Light Convolutional Neural Network with Feature Genuinization for Detection of Synthetic Speech Attacks

Zhenzong Wu[1], *Rohan Kumar Das*[1,*], *Jichen Yang*[1,*] *and Haizhou Li*[1,2]

[1]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2]Kriston AI Lab, China

wuzhenzong@u.nus.edu.sg, {rohankd, eleyji, haizhou.li}@nus.edu.sg

Abstract

Modern text-to-speech (TTS) and voice conversion (VC) systems produce natural sounding speech that questions the security of automatic speaker verification (ASV). This makes detection of such synthetic speech very important to safeguard ASV systems from unauthorized access. Most of the existing spoofing countermeasures perform well when the nature of the attacks is made known to the system during training. However, their performance degrades in face of unseen nature of attacks. In comparison to the synthetic speech created by a wide range of TTS and VC methods, genuine speech has a more consistent distribution. We believe that the difference between the distribution of synthetic and genuine speech is an important discriminative feature between the two classes. In this regard, we propose a novel method referred to as feature genuinization that learns a transformer with convolutional neural network (CNN) using the characteristics of only genuine speech. We then use this genuinization transformer with a light CNN classifier. The ASVspoof 2019 logical access corpus is used to evaluate the proposed method. The studies show that the proposed feature genuinization based LCNN system outperforms other state-of-the-art spoofing countermeasures, depicting its effectiveness for detection of synthetic speech attacks.

Index Terms: Feature genuinization, synthetic speech detection, ASVspoof 2019, logical access attacks

1. Introduction

In the recent years, automatic speaker verification (ASV) systems are deployed in different real-world applications [1–3]. These systems are exposed to spoofing attacks for unauthorized access, hence detection of such attacks attracts much attention [4, 5]. Various spoofing attacks are broadly classified into replay, impersonation, voice conversion (VC) and text-to-speech synthesis (TTS) attacks [6]. The latest progress in VC and TTS systems can produce perceptually natural sounding speech, which poises a threat to fool the ASV systems [7–9].

The research on spoofing countermeasures grew in the last decade since the inception of ASVspoof[1] challenge series. The challenge provided a platform to the researchers across different domains to explore fake speech detection using a common benchmarked corpus [10, 11]. Its recent edition ASVspoof 2019 is devoted to detection of both synthetic and replay speech with two subtasks [12]. The logical access track focuses on detection of synthetic speech created using state-of-the-art VC and TTS systems, which is the focus of this paper.

The explorations on spoofing attack detection cover two directions from the perspective of a detection task. The spoofing countermeasures either focus on novel front-end features or effective classifiers. Some of the former studies focused on the importance of robust features such as cochlear filter cepstral coefficient and instantaneous frequency (CFCCIF) [13], linear frequency cepstral coefficients (LFCC), subband spectral flux coefficients and spectral centroid frequency coefficients [14]. Later, the long-term constant-Q transform (CQT) based constant-Q cepstral coefficients (CQCC) proved to be one of the strong front-ends for synthetic speech detection [15]. The recent explorations with features derived from CQT are also found to effective for spoof detection [16–18].

With the advent of deep learning methods, robust classifiers are investigated for detection of spoofing attacks. Some of these include end-to-end systems with light convolutional neural networks (LCNN) [19, 20], squeeze excitation and residual networks [21, 22]. The end-to-end systems have much difference with the works that focus on novel features. The former are data driven deep learning methods, while the latter emphasize on hand-crafted feature, which require prior knowledge. Further, we note that the same neural network based system can perform differently for a range of features [19]. Therefore, a robust spoofing countermeasure is required to have a strong feature extractor that captures the discriminative artifacts along with an effective classifier.

The synthetic speech attacks can be created with a wide range of TTS and VC algorithms [6]. In general, spoofing countermeasures do not handle synthetic speech from unseen sources because of lack of generalization ability [23]. We note that genuine examples have a comparatively lower variance than synthetic speech. We believe that the consistent characteristics of genuine speech set genuine speech apart from a variety of different synthetic speech. A recent study using temporal domain information shows that spoofing detection can be improved by modifying the probability mass function of spoofed speech close to that of the genuine speech [24]. This process is termed as genuinization, which is found to be effective when applied to both train and test examples for synthetic speech detection.

In a similar direction, we hypothesize that, if we are able to derive a model that fits well the distribution of the genuine speech, such a model will take genuine speech as the input and generate genuine speech as the output following the same distribution of the genuine speech. However, when the model takes spoof speech as input, it will generate very different output, that amplifies the difference to genuine speech. With this hypothesis, we propose to derive a model using the genuine speech features with convolutional neural network (CNN) that is referred to as genuinization transformer. Further, the process is referred

[0]*Corresponding Author

[1]http://www.asvspoof.org/

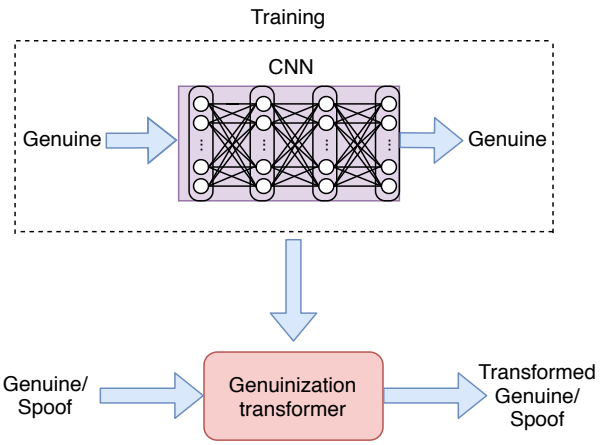

Figure 1: *The block diagram of feature genuinization process.*

to as feature genuinization as a given feature representation is projected on a domain learned using the genuine features. The genuinization transformer is then used together with an LCNN system for detection of synthetic speech attacks.

The rest of the paper is organized as follows. Section 2 introduces the details of proposed feature genuinization. Section 3 describes the feature genuinization based LCNN system for detection of spoofing attacks. The experiments and their results with discussion are reported in Section 4 and Section 5, respectively. Finally, the paper is concluded in Section 6.

2. Feature Genuinization

We aim to learn a transformer that does not change the characteristics of genuine speech features, whereas it projects spoof speech to a different output, maximizing the difference between genuine and spoof speech. Figure 1 shows the block diagram of the proposed feature genuinization process. It can be observed that there are two stages of the process. The first stage basically focuses on training a feature genuinization transformer using the characteristics features derived from only genuine speech. During the second stage, this trained feature genuinization transformer is used to convert any given features that enhances the discrimination of genuine and spoof speech.

The CNN based architectures have shown their effectiveness in the field of anti-spoofing research [25]. In this regard, we use CNN for training the genuinization transformer as shown in Figure 1. The detailed architecture of the CNN used in this framework can be seen from Figure 2. It can be observed that the functionality of the proposed genuinization transformer is similar to that of an autoencoder. However, the output of genuinization transformer is considered as the final transformer result. In addition, we apply a full convolutional layer and therefore, there is no fully connected layers in the transformer. This can thereby force the network to focus on the temporal correlation between the input signal and the whole stratification process. Further, it reduces the number of training parameters, which significantly results in less training period.

A study in [26] shows that it is a good practice to use strided convolution rather than pooling to downsample as it allows the network to learn its own pooling function. Therefore, we use this method during the training of genuinization transformer. In addition, batchNorm2d and leaky rectified linear unit (ReLU) activation function are used in the training because they can

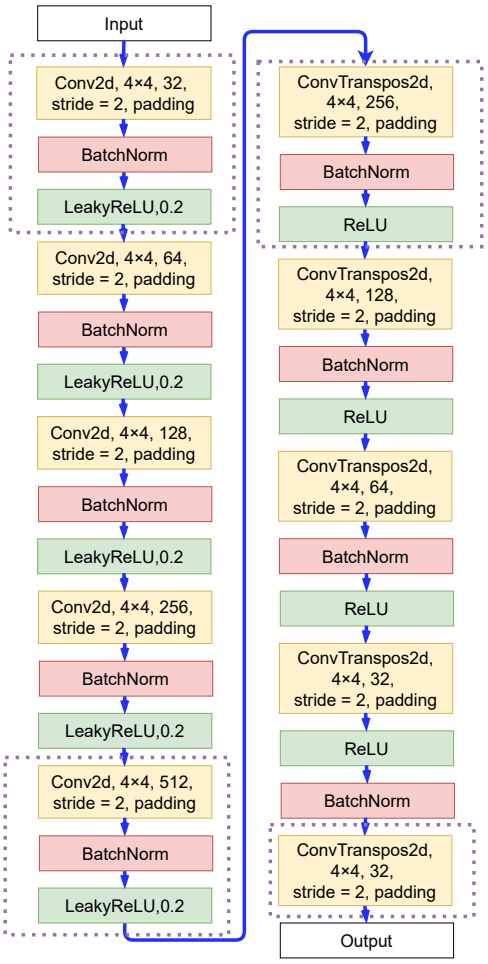

Figure 2: *The architecture of genuinization transformer.*

promote healthy gradient flow, which is critical for the learning process.

The architecture of proposed genuinization transformer shown in Figure 2 consists of two functionalities: encoding and decoding. During the encoding phase, the input signal is compressed through a number of strided convolutional layers, and then the convolution result is obtained by leaky ReLU. In the decoding phase, the encoding process is reversed by deconvolution, and then by ReLU. In this way, the transformer works as an autoencoder that learns the characteristics of genuine speech [27]. As a result of this, it amplifies the discrimination of genuine and spoof speech in the transformed domain.

Once the genuinization transformer is trained, it can be used to transform any given genuine/spoofed features to a transformed domain that is learned using the only genuine feature characteristics. This novel way of transforming the feature is referred as feature genuinization as mentioned earlier. Next, we discuss about the LCNN system using the feature generalization for detection of spoofing attacks.

3. LCNN with Feature Genuinization

Various deep learning systems have shown their effectiveness for spoofing attack detection [19, 21, 22, 28, 29]. Therefore, we plan to use the proposed feature genuinization with a deep learning system. The LCNN is one of the strongest systems

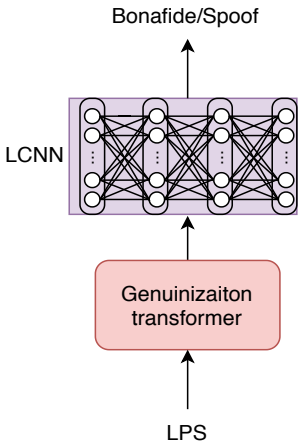

Figure 3: *Block diagram of the proposed feature genuinization based LCNN system.*

that has proven to be useful for its compactness and efficacy for anti-spoofing [19, 30]. In this work, we use LCNN based system with the transformed features obtained using genuinization transformer.

Figure 3 shows the block diagram of the proposed feature genuinization based LCNN system. We consider log power spectrum (LPS) of a given speech as the input feature to the genuinization transformer. It transforms the given input LPS to a genuinized feature, which is an input to the LCNN. During training, the training data and their corresponding label information is fed to the LCNN system. Once the training is completed, the detection result for a given input to the system can be obtained to identify the spoofing attacks.

We used Max-Feature-Map (MFM) activation function instead of commonly used ReLU function for the LCNN system similar to that in [19]. The main advantage of MFM is that it can learn compact features instead of sparse high-dimensional ones like ReLU. Further, MFM resorts to max function to suppress the activations of a small number of neurons so that MFM based CNN models are light and robust. Therefore, these are applied to reduce the dimensionality of the output and obtain more discriminative feature maps.

4. Experiments

In this section, we discuss the database and experimental setup for the studies.

4.1. Database

We consider the ASVspoof 2019 logical access corpus[2] for the studies of synthetic speech detection in this work [12, 31]. The corpus has three partitions, which are train, development and evaluation set. The genuine examples of the ASVspoof 2019 corpus are part of VCTK[3] database, which is a standard corpus for speech synthesis. It contains 107 speakers data that includes 46 male and 61 female speakers. It is to be noted that there is no overlap of speakers across different subsets. The synthetic speech attacks for the development set are created with two VC and four TTS state-of-the-art methods. However, the spoofed examples of evaluation set are derived from unseen methods.

[2]https://datashare.is.ed.ac.uk/handle/10283/3336
[3]http://dx.doi.org/10.7488/ds/1994
[4]https://pytorch.org

Table 1: *Summary of ASVspoof 2019 logical access corpus.*

Subset	#Male	#Female	#Bonafide	#Spoofed
Train	8	12	2,580	22,800
Development	4	6	2,548	22,296
Evaluation	21	27	7,355	63,882

The ASVspoof 2019 uses an ASV-centric metric given by tandem detection cost function (t-DCF) as the primary metric and equal error rate (EER) as a secondary metric for benchmarking the systems [31, 32]. We considered the scores of ASV system given along with the ASVspoof 2019 logical access corpus to combine with that from spoofing countermeasure system for computation of t-DCF measure. Table 1 presents a summary of the ASVspoof 2019 logical access corpus.

4.2. Experimental Setup

The long-term CQT based features are found to capture useful artifacts for spoofing attack detection [33]. Therefore, we use LPS derived from CQT as the input feature for the studies. The parameters for CQT computation are set based on following those in [15]. The number of octaves and frequency bins in every octaves are set at 9 and 96, respectively. In addition, the static dimension of LPS is 863. For LPS extraction from CQT, the length of every file is set as 256 frames by either padding and cropping. In particular, the examples with frame-length over than 256 frames are truncated, while the examples with frame-length less than 256 frames are filled with the last frame value. Thus, the we have an input feature of 863×256 for every example.

During training of the LCNN system, an additional batch normalization step is used after max pooling layer to increase the stability and convergence speed. As such models are prone to overfitting, we consider dropout and weight decay methods to avoid such issue. The dropout is used for fully connected layers with the ratio 0.4 and the weight decay is set to 2×10^{-4}. In addition, the parameters like number of layers and nodes are optimized on the development set. The proposed feature genuinization based LCNN system is implemented using PyTorch[4] toolkit.

5. Results and Discussion

The proposed system is a pipeline with a feature genuinization followed by LCNN. We compare the proposed system with LCNN baseline without feature genuinization. Further, we also consider the two baseline spoofing countermeasures of ASVspoof 2019 challenge. They are based on CQCC and LFCC features with Gaussian mixture model (GMM) classifier [12, 31].

Table 2 shows the results of proposed feature genuinization based LCNN system, that we refer as FG-LCNN, on ASVspoof 2019 logical access corpus and its comparison to the baseline systems. We observe that introducing feature genuinization module in the baseline LCNN system improves the detection of spoofing attacks. While the results on the development set are close, the improvement from the proposed system is evident from the results on the evaluation set, which contains more challenging spoofing attacks of unseen nature. This confirms our hypothesis to use a feature genuinization model exploiting the characteristics of genuine speech. Further, we find that the performance of the proposed system is much better than the two ASVspoof 2019 challenge baselines.

Table 2: *Performance of proposed feature genuinization based LCNN (FG-LCNN) and its comparison to baseline systems on ASVspoof 2019 logical access corpus.*

System	Development Set		Evaluation Set	
	t-DCF	EER (%)	t-DCF	EER (%)
Baseline: LCNN	0.002	0.080	0.111	4.448
FG-LCNN	**0.000**	**0.002**	**0.102**	**4.070**
ASVspoof 2019 Baseline [12]				
CQCC-GMM	0.0123	0.43	0.2366	9.57
LFCC-GMM	0.0663	2.71	0.2116	8.09

Table 3: *Performance of proposed feature genuinization based LCNN (FG-LCNN) and its comparison to feature spoofing based LCNN (FS-LCNN) contrast system on ASVspoof 2019 logical access corpus evaluation set.*

System	t-DCF	EER (%)
Baseline: LCNN	0.111	4.448
Prposed: FG-LCNN	**0.102**	**4.070**
Contrast: FS-LCNN	0.138	4.860

We further perform a justification experiment for validation of our proposed method. The idea behind feature genuinization process is based on the assumption that the genuine speech examples are considered to be less varied than the synthetic speech attacks created using a wide range of methods. We perform a contrast experiment, where we learn a transformation model using CNN by only considering the spoofed speech features. We refer this process as feature spoofing and the model as spoofing transformer, similar to the case of our proposed method. This spoofing transformer is then used to transform any given feature of genuine or spoofed speech to another domain, which is then used in the LCNN system pipeline, that we call FS-LCNN. The rest of the experimental setup remains the same to that our proposed method.

Table 3 shows the performance comparison of the FS-LCNN contrast system with our proposed FG-LCNN and the baseline LCNN system. We consider the results of evaluation set for the comparison as the results of development set can show very accurate detection of synthetic speech attacks. We find that the FS-LCNN contrast system does not perform better than our proposed FG-LCNN system, but rather degrades from the baseline LCNN system. This further strengthens our proposed idea of using feature genuinization process with LCNN system for detection of spoofing attacks.

We are now interested in comparing the proposed system to various single system based results available of ASVspoof 2019 logical access corpus. In this regard, we consider some of the well performing front-end as well as back-ends that have shown their effectiveness for spoofing attack detection in ASVspoof 2019 challenge. Some of the those front-ends are zero time windowing cepstral coefficients (ZTWCC), single frequency cepstral coefficients (SFCC) and instantaneous frequency cepstral coefficients (IFCC) that are implemented with GMM based classifier [34]. Further, deep learning based classifiers such as deep neural network (DNN), ResNet and LCNN are used for detection of spoofing attacks using front-ends like mel frequency cepstral coefficient (MFCC), constant-Q statistics-plus-principal information coefficients (CQSPIC), CQCC, LFCC, LPS of discrete Fourier transform (DFT) and fast Fourier transform (FFT) in ASVspoof 2019 challenge [19,35–37]. We report the respective system results from their published works for the comparison on the evaluation set of ASVspoof 2019 logical access corpus.

Table 4: *Performance comparison of the proposed feature genuinization based LCNN system to some known single systems on ASVspoof 2019 logical access evaluation set.*

System	t-DCF	EER (%)
ZTWCC-GMM [34]	0.141	6.13
IFCC-GMM [34]	0.357	15.59
SFFCC-GMM [34]	0.323	13.97
CQCC-DNN [35]	0.308	12.79
LFCC-DNN [35]	0.234	9.65
MFCC-ResNet [36]	0.204	9.33
LPS-DFT-ResNet [36]	0.274	9.68
CQCC-ResNet [36]	0.217	7.69
CQSPIC-DNN [35]	0.183	7.81
CQSPIC-GMM [35]	0.164	7.74
LFCC-LCNN [19]	0.100	5.06
LPS-FFT-LCNN [19]	0.103	4.53
Proposed: FG-LCNN	**0.102**	**4.07**

Table 4 reports the performance comparison of the proposed FG-LCNN system to some of the single systems reported in ASVspoof 2019 challenge discussed above. It is observed that the LCNN based systems represent the best performing single system, that justifies its use as the baseline LCNN in this work. Further, the effectiveness of proposed feature genuinization is evident on using it with the LCNN system, which outperforms other reported single systems in terms of EER on ASVspoof 2019 logical access corpus.

6. Conclusion

This work proposes a novel feature genuinization based LCNN system for detection of synthetic speech attacks. The characteristics of genuine speech are exploited to learn a model using CNN. It transforms a genuine feature distribution more close to that of the genuine speech, whereas leads to a different output for features of spoof speech, thereby maximizing their difference. The transformed features are then used with an LCNN system. The studies conducted on ASVspoof 2019 logical access corpus show the effectiveness of the feature genuinization based LCNN system for detecting synthetic speech attacks. The comparison of the proposed system to various state-of-the-art spoofing countermeasures showcases it as one of the strong single anti-spoofing system. The future work will focus on extending the studies to replay attack detection.

7. Acknowledgements

This research work is partially supported by Programmatic Grant No. A1687b0033 from the Singapore Government's Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain), Human-Robot Interaction Phase 1 (Grant No. 192 25 00054) by the National Research Foundation, Prime Minister's Office, Singapore under the National Robotics Programme. This work is also part of a collaboration with Kriston AI Lab, China in 2019.

8. References

[1] K. A. Lee, B. Ma, and H. Li, "Speaker verification makes its debut in smartphone," in *SLTC Newsletter*, February 2013.

[2] R. K. Das, S. Jelil, and S. R. M. Prasanna, "Development of multi-level speech based person authentication system," *Journal of Signal Processing Systems*, vol. 88, no. 3, pp. 259–271, Sep 2017.

[3] S. Jelil, A. Shrivastava, R. K. Das, S. R. M. Prasanna, and R. Sinha, "SpeechMarker: A voice based multi-level attendance application," in *Interspeech 2019*, 2019, pp. 3665–3666.

[4] Z. Wu and H. Li, "On the study of replay and voice conversion attacks to text-dependent speaker verification," *Multimedia Tools and Applications*, vol. 75, no. 9, pp. 5311–5327, May 2016.

[5] R. K. Das, X. Tian, T. Kinnunen, and H. Li, "The attacker's perspective on automatic speaker verification: An overview," in *Interspeech 2020*, 2020. [Online]. Available: https://arxiv.org/abs/2004.08849

[6] 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.

[7] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Odyssey 2018*, 2018, pp. 195–202.

[8] T. Kinnunen, J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, and Z. Ling, "A spoofing benchmark for the 2018 voice conversion challenge: Leveraging from spoofing countermeasures for speech artifact assessment," in *Odyssey 2018*, 2018, pp. 187–194.

[9] J. Lorenzo-Trueba, F. Fang, X. Wang, I. Echizen, J. Yamagishi, and T. Kinnunen, "Can we steal your vocal identity from the internet?: Initial investigation of cloning Obama's voice using GAN, WaveNet and low-quality found data," in *Odyssey 2018*, 2018, pp. 240–247.

[10] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *Interspeech 2013*, 2013, pp. 925–929.

[11] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado, "ASVspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, June 2017.

[12] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Interspeech 2019*, 2019, pp. 1008–1012.

[13] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *Interspeech 2015*, 2015, pp. 2062–2066.

[14] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Interspeech 2015*, 2015, pp. 2087–2091.

[15] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Odyssey 2016*, 2016, pp. 283–290.

[16] J. Yang, R. K. Das, and N. Zhou, "Extraction of octave spectra information for spoofing attack detection," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 27, pp. 2373–2384, 2019.

[17] J. Yang, R. K. Das, and H. Li, "Significance of subband features for synthetic speech decetion," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2160–2170, 2020.

[18] J. Yang and R. K. Das, "Long-term high frequency features for synthetic speech detection," *Digital Signal Processing*, vol. 97, p. 102622, 2020.

[19] G. Lavrentyva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlos, "STC antispoofing systems for the ASVspoof2019 challenge," in *Interspeech 2019*, Graz, Austria, 2019, pp. 1033–1037.

[20] Y. Yang, H. Wang, H. Dinkel, Z. Chen, S. Wang, Y. Qian, and K. Yu, "The SJTU robust anti-spoofing system for the ASVspoof 2019 challenge," in *Interspeech 2019*, 2019, pp. 1038–1042.

[21] C.-I. Lai, N. Chen, J. Villaba, and N. Dehak, "ASSERT:Anti-spoofing with squeeze-excitation and residual networks," in *Interspeech*, Graz, Austria, 2019, pp. 1013–1017.

[22] J. Monteiro and J. Alam, "Development of voice spoofing detection systems for 2019 edition of automatic speaker verification and countermeasures challenge," in *IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop 2019*, 2019, pp. 1003–1010.

[23] R. K. Das, J. Yang, and H. Li, "Assessing the scope of generalized countermeasures for anti-spoofing," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2020*, 2020, pp. 6589–6593.

[24] I. Lapidot and J.-F. Bonastre, "Effects of wavform PMF on anti-spoofing detection," in *Interspeech 2019*, 2019, pp. 2853–2857.

[25] C. Zhang, S. Ranjan, M. K. Nandwana, Q. Zhang, A. Misra, G. Liu, F. Kelly, and J. H. L. Hansen, "Joint information from nonlinar and linear features for spoofing detection: an i-vector based approach," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5035–5038, 2016.

[26] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *CoRR*, vol. abs/1511.06434, 2015. [Online]. Available: http://arxiv.org/abs/1511.06434

[27] J. Gehring, Y. Miao, F. Metze, and A. Waibel, "Extracting deep bottleneck features using stacked auto-encoders," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2013*, 2013, pp. 3377–3381.

[28] J. weon Jung, H. jin Shim, H.-S. Heo, and H.-J. Yu, "Replay attack detection with complementary high-resolution information using end-to-end DNN for the ASVspoof 2019 challenge," in *Proc. Interspeech 2019*, 2019, pp. 1083–1087.

[29] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection," in *Proc. Interspeech 2019*, 2019, pp. 1068–1072.

[30] X. Wu, R. He, Z. Sun, and T. 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.

[31] "ASVspoof 2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," 2019.

[32] T. Kinnunen, K. A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," in *Odyssey 2018*, 2018, pp. 312–319.

[33] R. K. Das, J. Yang, and H. Li, "Long range acoustic features for spoofed speech detection," in *Interspeech 2019*, 2019, pp. 1058–1062.

[34] K. N. R. K. R. Alluri and A. K. Vupala, "IIIT-H spoofing countermeasures for automatic speaker verification spoofing and countermesures challenge 2019," in *Interspeech 2019*, Graz, Austria, 2019, pp. 1043–1047.

[35] R. K. Das, J. Yang, and H. Li, "Long range acoustic and deep features perspective on ASVspoof 2019," in *Automatic Speech Recognition and Understanding (ASRU) Workshop*, 2019, pp. 1018–1025.

[36] M. Alzanto, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," in *Interspeech 2019*, Graz, Austria, 2019, pp. 1078–1082.

[37] J. Yang and R. K. Das, "Improving anti-spoofing with octave spectrum and short-term spectral statistics information," *Applied Acoustics*, vol. 157, 2020.