INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

A Light Convolutional GRU-RNN Deep Feature Extractor for ASV Spoofing Detection

Alejandro Gomez-Alanis[1], Antonio M. Peinado[1], Jose A. Gonzalez[2], and Angel M. Gomez[1]

[1]University of Granada, Granada, Spain
[2]University of Malaga, Malaga, Spain
{agomezalanis,amp,amgg}@ugr.es, j.gonzalez@uma.es

Abstract

The aim of this work is to develop a single anti-spoofing system which can be applied to effectively detect all the types of spoofing attacks considered in the ASVspoof 2019 Challenge: text-to-speech, voice conversion and replay based attacks. To achieve this, we propose the use of a Light Convolutional Gated Recurrent Neural Network (LC-GRNN) as a deep feature extractor to robustly represent speech signals as utterance-level embeddings, which are later used by a back-end recognizer which performs the final genuine/spoofed classification. This novel architecture combines the ability of light convolutional layers for extracting discriminative features at frame level with the capacity of gated recurrent unit based RNNs for learning long-term dependencies of the subsequent deep features. The proposed system has been presented as a contribution to the ASVspoof 2019 Challenge, and the results show a significant improvement in comparison with the baseline systems. Moreover, experiments were also carried out on the ASVspoof 2015 and 2017 corpora, and the results indicate that our proposal clearly outperforms other popular methods recently proposed and other similar deep feature based systems.

Index Terms: spoofing detection, automatic speaker verification, deep learning, ASVspoof.

1. Introduction

Automatic Speaker Verification (ASV) aims to authenticate the identity claimed by a given individual based on the provided speech samples [1]. This technology has gained significant interest in recent years due to its commercial applications. As the importance of this technology grows, so does the concerns about its security. Four types of spoofing attacks have been identified [2]: (i) replay (i.e. using pre-recorded voice of the target user), (ii) impersonation (i.e. mimicking the voice of the target voice), and, also, either (iii) text-to-speech synthesis (TTS) or (iv) voice conversion (VC) systems to generate artificial speech resembling the voice of a legitimate user. The aim of this work is the development of a common framework aiming at detecting different spoofing attacks, namely TTS, VC and replay attacks.

One popular approach for spoofing detection is to deploy machine learning techniques in order to learn to discriminate genuine vs. spoofed speech, using a training dataset. It is desirable that the anti-spoofing system learns to detect not only the attacks observed in the training dataset, but also be able to generalize to unseen attacks. To address this issue, deep feature extraction has been proposed in [3], where feature embeddings are extracted from an inner layer of a deep neural network to represent every temporal frame of the voice signal, or even the whole utterance.

Deep neural networks have shown to be very effective for feature engineering in several speech-based applications [4]. Their nonlinear modeling and discriminative capabilities make them not only a powerful back-end classifier [5, 6], but also advantageous for feature extraction [7]. The architecture of these deep feature extractors has shown to be determinant for the performance of the anti-spoofing system.

This paper presents a novel neural network architecture for ASV-based spoofing detection. We propose a hybrid light convolutional neural network (LCNN) [8] plus recurrent neural network (RNN) architecture which combines the ability of the LCNNs for extracting discriminative features at frame level with the capacity of gated recurrent unit (GRU) based RNNs for learning long-term dependencies of the subsequent deep features. The resulting architecture will be referred to as Light Convolutional Gated Recurrent Neural Network (LC-GRNN). Despite the fact that similar deep learning frameworks have been applied in learning video representations [9], audio tagging [10] and optical character recognition [11], to the best of our knowledge our work constitutes the first adaptation of such architecture to the problem of spoofing detection.

Our system has participated in the ASVspoof 2019 Challenge, whose results will be presented at a special session of Interspeech 2019, to address the issue of detecting: (i) logical access attacks (generated by TTS or VC algorithms), and (ii) physical access attacks (replay). Furthermore, we also evaluate our proposal on the ASVspoof 2015 [12] and 2017 [13] datasets in order to provide a comparison with other state-of-the-art systems.

This paper is organized as follows. Section 2 describes the proposed deep feature extractor employed along the work. Then, in Section 3, we outline the speech corpora, the network training and the system details. Section 4 discusses the performance of our system on the ASVspoof 2015 [12], 2017 [13] and 2019 [14] databases. Finally, we summarize the conclusions derived from this research in Section 5.

2. System Description

In our previous work [15], we proposed a hybrid CNN-RNN architecture to compute spoofing embeddings at utterance level. When evaluated on standard spoofing databases, our architecture was able to outperform other similar deep feature extractors which average frame-level features for getting the spoofing identity vector of the utterance.

Since our preliminary system of [15], our work has focused on building a better integration of convolutional and recurrent layers. In particular, in this work we take one step forward and propose to replace the fully connected layers inside the recurrent cells with LCNN layers in order to: (1) extract discriminative features at frame level, (2) learn long-term dependencies,

http://dx.doi.org/10.21437/Interspeech.2019-2212

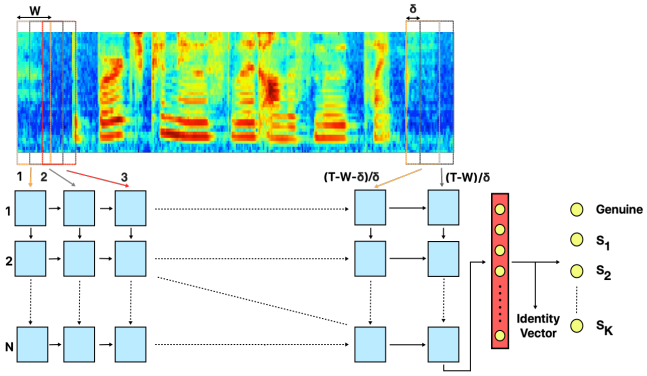

Figure 1: *Block diagram of the proposed LC-GRNN utterance-level identity vector extractor.*

and (3) integrate the extraction of frame-level deep features and the utterance-level identity vector into a single network. The Light term of the convolutional layers stands for the usage of Max-Feature-Map (MFM) activations. These are applied to reduce the dimension of the output and obtain more discriminative feature maps, as it has been shown for face recognition [8] and replay spoofing detection [16].

A block diagram of the proposed LC-GRNN architecture is shown in Fig. 1. At each time step, the LC-GRNN processes a context window of W consecutive frames. This context window moves forward δ frames on every time step[1], so that the total number of time steps of the LC-GRNN is $(T-W)/\delta$, where T is the number of frames of the utterance being processed. Moreover, the LC-GRNN has N recurrent layers. This architecture acts as a classifier whose task is to determine whether the input utterance is either genuine or belongs to one of the K spoofing attacks included in the training set $(S_1, S_2, ..., S_K)$. In order to do this, the output of the last time step and last recurrent layer is fed to a fully connected layer with MFM activation to obtain the spoofing identity vector of the whole utterance. During the training phase, this identity vector is finally passed through another fully connected layer with softmax activation of $K+1$ neurons to discriminate between the genuine and the K spoofing classes.

Unlike classical RNNs, the hidden state $\mathbf{h}_t^n$ $(t = 1, ..., (T-W)/\delta;\ n = 1, ..., N)$ of the LC-GRNN model is computed by convolving the current input features $\mathbf{x}_t^n$ and the previous state $\mathbf{h}_{t-1}^n$ with multiple filters. Due to the fact that most of the cues that enable the detection of spoofing attacks can be found in certain frequency bands [17], we embed such a prior in our deep feature extractor architecture by replacing the fully-connected operations in the GRU with convolutions. This has the potential advantage that more discriminative features can be extracted at the frame level [18].

As shown in Fig. 2, similarly to a GRU cell, our LC-GRU cell defines three gates, each one implemented by means of a LCNN. Each LCNN block in Fig. 2 consists of either one or two LCNN layers as shown in Fig. 3. Every LCNN layer performs convolutions (one or two) followed by an MFM operation intended to reduce the output feature maps by a $1/2$ factor via

[1]Instead of moving forward the context window one frame on every time step, we propose to move it δ frames ($\delta <$ W) in order to reduce the processing time, while maintaining the classification performance.

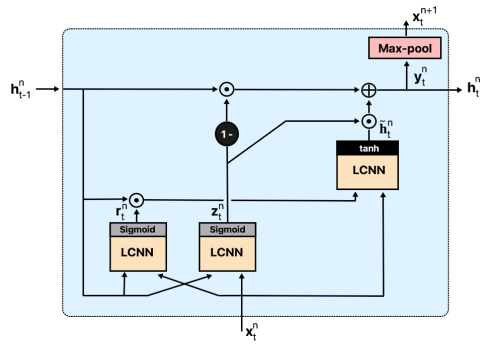

Figure 2: *Light Convolutional Gated Recurrent Unit cell (LC-GRU).*

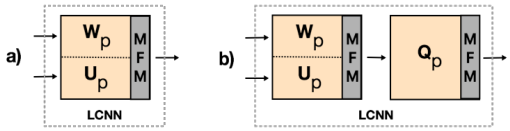

Figure 3: *Possible LCNN block configurations inside the LC-GRU cell. a) 1-layer LCNN, b) 2-layer LCNN. ($p = r, z, \tilde{h}$).*

a competitive relationship [8]. In this way, each time step of the LC-GRNN plays the role of a frame-level deep feature extractor providing N state (feature) vectors for each context window of W consecutive frames.

The update $\mathbf{z}_t^n$ and reset $\mathbf{r}_t^n$ gates determine which information from the previous frames needs to be passed along the next time steps, avoiding the risk of the vanishing gradient problem [19]. In the case of a single-layer LCNN block (Fig. 3a), they are computed as

$$\mathbf{z}_t^n = \sigma(\text{MFM}(\mathbf{W}_z^n * \mathbf{x}_t^n + \mathbf{U}_z^n * \mathbf{h}_{t-1}^n)), \quad (1)$$

$$\mathbf{r}_t^n = \sigma(\text{MFM}(\mathbf{W}_r^n * \mathbf{x}_t^n + \mathbf{U}_r^n * \mathbf{h}_{t-1}^n)), \quad (2)$$

where the operator $*$ denotes a convolution operation. These convolutional layers can be interpreted as filter banks which are trained and optimized to detect artifacts from the spoofed speech. The main advantage of employing these filters is the extraction of frame-level features at every time step which are more discriminative than those extracted by using fully connected units [20]. Similarly, the update activation gate

$$\tilde{\mathbf{h}}_t^n = tanh(\text{MFM}(\mathbf{W}_{\tilde{h}}^n * \mathbf{x}_t^n + \mathbf{U}_{\tilde{h}}^n * (\mathbf{r}_t^n \odot \mathbf{h}_{t-1}^n))), \quad (3)$$

uses the reset gate to store the relevant information from the past frames, removing firstly the non-relevant information through an element-wise multiplication (denoted as $\odot$) with the previous state. In these equations, MFM$(\cdot)$ is the Max-Feature-Map function, and $\sigma(\cdot)$, $tanh(\cdot)$ are the activation functions of the gates. The model parameters $\mathbf{W}_z^n$, $\mathbf{W}_r^n$, $\mathbf{W}_{\tilde{h}}^n$, $\mathbf{U}_z^n$, $\mathbf{U}_r^n$, $\mathbf{U}_{\tilde{h}}^n$, and $\mathbf{Q}_z^n$, $\mathbf{Q}_r^n$, $\mathbf{Q}_{\tilde{h}}^n$ are the filters of the 3 described LCNNs, which are shared in each time step of the LC-GRNN.

3. Experimental Framework

This section briefly describes the databases employed in our experiments, as well as the details of the proposed system.

Table 1: *LC-GRNN architecture ($p = r, z, \tilde{h}$).*

LC-GRNN	Type	Filter / Stride	Output
Layer 1	Conv ($\mathbf{W}_p^1$, $\mathbf{U}_p^1$)	5×5 / 1×1	$16 \times 256 \times 32$
	MFM	-	$8 \times 256 \times 32$
	MaxPool	2×1 / 2×1	$8 \times 128 \times 32$
Layer 2	Conv ($\mathbf{W}_p^2$, $\mathbf{U}_p^2$)	1×1 / 1×1	$16 \times 128 \times 32$
	MFM	-	$8 \times 128 \times 32$
	Conv ($\mathbf{Q}_p^2$)	3×3 / 1×1	$32 \times 128 \times 32$
	MFM	-	$16 \times 128 \times 32$
	MaxPool	2×1 / 2×1	$16 \times 64 \times 32$
Layer 3	Conv ($\mathbf{W}_p^3$, $\mathbf{U}_p^3$)	1×1 / 1×1	$32 \times 64 \times 32$
	MFM	-	$16 \times 64 \times 32$
	Conv ($\mathbf{Q}_p^3$)	3×3 / 1×1	$16 \times 64 \times 32$
	MFM	-	$8 \times 64 \times 32$
	MaxPool	2×1 / 2×1	$8 \times 32 \times 32$
-	FC1	-	512×2
	MFM	-	512
-	FC2	-	$K + 1$

3.1. Speech Corpora

We evaluated the proposed anti-spoofing system on both logical access (LA) and physical access (PA) attacks with the corpora described below.

3.1.1. Logical Access

A total of 10 and 19 TTS/VC attacks were generated for the ASVspoof 2015 [12] and 2019 LA [14] databases, respectively. However, only $K = 5$ and $K = 6$ attacks have been employed for training the ASVspoof 2015 and 2019 LA models, respectively, since the rest of attacks belong to the evaluation sets (unknown attacks).

3.1.2. Physical Access

The replay attacks of the ASVspoof 2017 version 1 [13] were generated using 3 categories (low, medium, high) of recording and playback devices. For a balanced training, we have considered $K = 4$ types of replay attacks as a result of combining low/medium and high qualities of both playback and recording devices.

On the other hand, ASVspoof 2019 PA [14] database includes a total of 9 different replay configurations, comprising 3 categories of attacker-to-speaker recording distances, and 3 categories of loudspeaker quality. The evaluation data was generated with different randomly acoustic and replay configurations. Each replay configuration has been considered a different type of replay attack, so that $K = 9$ replay attacks have been used for training.

3.2. System

This section details the methodology followed to train our proposed system. First, speech signals were segmented using a Blackman analysis window of 16 ms length with 4 ms of frame shift. Log magnitude spectrogram features with $F = 256$ bins were obtained to feed the proposed deep feature extractor described in Section 2. It processes context windows of $W = 32$ frames with a shift of $\delta = 12$ frames. For every experiment on each of the 4 described speech corpora, the system was trained employing only the training set of the corresponding database.

Table 1 shows a summary of the employed LC-GRNN architecture. It is composed by $N = 3$ recurrent layers, where each one has different light convolutional layers followed by a max pooling operation which reduces the frequency dimension. The LCNN block architectures are inspired by the ones proposed in [8, 16], which were able to extract very discriminative features. Once all the frame-level context windows are processed by the convolutional and recurrent layers, 8 feature maps of size 32×32 are flattened to make up a feature vector of 8192 components. Then, this vector is fed to a fully connected layer (FC1) with MFM activation to obtain the spoofing identity vector of the utterance of 512 components.

The proposed deep feature extractor was trained using the Adam optimizer [21] with a learning rate of $3 \cdot 10^{-4}$, early stopping, 60% dropout (FC1), and normalizing the input in mean and variance. All the specified hyperparameters were optimized using the development sets of the ASVspoof 2019 data corpora.

For the back-end, we evaluated three different classifiers: support vector machine (SVM), linear discriminant analysis (LDA), and its probabilistic version (PLDA). The objective of the classifier is to assign a score indicating whether the utterance is genuine or spoofed. In some models, we also applied a posterior normalization of the scores. Provided the prior of the different classes is uniform, the normalized score of the spoofing identity vector $\mathbf{x}$ is

$$p(genuine|\mathbf{x}) = log \frac{p(\mathbf{x}|genuine)}{\sum_{j=1}^{K+1} exp(p(\mathbf{x}|j))}, \quad (4)$$

where $p(\mathbf{x}|j)$ is the log posterior predictive probability of the spoofing identity vector $\mathbf{x}$ given class j ($j = 1, ..., K + 1$, including the genuine class).

4. Results

4.1. Results on ASVspoof 2015

Table 2 compares the performance of our proposed anti-spoofing system with different state-of-the-art and deep feature extractor based systems on the ASVspoof 2015 database. Our proposed system with PLDA classifier achieves the best performance in the known and unknown attacks, outperforming other state-of-the-art systems such as the CQCC + GMM [22] and LTSS + MLP [23]. It clearly outperforms other similar deep feature extractors of the literature such as the CNN + RNN [20] and FBANK + Best RNN [24], as well as our previous system FBANK + CNN + RNN [15]. In particular, the performance of our proposals for S10 attack is quite meaningful. Regarding the classifiers employed to score the spoofing identity vectors extracted by our proposed LC-GRNN, PLDA without scoring normalization yields the lowest Equal Error Rate (EER), outperforming the other 2 classifiers (SVM and LDA) in the S10 attack.

4.2. Results on ASVspoof 2017

Table 3 shows a comparison of the performance of our anti-spoofing system with different state-of-the-art single systems on the ASVspoof 2017 database. Our proposed system with PLDA and scoring normalization achieves the best performance. It outperforms other state-of-the-art single systems such as the SCMC + GMM [25], LCNN + GMM [16] and CNN + RNN [16], which were presented to the ASVspoof 2017 Challenge. In fact, our proposal achieves an EER 0.32% lower than the Siamese CNN + GMM [26], which is one of the state-of-the-art single deep feature extractors for this corpus.

Table 2: *Comparison between classifiers and with other systems on evaluation set of ASVspoof 2015 in terms of (%) EER*

System	Known	Unknown	
		S6 - S9	S10
Spectro + CNN + RNN [20]	0.40	0.60	14.27
FBANK + Best RNN [24]	0.20	0.50	10.70
FBANK + CNN + RNN [15]	0.03	0.13	9.34
CQCC + GMM [22]	0.05	0.31	1.07
LTSS + MLP [23]	0.10	0.11	1.56
LC-GRNN + SVM	**0.00**	**0.00**	1.01
LC-GRNN + LDA	**0.00**	0.01	0.82
LC-GRNN + PLDA (Norm.)	**0.00**	0.03	2.83
LC-GRNN + PLDA	**0.00**	**0.00**	**0.69**

Table 3: *Comparison between classifiers and with other systems on ASVspoof 2017 database in terms of (%) EER*

System	Development	Evaluation
Baseline: CQCC + GMM	10.35	30.60
SCMC + GMM [25]	9.32	11.49
LCNN + GMM [16]	4.53	7.37
CNN + RNN [16]	7.51	10.69
Siamese CNN + GMM [26]	-	6.40
LC-GRNN + SVM	4.62	8.12
LC-GRNN + LDA	4.10	7.53
LC-GRNN + PLDA	3.42	6.35
LC-GRNN + PLDA (Norm.)	**3.26**	**6.08**

Regarding the scoring normalization, we can conclude that it is beneficial when the nature of the unseen attacks is similar to the seen ones, such as the different replay configurations considered for training the ASVspoof 2017 model. However, the genuine class can be easily mistaken for an attack when it is generated with a new technique not seen during training (as it happens with the S10 attack based on MaryTTS [27] in the ASVspoof 2015 corpus).

4.3. Results on ASVspoof 2019 LA

Because of the good results obtained on the ASVspoof 2015 database at detecting logical access attacks, we decided to submit the LC-GRNN + PLDA as primary and single system, the LC-GRNN + LDA as contrastive 1 system, and the LC-GRNN + SVM as contrastive 2 system, to the ASVspoof 2019 Logical Access Challenge [14]. We can only compare their performance with the baseline systems because the participating systems have not been published yet.

Table 4 shows the results on the ASVspoof 2019 LA database obtained by the baseline systems (CQCC + GMM [22] and LFCC + GMM [28]) and our submitted systems. The proposed LC-GRNN system outperforms the baseline systems on the development and evaluation sets independently of the scoring classifier. Specifically, our contrastive 1 system (LC-GRNN + LDA) achieves a relative 22.37% and 34.38% better performance than CQCC + GMM and LFCC + GMM on the evaluation set, respectively. It is worth noticing that although our contrastive 1 system outperforms our primary/single system (LC-GRNN + PLDA), the difference of EER is only 0.06%. Taking into account that PLDA only performed 0.01% better on the overall EER of ASVspoof 2015 evaluation set, we can conclude that LDA and PLDA classifiers have a similar performance at scoring the LA spoofing identity vectors extracted by the proposed LC-GRNN system.

Table 4: *Results on ASVspoof 2019 Logical Access in terms of min-tDCF and EER (%)*

System	min-tDCF		EER (%)	
	Dev.	Eval.	Dev.	Eval.
Baseline 1: CQCC + GMM	0.0123	0.2366	0.43	9.57
Baseline 2: LFCC + GMM	0.0663	0.2116	2.71	8.09
LC-GRNN + SVM	0.0002	0.1873	0.01	7.12
LC-GRNN + PLDA	**0.0000**	0.1552	**0.00**	6.34
LC-GRNN + LDA	**0.0000**	**0.1523**	**0.00**	**6.28**

Table 5: *Results on ASVspoof 2019 Physical Access in terms of min-tDCF and EER (%)*

System	min-tDCF		EER (%)	
	Dev.	Eval.	Dev.	Eval.
Baseline 1: CQCC + GMM	0.1953	0.2454	9.87	11.04
Baseline 2: LFCC + GMM	0.2554	0.3017	11.96	13.54
LC-GRNN + LDA	0.0469	0.0946	1.59	3.49
LC-GRNN + PLDA	0.0306	0.0747	1.18	2.68
LC-GRNN + PLDA (Norm.)	**0.0203**	**0.0614**	**0.73**	**2.23**

4.4. Results on ASVspoof 2019 PA

According to the results obtained on the ASVspoof 2017 database at detecting replay attacks, we decided to submit the LC-GRNN + PLDA with scoring normalization as primary and single system, the LC-GRNN + PLDA without normalization as contrastive 1 system, and the LC-GRNN + LDA as contrastive 2 system, to the ASVspoof 2019 Physical Access Challenge [14].

Table 5 shows the results on the ASVspoof 2019 PA database obtained by the baseline systems and our submitted systems. The proposed LC-GRNN system clearly outperforms the baseline systems on the development and evaluation sets independently of the scoring classifier. Specifically, our primary/single system (LC-GRNN + PLDA with scoring normalization) achieves a relative 74.69% and 79.80% better performance than CQCC + GMM and LFCC + GMM on the evaluation set, respectively.

5. Conclusions

This paper has proposed a novel technique for the extraction of utterance-level identity vectors for an efficient detection of TTS/VC and replay attacks. In our system, a gated recurrent unit based RNN learns long-term dependencies of the subsequent deep features, while several integrated light convolutional neural networks extract discriminative features at frame level. This proposal has been submitted as a single system to the ASVspoof 2019 Challenge [14].

The results show that our proposed system notably outperforms the baseline systems of the ASVspoof 2019 challenges (CQCC + GMM and LFCC + GMM). Moreover, it also yields very remarkable results as single system on the ASVspoof 2015 and 2017 databases, outperforming other popular methods such as the CQCC + GMM [22] and even the fusion of systems (winner of the 2017 challenge) presented in [16].

6. Acknowledgements

This work has been supported by the Spanish MINECO/FEDER Project TEC2016-80141-P and the Spanish Ministry of Education through the National Program FPU (grant reference FPU16/05490). We also acknowledge the support of NVIDIA Corporation with the donation of a Titan X GPU, as well as the organizers of the ASVspoof 2019 Challenge.

7. References

[1] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[2] Z. Wu et al., "Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 768–783, 2016.

[3] N. Chen, Y. Qian, H. Dinkel, B. Chen and K. Yu, "Robust Deep Feature for Spoofing Detection - The SJTU System for ASVspoof 2015 Challenge," *Proc. Interspeech*, 2015.

[4] S. Yadav, and A. Rai, "Learning Discriminative Features for Speaker Identification and Verification," *Proc. Interspeech*, 2018.

[5] X. Tian, Z. Wu, X. Xiao, E.S. Chng, and H. Li, "Spoofing detection from a feature representation perspective," *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[6] C. Zhang, S. Ranjan, M. Nandwana, Q. Zhang, A. Misra, G. Liu, F. Kelly, and J.H., "Joint information from nonlinear features for spoofing detection," *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[7] A. Gomez-Alanis, A.M. Peinado, J.A. Gonzalez, and A.M. Gomez, "Performance evaluation of front- and back-end techniques for ASV spoofing detection systems based on deep features," *Proc. Iberspeech*, 2018.

[8] Xiang Wu, Ran He, Zhenan Sun and Tieniu Tan, "A Light CNN for Deep Face Representation with Noisy Labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[9] N. Ballas, L. Yao, C. Pal, and A. Courville, "Delving Deeper into Convolutional Networks for Learning Video Representations," *Proc. International Conference on Learning Representations (ICLR)*, 2016.

[10] Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. Plumbley, "Convolutional Gated Recurrent Neural Network Incorporating Spatial Features for Audio Tagging," *Proc. International Joint Conference on Neural Networks (IJCNN)*, 2017.

[11] J. Wang, and X. Hu, "Gated Recurrent Convolutional Neural Network for OCR," *Proc. Neural Information Processing System (NIPS)*, 2017.

[12] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," *Proc. Interspeech*, 2015.

[13] H. Delgado, M. Todisco, Md Sahidullah, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamigishi, "The ASVspoof 2017 Challenge: Asessing the Limits of Replay Spoofing Attack Detection," *Proc. Interspeech*, 2017.

[14] H. Delgado, M. Todisco, Md Sahidullah, N. Evans, T. Kinnunen, K. A. Lee, J. Yamigishi, et al. (2019, March). ASVspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge. [Online] Available: http://www.asvspoof.org

[15] A. Gomez-Alanis, A.M. Peinado, J.A. Gonzalez, and A.M. Gomez, "A Deep Identity Representation for Noise Robust Spoofing Detection," *Proc. Interspeech*, 2018.

[16] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," *Proc. Interspeech*, 2017.

[17] M. Witkowski, S. Zacprzak, P. Zelasko, K. Kowalczyk, and J. Galka, "Audio Replay Attack Detection Using High-Frequency Features," *Proc. Interspeech*, 2017.

[18] Y. Qian, N. Chen, H. Dinkel, and Z. Wu, "Deep Feature Engineering for Noise Robust Spoofing Detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1942–1955, 2017.

[19] Y. Bengio, P. Simard, and P. Frascani, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, pp. 157–166, 1994.

[20] C. Zhang, C. Yu, and J. H. L. Hansen, "An investigation of Deep-Learning Frameworks for Speaker Verification Antispoofing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, pp. 684–694, 2017.

[21] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6890*, 2014.

[22] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech and Language*, vol. 45, pp. 516–535, 2017.

[23] H. Muckenhirn, P. Korshunov, M. Magimai-Doss, and S. Marcel, "Long-Term Spectral Statistics for Voice Presentation Attack Detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2098–2111, 2017.

[24] Y. Qian, N. Chen, and K. Yu, "Deep features for automatic spoofing detection," *Speech Communication*, vol. 85, pp. 43–52, 2016.

[25] R. Font, J. M. Espin, and M. J. Cano, "Experimental analysis of features for replay attack detection–Results on the ASVspoof 2017 Challenge," *Proc. Interspeech*, 2017.

[26] K. Sriskandaraja, V. Sethu, and E. Ambikairajah, "Deep Siamese Architecture Based Replay Detection for Secure Voice Biometric," *Proc. Interspeech*, 2018.

[27] DFKI's Language Technology Lab and Multimodal Speech Processing, MMCI (2019, March). The Mary Text-to-Speech System (MaryTTS). [Online] Available: http://mary.dfki.de

[28] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," *Proc. Interspeech*, 2015.