Ghost-in-Wave: How Speaker-Irrelative Features Interfere DeepFake Voice Detectors

1st Xuan Hai
School of Information Science and Engineering
Lanzhou University
Lanzhou, China
haix21@lzu.edu.cn

2nd Xin Liu
School of Information Science and Engineering
Lanzhou University
Lanzhou, China
bird@lzu.edu.cn

3rd Zhaorun Chen
Elmore Family School of Electrical and Computer Engineering
Purdue University
IN, USA
chen4399@purdue.edu

4th Yuan Tan
School of Information Science and Engineering
Lanzhou University
Lanzhou, China
tany2023@lzu.edu.cn

5th Song Li
The State Key Laboratory of Blockchain and Data Security
Zhejiang University
Hangzhou, China
songl@zju.edu.cn

6th Weina Niu
School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu, China
vinusniu@uestc.edu.cn

7th Gang Liu
School of Information Science and Engineering
Lanzhou University
Lanzhou, China
andyliu@lzu.edu.cn

8th Rui Zhou
School of Information Science and Engineering
Lanzhou University
Lanzhou, China
zr@lzu.edu.cn

9th Qingguo Zhou
School of Information Science and Engineering
Lanzhou University
Lanzhou, China
zhouqg@lzu.edu.cn

***Abstract*—Recent speech synthesis technology can generate high-quality speech indistinguishable from human speech, thus introducing various security and privacy risks. Numerous recent studies have focused on fake voice detection to address these risks, with many claiming to achieve ideal performance. However, is this really the case? A recent research work introduced Speaker-Irrelative-Features (SiFs), unrelated to the information in speech files but capable of influencing fake detectors. This means that existing detectors may rely on SiFs to a certain extent to distinguish real and fake speech. In this paper, we introduce an evaluation framework to evaluate the influence of SiFs in existing fake voice detectors in depth. We evaluate three SiFs which include background noise, the mute parts before and after voice, and the sampling rate on ASVspoof2019 and FoR. Our results confirm the substantial influence of SiFs on fake voice detection performance, and we delve into the analysis of the underlying mechanisms.**

***Index Terms*—Deepfake, AI-Synthesized Speech, Fake Voice Detection**

I. Introduction

Speech serves as a predominant communication modality for information propagation within human society. Moreover, its utility has extended significantly into digital systems, encompassing data transmission, authentication processes, and various other applications. Speech synthesis is the technology that can generate speech for specific target sounds. Traditional speech synthesis methods use splicing and editing to generate speech and result in a distinctly unnatural sound that is easily detectable by the human ear. With the profound integration of deep learning technology, the performance of recent speech synthesis researches [1]–[3] have witnessed notable enhancements. These technologies have been widely used in various scenarios. However, these speech synthesis technologies can also generate high-quality fake voices and create a series of security and privacy risks such as telecom fraud, political smear campaigns, etc. A number of examples have been reported. As an illustration, a noteworthy incident highlighted in The Wall Street Journal involved cybercriminals employing AI to replicate the voice of a CEO in an unusual cybercrime case [4]. In this scenario, the malefactors exploited AI-based software with remarkable proficiency, successfully mimicking the CEO's voice. They utilized this deceptive technology to orchestrate a sophisticated phone scam, ultimately defrauding a substantial amount exceeding $243,000. Therefore, it is very significant to develop powerful fake voice detection technologies to mitigate security and privacy risks.

Over the past few years, fake voice detection has experienced rapid development and achieved significant progress. Most recent research works claim that their method is very effective and presents an ideal performance in their experiment. It seems that fake voice detection is no longer a challenging problem. However, a recent work [5] shows that these fake voice detectors are easily disturbed by a set of factors named

Xuan Hai and Xin Liu contributed equally to this work
Corresponding author: Xin Liu; Gang Liu.

Speaker-irrelative Features (SiFs) which are not related to speech. This suggests that these detectors partly depend on beyond the characteristics of the speech itself to distinguish between real and fake speech. Moreover, the performance of detectors may no longer be optimal if there are changes in the SiFs. It will seriously affect the usability of these detectors in the real world.

In this paper, we design an evaluation framework and conduct a series of experiments to further analyze the impact of SiFs on the existing detectors. We remove a part of SiFs in the dataset separately to create new datasets. And then, we retrain the target detectors with these new datasets to evaluate the performance. We select the two most widely used datasets ASVspoof2019 and FoR to evaluate the performance and discuss the specific effects of different SiFs on the detectors. The result of experiments shows that all of the SiFs selected in this paper have significant effects on the performance of the detectors.

The main contributions of this paper can be summarized as follows:

- We analyze the defects of existing fake voice detection methods and discuss the negative influence of SiFs.
- We design an evaluation framework to further evaluate the performance of existing fake voice detectors. The result indicates that all of the existing detectors are significantly influenced by SiFs and get poor performance after removing some kinds of SiFs in most instances.
- We discuss the mechanism of action of SiFs in the decision of the detectors.

II. Related Works

Existing research on fake voice detection has employed diverse methodologies. Most of them are deep learning based technology. Deep learning methodologies leverage sophisticated Deep Neural Network (DNN) models to extract nuanced high-level features, enabling the identification of subtle differences between fake and authentic vocalizations.

Alzantot et al. [6] introduced a detection system built upon a Residual Convolutional Neural Network, utilizing Mel-frequency cepstrum (MFCC), Constant Q-transform cepstral coefficients (CQCC), and short-time Fourier transform (STFT) as input features. Ballesteros et al. [7] proposed Deep4SNet, a computer vision-based method that employs histogram visualization of time-domain waveforms for the classification of voice conversion speech. Tak et al. [8] innovatively designed RawNet2, an end-to-end approach that processes raw waveforms as input, leveraging neural networks to extract frequency-domain features for detection. Wang et al. [9] presented DeepSonar, a system based on neural network features extracted from a DNN-based speaker recognition system, demonstrating optimal performance in their experiments. Monteiro et al. developed an end-to-end Lookup-based Convolutional Neural Network (LCNN) ensemble model for fake voice detection [10]. Tak et al. [11] using wav2vec 2.0 which is a pre-trained speech representation model to extract the implicit representation of speech and get an ideal performance.

III. Evaluation Method

A. Insight

Several recent researches on fake voice detection got ideal performance in their experiments which utilized deep learning models. Most of them are trained and evaluated on the ASVspoof2019 dataset. However, there are several studies [12]–[14] indicate that the obvious difference of mute parts before and after human voice between bonafide and spoof samples in ASVspoof2019 dataset may result in all detectors trained and evaluated using this dataset easily distinguishing spoof samples based on the feature. This situation can create an illusion for the designer that the model has achieved ideal performance. It means that the real performance of these methods may not be excellent as shown in their experiments. A recent study [5] shows that existing detectors are sensitive to a series of features named Speaker-Irrelative-Features (SiFs) which are not related to the information expressed by the speech files, such as mute parts before and after the speaker's voice, background noise of speech, etc. The work [15] proposed an adversarial attack method by modifying the background noise and mute parts before and after the speech to trick the fake voice detectors into making wrong decisions. This implies that these detection methods might not effectively capture the fundamental distinctions between real and fake voices but instead focus more on SiFs. To delve deeper into the influence of SiFs on fake voice detection, we have devised a series of evaluations.

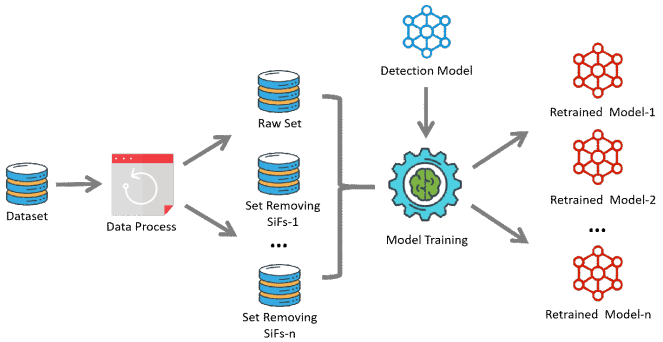


Fig. 1: Architecture of Evaluation Framework

B. Evaluation Framework

This evaluation of the paper has two primary goals: analyzing the impact of SiFs on existing methods and exploring the specific ways SiFs affect the existing detectors. To achieve these objectives, we design an evaluation framework shown in Figure 1 by processing the dataset and retraining existing detection detectors. Firstly, we preprocess the dataset to eliminate a portion of SiFs, thereby eliminating their influence. Then, we retrain the existing detectors by using the preprocessed dataset and evaluate their performance. If the detectors do not learn SiFs during the training process, retraining the

detector after removing SiFs is unlikely to cause significant performance fluctuations. Conversely, if noticeable changes occur in performance, it indicates that SiFs are crucial features learned by the detector to distinguish between fake and real speech.

C. Dataset Process

The dataset is the basis of deep learning model training. We select the background noise, mute parts before and after the speaker's voice, and the sampling rate of speech files as the evaluation features which are the most representative features of SiFs. To evaluate the impact of these features, we remove these features from the datasets separately using Sound eXchange (SoX) which is a cross-platform audio editing software to create new evaluation datasets. We get four datasets after the removal process: 1) **Raw Set:** This is the raw dataset without any processing. 2) **Denoised Set:** This is the dataset that has removed the background noise by using sox {file} -n noiseprof | sox {file} {new file} noisered - 0.01 and other information is consistent with the raw set. 3) **Silence Set:** This is the dataset that has removed the mute parts before and after the speaker's voice by using sox {file} {new file} silence 1 0.00001 1% -1 0.00001 1% lowpass 4000 and other information is consistent with the raw set. 4) **Msr Set:** This is the dataset that modifies the sampling rate from 16000 to 8000 by using sox {} -r 8000 {} and other information is consistent with the raw set.

D. Model Retraining

To analyze the performance difference, we retrain the target models in our evaluation. Every model is retrained with the raw dataset and three processed datasets. All models we select are collected from open-source repositories released by the authors and all of the settings are the same as the authors' version besides datasets. We retrain these models by using the raw dataset and three processed datasets with the same setting and get the three versions of these models: 1) **Raw Model:** This model is retrained with the training subset of the raw set. 2) **Denoised Model:** This model is retrained with the training subset of the denoised set. 3) **Silence Model:** This model is retrained with the training subset of the silence set. 4) **Msr Model:** This model is retrained with the training subset of the msr set.

IV. Evaluation Result

A. Evaluation Setup

1) Detectors: In the evaluation of this paper, we choose seven recent detectors published in top conferences or related challenges: AASIST [16], RawGAT-ST [17], RawNet2 [8], SAMO [18], MTLISSD [19], SSL [11] and FastAudio [20]. RawNet2 is the baseline for DeepFake in ASVspoof 2021 and other detectors get ideal performance in ASVspoof2019 datasets. We obtained the implementations of these detectors from open-source repositories made available by the authors.

2) Datasets: We select the two most commonly used datasets ASVspoof2019 LA subset [21] and Fake-or-Real (FoR) [22]. The detailed information of ASVspoof2019 is shown in Table I and FoR is a more simple dataset that contains 42260 bonafide samples and 42463 spoof samples generated by 7 synthetic-based algorithms. We process the ASVspoof2019 using the method in Section III and get three new datasets. The FoR has been processed to remove the mute parts before and after voice so we didn't do this again. The training and evaluation subsets of each dataset are processed identically.

TABLE I: Statistics of ASVspoof 2019 LA Dataset

Subset	Speech Samples		Speakers		Spoofing Algorithms		
	Logic Access		Total:107		VC	TTS	VC+TTS
	Bonafide	Spoof	Male	Female			
Training	2,580	22,800	8	12	2	4	0
Development	2,548	22,296	8	12			
Evaluation	7,335	63,882	30	37	5	11	3

B. Evaluation on ASVspoof2019

The performance of retrained models is presented in Table II and the difference in performance between the models trained with raw set and those trained with the processed set is shown in Table III. It is obvious that the performance of most of the target detectors shows significant fluctuations after removing specific SiFs and the change in the silence set is particularly obvious. All of the average performance differences shown in Table III are negative. **The performance fluctuations will be further amplified if the detectors trained on the raw set are employed to distinguish the samples in processed datasets.** The result may indicate that all of the SiFs we select in this paper are one of the bases for the detector to distinguish real and fake speech.

TABLE II: The result of the performance of retrained models on ASVspoof2019 eval subset. The evaluation indicator is Equal-Error-Rate (EER)

Models	Dataset			
	Raw Set	Denoised Set	Silence Set	Msr Set
AASIST	1.13%	2.50%	24.45%	1.37%
RawGAT-ST	1.39%	1.39%	22.50%	1.51%
RawNet2	5.49%	5.97%	23.64%	5.91%
SAMO	1.10%	1.99%	18.34%	1.40%
MTLISSD	2.58%	6.47%	23.43%	13.66%
SSL	0.22%	0.46%	7.97%	0.32%
FastAudio	1.78%	2.30%	19.70%	3.41%

To further analyze the influence of the mute parts, we also extract the data of performance about every spoof algorithm which is shown in Table IV. The result of the silence set shows an obvious pattern that the Equal-Error-Rate (EER) of raw models in the synthetic-based algorithms (A07-A16) is better than the one in the voice conversion-based algorithms(A17-A19) in the raw set but opposite in the silence set. We compared the difference in duration between samples in the raw set and silence set shown in Figure 2. The result shows that the duration of the mute parts before and after the speaker's voice of speeches generated by synthetic-based algorithms

is significantly shorter than speeches generated by voice conversion-based algorithms and real speeches in all three subsets. It means that the mute parts are one of the important bases for the detectors to make judgments. Results on other data sets also show that various types of SiFs have a significant impact on detector performance but their impact on different algorithms is not obvious.

TABLE III: The performance difference between raw model and other models trained with processed dataset

Models	Dataset		
	Denoised Set	Silence Set	Msr Set
AASIST	-121.24%	-2063.72%	-21.24%
RawGAT-ST	0.00%	-1518.71%	-8.63%
RawNet2	-8.74%	-330.60%	-7.65%
SAMO	-80.91%	-1567.27%	-27.27%
MTLISSD	-150.78%	-808.14%	-429.46%
SSL	-109.09%	-3522.73%	-45.45%
FastAudio	-29.21%	-1006.74%	-91.57%
Average	**-71.40%**	**-1545.00%**	**-90.18%**

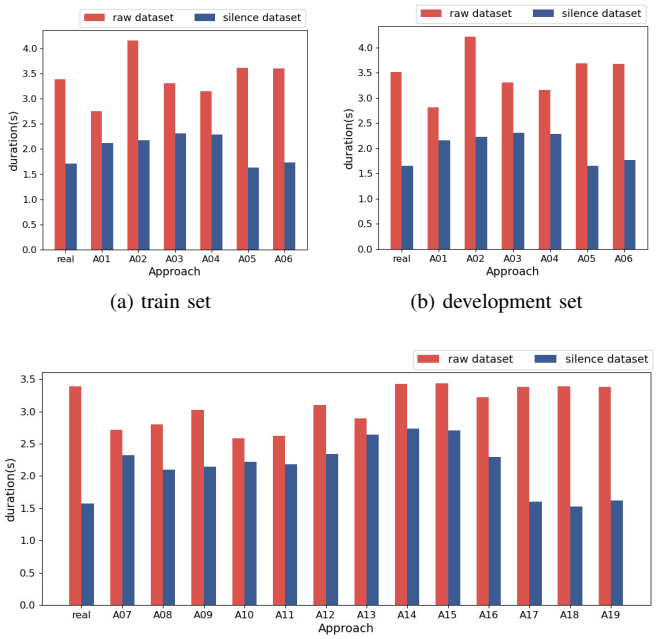
C. Evaluation on FoR

All of the target detectors we select are designed based on ASVspoof2019. The above experiments indicate that SiFs in ASVspoof2019 are an important factor affecting the detectors' judgment. In this evaluation, we use the FoR which is the dataset that was not considered during the detector design stage to evaluate the influence of SiFs across datasets to gain a comprehensive understanding of their impact on fake voice detection. The SAMO requires the speaker's information which is not given by FoR, so we can't evaluate the performance of it in this dataset.

The performance of these detectors on FoR is shown in Table V. While the performance of most detectors on the raw set appears to be acceptable, it falls short of the levels achieved on ASVspoof2019. Given that the spoof algorithms in the evaluation set of ASVspoof2019 differ from those in the training set, while the algorithms in FoR remain the same, the obtained result is still notably distant from the ideal scenario. There is another interesting phenomenon that SiFs still have a significant impact on detector performance but the impact is positive in half of the results. We speculate that the difference may be due to the mute parts. The denoise and modifying sampling rate process will impact the mute parts of the samples in ASVspoof2019 but the samples in FoR have no mute parts. This may mean that multiple SiFs can interact with each other and amplify the effects.

D. Analysis of the influence mechanism of SiFs

The above evaluations prove that SiFs have a significant impact on the performance of fake voice detectors. In this subsection, we try to visualize the characteristics of different type of samples from four datasets in ASVspoof2019 evaluation to analyze the influence mechanism of SiFs. We select AASIST as the target and extract the output of sinc convolution layer after selu activation function during model inference. The

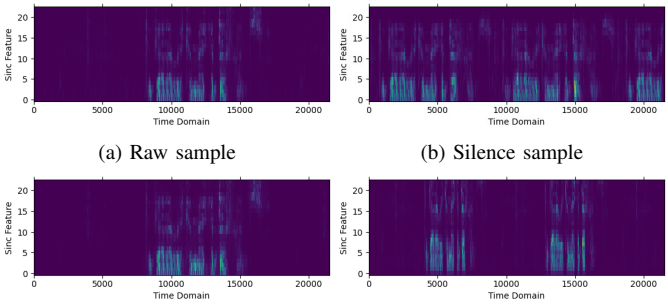


(a) train set (b) development set

(c) evaluation set

Fig. 2: Average duration of raw set and silence set. The difference represents the duration of mute parts

sample LA_E_4633286 from the ASVspoof2019 evaluation subset is visualized in Figure 3.



(a) Raw sample (b) Silence sample

(c) Denoised sample (d) Msr sample

Fig. 3: Visualized feature of LA_E_4633286. The X-axis represents the time domain and the Y-axis represents the value of sinc convolution.

The sinc convolution can be viewed as a set of bandpass filters and the output of it represents the extraction results of frequency-domain features by AASIST. The removing silence process changes the duration of the sample, causing the speaker's voice to appear multiple times in the figure because of the padding process of AASIST. The denoised mainly impacts the characteristic information of mute parts before and after the voice and makes some samples smoother in hearing. The modifying sampling rate process makes low-frequency information more prominent in the feature map and indirectly changes the time domain length in the feature map because the number of sampled data points becomes half of the original. The impact of this feature on detection performance may vary

TABLE IV: All spoof algorithms performance of the retrained models on ASVspoof2019. The evaluation indicator is EER (%)

Model	Dataset	Spoof Algorithm													Average
		A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	**EER**
AASIST	Raw	0.80	0.44	0.00	1.16	0.31	0.91	0.10	0.14	0.65	0.72	1.52	3.40	0.62	1.13
	Denoised	0.33	1.26	0.04	0.46	0.26	0.47	0.18	0.16	0.34	1.38	2.88	9.23	1.49	2.50
	Silence	39.23	3.29	1.10	50.40	44.58	40.74	7.69	15.59	33.27	4.31	2.19	5.92	1.08	24.45
	Msr	0.24	1.53	0.11	0.35	0.33	0.29	0.33	0.20	0.22	0.55	1.10	3.77	0.53	1.37
RawGAT-ST	Raw	1.19	0.44	0.00	1.06	0.31	0.91	0.10	0.14	0.65	0.72	1.52	3.40	0.62	1.39
	Denoised	0.67	1.28	0.24	0.83	0.50	1.14	0.43	0.20	0.59	1.12	2.62	6.90	0.99	2.06
	Silence	38.69	3.03	1.93	50.40	35.00	39.49	5.79	9.56	32.94	3.78	1.52	4.82	0.89	22.50
	Msr	0.63	2.12	0.42	0.67	0.59	0.45	1.26	0.20	0.30	0.61	1.30	3.82	0.42	1.51
RawNet2	Raw	2.91	4.62	0.15	2.93	1.35	2.93	0.23	0.92	2.67	1.30	11.37	16.92	1.85	5.49
	Denoised	1.59	4.60	0.27	2.05	1.25	2.95	0.39	1.11	2.44	1.51	9.93	24.34	2.81	5.97
	Silence	40.03	17.17	2.91	43.96	35.98	35.80	9.08	17.36	27.48	7.65	9.12	16.30	2.42	23.64
	Msr	0.69	3.11	0.20	1.16	0.61	1.14	1.01	0.41	0.59	1.34	11.64	20.65	4.31	5.91
SAMO	Raw	0.55	1.90	0.54	0.63	0.54	0.59	0.53	0.55	0.63	0.84	1.26	3.85	0.93	1.10
	Denoised	1.49	1.99	1.47	1.79	1.51	1.73	1.47	1.47	1.60	1.92	2.81	5.82	1.99	1.99
	Silence	25.99	2.57	2.68	50.51	28.16	43.06	5.86	5.68	15.12	4.34	2.68	4.60	2.83	18.34
	Msr	0.78	3.39	0.77	0.79	0.77	0.78	0.77	0.77	0.80	0.80	1.18	4.86	0.95	1.40
MTLISSD	Raw	0.19	0.14	0.04	1.53	0.1	0.45	0.65	1.65	2.16	0.26	13.12	0.86	1.43	2.58
	Denoised	0.14	1.24	0.19	0.41	0.14	1.26	0.12	0.11	0.11	0.65	5.45	41.11	5.96	6.47
	Silence	36.67	5.48	5.76	48.62	18.8	36.63	18.97	14.82	19.02	24.01	6.88	32.42	9.97	23.43
	Msr	0.12	1.44	0.12	0.33	0.38	0.12	0.12	0.34	0.38	0.47	46.12	43.26	25.29	13.66
SSL	Raw	0.02	0.18	0.00	0.26	0.15	0.11	0.00	0.02	0.06	0.06	0.33	0.55	0.33	0.22
	Denoised	0.10	0.31	0.04	0.26	0.18	0.15	0.04	0.06	0.10	0.12	0.33	1.81	0.43	0.46
	Silence	2.00	0.18	0.02	28.08	19.97	6.61	0.10	0.41	2.14	0.45	0.04	0.77	0.29	7.97
	Msr	0.02	0.16	0.00	0.23	0.20	0.06	0.00	0.02	0.06	0.06	0.45	0.77	0.43	0.32
FastAudio	Raw Set	0.23	1.12	0.02	0.43	0.34	0.23	0.10	0.19	0.19	0.16	1.04	7.55	0.23	1.78
	Denoised	0.10	0.81	0.08	0.30	0.26	0.37	0.12	0.12	0.18	0.12	0.75	9.24	0.19	2.30
	Silence	17.28	3.36	1.06	46.66	44.50	26.92	2.30	8.47	25.62	4.13	3.14	20.29	1.00	19.70
	Msr	0.23	0.81	0.06	0.38	0.42	0.24	0.12	0.24	0.24	0.12	3.15	11.97	0.53	3.41

TABLE V: The result of the performance of retrained models on FoR. Diff represents the performance difference between the raw model and other models trained with processed dataset

Models	Raw	Csr		Denoised	
	EER	EER	Diff	EER	Diff
AASIST	5.65%	3.63%	35.75%	5.39%	4.60%
RawGAT-ST	3.79%	3.37%	11.08%	11.57%	-205.28%
RawNet2	9.95%	5.65%	43.22%	15.36%	-54.37%
MTLISSD	26.50%	43.16%	-62.87%	14.26%	46.19%
SSL	1.81%	1.40%	22.65%	0.58%	67.96%
FastAudio	0.00%	1.31%	-	9.80%	-

if the detector reads audio files at a fixed sample rate. It means that the method of file reading may also influence the detection results and also shows the diversity of the way SiFs influence the result of detectors.

V. Discussion

The result of evaluations in this paper proves that the SiFs have a significant impact on the detection performance of fake voice detectors and SiFs serve as one of the foundations for the detector to differentiate between fake and real speech. The poor performance of the detectors in the evaluations means that the actual performance of existing detectors is not as good as claimed. We believe that the existing perspective of design has not captured the essential difference between the real and fake voice. In this paper, we just evaluate the influence of a part of SiFs. It is highly probable that there are additional SiFs that will similarly exert a substantial influence on detector performance. The detectors are vulnerable to attacks if they rely on SiFs to distinguish between real and fake speech because modifying specific SiFs is easy to do. It is hard to remove all of the SiFs from speech files. Therefore, we need to explore a new perspective on design to build practical detectors for real-world environments.

VI. Conclusion

In this paper, we first analyze the negative influences of SiFs on the existing fake voice detection methods. Then, we propose an evaluation framework to evaluate the influence of background noise, mute parts before and after the speaker's voice, and sampling rate of speech files which are representative SiFs. The result of the evaluation shows that existing detectors are greatly influenced by SiFs and we also analyze the influence mechanism of these features. Finally, we discuss the direction of high-quality detector design.

Acknowledgment

This work is supported by HY-Project under No.4E49EFF3 and Supercomputing Center of Lanzhou University.

References

[1] R. Valle, K. Shih, R. Prenger, and B. Catanzaro, "Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis," *arXiv preprint arXiv:2005.05957*, 2020.

[2] M. W. Lam, J. Wang, D. Su, and D. Yu, "Bddm: Bilateral denoising diffusion models for fast and high-quality speech synthesis," *arXiv preprint arXiv:2203.13508*, 2022.

[3] H. Siuzdak, P. Dura, P. van Rijn, and N. Jacoby, "Wavthruvec: Latent speech representation as intermediate features for neural speech synthesis," *arXiv preprint arXiv:2203.16930*, 2022.

[4] C. Stupp, "Fraudsters used ai to mimic ceo's voice in unusual cybercrime case," *The Wall Street Journal*, 2019. [Online]. Available: https://www.wsj.com/articles/fraudsters-use-ai-to-mimicceos-voice-in-unusual-cybercrime-case-11567157402

[5] X. Liu, Y. Tan, X. Hai, Q. Yu, and Q. Zhou, "Hidden-in-wave: A novel idea to camouflage ai-synthesized voices based on speaker-irrelative features," in *2023 IEEE 34th International Symposium on Software Reliability Engineering (ISSRE)*. IEEE, 2023, pp. 786–794.

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

[7] D. M. Ballesteros, Y. Rodriguez-Ortega, D. Renza, and G. Arce, "Deep4snet: deep learning for fake speech classification," *Expert Systems with Applications*, vol. 184, p. 115465, 2021.

[8] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6369–6373.

[9] R. Wang, F. Juefei-Xu, Y. Huang, Q. Guo, X. Xie, L. Ma, and Y. Liu, "Deepsonar: Towards effective and robust detection of ai-synthesized fake voices," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1207–1216.

[10] J. Monteiro, J. Alam, and T. H. Falk, "Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers," *Computer Speech & Language*, vol. 63, p. 101096, 2020.

[11] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," *arXiv preprint arXiv:2202.12233*, 2022.

[12] Y. Zhang12, W. Wang12, and P. Zhang12, "The effect of silence and dual-band fusion in anti-spoofing system," in *Proc. Interspeech*, 2021.

[13] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch *et al.*, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[14] Y. Zhang, Z. Li, J. Lu, H. Hua, W. Wang, and P. Zhang, "The impact of silence on speech anti-spoofing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[15] X. Hai, X. Liu, Y. Tan, and Q. Zhou, "Sifdetectcracker: An adversarial attack against fake voice detection based on speaker-irrelative features," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 8552–8560.

[16] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6367–6371.

[17] H. Tak, J.-w. Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," *arXiv preprint arXiv:2107.12710*, 2021.

[18] S. Ding, Y. Zhang, and Z. Duan, "Samo: Speaker attractor multi-center one-class learning for voice anti-spoofing," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[19] Y. Mo and S. Wang, "Multi-task learning improves synthetic speech detection," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 6392–6396.

[20] Q. Fu, Z. Teng, J. White, M. Powell, and D. C. Schmidt, "Fastaudio: A learnable audio front-end for spoof speech detection," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

[21] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[22] A. P. T. L. at York, "Datasets," https://bil.eecs.yorku.ca/datasets/, accessed September 11, 2023.