On the robustness of audiovisual liveness detection to visual speech animation

Jukka Komulainen[1], Iryna Anina[1], Jukka Holappa[1], Elhocine Boutellaa[2] and Abdenour Hadid[1]
[1]Center for Machine Vision and Signal Analysis, University of Oulu, Finland
[2]Telecom Division, Centre de Développement des Technologies Avancées, Algiers, Algeria
{jukmaatt,isavelie,jukkaho,eboutell,hadid}@ee.oulu.fi

Abstract

Audiovisual speech synchrony detection is an important liveness check for talking face verification systems to make sure that the (pre-defined) content and timing of the given audible and visual speech samples match. Nowadays, there exists virtually no technical limitations for combining transferable facial animation and voice conversion (or synthesis) to create an ultimate audiovisual artifact that is able to spoof even advanced random challenge-response based liveness detection. In this study, we investigate the performance of the state-of-the-art text-independent lip-sync detection techniques under presentation attacks consisting of audio recordings of the targeted person and corresponding animated visual speech. Our experimental analysis with three different photo-realistic visual speech animation techniques reveals that generic synchrony models can be fooled even with underarticulated but synchronized lip movements. Thus, measuring audio-video synchrony or content alone is not enough for securing audiovisual biometric systems. Our preliminary findings suggest though that adaptation of person-specific audiovisual speech dynamics is one possible approach to tackle these kinds of high-effort attacks.

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. Although audiovisual biometric systems considering late multi-modal integration of face and voice increases the recognition performance compared to the individual modalities, they are also very vulnerable to presentation 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 talking face verification considering late fusion [3].

One approach to counter audiovisual presentation attacks is to apply dedicated countermeasures for each of the two modalities [15] in order to determine if the presented face and voice traits originate from a living legitimate user. Unfortunately, these kind of techniques have shown to have problems in generalizing their great performance beyond the development data (laboratory conditions) [15]. Thus, much work is still needed to get them working in the open environments of practical use case scenarios. 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 passphrase is pronounced. The measurement of correlation and joint dynamics can be considered as liveness detection of the recording process as it determines whether the content and timing of captured audible and visual speech match, i.e. if they were recorded from the same source at the same time.

The audiovisual synchrony detection can be performed using text-independent [1, 2, 3, 7, 20], and text-dependent [14, 17] approaches. Text-independent approaches are effective in detecting crude attacks in which the attacker has managed to acquire only separate audio and video recordings (or photo) of the targeted person that are presented to the biometric system. However, they are naturally powerless under pre-recorded video replay attacks with synchronized audiovisual speech. Text-dependent synchrony assessment methods overcome this issue by utilizing challenge-response approach in which the biometric system prompts the user a randomly selected sentence or sequence of digits [14, 17] (challenge) and then verifies whether the preassigned utterance can be recognized in both modalities within the specified time window (response).

The state-of-the-art real-time voice conversion techniques are capable of fooling both humans and automatic systems [6]. The recent advances in facial reenactment [21] have enabled real-time re-rendering the facial expressions and visual speech of the source actor on top of a video stream of the targeted person in a photo-realistic manner such that it seamlessly blends even with the real-world illumination. As a consequence, there exists nowadays no virtual technical limitations for combining transferable facial animation and voice conversion (or synthesis) to create an interactive audiovisual artifact mimicking both voice and face biometrics including matching visual speech. Therefore, the use of generic lip-sync model or random challenge-response based liveness check alone for securing audiovisual biometric systems can be questioned. It is worth highlighting that Wells Fargo is experimenting with talking face verification for mobile banking and the pilot system uses only text-dependent random challenge-response based liveness check provided by SpeechPro's VoiceKey.OnePass[1].

The literature on facial animation in spoofing is scarce. In [3], it was suggested that subject-specific synchrony models might be robust to higher-effort forgeries like face animation but no experimental validation was conducted. Few studies [10, 22] have investigated the performance of talking face verification systems under synthetic audiovisual artifacts and demonstrated that they are indeed vulnerable to these kind of attacks. However, these prior works did not consider any kind of presentation attack detection (PAD) in their experiments. Furthermore, while the commercial facial animation software[2] used in [22] is practical in creating inveractive avatars, it is not capable of producing photo-realistic synthetic talking face with natural motion. In [10], the only association of the facial animation with the corresponding synthesized speech was length of the output video, i.e. only different facial expressions were animated instead of (synchronized) visual speech.

In this present study, we address these issues and investigate the performance of the state-of-the-art text-independent lip-sync detection methods under audiovisual presentation attacks combining audio recordings of the targeted person with corresponding animated visual speech. Our experimental analysis with three different photo-realistic image-based visual speech animation techniques [4, 8, 24] reveals that generic synchrony assessment models can be fooled even with underarticulated animated speech. Therefore, measuring audio-video synchrony or content alone is indeed not enough for securing audiovisual biometric systems. Our preliminary investigations show though that adaptation of person-specific audiovisual speech dynamics is one possible approach to tackle these high-effort attacks, thus confirming the intuition of [3].

The rest of the paper is organized as follows. In Section 2, we introduce the lip-sync detection methods investigated in this study. The visual speech synthesis techniques used for fooling the synchrony measures are described in Section 3. The experimental analysis is provided in Section 4. Finally, our conclusions are presented in Section 5.

2. Lip-sync detection

Fig. 1 depicts an overview of widely used generic lip-sync detection pipeline. First, acoustic and visual features are extracted separately from the given audiovisual speech sequence. Then, the two different features are projected into a common space in which their correlation can be evaluated as a function of time [20]. Finally, a synchrony measure is applied to determine whether the observed audiovisual signals originate from the same source, i.e. a talking face. The following sections introduce the audio-video synchrony method investigated in this paper and describe briefly the configurations of each phase in the pipeline.

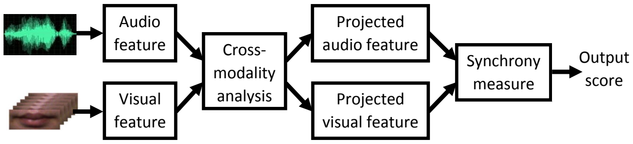


Figure 1: Generic lip-sync assessment pipeline.

2.1. Features

We chose to use the Mel-frequency cepstral coefficients (MFCC) as audio features because they are used almost without an exception in state-of-the-art lip-sync detection methods. For the visual speech representation, we considered two fundamentally different approaches, static features describing appearance of the mouth region, discrete cosine transform (DCT), and dynamic features modeling the actual motion between consecutive video frames, space-time auto-correlation of gradients (STACOG) [13]. DCT was selected because it has been the most used feature in related works, e.g. [1, 3], whereas STACOG are the state-of-the-art features in audiovisual speech synchrony assessment [2].

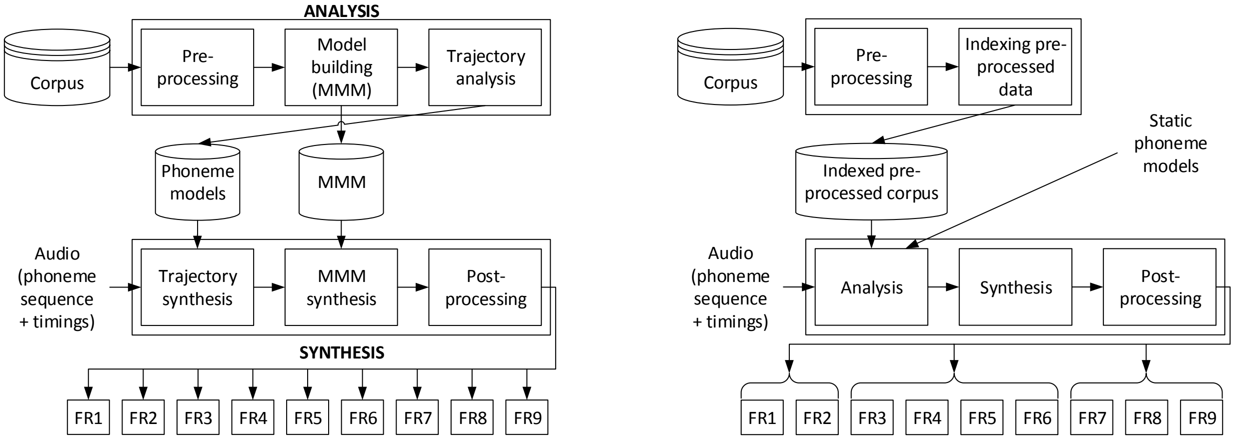
2.2. Joint space analysis

We considered canonical correlation analysis (CCA) based cross-modality mapping[3] that was originally proposed for audiovisual speech synchrony assessment in [20]. Given two signals X and Y, CCA finds a linear projection that maximizes their cross-correlation in the resulting common space, thus the first pair of basis vectors (w_1, z_1) gives the direction along which the signals are maximally correlated. The second pair (w_2, z_2) of CCA basis vectors is obtained by maximizing the same correlation but subject to the constraint that the projections are to be uncorrelated with the first pair of canonical components. This procedure is iterated in order to find the remaining CCA basis vectors w_i and z_i that form an orthonormal basis for the joint space.

[1] http://speechpro-usa.com/media/news/2015-07-29

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

[3] We also tried co-inertia analysis (CoIA) [5] as cross-modality mapping method because it has shown to be effective in many related studies, e.g. [1, 3, 7]. However, we were not able to achieve comparable performance to CCA with CoIA in our experiments.



(a) (b)

Figure 2: Overview of the considered a) generative model based [8] and b) concatenative visual speech synthesis [24].

2.3. Synchrony measure

The lip-sync detection is based on the synchrony measure proposed in [3]. Given the acoustic and visual speech features X and Y of an audiovisual sequence, their synchrony S is estimated by computing the overall correlation of the two projected speech signals along the first K dimensions of the joint space:

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

3. Visual speech synthesis

This section gives a brief overview on the three different image-based visual speech synthesis techniques used for evaluating the robustness of generic lip-sync detection. The animation methods considered in this work can be divided into model based and concatenative approaches [16].

3.1. Generative model based visual speech synthesis

One way to synthesize novel speech videos is to parametrize the original visual speech and build a statistical model capable of generating novel speech from text input [16]. Ezzat *et al.* [8] proposed to build a multidimensional morphable model (MMM) based on the given visual speech corpus and then learning the trajectories of the original speech in the resulting MMM space (see, Fig. 2a).

The MMM is built by selecting a reference image and a set of images containing key mouth shapes and computing the optical flows that morph each key image to the reference image. Any novel video frame can be then represented in the model space by a set of parameters that are used to synthesize the target frame based on the reference image and the pre-computed optical flow vectors. During trajectory analysis, the entire corpus is projected into the model space to build phoneme models, i.e. time series of parameters. The synthesis of novel speech based on a target phoneme sequence is solved as a regularization problem by minimizing both target and smoothness terms of the trajectory in the model space. Given a time series of parameters, MMM is then able to synthesize individual images that form the final synthesized visual speech video.

The MMM requires relatively large video corpus for every novel speaker (about 8 minutes of speech was used in [8]). In [4], a matching-by-synthesis approach was proposed to transfer an original MMM trained from a large speech corpus to a novel person with limited training data (only about 15 seconds of video). A semi-automatic (i.e. manually initialized) approach was used for replacing all the original prototype images with ones of the novel subject based on the similarity of mouth appearance. The transferable MMM (T-MMM) can already produce animated speech that resembles the appearance of the novel person but with the speaking style of the original user. Thus, an adaptation method was also introduced in [4] to refine the MMM phoneme model to match the speaking style of the novel person.

3.2. Concatenative visual speech synthesis

Instead of synthesizing images using a generative model like MMM, novel visual speech can be generated by concatenating a set of original video segments of the target speaker that match (partially) the target phoneme sequence. These kind of concatenative methods have to deal with the trade-off between the continuity of the synthesized visual speech and the quality of lip-syncing as stitching of longer segment leads to more natural motion while the resulting animation may not match the phonetic context that well.

Zhou *et al.* [24] proposed a concatenative visual speech animation system (see, Fig. 2b) that aims at minimizing this trade-off. The given visual speech corpus is pre-processed and indexed in order to speed up the analysis stage that needs to be performed separately for every target phoneme sequence. During analysis, an optimal set of variable-length phoneme segments are selected with respect to concatenation penalty (continuity) and cost for replacing one particular phoneme by some other (lip-sync) based on external phoneme models. Novel visual speech is then synthesized by concatenating the chosen overlapping video segments so that their appearance similarity is maximized at the transition point.

4. Experimental analysis

The following section introduces first the experimental setup. The experimental analysis itself consist of three main steps: evaluating 1) the effectiveness of generic lip-sync detection under audiovisual replay attacks, and 2) its robustness under audiovisual replay attacks and the different visual speech animation based attacks in cross-database scenario, and 3) exploring the effectiveness of subject-specific audiovisual synchrony detection to counter facial animation based attacks.

4.1. Experimental setup

In the literature, different configurations of MFCC features have been considered for describing the audible speech. Inspired by [1, 2, 3], three variations combining the 13 first MFCC coefficients and their first (Δ) and second-order (Δ^2) derivatives are used in our experiments (referred to as MFCC, MFCC-Δ and MFCC-Δ-Δ^2). Accurate characterization of the speech is not the objective when evaluating the degree of synchrony in the observed audiovisual speech [2]. Thus, we compute the audio features at video frame rate to simplify visual speech feature extraction.

We aimed at mitigating the effect of inaccurate mouth detection in visual speech feature extraction while still keeping the pipeline automatic. We selected the face landmark detector proposed in [11] because it performs robustly on the used audiovisual datasets. We used the implementation available in the dlib library [12] for determining eye and mouth locations in every video frame and followed the strategy proposed in [23] to get a good approximation of the whole mouth region (see, Fig. 3). The resulting rectangular mouth image is the resized to 70×40 pixels from which we extract the first 35 DCT coefficients corresponding to the low spatial frequencies in a zigzag way and the 1584-dimensional STACOG feature vector using the default parameters of the publicly available implemention [13].

This study focuses mainly on investigating whether audio-video synchrony detection techniques are able to tell a difference between the visual speech animation based attacks and the corresponding original videos. Therefore, the number of CCA dimensions K used in synchrony measure (see, Equation 1) is tuned separately for each dataset and feature configuration so that best possible performance is obtained in detecting unsynchronized audiovisual speech. We use equal error rate (EER) to measure how well the lip-sync detection methods are able to determine whether the observed audiovisual speech is originated from a genuine subject or an attack.

4.2. Generic lip-sync model for audiovisual replay attack detection

We begin our experiments by evaluating the effectiveness of the different lip-sync detection methods under replay attacks combining audio and video recorded from different sources. We reproduced the experiments of [2] by following the same evaluation protocol on the the XM2VTS database [18]. The dataset consisting of audiovisual speech sequences of 295 subjects is split into two subject-disjoint halves. The audiovisual replay attacks are created by switching the audio tracks between sequences of the same person recorded during four sessions. Every subject is pronouncing the same sentence (Joe took fathers green shoe bench out), thus high level of synchrony is probably perceived between the observed audio and video. The lip-sync detection models are trained on the real videos of one group and the resulting model is the evaluated on the other group. This process is repeated by alternating the role of the two folds and reported EER is the average of the two tests.

Method	Improved	Original [2]
DCT+MFCC	10.52	21.10
DCT+MFCC-Δ	10.86	-
DCT+MFCC-Δ-Δ^2	**9.64**	-
STACOG+MFCC	6.21	10.70
STACOG+MFCC-Δ	**5.58**	-
STACOG+MFCC-Δ-Δ^2	5.75	-

Table 1: Baseline performance of different generic lip-sync detection methods in terms of EER (%) under audiovisual replay attacks with unsynchronized audio and video tracks.

The baseline performance of the different audio-video synchrony assessment methods under the traditional replay attacks is reported in Table 1. The results are consistent [2] as the configurations using STACOG as visual features are more robust than the ones using DCT. However, it is worth noting that both visual features achieve better performance than the best results reported in [2], thus confirming clearly the benefits of the more advanced mouth localization.

4.3. Generic lip-sync model under animated visual speech based attacks

It is reasonable to assume that in real-world applications it is not feasible to train the synchrony models on the audio-

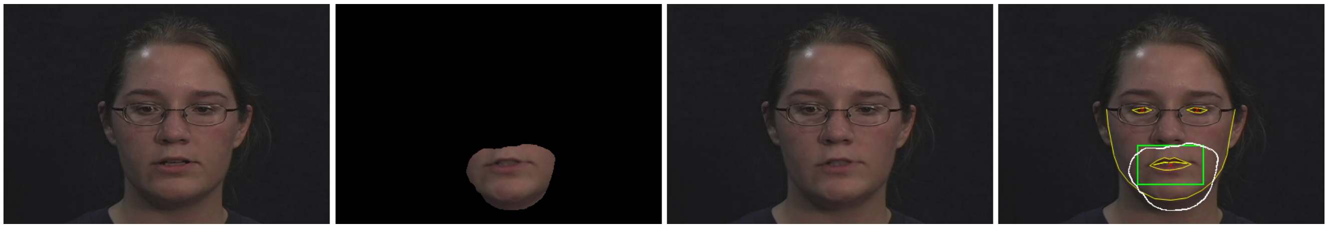


Figure 3: The background compositing process: a background video frame (with natural head and eye movement), a mask with generated visual speech and the final composite video frame, respectively; and used mouth detection strategy (right) where mouth analysis window is visualized with green and masking used in animation with white outline.

visual data of the end-users captured with the target device because time-consuming data collection is not particularly a desirable property in biometric systems. Therefore, generic lip-sync models trained on large-scale representative development datasets are applied in practical applications.

In the following, we conduct a set of cross-database experiments to simulate this condition by using the generic synchrony models of trained on the XM2VTS corpus and evaluating their robustness on three additional datasets consisting of traditional audiovisual replay attacks and three different photo-realistic facial animation based attacks. From now on, all audiovisual replay attacks are created by switching the audio and video tracks between original audiovisual speech sequences. This corresponds to an attack scenario in which audio and video originate from the same person uttering different sentences.

All used facial animation techniques require head pose normalization prior superimposing a mask to extract the region of interest (ROI), e.g. mouth-chin area, used for creating the visual speech synthesis. The final output videos of talking faces in every dataset have been created by compositing the synthesized mouth onto a background sequence containing natural head and eye movements (see, Fig. 3). The stitching is conducted by replacing the mouth-chin area normalized from the original image with the synthesized one and undoing the pose correction. The mouth mask is smoothed at the edges to perform a seamless blend between the background image and the synthesized mouth. The interested reader is referred to see [8, 24] for more details.

There is some overlap between the background video frame and the mouth region used for computing the visual speech features (see, Fig. 3). It is worth highlighting that we do not exploit the possibly visible artifacts introduced in making the composite videos, e.g. boundaries between the mouth mask and the background sequence. First, the overlap corresponds to facial regions whose motion is insignificant compared with that of the mouth area. Furthermore, the original and the corresponding animated videos undergo through the same video compression pipeline, in order to mitigate the effect of factors unrelated to the audio-video synchrony, e.g. video codec and resolution, and possibly the composite artifacts as well. Only genuine audiovisual speech samples are seen during training, i.e. no prior knowledge of the attacks is introduced, thus the lip-sync models cannot describe any other cues besides natural synchrony.

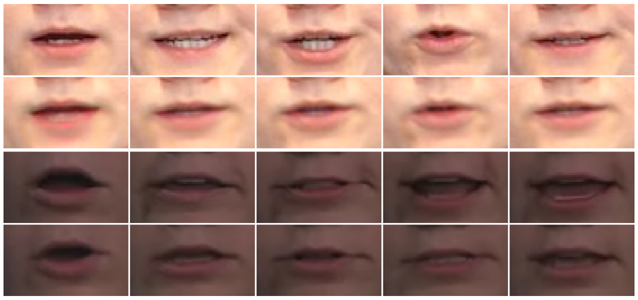


Figure 4: Examples of original video frames and corresponding synthetic images generated with MMM (top) and concatenative visual speech synthesis datasets (bottom).

4.3.1 Underarticulated visual speech animation

We implemented the MMM based visual speech animation method [8] and collected a visual speech corpus consisting of 196 videos of one subject uttering naturally phonetically balanced TIMIT sentences [9]. 186 videos of the dataset were used for building the MMM, while the remaining ten sequences were left for testing. The resolution of the videos is 640 × 480 and frame rate 30 fps. The gradient descent learning method used for improving the articulation of MMM [8] does not perform well on our visual speech corpus. The untrained MMM tends to average out the mouth movements so that it looks underarticulated but the resulting speech is still synchronized with the audio. This crude facial animation can be considered as the first challenge to the lip-sync detection methods. Fig. 4 shows some snapshots of both real and corresponding animated video frames.

Table 2 presents the performance of the different audiovisual synchrony assessment techniques under the replay attacks and attacks with underarticulated visual speech animation of the MMM dataset. The results depict that the

Method	Replay	Animation
DCT+MFCC	25.56	40.00
DCT+MFCC-Δ	20.00	40.00
DCT+MFCC-Δ-Δ^2	**20.00**	**30.00**
STACOG+MFCC	**3.33**	**10.00**
STACOG+MFCC-Δ	10.00	30.00
STACOG+MFCC-Δ-Δ^2	13.33	20.00

Table 2: Cross-database performance of the generic lip-sync detection methods in terms of EER (%) under the replay and underarticulated facial animation based attacks of the MMM dataset.

generic lip-sync detection is able to generalize in detecting audiovisual replay attacks beyond the development set. The more important finding is, however, that the performance of all lip-sync based liveness detection techniques degrades dramatically when facing the new attack type based on the underarticulated visual speech synthesized based on the original audio track. Thus, it seems that even crude but smooth facial animation synchronized with audio content is enough for fooling liveness detection based on generic lip-sync models.

4.3.2 Transferable visual speech animation

While the excessive requirement of training data of the targeted person limits the use of MMM [8], the transferable MMM [4] is more realistic in real-world spoofing scenarios. We experimented also with a small T-MMM dataset provided by the authors of [4]. Each of the ten short video clips includes a person uttering one digit from one to ten. The animated videos were generated using the phoneme model of the original speaker and the adapted model mimicking the speaking style of the novel person, thus the dataset consists of ten real and 20 animated videos. The resolution of videos is 720 × 480 and frame rate 30 fps. Both real and animated visual speech are well-articulated. In real-world applications, the use of single digits is not probably enough for secure and robust authentication, thus we combine the digits into variable length passphrases like in [17]. The length of the PIN codes varies from three to five and the different permutations contain the same digit only once.

The results of the experiments on the T-MMM dataset can be seen in Table 3. Again, while the generic audio-video synchrony models are able to generalize in detecting audiovisual replay attacks very well, their performance drops significantly when the transferable facial animation is introduced to the systems. Similarly to [17], increasing the number of digits in passphrase decreases the EER suggesting that longer PIN codes improve the robustness of the system under replay attack detection.

4.3.3 Concatenative visual speech animation

The concatenative visual speech model [24] was build using the publicly available Audiovisual Database of Spoken American English [19] where the participants were ask to speak 238 words and 166 TIMIT sentences [9]. The subject labeled as ”F05” was chosen as the animation character. In order to maximize the quality visual speech synthesis, we used leave-one-out method for training the concatenative visual speech model, i.e. 154 videos is used for training the visual speech model for creating an animation test sample and the process is repeated for all 155 videos. Thus, the resulting dataset contains in total 155 real and 155 animated videos. The resolution of output videos is 450 × 300 and frame rate 30 fps. Fig. 4 shows some snapshots of both real and corresponding animated talking faces.

The experimental results on this dataset are shown in Table 4. Also in this case, the findings are consistent with the previous experiments as the lip-sync based liveness detection methods cannot tell a difference between the facial animation and original visual speech sequences but performs well in detecting audiovisual replay attacks.

4.4. Subject-specific tuning

Intuitively, the source actor or model directing the facial reenactment or voice conversion (or synthesis) process is unlikely to be able to mimic the speaking style of the targeted person. Therefore, it was suggested in [3] that use of client-specific synchrony models could possibly still show strong robustness to high-effort impostor attacks such as voice conversion and facial animation.

Next, we perform preliminary experiments and try to find out if adaptation of subject-specific visual and audible speech dynamics improve the robustness of lip-sync based liveness detection. Since the data provided by the authors of [4] does not provide enough data for both training and testing the synchrony models, we consider only the MMM (see, Section 4.3.1) and concatenative visual speech synthesis datasets (see, Section 4.3.3). We follow the experimental protocol used in Section 4.2 and divide the datasets into two halves, which are used for training and testing in turns. For the sake of simplicity, we chose only the best-performing synchrony assessment methods for this experiment.

The MMM dataset is too small for training the high-dimensional STACOG features, thus the combination of DCT+MFCC-Δ-Δ^2 was chosen on this corpus. Table 5 depicts that the subject-specific lip-sync detection method is able to achieve perfect performance under both attack types of the MMM dataset. This can be explained with two factors: 1) the videos in MMM dataset are quite long and the utterances vary between videos when the asynchrony in the replay attacks is obvious (for human observer), and 2) even though the synthesized visual speech is synchronized with the original audio content, the visual speech does not

	3 digits		4 digits		5 digits	
Method	Replay	Animation	Replay	Animation	Replay	Animation
DCT+MFCC	**23.00**	**38.33**	**16.67**	**34.76**	**14.05**	**32.74**
DCT+MFCC-Δ	28.17	48.75	23.33	52.86	18.33	50.40
DCT+MFCC-Δ-Δ^2	29.17	55.42	20.48	55.24	16.27	54.76
STACOG+MFCC	19.17	32.50	16.19	36.43	13.10	38.10
STACOG+MFCC-Δ	**11.17**	**20.00**	**8.48**	**16.43**	**5.32**	**13.10**
STACOG+MFCC-Δ-Δ^2	18.33	35.00	13.81	39.52	8.10	35.32

Table 3: Cross-database performance of the generic lip-sync detection methods in terms of EER (%) under the replay and well-articulated facial animation based attacks of the T-MMM dataset.

Method	Replay	Animation
DCT+MFCC	**14.15**	33.33
DCT+MFCC-Δ	17.13	35.95
DCT+MFCC-Δ-Δ^2	17.65	**32.03**
STACOG+MFCC	12.42	**28.76**
STACOG+MFCC-Δ	12.57	31.37
STACOG+MFCC-Δ-Δ^2	**11.39**	32.03

Table 4: Cross-database performance of the generic lip-sync detection methods in terms of EER (%) under the replay and facial animation based attacks of the concatenative visual speech synthesis dataset.

Method	Replay	Animation
Generic model	20.00	30.00
Subject-specific model	**0.00**	**0.00**

Table 5: Performance comparison between generic and subject-specific lip-sync models in terms of EER (%) under the two different attack scenarios of the MMM dataset.

resemble the speaking style of the targeted person due to the underarticulation issue (see, Section 4.3.1). Even if the MMM based model was trained to improve the articulation, the subtle dynamics of a talking mouth may be lost after smoothing the synthesized trajectories [25].

The combination of STACOG+MFCC-Δ-Δ^2 was selected for the concatenated visual speech synthesis dataset. From the results presented in Table 6, we can see that this dataset is more challenging but still the subject-specific lip-sync model is able to separate both audiovisual replay and facial animation based attacks quite well from real videos. The concatenated visual speech synthesis preserves the true dynamics of the original speaker much better than MMM. However, the synthesized videos suffer occasionally of abrupt changes of facial textures or jerky motions between two frames that contain phonemic transitions. Even though the method tries to find positions in the original videos where a transition can be made to other positions without introducing these noticeable discontinuities, the training speech corpus is unlikely to contain smooth transitions between all possible phoneme combinations. Furthermore, variations in head pose might introduce similar effects even if pre-processing (pose normalization) is applied. The animation artifacts explain why the client-specific synchrony model is able to separate animated visual speech based attacks from the original videos quite well.

The results on both datasets are very promising as the subject-specific synchrony models are able to perform robustly on both audiovisual attacks. These preliminary findings suggest that adaptation of person-specific audiovisual speech dynamics might be indeed one possible approach to tackle these sophisticated high-effort attacks, thus confirming the intuition proposed in [3].

Method	Replay	Animation
Generic model	11.39	32.03
Subject-specific model	**2.61**	**5.89**

Table 6: Performance comparison between generic and subject-specific lip-sync models in terms of EER (%) under the two different attack scenarios of the concatenated visual speech synthesis dataset.

5. Conclusion

We investigated the performance of the text-independent state-of-the-art lip-sync based liveness detection techniques under presentation attacks consisting of audio recordings of the targeted person and corresponding photo-realistic animated visual speech. Our experimental analysis with three different image-based visual speech synthesis techniques reveals that generic synchrony measures can be fooled even with underarticulated lip movements. Therefore, measuring audio-video synchrony or content alone is not enough for securing audiovisual biometric systems.

Our preliminary investigations suggest though that adaptation of person-specific audiovisual speech dynamics is one possible approach to tackle these high-effort forgeries because speech dynamics of a targeted person are hard to synthesize. Since time-consuming data collection during enrollment phase is undesirable property for biometric systems, the visual speech model could be gradually tuned with data captured during successful verification attempts.

In future, more representative audiovisual databases are needed for conducting comprehensive follow-up studies on the use of person-specific lip dynamics for recognition and presentation attack detection purposes and developing methods explicitly describing the facial animation related artifacts, e.g. image/video quality defects and unnatural motion. The datasets should consider logical or presentation attacks combining both synthetic voice and visual speech.

Acknowledgments

The financial support of Academy of Finland is gratefully acknowledged.

References

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

[2] E. Boutellaa, Z. Boulkenafet, J. Komulainen, and A. Hadid. Audiovisual synchrony assessment for replay attack detection in talking face biometrics. *Multimedia Tools and Applications*, pages 1–15, 2015.

[3] H. Bredin and G. Chollet. Making talking-face authentication robust to deliberate imposture. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1693–1696, March 2008.

[4] Y.-J. Chang and T. Ezzat. Transferable videorealistic speech animation. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '05, pages 143–151. ACM, 2005.

[5] S. Doledec and D. Chessel. Co-inertia analysis: an alternative method for studying species-environment relationships. *Freshwater Biology*, 31(3):277–294, 1994.

[6] N. Evans, F. Alegre, Z. Wu, and T. Kinnunen. *Encyclopedia of Biometrics*, chapter Anti-spoofing: Voice Conversion, pages 1–10. Springer, 2014.

[7] N. Eveno and L. Besacier. Co-inertia analysis for "liveness" test in audio-visual biometrics. In *International Symposium on Image and Signal Processing and Analysis, (ISPA)*, pages 257–261, Sept 2005.

[8] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '02, pages 388–398. ACM, 2002.

[9] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. Darpa timit acoustic phonetic continuous speech corpus, 1993.

[10] W. Karam, H. Bredin, H. Greige, G. Chollet, and C. Mokbel. Talking-face identity verification, audiovisual forgery, and robustness issues. *EURASIP Journal on Advances in Signal Processing*, 4, 2009.

[11] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, 2014.

[12] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[13] T. Kobayashi and N. Otsu. Motion recognition using local auto-correlation of space-time gradients. *Pattern Recognion Letters*, 33(9):1188–1195, July 2012.

[14] K. Kollreider, H. Fronthaler, M. I. Faraj, and J. Bigun. Real-time face detection and motion analysis with application in "liveness" assessment. *IEEE Transactions on Information Forensics and Security*, 2(3):548–558, 2007.

[15] S. Marcel, M. S. Nixon, and S. Z. Li. *Handbook of Biometric Anti-Spoofing: Trusted Biometrics Under Spoofing Attacks*. Springer, 2014.

[16] W. Mattheyses and W. Verhelst. Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Communication*, 66:182 – 217, 2015.

[17] A. Melnikov, R. Akhunzyanov, O. Kudashev, and E. Luckyanets. *International Conference on Image Analysis and Processing (ICIAP)*, chapter Audiovisual Liveness Detection, pages 643–652. Springer, 2015.

[18] K. Messer, J. Matas, J. Kittler, and K. Jonsson. Xm2vtsdb: The extended m2vts database. In *Second International Conference on Audio and Video-based Biometric Person Authentication*, pages 72–77, 1999.

[19] C. Richie, S. Warburton, and M. Carter. Audiovisual database of spoken american english, 2009.

[20] M. Slaney and M. Covell. Facesync: A linear operator for measuring synchronization of video facial images and audio tracks. In *Neural Information Processing Systems Conference*, pages 814–820, 2000.

[21] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2016.

[22] F. Verdet and J. Hennebert. Impostures of Talking Face Systems Using Automatic Face Animation. In *IEEE Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2008.

[23] G. Zhao, M. Pietikäinen, and A. Hadid. Local spatiotemporal descriptors for visual recognition of spoken phrases. In *Proceedings of the International Workshop on Human-centered Multimedia (HCM)*, pages 57–66. ACM, 2007.

[24] Z. Zhou, G. Zhao, Y. Guo, and M. Pietikäinen. An image-based visual speech animation system. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(10):1420–1432, 2012.

[25] Z. Zhou, G. Zhao, and M. Pietikäinen. Synthesizing a talking mouth. In *Proc. Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, pages 216–223. ACM, 2010.