DETECTING REPLAY ATTACKS IN AUDIOVISUAL IDENTITY VERIFICATION

Hervé BREDIN [1]*, Antonio MIGUEL* [2]*, Ian H. WITTEN* [3] *and Gérard CHOLLET* [1]

[1] Ecole Nationale Supérieure des Télécommunications, Dept. TSI, Paris, France
[2] Communication Technologies Group (GTC), I3A, University of Zaragoza, Spain
[3] University of Waikato, Dept. Computer Science, Hamilton, New Zealand

ABSTRACT

We describe an algorithm that detects a lack of correspondence between speech and lip motion by detecting and monitoring the degree of synchrony between live audio and visual signals. It is simple, effective, and computationally inexpensive; providing a useful degree of robustness against basic replay attacks and against speech or image forgeries. The method is based on a cross-correlation analysis between two streams of features, one from the audio signal and the other from the image sequence.

We argue that such an algorithm forms an effective first barrier against several kinds of replay attack that would defeat existing verification systems based on standard multimodal fusion techniques. In order to provide an evaluation mechanism for the new technique we have augmented the protocols that accompany the BANCA multimedia corpus by defining new scenarios. We obtain 0% equal-error rate (EER) on the simplest scenario and 35% on a more challenging one [1].

1. INTRODUCTION

Numerous studies have exposed the limits of biometric identity verification based on a single modality (such as fingerprint, iris, hand-written signature, voice, face). Consequently many researchers are exploring whether the coordinated use of two or more modalities can improve performance. The "talking-face" modality, which includes both face recognition and speaker verification, is a natural choice for multimodal biometrics in many practical applications—including face-to-face scenarios, remote video cameras, and even future personal digital assistants.

Talking faces provide richer opportunities for verification than does 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, most work on audiovisual speech-based biometrics ignores this third information source: it uses the audio and video streams separately and performs fusion at the score level [1] [2]. Nevertheless, some research in speech recognition has shown that it is helpful to take into account the synchronized lip motion, particularly in noisy environments [3] [4].

The aim of this paper is to exploit this novel characteristic of the talking-face modality within the specific framework of identity verification. Section 4 presents a simple method for detecting and quantifying the synchronization between speech and lip motion, based on the correlation between primitive measures of audiovisual activity. The technique can be used to augment an existing audio-visual verification system without excessive computational cost. Doing so thwarts a number of deliberate (so-called "high-effort") attacks that would defeat a standard system.

Many databases are available to the research community to help evaluate multimodal biometric verification algorithms, such as BANCA [5], XM2VTS and BIOMET [6]. Different protocols have been defined for evaluating biometric systems on each of these databases, but they share the assumption that impostor attacks are zero-effort attacks. For example, in the particular framework of the BANCA database, each subject records one client access and one impostor access per session. However, the only difference between the two is the particular message that the client utters—their name and address in the first case; the target's name and address in the second. Thus the "impersonation" takes place without any knowledge of the target's face, age, and voice. These zero-effort impostor attacks are unrealistic—only a fool would attempt to imitate a person without knowing anything about them. In this work we adopt more realistic scenarios in which the impostor has more information about the target.

This article is organized as follows. The next section presents the deliberate (as opposed to "zero-effort") impostor attacks that we have defined. The following section describes the features that our new algorithm uses, while the one after that describes the algorithm itself. Section 5 describes the evaluation methodology, followed by a presentation of performance results for the algorithm. The final section summarizes the results and draws some conclusions.

2. DELIBERATE IMPOSTOR ATTACKS

A major drawback of using the talking-face modality for identity verification is that an impostor can easily obtain a sample of any client's audiovisual identity. Contrast this with iris

[1] This work was initiated in the framework of the First Biosecure Residential Workshop - http://www.biosecure.info

1-4244-0469-X/06/$20.00 ©2006 IEEE

recognition: it is quite difficult to acquire a sample of another person's iris. But numerous small devices allow an impostor to take a picture of the target's face without being noticed, and some mobile phones are even able to record movies. Of course, it is even easier to acquire a recording of the target's voice. Therefore, protocols to evaluate audiovisual identity verification systems should recognize this fact, for example by adding replay attacks to their repertoire of envisaged impostor accesses.

2.1. *Paparazzi* scenario

In this scenario, prior to the attack the impostor takes a still picture of the target's face and acquires an audio recording of their voice. Then, when trying to spoof the system, the impostor simply places the picture in front of the camera and plays the audio recording. The purpose of this scenario is to illustrate the limits of a system that does not take into account the dynamics of lips motion. It has already been tackled in [7].

2.2. *Big Brother* scenario

In this scenario, prior to the attack the impostor records a movie of the target's face, instead of a still picture, and acquires a voice recording as before. However, the audio and video do not come from the same utterance, so they are not synchronised. This is a realistic assumption in situations where the identity verification protocol chooses an utterance for the client to speak. Using the same process as in the *Paparazzi* scenario, the impostor tries to spoof the system by a simple replay attack. In this paper, we address this kind of impostor attack by detecting lack of synchronisation between the audio and video streams.

2.3. Forgery scenarios

More elaborate impostor attacks can include voice and face forgery. Perrot *et al.* [8] use a recording of the target's voice, and automatically transform the impostor's voice so that it resembles the recorded utterance. Abboud and Chollet [9] track the impostor's lip motion throughout a video sequence, and then animate the target's face in a way that moves their lips to match the impostor's. A combination of these two forgeries would be a real threat for a talking-face-based identity verification system.

3. AUDIOVISUAL FEATURES

3.1. Audio features

Let y be the audio signal from a BANCA sequence. Every 10 ms, a 20 ms window is extracted on which the log-energy is computed:

$$e = \log \sum_{n=1}^{N} y(n)^2 \tag{1}$$

Therefore, 100 samples are extracted per second. Then, a simple voice activity detector based on a bi-gaussian modeling of signal energy distribution is applied: this gives the time stamps allowing to distinguish between silence and voice activity.

3.2. Visual features

For each frame, the lip area is manually located with a rectangle r of size proportional to 20x30 and centered on the mouth (as shown in figure 1) and converted to gray-level. Finally, the mean of the values of the pixels of the lip area (of width W and height H) is computed:

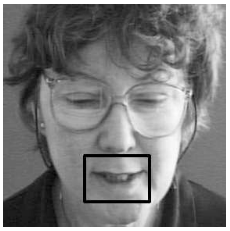

Fig. 1. Manual location of the lips

$$m = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} r(i,j) \tag{2}$$

Audiovisual sequences of the BANCA database are recorded at 25 frames per second. Therefore, 25 samples are extracted per second.

3.3. Different sample rates

As a result of these separate processes of features extraction, audio and visual features are sampled at two different rates. The proposed algorithm deals with audio and video features that must have the same sample rate. Three techniques are proposed to balance the sample rates:

Downsampling the audio signal Every 4 audio samples, only their average value is kept;

Duplicating samples of the visual signal After every sample, 3 identical samples are added;

Linearly interpolating the visual signal Between two samples, 3 linearly interpolated samples are added.

4. AUDIOVISUAL SYNCHRONY MODELLING

4.1. State of the art

Very few previous works on the particular subject of liveness detection based on speech/lips synchronisation were found in the literature. In [7], a Gaussian Mixture Model is learnt on the concatenated audio (MFCC coefficients) and visual (eigenlips projection) features. An Equal Error Rate (EER) of 2% is reached on the equivalent of the *Paparazzi* scenario.

4.2. Preliminary observation

The initial observation that led to a simple model based on correlation between audio and video features is presented in Figure 2. 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. In our particular case, we chose the mean of pixels value instead of the openness because it is easier and faster to compute, supposing that when the mouth is open, pixels are darker and *vice versa*.

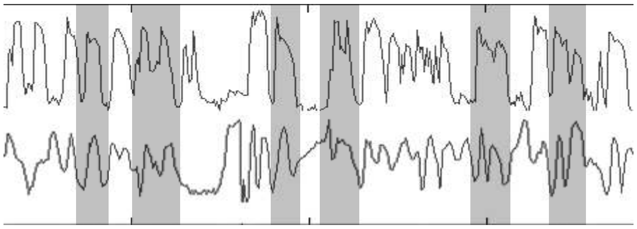

Fig. 2. Speech energy vs. Mouth openness

4.3. Cross-correlation

Let $A(t)$ and $V(t)$ be two one-dimensional random variables representing respectively the audio and the visual samples. The cross-correlation $X(d)$ ($d \in [-L, L]$) between A and V is defined as follows:

$$X(d) = \mathbf{E}(\tilde{A}(t) \cdot \tilde{V}(t-d)) \qquad (3)$$

where $\tilde{S}$ is the centered and variance-normalized version of $S \in \{A, V\}$. In our case, where $A(t)$ and $V(t)$ are only defined for $t \in [1, T]$, we can approximate X by:

$$\hat{X}(d) = \frac{1}{T-d} \sum_{t=1}^{T} \tilde{A}(t) \cdot \tilde{V}(t-d) \qquad (4)$$

assuming that $\tilde{V}(t) = 0$ for $t < 1$ and $t > T$.

4.4. Training

$$L_{max}(X) = \text{argmax}_{d \in [-L,L]} |X(d)| \qquad (5)$$

is the delay for which the correlation between A and V is maximum. Figures 3 and 4 show how it is computed and what is its distribution on two training sets: synchronised and artificially desynchronised (audio from one sequence, video from another one). Then, L_{sync} is defined as the delay corresponding to the peak in the *synchronised* training set distribution.

4.5. Testing

When testing the synchronisation of a new sequence $AV = \{A, V\}$, the score s of AV is computed as follows:

$$s(AV) = 1 - \frac{|L_{max}(X) - L_{sync}|}{L} \qquad (6)$$

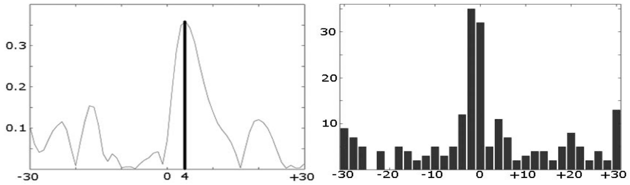

Fig. 3. Example of $L_{max}(X)$ and its distribution on 208 synchronised sequences

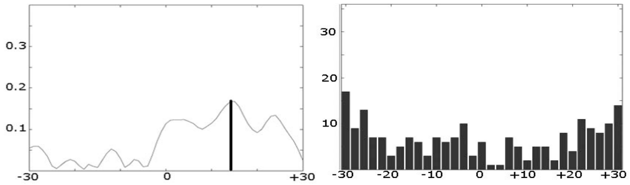

Fig. 4. Example of $L_{max}(X)$ and its distribution on 208 not-synchronised sequences

According to a given threshold $\theta \in [0, 1]$, the sequence AV is decided to be synchronised if $s(AV) \geq \theta$ and not synchronised if $s(AV) < \theta$.

5. EXPERIMENTS

The protocols we used are inspired by the original BANCA Mc protocol [5]. Thus the 52 speakers are divided into two groups (G1 and G2) with 13 females and 13 males in each one. Each speaker recorded four sessions (S1 to S4) during which two accesses were performed (client and impostor). These two groups are completely independent: when G1 is used for training tests are performed on G2, and *vice versa*.

For reasons stated in the introduction, we adapted them to simulate more realistic scenarios. Two new protocols were designed in which training and client access sequences are identical to the original BANCA Mc protocol, but with modified impostor access sequences:

***Paparazzi* protocol** The video is made of only one repeating frame, while the audio is kept unchanged;

***Big Brother* protocol** The video is taken from a different sequence, while the audio is kept unchanged.

6. RESULTS

The system obtained 0% equal-error rate (EER) on the *Paparazzi* scenario, because the visual signal for the impostor was constant and thus completely uncorrelated with the audio signal. In the more challenging *Big Brother* scenario, the system with the best tuned parameters obtained 35% EER.

Figure 5 shows the influence of parameter L (which was introduced in section 4.3). It appears that the best value lies between 20 and 50, which corresponds to a delay of between 1 and 2 seconds.

Using time-stamps of voice activity, silence frames were deleted in the audio and visual signals. Indeed, it has been noticed that when people are taking breath between two utterances,

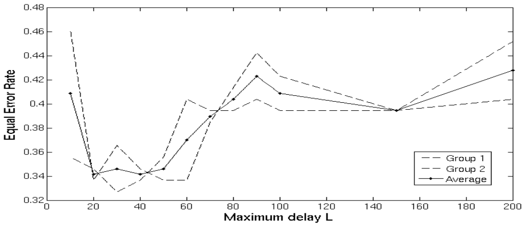

Fig. 5. Influence of maximum delay L on Equal Error Rate

they sometimes open their mouths: this fact is an obvious potential source of error for our system. Figure 6 shows that deleting silence frames gives better performance.

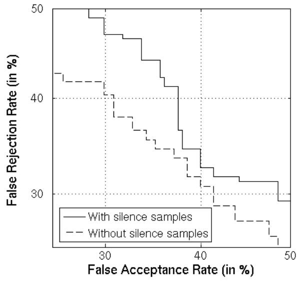

Fig. 6. Influence of silence frames deletion

Figure 7 shows the influence on performance of the method used to balance sample rates. The left curve compares linear interpolation with the duplication of visual samples. It appears that the latter is slightly better, probably because no artificial data is produced. The right curve shows that upsampling the visual samples or downsampling the audio samples does not cause any significant difference.

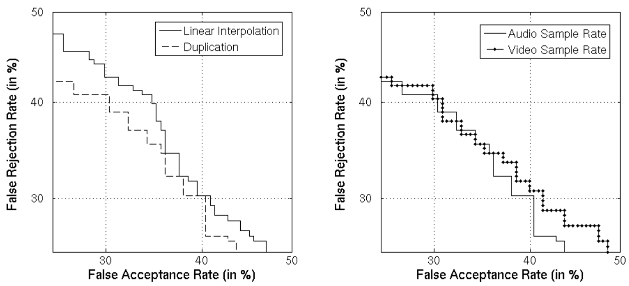

Fig. 7. Influence of sample rate balance

7. CONCLUSION

This paper has argued that account should be taken of the synchronization between the audio and video signals in audiovisual identify verification, in order to defeat sophisticated attacks and forgeries. Since the problem of skilled attacks is not treated by standard evaluation techniques, we have defined new protocols for the BANCA database in order to augment the existing evaluation methodology.

A simple algorithm has been developed to detect and measure synchrony, and tested against two realistic attack scenarios using the BANCA database. An error rate of 0% was reached on the simplest scenario, *Paparazzi*, where a still picture is placed before the camera. Preliminary work using features related to a more accurate shape of the mouth (such as its openness), instead of the simple features we have described, suggest encouraging results. However, robust automatic lip tracking is still needed to further improve the method, and we plan work in this area in order to further improve defences against higher-effort impostor attacks.

8. REFERENCES

[1] S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Fusion of Face and Speech Data for Person Identity Verification," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1065 – 1074, September 1999.

[2] A. Jain, L. Hong, and Y. Kulkarni, "A Multimodal Biometric System Using Fingerprint, Face, and Speech," in *Audio- and Video-based Biometric Person Authentication*, 1999.

[3] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent Advances in the Automatic Recognition of Audio-Visual Speech," *Proc. IEEE*, vol. 91, no. 9, September 2003.

[4] M. Heckmann, F. Berthommier, and K. Kroschel, "Noise Adaptive Stream Weighting in Audio-Visual Speech Recogntion," in *EURASIP J. Appl. Signal Processing*, November 2002, vol. 2002, pp. 1260 – 1273.

[5] 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*, January 2003, vol. 2688, pp. 625 – 638.

[6] S. Garcia-Salicetti, C. Beumier, G. Chollet, B. Dorizzi, J.-L. Jardins, J. Lunter, Y. Ni, and D. Petrovska-Delacretaz, "BIOMET: a Multimodal Person Authentication Database including Face, Voice, Fingerprint, Hand and Signature Modalities," *Audio- and Video-Based Biometric Person Authentication*, pp. 845 – 853, June 2003.

[7] G. Chetty and M. Wagner, ""Liveness" Verification in Audio-Video Authentication," in *8th International Conference on Spoken Language Processing*, October 2004.

[8] P. Perrot, G. Aversano, G. Chollet, and M. Charbit, "Voice Forgery Using ALISP: Indexation in a Client Memory," in *ICASSP 2005*, 2005.

[9] B. Abboud and G. Chollet, "Appearance based Lip Tracking and Cloning on Speaking Faces," in *ISPA 2005*, September 2005.