Replay Spoof Attack Detection using Deep Neural Networks for Classification

Salahaldeen Duraibi
Department of Computer Science
University of Idaho,
Moscow, USA
Department of Computer Science
University of Jazan,
Jazan, KSA
Email: dura6540@vandals.uidaho.edu

Wasim Alhamdani
Department of Computer and Information Sciences
University of the Cumberlands
Williamsburg, USA
Email:
wasim.alhamdani@ucumberlands.edu

Frederick T. Sheldon
Department of Computer Science
University of Idaho
Moscow, USA
Email: sheldon@ieee.org

***Abstract*— In this paper, we explore the use of the deep learning approach for replay spoof detection in speaker verification systems. Automatic speaker verifications (ASVs) can be easily spoofed by previously recorded genuine speech. In order to counter the issues of spoofing, detecting spoofing attacks play an important role. Hence, we consider the detection of replay attack spoofing that is the most easily accomplished spoofing attack. In this light, we propose a deep neural network-based (DNN) classifier using a hybrid feature from Mel-frequency cepstral coefficient (MFCC) and constant Q cepstral coefficient (CQCC). Several experiments were conducted on the latest version of ASVspoof 2017 dataset. The results are compared with a base line system that uses the Gaussian mixture model (GMM) classifier with different features that include MFCC, CQCC, and the hybrid feature of the two. The experiment results reveal that the DNN classifier outperforms the conventional GMM classifier. It was found that the hybrid-based features are superior to single features, such as CQCC and MFCC in terms of equal error rate (ERR). In addition, like many previous researchers have found, it turned out that high-frequency regions of speech utterance convey much more discriminative information for replay attack detection.**

Keywords— DNN, classifier, deep learning, replay spoof attack, biometric, speaker verification, Security

I. INTRODUCTION

The use of biometrics for human authentication has become an important part of securing applications [1]. Although, in the last decade, the world has experienced rapid adoption and increased effectiveness of biometrics-based authentication systems, the success of such biometric systems is not only measured in its effectiveness, but also the availability and affordability of the technology required to measure the human factors it uses [2, 3]. Voice is one of the most commonly used biometrics in the market nowadays, because it has the advantage that it can be easily captured with readily available personal devices [4]. Hence, speaker verification systems (ASVs) have gained the attention of both researchers and industry [5]. Nevertheless, ASVs are vulnerable to spoofing attacks where an impersonator claims the identity of someone that he is not [6]. Known spooking attacks associated with ASVs include mimicry attacks, conversion attacks, and replay attacks. A mimicry attacks is when a professional mimics the utterance of the voice of a genuine user to gain unauthorized access [7]. Conversion attacks happens when voice transformation is performed on an imposter's utterance so that it sounds more like that of a genuine user [8]. Finally, replay attack is a spoof technique where pre-recorded speech is provided to speech verification systems, with the help of record-and-play devices, to impersonate a genuine user [6]. Among the three spoof attacks, replay attacks pose the most serious threat to ASVs [9]. The technique does not require expertise and is difficult to detect. A high quality smartphone is sufficient for someone to employ the technique effectively. As a consequence, there is a pressing demand for robust replay attack detection methods [9, 10].

There exist some traditional replay prevention attack techniques, such as the use of a text-prompted ASV systems [11, 12], and the use of additional biometric modalities [13, 14], for use as a means of replay attack prevention. In these systems, there is a tradeoff between system security and system complexity [15]. Therefore, replay attack detection approaches have attracted researchers [16, 17]. The replay attack detection methods are of two approaches. One approach researchers opted for adds auxiliary information to the utterance; for instance, researchers in [18] proposed the addition of a time stamp in the utterance of genuine user thereby validating the time stamp in the replay attack. The other approach focuses on detecting the distortion associated with the recording and replay devices that are used to accomplish the replay attack [9].

A number of researchers have worked on the detection of distortion as a countermeasure to replay attacks. For instance, in [19], to detect the replay attack the research found that the spectral ratio, low-frequency ratio, and modulation index carry relevant information to differentiate between the replay speech from the genuine speech. A matching algorithm that detects the cut and paste of the voice is employed and the detection is achieved by comparing the pitch and MFCC contours using dynamic time warping (DTW). Several replay detection solutions on text-independent ASV systems are proposed in [20]. Conventional ASV systems are employed in this research. However, in later studies it became apparent that such systems have issues of precision and performance when detecting replay attacks. For example, in [21], the research proves that with the use of replay attack detection GMM-UBM-based ASV systems the equal error rate (EER) increases from 2.61% to 15.03% for 459 genuine male and 1.82% to 34.09% for 220 genuine female speeches. Hence, it can be concluded that GMM-UBM-based systems are highly vulnerable to replay attacks [22]. In this light, domain research started employing deep learning techniques to overcome the problem [23]. Neural networks for ASV systems provid more natural solutions than conventional algorithms.

978-1-7281-7624-6/20/$31.00 ©2020 IEEE
DOI 10.1109/CSCI51800.2020.00036

This inspired the successful application of deep learning frameworks in the domain ASV systems for replay attack detection. For example, some of the researchers have employed the conventional systems, such as GMM-UBM and i-vectors with DNN, for better performance and precision. These systems are usually employed in text-dependent SV systems [24, 25]. For instance, DNN is used for extracting frames from a speaker's speech and calculating its utterance-level information. Subsequently, the output of the DNN is converted to i-vectors, and at the backend, PLDA is used for a verification score [26, 27]. In such work, the DNN features are either used alone or combined with the conventional features (i.e., MFCC). Features extracted from DNN are employed in the posterior. These researchers use DNN to extract speaker-specific data and apply the similarity metrics check for the verification score [28]. However, there is always a room for improvement, and such domain researchers are still working obtaining systems with better precision and performance that use deep learning techniques. There is a number of studies using deep learning model for replay attack, and it still remains hot area of research. Different studies using deep learning models for replay attack detection are presented in the related work section.

In this paper, we propose a deep neural network-based classifier that uses hybrid features. Our contribution is twofold. The first is at the feature selection level, where we propose a hybrid feature that uses CQCC and MFCC to capture discriminative characteristics of both high frequency and low frequency regions of speech utterance, respectively. The second is that we propose a model level work that contributes a DNN-based classifier that is trained to discriminate between different conditions due to changes in playback, recording, and environmental natures.

The rest of the paper is as follows: Section II is the related work, Section III is the feature extraction for the replay attack detection, Section IV discusses the proposed DNN-classifier architecture, Section V presents the experiment and results, and Section VI concludes the paper.

II. Related Work

The conventional SV systems used GMM-UBM and i-vector as state-of-the art approaches of modeling for a long time. The modeling approaches rely on dimensional input features for extraction using MFCC. Nevertheless, MFCC has shown performance degradation [29]. As a result, the domain researchers have started investigating deep learning approaches for better feature extraction [28, 30]. Deep learning approaches have shown promising results with other verification systems using biometric trails, such as face and signature. For instance, deep neural network gained momentum for use in re-identification problems in face recognition [31] and the identification of forged a signature in signature verification systems [32]. Hence, using DNN, CNN, or RNN for feature extraction of biometric models has become a trend [33]. Particularly, DNN and CNN approaches have become more suitable for the extraction of speaker specific features to detect replay spoofing attacks [34].

Several replay attack detection approaches have been proposed following the ASVspoof attack challenge held in 2017 [35]. Lantain et al. [36] observed that the use of inverted Mel-frequency cepstral coefficients (IMFCC) for replay attack detection outperformed that of standard MFCC features. In [37], the researchers proposed a DNN-frontend architecture, which uses high frequency cepstral coefficients (HFCC) in tandem with CQCC features, for replay attack detection. They employed the ASVspoof dataset and used a SVM classifier. Some of the research has employed the conventional systems, such as GMM-UBM and i-vectors with DNN for better performance and precision. These systems are usually employed in text-dependent speaker verification systems [24, 25]. For instance, DNN is used for extracting frames from a speaker's speech and calculating its utterance-level information. Subsequently, the output of the DNN is converted to i-vectors, and at the backend, PLDA is used for a verification score [26, 27]. In such works, the DNN features are either used alone or combined with the conventional features (i.e., MFCC). Features extracted from DNN are employed in the posteriors. These researchers use DNN to extract speaker specific data and apply the similarity metrics check for the verification score [28]. In [38], the researchers used a DNN-based classifier using features extracted from the long-term average spectrum (LTAS). In the end, the proposed DNN classifier is compared with the GMM classifier that uses MFCC and CQCC features. The result has shown that the DNN outperforms the GMM classifier.

The researchers in [33] proposed an end-to-end deep learning approach for the detection of replay spoof attacks. The approach consists of a lite convolutional neural network (LCNN) (at the frontend) and recurrent neural network (RNN) at the backend. Observations reported in this paper included that frequency range analysis for replay attack detection performs better with 4-8 kHz and 6-8 kHz sub-bands, which carry the most discriminative pieces of information. The researchers in [39, 40] have also employed CNN for replay attack detection using a hybrid feature of MFCC and CQCC extractions. They used the CNN as the classifier in the backend, while using GMM for the frontend of an automatic speaker verification system. The researchers have built a frontend feature extractor using CNN; they used CQCC features for the system. In [41], the researchers propose a system with an improved spoofing detection approach. The approach uses Gated Recurrent CNN as a deep feature extractor, which is later used as the backend classifier. They propose a new input feature, signal-to-noise masks (SNMs). An end-to-end system that employs CNN as feature extractor and RNN as the classifier is proposed for spoof detection in [23].

III. Feature for Raplay Attack Detection

Detecting a replay attack needs systems to identify whether the speech segment is played live or recorded-and-replayed. Naturally, in replay attacks, the tone and context of the speech of the speaker does not change. As depicted in Figure 2, the difference between the genuine and replay speeches is obvious. Since the replay speech uses recording and replay devices, the replay speech carries information about the two devices on top of the genuine speech. As such, replay attack detection not only needs to focus on the content of the speech, but also whether the sound characteristics of the recording and replay devices exist. It became difficult for one type of feature, such as constant Q cepstral coefficient (CQCC), standard Mel-frequency cepstral coefficient (MFCC), or long-term average spectrum (LTAS), to

fit in the representation of replay spoof detection [21]. To that end, the hybrid features of CQCC and MFCC are used as an input to the DNN classifier in this research. Previous studies conducted on replay attack detection have shown promising results by employing the hybrid feature extraction method to train the classifiers [38]. The hybrid features work better in determining the differences between the genuine and replay speech. Some researchers have used the deep learning for classifiers but with different combinations of features. For instance, the hybrid features of high frequency cepstral coefficients (HFCC) with CQCC is used in [37], and researchers in [33] used CQCC for replay detection.

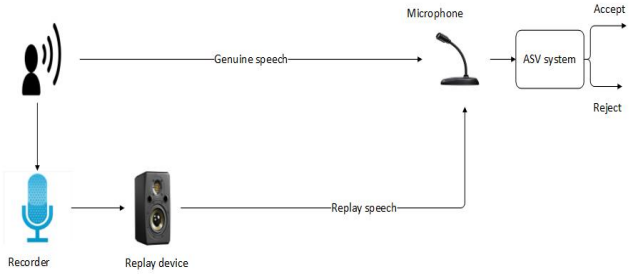

Fig. 2. Replay attack process on ASV systems

In the extraction of the MFCC features, as one of the best known features of speech processing, the power-spectrum of the framed speech signal is transformed by a filter bank for dimensionality reduction. In speech processing, this research uses MFCC to process high frequency regions that are most affected by the spoofing artifacts, while CQCC is used to extract low frequency regions that may carry minor, but additional information to distinguish between genuine and replayed speeches. These two features are used to exploit possible complementary characteristics in feature space. As shown in Figure 3, the resulting hybrid features are used for training the DNN classifier.

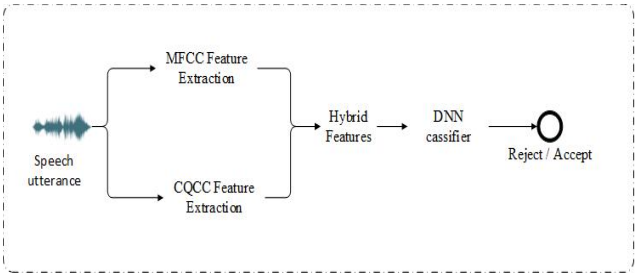

Fig. 3. Hybrid feature extraction process

In contrast to the hybrid features employed in [40], our research uses zero-mean and unit variance normalization 30-dimentional MFCC and CQCC features, along with first and second derivatives. The mean and variances are computed on the full training data.

IV. THE DNN-ARCHITECTURE CLASSIFIER

In this paper, a classification model in the form of DNN is proposed. In most cases, for replay spoof attack detection, DNN is used for feature extraction. The proposed DNN architecture is fully connected to a feed-forward neural network. It consists of multiple hidden layers. When using deep learning, it is necessary to have enough samples to improve the performance of the deep network. There are a number of problems associated with DNN. The first problem is the need for more computational capacity; this occurs from the input of every layer and the parameters of preceding layers that change in every training epoch. The second problem associated with DNN is over-fitting. Nevertheless, these issues can be tackled with batch training (dividing data into groups), which helps overcome the memory requirement. Batch normalization is also useful to accelerate the learning process and help reduce the effect of hyper-parameters. In the end, the dropout technique is a solution to prevent over-fitting.

Spoofing-discriminant DNNs with five hidden layers are used to distinguish the genuine from the replay spoof speech in this paper. Each of the layers has 2048 nodes with a sigmoid activation function. The nodes at the output layer is K + 1 and the softmax activation function is used in that layer. The K + 1 output nodes correspond to one genuine speech and K spoofing attacks. Batch normalized super vectors F_i that are composed of n hybrid features vectors are used for DNN training.

$$F_i = \{x_1, x_2, \ldots\ldots, x_T\}$$

Using Tensorflow/keras, the DNN is built and trained. Stochastic gradient descent methods are used and the cross entropy function is selected as the cost function. The maximum training epoch is chosen as 120, while the mini-batch size is set as 128. By counting how many frames are similar to the genuine speech, the replay speech is detected. The posterior obtained at the output layer of DNN is transformed into Log-Likelihood (LLR) score as:

$$LLR_{DNN} = \log p(genuine|F_i) - \log p(replay|F_i)$$

V. EXPERIMENT EVALUATION AND RESULTS

The research uses the ASVspoof 2017 dataset [42] in the experiment. The dataset is produced for spoof attacks, most of which are replay spoof attacks. The resolution of all the audio signals in the dataset is 16 bits; the sampling rate is 16 kHz. The duration of the utterance lasts approximately 3 to 5 seconds. The summary of the dataset is depicted in Table 1.

TABLE I. SUMMARY OF DATASET USED IN THE EXPERIMENT

subsets	#Speakers	#Genuine utterances	#Spoofed utterances
Training	10	1508	1508
Development	8	760	950
Evaluation	24	1298	12008
Total	42	3566	14466

A hamming window with the size of 10ms is used to calculate the short-time zero-crossing rate and short-term energy. Then the threshold is set to find the silent segment and speech segment. The silent segment is of the length of 512 ms, while the speech segment has a length of 1525 ms. A segment with insufficient length is duplicated. The silent segment is represented with a MFCC feature and, after setting at a

frequency of 3-8kKz, it gets 60*78 features. The speech segment extracts are represented with a CQCC feature and the selected hamming window. The window size is 50 ms; and the step size is 25 ms, that is, there is a 25 ms overlap, and 48 coefficients selected.

After CQCC and MFCC features are extracted and combined, a feature of 60 x 126 is obtained. The feature is then used as an input to the DNN classifier. All three subsets were employed for the model training. The result has shown that the hybrid features with the DNN classifier produces the best performance. We have used a single feature as input to baseline GMM system and DNN syatem as well. The results are shown in Table 2.

TABLE II. EVALUATION RESULTS

Systems	EER%		
	Dev	Eval	Eval + Dev
CQCC - GMM	8.18	27.79	15.54
MFCC - GMM	5.54	25.54	13.56
CQCC - DNN	10.8	23.57	10.8
MFCC - DNN	7.36	20.84	14.78
MFCC + CQCC - GMM	8.67	22.67	18.25
MFCC + CQCC - DNN	**2.68**	**7.65**	**5.64**

We used CQCC as input features to the DNN. The system got very good results with EERs of 10.8%, 23.57%, and 10.8% on the Dev, Eval, and Dev+Eval sets respectively. Likewise, we used MFCC as input features to the DNN and shown good results with ERRs of 5.54%, 25.54%, and 13.54% on the Dev, Eval, and Dev+Eval sets respectively. This shows that our DNN classifier has outperformed the GMM based classifier. Using the hybrid features (MFCC+CQCC) as input to the DNN system, the result shows ERRs of 2.68% on Dev set, 7.65% on Eval set, and 5.64% on Dev+Eval set. The result also shows the benefit of combining CQCC and MFCC features. The overall results show that the use of the Hybrid features with our DNN system outperforms the GMM systems.

VI. CONCLUSION

In this study, we explored the application of deep learning frameworks as classifier for the detection of replay spoofing attacks in automatics speaker verification systems. The study investigates implementation DNN at the back of ASV systems which is not common in the literature. The model has been shown effective in the experiment and the proposed solution outperformed existing classifications. The ASVspoof 2017 dataset was employed for the experiment where hybrid features of CQCC and MFCC extractions to train the model. The hybrid features used as the input to the DNN model shown that good discriminative characteristics were extracted from the speech utterance to differentiate between genuine and replay speeches. In the future, will be working an end to end system that uses deep learning for both frontend and backend of ASV systems for replay spoof detection with improved performance.

REFERENCES

[1] Evans, N., et al., *Speaker recognition anti-spoofing*, in *Handbook of biometric anti-spoofing*. 2014, Springer. p. 125-146.

[2] Jain, A.K., R. Bolle, and S. Pankanti, *Biometrics: personal identification in networked society*. Vol. 479. 2006: Springer Science & Business Media.

[3] Gayathri, M., C. Malathy, and M. Prabhakaran. *A Review on Various Biometric Techniques, Its Features, Methods, Security Issues and Application Areas*. in *International Conference On Computational Vision and Bio Inspired Computing*. 2019. Springer.

[4] Zhang, X., et al. *Voice Biometric Identity Authentication System Based on Android Smart Phone*. in *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*. 2018. IEEE.

[5] Duraibi, S., F.T. Sheldon, and W. Alhamdani, *VOICE BIOMETRIC IDENTITY AUTHENTICATION MODEL FOR IOT DEVICES.*

[6] Wu, Z., et al., *Spoofing and countermeasures for speaker verification: A survey.* speech communication, 2015. **66**: p. 130-153.

[7] Vestman, V., et al., *Voice mimicry attacks assisted by automatic speaker verification.* Computer Speech & Language, 2020. **59**: p. 36-54.

[8] Pal, M. and G. Saha, *On robustness of speech based biometric systems against voice conversion attack.* Applied Soft Computing, 2015. **30**: p. 214-228.

[9] Wang, X., et al., *ASVspoof 2019: a large-scale public database of synthetized, converted and replayed speech.* Computer Speech & Language, 2020: p. 101114.

[10] Alegre, F., A. Janicki, and N. Evans. *Re-assessing the threat of replay spoofing attacks against automatic speaker verification*. in *2014 International Conference of the Biometrics Special Interest Group (BIOSIG)*. 2014. IEEE.

[11] Hong, Q., S. Wang, and Z. Liu. *A robust speaker-adaptive and text-prompted speaker verification system*. in *Chinese Conference on Biometric Recognition*. 2014. Springer.

[12] De Leon, P.L., et al., *Evaluation of speaker verification security and detection of HMM-based synthetic speech.* IEEE Transactions on Audio, Speech, and Language Processing, 2012. **20**(8): p. 2280-2290.

[13] Bredin, H., et al. *Detecting replay attacks in audiovisual identity verification*. in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. 2006. IEEE.

[14] Nematollahi, M., et al. *Digital speech watermarking for anti-spoofing attack in speaker recognition*. in *2014 IEEE REGION 10 SYMPOSIUM*. 2014. IEEE.

[15] Shang, W. and M. Stevenson, *Detection of speech playback attacks using robust harmonic trajectories.* Computer Speech & Language. **65**: p. 101133.

[16] Rafi B, S. and S.R.M. Kodukula, *Importance of analytic phase of the speech signal for detecting Replay attacks in automatic speaker verification systems.* 2019.

[17] Shang, W. and M. Stevenson. *A playback attack detector for speaker verification systems*. in *2008 3rd International Symposium on Communications, Control and Signal Processing*. 2008. IEEE.

[18] Faundez-Zanuy, M., M. Hagmüller, and G. Kubin, *Speaker verification security improvement by means of speech watermarking.* Speech communication, 2006. **48**(12): p. 1608-1619.

[19] Villalba, J. and E. Lleida. *Preventing replay attacks on speaker verification systems*. in *2011 Carnahan Conference on Security Technology*. 2011. IEEE.

[20] Korshunov, P., et al. *Overview of BTAS 2016 speaker anti-spoofing competition*. in *2016 IEEE 8th international conference on biometrics theory, applications and systems (BTAS)*. 2016. IEEE.

[21] Singh, M., J. Mishra, and D. Pati. *Replay attack: Its effect on GMM-UBM based text-independent speaker verification system*. in *2016 IEEE Uttar Pradesh Section International Conference on*

Electrical, Computer and Electronics Engineering (UPCON). 2016. IEEE.

[22] Shaikh, R. and M. Sasikumar, *Data Classification for achieving Security in cloud computing.* Procedia computer science, 2015. **45**: p. 493-498.

[23] Zhang, C., C. Yu, and J.H. Hansen, *An investigation of deep-learning frameworks for speaker verification antispoofing.* IEEE Journal of Selected Topics in Signal Processing, 2017. **11**(4): p. 684-694.

[24] Richardson, F., D. Reynolds, and N. Dehak, *A unified deep neural network for speaker and language recognition.* arXiv preprint arXiv:1504.00923, 2015.

[25] Snyder, D., et al. *Deep neural network-based speaker embeddings for end-to-end speaker verification*. in *2016 IEEE Spoken Language Technology Workshop (SLT)*. 2016. IEEE.

[26] Matějka, P., et al. *Analysis of DNN approaches to speaker identification*. in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. 2016. IEEE.

[27] Li, C., et al., *Deep speaker: an end-to-end neural speaker embedding system.* arXiv preprint arXiv:1705.02304, 2017. **650**.

[28] Dey, S., et al. *Deep neural network based posteriors for text-dependent speaker verification*. in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2016. IEEE.

[29] Xu, J., et al., *Deep multi-metric learning for text-independent speaker verification.* Neurocomputing, 2020. **410**: p. 394-400.

[30] Irum, A. and A. Salman, *Speaker Verification Using Deep Neural Networks: A.* International Journal of Machine Learning and Computing, 2019. **9**(1).

[31] Ustinova, E., Y. Ganin, and V. Lempitsky. *Multi-region bilinear convolutional neural networks for person re-identification*. in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2017. IEEE.

[32] Hafemann, L.G., R. Sabourin, and L.S. Oliveira. *Writer-independent feature learning for offline signature verification using deep convolutional neural networks*. in *2016 international joint conference on neural networks (IJCNN)*. 2016. IEEE.

[33] Lavrentyeva, G., et al. *Audio replay attack detection with deep learning frameworks*. in *Interspeech*. 2017.

[34] Chen, K. and A. Salman, *Learning speaker-specific characteristics with a deep neural architecture.* IEEE Transactions on Neural Networks, 2011. **22**(11): p. 1744-1756.

[35] Kinnunen, T., et al. *Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research*. in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017. IEEE.

[36] Li, L., et al., *A study on replay attack and anti-spoofing for automatic speaker verification.* arXiv preprint arXiv:1706.02101, 2017.

[37] Nagarsheth, P., et al. *Replay Attack Detection Using DNN for Channel Discrimination*. in *Interspeech*. 2017.

[38] Bakar, B. and C. Hanilçi. *An Experimental Study on Audio Replay Attack Detection Using Deep Neural Networks*. in *2018 IEEE Spoken Language Technology Workshop (SLT)*. 2018. IEEE.

[39] Huang, L., Y. Gan, and H. Ye. *Audio-replay Attacks Spoofing Detection for Automatic Speaker Verification System*. in *2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*. 2019. IEEE.

[40] Huang, L. and C.-M. Pun. *Audio replay spoof attack detection using segment-based hybrid feature and densenet-LSTM network*. in *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. 2019. IEEE.

[41] Gomez-Alanis, A., et al., *A gated recurrent convolutional neural network for robust spoofing detection.* IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019. **27**(12): p. 1985-1999.

[42] Kinnunen, T., et al., *The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection.* 2017.