recorded. The online estima-

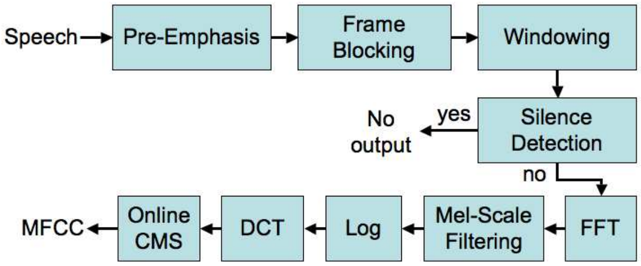

Figure 1. Audio front-end: GOSPEL

tion of cepstral mean is obtained by running-average over the cepstral vector sequence.

An online silence/voice detection function is also provided. Discrimination between speech and silence is based on energy thresholding, with either fixed or exponential adaptive thresholds.

Moreover, GOSPEL supports fixed-point arithmetic: fixed-point optimization, that can be chosen at compile time, is exploited to achieve faster processing throughput on those platforms which do not provide a floating-point unit (like nowadays SmartPhones and PDAs).

The GOSPEL library has been intensively tested and evaluated in speaker verification experiments on the BANCA database. Verification accuracy results, for all the features described above (running-CMS, online silence detection, fixed-point optimization), are given in section 8.

4. Visual Front-End

The visual front-end is divided into two parts, performing the features extraction for face and lips respectively, as shown in figure 2. First, face is tracked in the video and face features are extracted. Then, for each frame of the video, within the detected face, lips are located and lips features extraction is performed. It was developed using the *Open CV* open-source C/C++ library, freely available over the Internet [2].

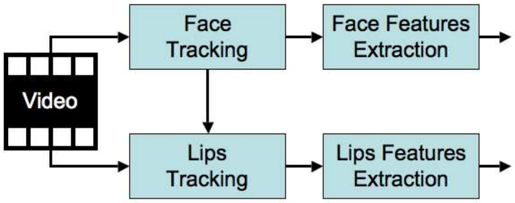

Figure 2. Visual front-end

4.1 Face module

4.1.1 Face detection

The *OpenCV* library face detector algorithm is first used to get a rough idea of the location of the face: it is an implementation of Viola's algorithm [14], well-known for being very efficient and very fast. The bounding box of the resulting face candidates is then used as a region of interest where to look for a face in the second step of the algorithm. Figure 3 shows the face candidates and the corresponding bounding box on a sample from the BANCA database [1]. Within this region of interest, a moving window scans every possible rectangle at every position with many sizes. For each candidate, the Distance From Face Space (DFFS) [11] is computed and the candidate with the lowest DFFS is chosen as the location of the face. It is defined as the distance between the face candidate and its projection in the eigenface space [13]. Figure 4 summarizes how this distance is computed. A temporal median filter is then applied on the location and size of the detected faces in the video in order to avoid local detection problems. Figure 5 shows the final result of face detection on the same example as before.

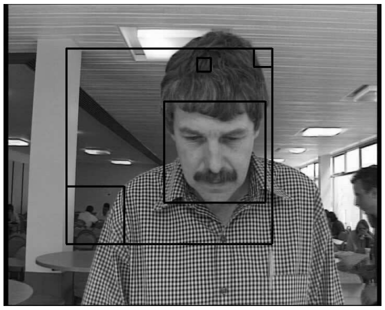

Figure 3. Face candidates bounding-box

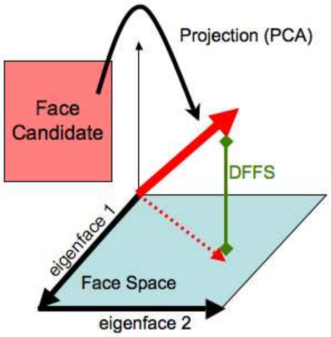

Figure 4. Distance From Face Space

These two steps need a preliminary training phase. We used the frontal-face Haar cascade available in *OpenCV* to

[2] http://www.intel.com/research/mrl/research/opencv/

find the first face candidates and its corresponding bounding box. The principal component analysis (PCA) needed for the computation of the DFFS is learned based on a training set extracted from the BANCA database (see section 7 for more details) and using the PCA-related functions available in *OpenCV*.

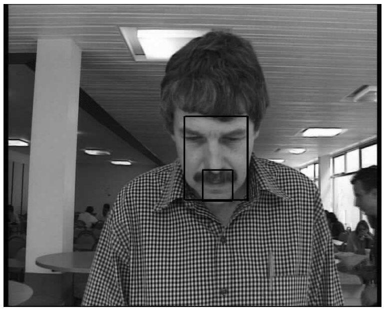

Figure 5. Face and lips tracking

4.1.2 Face tracking

Since this exhaustive search for the best face candidate is very CPU-consuming, a simple tracking algorithm is used: given the location and size of the face in frame $n-1$, the face in frame n is searched in its neighborhood, allowing only a small difference in size. To avoid any divergence in tracking, the algorithm is reinitialized every 20 frames.

4.1.3 Features extraction

Once the face is detected, it is size-normalized to the size of the previously learned eigenfaces. The decomposition of the detected face on the eigenfaces is computed and used as features for face recognition.

4.2 Lips module

4.2.1 Mouth Detection

The very same algorithm as in the first step of face tracking is applied for mouth detection. In each detected face, its lower part is searched for a mouth candidate using the Viola's algorithm. Thus, no real tracking of the lips is performed in this module: it would rather be considered as a *mouth area detector*. Hence, a Haar cascade is learned based on rectangle mouth images extracted from the BANCA database: the effective lips contour tracking is still being investigated since, in our knowledge, no open-source libraries or software for this task is available yet. As one can notice for face detection in figure 3, a lot of false mouth candidates may be detected. Then, a simple algorithm is applied: the biggest detected mouth candidate in the lower part of the face is chosen as the right one. A temporal median filter is then applied in order to avoid local detection problems. Figure 5 shows an example of the output of the mouth detection module.

4.2.2 Features extraction

Once the mouth area is detected, it is size-normalized to 64x64 and a Discrete Cosine Transform (DCT) is applied. Among these 4096 DCT features available, only 50 are kept as lips features. Their selection is performed based on a training set: the ones with the highest energy are chosen. DCT is performed using the *OpenCV* library.

5. Replay Attacks Detection

The talking-face modality is one of the biometrics the most likely to be defeated by replay attacks. As a matter of fact, it is based on the identification of a person using his/her voice and his/her face: two pieces of information which can easily be recorded (which is not that easy for iris or fingerprint, for example). Thus, an imposter could show to the camera a picture of the face of his/her target while playing a recording of the latter's voice previously acquired without any consent nor knowledge of the impersonated person.

This particular scenario (called *Paparazzi*) was proposed in [3], along with the *Big Brother* scenario where the imposter not only owns a picture of the face but a whole video of his/her target. In this paper, a replay attacks detection algorithm is also developed. It is based on a measure of correlation between two streams: one representing the voice dynamics and the other one representing the lip motion. The initial observation that led to this algorithm is presented in figure 6. The upper signal is the energy of speech and the bottom one is the openness of the mouth, both extracted from the same audiovisual sequence. The shadowed parts of the curves emphasize how similar and correlated these two signals can be.

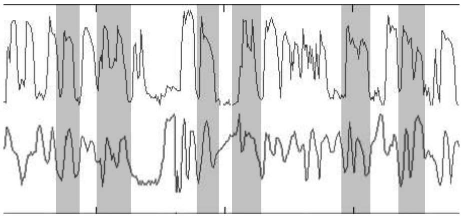

Figure 6. Speech energy vs. Mouth openness

Preliminary results with features as simple as the log-energy of the audio signal and the average value of gray level pixel of the mouth area for the visual signal give encouraging results for future improvements (see [3] for more details).

6. Fusion

6.1. Score fusion

Score fusion consists in the combination of the scores of two or more monomodal identity verification algorithms. In [6], this kind of fusion has already been studied using multiple face recognition algorithms on the BANCA database. We used the open-source Support Vector Machine (SVM) library *libSVM* [4] to perform fusion of speaker verification and face recognition scores. More precisely, a Support Vector Classifier with a linear kernel is learned and applied in the 2-dimensional bimodal score space, after a preliminary normalization step.

6.2. Feature fusion

Feature fusion consists in the combination of two or more monomodal feature vectors into one multimodal features vector to be used as the input of a common multimodal identity verification algorithm.

Audio and visual frame rates are different. Typically, 100 audio feature vectors are extracted per second whereas only 25 video frames are available during the same period. Therefore, one solution is to perform linear interpolation of the visual feature vectors. Another one is to downsample the audio features to reach the video frame rate.

Only simple concatenation of audio and visual feature vectors has been investigated so far. As expected (yet, it still had to be tested), the concatenation-based system is worse than the best monomodal system (see results in section 8). However, more elaborated combination methods still need to be investigated. For example, a transformation such as Principal Component Analysis, Independent Component Analysis or Linear Discriminant Analysis might intrinsically model the correlation between voice dynamics and lips motion. The open-source pattern classification libraries Torch [3] or LNKnet [4] will be used for this purpose.

7. Experiments

7.1 The BANCA database

The BANCA audiovisual database [1] contains 52 speakers divided into 2 groups G1 and G2 of 26 speakers each (13 females and 13 males). 12 sessions were recorded in 3 different conditions (controlled, adverse and degraded). In each session and for each speaker, 2 recordings were performed: one client access where the speaker pronounces digits and his/her (fake) own address and one impostor access where he/she pronounces digits and the address of another person.

7.2 The Pooled BANCA Protocol

The experiments are performed following the Pooled BANCA protocol [1]. In each modality (voice, face, lips and concatenation of voice and lips), for each subject, a GMM is adapted from a world GMM using only the features extracted from the client access of the first controlled session. 312 imposter (one per client per session) and 234 (one per client per session, except the first one used for training) client accesses are performed for each group. The world GMM is learned on the features extracted from the 20 controlled videos of the world model (more than 11000 visual samples).

The face space (used for the face tracking algorithm and the eigenface projection) is built using the manually located face from the world model of the English still images BANCA dataset (300 faces from 30 different subjects).

7.3 Extracted features

In our experiments the BANCA audio (from the first, high-quality, microphone) has been resampled to 16kHz. Speech preprocessing is performed on 20 ms Hamming-windowed frames, with 10 ms overlap. For each frame, 15 MFCC coefficients and their first-order deltas are extracted in the full frequency range, with 20 MEL-scaled triangular filters.

Automatic face tracking is performed on every video of the BANCA database and 80 eigenface coefficients are extracted from each frame (about 400 frames or more per video). Similarly, 50 DCT coefficients are extracted from the mouth area (size-normalized to 64x64), for each frame of each video. Among the 4096 DCT coefficients, the 50 with highest energy (in the world model) are kept as the most significant.

7.4 Score normalization and fusion

In order to achieve good results during the score fusion process, scores have to be normalized so that scores from different modalities vary in the same predefined range of values.

For that purpose, we used the fact that groups G1 and G2 are two completely distinct sets of subjects: no cross access is performed between them. A linear transformation is learned on scores from G2 to constrain them between -1 and 1 and the SVM classifier is trained on G2. Then, the

[3] http://www.torch.ch/

[4] http://www.ll.mit.edu/SST/lnknet/

same linear transformation is applied on scores from G1, on which the SVM classifier is applied.

8. Results

8.1 Voice

Several “online” speech preprocessing techniques (provided by our audio front-end and described in section 3) have been evaluated. All the speaker verification experiments have been performed following the BANCA P protocol and using our BECARS-based GMM classifier, with 128 gaussians. Speaker models are obtained by MAP adaptation (adapting just the mean of the distributions) from a gender-independent world model (trained on the “controlled” part of BANCA world model data). The GMM training/testing is done only on frames detected as speech, that is frames whose total energy exceeds a given threshold. Unless otherwise stated this threshold is fixed to a very low value (corresponding to an average signal power of about -70 dB compared to full saturation).

Validation of the GOSPEL library Some tests have been conducted to validate our new audio front-end GOSPEL against the previously adopted HTK speech parameterization module, using the same configuration for both of them. The GOSPEL module produced a small improvement for the equal error rate (+0.6% averaged on both BANCA speaker groups).

Thus, we concluded that the two software modules are equivalent for standard MFCC parameterization, within statistical errors.

Online CMS The “running-CMS” algorithm, implemented in GOSPEL, as been tested and compared to standard offline CMS. Results show that online cepstral mean estimation does not deteriorate verification performance. On the contrary an increased accuracy is obtained on both BANCA groups (5.8% EER against 6.4% EER on group G1, 7.4% EER against 7.7% EER on group G2). Figure 7 shows DET curves for the G1 case. The grey regions in the plot represent 95% confidence intervals for our tests. The darkest region correspond to a large-sample approximation of the confidence interval (which is optimistic, considering the size of the BANCA database). The lighter grey shading corresponds to the most pessimistic limit (the so-called Chernoff limit [9]), for the confidence interval.

Considering confidence intervals, we can conclude that running-CMS performs as well as standard CMS on BANCA protocol P.

Fixed-point processing Fixed-point voice feature extraction (using an approximated representation of fractional numbers on 16-bit integers) has also been tested. As figure 7 shows, we observe a degradation in terms of verification performance (about -3.5%, averaged on both BANCA groups), compared to the floating-point case. This difference has the same order of magnitude as the confidence intervals. This loss in accuracy corresponds to a considerable gain in terms of processing speed on limited resource devices. We have tested our library on a SmartPhone equipped with an Intel PXA263 processor that does not provide a floating-point unit. On this platform, the optimized part of the algorithm runs about 3.5 times faster than its floating-point correspondent.

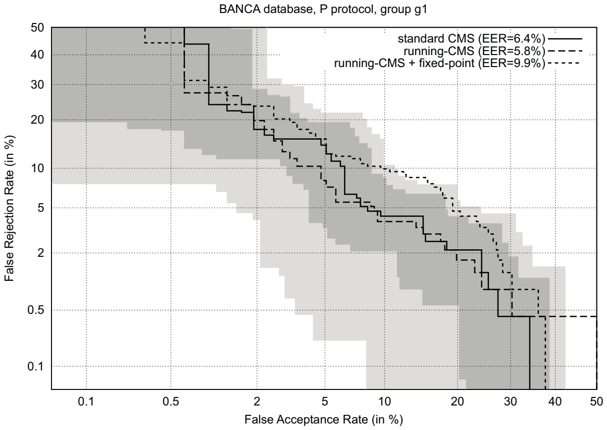

Figure 7. Running-CMS and fixed-point processing

Adaptive silence thresholding The experiments presented in this section compare verification accuracy for both fixed and adaptive silence thresholding. Firstly, a baseline threshold as been estimated on the world model data, by fitting the distribution of frame energy with two gaussians. Then, the energy threshold for silence deletion has been fixed to $\mu_s - 2\sigma_s$, where μ_s and σ_s are the mean and the standard deviation of the rightmost gaussian. This threshold value has been either kept fixed or used as an initialization for the adaptive thresholding (thresholding is reinitialized for each sentence). Figure 8 shows that, for the BANCA P protocol, adaptive thresholding performs significantly better than a fixed threshold approach, giving 5.4% EER on G1 and 4.8% EER on G2.

8.2 Face and lips

Figures 9 and 10 present the performance of GMM modeling for identity verification based on face and lips respec-

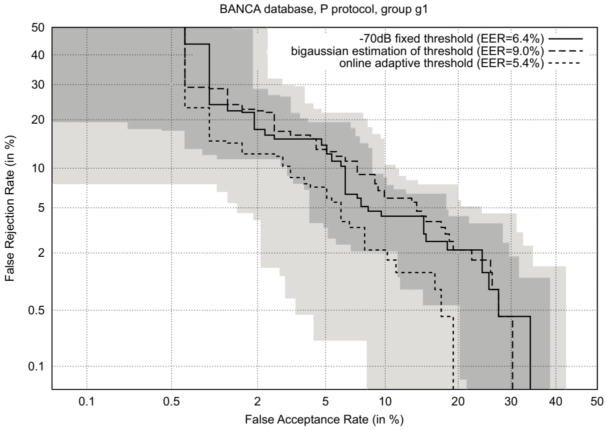

Figure 8. Different silence detection methods

tively. For face recognition, using 32 or 64 gaussians gives the best result: around 28% Equal Error Rate (EER). These relatively poor results can be explained by the simplistic features used to model face: eigenfaces with no normalization of any kind (rotation of the head, light changes, etc.). Lips-based recognition reaches at best 34% EER for 64 gaussians, whereas all the other systems with less or more gaussians stand around $38 - 39\%$. The same kind of performance was achieved on G2 (not plotted).

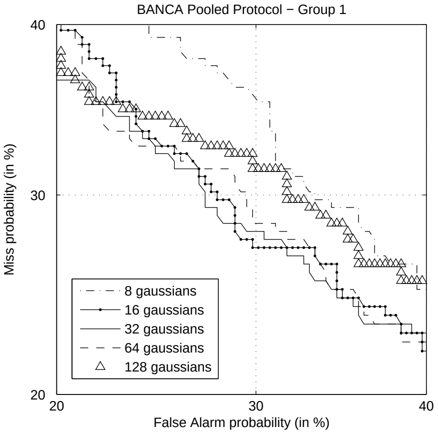

Figure 9. GMM on face features

8.3 Feature fusion

Figure 11 presents the result of the experiments we performed about feature fusion. Lips feature vectors were linearly interpolated to reach the audio frame rate. Then, a simple concatenation of lips feature vectors and voice feature vectors was performed. Voice only (with 64 gaussians) gives an EER of 8.5%, lips only (still with 64 gaussians) gives an EER of 34%. Combining them strongly degrades the performance (compared to voice only) and gives an EER of 32%.

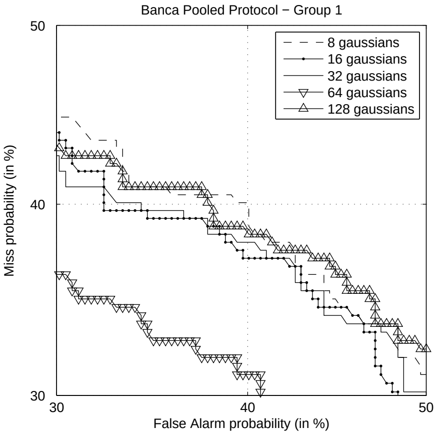

Figure 10. GMM on lips features

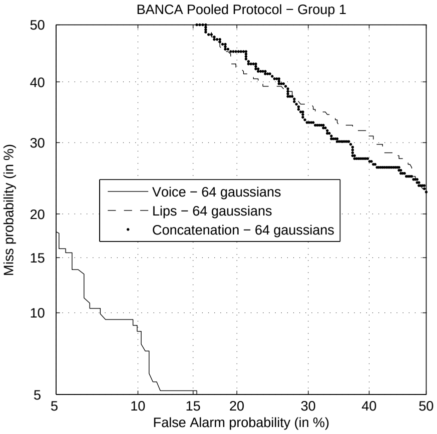

Figure 11. Fusion of voice and lips features

8.4 Score fusion

Following the process described in section 7, we performed score fusion using an SVM classifier with linear kernel: results are presented in figure 12. Voice only (with 256 gaussians) gives an EER of 8.1%, face only (with 256 gaussians) gives an EER of 31%. Performing score fusion brings a non-significant improvement over the voice only systems: 7.7% EER.

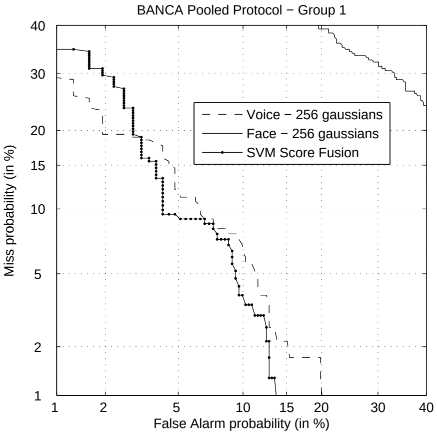

Figure 12. Score fusion with SVM

9. Conclusion and future work

The BioSecure Talking-Face reference system has been introduced in this paper. It is based on the open-source software BECARS initially developed for speaker verification. Our new audio front-end performs “live” feature extraction, including online CMS and silence deletion. Moreover it is implemented both with floating and fixed point operations, which makes it usable on portable devices such as PDA or SmartPhones. The usability of the GMM approach for face- and lips-based recognition was also demonstrated. The reference system makes extensive use of open-source libraries and is freely available on request to the authors. An original way of using the intrinsic bimodal nature of talking-faces has been reported: the detection of a lack of correspondence between the voice and the lips motion is of great help when dealing with simple replay attacks.

In the future, using this reference system as a basis, we plan to improve some of the modules. More precisely, much more efficient face tracking algorithms based on Active Appearance Modelling (AAM) [5] can be investigated. For that purpose, we plan to use the open-source AAM library available on the internet[5]. This might as well help to lead to an improved lips tracking algorithm and consequently the replay attacks detection module. Finally, though eigenface coefficients have been used as a reference in the field of face recognition, better features can be extracted: promising KCFA [15] shall be investigated, for example.

10. Acknowledgments

This work was supported by the EC BioSecure IST-2002-507634 project, the EC SecurePhone IST-2002-506883 project and by the Paris Institute of Technology (ParisTech). The authors would like to thank Santa Rossi for her contribution to lip tracking.

References

[1] E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran. The BANCA Database and Evaluation Protocol. In *Lecture Notes in Computer Science*, volume 2688, pages 625 – 638, January 2003.

[2] R. Blouet, C. Mokbel, H. Mokbel, E. Sanchez, and G. Chollet. BECARS: a Free Software for Speaker Verification. In *ODYSSEY 2004*, pages 145 – 148, 2004.

[3] H. Bredin, A. Miguel, I. H. Witten, and G. Chollet. Detecting Replay Attacks in Audiovisual Identity Verification. Accepted for ICASSP 2006, May 2006.

[4] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[5] T. Cootes, G. Edwards, and C. Taylor. Active Appearance Models. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pages 681 – 685. June 2001.

[6] J. Czyk, M. Sadeghi, J. Kittler, and L. Vandendorpe. *Decision Fusion for Face Authentication*, volume 3072/2004, chapter Biometric Authentication: First International Conference, ICBA 2004, pages 686 – 693. Springer-Verlag GmbH, July 2004.

[7] A. Dempster, N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. of Royal Statistical Society*, 39(1):1 – 22, 1977.

[8] N. A. Fox, R. Gross, J. F. Cohn, and R. B. Reilly. Robust Automatic Human Identification using Face, Mouth, and Acoustic Information. In *AMFG 2005*, pages 264 – 278, 2005.

[9] I. Guyon, J. Makhoul, R. Schwartz, and V. Vapnik. What Size Test Set Gives Good Error Rate Estimates? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):52 – 64, January 1998.

[10] C. Mokbel. Online Adaptation of HMMs to Real-Life Conditions: A Unified Framework. In *IEEE Transactions on Speech and Audio Processing*, volume 9, pages 342 – 357. 2001.

[11] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior. Recent Advances in the Automatic Recognition of Audiovisual Speech. In *IEEE*, volume 91, pages 1306 – 1326, September 2003.

[12] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10:19 – 41, 2000.

[13] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71 – 86, 1991.

[14] P. Viola and M. Jones. Robust Real-Time Object Detection. *Int. Journal of Computer Vision*, 2002.

[15] C. Xie, M. Savvides, and B. V. Kumar. Kernel Correlation Filter Based Redundant Class-Dependence Feature Analysis (KFCA) on FRGC2.0 Data. In *AMFG 2005*, pages 32 – 43, 2005.

[5] `http://www.imm.dtu.dk/~aam/`