SPOOFING DETECTION VIA SIMULTANEOUS VERIFICATION OF AUDIO-VISUAL SYNCHRONICITY AND TRANSCRIPTION

Lea Schönherr, Steffen Zeiler, and Dorothea Kolossa

Institute of Communication Acoustics, Ruhr-Universität Bochum, Germany

lea.schoenherr@rub.de, steffen.zeiler@rub.de, dorothea.kolossa@rub.de

ABSTRACT

Acoustic speaker recognition systems are very vulnerable to spoofing attacks via replayed or synthesized utterances. One possible countermeasure is audio-visual speaker recognition. Nevertheless, the addition of the visual stream alone does not prevent spoofing attacks completely and only provides further information to assess the authenticity of the utterance. Many systems consider audio and video modalities independently and can easily be spoofed by imitating only a single modality or by a bimodal replay attack with a victim's photograph or video.

Therefore, we propose the simultaneous verification of the data synchronicity and the transcription in a challenge-response setup. We use coupled hidden Markov models (CHMMs) for a text-dependent spoofing detection and introduce new features that provide information about the transcriptions of the utterance and the synchronicity of both streams. We evaluate the features for various spoofing scenarios and show that the combination of the features leads to a more robust recognition, also in comparison to the baseline method. Additionally, by evaluating the data on unseen speakers, we show the spoofing detection to be applicable in speaker-independent use-cases.

Index Terms— spoofing detection, liveness detection, audio-visual speaker recognition, multimodal biometrics, coupled hidden Markov models

1. INTRODUCTION

In a spoofing attack against biometrics, a malicious party tries to imitate another person's biometrics. One approach to increase the robustness against spoofing attacks is to use multimodal biometrics, e. g., audio-visual speaker verification [1]. However, if no countermeasure is implemented, multimodal systems have shown to be vulnerable against spoofing attacks imitating only one trait successfully, since they often focus on the trait with the least distortions [2, 3].

In general, different kinds of spoofing attacks against audio-visual authentication need to be considered. This encompasses replay attacks, where an impostor uses a previously recorded utterance of the victim, e. g., a video of the entire identification process or a recording of the audio channel and an additional visual input, like an arbitrary image or video of the victim.

Playing back a synthesized version of the utterance constitutes another spoofing attack. Such synthesized information has the advantage that a response to a challenge can be imitated as well [4, 5]. However, in contrast to recordings, synthesized videos are more difficult to access and the movements of the lips often appear artificial [6].

To prevent such attacks, the task is to distinguish between a spoofing attack and a genuine speaker such that a sophisticated attack can be detected, but a genuine speaker is not rejected.

For speaker recognition using audio data only, many different approaches for spoofing and liveness detection exist. To counter replay attacks, text-dependent recognition and different phrases for each identification process can be used [7]. For the classification of synthesized utterances, the constant Q transform (CQT) has been shown to achieve robust results for an audio-only recognition [8].

Audio-visual speaker recognition [9, 10] provides much more information to verify a response, but only a few recent works have investigated audio-visual spoofing detection. All of these works use synchronicity measures either for a single utterance [11, 12] or multiple utterances [13, 14, 15]. In [11] the difference to stored sample utterances is calculated for the audio and the video channel separately, using dynamic time warping (DTW). The resulting time differences of both modalities are then compared to verify the utterance. Further approaches use canonical correlation analysis (CCA) to maximize the cross-correlation between matching audio and video frames [12, 13, 15]. Similar to these works, in [14] a co-inertia analysis (CoIA) is applied to the audio and video data to calculate features for a correlation.

All these approaches are still vulnerable to video replay attacks since they only measure the synchronicity. Particularly the approaches in [11] and [12] may not be able to distinguish between a recorded video and a genuine utterance, since the challenge does not change.

While hidden Markov models (HMMs) have been used for audio-visual speaker verification [16, 17, 18], the authors do not verify their method for spoofing attacks or again only detect the synchronicity but do not consider the transcription. Especially, in [16] where the authors also use CHMMs, only an asynchrony detection is applied by considering major signal changes (e. g., starts or ends of words). Hence, a video replay attack is impossible to detect with this approach.

In contrast to these works, we use coupled hidden Markov models (CHMMs) to simultaneously verify the audio-visual synchronicity and transcription in a challenge-response setup. CHMMs have proven successful in audio-visual speech recognition, increasing the robustness of speech recognition in adverse conditions [19, 20]. For audio-visual speech recognition, CHMMs are more appropriate than HMMs with early feature fusion, as they allow slight asynchronicities between feature streams, which gives them a significant advantage regarding the recognition performance.

We will use and expand their capability to handle and detect asynchronicity in the following, which will allow us to employ them for simultaneous verification of audio-visual synchronicity and spoken content of the utterance. The proposed spoofing can be deployed in a deep-learning-based approach for the speech recognition in an equivalent manner. However, this work focuses on spoofing detection and due to the limited data available here, a GMM/HMM-based CHMM system is a good starting point that already allows us to explore the applicability of different feature sets for verification purposes.

The paper is organized as follows: After a brief introduction of CHMMs in Section 2, we explain how the CHMMs can be constructed for a spoofing detection task with changing utterances. In Section 3 the calculation of different synchronicity and transcription features via CHMMs is described. The results for different spoofing scenarios, a comparison with a baseline method, and a cross-speaker verification are presented in Section 4 before concluding in Section 5.

2. SPOOFING DETECTION

In order to recognize a synchronization mismatch between the audio and the video data, we use CHMMs, so that we can simultaneously recognize both the audio and the video transcription, and any time difference between the audio and the video stream.

2.1. Coupled HMMs

CHMMs are an extension of HMMs that is particularly useful for combining different streams in a multimodal system without the necessity of fusion on the feature level. In this work, CHMMs are used to verify whether the bi-modal data is genuine. For the construction of the CHMMs, it is necessary initially to represent each single word as a uni-modal HMM. In general, for audio-visual speech recognition with CHMMs, the two marginal HMMs, one for each stream, are trained separately for each word. During the training of the HMMs, the conditional observation likelihoods

$$b(i,t) = P(\boldsymbol{o}(t)|q(t) = i), \tag{1}$$

for state $q(t) = i$ are calculated based on the observations $\boldsymbol{o}(t)$ of the stream for each time frame t. Additionally, the state transition probabilities $a(i,j)$ are obtained during training. The probability of going from state $q_t = i$ to state $q_{t+1} = j$ in a discrete time step $t \to t+1$ is

$$a(i,j) = P(q_{t+1} = j|q_t = i). \tag{2}$$

As in a speech recognition application, the state transitions are defined such that the model can not step back into a previous state:

$$a(i,j) = 0, \quad \forall i > j. \tag{3}$$

For a CHMM, all states $Q^A = \{q_1^A, \ldots, q_{N^A}^A\}$ of the audio HMM are combined with all states $Q^V = \{q_1^V, \ldots, q_{N^V}^V\}$ of the visual HMM such that the resulting CHMM has $N = N^A \cdot N^V$ states. The new conditional observation likelihoods $b(\boldsymbol{i}, t)$ for each coupled state are combinations of the corresponding conditional observation likelihoods of the audio A and the video V stream

$$\begin{aligned} b(\boldsymbol{i},t) &= b^A(i^A,t) \cdot b^V(i^V,t) \\ &= P(\boldsymbol{o}^A(t)|q^A(t) = i^A)P(\boldsymbol{o}^V(t)|q^V(t) = i^V), \end{aligned} \tag{4}$$

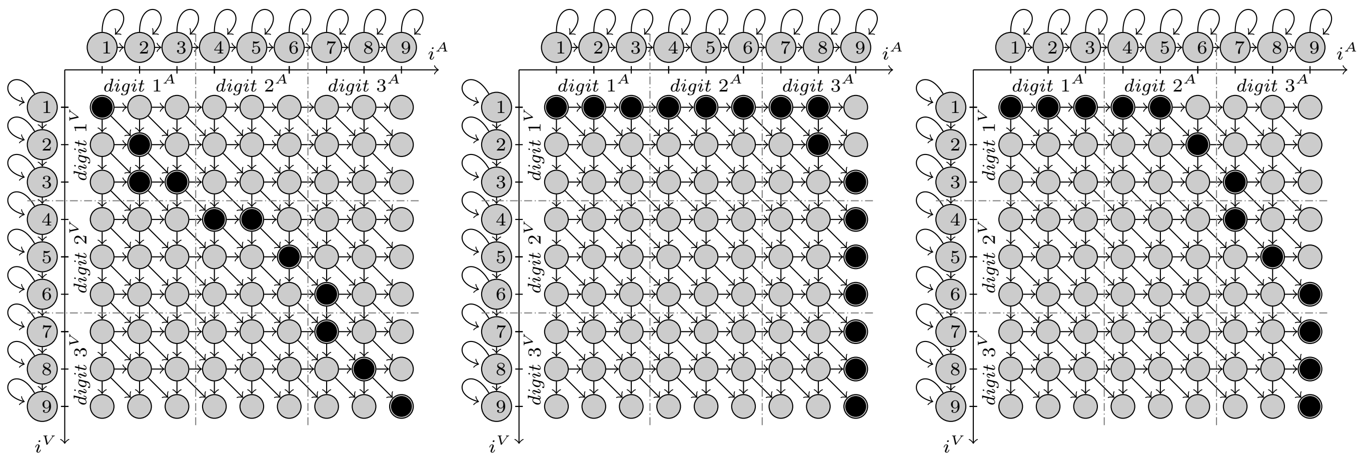
with $\boldsymbol{i} = [i^A, i^V]$ describing the coupled state as a combination of the single-modality states i^A and i^V. The feature vectors $\boldsymbol{o}^A(t)$ and $\boldsymbol{o}^V(t)$ are obtained from the audio and the video stream, respectively. The state transitions for the CHMM are calculated by:

$$a(\boldsymbol{i},\boldsymbol{j}) = a(i^A,j^A) \cdot a(i^V,j^V), \tag{5}$$

with the coupled states $\boldsymbol{i} = [i^A, i^V]$ and $\boldsymbol{j} = [j^A, j^V]$.

2.2. CHMMs for Spoofing Detection

For audio-visual speech recognition, the corresponding audio HMM and video HMM are, like in [16] and [19], used as marginal HMMs creating a combined word CHMM as a Cartesian product model, cf. Equations (4) and (5). The single word models may then be combined according to a task grammar. With this approach, the audio and the visual streams can be asynchronous within one word, but not across different words. This is sufficient for audio-visual speech recognition since the audio and the visual stream can usually be assumed to be synchronous.



(a) Synchronous streams

(b) Invalid visual stream

(c) Asynchronous streams

Fig. 1: The CHMMs for different scenarios and possible paths through the CHMM with $M = 3$ digits as the challenge: Figure 1a shows a synchronous utterance, Figure 1b is an example of an utterance with an invalid visual stream, and Figure 1c shows an utterance with asynchronous streams.

2.2.1. CHMM Construction

In the case of a spoofing attack, where either the two streams do not match or one stream is missing completely, the synchronicity may not be given. For a spoofing detection, this asynchrony can be assessed. Thus, in the following, grammar models are built on the level of the single HMMs and combined to one CHMM for the entire utterance. With this approach, the audio and the visual streams may be in completely different words at the same time step.

For our experiments, a sequence of random digits is used as the utterance. With ten different digits ('zero' to 'nine'), 10^M combinations of different utterances are possible, where M is the number of digits in the sequence.

In Figure 1 different spoofing scenarios and their possible paths through a CHMM are sketched for an utterance with $M = 3$ digits. For easier visualization, the CHMMs in the figure are simplified. In general, the audio HMM requires more states than the video HMM, but here, they are depicted for $N^A = N^V$, and we show only three digits. The CHMM used for the spoofing detection has many more states and thus possible paths. Further, the different digits do not necessarily have the same number of states and not all possible state transitions are sketched. In general, transitions are only possible top-to-bottom and left-to-right, according to Equations (3) and (5).

In Figure 1a, a possible path for a synchronous utterance is shown. Although the streams are synchronous, the recognized digits still have to be compared to the challenge. Figure 1b depicts a spoofing scenario with an invalid visual stream (e. g., a still image of the victim). In this example, the visual recognition stays in the first visual state, while the audio recognition proceeds. Due to the structure of the CHMM, for the video, any arbitrary transcription will be recognized as well. Figure 1c represents a spoofing scenario where the two streams are not synchronous. In the latter examples, an analysis of the coupled state sequence provides useful information about the synchronicity.

2.2.2. Optimization of Resource Use

Due to the combinatorial nature of our CHMM construction scheme and the resulting high number of coupled states, the computations would get infeasible, if we were to evaluate all $(10 \cdot M)^2$ possible combinations of digits in one compound CHMM. Therefore, we limit the construction of the marginal HMMs for asynchrony detection to only the most likely digits at each of the M positions. To obtain these digits, the K best digits are determined for each position for the audio and the video stream. These resulting $2K$ digits per position are considered to construct the two marginal HMMs. Since some of the $2K$ digits of each position will often be recognized by both, the audio and the video model, among K best digits, the redundant digits are discarded for the construction of the marginal HMMs. The resulting CHMM has at most $(2K \cdot M)^2$ combinations of digits. This reduces the computational cost significantly.

3. PROPOSED FEATURES

For the recognition, the forward-backward algorithm is used to obtain the matrix Γ with $N \times T$ values, describing the probabilities for being in all coupled states at time $t = 1, \ldots, T$. With the Viterbi algorithm and Γ, the most likely path through the CHMM is calculated. The resulting path is a sequence of coupled states

$$\boldsymbol{q} = \Big[q(1) = [i^A(1), i^V(1)], \dots, \\ q(T) = [i^A(T), i^V(T)]\Big], \tag{6}$$

describing the recognized coupled states in the order of recognition.

3.1. Synchronicity Features

The audio HMMs are defined with three states per phoneme, whereas the video HMMs use only one state per phoneme. Thus, the time alignment difference between the audio and the video stream is calculated via:

$$\lambda(t) = \left\lceil \frac{i^A(t)}{3} \right\rceil - i^V(t). \tag{7}$$

In the case of a genuine utterance, the values of $|\lambda|$ should be small. In contrast, a spoofed utterance with non-matching streams will typically show larger values. As features for the recognition, two different values have shown to be useful, the entropy E of the time alignment difference λ and the mean value of $\lambda(t)$, denoted by Λ, over all time steps $t = 1, \dots, T$.

However, using only these features, the audio and the video stream may appear synchronous, even if the recognized digits of the streams are different, especially, if the two different digits have the same number of states. As a proposed countermeasure, an additional CHMM is constructed containing only the challenged digits. With this CHMM, the distances $\lambda_\kappa(t)$ are calculated according to Equation (7). In a genuine scenario, both distance vectors λ and λ_κ should be very similar. Therefore, we propose two more features:

$$\Lambda_\kappa = \frac{1}{T} \sum_{t=1}^{T} \lambda(t) - \lambda_\kappa(t), \tag{8}$$

$$\Lambda_{|\kappa|} = \frac{1}{T} \sum_{t=1}^{T} |\lambda(t) - \lambda_\kappa(t)|. \tag{9}$$

Although the measures are similar, the combination of Λ_κ and $\Lambda_{|\kappa|}$ leads to a more robust recognition.

3.2. Transcription Features

As additional features for the spoofing detection, the transcriptions of both streams, obtained with the CHMM-based recognition, are used. This is necessary to prevent replay attacks with videos, where the streams may be synchronous, but do not contain the utterance of the challenge.

To detect this situation, the differences of the audio transcription $\tau_A(m)$ and video transcription $\tau_V(m)$ to the challenged digits $\tau(m)$ over all positions $m = 1, .., M$ are calculated with the Hamming distance, such that each substituted digit increases the calculated distance by 1. Thus, the resulting distances may be between 0 and M and are considered as features Θ_A and Θ_V for the distance of the audio transcription and the video transcription, respectively.

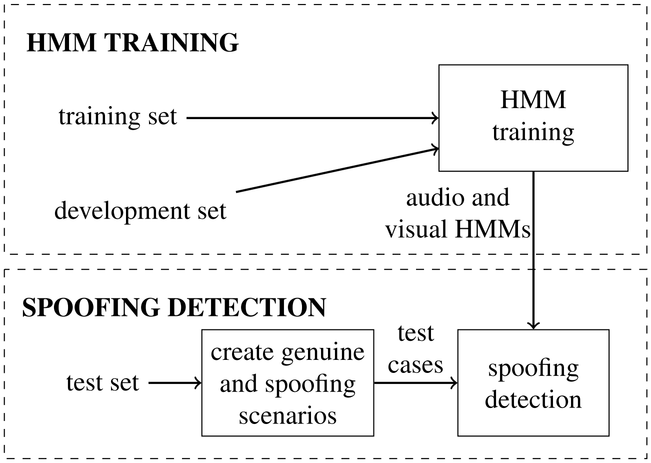


Fig. 2: The training set is used for the training of the audio and the visual HMMs of the single digits. The development set is considered to verify the performance of the speech recognition system. With the test set, the genuine, and spoofing scenarios are built for the evaluation of the spoofing detection, which uses the trained HMMs. Utterances from all 30 speakers are used for the training, development, and test set.

4. EXPERIMENTAL RESULTS

For the experimental evaluation, we have used a set of 30 speakers (15 females and 15 males) with utterances of single digits from 0–9. From each speaker, 270 utterances are used. The dataset is recorded with Microsoft's Kinect sensor that provides also reliable information about the location of the mouth region. In the following experiments, we have used the first of the four available microphone channels. The data set deployed for the experiments is accessible online. [1]

As audio features, the first 13 mel-frequency cepstral coefficients (MFCCs) and their first and second derivatives have been considered. As video features, we have used the first 8×8 coefficients of a two-dimensional discrete cosine transformation (DCT) of the cropped mouth region.

In Figure 2 the training and spoofing detection is sketched. 170 recordings for each of the 30 speakers have been used in HMM training to get a speaker-independent audio-visual speech model. Further 40 recordings per speaker have been used as the development set to verify the speech recognition performance. The remaining 60 recordings per speaker have been deployed to create the genuine and spoofing scenarios together with the scenarios and the trained HMMs, the spoofing detection is tested.

[1] doi.org/10.5281/zenodo.823531

$M = 3$ digits were concatenated per utterance to build 20 utterances per speaker for each spoofing scenario and the genuine utterances. We have used $K = 3$ for all experiments, such that 3–6 digits per position are considered to construct the marginal HMMs. This is a trade-off between the complexity of the resulting CHMM and the probability of obtaining the uttered digit in the CHMM.

The spoofing detection has been applied in a challenge-response setup. Thus, a recorded video of the victim with the correct utterance is hard to access, since the utterance changes for each verification process. However, it is still possible to use recorded videos with different utterances or to use a modified or synthesized video. Such artificial constructed videos may show a delay which can be detected, even if the utterance is correct.

To create test scenarios, different combinations of the audio and video stream have been created. In all cases the audio and the video stream is from the same speaker for one scenario:

- Scenario #1 (still image): The audio stream is the genuine utterance corresponding to the challenge and the video stream is only one image for the entire utterance.
- Scenario #2 (cross-video): The audio stream is the genuine utterance corresponding to the challenge and the video stream is replaced by an arbitrary other one.
- Scenario #3 (wrong utterance): The audio stream and the video stream do match, but do not correspond to the digits of the challenge.
- Scenario #4 (delayed): The audio and the video stream correspond with the challenge, but they have a delay (±1 s, ±0.5 s, and ±0.25 s).

4.1. Baseline System Description

For comparison, we also have implemented one of the latest spoofing detection approaches for speaker verification [13]. This approach can also be used in a challenge-response setup. Additionally, it also does not need specifically enrolled utterances for the spoofing detection. However, in contrast to our method, the baseline method is speaker-dependent. Thus, for training it needs utterances from each enrolled speaker.

For the spoofing detection, the baseline system uses CCA to compare the audio and the video stream [21]. For this purpose, the projection matrices W and Z (canonic correlation matrices) are calculated with the training set for each speaker separately. The score for the spoofing detection is calculated by

$$S(X, Y) = \frac{1}{N} \sum_{n=1}^{N} corr(Xw_n, Yz_n), \qquad (10)$$

where w_n and z_n are the n^{th} column of the projection matrices W and Z, respectively, and X and Y are the audio and the video stream of the test scenario, respectively. The parameter N is tuned on the development set. Hence, only the synchronicity, but not the transcription is verified. Therefore, it is not possible to detect scenario #3 with synchronous video, but the wrong utterance, so this scenario has not been considered in our evaluation of the baseline method.

	SF	TF	all	baseline
Scenario #1	3.25	13.75	**1.50**	3.85
Scenario #2	10.50	12.25	**6.25**	21.49
Scenario #3	10.75	**1.75**	2.00	
Scenario #1–2	5.68	14.45	**5.18**	16.42
Scenario #1–3	8.83	12.50	**5.50**	
Scenario #4 (±1 s)	**2.34**	43.28	2.71	14.52
Scenario #4 (±0.5 s)	3.10	52.27	**2.86**	14.50
Scenario #4 (±0.25 s)	**8.88**	51.25	9.40	10.85
Scenario #4	6.96	54.88	**6.46**	13.45
Scenario #1,2,4	**5.89**	55.06	7.11	14.13
Scenario #1–4	8.69	56.72	**6.83**	

Table 1: EER (in %) of different features for the spoofing scenarios and their combinations. The input features are synchronicity features ($SF = [E, \Lambda, \Lambda_\kappa, \Lambda_{|\kappa|}]$), transcription features ($TF = [\Theta_A, \Theta_V]$), and both together (all). As the baseline, the approach in [13] has been used.

As features for the baseline method, the same ones as in [13] have been used in the evaluation since these provided the best results. These are MFCCs for the audio data and space-time auto-correlation of gradients (STACOG) for the visual data [22].

4.2. Results

To build a spoofing detection with the different proposed synchronicity and transcription features, support vector machines (SVMs) have been employed in the following. For this purpose, we have evaluated the synchronicity features $SF = [E, \Lambda, \Lambda_\kappa, \Lambda_{|\kappa|}]$ and the transcription features $TF = [\Theta_A, \Theta_V]$ separately, and all of these features together. Table 1 provides an overview of the equal error rate (EER) for the different spoofing scenarios and combinations of those. Scenario #4 is separated into different groups, considering different delays. Scenario #4 is a combination of all three delays.

In some cases, the different features lead to very different EERs. In Table 1, the best results are marked in bold. For most spoofing scenarios, a combination of all features leads to the best EER, and especially if the different spoofing scenarios are averaged, a combination of all features clearly provides the best results. In general, the difference of the EER of the single feature groups and the combination never exceeds 1.22 % and in many cases, the combination is better by a large margin.

	S1–S5	S6–S10	S11–S15	S16–S20	S21–S25	S26–S30	average
Scenario #1	1.00	4.50	2.00	2.00	2.50	3.00	2.50
Scenario #2	8.00	10.00	10.00	7.00	8.00	7.50	8.42
Scenario #3	1.00	3.00	2.00	1.00	4.50	3.00	2.42
Scenario #4	4.50	11.17	9.00	7.42	5.50	12.25	8.31
Scenario #1–4	3.56	14.44	12.67	7.78	6.78	10.17	9.23

Table 2: The EER (in %) for the cross-speaker verification. The speakers in the first row are left out and used to evaluate the spoofing detection.

Scenario #2 benefits the most from the combination of both features, since the video-only speech recognition is not as reliable as the audio-only recognition, such that the mismatch of the visual stream is not always detected, and a genuine, matching visual stream can sometimes be falsely classified as spoofed. Additionally, the synchronicity measure is not as robust here as for scenario #1 were only one image is used for the whole visual stream. Since, overall, both features perform about equally well in this scenario, large improvements are possible due to their complementary information. For scenario #4 the synchronicity features are more valuable than the transcription features. This is no surprise, due to the capability of CHMMs to achieve a reliable recognition for asynchronous data, which clearly distinguishes them from early-integration-based approaches for this task. Therefore, the introduced distance features provide a robust measure of classification.

4.2.1. Comparison with Baseline System

In all cases, the combination of the synchronicity features and the transcription features leads to a lower EER in comparison to the results of the baseline system. Especially for scenario #2, much better results can be achieved with the proposed approach. This indicates that the CCA-based classification focuses on the assignment of major signal changes. Thus, a wrong transcription in one stream is more difficult to detect.

Interestingly, the EERs for the different delays in scenario #4 are similar. This may also be the result of the synchronicity-based spoofing detection which relies only on the assignment of silent/non-silent and non-movement/movement parts.

4.2.2. Cross-Speaker Verification

For many use-cases of spoofing detection it is not feasible to collect enough data from each enrolled speaker to train the spoofing detection. Therefore, a 6-fold cross-verification has been performed by leaving out a group of speakers during training of the spoofing classification. The utterances of this held-out group of speakers are used for the evaluation in Table 2. As input features for all cases, all introduced synchronicity and transcription features $[E, \Lambda, \Lambda_{\kappa}, \Lambda_{|\kappa|}, \Theta_A, \Theta_V]$ are considered.

The results show that the EER is similar for unseen speakers in most of the cases. It can also be observed that some speakers seem to be easier to spoof (group S6-S10), while for some speakers (group S1-S5), the EER is even lower than the corresponding results of Table 1 (all features). Thus, some speakers seem to be more vulnerable to spoofing attacks, although all results point to a good performance in general.

5. CONCLUSIONS

We have proposed a text-dependent audio-visual spoofing detection for speaker verification. For its evaluation, we have considered different spoofing scenarios, which can be used in a real attack. We have introduced a CHMM-based synchronicity measure, which is available for spoofing scenarios with non-matching streams. Additionally, the assessment of the transcription with the CHMM-setup also provides a simultaneous verification of both features groups, which can improve the classification in many cases where synchronicity metrics alone are not sufficient. Additionally, it is also possible to detect spoofing attacks that are synchronous but contain the wrong utterance. This shows the great advantage in contrast to approaches using synchronicity-based methods only.

Furthermore, via a cross-speaker validation, we have shown that the proposed spoofing detection can be used speaker-independently so that new speakers can be enrolled with no extra effort.

The introduced approach needs an additional step for the training of the CHMMs. However, a speaker verification with changing utterances is much harder to spoof, since an attacker either needs to produce all possible utterances or has only limited time to produce a spoofing attack. Therefore, this additional training step should often be justified, and it only needs to be performed once, before system deployment. In combination with an audio-visual speaker identification system, like in [9], the both systems can benefit from each other. Thus, a more secure and robust audio-visual speaker recognition can be achieved.

The proposed approach is not limited to digits and can be used for arbitrary words or sentences as long as speaker-independent HMMs for speaker recognition can be trained. For future work, a large-vocabulary version, based on triphone-level CHMMs including a deep learning approach will be investigated, to achieve a higher diversity of possible utterances for a still greater resilience against playback or synthesis.

6. REFERENCES

[1] Petar S. Aleksic and Aggelos K. Katsaggelos, "Audio-visual biometrics," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 2025–2044, 2006.

[2] Zahid Akhtar, "Security of multimodal biometric systems against spoof attacks. PhD thesis," *Department of Electrical and Electronic Engineering, University of Cagliari, Italy*, 2012.

[3] Thomas Kraft, "Analyzing state-of-the-art audio-visual authentication systems on the web. Bachelor thesis," *Department of Electrical Engineering and Information Technology, Ruhr-Universität Bochum, Germany*, 2017.

[4] Walid Karam, Hervé Bredin, Hanna Greige, Gérard Chollet, and Chafic Mokbel, "Talking-face identity verification, audiovisual forgery, and robustness issues," *EURASIP Journal on Advances in Signal Process*, pp. 4:1–4:15, 2009.

[5] Florian Verdet and Jean Hennebert, "Impostures of talking face systems using automatic face animation," in *IEEE Second International Conference on Biometrics: Theory, Applications and Systems*, 2008.

[6] Dietmar Schabus, Michael Pucher, and Gregor Hofer, "Joint audiovisual hidden semi-Markov model-based speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 336–347, 2014.

[7] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[8] Massimiliano Todisco, Héctor Delgado, and Nicholas Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Speaker Odyssey Workshop*, 2016, vol. 25, pp. 249–252.

[9] Lea Schönherr, Dennis Orth, Martin Heckmann, and Dorothea Kolossa, "Environmentally robust audio-visual speaker identification," in *IEEE Spoken Language Technology Workshop*, 2016, pp. 312–318.

[10] Mohammad Rafiqul Alam, Mohammed Bennamoun, Roberto Togneri, and Ferdous Sohel, "A deep neural network for audio-visual person recognition," in *IEEE International Conference on Biometrics Theory, Applications and Systems*, 2015.

[11] Amit Aides and Hagai Aronowitz, "Text-dependent audiovisual synchrony detection for spoofing detection in mobile person recognition," in *Interspeech*, 2016, pp. 2125–2129.

[12] Jukka Komulainen, Iryna Anina, Jukka Holappa, Elhocine Boutellaa, and Abdenour Hadid, "On the robustness of audiovisual liveness detection to visual speech animation," in *IEEE International Conference on Biometrics Theory, Applications and Systems*, 2016.

[13] Elhocine Boutellaa, Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid, "Audiovisual synchrony assessment for replay attack detection in talking face biometrics," *Multimedia Tools and Applications*, vol. 75, no. 9, pp. 5329–5343, 2016.

[14] Hervé Bredin and Gérard Chollet, "Making talking-face authentication robust to deliberate imposture," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 1693–1696.

[15] Hervé Bredin and Gérard Chollet, "Audiovisual speech synchrony measure: Application to biometrics," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 179–184, 2007.

[16] Enrique Argones Rúa, Hervé Bredin, Carmen García Mateo, Gérard Chollet, and Daniel González Jiménez, "Audio-visual speech asynchrony detection using co-inertia analysis and coupled hidden Markov models," *Pattern Analysis and Applications*, vol. 12, no. 3, pp. 271–284, 2009.

[17] David Dean, Sridha Sridharan, and Tim Wark, "Audio-visual speaker verification using continuous fused HMMs," in *HCSNet Workshop on the Use of Vision in HCI*, 2006, vol. 56, pp. 87–92.

[18] Tieyan Fu, Xiao Xing Liu, Lu Hong Liang, Xiaobo Pi, and Ara V. Nefian, "Audio-visual speaker identification using coupled hidden Markov models," in *International Conference on Image Processing*, 2003, vol. 3, pp. III–29–32.

[19] Ara V. Nefian, Luhong Liang, Xiaobo Pi, Liu Xiaoxiang, Crusoe Mao, and Kevin Murphy, "A coupled HMM for audio-visual speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, vol. 2, pp. II–2013–II–2016.

[20] Ahmed Hussen Abdelaziz, Steffen Zeiler, and Dorothea Kolossa, "Learning dynamic stream weights for coupled-HMM-based audio-visual speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 5, pp. 863–876, 2015.

[21] David Hardoon, Sandor Szedmak, and John Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[22] Takumi Kobayashi and Nobuyuki Otsu, “Motion recognition using local auto-correlation of space–time gradients,” *Pattern Recognition Letters*, vol. 33, no. 9, pp. 1188–1195, 2012.