Impostures of Talking Face Systems Using Automatic Face Animation

Florian Verdet and Jean Hennebert

Abstract— **We present in this paper a new forgery scenario for the evaluation of talking-face verification systems. The scenario is a replay-attack where we assume that the forger has got access to a still picture of the genuine user. The forger is then using a dedicated software to realistically animate the face image, reproducing head and lip movements according to a given speech waveform. The resulting forged video sequence is finally replayed to the sensor. Such attacks are nowadays quite easy to realize for potential forgers and can be opportunities to attempt to forge text-prompted challenge-response configurations of the verification system. We report the evaluation of such forgeries on the BioSecure BMEC talking face database where a set of 430 users are forged according to this face animation procedure. As expected, results show that these forgeries generate much more false acceptation in comparison to the classically used random forgeries. These results clearly show that such kind of forgery attack potentially represents a critical security breach for talking-face verification systems.**

I. INTRODUCTION

One of the direct advantages of biometric systems consists in the fact that the user doesn't have to remember passwords or keep all the different access keys. The other advantage of biometric systems is often claimed to be enhanced security. It is indeed generally believed that biometric data is difficult to steal, imitate or generate. The work reported in this paper is challenging this last statement in the framework of talking face systems. Talking face systems are multi-modal systems combining voice and face.

We usually evaluate the security performance of biometric verification systems by measuring their ability to accept true users and to reject impostures. This is classically done using a pre-recorded biometric database, preferably composed of a large collection of users. The ability of the system to accept true clients can be quite precisely estimated provided that the recording conditions of the database match the ones of the deployed system. On the other hand, it is hard to estimate reliably the ability to reject impostures as the behavior of the forgers is typically never known in advance.

Frequently, evaluations are carried out using so-called *random* impostures. This means that the data of a randomly chosen user is presented to the system using the claimed identity of a true user. In practice, such a scenario would correspond to very weak attacks, not very much realistic when considering intentional forgers. Random forgery evaluations are mainly used due to their simplicity of implementation. However, performances measured with such scenarios should be considered as quite optimistic, at least for talking face verification systems.

F. Verdet (2)(3) and J. Hennebert (1)(3) are affiliated with:
(1) University of Applied Science, HES-SO // Wallis, Switzerland,
(2) Laboratoire Informatique, University of Avignon, France,
(3) Department of Informatics, University of Fribourg, Switzerland,
{florian.verdet,jean.hennebert}@unifr.ch

Numerous talking face based verification systems have been developed and studied. We refer to [1] and especially [2] for a description of state-of-the-art talking face verification systems and for more pointers to reference papers. Talking face based systems use two channels of biometric information to verify the claimed identity of a user: the speech signal and facial features. Without being exhaustive, the talking face approach shows certain advantages:

- The fact to base the verification decision on two biometric modalities makes the system more accurate than the corresponding mono-modal systems based on speech or face only.
- The sensors are common video cameras equipped with a microphone and thus available at low cost.
- The usage difficulty is low, the user has simply to speak in front of a camera and hasn't any direct contact with the sensor.
- Lip movement detection could be used for rejecting replay attacks which are simply based on presenting a photo of the user to the camera.

In this paper, we focus on ways to evaluate the robustness of such talking face systems against intentional impostures that are more evolved than the classical random impostures. More specifically, we present here a new kind of forgery scenario which is based on automatic face animation of a still picture of the true user. There are two advantages of such a scenario. First, the scenario is more realistic than random impostures as the only hypothesis is that the forger has stolen a picture of the true user. It can also be a good attack candidate in face of stronger challenge-response configurations of the verification system. Second, the scenario is quite easy to implement on a large number of users as it can be automated.

This article is organized as follows. In Section 2, we present the new forgery type based on automatic face animation. In Section 3, we present the BMEC database used in this project together with the test protocols defined on this new type of forgery. In Section 4, we describe the talking face verification system we used as reference in the present imposture evaluation. In Section 5, we discuss the results obtained by this system according to the different forgery types. Finally, some conclusions and hints to future work are presented.

II. IMPOSTURES WITH FACE ANIMATION

In this paper, we are interested in replay attacks where the impostor has got access to the sensor and presents the

IEEE Conference on Biometrics: Theory, Applications and Systems (BTAS 08), Arlington, Virginia, USA. 2008. pp. 1-4

biometric data that was created by means of data processing tools. We consider that he was able to steal or to intercept a part of real users' data. This kind of attack probably represents, for talking face systems, a realistic attack and is simple enough to put into practice. More specifically, this replay attack scenario makes the following assumptions.

As illustrated in Figure 1, the forger succeeds in getting a still picture of the person he wants to imitate. Starting with this picture, he uses a dedicated software to animate the face in function of a speech recording. Such software are available nowadays from different suppliers. For this project, we used CrazyTalk from Reallusion [3]. This software offers the possibility to animate a face by giving a limited number of reference points that mark the lips, the eyes, the nose and the contour of the face. With this software, the time required to build an animation and to generate the impostor video file is inferior to one minute.

For the speech part, we used a free speech synthesis tool [4] with gender dependent voices (according to the gender of the user to forge). In future work, more advanced forgeries could be built on the speech side, either by recording the true user without his knowledge (for text independent systems), or using speech modification tools (for text dependent or challenge response systems).

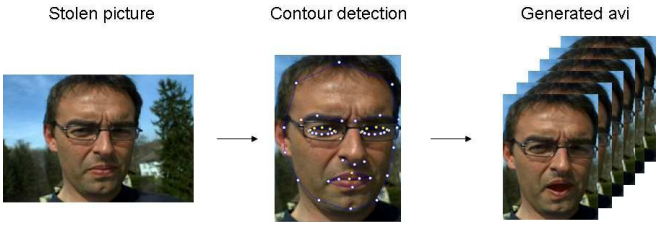

Fig. 1. Generation of face animation forgery.

This attack scenario offers the double advantage of being relatively simple to put into practice within large scale evaluations and of being more realistic than random forgery scenarios. Furthermore, lip animation is done according to the acoustic events automatically detected in the audio stream (plosives, occlusives etc), creating an additional difficulty for liveness detectors, which try to exploit lip-speech synchrony [5]. Finally, the evolution of equipment performance excites the imagination to synthesize real-time animations to attack challenge-type (question-answer) biometric systems.

III. BMEC DATABASE AND EVALUATION PROTOCOLS

The *mobile scenario* part of BioSecure BMEC database contains the following modalities [6] [2]: fingerprint, signature and talking-face. It is on this last modality that the experiences described in this paper are carried out. The talking-face data has been recorded by means of a webcam on a mobile computer of type SAMSUNG Q1 in two sessions, separated by an average timespan of one month. Different contents and different recording conditions have been simulated in the BMEC database to allow the evaluation of multiple scenarios: text-dependent, text-independent, text-prompted, office condition (inside) and street condition (outside). In our test protocol, we use, for each client in the database, 4 video files:

- 2 files recorded inside of a building during the first session and containing a short English phrase in text-prompted mode each
- 2 files recorded outside during the second session and containing a short English phrase in text-prompted mode each

The recording duration of each file is limited to 10 seconds and the average duration of speech activity is around 4 seconds. The video part of these files is in DV format .

For training a client model, we use **one** video sequence of first session's *interior* condition. Two models are thus built for each client using each of the two available files. These models are handled independently (as if they were from two different users) at the time of testing. The tests are carried out on the 2 sequences of second session's *exterior* condition. This gives a total of 4 client access tests per user.

Random impostor access tests are carried out by randomly selecting 10 other users from the database (from the development set while tuning the system and from the evaluation set upon final scoring). In a manner similar to the client accesses, 4 tests are done for each user-impostor pair, which gives a total of 40 random impostor tests per user.

The production of face animation forgeries has been done by generating one forgery per user which then is tested against the 2 available client models.

The protocol is organized into a **development** and an **evaluation** phase on two distinct user sets. These user sets correspond to BMEC's *DEV* and *TEST* sets [2]. The development phase is carried out on 50 users which are provided for verification system parameter optimization. So to keep a realistic forgery evaluation scenario, no example of face animation forgeries was available during development phase. The evaluation set includes 430 users and yields following number of tests:

- 1720 client access tests (430 users * 2 models * 2 accesses)
- 17200 unintended impostor access tests of random type
- 860 intentional access tests of face animation type

IV. VERIFICATION SYSTEM BASED ON GMMs

Talking-face verification systems should be able to build models for each client that capture its voice structure for the speech part and its facial traits for the face part.

Figure 2 illustrates our verification system. In perspective of evolving a reference system that is simple and generic enough, we chose to use standard multi-modal Gaussian Mixtures (GMMs) to model the two signal streams (speech and face) in an independent way. The fusion is done at likelihood score level. This approach allows also to measure the performances of the sub-systems speech only (1), face only (2) and on the fused system (3).

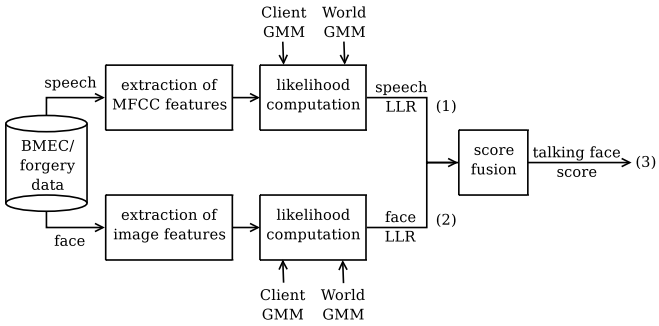

Fig. 2. Talking face verification system.

A. Feature Extraction

For speech, we use Mel frequency cepstral features (MFCC) [7] calculated on 25.625 ms windows shifted by 10 ms. For each window, 12 MFCC parameters are extracted as well as the signal energy present in the window. A voice activity detection module based on a bi-gaussian model is used to remove silence parts from the signal. The MFCCs are then normalized by mean and variance, which are estimated for each file on its signal part after silence removal.

For face, we extract one image every second from the video sequence and the following steps are applied to each of those extracted frames. Spotting of eyes' position and face cropping is done similar to the reference talking-face system of the BioSecure project [8]. After rescaling to a uniform format of 120×160 pixels, each image is converted to gray scale and histogram normalization is applied. The image is then chopped into windows of 15×15 pixels with a consecutive shift of 7 pixels between each window. Then, a feature extraction of DCT-Mod2 type is applied to each chop [9].

B. Multi-Gaussian Modeling

GMMs are used to estimate the feature vector likelihoods. GMMs are often considered as reference algorithm for speaker modeling [10] and generally have given good results for face verification tasks [9]. GMMs are also modeling tools that are simple to use, very flexible and able to estimate relatively complex probability densities. Here, GMMs estimate the probability density $p(x_n|M_{client})$ of a vector x_n of D dimensions (the number of features in each vector). This probability density, also called the *likelihood*, is estimated by means of a weighted sum of multivariate Gaussian densities:

$$p(x_n|M_{client}) = \sum_{i=1}^{I} w_i \mathcal{N}(x_n, \mu_i, \Sigma_i) \qquad (1)$$

where n is the vector's index (MFCC or DCT-Mod2 vectors in our case), I is the number of Gaussian distributions and w_i is the weight of the Gaussian i. The Gaussian densities $\mathcal{N}$ are parametrized by a vector of means μ_i $(D \times 1)$ and a covariances matrix Σ_i $(D \times D)$. The weights w_i satisfy the constraint $\sum_{i=1}^{I} w_i = 1$.

We assume here that the coefficients of the feature vectors are decorrelated, so that we can use a diagonal covariance matrix. Making the extra assumption on temporal independence of the speech observations (spatial independence for face observations), the global likelihood score of the client model for the feature vector sequence $X = \{x_1, x_2, ..., x_N\}$ can be calculated as follows:

$$S_c = p(X|M_{client}) = \prod_{n=1}^{N} p(x_n|M_{client}) \qquad (2)$$

In a similar way, we also estimate the likelihood S_w that X doesn't belong to the client by using a world model M_{world}. In our case, this world model is unique and is trained on the development data set. In function of the preceding hypothesis, the optimal decision to accept or reject the claimed user identity is taken by comparing the ratio of the likelihood scores of the client and the world models with a global threshold T. Here, this ratio is calculated in the logarithmic domain with $R_c = \log(S_c) - \log(S_w)$. For evaluating system performance or its resistance against forgeries, a DET (or ROC) curve [11] is drawn by moving the threshold T over all possible values and measuring the rates of false acceptation and false rejection.

Training of the world model is done via the *Expectation-Maximization* (EM) algorithm, which iteratively refines the different parameters of the model so that its likelihood is maximized until its parameters converge, typically after some iterations [12]. We use a binary splitting procedure to increase the number of Gaussians until it reaches a predefined value. This method allows us to obtain GMMs with a high number of Gaussians while limiting the risks to fall into local maxima during EM procedure. The world model is trained using all available client accesses of the development data set.

The client models on both, the speech and the face side are then obtained through an adaptation algorithm of *Maximum A-Posteriori* (MAP) type, applied on the world model [13].

For voice modeling, a mixture of 32 Gaussians is used. Since only a few seconds of speech are available for training each model, it is rather necessary to use MAP adaptation from a world model to try one's best. For the face, a mixture of 128 Gaussians is used. These values have been optimized on the development set.

At time of a testing, the scores obtained for the images extracted at a one-second interval on the whole length of the file are fused together calculating their average.

C. Score Fusion

Before fusing the scores of the two modalities, we normalize them by the mean and the variance that were estimated on a set of scores obtained from a batch of "world" users, scored against the client model (also called *Znorm*). This normalization batch is made up of only one sequence of each user of the development set.

The two modalities are then fused together by simple addition of the normalized likelihood scores $R_{c,talkingFace} = R_{c,speech} + R_{c,face}$. More elaborated approaches to accomplish fusion could be thought of, as for instance a weighted

sum of the scores or even using classifiers trained for the fusion task [14]. The objective of the work described in this paper isn't to optimize the performance of a system, but rather to estimate the impact of forgeries.

V. SYSTEM EVALUATION

Figure 3 shows the results of the reference system described above. On random impostures, the equal error rate (EER, at whose threshold T, the false acceptation and false rejection rates are equal) is of 23.3 %. On face animation forgeries, the error rate climbs to 40.0 %.

An analysis of the individual modalities (speech and face sub-systems) is presented in Figure 4. It shows that the sole face part of those face animation forgeries creates a severe raising of the EER up to 49.7 %, which means that almost every second forger access would succeed in entering the system (while, with the same threshold, rejecting almost every second client).

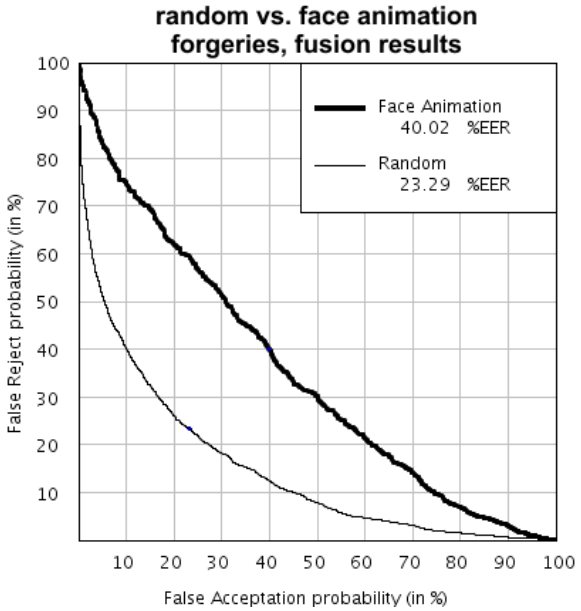

Fig. 3. Results (ROC curve and EER) of the reference system on random and face animation forgeries.

VI. CONCLUSIONS

In this paper, we introduced a new forgery type based on face animation where the underlying hypotheses are quite realistic. Publicly available voice synthesis and face animation tools were used for these experiments. The evaluation reported in this article shows the efficiency of these forgeries with a significant increase of the equal error rate measured with a reference system. More realistically, this forgery scenario presents an interesting alternative to the random forgery scenario and is relatively easy to implement in the framework of large-scale evaluations.

VII. ACKNOWLEDGMENTS

This work was partly supported by the Swiss NSF program "Interactive Multimodal Information Management (IM2)", as part of NCCR and by the EU BioSecure IST-2002-507634 NoE project.

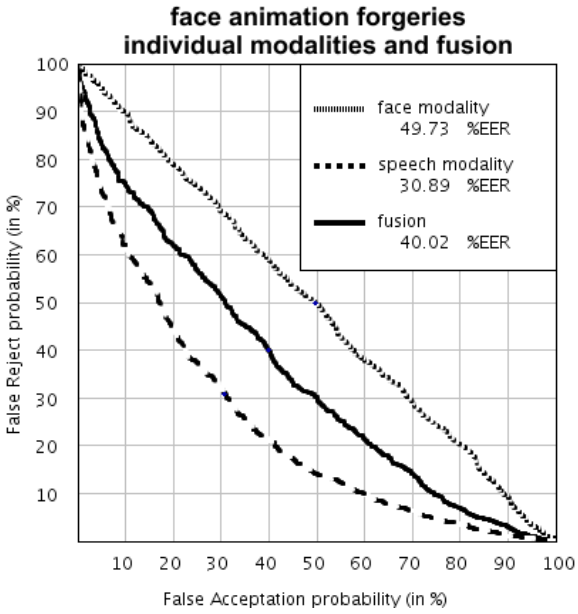

Fig. 4. ROC curve and EER for individual modalities and fusion of the reference system on face animation forgeries.

REFERENCES

[1] A. Jain, A. Ross, and S. Prebhakar. An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image- and Video-Based Biometrics*, 14(1), 2004.

[2] B. Fauve, H. Bredin, W. Karam, F. Verdet, A. Mayoue, G. Chollet, J. Hennebert, R. Lewis, J. Mason, C. Mokbel, and D. Petrovska. Some results from the biosecure talking face evaluation campaign. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Las Vegas, USA, 2008.

[3] Reallusion CrazyTalk software. http://crazytalk.reallusion.com/.

[4] Festival Text-To-Speech System. http://www.cstr.ed.ac.uk/projects/festival/.

[5] Hervé Bredin. *Vérification de l'identité d'un visage parlant. Apport de la mesure de synchronie audiovisuelle face aux tentatives délibérées d'imposture*. PhD thesis, École Nationale Supérieure des Télécommunications (ENST), Paris, 2007.

[6] BioSecure Network of Excellence Biometric Multimodal Evaluation Campaign. http://diuf.unifr.ch/go/bmec, 2007.

[7] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals Of Speech Recognition*. Prentice Hall, 1993.

[8] H. Bredin, G. Aversano, C. Mokbel, and G. Chollet. The biosecure talking-face reference system. In *Proc. Workshop on Multimodal User Authentication (MMUA)*, 2006.

[9] F. Cardinaux, C. Sanderson, and S. Bengio. User authentication via adapted statistical models of face images. *IEEE Transactions on Signal Processing*, 54(1):361–373, 2006.

[10] Douglas Reynolds. Automatic speaker recognition using gaussian mixture speaker models. *The Lincoln Laboratory Journal*, 8(2):173–191, 1995.

[11] Alvin Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The det curve in assesment of detection task performance. In *Eurospeech 1997*, pages 1895–1898, Rhodes, Greece, 1997.

[12] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of Royal Statistical Society*, 39(1):1–38, 1977.

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

[14] Anil Jain, Karthik Nandakumar, and Arun Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38:2270–2285, 2005.