INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Generalized Spoofing Detection Inspired from Audio Generation Artifacts

Yang Gao[1,⋆] , *Tyler Vuong*[1], *Mahsa Elyasi*[2], *Gaurav Bharaj*[2], *Rita Singh*[1]

[1]Carnegie Mellon University, USA
[2]AI Foundation, USA

{yanggao,tvuong,rsingh}@andrew.cmu.edu {gaurav,masha}@aifoundation.com

Abstract

State-of-the-art methods for audio generation suffer from fingerprint artifacts and repeated inconsistencies across temporal and spectral domains. Such artifacts could be well captured by the frequency domain analysis over the spectrogram. Thus, we propose a novel use of long-range spectro-temporal modulation feature – 2D DCT over log-Mel spectrogram for the audio deepfake detection. We show that this feature works better than log-Mel spectrogram, CQCC, MFCC, as a suitable candidate to capture such artifacts. We employ spectrum augmentation and feature normalization to decrease overfitting and bridge the gap between training and test dataset along with this novel feature introduction. We developed a CNN-based baseline that achieved a 0.0849 t-DCF and outperformed the previously top single systems reported in the ASVspoof 2019 challenge. Finally, by combining our baseline with our proposed 2D DCT spectro-temporal feature, we decrease the t-DCF score down by 14% to 0.0737, making it a state-of-the-art system for spoofing detection. Furthermore, we evaluate our model using two external datasets, showing the proposed feature's generalization ability. We also provide analysis and ablation studies for our proposed feature and results.

Index Terms: ASVspoof challenge, spoofing detection, 2D-DCT, modulation feature

1. Introduction

Audio deepfakes use deep learning and machine learning algorithms to generate or manipulate audio content with an intent to deceive. Such audio deepfakes are especially dangerous due to their innate embedding of biometrics, used in speech-based identity verification systems. State-of-the-art audio deepfake methods rely on voice conversion, text-to-speech synthesis, generative models, and neural vocoders [1, 2, 3, 4, 5]. With these advances, the quality of deepfakes has significantly improved, making them a pernicious means to commit a wide variety of fraudulent activities – identity theft and misinformation spread by untrained bad actors. Such techniques even outperform professional human impersonators and threaten automatic speaker verification (ASV) systems [6].

For better spoof attack detection in ASV systems, ASV spoof challenges [7, 8, 9, 10, 11] have been created. In such challenges, the logical access (LA) consists of synthetically spoofed audio, which uses conventional signal processing and generative techniques that [12, 13, 14] propose the use of feature selection (e.g., Constant Q cepstral coefficients [15], MFCC, spectrogram, etc.) to search for the best features for spoof detection. However, these features have been developed for generic tasks, such as automatic speech recognition (ASR) and sound-based event detection, etc. They may not capture the fundamental differences between real and fake speech well. Further, the choice of feature selection can be influenced by audio datasets and is inconsistent. For better generalization, as noted in [6], unlike real speech, machine-generated speech consists of signature artifacts that can be leveraged for spoof detection. They propose a lightweight model with several human speech characteristics features and achieve comparably higher accuracy.

In computer vision, generative adversarial networks (GANs) [16] are a popular choice for image generation. Such methods have associated "fingerprint" [17] and signal-domain [18] artifacts that can be leveraged for detection and attribution studies. In speech synthesis, generative methods are used for feature learning from input linguistic features, while neural vocoders convert generated features into waveform outputs. Here, the audio is usually synthesized in frames or blocks of frames and has no cross-frame temporal consistency. This can lead to temporal modulation artifacts. Additionally, such methods are typically trained with element-wise mean-square-error losses in the Mel-Spectrogram domain [4, 19] and do not account for cross-frame consistency. Furthermore, speech is mainly encoded in the frequency ranges 0-4 kHz of auditory perception (based on the learning principles). There are associated artifacts with the generated outputs [20], especially at high frequencies [21].

Based on these observations for feature artifacts, we propose using long-range frequency analysis on log-Mel-Spectrogram (in feature domain) for spoof detection. Since 2D-DCT features capture repeated patterns/artifacts by analyzing the joint spectro-temporal modulation frequencies, we introduce the novel use of global 2D-DCT on log-Mel-Spectrograms, a long-range spectro-temporal feature, to capture audio deepfake artifacts. The spoof detection convolutional neural network (CNN) classifier that operates on log-Mel Spectrum consists of the features with limited receptive fields and focuses on finding local short/medium time patterns/correlations in the input audio. The proposed global 2D-DCT feature essentially forces the CNN classifier to learn from the input audio's long-term/global modulation patterns. These 2D-DCT features correspond to the long-term spectro-temporal modulations rather than localized ones. Therefore, we call this proposed feature *global modulation* (Global M) feature. We show that the proposed feature detects deepfakes at a higher accuracy compared with the standard log-Mel features and could compensate our strongest baseline model to improve the overall detection performance further.

To summarize, in this paper, we compare the proposed global modulation features with traditional features such as MFCC, log-Mel, and Constant Q cepstral coefficients (CQCC) and present the following novel contributions:

1. We propose a novel long-range spectro-temporal feature – global modulation feature, for audio deepfake detection.
2. We further implement SpecAugment [22] and feature normalization to reduce over-fitting and bridge the gap between training and test dataset from unseen attacks.
3. The resulting baseline system achieves the best tandem detection cost function (t-DCF) scores as single systems according to [10]. Furthermore, our proposed feature can

[⋆] Work performed during internship at AI Foundation

http://dx.doi.org/10.21437/Interspeech.2021-1705

compensate for this strong baseline to bring the t-DCF and the equal error rate (EER) down and achieve state-of-the-art performance on the ASVspoof challenge 2019 logical access (LA).

Finally, the proposed global modulation feature also achieves a higher accuracy on general tasks, such as speaker verification, shown in Section 4.3.

2. Related works

2.1. Audio deepfake detection

The ASVspoof challenges [7, 9, 11] have raised efforts in fake speech spoofing attack countermeasures on ASV systems. Previous studies on anti-spoofing attacks on ASV systems and synthetic speech detection evaluate various features [14, 23] and deep learning models [24] for detection performance. However, with the fast evolution of deepfake techniques, developing a detection system that is not constrained by the training data and can accurately detect new spoofed data generated from different or unseen deepfake algorithms is still challenging.

In the ASVspoof challenge 2019 dataset, the logical access (LA) contains fake audio generated by multiple methods as in Table 5. As reported in [10], the best single system for LA data achieves a t-DCF metric [10] score of about 0.13 and an EER score of 5%. The top-3 primary system (a weighted voting of multiple systems) achieves a t-DCF score of less than 0.1 and an EER of smaller than 3%.

There are also datasets for audio deepfake detection like FoR dataset [25] and RTVCspoof dataset created using neural generation models as in [26]. In our work, we also use these external datasets effectively as unseen test attacks to our proposed detection system.

2.2. Modulation features

The modulation feature captures the longer time patterns in the signal, which are often ignored in MSE-based generation [27, 19]. Not only inspired by the generation artifacts, moreover, but the proposed feature is also global modulation feature that analyzes the joint long-range spectro-temporal modulation information.

In [28], the importance of spectral and temporal modulation content of the auditory spectrogram is discussed. Here, filter banks selecting different spectro-temporal modulation parameters range from slow to fast rates temporally and from narrow to broad scales spectrally. The spectro-temporal receptive fields (STRFs) of these filters are related to human perception's auditory system. We also note that, from a physiological point of view, neurons in the primary auditory cortex of mammals are explicitly tuned to spectro-temporal patterns, e.g., spectro-temporal features, [29]. Suthokumar et al. [30] analyze the temporal modulation by performing FFT analysis in each sub-band, and show the effectiveness of temporal dynamics for replay spoofing detection.

However, in previous studies, 2D-DCT was only used to calculate **local** spectro-temporal modulation, such as for robust automatic speech recognition (ASR) [31]. Medium range modulation features were discussed in [32, 33] and long-range modulation was proposed in [34] – but both only for the temporal domain. Our **global** modulation feature combines spectral (as MFCC) and temporal modulation information for better long-range feature modeling. To the best of our knowledge, such a long-range feature modeling has not been carried out in previous studies in speech.

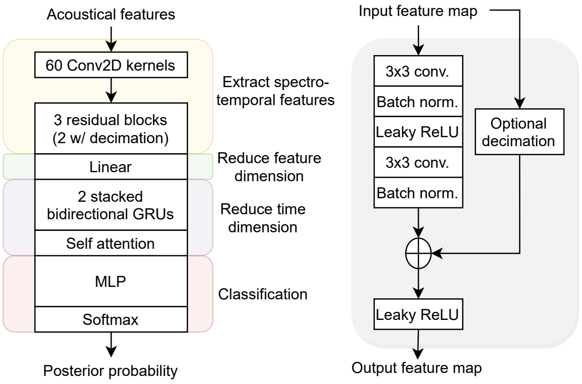

Figure 1: *Block diagram of the baseline system (left) and the zoomed-in view of one residual block (right).*

3. Experiments

3.1. Baseline model

The baseline we use is a CNN-based model, similar to the baseline CNN model in [27]. As shown in Figure 1, the baseline model first consists of an initial convolutional layer followed by three residual blocks. Next, the output is passed through bidirectional Gated Recurrent Units (GRUs) and a self-attentive pooling layer. After temporal modeling and the self-attentive pooling, the feature vector is passed through a one-hidden-layer multi-layer perceptron (MLP) with two dimensions for the output. Finally, softmax is applied to obtain the prediction probability of genuine speech.

3.2. Proposed feature

The proposed feature is a simple and effective spectro-temporal feature: the 2D-DCT on log-Mel-spectrogram. This is actually similar to the computation of Mel-frequency cepstral coefficients (MFCC) with the difference that we are applying a 2-dimensional (2D) discrete cosine transform (DCT) globally on both the temporal dimension and frequency dimension of the log-Mel spectrogram. The detailed computation steps are described as following:

a) Employ the fast Fourier transform (FFT) to compute the spectrum $X(w)$ of $x(n)$.

b) Compute power spectrum $|X(w)|^2$ and obtain the Mel-spectrum M by applying a Mel-frequency filter bank.

c) Apply multi-dimensional discrete cosine transform (DCT) to log-Mel to obtain $dctn_M$.

d) (Optional) Apply $l1$-normalization or standardization normalization on the obtained $dctn_M$.

Figure 2 shows the proposed 2D-DCT features for different spoofing types. The 2D-DCT features are in log-scale. From the visualization, we can see the proposed feature could obtain the differences in their patterns across different spoofing types. A17 and A19 use signal processing methods to generate fake audios, and the proposed features of these two are similar to the bonafide. In contrast, other methods give more complex changes compared to the bonafide (real audio) type.

3.3. Implementation details

For experiments with conventional and proposed features, we verify the spoofing countermeasures in performance improvements. We use the detection model that is modified from the residual net

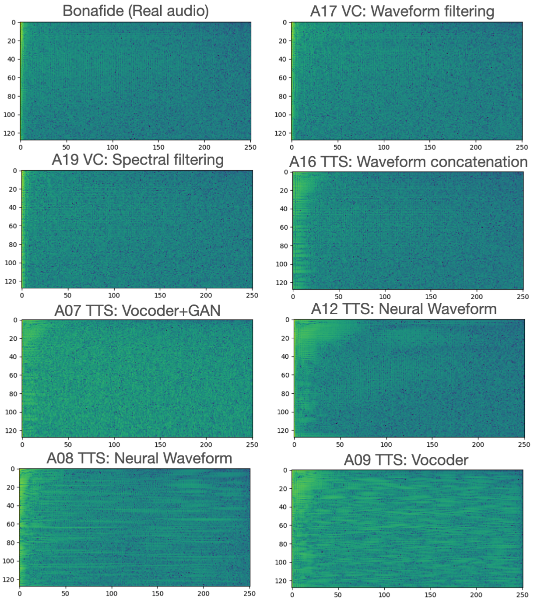

Figure 2: *Visualization of the proposed features averaged within different spoofing types. Vertical axis is from mel-filters domain as spectro-modulation axis, and horizontal axis is from time frames as temporal-modulation axis. (Best viewed zoomed in)*

Table 1: *SpecAugment (SA) and Normalization approaches*

Features	**t-DCF**	**EER (%)**
log-Mel (Baseline$_1$)	0.0902	6.551
log-Mel w/ SA (Baseline$_2$)	**0.0849**	**5.139**
2D-DCT of log-Mel (Global M)	0.2851	12.40
Normalized Global M	0.1358	6.852
Normalized Global M w/ SA (Ours)	**0.1387**	**6.325**
T32 (Best single system [10])	0.1239	4.92

architectures proposed in [24]. To evaluate the proposed features, this model is similar to our baseline model without the attention layer since the temporal information is already condensed into the global DCT domain. The audio sequences are cut or padded to 4 seconds, as the temporal range. The sampling rate is 16k, the FFT size is 1024, the window size is 512 and the hop size is 256, and the mel-filter number is 128. The details of the model implementation are in section 3.2 of [6].

Furthermore, we found the spectrum augmentation on the input features, and the normalization of the 2D-DCT features could improve the performance significantly, as shown in Table 1. We implemented the SpecAugment (SA) [22] approach on log-Mel-spectrogram with torchaudio. The randomly masking on the frequency channels and time steps of the spectrogram helps preventing overfitting and increases the model's performance [22]. For the SA on the proposed global modulation feature, a randomly zeroing-out manner is implemented to generate blank areas on both dimensions. This augmentation is only applied to the training data on the fly during training. For normalization on the 2D-DCT is applied using two approaches for comparison. Two normalization approaches, the l1-norm normalization and the mean/std standardization normalization, implemented using sklearn toolbox in PYTHON, achieve similar results. In contrast, the normalization does not help (much) for the other traditional features since the values are already in reasonable ranges and the l1-norm will break the spectral and temporal dynamics across the frames.

Table 2: *Single systems comparisons as ASV countermeasures*

	Countermeasure EER%		t-DCF	
Features	DEV	EVAL	DEV	EVAL
Aperiodic parameters (AP)	21.19	20.65	0.4374	0.4445
Spectral envelope (SP)	10.55	9.31	0.3520	0.2453
MFCC	7.14	11.64	0.1942	0.2663
CQCC	1.37	10.89	0.0407	0.2746
log-Mel spectrogram	0.48	9.39	0.0132	0.1954
Normalized Global M	0.23	6.85	0.0067	0.1358
Normalized Global M w/ SA	0.17	6.32	0.0043	0.1387

Table 3: *Weighted voting scores with different voting mechanisms*

	Global Modulation + Baseline$_1$		Global Modulation + Baseline$_2$	
Ratios	t-DCF	EER	t-DCF	EER
min	0.1306	7.098	0.1230	6.636
0.0	0.1397	6.325	0.1387	6.325
0.1	0.1207	5.92	0.1253	5.778
0.2	0.1063	**5.89**	0.1141	5.780
0.3	0.0984	5.90	0.1057	5.631
0.4	0.0923	5.98	0.0994	5.520
0.5	0.0883	6.07	0.0930	5.542
0.6	0.0867	6.17	**0.0890**	**5.301**
0.7	**0.0865**	6.27	0.1057	5.563
0.8	0.0870	6.35	0.1142	5.778
0.9	0.0875	6.45	0.1253	5.929
1.0	0.0902	6.55	0.0849	5.139
max	**0.0737**	**4.03**	**0.0864**	**4.216**

4. Results

4.1. Single systems and weighted voting scores

We evaluated the single system model taking in one type of feature and compared the proposed global modulation feature with the previously proposed feature aperiodic signal (AP), spectral envelope (SP) [6], and other conventional features such as MFCC, CQCC, and spectrogram. To have a fair comparison, the model is the same ResNet model as in Section 3.2 of [6] with the last layer's dimension change to facilitate the feature size difference. From the results in Table 2, we can see the proposed feature is significantly better in both the EER and the t-DCF scores than the other features.

We further evaluate the joint performance of our proposed feature with the strong baseline models. We use different voting mechanisms for the joint scores between the Global Modulation feature and the baseline models as following: For the prediction probability outputs of both systems, we weighted the prediction score using a ratio of 0.1 to 0.9. We use a max metric to keep the most confidence voting among the two systems, which gives us the best performance. In contrast, the min-metric keeps the lower confidence prediction of the two joint systems. From the results in Table 3, we can see the joint scores improve the overall countermeasure performance.

4.2. Audio type analysis

To evaluate the detection performance on different spoofing audio types, we do a comprehensive analysis on the t-DCF and EER scores for all spoofing audio types in the LA evaluation set, as shown in Table 5. The A17 type, generated with waveform filtering manipulations on the real audios, is visualized in Figure 2. It has a very similar modulation pattern to the bonafide audios and is the hardest type according to [11]. Our baselines and the proposed feature achieve top performance, compared to the EERs of single systems reported in [11]. And our joint system achieves one of the best three compared to all the other systems that use an ensemble of classifiers [10].

Table 4: *EERs of evaluation set for ASVspoof2019 LA for speaker verification*

	ASV EER%													
Spoofing ID	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	ALL
STFT	2.33	2.65	3.75	47.56	40.89	47.59	37.01	29.09	35.48	4.09	12.07	28.61	1.88	22.24
MFCC	7.12	5.08	8.12	39.76	28.99	49.01	33.81	19.04	41.39	9.08	18.00	16.47	2.09	15.99
AP	38.93	32.46	32.59	42.37	38.29	43.28	37.02	33.96	41.12	49.06	40.05	34.57	44.53	39.25
SP	50.97	49.94	40.07	49.75	49.25	52.04	52.30	51.03	51.74	51.99	41.49	46.16	45.78	42.08
Global M	1.45	8.01	8.35	31.97	32.85	38.92	20.64	14.10	28.22	2.91	23.93	27.79	1.11	18.69

Table 5: *Breakdown analysis of the performance on different Spoofing audio types*

	Info		Baseline$_1$		Baseline$_2$		Proposed feature		Joint w/ Baseline$_1$		Joint w/ Baseline$_2$	
ID	System	Details	t-DCF	EER	t-DCF	EER	t-DCF	EER	t-DCF	EER	t-DCF	EER
A07	TTS	Vocoder+GAN	0.0000	0.0000	0.0000	0.0000	0.0054	0.1799	0.0014	0.0407	0.0020	0.0645
A08	TTS	Neural waveform	0.0463	1.4901	0.0163	0.5297	0.0521	1.9727	0.0147	0.5297	0.0254	0.7911
A09	TTS	Vocoder	0.0015	0.0577	0.0003	0.0170	0.0093	0.2852	0.0028	0.0815	0.0035	0.1392
A10	TTS	Neural waveform	0.0084	0.3022	0.0058	0.2445	0.0417	1.3208	0.0080	0.2852	0.0164	0.5059
A11	TTS	Griffin lim	0.0102	0.3667	0.0072	0.2852	0.0407	1.3038	0.0083	0.2682	0.0152	0.4720
A12	TTS	Neural waveform	0.0041	0.1222	0.0020	0.0645	0.0635	1.9557	0.0090	0.2852	0.0193	0.6111
A13	TTS-VC	WC + waveform filtering	0.0029	0.0985	0.0003	0.0170	0.0650	2.0372	0.0113	0.3429	0.0218	0.6689
A14	TTS-VC	Vocoder	0.0079	0.2445	0.0037	0.1222	0.0270	0.8149	0.0069	0.2274	0.0095	0.3022
A15	TTS-VC	Neural waveform	0.0186	0.5942	0.0061	0.1799	0.0248	0.7911	0.0069	0.2275	0.0097	0.3259
A16	TTS	Waveform concatenation (WC)	0.0007	0.0407	0.0005	0.0169	0.0062	0.1867	0.0010	0.0407	0.0016	0.0578
A17	VC	Waveform filtering	0.9760	44.486	0.7670	26.538	0.9017	36.286	0.8004	28.324	0.6218	28.405
A18	VC	Vocoder	0.0061	0.2037	0.0098	0.3259	0.1985	6.1286	0.0201	0.6111	0.0602	1.7927
A19	VC	Spectral filtering	0.0040	0.1222	0.0051	0.1630	0.0151	0.5297	0.0050	0.1799	0.0058	0.2037

4.3. Speaker verification using the proposed features

To evaluate our proposed feature's effectiveness, we evaluate the feature under the automatic speaker verification scenario, as in [6]. The ASV model is trained with the ASVspoof 2019 data LA training set. We assign each spoofed utterance an identity that uniquely incorporates both speaker and attack. The 20 speakers and 6 types of attack in the ASVspoof2019 LA training set are combined into 120 "spoofed identities". With the bonafide audios, we have positive pairs, and negative pairs generated randomly in a balanced 1:1 ratio. The results are shown in Table 4. The proposed feature is compared with other features' results of [6]. Unlike AP and SP, the proposed 2D modulation feature is not only more powerful as in a detection model but also effective in the audio type and speaker verification tasks. This clearly shows the potential of this proposed feature for several applications.

5. Discussions

As the above results show, our proposed global modulation feature has a strong performance compared to other conventional features. We also test our best model's detection accuracy on the other external datasets FoR [25] and RTVCspoof collected in [26]. For each dataset, 200 fake and 200 real samples are selected randomly from their test sets. Our Global modulation feature model could also predict the class of the randomly selected test data with reasonable accuracy of 90% to 98%.

We also compared the global modulation feature on the high-frequency section of the log-Mel spectrogram compared with the low-frequency section. Consistent with [21], the high-frequency section gives higher detection performance compared to the low-frequency section, although still not as good as using the global information altogether. Finally, we compare a blocked version of the modulation feature with our proposed global modulation feature. We did a simple 2x2 division on the log-Mel spectrogram and computed the 2D-DCT features separately for each block. The resulting localized modulation features give a significantly lower detection performance of around 20% EER. This shows the importance of taking long-range frequency computation to obtain the global inconsistencies for the audio detection leanings. Interestingly, in [27], their proposed spectro-temporal receptive fields (STRFs) is a localized modulation feature. And in their experiments for the ASVspoof challenge, they concluded that 'the STRFs effectively reject distractor noise, but are by themselves not sufficient for discriminating real from synthetic speech'. Their results, in comparison, give another evidence for the importance of computing the modulation features globally.

Also, it needs to be noted that the eval results for each feature are averaged across the eval EERs and t-DCFs of multiple runnings' best validation models for the soundness of the scores. The best eval score we have from a single running may be lower (E.g. the best baseline we have has an EER of 4.03%). The t-DCF score is evaluated using the same metric as in [10].

6. Conclusions

In this paper, we propose a simple yet effective feature, the global modulation feature, inspired by the fake audios' artifacts. We show that this proposed feature could improve the strongest baseline we have to further increase the countermeasure system's detection performance for the ASV system. Furthermore, we use this proposed feature to train our own ASV system and show that it also works very well for speaker verification tasks. This shows the broader potentials of the proposed global modulation feature.

In future works, we could embrace more data augmentation approaches, e.g., adding noise, etc. Moreover, with the future-released evaluation plan from ASVspoof challenge 2021, we would also evaluate the proposed feature's robustness to channel variations and its performance with the physical access (PA) dataset in ASVspoof Challenges [7, 8, 10, 11].

7. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[2] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron," in *international conference on machine learning*. PMLR, 2018, pp. 4693–4702.

[3] Y. Gao, W. Zheng, Z. Yang, T. Köhler, C. Fuegen, and Q. He, "Interactive Text-to-Speech system via joint style analysis," *Proc. Interspeech 2020*, pp. 4447–4451, 2020.

[4] Y. Gao, R. Singh, and B. Raj, "Voice impersonation using generative adversarial networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2506–2510.

[5] T. Kaneko and H. Kameoka, "CyclegGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2100–2104.

[6] Y. Gao, J. Lian, B. Raj, and R. Singh, "Detection and evaluation of human and machine generated speech in spoofing attacks on automatic speaker verification systems," *arXiv preprint arXiv:2011.03689*, 2020.

[7] 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," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] 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, 2017.

[9] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," 2017.

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

[11] A. Nautsch, X. Wang, N. Evans, T. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "Asvspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2021.

[12] H. Yu, Z.-H. Tan, Y. Zhang, Z. Ma, and J. Guo, "DNN filter bank cepstral coefficients for spoofing detection," *Ieee Access*, vol. 5, pp. 4779–4787, 2017.

[13] B. Balamurali, K. E. Lin, S. Lui, J.-M. Chen, and D. Herremans, "Toward robust audio spoofing detection: A detailed comparison of traditional and learned features," *IEEE Access*, vol. 7, pp. 84 229–84 241, 2019.

[14] M. R. Kamble, H. B. Sailor, H. A. Patil, and H. Li, "Advances in anti-spoofing: from the perspective of ASVspoof challenges," *APSIPA Transactions on Signal and Information Processing*, vol. 9, 2020.

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

[16] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2*, 2014, pp. 2672–2680.

[17] N. Yu, L. S. Davis, and M. Fritz, "Attributing fake images to GANs: Learning and analyzing GAN fingerprints," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7556–7566.

[18] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *International Conference on Machine Learning*. PMLR, 2020, pp. 3247–3258.

[19] T. Vuong, Y. Xia, and R. M. Stern, "A Modulation-Domain loss for neural-network-based real-time speech enhancement," *arXiv preprint arXiv:2102.07330*, 2021.

[20] J. Pons, S. Pascual, G. Cengarle, and J. Serrà, "Upsampling artifacts in neural audio synthesis," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3005–3009.

[21] X. Tian, Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Spoofing detection from a feature representation perspective," in *2016 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 2119–2123.

[22] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

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

[24] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," *arXiv preprint arXiv:1907.00501*, 2019.

[25] R. Reimao and V. Tzerpos, "FoR: A Dataset for synthetic speech detection," in *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE, 2019, pp. 1–10.

[26] N. Subramani and D. Rao, "Learning efficient representations for fake speech detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 5859–5866.

[27] T. Vuong, Y. Xia, and R. Stern, "Learnable spectro-temporal receptive fields for robust voice type discrimination," *arXiv preprint arXiv:2010.09151*, 2020.

[28] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[29] M. R. Schädler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 131, no. 5, pp. 4134–4151, 2012.

[30] G. Suthokumar, V. Sethu, C. Wijenayake, and E. Ambikairajah, "Modulation dynamic features for the detection of replay attacks." in *Interspeech*, 2018, pp. 691–695.

[31] B. T. Meyer, S. V. Ravuri, M. R. Schädler, and N. Morgan, "Comparing different flavors of spectro-temporal features for ASR," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[32] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE transactions on speech and audio processing*, vol. 2, no. 4, pp. 578–589, 1994.

[33] H. Hermansky and S. Sharma, "Temporal patterns (TRAPS) in ASR of noisy speech," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, vol. 1. IEEE, 1999, pp. 289–292.

[34] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7234–7238.