AUDIO-VISUAL SPEECH SYNCHRONY MEASURE FOR TALKING-FACE IDENTITY VERIFICATION

Hervé Bredin and Gérard Chollet

CNRS-LTCI, GET-ENST, 46 rue Barrault, 75013 PARIS, France

ABSTRACT

We investigate the use of audio-visual speech synchrony measure in the framework of identity verification based on *talking faces*. Two synchrony measures based on *Canonical Correlation Analysis* and *Co-Inertia Analysis* respectively are introduced and their performances are evaluated on the specific task of detecting synchronized and not-synchronized audio-visual speech sequences. The notion of *high-effort impostor attacks* is also introduced as a dangerous threat for current biometric system based on speaker verification and face recognition. A novel biometric modality based on synchrony measures is introduced in order to improve the overall performance of identity verification, and more specifically its robustness to *replay attacks*.

Index Terms— Identification of persons, Speech processing, Video signal processing

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). 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, visiophony and even future personal digital assistants.

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.

The aim of this paper is to exploit this novel characteristic of the talking-face modality within the specific framework of identity verification. In Sec. 2, two algorithms for measuring a degree of synchrony between two multidimensional random variables are overviewed and their application to audio-visual speech is introduced. Sec. 3 specifically deals with the *replay attacks* issue.A novel approach for identity verification using client-dependent synchrony models is then presented in Sec. 4. Finally, attempts to integrate this modality in an existing audiovisual identity framework are presented in Sec. 5.

2. AUDIO-VISUAL SPEECH SYNCHRONY MEASURE

A comprehensive overview of the literature on how to measure the degree of correspondence between audio and visual speech can be found in [1].

2.1. Canonical Correlation Analysis (CCA)

Given two random variables X and Y in $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively, the CCA is a set of two linear projections $\dot{\mathbf{A}}$ and $\dot{\mathbf{B}}$ (called canonic correlation matrices) that aim at whitening X and Y under the constraint of making their cross-correlation diagonal and maximally compact, in the projected spaces. Details for $\dot{\mathbf{A}}$ and $\dot{\mathbf{B}}$ calculation can be found in [2].

Using the first K vectors of $\dot{\mathbf{A}}$ and $\dot{\mathbf{B}}$, we define an audio-visual speech synchrony measure in Eq. 1.

$$\dot{S}_{\dot{\mathbf{A}},\dot{\mathbf{B}}}(X,Y) = \frac{1}{K}\sum_{k=1}^{K}\left|\text{corr}\left(\dot{a}_k^T X, \dot{b}_k^T Y\right)\right| \quad (1)$$

2.2. Co-Inertia Analysis (CoIA)

CoIA was first introduced in biology [3] to find hidden relationships between species and their living environment, and is relatively new in our domain (though it was recently used for liveness test or replay attacks detection in [4]). The difference with CCA stays in the fact that the involved linear projections $\ddot{\mathbf{A}}$ and $\ddot{\mathbf{B}}$ aim at maximizing the *covariance* of X and Y in the projected spaces. Details for $\ddot{\mathbf{A}}$ and $\ddot{\mathbf{B}}$ calculation can be found in [3].

Using the first K vectors of $\ddot{\mathbf{A}}$ and $\ddot{\mathbf{B}}$, we define an audio-visual speech synchrony measure in Eq. 2.

$$\ddot{S}_{\ddot{\mathbf{A}},\ddot{\mathbf{B}}}(X,Y) = \frac{1}{K}\sum_{k=1}^{K}\left|\text{cov}\left(\ddot{a}_k^T X, \ddot{b}_k^T Y\right)\right| \quad (2)$$

2.3. Application to Audio-Visual Speech

Given an audiovisual sequence AV, let us denote by X a random variable that corresponds to the acoustic speech parameters and by Y another random variable for the visual speech parameters.

Audio-Visual Speech Features The first step is to define the random variables X and Y that represent respectively the acoustic and visual speech. Classical Mel-Frequency Cepstral Coefficients (MFCC) are extracted every 10 ms from the audio signal. In our case, we only kept the first 15 MFCCs (the first and second order derivatives were found not to bring any improvement in our preliminary experiments) as the random variable X. For each frame of the video (25 images per second), visual speech features are computed by performing a Discrete Cosine Transform (DCT) of the lip area that is tracked throughout the video (using the algorithm described in [5]). Only the first 30 DCT coefficients (low spatial frequencies) are kept as the random variable Y. Linear interpolation of Y is performed in order to balance the audio and visual sample rates (100Hz and 25Hz respectively, before interpolation).

Synchrony Measure Using transformation matrices $\mathbf{A}$ and $\mathbf{B}$ previously learned by CCA and/or CoIA, it is therefore possible to measure the degree of synchrony between X and Y. We will discuss more precisely how to choose the training set used to learn the matrices $\mathbf{A}$ and $\mathbf{B}$ in Sec. 3.2. Eq. (1,2) are used to obtain a measure

of audio-visual speech synchrony: the higher the more synchronous. Setting a threshold θ finally allows to decide on the synchrony of X and Y: they are synchronized if $S_{\mathbf{A},\mathbf{B}}(X,Y) > \theta$ and not synchronized otherwise.

3. REPLAY ATTACKS

The major weakness of existing audiovisual identity verification systems is that it can easily be fooled by an impostor who *replays* biometric data (recording of the voice, picture of the face, etc.) of his/her target in front of the sensors.

3.1. Impersonation Scenarios

Many databases are available to the research community to help evaluate multimodal biometric verification algorithms, such as BANCA, XM2VTS, BT-DAVID, BIOMET, MyIdea and IV2. 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, i.e. that the impostors use their own voice and face to perform the impersonation trial; which is quite unrealistic.

In this section, we will tackle the *Big Brother* scenario (introduced in [6]): prior to the attack the impostor records a movie of the target's face and acquires a recording of his/her voice. However, the audio and video do not come from the same utterance, so they may not be synchronized. This is a realistic assumption in situations where the identity verification protocol prompts a text for the client to speak.

3.2. Training

As mentioned in Sec. 2, a preliminary training step is needed to learn the projection matrices $\mathbf{A}$ and $\mathbf{B}$ (both for CCA and CoIA) and –then only– the synchrony measures can be computed. This training step can be done using different training sets depending on the targeted application.

World model In this configuration, a large training set of synchronized audio-visual sequences is used to learn $\mathbf{A}$ and $\mathbf{B}$.

Client model The use of a client-dependent training set (of synchronized audio-visual sequences from one particular person) will be more deeply investigated in Sec. 4.

No training One could also avoid the preliminary training set by learning (at test time) $\mathbf{A}$ and $\mathbf{B}$ on the tested audio-visual sequence (X,Y) itself.

Self-training This method is an improvement brought to the above and was driven by the following intuition: *It is possible to learn a* synchrony model *between synchronized variables, but* nothing *can be learned from not-synchronized variables.* Given a tested audio-visual sequence (X,Y), with $X = \{\mathbf{x_1}, ..., \mathbf{x_N}\}$ and $Y = \{\mathbf{y_1}, ..., \mathbf{y_N}\}$, one can therefore try to learn the projection matrices $\mathbf{A}$ and $\mathbf{B}$ from a sub-sequence $(X_{\text{train}} = \{\mathbf{x_1}, ..., \mathbf{x_L}\}, Y_{\text{train}} = \{\mathbf{y_1}, ..., \mathbf{y_L}\})$, with $L < N$ and compute the synchrony measure S on what is left of the sequence: $(X_{\text{test}}, Y_{\text{test}})$ with $X_{\text{test}} = \{\mathbf{x_{L+1}}, ..., \mathbf{x_N}\}$ and $Y_{\text{test}} = \{\mathbf{y_{L+1}}, ..., \mathbf{y_N}\}$. In order to improve the robustness of this method, a cross-validation principle is applied: the partition between training and test set is performed P times by randomly drawing samples from (X,Y) to build the training set (keeping the others for the test set). Each partition p leads to a measure S_p and the final synchrony measure S is computed as their mean: $S = \frac{1}{P}\sum_{p=1}^{P} S_p$.

3.3. Experiments

Experiments are performed on the BANCA database [7], which is divided into two disjoint groups (G1 and G2) of 26 persons. Each person recorded 12 videos where he/she says his/her own text (always the same) and 12 other videos where he/she says the text of another person from the same group: this makes 624 synchronized audio-visual sequences per group. On the other side, for each group, 14352 not-synchronized audio-visual sequences were artificially recomposed from audio and video from two different original sequences with one strong constraint: that the person heard and the person seen pronounce the same utterance (in order to make the boundary decision between synchronized and not-synchronized audio-visual sequences even more difficult to define).

3.4. Results

Fig. 1 are DET curves [8] showing the performance of the CCA (left) and CoIA (right) measures using the different training procedures described in Sec. 3.2. The best performance is achieved with the

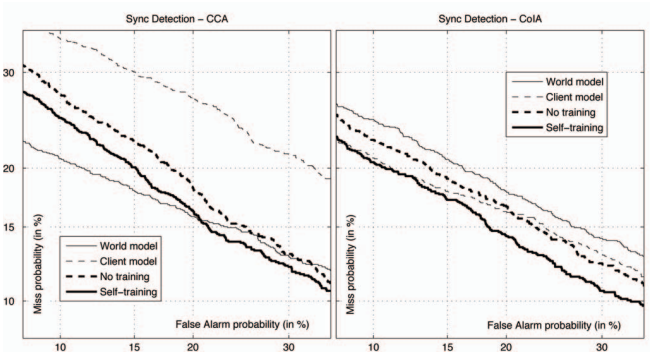

Fig. 1. Synchrony detection with CCA and CoIA

novel *Self-training* we introduced, both for CCA and CoIA, as well as with the CCA using *World model*: it gives an equal error rate (EER) of around 17%. It is noticeable that *World model* works better with CCA whereas *Client model* gives poor results with CCA and works nearly as good as *Self-training* with CoIA. This latter observation confirms what was previously noticed in [9]. The CoIA is much less sensitive to the number of training samples available: the CoIA works fine with little data (*Client model* only uses one BANCA sequence to train $\mathbf{A}$ and $\mathbf{B}$ [7]) and the CCA needs a lot of data for robust training.

Finally, Fig. 2 shows that one can improve the performance of the algorithm for synchrony detection by fusing two scores (based on CCA and based on CoIA). After a classical step of score normalization, a Support Vector Machine (SVM) with linear kernel is trained on one group (G1 or G2) and apply on the other one. The fusion of CCA with *World model* and CoIA with *Self-training* lowers the EER to around 14%. This final EER is comparable to what was achieved in [4].

4. IDENTITY VERIFICATION

According to the results obtained in Fig. 1, not only can synchrony measures be used as a first barrier against replay attacks, but it also led us to investigate the use of audio-visual speech synchrony measure for identity verification (see performance achieved by the CoIA with *Client model*).

Some previous work have been done in identity verification using fusion of speech and lip motion. In [10] the authors apply classical linear transformations for dimensionality reduction (such as Principal Component Analysis - PCA, or Linear Discriminant Analysis

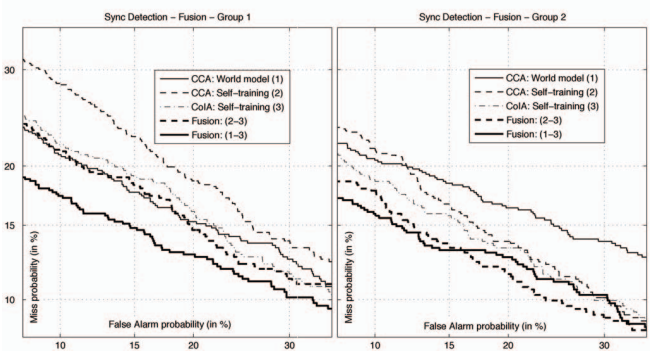

Fig. 2. Fusion of CoIA and CCA

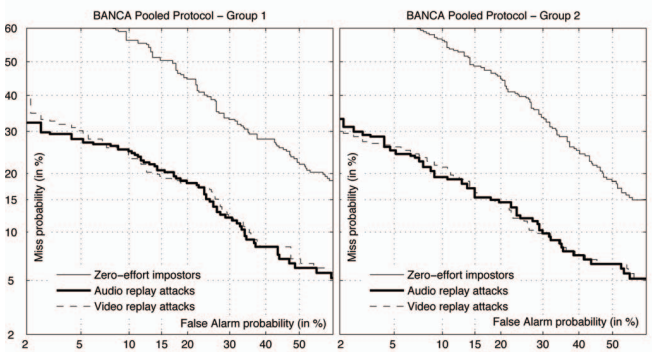

Fig. 3. Identity verification with speech synchrony

- LDA) on feature vectors resulting from the concatenation of audio and visual speech features. CCA is used in [11] where projected audio and visual speech features are used as input for client-dependent HMM models.

Our novel approach uses CoIA with *Client model* (that achieved very good results for synchrony detection) to identify people with their personal way of synchronizing their audio and visual speech.

4.1. Principle

Given an enrollment audio-visual sequence AV_λ from a person λ, one can extract the corresponding synchronized variables X_λ and Y_λ as described in Sec. 2.3. Then, using (X_λ, Y_λ) as the training set, client-dependent CoIA projection matrices $\ddot{\mathbf{A}}_\lambda$ and $\ddot{\mathbf{B}}_\lambda$ are computed and stored as the model of client λ.

At test time, given an audio-visual sequence AV_ϵ from a person ϵ pretending to be the client λ, one can extract the corresponding variables X_ϵ and Y_ϵ. $\dot{S}_{\ddot{\mathbf{A}}_\lambda, \ddot{\mathbf{B}}_\lambda}(X_\epsilon, Y_\epsilon)$ (defined in Eq. 2) finally allows to get a score which can be compared to a threshold θ. The person ϵ is accepted as the client λ if $\dot{S}_{\ddot{\mathbf{A}}_\lambda, \ddot{\mathbf{B}}_\lambda}(X_\epsilon, Y_\epsilon) > \theta$ and rejected otherwise.

4.2. Experiments

Experiments are performed on the BANCA database following the *Pooled* protocol [7]. The impostor accesses are *zero-effort* impersonation attacks since the impostor uses his/her own face and voice when pretending to be his/her target. Therefore, we also investigated replay attacks. The client accesses of the Pooled protocol are not modified, only the impostor accesses are, to simulate replay attacks:

Video replay attack A video of the target is shown while the original voice of the impostor is kept unchanged.

Audio replay attack The voice of the target is played while the original face of the impostor is kept unchanged.

Notice that, even though the acoustic and visual speech signals are not synchronized, the same utterance (a digit code and the name and address of the claimed identity) is pronounced.

4.3. Results

Fig. 3 shows the performance of identity verification using the client-dependent synchrony model on these three protocols. On the original *zero-effort* Pooled protocol, the algorithm achieves an EER of 32%. This relatively weak method might however bring some extra discriminative power to a system only based on the speech and face modalities, which we will study in the following section. We can also notice that it is intrinsically robust to replay attacks: both audio and video replay attacks protocols lead to an EER of around 17%. This latter observation also shows that this new modality is very little correlated to the speech and face modality, and mostly depends on the actual correlation for which it was originally designed.

5. TALKING-FACE FUSION

The system detailed in [5] was used as the basis of this last set of experiments. It consists of the score fusion of two mono-modal biometric recognition algorithms: speaker verification and face recognition. It is not the aim of this paper to describe precisely the algorithms at stake for these two modalities: the interested reader might want to have a look at [5]. Nevertheless, their respective performances are shown in Fig.4. Once again, a SVM with linear kernel is used to discriminate (in the score space) between client and impostor accesses.

Two talking-face systems can then be compared: the original one, based on the fusion of speaker verification and face recognition scores and the new one, based on the fusion of speaker verification, face recognition and client-dependent synchrony scores.

5.1. SVM training

An important point has to be considered regarding the training set used for SVM training. It must contain samples from two sets: scores from genuine client accesses and scores from impostor accesses. Since only *zero-effort* impersonation trials were performed until now, it seemed natural to gather the training set using scores coming exclusively from this type of scenario.

But is it really adapted to the case where we have to tackle with higher effort impostors (with audio and video replay attacks for instance)? Isn't it necessary to take this kind of attacks into account when gathering the training set?

In the following, we will therefore use two types of SVM training set. They share common scores for the client class. They only differ in the samples contained in the impostor class: the first one (which we call *zero-effort training set*) only contains *zero-effort* impostor scores, the second one (called *replay attacks training set*) contains *zero-effort* impostor scores as well as audio and video replay attacks impostor scores.

5.2. Results

Fig. 4 shows the relative performance (on the original *zero effort* BANCA Pooled protocol) of the Speaker-Face system and Speaker-

Face-Sync system. As expected, the latter brings in average an improvement of the EER of about 0.8%.

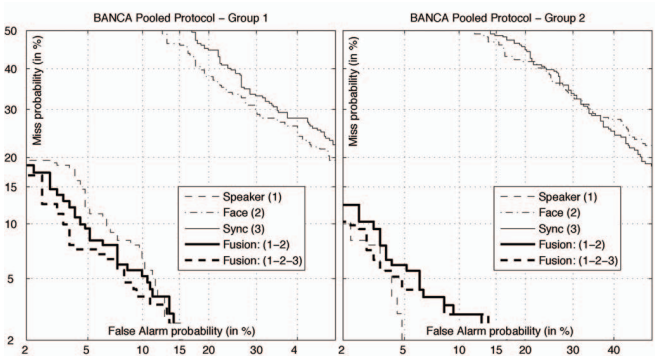

Fig. 4. Zero-effort impostors

Fig. 5 shows the influence of the choice of the SVM training set: *zero-effort training set* on the left and *replay attacks training set* on the right. One can notice on the left that the original Speaker-Face system can be completely fooled with an audio replay attack (46% EER), and that the addition of the Sync module only improves the EER of 1%. However, in the case where high-effort impostors are taken into account during the SVM training process (*replay attacks training set*, right curves), the improvement brought by the Sync module is much more significant: 16%, reducing the EER from 37% to 21% (and even 25% improvement if we consider the original Speaker-Face system). However, note that this type of training

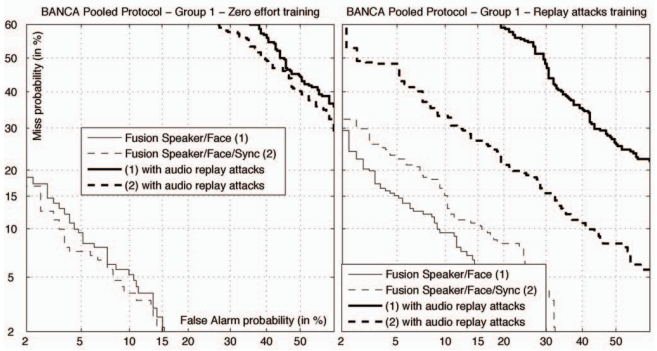

Fig. 5. Replay attacks and training set

degrades the performance of the Speaker-Face-Sync system on the (unrealistic) original *zero-effort* scenario from 6% to 11% EER.

6. CONCLUSION

We have investigated the use of audio-visual speech synchrony measure in the framework of identity verification with *talking faces*. The best algorithm for detection of not-synchronized sequences (based on the fusion of two measures with CCA and CoIA) achieves 14% equal error rate (EER). It might be used as a first barrier against *replay attacks*.

The main contribution of this paper is the introduction of a novel biometric modality based on synchrony measures, achieving 32% EER on the BANCA Pooled protocol. Though it is a weak modality, it has two interesting characteristics. Firstly, it is complementary to the other *talking face* modalities (Speaker and Face) and adds a small (0.8%) improvement on the performance. Secondly, it is intrinsically robust to replay attacks since it is based on the synchrony between audio and visual speech: fused with a Speaker/Face system, it strongly reduces the degradation resulting from audio replay attacks (from 46% EER to 21% EER).

7. ACKNOWLEDGMENTS

This work was partially supported by the EEC through the NoE-BioSecure and the NoE-KSpace.

8. REFERENCES

[1] Hervé Bredin and Gérard Chollet, "Measuring Audio and Visual Speech Synchrony: Methods and Applications," in *International Conference on Visual Information Engineering*, 2006.

[2] Malcolm Slaney and Michele Covell, "FaceSync: A Linear Operator for Measuring Synchronization of Video Facial Images and Audio Tracks," *Neural Information Processing Society*, vol. 13, 2000.

[3] Sylvain Dolédec and Daniel Chessel, "Co-Inertia Analysis: an Alternative Method for Studying Species-Environment Relationships," *Freshwater Biology*, vol. 31, pp. 277–294, 1994.

[4] Nicolas Eveno and Laurent Besacier, "Co-Inertia Analysis for "Liveness" Test in Audio-Visual Biometrics," *International Symposium on Image and Signal Processing Analysis*, pp. 257–261, 2005.

[5] Hervé Bredin, Guido Aversano, Chafic Mokbel, and Gérard Chollet, "The Biosecure Talking-Face Reference System," in *2nd Workshop on Multimodal User Authentication*, May 2006.

[6] Hervé Bredin, Antonio Miguel, Ian H. Witten, and Gérard Chollet, "Detecting Replay Attacks in Audiovisual Identity Verification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2006.

[7] Enrique Bailly-Baillière, Samy Bengio, Frédéric Bimbot, Miroslav Hamouz, Josef Kittler, Johnny Mariéthoz, Jiri Matas, Kieron Messer, Vlad Popovici, Fabienne Porée, Belen Ruiz, and Jean-Philippe Thiran, "The BANCA Database and Evaluation Protocol," in *Lecture Notes in Computer Science*, January 2003, vol. 2688, pp. 625 – 638.

[8] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET Curve in Assessment of Detection Task Performance," in *European Conference on Speech Communication and Technology*, 1997, pp. 1895 – 1898.

[9] Roland Goecke and Bruce Millar, "Statistical Analysis of the Relationship between Audio and Video Speech Parameters for Australian English," *International Conference on Audio-Visual Speech Processing*, 2003.

[10] Claude C. Chibelushi, John S.D. Mason, and Farzin Deravi, "Integrated Person Identification Using Voice and Facial Features," *IEE Colloquium on Image Processing for Security Applications*, , no. 4, pp. 1–5, 1997.

[11] Mehmet Emre Sargin, Engin Erzin, Yucel Yemez, and A. Murat Tekalp, "Multimodal speaker identification using canonical correlation analysis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2006, vol. 1, pp. 613–616.