MAKING TALKING-FACE AUTHENTICATION ROBUST TO DELIBERATE IMPOSTURE

Hervé Bredin and Gérard Chollet

{bredin, chollet}@tsi.enst.fr
GET-ENST, Dept. TSI, CNRS LTCI, Paris, France

ABSTRACT

We expose the limitations of existing frameworks designed for the evaluation of audiovisual biometric authentication algorithms. The weakness of a classical audiovisual authentication system is uncovered when confronted to realistic deliberate impostors. A client-dependent audiovisual synchrony measure is used in order to deal with deliberate impostors and three new fusion strategies and their performance against random and deliberate impostors are studied.

Index Terms— Robustness, Speaker recognition, Face recognition

1. INTRODUCTION

Numerous studies have exposed the limits of biometric identity verification based on a single modality (such as fingerprint, iris, handwritten signature, voice, face). The talking face modality, that includes both face recognition and speaker verification, is a natural choice for multimodal biometrics. Talking faces provide richer opportunities for verification than does any ordinary multimodal fusion. The signal contains not only voice and image but also a third source of information: the simultaneous dynamics of these features. Natural lip motion and the corresponding speech signal are synchronized.

However, this specificity is often forgotten and most of the existing talking-face authentication systems are based on the fusion of the scores of two separate modules of face verification and speaker verification, as described in section 2. Even though this prevalent paradigm may lead to the best performance on widespread evaluation frameworks based on random impostor scenarios, the question of its performance against real life impostor attacks is studied in section 3. Section 4 is devoted to the description of the client-dependent audiovisual synchrony measure introduced in order to tackle these attacks and deals with the highlighted weakness of the current talking-face authentication systems. Three fusion strategies are finally introduced and compared in section 5.

2. AUDIO-VISUAL BIOMETRICS

Most of the existing talking-face authentication systems are based on the fusion of two scores, obtained through speaker and face verification [1]. In order to show the limits of this kind of approach, the first mandatory step is to develop such a system.

2.1. Face verification

The classical *eigenface* approach, combined with the Mahalanobis distance, is used in our implementation of the face verification module [2]. We will mostly focus on its specificities: using every *reliable* detected face available in a video sequence.

Once face detection is applied on each frame of the video sequence (using *Fasel et al.*'s algorithm [3]), *distance from face space* (DFFS) is computed for every detected face as the distance between the face and its projection on the face space (obtained via principal component analysis) [2]. We define a *reliability* coefficient r as the inverse of the DFFS ($r = 1/\text{DFFS}$): the higher, the more reliable. Finally, a detected face is kept as correct if its r coefficient is higher than a threshold $\theta_r = 2/3 \cdot r_{\max}$, where $r_{\max}$ is the maximum value of r on the current video sequence. Figure 1 shows (from left to right) the face corresponding to $r = r_{\max}$, an example of correctly detected face and an example of rejected face. Only *eigenface*

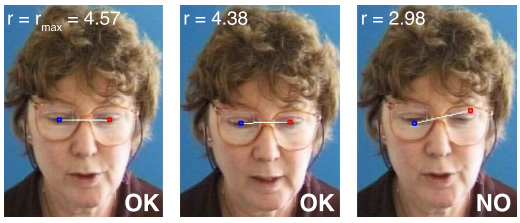

Fig. 1. Selection of *reliable* faces

features corresponding to correctly detected face are kept to describe the face appearing in the video sequence. Finally, at test time, the *Mahalanobis* distance is computed between the *eigenface* features (of dimension 80, in our case) of each of the N correctly detected faces of the enrollment video sequence and each of the M correctly detected face of the test video sequence, leading to $N \times M$ distances. The opposite of the mean of these $N \times M$ distances is taken as the score S_{f} of the face verification module.

2.2. Speaker verification

The speaker verification module is also based on a very classical approach: gaussian mixture model with universal background model (GMM-UBM) [4].

First, silence detection is performed, based on a bi-gaussian modeling of the acoustic energy distribution. Then, MFCC features are extracted on a 20 ms long window every 10 ms (12 MFCC with first and second derivatives), while only keeping the features corresponding to non-silent windows. Using the Expectation-Maximization (EM) algorithm, a 256 gaussians mixture model (GMM) – called Universal Background Model (UBM) – is trained using a set of recordings of numerous speakers in order to maximally cover the variability among speakers. Using the MFCC features extracted from the enrollment sequence, Maximum A Posteriori (MAP) adaptation is applied in order to get a client-dependent GMM from the UBM [4]. At test time, MFCC features are extracted from the test sequence and are compared to both the client-dependent GMM and the UBM:

the likelihood ratio is finally taken as the score S_{s} of the speaker verification module.

2.3. Fusion

The fusion module is also very classical [5]: it is based on the weighted sum of normalized speaker and face verification scores S_{s} and S_{f}.

Weighted sum of face and speaker verification scores is meaningful only if S_{s} and S_{f} vary in the same range of values. Therefore, a first step of σ/μ normalization is applied in order to make sure that normalized scores $\overline{S_{\mathrm{s}}}$ and $\overline{S_{\mathrm{f}}}$ have comparable values. For that purpose, a development set should be available, that contains a collection of actual test scores S_{s} and S_{f} allowing to estimate the mean μ_{s} and μ_{f} and standard deviation σ_{s} and σ_{f} of false claim scores. Normalization of test scores is then performed:

$$\overline{S_{\mathrm{s}}} = \frac{S_{\mathrm{s}} - \mu_{\mathrm{s}}}{\sigma_{\mathrm{s}}} \quad \text{and} \quad \overline{S_{\mathrm{f}}} = \frac{S_{\mathrm{f}} - \mu_{\mathrm{f}}}{\sigma_{\mathrm{f}}} \tag{1}$$

From this point, in order to make the rest of the paper more readable, we will assume that scores are already normalized and we will therefore denote the normalized score S instead of $\overline{S}$.
The fusion is then performed by computing a weighted sum of the normalized speaker and face verification scores S_{s} and S_{f}:

$$S_0 = w_{\mathrm{s}} S_{\mathrm{s}} + w_{\mathrm{f}} S_{\mathrm{f}} \quad \text{with} \quad w_{\mathrm{f}} + w_{\mathrm{s}} = 1 \tag{2}$$

Optimal weights w_{s} and w_{f} are estimated on the development set by minimizing the weighted error rate defined in next section.

3. LIMITATIONS OF AV BIOMETRICS

In order to evaluate and compare performances of biometric authentication systems, reproducible evaluation frameworks are set up. They usually consist of a database (containing biometric samples of a large collection of people) and associated evaluation protocols that describe the list of true and false claims that have to be tested. In the particular case of audiovisual biometrics, one can refer to the XM2VTS, MyIDEA, BioSecure, IV^2 or BANCA databases [6]. In the rest of the paper, we will make use of the latter and its associated Pooled (P) protocol. However, the limitations that we will expose are common to all these evaluation frameworks.

3.1. Evaluation

The BANCA audiovisual database contains 52 speakers divided into 2 groups G1 and G2 of 26 speakers each (13 females and 13 males). This division into 2 disjoint groups allows to use G2 as a development set when testing on G1 (and reciprocally). Twelve sessions were recorded in three different conditions (controlled, adverse and degraded). In each session and for each speaker, two recordings were acquired: one true claim access where the speaker pronounces digits and his/her (fake) own address and one false claim access where he/she pronounces digits and the address of another person of the same group.

According to the P protocol, for each person, the true claim access of the controlled session #1 is used as the enrollment data. True claim accesses of sessions #2 to #4, #6 to #8 and #10 to #12 are used as client accesses, while every twelve false claims accesses are used as impostor accesses. Therefore, this makes 234 client and 312 impostor accesses per group.

In order to proceed with the comparison of our different systems, we will evaluate their performance using the weighted error rate (WER) as defined in equation (3): the cost of false acceptance (FAR stands for false acceptance rate) is ten times higher than the cost of false rejection (FRR, for false rejection rate). Therefore, minimizing WER makes the system more difficult to fool but also more likely to reject genuine users.

$$\mathrm{WER} \quad = \quad \frac{1}{11} \left(\mathrm{FRR}\left(\theta\right) + 10 \cdot \mathrm{FAR}\left(\theta\right) \right) \tag{3}$$

The development set is used to optimize the decision threshold θ in order to minimize the error rate and it is consequently applied on the test set. Reported error rates will be complemented by confidence intervals at 95%, as defined in [7].

3.2. Deliberate impostors

The only information about his/her target that is known by the impostor is his/her name and address. No real effort is performed by the impostor while trying to impersonate his/her target. Therefore, these impostor accesses (that we will refer to as random impostor accesses afterwards) appear to be quite unrealistic. Only a fool would attempt to imitate a person knowing so little about them.

In real life, an impostor would have acquired some information about his/her target before trying to impersonate him/her. In the case of audiovisual biometrics, it should be very easy to acquire a picture of the face of the target and a recording of his/her voice (thanks to a telephone call, for example). Showing the picture of the face of the target while playing the audio recording of his/her voice would then be enough to completely fool a talking-face authentication system based on the score fusion of two modules of face and speaker verification.

Such deliberate impostor attacks were simulated. Every original BANCA false claim access was modified into a sequence made of the combination of a video sequence showing a moving picture (as shown in figure 2) and the audio of a genuine access of the target. We will refer to these new impostor accesses as deliberate impostor accesses afterwards.

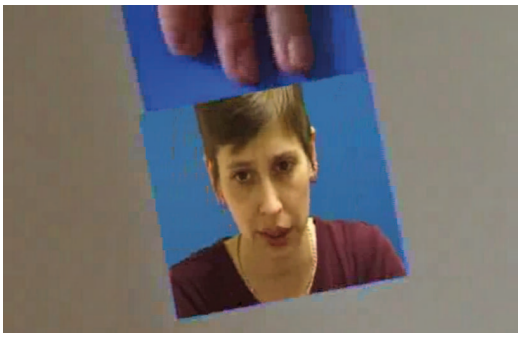

Fig. 2. Deliberate impostor attack

3.3. Results

Table 1 shows the effect of deliberate impostor attacks on the initial systems described in section 2. Speaker and face verification modules are completely fooled by deliberate impostors and so it goes for their fusion (for which the weighted error rate increases from 6% to 90%).

Modality		Random imp.	Deliberate imp.
Speaker	S_s	$6.0 \pm 2.0\%$	$92.1 \pm 1.6\%$
Face	S_f	$7.9 \pm 1.3\%$	$72.7 \pm 4.8\%$
Fusion	S_0	$6.3 \pm 2.2\%$	$90.2 \pm 1.9\%$

Table 1. Weighted error rate for the initial systems

4. SYNCHRONY VERIFICATION

There are many ways of dealing with this kind of deliberate impostor attacks. The first solution is to ask for the enunciation of a different utterance (chosen randomly) for each new access, thus preventing the prior recording of the voice of the target. An alternative is to analyze the motion of the detected face and look for a suspicious behavior [8]. The third solution is to study the degree of correspondence between the shape and motion of the lip and the acoustic signal [9].

4.1. Client-dependent synchrony measure

In [10] we introduced a new biometric modality based on a client-dependent measure of the synchrony between acoustic and visual speech features. We will quickly overview its main principle and performance.

Every 10 ms, a 24-dimensional acoustic feature vector (12 MFCC coefficients and their first derivatives) is extracted and will be denoted as $X \in \mathbb{R}^n$ in the rest of the paper. As far as visual feature vectors are concerned, we chose to extract Discrete Cosine Transform (DCT) of the mouth area. Figure 3 summarizes this process. For each frame of the sequence, face is detected and a *Viola & Jones* mouth detector is applied to locate the mouth area [11], from which 28 DCT coefficients corresponding to the low spatial frequencies are computed. In order to equalize the sample rates of acoustic and visual features (initially 100 Hz and 25 Hz respectively), visual features are linearly interpolated. First derivatives are then appended, leading to 56-dimensional visual feature vectors $Y \in \mathbb{R}^m$.

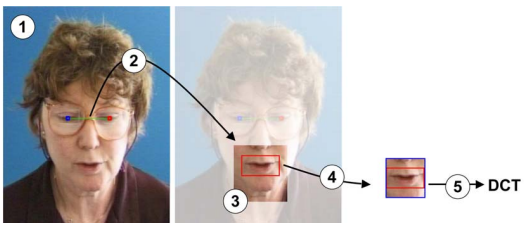

Fig. 3. Visual features extraction

Using the acoustic and visual features X and Y extracted from the enrollment sequence, co-inertia analysis (CoIA) is applied in order to compute the client-dependent synchrony model $(\mathbf{A}, \mathbf{B})$. The columns of matrices $\mathbf{A}$ and $\mathbf{B}$ are vectors $\mathbf{a}_k$ and $\mathbf{b}_k$, $k \leq \min(n, m)$, that are defined recursively as the projection vectors maximizing the covariance between X and Y:

$$(\mathbf{a}_1, \mathbf{b}_1) = \underset{(a,b)}{\operatorname{argmax}} \operatorname{cov}\left(a^t X, b^t Y\right) \tag{4}$$

The same maximization in the orthogonal subspaces to $\mathbf{a}_1^t X$ and $\mathbf{b}_1^t Y$ allows the computation of $\mathbf{a}_2$ and $\mathbf{b}_2$, and so on for the other $\mathbf{a}_k$ and $\mathbf{b}_k$. More information on CoIA can be found in [10].

At test time, acoustic and visual feature vectors X^Γ and Y^Γ of the test sequence Γ are extracted and a measure S_c of their synchrony is computed using the synchrony model $\left(\mathbf{A}^\lambda, \mathbf{B}^\lambda\right)$ of the claimed identity λ:

$$S_\mathrm{c} = \frac{1}{D} \sum_{k=1}^{D} \operatorname{corr}\left(\mathbf{a}_k^{\lambda t} X^\Gamma, \mathbf{b}_k^{\lambda t} Y^\Gamma\right) \tag{5}$$

where D is the number of dimensions actually used to compute the correlation. In our case, $D = 18$ showed to be the most efficient value for D, and will be used in the rest of the paper [10].

4.2. Results

Table 2 shows the effect of deliberate impostor attacks on the synchrony verification module. It clearly shows that this modality is intrinsically robust to deliberate impostors though its performance on random impostors is worse than the fusion of speaker and face verification.

Modality		Random imp.	Deliberate imp.
Synchrony	S_c	$7.7 \pm 1.1\%$	$6.9 \pm 0.1\%$

Table 2. Weighted error rate for the synchrony modality

5. FUSION

Therefore, we propose to make use of the obvious complementarity between the initial audiovisual biometric system and this new modality based on client-dependent audiovisual synchrony measure. Three strategies of fusion will be presented and their performance compared.

5.1. Strategies

The first fusion strategy is the direct extension of the one presented in section 2. As shown in equation (6), the fused score S_1 is a weighted sum of three normalized monomodal scores (those weights being optimized on the development set in order to minimize the error rate):

$$S_1 = w_\mathrm{s} S_\mathrm{s} + w_\mathrm{f} S_\mathrm{f} + w_\mathrm{c} S_\mathrm{c} \quad \text{with} \quad \sum w = 1 \tag{6}$$

As seen in figure 4, impostor synchrony scores (either random or deliberate) are globally lower than true claim scores. The second fusion strategy benefits from this property and is designed to decrease the value of score S_1 for claims with low synchrony verification score S_c, as shown in equation (7):

$$S_2 = \alpha\left(S_\mathrm{c}\right) S_1 \tag{7}$$

where α is the cumulative distribution function of true claims synchrony scores:

$$\alpha\left(S_\mathrm{c}\right) = p\left(s \leq S_\mathrm{c} | \text{ true claim}\right) \tag{8}$$

The function α is drawn as a thick black line in the top right corner of figure 4. It is estimated using the true claims synchrony scores of the development set.

The third fusion strategy benefits from the complementarity of the first fusion strategy and the synchrony verification module – the former being very sensitive to deliberate impostors but more efficient

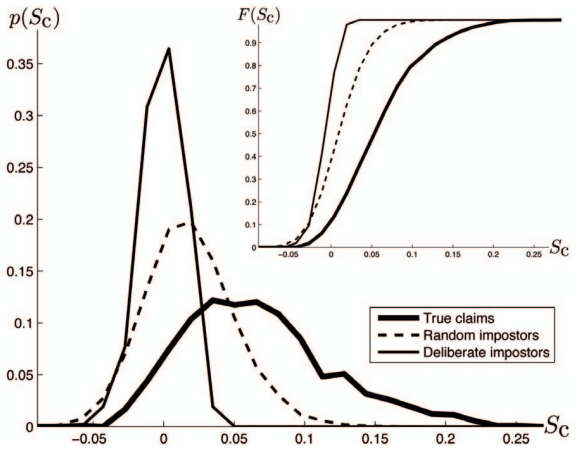

Fig. 4. Synchrony scores distributions and their corresponding cumulative distribution functions (in the top right corner)

against random impostors, while the latter is very robust to attacks, though it is less efficient against random impostors:

$$S_3 = \alpha(S_c)\, S_1 + [1 - \alpha(S_c)]\, S_c \quad (9)$$

As shown in equation (9), this last strategy is based on an adaptive weighted sum of (normalized) scores. More weight is given to the synchrony verification module if the synchrony measure is low; and reciprocally its weight is decreased when the synchrony measure is higher and the first strategy fusion can actually be trusted.

5.2. Results

Table 3 summarizes the performance of the original fusion and the three proposed strategies. While the first one (weighted sum of the three modalities – speaker, face and synchrony) does not bring any improvement, the second and third ones make the audiovisual biometric system robust to deliberate impostors, while maintaining its raw performance against random impostors.

Modality		Random imp.	Deliberate imp.
Fusion	S_0	$6.3 \pm 2.2\%$	$90.2 \pm 1.9\%$
Fusion	S_1	$7.5 \pm 2.5\%$	$90.0 \pm 1.9\%$
Fusion	S_2	$6.6 \pm 1.8\%$	$24.2 \pm 4.6\%$
Fusion	S_3	$6.4 \pm 1.8\%$	$14.1 \pm 3.5\%$

Table 3. Weighted error rate for the three new fusion strategies

6. CONCLUSION AND PERSPECTIVES

We exposed the limitations of existing evaluation frameworks for audiovisual biometric authentication algorithms and proposed more realistic scenarios where impostors no longer perform random attacks and, instead, use material about the target that they acquired previously to fool the system. A client-dependent audiovisual synchrony measure is presented as a solution to this major issue. Three different fusion strategies try to benefit from its high robustness to deliberate impostors and we show that it is possible to drastically diminish the sensitiveness of a classical audiovisual biometric authentication algorithm to deliberate imposture while maintaining its performance in the genuine evaluation framework.

Our approach based on audiovisual synchrony measure should also be evaluated on higher-effort forgeries such as voice conversion and face animation that would lead to synchronous audiovisual speech that resemble the target. As a matter of fact, voice conversion would allow an impostor to fool a system based on a random prompt and face animation techniques could possibly defeat algorithms checking the liveness from image motion: commercial softwares are already available, allowing to realistically animate a picture according to a given voice recording. Nonetheless, since our approach is based on a *client-specific* synchrony model, it could possibly still show strong robustness to this high-effort impostor attacks.

Acknowledgments This work was partially supported by the European Community through the BioSecure NoE and the KSpace NoE.

7. REFERENCES

[1] Petar S. Aleksic and Aggelos K. Katsaggelos, "Audio-Visual Biometrics," in *Proceedings of the IEEE*, November 2006, vol. 94, pp. 2025–2044.

[2] Matthew Turk and Alex Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[3] Ian Fasel, Bret Fortenberry, and J. R. Movellan, "A Generative Framework for Real-Time Object Detection and Classification," *Computer Vision and Image Understanding - Special Issue on Eye Detection and Tracking*, vol. 98, no. 1, pp. 182–210, 2004.

[4] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, pp. 19 – 41, 2000.

[5] Arun A. Ross, Karthik Nandakumar, and Anil K. Jain, *Handbook of Multibiometrics*, Springer, 2006.

[6] Enrique Bailly-Baillière et al., "The BANCA Database and Evaluation Protocol," in *4th International Conference on Audio-and Video-Based Biometric Person Authentication (AVBPA'03)*, Guildford, UK, January 2003, vol. 2688 of *Lecture Notes in Computer Science*, pp. 625 – 638, Springer.

[7] Samy Bengio and Johnny Mariéthoz, "A Statistical Significance Test for Person Authentication," in *ODYSSEY 2004 - The Speaker and Language Recognition Workshop*, Toledo, Spain, May 2004, pp. 237–244.

[8] Hyung-Keun Jee, Sung-Uk Jung, and Jang-Hee Yoo, "Liveness Detection for Embedded Face Recognition System," *International Journal of Biomedical Sciences*, vol. 1, no. 4, pp. 235–238, 2006.

[9] Girija Chetty and Michael Wagner, ""Liveness" Verification in Audio-Video Authentication," in *10th Australian International Conference on Speech Science and Technology (SST'04)*, Sydney, Australia, December 2004, pp. 358–363.

[10] Hervé Bredin and Gérard Chollet, "Audio-Visual Speech Synchrony Measure for Talking-Face Identity Verification," in *32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'07)*, Honolulu, USA, April 2007.

[11] M. Castrillón Santana, J. Lorenzo Navarro, O. Déniz Suárez, and A. Falcón Martel, "Multiple Face Detection at Different Resolutions for Perceptual User Interfaces," in *2nd Iberian Conference on Pattern Recognition and Image Analysis*, Estoril, Portugal, June 2005.