Multimed Tools Appl (2016) 75:5329–5343
DOI 10.1007/s11042-015-2848-2

Audiovisual synchrony assessment for replay attack detection in talking face biometrics

Elhocine Boutellaa[1,2] · Zinelabidine Boulkenafet[1] · Jukka Komulainen[1] · Abdenour Hadid[1]

Received: 3 May 2015 / Revised: 9 July 2015 / Accepted: 30 July 2015 /
Published online: 18 August 2015

Abstract Audiovisual speech synchrony detection is an important liveness check for talking face verification systems in order to make sure that the input biometric samples are actually acquired from the same source. In prior work, the used visual speech features have been mainly describing facial appearance or mouth shape in frame-wise manner, thus ignoring the lip motion between consecutive frames. Since also the visual speech dynamics are important, we take the spatiotemporal information into account and propose the use of space-time auto-correlation of gradients (STACOG) for measuring the audiovisual synchrony. For evaluating the effectiveness of the proposed approach, a set of challenging and realistic attack scenarios are designed by augmenting publicly available BANCA and XM2VTS datasets with synthetic replay attacks. Our experimental analysis shows that the STACOG features outperform the state of the art, e.g. discrete cosine transform based features, in measuring the audiovisual synchrony.

Keywords Audiovisual speech synchrony · Replay attack · Liveness detection · Talking face biometrics

✉ Elhocine Boutellaa
eboutell@ee.oulu.fi

Zinelabidine Boulkenafet
zboulken@ee.oulu.fi

Jukka Komulainen
jukmaatt@ee.oulu.fi

Abdenour Hadid
hadid@ee.oulu.fi

[1] Center for Machine Vision Research, Computer Science and Engineering, University of Oulu, Oulu, Finland

[2] Telecom Division, Centre de Développement des Technologies Avancées, Algiers, Algeria

1 Introduction

Nowadays, almost every mobile device is equipped with a microphone and a front-facing video camera (e.g. laptops and camera phones) while fingerprint and iris sensors are only just emerging in consumer level devices. Therefore, it is appealing to perform multi-modal person verification combining two natural and non-intrusive biometric modalities, namely face and voice. Joint analysis of face and voice biometrics (usually referred to as talking face), exploiting the synchronization between the speech signal and the corresponding lip motion and their dynamics, provides a unique advantage over ordinary multi-modal fusion techniques. Hence, the synchronization and dynamic properties can be utilized as a?third cue? in addition to the individual voice and face modalities.

Most of the existing audiovisual fusion authentication systems, e.g. [3, 10], however, ignore this special correlation and consider only late integration techniques that operate at score or decision levels of the two individual modalities (i.e. face and voice). Although this kind of late multi-modal fusion of face and voice biometrics increases the recognition performance, it is also very vulnerable to spoofing attacks in which a person tries to masquerade as another one by falsifying the biometric data of the targeted person and thereby gaining an illegitimate advantage. For instance, presentation of pre-recorded audio clip (replay attack) together with a still photograph is already enough to fool the system. It has been also shown that multi-biometrics itself is not inherently robust to spoofing attacks because successfully spoofing one unimodal subsystem is enough to break naively tuned fusion strategies [18].

One approach to counter audiovisual spoofing attacks is to apply dedicated countermeasures for each of the two modalities [16] in order to determine if the presented traits originate from a living legitimate user. Another way to ensure the liveness of a subject is to exploit the intrinsic property of speech and to analyse the synchronization and dynamics of lip movements and voice when a sentence is pronounced. The measurement of correlation and joint dynamics reveals whether or not they come from the same person and whether or not they were recorded at the same time. The same audiovisual synchrony based countermeasure can be used also for detecting video replay attacks consisting of audio and video that have been recorded simultaneously. This can be performed by utilizing challenge-response approach in which the user is prompted to utter a randomly selected sentence or sequence of digits, for instance. Since state-of-the-art facial animation techniques[1] are currently not able to mimic the facial dynamics of the targeted person, enhanced robustness to audiovisual imposture (combination of facial animation and voice transformation) [12] could be obtained if the person-specific motion models are accurate enough. Joint bi-modal analysis of person-specific dynamics of voice and lips does not only provide proof of liveness and, therefore, a first level of robustness against spoofing attacks but it has also shown to improve the performance of audiovisual biometric systems [9].

Some preliminary work (e.g. [1, 4, 5, 20]) has been conducted for investigating the correlation between audio and visual signals by jointly modelling acoustic features and appearance or shape based lip features. However, the used visual speech features in prior works have been limited to static descriptors when the actual motion modelling between consecutive video frames have been ignored. In this paper, we concentrate on the visual speech features. More specifically, we propose to consider also the dynamics between video frames using space-time auto-correlation of gradients (STACOG) [14]. The STACOG description is coupled with Mel-frequency cepstral coefficient (MFCC) features using

[1] http://www.reallusion.com/crazytalk/

canonical correlation analysis (CCA). The effectiveness of the proposed visual speech features is evaluated on realistic and challenging synthetic replay attacks that have been generated by augmenting publicly BANCA [2] and XM2VTS [17] databases. Our experimental analysis depicts that STACOG features outperform discrete cosine transform (DCT) based features [1, 4] that have been among the state-of-the-art visual speech features for measuring correlation between audio and visual signals.

The rest of the paper is organized as follows. In Section 2, we review the related works. In Section 3, we overview the audiovisual replay attack problem and define the attack scenarios addressed in our case. The detail of our approach is presented in Section 4. The experimental analysis, results and discussion are provided in Section 5. Section 6 concludes the present paper and discusses the future work.

2 Related work

Audiovisual synchrony assessment based liveness detection in talking face verification has not received much attention in the literature. Salney and Covell [20] measured the audiovisual synchrony using CCA on facial image intensity and various audio features. They investigated MFCCs, linear-predictive coding (LPC), line spectral frequencies (LSF), spectrograms, and raw audio signal energy. The authors concluded that MFCC are the most appropriate audio features. The synchrony was estimated by rotating audio and visual features using the first direction of CCA and computing the Pearson correlation. Eveno and Besacier [8] used LPC audio features and mouth shape (width, height and area) with co-inertia analysis (Co-IA) to compute a liveness score for the talking face video. They shifted the audio within a time window and derived the synchrony measure based on the maximum audiovisual correlation.

Bredin and Chollet [4] estimated the audiovisual synchrony using DCT visual features, MFCC audio features and Co-IA based method. This synchrony measure was then fused with face and voice authentication scores to make the system robust to spoofing attacks. Score level fusion of independent visual and acoustic components with their correlation was also performed by Chetty [5] to check the biometric liveness. The correlation between the acoustic and the visual speech features was modeled using kernel canonical correlation analysis (kCCA). In this approach, the visual speech features are extracted by a variant of the optical flow algorithm called multi-channel gradient model (MCGM). A comparative study of various visual features and synchrony methods for liveness check was conducted in [1]. The best results were reported using DCT based visual features. The audiovisual synchrony was estimated by fusing Co-IA and coupled hidden Markov models.

Estimating the audiovisual synchrony is also important for applications in multimedia industry. For instance, drift between audio and video tracks in a multimedia show will definitely degrade a viewing experience. Recently, some attempts [7, 15] have been made to detect audiovisual asynchrony and recover the occurred drift in order to render high quality videos. Liu and Sato [15] tried to determine the drift that maximizes the audiovisual correlation within a time window by shifting audio and computing the correlation for different drift sizes. The audiovisual correlation was estimated by kernel density estimation and quadratic mutual information. The detected drift was used to recover the audiovisual synchrony. Also El-Sallam and Mian [7] proposed an algorithm for synchronizing the audio and video signals. In their approach, the speech signal is modeled using Taylor series taking into account the relationship between the speech amplitude and lips movements. The horizontal and vertical lips movement were estimated as change in mouth width and height,

respectively. Then, the delay between the video and audio was determined by inspecting the correlation between the two signals.

Among the aforementioned works, only in paper [5] the used visual speech features were modeling the motion between frames. All other approaches extracted the visual features (such as DCT coefficients, mouth region intensity, mouth shape etc.) frame-wise manner, thus ignoring the actual lip motion between consecutive frames. However, the tight relation between the lip movements and the uttered speech may be lost if the visual features are extracted from consecutive frames separately. Thus, we argue that the use of spatio-temporal features to model the visual speech dynamics is crucial for audiovisual synchrony assessment. Therefore, in our proposed approach, the visual speech is characterized using a motion feature.

3 Attacks against talking face verification systems

In general, the evaluation of audiovisual biometric systems has been conducted on publicly available databases that consider only impostor attacks (zero-effort attacks). In such evaluation protocols, the imposter is assumed to know little information about the authorized client to impersonate. Thus, the imposter uses his own biometrics (voice, face, etc.) to attack the system. These kind of attacks can be easily addressed by increasing the performance of face and speech recognition modules in terms of false acceptance rate.

In more challenging scenarios, an attacker would collect enough information about the targeted person and put more effort to impersonate him/her. For example, an imposter could mimic the voice of the target. In the most challenging case, i.e. a spoofing attack, the attacker forges the biometric traits of the targeted person and uses synthetic artifacts (e.g. a mask, audio, video, etc.) to fool the biometric system.

An audiovisual replay attack is a spoofing attack where the attacker makes use of pre-recorded biometrics (face and voice) of the targeted person. While good quality speech samples might be difficult to acquire covertly together with synchronized frontal face video, the attacker may easily obtain high quality audio and video of the client from separate recordings. For instance, the attacker can use a photo and a voice recording of the targeted person to fool the biometric system. Fortunately, attacks using a still photo can be easily tackled by examining the synchrony between audio and video [4].

More advanced replay attack scenarios can be performed by combining audio and video from different sequences. Various combinations according to the person identity and the uttered sentence can be considered. Zhu et al. [24] and Eveno et al. [8] considered two replay attack scenarios combining audio and video taken from different persons pronouncing different and same sentences. Combining audio and visual speech of the same utterance may lead to a highly synchronized audiovisual sequence. However, the identity of the person could be naturally recognized solely either from the captured voice or face sample. Zhu et al. [24] and Argones Rúa et al. [1] studied the case of combining audio and video taken from recordings of the same person who is uttering different sentences. These three scenarios form a group of realistic fraudulent replay attacks that can be easily produced by an attacker. In this paper, we introduce an even more challenging attack scenario where the audio and video are taken from two different sequences of the same person, who is pronouncing the same sentence. In other words, high level of synchrony is probably perceived between the audio and video but also the biometric samples correspond to the targeted person when conventional audiovisual biometric systems, e.g. [3, 10], can be fooled.

The replay attack scenarios tackled in this paper are synthesized by associating audio and visual speech from different sequences. Let $A_{P,S}V_{P,S}$ be a synchrony audiovisual sequence composed of the acoustic speech A and visual speech V where P is the identity of the speaker and S is the uttered speech. The possible asynchrony combinations of an audio track and a video track from different sequences that may form a significant replay attack are:

- Scenario #1 $A_{P,S}V_{P\prime,S\prime}$: audio and video are from different persons uttering different sentences.
- Scenario #2 $A_{P,S}V_{P\prime,S}$: audio and video are from different persons uttering the same sentence.
- Scenario #3 $A_{P,S}V_{P,S\prime}$: audio and video are from the same person uttering different sentences.
- Scenario #4 $A_{P,S}V_{P,S}$: audio and video are from different sequences of the same person uttering the same sentence.

In order to ensure the liveness of a talking face, we describe below a novel approach and evaluate it under these four audiovisual replay attack scenarios.

4 Proposed approach

Figure 1 depicts an overview of the proposed approach. Given an audiovisual speech sequence, the acoustic and visual features are first extracted separately. Then, a cross-modality analysis is performed on the two features to map them into an audiovisual feature space where their relation will be clearer. Finally, the synchrony between the audio and video is assessed in the audiovisual feature space to decide whether the sequence is a live one or not. In the following, we go through the details of each step.

4.1 Audiovisual speech features

In order to measure the audiovisual synchrony, we need to describe the two speech modalities by extracting appropriate features. For the proposed approach, we selected Mel-frequency cepstral coefficients (MFCC) and space-time auto-correlation of gradients (STACOG) which are introduced in this section.

4.1.1 Audio features

The audio features are represented by the Mel-frequency cepstral coefficients which are among the state-of-the-art features used in the automatic speech and speaker recognition systems. To obtain the MFCC features, the voice signal is first divided into statistically stationary frames. Then, the power spectrum of each frame is computed using the fast Fourier transform. To simulate the selectivity of the human's aural perception, the calculated spectrums are mapped into the Mel-frequency domain using a set of triangular Mel filter bank windows. Finally, the DCT is applied on the log filter bank energies. The MFCCs correspond to the amplitude of the resulting DCT coefficients. Most of the human voice information is captured by the low frequency DCT coefficients while the high coefficients represent the noise in the signal. Thus, only the first coefficients are usually used to represent the speech.

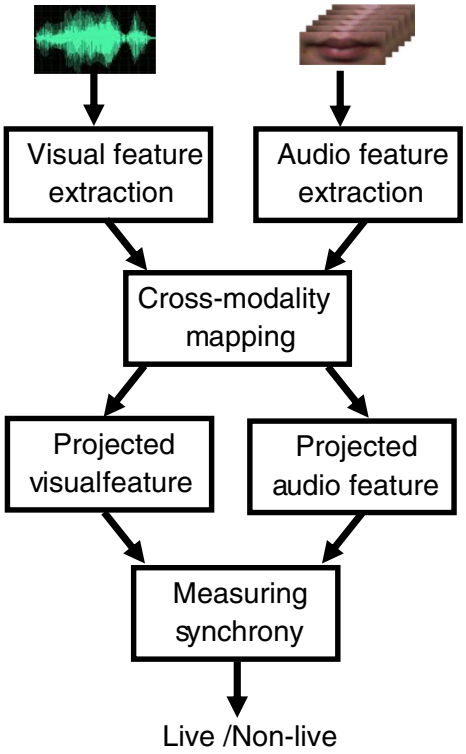

Fig. 1 Overview of the proposed approach

4.1.2 Visual features

In order to extract the visual speech dynamics, the used visual features should take into account the spatio-temporal information in addition to appearance. Therefore, we consider the features called space-time auto-correlation of gradients [14]. STACOG is a motion feature extraction method that encodes the local auto-correlations of space-time gradients in a video for capturing the local geometric characteristics of the moving objects. STACOG has been successfully used for motion recognition (e.g. hand gesture and human action) [14] where it demonstrated superior performance and computation efficiency over similar methods.

Formally, let V be a video and (d_x, d_y, d_t) be the derivative at the point (x, y, t) of V. The space-time gradient vector at each point of the three dimensional plane XYT in V is geometrically represented by the magnitude m, the spatial orientation angle θ and temporal elevation angle ϕ:

$$m = \sqrt{(d_x^2 + d_y^2 + d_t^2)}, \tag{1}$$

$$\theta = arctan(d_x, d_y), \tag{2}$$

$$\phi = arcsin(d_t/m). \tag{3}$$

The two angles θ and ϕ, estimated on a unit hemisphere, define the space-time gradient orientation. This orientation is coded into a histogram by voting weights (the magnitude) to the corresponding bin. This representation considers four orientations along the longitude

Table 1 Equal error rate (%) of the audiovisual synchrony detection for different configurations of the proposed approach under scenarios #1 and #3 on BANCA database

Method	Scenario #1	Scenario #3
DCT-PLS	27.4	27.1
DCT-CCA	21.8	22.5
STACOG-PLS	13.8	15.9
STACOG-CCA	5.6	6.5

by five orientations along the latitude and one orientation for the pole, yielding in a sparse histogram h of $B = 21$ bins.

Once the space-time orientation of the video is encoded, the motion is described by the local auto-correlations. The zeroth and first order auto-correlations defined by (4) and (5) are selected to characterize the local motion:

$$F_0 = \sum_r m(r)h(r), \tag{4}$$

$$F_1(a) = \sum_r min[m(r), m(r+a)]h(r)h(r+a)^T, \tag{5}$$

where $(\cdot)^T$ is the transpose operator and $a = (\Delta x, \Delta y, \Delta t)$ defines the displacement from the point $r = (x, y, t)$. Assuming that the adjacent gradients are highly correlated, the displacements are limited to local neighbors. The STACOG features are finally obtained by stacking the elements of F_0 and F_1 in a single vector.

The presented formulation of the STACOG features correspond to the parameters used in our experiments. For more detailed presentation of the STACOG features considering arbitrary high order auto-correlations, we refer the reader to [13, 14].

4.2 Measuring audio and visual speech correlation

In order to estimate the degree of synchrony between audio and visual signals in a video, we use projection methods that map the two modalities into new space where their relation become clearer. We introduce below the two cross-modality mapping approaches used in our experiments and present the synchrony measure.

4.2.1 Partial least squares analysis

Partial least squares (PLS) [19] is a method for modeling the relation between sets of observed variables. The intuition behind PLS is that the observed data is generated by a small number of latent variables. As a regression model, PLS projects the regressors (input) and responses (output) onto a low dimensional latent linear subspace. The linear projections are chosen such that the covariance between latent scores of regressors and responses

Table 2 Equal error rate (%) of the audiovisual synchrony detection for different configurations of the proposed approach under scenarios #2 and #4 on XM2VTS database

Method	Scenario #2	Scenario #4
DCT-PLS	25.3	27.9
DCT-CCA	18.7	21.0
STACOG-PLS	13.9	17.2
STACOG-CCA	6.9	10.7

is maximized. A linear mapping from the regressors' latent score to response's latent score is then defined.

Formally, let X be the regressor matrix and Y be the response. PLS models X and Y as follows:

$$X = TP^T + E, \tag{6}$$

$$Y = UQ^T + F, \tag{7}$$

$$U = TD + H. \tag{8}$$

T and U are matrices of the extracted PLS latent projections, P and Q are the loading matrices and E, F and H are the residual matrices. D is a diagonal matrix which relates the latent scores of X and Y. PLS finds a one-dimensional projection of X and Y iteratively in a greedy way. It estimates the normalized basis vectors w_i and z_i such that the covariance between the score vectors t_i and u_i (rows of T and U) is maximized:

$$max([cov(t_i, u_i)]^2) = max([cov(Xw_i, Yz_i)]^2), \tag{9}$$

s.t. $\|w_i\| = \|z_i\| = 1$.

4.2.2 Canonical correlation analysis

Canonical correlation analysis (CCA) [11] is a statistical method that measures the relationship between two multidimensional data sets. Given two random variables X and Y, CCA finds the set of two linear projections W and Z (called canonic correlation matrices) that transform X and Y to make their cross-correlation diagonal and maximally compact in the projected spaces:

$$(w_i, z_i) = argmax\, Corr(Xw_i, Yz_i). \tag{10}$$

The vectors w_i and z_i, referred to as CCA basis vectors, form an orthonormal basis for the corresponding transform space. The first pair of these basis vectors (w_1, z_1) is given by the directions along which the projections are maximally correlated. The second pair (w_2, z_2) of CCA basis vectors is obtained by maximizing the same correlation, but this time subject to the constraint that the projections are to be uncorrelated with the first pair of canonical components. This procedure is iterated to extract the remaining canonical pairs.

4.2.3 Synchrony measure

Given the acoustic and visual speech features X and Y of an audiovisual sequence, their synchrony S is computed by projecting them using the first K vectors of matrices W and Z (projection matrices of PLS or CCA) and estimating their correlation in the audiovisual feature space. The synchrony measure used in our approach is similar to that of [4] defined by (11):

$$S_{W,Z}(X, Y) = 1/K \sum_{k=1}^{K} |corr(Xw_k, Yz_k)|\,. \tag{11}$$

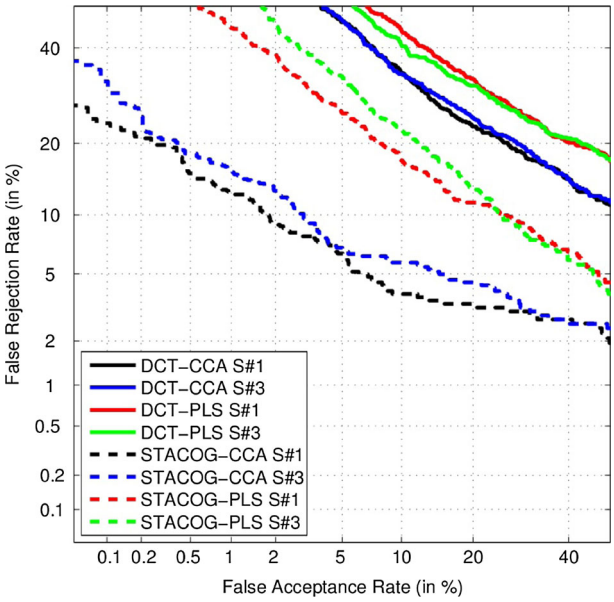

Fig. 2 DET curves of the proposed approach under Scenarios #1 and #3 on BANCA database

5 Experimental evaluation

In this section, we evaluate the performance of two visual features (DCT and STACOG) using the two cross-modality methods (PLS and CCA) under the four attack scenarios described in Section 3. For the experimental validation, we selected two publicly available databases, BANCA and XM2VTS.

5.1 Databases

BANCA database [2] contains 52 speakers divided into two groups of 26 exclusive speakers each. The recording of BANCA was performed over 12 sessions under three conditions (controlled, adverse and degraded). In each session, two recordings were acquired for each subject: one where the speaker pronounces random 12 digit number, his/her information (name, address and date of birth) and a second where he/she pronounces the sentence of another person of the same group. The audio was captured at a frequency of 32 kHz and the video at 25 fps. As the uttered numbers in each recording are random, all the sentences are different. Thus, only scenarios #1 and #3 are evaluated on BANCA database.

The second database is XM2VTS [17] which contains audiovisual sequences of 295 subjects recorded in four sessions over a period of five months. The audio was recorded at 32 KHz frequency and video at 25 fps. The pronounced sentence is the same (Joe took fathers green shoe bench out) in the four sessions for all the speakers. Thus, only scenario #2 and scenario #4 (defined in Section 3) are considered on XM2VTS database as it does not contain persons uttering different sentences. Following the same splitting as in BANCA, we divide the XM2VTS database into two subject-disjoint folds of which each contains half of the subjects.

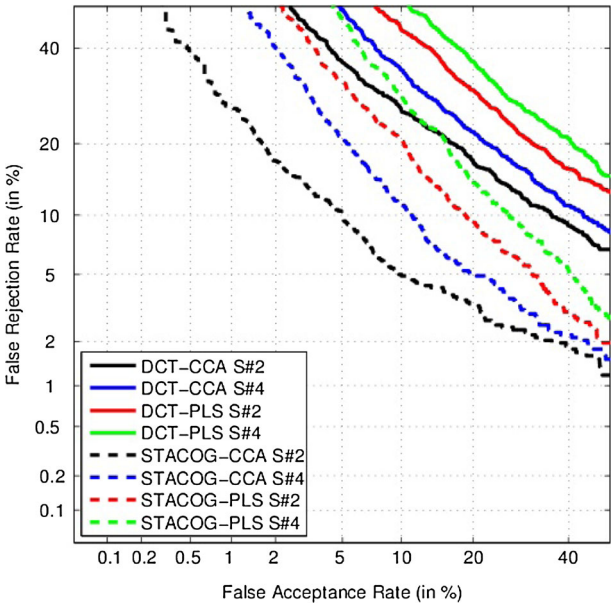

Fig. 3 DET curves of the proposed approach under Scenarios #2 and #4 on XM2VTS database

5.2 Setup

In our experiments, the audible speech is characterized by the 13 first MFCC coefficients. Our aim is not to highly characterize the speech but to evaluate the degree of correspondence between audio and video. Thus, we extract the MFCC features at the video frame rate. Since audio and visual speech features have to be computed at the same rate, the downsampling of audio or upsampling of video features can be avoided in this manner.

Prior to the visual speech feature extraction, we detect and segment the lip region from the videos. A number of robust face landmark estimation approaches were proposed in the literature [21–23]. To be consistent with the related works, we use the Viola-Jones detector [22] for lip detection and segmentation. Therefore, we apply the detector to each frame of the video and resize the detected lip region to 40×70 pixels. For the sake of fair comparison, we have extracted also DCT features used in previous works, e.g. [4, 24]. We apply the DCT to each frame of the video and take the top 35 coefficients, in a zigzag way, as the visual features of the frame.

The liveness of the videos is evaluated in terms of equal error rate (EER) and detection error tradeoff (DET) curves. For both databases, the two groups are alternatively used for training (computation of PLS and CCA projection matrices) and test (synchrony evaluation). For each configuration of our experiments, we tune the parameter K of (11) on one group and use it to evaluate the performance on the other group. Then, we repeat this process alternating the role of the two groups. The reported EER is the average of those of two groups. We have also computed the half total error rate (HTER) in order to compare our results against those of the previous works.

Table 3 Comparison against the state-of-the-art methods

Reference	Replay attack	Database	EER (%)	HTER (%)
eigLip-kCCA+ICA [6]	Still photo	VidTIMIT	8.2	-
		Dafex	9.2	-
DCT-CoIA [24]	Scenario #1	VidTIMIT	13.1	-
	Scenario #2		14.6	-
	Scenario #3		18.3	-
DCT-CoIA [1]	Scenario #3	BANCA	-	16.4
DCT-CCA [1]	Scenario #3		-	13.3
DCT-CHHM [1]	Scenario #3		-	10.6
Lip shape-CoIA [8]	Scenario #1	XM2VTS	12.5	-
	Scenario #2		14.5	-
Our best approach STACOG-CCA	Scenario #1	BANCA	5.6	5.1
	Scenario #3		6.5	6.1
	Scenario #2	XM2VTS	6.9	5.8
	Scenario #4		10.7	10.4

5.3 Results and discussion

Equal error rates for the different configurations of our experiments are reported in Table 1 for BANCA database and in Table 2 for XM2VTS database. The DET curves are depicted in Figs. 2 and 3 for BANCA and XM2VTS databases, respectively.

The results confirm our hypotheses on the importance of visual speech features for measuring the synchrony between voice and lip motion. As it can be seen, the STACOG features outperform DCT features on both datasets and under all scenarios. As we expected, capturing the dynamics of the visual speech using the spatio-temporal motion descriptor is indeed a key cue in mapping visual to acoustic speech. Thus, important visual speech characteristics are lost when extracting visual features in a frame-by-frame manner ignoring the relation between consecutive frames, as it is done in the case of DCT features. In overall, the EER using STACOG compared to DCT is reduced by more than 10 % for PLS and more than 15 % for CCA across scenarios #1 and #3 on the BANCA database (See Table 1). As it can be noticed from Fig. 3 and Table 2, this observation is also consistent for XM2VTS database for scenarios #2 and #4 where significant improvement in terms of EER is achieved using STACOG compared to DCT.

Considering the cross-modality methods, CCA outperforms PLS with the two visual features and across the four scenarios. The gain in EER using CCA compared to PLS is clear on both BANCA and XM2VTS. This means that projecting the audio and visual modalities into a new space where their correlation is maximized captures their relation better than searching for latent variables with maximum covariance in the two modalities.

Regarding the performance over various attack scenarios, the difference is less significant on BANCA database than it is in the case of XM2VTS database. This is due to the nature of data in the two databases. XM2VTS contains shorter videos (about 4 s) than those of BANCA (about 20 s). Moreover, some videos on BANCA dataset include different amount of silence in the beginning whereas on XM2VTS the difference of amount of silence across videos is not that huge. The silence plays important role as the alignment

between speech and video combined from different sequences will not be perfect. Indeed, we have visually examined the attack scenarios using some videos from both BANCA and XM2VTS datasets. We noticed that the asynchrony of XM2VTS synthesized videos from different users was hard to perceive, whereas in the case of BANCA even when associating video and audio of the same person the asynchrony was obvious.

In Table 3, we summarize the synchrony detection errors reported by the state-of-the-art works along with the best performing variants of our proposed approach. The results presented in [1, 8] can be directly compared with the performance of our method because the same experimental data and evaluation protocol were utilized. For completeness, additional results on other datasets were included as indicative performance in other attack scenarios. The superiority of our method is clearly noticed when comparing with the state of the art. This is attributed to the use of STACOG motion feature which allows capturing the dynamics of the visual speech instead of exploiting only lip shape [8] or appearance [1].

6 Conclusion

In this paper, we addressed the problem of detecting audiovisual synchrony for replay attack detection in talking face verification systems. It was shown that modeling the visual speech dynamics is a key factor in detecting the audiovisual asynchrony and replay attacks made by associating audio and video tracks from different recordings.

The best results of our study were obtained using the STACOG for visual speech feature extraction and CCA for joint-modeling of acoustic and visual speech. Under the most challenging scenario, i.e. when combining visual and acoustic speech from different sequences of the same person, our approach achieved an EER of 10.7 % on XM2VTS database. The obtained results demonstrated the effectiveness of the proposed approach over the state of the art. It is also worth mentioning that the proposed approach can be used for other applications like sound source localization and film post production, e.g. dubbing and combining video and audio streams from different devices.

As future work, we plan to study more challenging replay attacks by aligning speech with lip dynamics and using facial animation according to the uttered speech, e.g. voice conversion and synthetic speech. We intend also to perform more analysis on the effect of the speech content on the synchrony. In addition, it is important to investigate better methods for joint modeling of audiovisual features and for mouth segmentation.

Acknowledgments E. Boutellaa is acknowledging the financial support of the Algerian MESRS and CDTA under the grant number 060/PNE/ENS/FINLANDE/2014-2015. The support of the Academy of Finland and Infotech Oulu Doctoral Program is also acknowledged.

References

1. Argones Rúa E, Bredin H, Garca Mateo C, Chollet G, Gonzlez Jimnez D (2009) Audio-visual speech asynchrony detection using co-inertia analysis and coupled hidden markov models. Pattern Anal Applic 12(3):271–284

2. Bailly-Baillire E, Bengio S, Bimbot F, Hamouz M, Kittler J, Marithoz J, Matas J, Messer K, Popovici V, Pore F, Ruiz B, Thiran JP (2003) The banca database and evaluation protocol. In: Kittler J, Nixon M (eds) Audio- and Video-Based Biometric Person Authentication, Lecture Notes in Computer Science, vol 2688, pp 625–638. Springer, Berlin
3. Ben-Yacoub S, Abdeljaoued Y, Mayoraz E (1999) Fusion of face and speech data for person identity verification. IEEE Transactions on Neural Networks 10(5):1065–1074
4. Bredin H, Chollet G (2008) Making talking-face authentication robust to deliberate imposture. In: International conference on acoustics, speech and signal processing (ICASSP), pp 1693–1696
5. Chetty G (2009) Biometric liveness detection based on cross modal fusion. In: 12th International Conference on Information Fusion, FUSION '09, pp 2255–2262
6. Chetty G (2010) Robust audio visual biometric person authentication with liveness verification. In: Sencar H, Velastin S, Nikolaidis N, Lian S (eds) Intelligent multimedia analysis for security applications, studies in computational intelligence, vol 282, pp 59–78. Springer, Berlin
7. EL-Sallam AA, Mian AS (2011) Correlation based speech-video synchronization. Pattern Recogn Lett 32(6):780–786
8. Eveno N, Besacier L (2005) Co-inertia analysis for "liveness" test in audio-visual biometrics. In: International symposium on image and signal processing and analysis, (ISPA), pp 257–261
9. Faraj MI, Bigun J (2007) Audio-visual person authentication using lip-motion from orientation maps. Pattern Recogn Lett 28(11):1368–1382
10. Fauve B, Bredin H, Karam W, Verdet F, Mayoue A, Chollet G, Hennebert J, Lewis R, Mason J, Mokbel C, Petrovska D (2008) Some results from the biosecure talking face evaluation campaign. In: IEEE international conference on acoustics, speech and signal processing, ICASSP 2008, pp 4137–4140
11. Hardoon DR, Szedmak SR, Shawe-taylor JR (2004) Canonical correlation analysis: An overview with application to learning methods. Neural Comput 16(12):2639–2664
12. Karam W, Bredin H, Greige H, Chollet G, Mokbel C (2009) Talking-face identity verification, audiovisual forgery, and robustness issues. EURASIP Journal on Advances in Signal Processing 4
13. Kobayashi T, Otsu N (2008) Image feature extraction using gradient local auto-correlations. In: Proceedings of the 10th European conference on computer vision: Part I, ECCV '08, pp 346–358. Springer, Berlin
14. Kobayashi T, Otsu N (2012) Motion recognition using local auto-correlation of space-time gradients. Pattern Recognit Lett 33(9):1188–1195
15. Liu Y, Sato Y (2010) Recovery of audio-to-video synchronization through analysis of cross-modality correlation. Pattern Recognit Lett 31(8):696–701
16. Marcel S, Nixon MS, Li SZ (2014) Handbook of Biometric Anti-Spoofing: Trusted Biometrics Under Spoofing Attacks. Springer
17. Messer K, Matas J, Kittler J, Jonsson K (1999) Xm2vtsdb: The extended m2vts database. In: 2nd international conference on audio and video-based biometric person authentication, pp 72–77
18. Rodrigues RN, Ling LL, Govindaraju V (2009) Robustness of multimodal biometric fusion methods against spoof attacks. Journal of Visual Language and Computing 20(3):169–179
19. Rosipal R, Krmer N (2006) Overview and recent advances in partial least squares. In: Saunders C, Grobelnik M, Gunn S, Shawe-Taylor J (eds) Subspace, Latent Structure and Feature Selection, Lecture Notes in Computer Science, vol 3940, pp 34–51. Springer, Berlin
20. Slaney M, Covell M (2000) Facesync: A linear operator for measuring synchronization of video facial images and audio tracks. In: Neural information processing systems conference, pp 814–820
21. Uřičář M, Franc V, Thomas D, Akihiro S, Hlaváč V (2015) Real-time multi-view facial landmark detector learned by the structured output svm. In: IEEE international conference on automatic face and gesture recognition conference and workshops. IEEE
22. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vis 57(2):137–154
23. Zhu X, Ramanan D (2012) Face detection, pose estimation, and landmark localization in the wild. In: IEEE conference on computer vision and pattern recognition, pp 2879–2886
24. Zhu ZY, He QH, Feng XH, Li YX, Feng Wang Z (2013) Liveness detection using time drift between lip movement and voice. In: International conference on machine learning and cybernetics (ICMLC), vol 02, pp 973–978

Elhocine Boutellaa received his M. Sc in computer science in 2011. He is working towards his PhD at école nationale supérieure d ′informatique in Algeria. He is currently a visiting student to the center for machine vision, University of Oulu. He has been working as research assistant with centre de development des technologies avancées since 2010. His research interests include biometrics, face analysis and biometric antispoofing.

Zinelabidine Boulkenafet received his engineering degree and master degree in computer science from the National School of Computer Science Algiers, Algeria. He is currently a PhD student in Computer Science and Engineering at the University of Oulu, Finland. His study and research interests include signal and image processing, biometrics and spoofing detection.

Jukka Komulainen received his M.Sc. degree in information engineering from the University of Oulu in May 2010. He is currently pursuing a Ph.D. degree in computer science and engineering at University of Oulu. His research interests include image processing, machine learning and pattern recognition with a particular focus on biometrics and especially biometric anti-spoofing.

Abdenour Hadid received the Doctor of Science in Technology degree in electrical and information engineering from the University of Oulu (Finland) in 2005 and the Professional Specialization Studies in Business Engineering from Oulu University of Applied Sciences in Finland in 2008. Since 2010, he is an Adjunct Professor (Docent) and senior researcher in the Center for Machine Vision Research of the University of Oulu. He made significant contributions to the state-of-the-art and his work is gaining increasing interest in the scientific community. According to Google scholar (April 2015), his h-index is 24 and his work is already cited more than 5500 times. He lectured many invited talks and tutorials in international events (e.g. FG2011, ICIAP2011, ACCV2009, ICPR2008 etc.). He was a work package leader in the FP7 EU project TABULA RASA dealing with vulnerabilities of different biometric systems under spoofing attacks. He co-organized the first Algerian summer school on biometrics in Algiers in 2010. He is the tutorial chair of IPTA2012 and IPTA 2014 conferences and the main organizer of many workshops (ACCV/LBP2012, ECCV/LBP2014, ECCV/Soft Biometrics 2014 etc.). He has been visiting the Institute of Automation at the Chinese Academy of Science (Prof. Stan Li) in spring 2006, the Institute of Industrial Science at the University of Tokyo (Prof. Y. Sato) in summer 2009, the Eurecom Institute at Sophia Antipolis (Prof. J-L Dugelay) in summer 2010, the University of Cagliari (Prof. F. Roli) in fall 2011, the university of Valenciennes in France (Prof. A. Taleb-Ahmed) in summer 2013 and the university of Montreal in Canada (Prof. E. Granger) in summer 2014.