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

The Effect of Silence and Dual-Band Fusion in Anti-Spoofing System

Yuxiang Zhang[12], *Wenchao Wang*[12], *Pengyuan Zhang*[12]

[1]Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, China
[2]University of Chinese Academy of Sciences, China

{zhangyuxiang, wangwenchao, zhangpengyuan}@hccl.ioa.ac.cn

Abstract

The current neural network based anti-spoofing systems have poor robustness. Their performance degrades further after voice activity detection (VAD) performed, making it difficult to be applied in practice. This work investigated the effect of silence at the beginning and end of speech, finding that silent differences are part of the basis for countermeasures' judgements. The reason for the performance deterioration caused by VAD is also explored. The experimental results demonstrate that the neural network loses the information about silent segments after the VAD operation removes them. This can lead to more serious overfitting. In order to solve the overfitting problem, the work in this paper also analyzes the reasons for system overfitting from different frequency sub-bands. It is found that the high-frequency part of the feature is the main cause of system overfitting, while the low-frequency part is more robust but less accurate against known attacks. Therefore, we propose the dual-band fusion anti-spoofing algorithm, which requires only two sub-systems but outperforms all but one primary system submitted to the logical access condition of the ASVspoof 2019 challenge. Our system has an EER of 3.50% even after VAD operations performed, thus can be put into practical application.
Index Terms: anti-spoofing, synthetic speech detection, sub-band countermeasures, system overfitting, speaker verification

1. Introduction

As one type of biometric technology, automatic speaker verification (ASV) [1] has made significant progress in recent years and has been widely used in financial, military and other fields. However, with the development of text-to-speech synthesis (TTS) [2], voice conversion (VC) [3] and the popularity of high-quality recording devices such as smart phones, spoofing attacks against ASV systems are gradually gaining the attention of researchers. In order to improve the performence of anti-spoofing systems, the ASVspoof challenges were successfully held in 2015 [4], 2017 [5] and 2019 [6], building datasets in two evaluation conditions: logical access (LA) considers spoofing attacks from TTS and VC, while physical access(PA) refers to attacks generated by replay [7]. A large number of systems with great capability emerged from these challenges, which promoted the development of ASV anti-spoofing research.

However, most works in the challenges had focused on performance improvements for the specific dataset, ignoring the basis on which the anti-spoofing systems make judgements. In addition, most countermeasures do not apply voice activity detection (VAD) without a detailed reason. But due to the wide application of VAD in real ASV systems, there is demand to try to introduce VAD into countermeasure.

In this work, the impact of silent segments that appear at the beginning and end of speech is explored. Furthermore, the reasons for the performance degradation due to VAD is explained. It is found that VAD exacerbates the overfitting of neural networks, so it is necessary to study the reasons for the overfitting of neural network and propose more robust algorithms.

On the one hand, since any single spoofing attacks may only be effectively detected in a certain sub-band frequency of features. There are many traditional contermeasurers that use Gaussian mixture models (GMM) as classifiers demonstrate the effectiveness of sub-band fusion [8, 9, 10, 11, 12]. On the other hand, a number of neural network architectures have been applied to anti-spoofing systems successfully, such as light convolutional neural network (LCNN) [13, 14], residual neural network (ResNet) [15, 16], siamese convolutional neural network [17] and so on. But discriminative deep neural networks are often plagued by overfitting because of the large number of parameters and depth, and few neural network based countermeasures utilize sub-band information to avoid overfitting. Thus, the performence of ResNet based anti-spoofing systems in different frequency sub-bands is investigated. In the experiments, the spectrogram is divided into two parts, high-frequency and low-frequency, as the features fed into the Squeeze-and-Excitation ResNet (SENet) [18]. It is found that overfitting is mainly caused by the high-frequency part of the features, while the low-frequency part can effectively avoid overfitting and obtain better performance on the evaluation partition. Therefore, the SENet based low-frequency anti-spoofing algorithm with an EER of 1.14% is proposed, and on this basis, the dual-band fusion anti-spoofing algorithm with small calculation amount and well robustness is proposed.

The contributions of this work include: (i) Exploring the impact of silence differences on neural network based anti-spoofing systems and summarizing the reasons for the negative impact of VAD; (ii) Investigating the effectiveness of sub-band spectrogram features in neural network based countermeasures; (iii) Developing the robust low-frequency anti-spoofing algorithm and outstanding dual-band fusion algorithm for logical access partition in ASVspoof 2019 database.

2. Silence and Dual-Band Fusion

In this section, the effect of silence on the performance of countermeasures is illustrated, and the dual-band fusion algorithm is proposed as well.

2.1. Silence difference

In the ASVspoof 2017 database, the silence differences are important artefacts that influence countermeasures [19]. The same deficiencies is also found in the ASVspoof 2019 logical access dataset.

The silence difference is defined as the difference between the silent segments that present at the beginning and end of the bonafide speech and the spoof class speech. As shown in Fig. 1: there are silent segments containing weak stochastic noise at the

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-1281

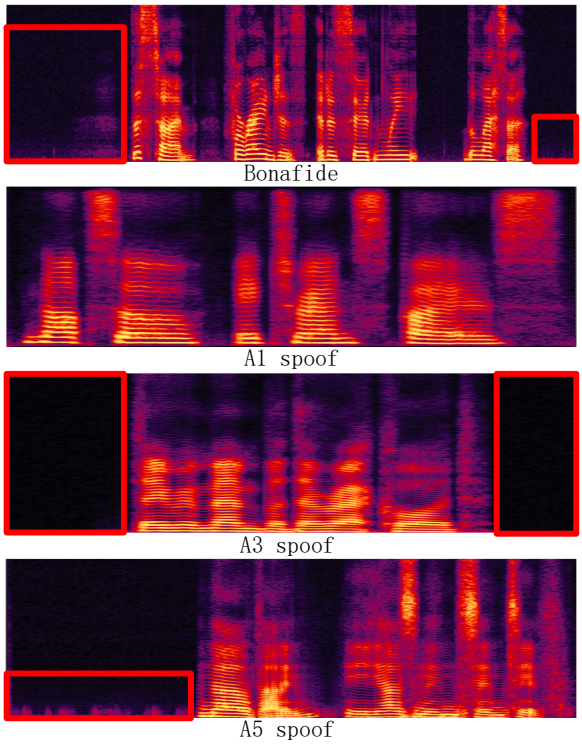

Figure 1: *Spectrogram of speech that have silence difference.*

beginning and end of the bonafide speech. However, the silent segments of the spoof speech differs from the bonafide speech: there is no silence, or the silent segments have no noise, or have abnormal noise. Therefore, these differences may be part of the foundation for the countermeasures' judgement.

In addition, performing VAD on speech usually leads to worse experimental results. Since the VAD operation removes all silent segments, the experimental results of cutting 100 ms of silent segments at the beginning and end of speech also help to analyze the effect of VAD on the anti-spoofing systems. On this basis, the experimental results generated after adopting the VAD operation were further analyzed.

Through the experiments that cutting the beginning and end of the data to remove the silence difference or performing VAD, it is possible to determine whether the countermeasures actually extract valid information, rather than making judgements based on silent segments. From this, the reliability of countermeasures and the possibility of practical application can be improved.

2.2. Dual-band fusion anti-spoofing algorithm

The anti-spoofing systems based on ResNet and its variants often suffer from the disadvantage of overfitting. It has been demonstrated that features in different sub-bands have different effects on different types of spoofing methods in traditional countermeasures. The impact of different subbands and their importance on replay spoofing detection has been investigated [20]. Yet, the performance of different sub-band features on neural network based countermeasures of LA has not been studied. So different sub-bands' information in the neural network is used to prevent overfitting. There is a clear gap between the speech synthesized by the neural-waveform TTS methods and the early vocoder methods, especially at high-frequency. Thus insufficient data in the training set may cause the neural network to focus too much on the high-frequency differences. As a result, countermeasures perform poorly against unknown attacks with good quality in high-frequency part. So the causes of overfitting from sub-band features are analyzed and the dual-band fusion anti-spoofing algorithm is proposed.

The features used in the dual-band fusion anti-spoofing system is the spectrogram that is splited into two parts: low-frequency part $(0-4\text{kHz})$ and high-frequency part $(4-8\text{kHz})$. The two parts of the spectrograms are fed into two SENets for training, and two anti-spoofing systems are obtained. The details of the features and models are described in Section 3. The fusion of the two sub-band systems is performed linearly at the score level with equal wights.

3. Experimental Setup

3.1. Dataset and metrics

Experiments were conducted under the ASVspoof 2019 LA database, in which spoofed speech is generated from different TTS and VC algorithms. Our models are trained with the training partition consists of 6 algorithms. The development partition containing the same algorithms as the training partition is used for model selection and determining whether the models are overfitted. The evaluation partition includes 2 known algorithms and 11 unseen algorithms which are different from those in the training and development partitions. The metrics used in this work is the equal error rate (EER), which defines an operating point where the false acceptance rate (FAR) and false rejection rate (FRR) of the system are equal, and the minimum normalized tandem detection cost function (min t-DCF) [21].

3.2. Front-End

The main acoustic feature used in our experiments is log power magnitude spectrogram. The Fast Fourier Transform (FFT) spectrogram is extracted with 1728 window length and 130 hop length while the Blackman window is used. Because that the voice length is different, the length of each spectrogram is truncated or concatenated into 600 frames. So the height of entire feature is 865 while the width is 600. Inspired by Kaldi [22] and Librosa [23], the spectrograms are flipped left and right before concatenating to prevent interference from discontinuities.

For the dual-band fusion system, since the spectrogram is cropped into two parts of the same size, high-frequency and low-frequency, the input feature of each sub-band system has a shape of 433×600.

The VAD operation used in the experiment is the VAD method based on the end-point detection.

3.3. Model architecture

In this work, LCNN and SENet are used as classifiers. The model architecture and training strategy of LCNN are as same as [14]. The SENet is integration of the ResNet with the squeeze-and-excition (SE) block [18]. The skip connections of the ResNet can deepen the depth of network without performance degradation. The SE block can adaptively acquire the importance of each feature channel, and explicitly model the interdependencies between feature channels by assigning weights to them. The features that are useful for the current task are boosted according to their importance, while the features that are less useful are suppressed. Due to these advantages, the SENet can effectively distinguish the bonafide speech from the

spoof speech. The detailed structure of SENet used in our experiments is shown in Table 1.

Table 1: *The architecture of SENet. The channels, strides, repeat times and number of parameters are specified.*

Stage	Block	Params
Conv1	conv2d, 16, 7×7, stride=2, padding=3 BatchNorm, ReLU maxpool, 3×3, stride=2, padding=1	0.8K
Conv2	[SENet Block, 16, stride=1]$\times3$	13.9K
Conv3	[SENet Block, 32, stride=2]$\times4$	57.9K
Conv4	[SENet Block, 64, stride=1]$\times6$	347.1K
Conv5	[SENet Block, 128, stride=2]$\times3$	694.2K
	Global Average Pooling, A-Softmax	0.2K
Total	-	1.1M

3.4. Training strategy

The angular margin based softmax loss (A-softmax) [24] is used as the loss function. Adam [25] is adopted as the optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$ and weigth decay 10^{-4}.

For SENet, the learning rate increases linearly for the first 1000 warm-up steps and then decreases proportionally to the inverse square root of the step number. The SENet models were trained with 32 epochs, while the LCNN models were trained with 16 epochs. The model with the lowest loss on development set is selected as the final model for evaluation.

The weights of all convolutional layers are initialized using Kaiming initialization [26].

4. Experimental Results

4.1. Silence difference and VAD results

Our experimental results in Table 2 suggest that silence differences are cues to anti-spoofing systems. Cutting the silent segments at the beginning and end of speech or performing VAD can cause degradation of system performance.

As expected, more test examples in both the development and evaluation partitions are mismatched by LCNN and SENet systems after cutting first and last 100 ms of speech. These results confirm the hypothesis that the silence differences are meaningful for neural network based countermeasures.

As for VAD, it is obvious that VAD operation barely affects the performance of countermeasures against known attacks, but aggravates the problem of overfitting. Combining with the conclusion that the silence differences are additional cues for countermeasures, it is believed that by using VAD to excise all silent segments, the anti-spoofing systems can focus more on the differences in the speech itself rather than the silent segments.

It also shows that the neural network based anti-spoofing systems have sufficient ability to distinguish bonafide speech from spoof speech with known methods by the characteristics of the speech itself. However, because that the information can only be extracted from the speech, the anti-spoofing systems pay more attention to the information that can easily discern spoof speech. Therefore, the unknown spoof methods can not be adequately recognized, which leads to more severe overfitting. As VAD is a common operation in practical applications, and the experimental evidence that VAD does not degrade the systems' performance facing known attacks has been found. VAD can be performed in practical applications of the anti-spoofing system as long as the overfitting problem is solved.

Table 2: *Results of silence difference. "VAD" means performing VAD, "cut" means cutting first and last 100ms of speech. "FN" and "FP" denote false negative and false positive counts.*

Model	Set	FN	FP	FRR/%	FAR/%
LCNN	dev	0	1	0	0
	eval	269	2344	3.66	3.67
LCNN-cut	dev	23	201	0.90	0.90
	eval	284	2469	3.86	3.87
LCNN-VAD	dev	0	0	0	0
	eval	939	8138	12.77	12.74
SENet	dev	0	0	0	0
	eval	747	6490	10.15	10.16
SENet-cut	dev	0	2	0	0
	eval	909	7897	12.36	12.36
SENet-VAD	dev	0	0	0	0
	eval	1571	13613	21.36	21.31

4.2. Sub-band results and fusion results

The results shown in Table 3 demonstrate the EER of the FFT-SENet anti-spoofing system before and after VAD operation. The first column illustrates the bandwidth of each system whether VAD is performed.

Through observing the results before and after VAD, it can be found that before VAD operation, the spoofing method most difficult to be detected is A17, a VC algorithm. However, after VAD, it is harder to detect three TTS methods A10, A11 and A15 with features of all three frequency bands. Meanwhile high-frequency features are still difficult to detect A17. Through observing the spectrogram and listening, it can be found that the speech synthesized by these TTS methods has higher quality, but there is no silence at the beginning and end of speech. However, the spoof speech produced by A17 is based on the real voice, so the silent segments in the speech are similar to the real speech. And because A17 is a VC method based on waveform filtering, the speech generated by this method is not smooth enough, resulting in discontinuous fundamental frequency. These results are consistent with the conclusions in the previous section, indicating that the anti-spoofing systems trained with the data after VAD has research significance .

For single system, the results of the development set (A01-A06) indicate that in the face of known attacks, the full-band FFT features perform best with the SENet model in all metrics. While the performence of high-frequency features are slightly degraded but also excellent. And the low-frequency features have the worst performance. However, the results of the evaluation set (A07-A19) show that the robustness of the different sub-band features are almost opposite to the accuracy of the development set: low-frequency features have more stable and great performance for all unknown attack methods in the evaluation partition, while full-band features and high-frequency features are overfitting. The single system with the best performance on the evaluation partition is the low-frequency system before VAD, with a min t-DCF of 0.04 and an EER of 1.14%.

Table 3: *Results in terms of EER/% for development (A01-A06) and evaluation (A07-A19) subsets and respective pooled min t-DCF(P1) and pooled EER/% (P2). The EER of each spoof method is calculated separately.*

Freq-bands	A01	A02	A03	A04	A05	A06	P1	P2	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	P1	P2
FFT 0-8k	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.04	0.00	0.53	0.02	2.09	1.70	0.37	0.38	1.33	9.24	0.00	32.5	35.8	0.00	0.21	10.2
FFT 0-4k	0.24	0.51	0.24	0.48	0.54	1.02	0.02	0.55	0.19	1.77	0.06	0.37	0.53	0.08	0.16	0.08	0.27	0.35	2.50	1.24	1.48	**0.04**	**1.14**
FFT 4-8k	0.00	0.03	0.03	0.03	0.19	0.00	0.00	0.06	0.00	0.59	0.63	5.64	36.1	16.2	22.6	2.58	9.22	0.04	35.5	4.39	0.12	0.36	14.0
FFT fusion	0.00	0.00	0.00	0.00	0.00	0.00	**0.00**	**0.00**	0.00	0.04	0.00	0.45	2.03	0.38	0.81	0.20	0.42	0.00	4.34	0.62	0.00	**0.05**	**1.56**
VAD 0-8k	0.00	0.00	0.00	0.00	0.00	0.00	**0.00**	**0.00**	0.00	0.38	0.27	48.2	62.5	19.9	14.9	0.44	50.1	0.04	9.34	10.1	0.04	0.39	21.4
VAD 0-4k	1.23	0.32	0.19	2.64	1.26	2.92	0.05	1.85	3.96	3.50	0.32	4.65	5.62	2.38	0.47	1.22	5.65	3.76	8.64	9.92	5.99	0.17	5.22
VAD 4-8k	0.00	0.08	0.13	0.08	0.32	0.03	0.00	0.20	0.00	0.34	0.09	10.4	41.5	18.2	39.3	1.51	11.6	0.22	38.6	10.8	0.14	0.36	17.5
VAD fusion	0.19	0.00	0.03	0.82	0.27	0.90	0.01	0.50	0.04	0.06	0.00	1.78	6.97	2.88	1.16	0.12	2.61	0.14	8.21	3.28	0.14	0.11	3.50

In summary, high-frequency features offer better performence against known spoofing attack methods, but lead to severe overfitting. Low-frequency features can effectively detect most spoofing methods, while the accuracy in the face of known attacks is slightly lower. Therefore, fusion of the two sub-band systems will improve the performance of countermeasure. The results obtained by the fusion system before VAD on the development partition are all correct. But the performance of the system fusion could not be fully reflected on the evaluation set because the system before VAD makes judgements based on the silence. Its performance decreases slightly compared to the low frequency system. And after performing VAD, the performance of the fusion system improves because effective information is extracted from the speech. The EER of the fusion system is 3.50% and the min t-DCF is 0.11, which are reduced by 33% and 35% respectively compared to the low-frequency system.

Table 4: *Performance for the ASVspoof 2019 evaluation partitions in terms of pooled min t-DCF and pooled EER for top-performing single systems and primary systems.*

(a) *Single systems*

System	t-DCF	EER%
FFT-SENet [15]	0.2160	11.75
MFCC-ResNet [27]	0.2042	9.33
CQCC-ResNet [27]	0.2166	7.69
LFCC-LCNN [14]	0.1000	5.06
FFT-LCNN [14]	0.1028	4.53
LFCC-Siamese CNN [17]	0.093	3.79
FFT-LCGRNN [28]	0.0776	3.03
FFT-L-SENet-VAD	0.1679	5.22
FFT-L-SENet	**0.0368**	**1.14**

(b) *Primary systems*

System	t-DCF	EER%
T05 [6]	0.0069	0.22
T45 [14]	0.0510	1.84
T60 [29]	0.0755	2.64
GMM fusion [12]	0.0740	2.92
T24 [6]	0.0953	3.45
T50 [30]	0.1671	3.56
FFT-VAD SENet dual-band fusion	0.1056	3.50
FFT SENet dual-band fusion	**0.0498**	**1.56**

4.3. Performence comparison and discussion

Table 4 shows the results of seven top-performing single systems, six primary systems and our best system before and after performing VAD on the ASVspoof 2019 LA evaluation set.

The results of single systems are shown in Table 4a. All systems are represented by a name consisting of input features and system architectures. Our FFT Low-frequency SENet system before VAD achieves the best result in both metrics. The min t-DCF is reduced to less than 0.05, which is a great improvement compared to other single systems. In addition, our system is more robust and maintains a good performance even VAD are performed. It means that the possibility of applying our system in real-world scenarios is greatly increased.

The results of primary systems are shown in Table 4b, in which T05, T45, T60, T24 and T50 represent the anonymous identifiers of the teams in the ASVspoof 2019 challenge. These primary systems are based on the ensembles of at least four subsystems, which may contain different front-end features and neural network architectures. The GMM fusion system consists of a nonlinear fusion of six sub-band GMM based systems. However, benefiting from the excellent single-system performance, our dual-band fusion system outperforms all primary systems but the T05 by simply fusing the two sub-band systems. Moreover, even after VAD, the min t-DCF of our dual-band fusion system still reaches about 0.1.

5. Conclusions

In this work, the effect of silence and dual-band fusion on neural network based anti-spoofing systems is explored. The silence differences in speech can help countermeasures distinguish the spoof speech. So the overfitting of the anti-spoofing system becomes more serious after the VAD operation removes all silent segments. Analyzing the causes of overfitting from the perspective of frequency bands, we found that the high-frequency part of the features is the main reason for the overfitting of the neural network based countermeasures. Therefore, using the low-frequency spectrogram as the feature, we obtained the single system with min t-DCF of 0.04 and an EER of 1.14% on the ASVspoof 2019 LA database. Furthermore, our dual-band fusion system can accurately identify known attacks and remain robust to unknown attacks by linearly fusing two systems that separately trained with two sub-band features. Even after VAD, our dual-band fusion system still maintains good robustness and credibility, thus can be applied to practical applications.

6. References

[1] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

[2] V. Shchemelinin and K. Simonchik, "Examining vulnerability of voice verification systems to spoofing attacks by means of a tts system," in *International Conference on Speech and Computer*. Springer, 2013, pp. 132–137.

[3] T. Kinnunen, Z.-Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4401–4404.

[4] 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*, Dresden, Germany, September 2015, pp. 2037–2041.

[5] 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," in *INTERSPEECH 2017 Annual Conference of the International Speech Communication Association*, 2017, pp. 2–6.

[6] 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 - 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, September 2019.

[7] 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.

[8] K. Sriskandaraja, V. Sethu, P. N. Le, and E. Ambikairajah, "Investigation of sub-band discriminative information between spoofed and genuine speech." in *Interspeech*, 2016, pp. 1710–1714.

[9] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Galka, "Audio replay attack detection using high-frequency features." in *Interspeech*, 2017, pp. 27–31.

[10] J. Yang, R. K. Das, and H. Li, "Significance of subband features for synthetic speech detection," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2160–2170, 2019.

[11] S. Garg, S. Bhilare, and V. Kanhangad, "Subband analysis for performance improvement of replay attack detection in speaker verification systems," in *2019 IEEE 5th International Conference on Identity, Security, and Behavior Analysis (ISBA)*. IEEE, 2019, pp. 1–7.

[12] H. Tak, J. Patino, A. Nautsch, N. Evans, and M. Todisco, "Spoofing attack detection using the non-linear fusion of sub-band classifiers," in *Interspeech*, 2020, pp. 1106–1110.

[13] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks." in *Interspeech*, 2017, pp. 82–86.

[14] G. Lavrentyeva, A. Tseren, M. Volkova, A. Gorlanov, A. Kozlov, and S. Novoselov, "Stc antispoofing systems for the asvspoof2019 challenge," in *Interspeech*, 2019, pp. 1033–1037.

[15] C.-I. Lai, N. Chen, J. Villalba, and N. Dehak, "Assert: Anti-spoofing with squeeze-excitation and residual networks," in *Interspeech*, 2019, pp. 1013–1017.

[16] W. Cai, H. Wu, D. Cai, and M. Li, "The dku replay detection system for the asvspoof 2019 challenge: On data augmentation, feature representation, classification, and fusion," in *Interspeech*, 2019, pp. 1023–1027.

[17] Z. Lei, Y. Yang, C. Liu, and J. Ye, "Siamese convolutional neural network using gaussian probability feature for spoofing speech detection," in *Interspeech*, 2020, pp. 1116–1120.

[18] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[19] B. Chettri, E. Benetos, and B. L. Sturm, "Dataset artefacts in anti-spoofing systems: a case study on the asvspoof 2017 benchmark," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 3018–3028, 2020.

[20] B. Chettri, T. Kinnunen, and E. Benetos, "Subband Modeling for Spoofing Detection in Automatic Speaker Verification," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 341–348. [Online]. Available: http://dx.doi.org/10.21437/Odyssey.2020-48

[21] T. Kinnunen, H. Delgado, N. Evans, K. A. Lee, V. Vestman, A. Nautsch, M. Todisco, X. Wang, M. Sahidullah, J. Yamagishi *et al.*, "Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2195–2210, 2020.

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[23] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8. Citeseer, 2015, pp. 18–25.

[24] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[27] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," in *Proc. Interspeech 2019*, 2019, pp. 1078–1082.

[28] A. Gomez-Alanis, J. A. Gonzalez-Lopez, and A. M. Peinado, "A kernel density estimation based loss function and its application to asv-spoofing detection," *IEEE Access*, vol. 8, pp. 108 530–108 543, 2020.

[29] B. Chettri, D. Stoller, V. Morfi, M. Ramírez, E. Benetos, and B. Sturm, "Ensemble models for spoofing detection in automatic speaker verification," in *20th Annual Conference of the International Speech Communication Association: Crossroads of Speech and Language, INTERSPEECH 2019; Graz; Austria; 15 September 2019 through 19 September 2019*. International Speech Communication Association, 2019, pp. 1018–1022.

[30] 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, pp. 1038–1042.